О.Г. Додонов, Д.В. Ланде,
В.В. Прищепа

КОМП'ЮТЕРНА КОНКУРЕНТНА РОЗВІДКА

Монографія

Друге видання

Київ – 2026

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ

О.Г. Додонов, Д.В. Ланде,

В.В. Прищепа

КОМП'ЮТЕРНА КОНКУРЕНТНА РОЗВІДКА

Монографія

Друге видання

Київ – 2026

УДК 004.5
ББК 22.18, 32.81, 60.54
С95

О.Г. Додонов, Д.В. Ланде, В.В. Прищепа
Комп'ютерна конкурентна розвідка : Монографія. Друге видання – Київ: ТОВ «Інжиніринг», 2026. – 384 с.

Книгу присвячено розгляду питань комп'ютерної конкурентної розвідки та розвідки у відкритих ресурсах мережі Інтернет. Комп'ютерна конкурентна розвідка охоплює автоматизовані процедури збору та аналітичної обробки інформації, які проводяться з метою підтримки прийняття управлінських рішень, підвищення конкурентоспроможності виключно з відкритих джерел у комп'ютерних мережах – вебсайтів, блогосфери, соціальних мереж, месенджерів, баз даних. У книзі розглядаються різні аспекти інформаційно-аналітичної діяльності у мережевому середовищі. Як теоретичні основи комп'ютерної конкурентної розвідки розглядаються елементи теорії інформації, аналізу соціальних мереж, інформаційного та математичного моделювання. Друге видання включає додаткові розділи, присвячені використанню штучного інтелекту, мережевого та геопросторовому аналізу.

Для широкого кола фахівців у сфері інформаційних технологій та безпеки.

Рекомендовано до друку Вченою радою Інституту проблем реєстрації інформації НАН України
(протокол № 3 від 3 лютого 2026 року)

Рецензенти:
Член-кор. НАН України, д.т.н., професор *В.В.Мохор*
Д.ю.н., професор *К.І. Бєляков*

© О.Г. Додонов, Д.В. Ланде,
В.В. Прищепа
2026

Зміст

Вступ

Комп'ютерна конкурентна розвідка (Computer Competitive Intelligence) охоплює процедури збирання та оброблення інформації, що проводяться з метою підтримки ухвалення управлінських рішень, підвищення конкурентоспроможності організацій виключно з відкритих джерел у комп'ютерних мережах, більшість із яких є надбудованими над мережею Інтернет – так званими оверлейними. Тому часто як синонім конкурентної розвідки вживають термін інтернет-розвідка. Таким чином, ця книга фактично присвячена проблематиці конкурентної розвідки, але з одним істотним обмеженням: усі джерела інформації, необхідні для проведення розвідувальної діяльності, є відкритими та доступними в комп'ютерних мережах. Більше того, значна частина інструментарію, програмного забезпечення для оброблення інформації також вільно доступна через сучасні комп'ютерні мережі. В англомовній літературі такий вид розвідки прийнято називати розвідкою за відкритими джерелами (Open Sources INTelligence, OSINT)[1], що також можна вважати синонімом терміна «конкурентна розвідка». Однак слід зазначити, що в зарубіжній літературі вживання OSINT значною мірою обмежене застосуванням у державній сфері. Але саме для технологій OSINT створено найбільшу кількість методик, технік і технологій.

Розвідувальна інформація може бути отримана з офіційних джерел, інших відкритих джерел, засобів масової інформації, оголошень, реклами, внутрішньофірмових, банківських, урядових звітів, баз даних, від експертів – шляхом добування (збирання), аналізу або спеціальної обробки даних, текстів[2]. Щоправда, при цьому обсяг різнорідних відомостей, які необхідно переробити, щоб отримати крупиці знань, є величезним, а тому нині конкурентна розвідка немислима без використання спеціалізованих інформаційних технологій, практичного застосування сучасної концепції великих даних (Big Data).

[1] Hwang, Y.W., Lee, I.Y., Kim, H., Lee, H. and Kim, D., 2022. Current

[2] Zanasi, A., 1998. Competitive intelligence through data mining public sources. Competitive Intelligence Review: Published in Cooperation with the Society of Competitive Intelligence Professionals, 9(1), pp.44-54.

На думку колишнього директора Центрального розвідувального управління США (ЦРУ) Р. Хілленкерта, «80 % розвідувальної інформації отримується з таких джерел, як книги, журнали, науково-технічні огляди, фотографії, комерційні аналітичні звіти, газети, теле- та радіопередачі…».

За іншими оцінками, у будь-якій розвідці від 35 до 95 % усієї інформації добувається з відкритих джерел.

Відомо, що для бізнес-структур 95 % корисної інформації надає конкурентна розвідка, 4,1 % інформації можна легально отримати з державних структур. Дозволити собі повноцінне проведення бізнес-розвідки на ринках можуть лише великі компанії, проте можливості конкурентної розвідки доступні практично всім[3].

Значущість розвідки за відкритими джерелами відзначив ще президент США Ліндон Джонсон (Lyndon Baines Johnson) 30 червня 1966 р., коли виголошував промову на церемонії складення присяги директором ЦРУ Річардом М. Гелмсом (Richard McGarrah Helms): «Найвищі досягнення не є результатом потай переказаної таємної інформації, а випливають із терплячого, щогодинного вивчення друкованих джерел»[4].

За усталеною помилковою думкою, вся корисна розвідувальна інформація добувається з секретних джерел агентурним або оперативним шляхом – насправді це не так. Відоме визнання адмірала Захаріаса – заступника начальника розвідки Військово-морських сил США у роки Другої світової війни – спростовує це. Так, за його оцінкою, 95 % інформації розвідка ВМС черпала з відкритих джерел, 4 % – з офіційних, і лише 1 % – з конфіденційних джерел. Заради справедливості слід зазначити, що часто саме цей один відсоток є тим золотим відсутнім елементом, який дозволяє скласти цілісну картину розрізненої мозаїки всіх розвідданих. І якщо таке співвідношення справедливе для військової розвідки, то тим більше воно буде правильним для конкурентної розвідки для бізнес-структур.

---

[3] Ланде Д.В. Правові питання конкурентної розвідки. «Інформація і право», 2020. – № 2(33)ю – С. 51-68. DOI: 10.37750/2616-6798.2020.2(33).208089

[4] Johnson, L.K. ed., 2007. Handbook of intelligence studies (Vol. 2). London: Routledge.

Водночас аналіз розсекреченого звіту ЦРУ за 1987 рік «Enterprise-Level Computing in Soviet Economy»[5] (C87-10043) дає уявлення про те, який колосальний обсяг даних необхідно було обробляти аналітикам. Для складання звіту постійно протягом року сканувалося 347 відкритих джерел; для створення зведення обсягом в одну сторінку щоденно оброблявся інформаційний масив обсягом приблизно 7 млн слів.

Загальновідомо, що основна відмінність конкурентної розвідки від промислового шпигунства – це легітимність і дотримання етичних норм[6]. Тут це положення доведено до абсолюту: виключно всі джерела інформації в цьому разі є доступними та легальними.

Інтернет-розвідка, розвідка за інтернет-джерелами, як, власне, і вся конкурентна розвідка, являє собою особливий вид інформаційно-аналітичної роботи, що дозволяє збирати різнобічну бізнес-інформацію без застосування тих специфічних методів оперативно-розшукової діяльності, які є винятковою прерогативою правоохоронних органів.

Разом із тим методи ведення інтернет-розвідки, техніки та технології її проведення дуже близькі до тих, що використовуються в традиційній розвідувальній діяльності спецслужбами.

Застосування інтернет-розвідки в комерційній компанії виправдовується не лише міркуваннями інформаційної безпеки, але є важливим і для розв'язання завдань менеджменту та маркетингу, оскільки забезпечує:

- спостереження за репутацією компанії (з погляду клієнтів, конкурентів, державних органів);
- активну участь у формуванні іміджу компанії, інформаційного поля навколо компанії;
- відстеження появи нового конкурента, технології або каналу збуту;
- виявлення можливих злиттів і поглинань;
- оцінку потенційних ризиків під час інвестицій;

[5] https://www.cia.gov/readingroom/docs/DOC_0000500558.pdf

[6] Antonina Mytko & Tetiana Mishchuk. Characteristics of Competitive Intelligence and Espionage in Enterprises. International relations, public communications and regional studies, 2017. – № 2. DOI: https://doi.org/10.29038/2524-2679-2017-02-90-103

- випередження кроків конкурентів у рамках маркетингових кампаній;
- випередження конкурентів у тендерах;
- виявлення каналів витоку інформації.

Засновником сучасної бізнес-розвідки вважається компанія Xerox, яка зіткнулася з конкуренцією з боку японських виробників[7]. На початку 70-х років XX ст., після виходу японців на американський ринок, менеджери Xerox помітили, що компанія почала втрачати позиції на ринку. Ситуацію виправили зміни, засновані на зборі актуальної інформації про ринок і конкурентів. Xerox, завдяки своєму японському філіалу, створив систему оцінювання та аналізу роботи (бенчмаркінг), а потім адаптував і застосував до бізнесу розвідувальні технології. При цьому однією з основних умов організації цього процесу було неухильне дотримання закону, оскільки репутація компанії могла б зруйнуватися набагато раніше, ніж можна було б скористатися економічними перевагами промислового шпигунства. Незабаром ці методи роботи почали застосовувати й інші американські компанії. Згодом бізнес-розвідка почала застосовуватися в Європі, а згодом – і в усьому світі.

Ігнорування можливостей бізнес-розвідки на початковому етапі дорого обходилося навіть для найбільших компаній[8]. Так, після створення фотоапарата, який видавав готовий знімок, компанія Polaroid почала почивати на лаврах. Коли аналітичний відділ компанії представив звіт, у якому вказав на перспективи розвитку фотоіндустрії та зародження цифрової ери, керівництво компанії назвало цю інформацію «футуристичною нісенітницею». Минуло деякий час, і в жовтні 2001 року компанія Polaroid ініціювала першу процедуру банкрутства.

Аналогічно в 70-х роках XX ст. «Велика трійка» американських виробників автомобілів не відреагувала на появу на ринку японських автовиробників. Однак самі американці обрали невеликі, економні та надійні японські автомобілі, і американські корпорації зазнали значних збитків.

---

[7] Prescott, J.F. and Miller, S.H. eds., 2001. Proven strategies in competitive intelligence: lessons from the trenches. John Wiley & Sons.

[8] Gilad, B., 2004. Using competitive intelligence to anticipate market shifts. Control Risk and Create Powerful Strategies, Amacom, New York.

Бізнес-розвідники з корпорації Samsung дізналися з відкритої преси, що останній американський завод із виробництва гітар може закритися через дешевші корейські інструменти, а уряд США готується захистити своїх виробників за допомогою митних пошлин. Вчасно дізнавшись про це, представники Samsung встигли ввезти до США велику кількість гітар, а в результаті запровадження ввізних пошлин – ще й підняти ціни на цей музичний інструмент.

Сучасний розвиток інформаційних технологій зробив комп'ютерну конкурентну розвідку доступною навіть для відносно невеликих компаній; нині вона поширена на всіх рівнях економіки.

На практиці понятєва база конкурентної розвідки ще остаточно не сформована: поки що не робиться різниці між термінами «ділова» або «економічна» розвідка, і під конкурентною розвідкою помилково розуміють увесь комплекс заходів, пов'язаний з інформаційно-аналітичним забезпеченням управління підприємницькими ризиками, виявлення загроз, можливостей та інших факторів, що впливають на отримання конкурентних переваг у бізнесі.

В арсеналі тих, хто нині повноцінно займається конкурентною розвідкою, немає спеціальної апаратури, шпигунської техніки. Їхній основний інструмент – комп'ютер, підключений до мережі Інтернет. Діяльність підрозділів (служб) конкурентної розвідки компаній дедалі більше ґрунтується на останніх досягненнях у галузі штучного інтелекту в поєднанні з напрацюваннями в галузях психології, соціології, економіки.

Нині створюються численні професійні об'єднання (спільноти) фахівців у галузі конкурентної розвідки. Найвідоміші з таких спільнот, що займаються організацією конференцій, тренінгів, – це Strategic and Competitive Intelligence Professionals (SCIP) у США (www.scip.org) (рис. 1) та Competia у Канаді (www.competia.com).

До 2014 року конкурентна розвідка в Україні розвивалася переважно в комерційному секторі, зосереджуючись на аналізі ринків, моніторингу конкурентів та оцінці інвестиційних ризиків. Після анексії Криму та початку війни на Донбасі виникла потреба в інтеграції розвідувальних методологій відкритих джерел у сферу безпеки та оборони.

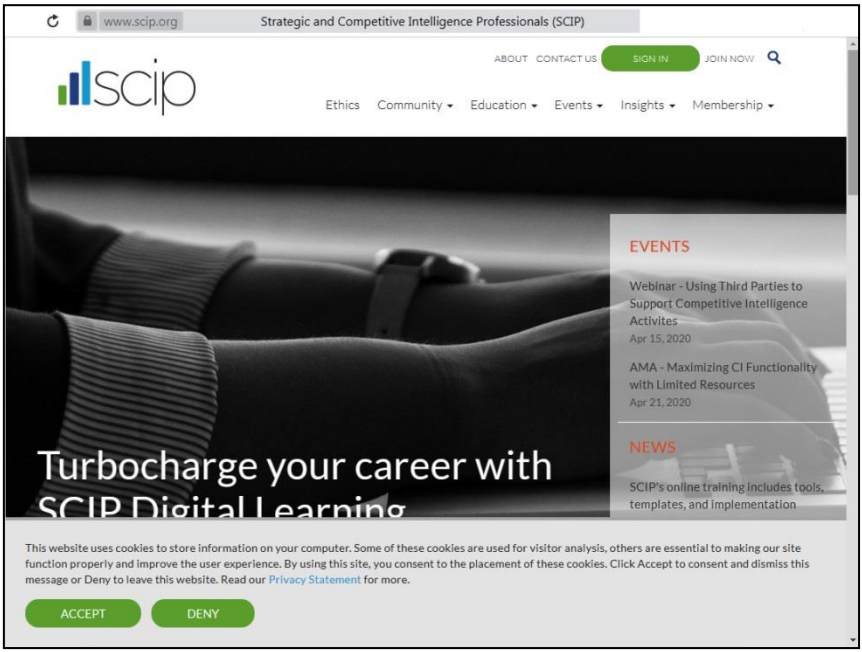


Рисунок 1 – Фрагмент вебсайту організації SCIP (www.scip.org)

Повномасштабне вторгнення 2022 року прискорило цей процес: OSINT перетворився з допоміжного інструменту на основний механізм верифікації та документування воєнних злочинів, моніторингу переміщення військ та аналізу ворожих наративів.

Українська OSINT-спільнота виробила неформальні кодекси поведінки, що балансує між:

- необхідністю оперативного інформування суспільства;
- запобіганням розголошення чутливих даних (координати, тактичні прийоми, особовий склад);
- дотриманням принципів верифікації та мінімізації шкоди від поширення графічного контенту.

Сучасні українські OSINT-фахівці використовують широкий спектр інструментів:

- Геолокація та аналіз супутникових знімків: Google Earth, OpenStreetMap (OSM), NASA FIRMS.
- Соціальні мережі та месенджери: системний моніторинг Telegram-каналів, Twitter/X, VK для виявлення патернів ворожої активності.
- Штучний інтелект: автоматизована класифікація зображень, розпізнавання облич (у межах етичних обмежень), аналіз великих масивів текстових даних.

Українська школа OSINT на цейчас поєднує:

- Традиційні розвідувальні методики (оцінка джерел, перехресна верифікація);
- Цифрову криміналістику (аналіз метаданих, ланцюжків поширення);
- Конкурентно-аналітичні методи (SWOT-аналіз ворога, сценарне моделювання)

OSINT-методи стали основою для:

- ідентифікації підрозділів ворога, причетних до масових вбивств (Буча, Ірпінь);
- відстеження депортації українських дітей до рф;
- фіксації ударів по цивільній інфраструктурі з прив'язкою до конкретних військових формувань.

Публічні OSINT-звіти (наприклад, Oryx[9]) надають верифіковані дані про втрати техніки, що використовується для оцінки ефективності ЗСУ та планування постачання озброєнь.

Волонтерські групи оперативно інформують про переміщення ворожих колон, що дозволяє коригувати вогневе ураження.

OSINT-докази стали ключовими матеріалами для:

- Міжнародного кримінального суду (ICC);
- Міжнародного суду ООН;
- національних прокуратур країн, що застосовують універсальну юрисдикцію.

[9] https://www.oryxspioenkop.com/

Конкурентна розвідка та OSINT в Україні еволюціонували від нішевих бізнес-інструментів до стратегічних компонентів національної стійкості. Унікальність українського досвіду полягає у:

- гібридній моделі співпраці держави, громадянського суспільства та міжнародних партнерів;
- високій адаптивності методологій до динамічних умов гібридної війни;
- етичній зрілості спільноти, що балансує між відкритістю та безпекою.

Для подальшого розвитку критично важливо:

- формалізувати правовий статус OSINT-діяльності;
- інвестувати в підготовку фахівців нового ґатунку – «аналітик-дослідник-технолог»;
- інтегрувати українські напрацювання в міжнародні системи раннього попередження та документування агресії.

1. Конкурентна розвідка та OSINT

1.1. Завдання конкурентної розвідки

Основними завданнями інтернет-розвідки, як сегмента конкурентної розвідки, є:

1. Інформаційне забезпечення процесу вироблення управлінських рішень на стратегічному та тактичному рівнях;
2. Раннє попередження, тобто звернення уваги осіб, що ухвалюють рішення, на загрози, які потенційно можуть завдати шкоди бізнесу;
3. Прогноз і запобігання можливим загрозам бізнесу;
4. Виявлення (спільно зі службою безпеки) спроб конкурентів отримати доступ до закритої інформації компанії;
5. Визначення сприятливих можливостей для бізнесу;
6. Управління ризиками, забезпечення ефективного реагування компанії на швидкі зміни навколишнього середовища, інтернет-простору;
7. Промислова контррозвідка, випередження розвідувальної діяльності конкурентів у мережевому середовищі, аналітична підтримка служби безпеки компанії.

Наведені вище завдання конкурентної розвідки є ключовими; вони слугують досягненню фундаментальної мети конкурентної розвідки – забезпечити захищеність компанії, усвідомленню того факту, що її доля перебуває в руках осіб, що ухвалюють рішення, що компанія не стане раптово жертвою чиїхось ворожих дій.

Крім того, в рамках комп'ютерної конкурентної розвідки мають бути вирішені такі завдання:

- збір і своєчасне забезпечення керівництва та бізнес-підрозділів компанії надійною та всебічною інформацією з мережевих джерел про «зовнішнє» та «внутрішнє» середовище підприємства;
- виявлення факторів ризику, загроз, які можуть зачіпати економічні інтереси бізнесу або завадити його нормальному функціонуванню;

- виявлення нових можливостей та інших факторів, що впливають на отримання конкурентних переваг;
- посилення сприятливих та локалізація несприятливих факторів конкурентного середовища на діяльність бізнес-структури;
- вироблення прогнозів і рекомендацій щодо впливу конкурентного середовища на діяльність бізнес-структури.

Конкурентна розвідка стає сучасним напрямом дослідження поведінки конкурентів на ринку, що дозволяє створювати моделі ринку, його учасників, визначати характеристики та оптимізувати тактику і стратегію розвитку суб'єктів господарювання на певних ринках. Для вирішення її завдань вимагається використання ефективних прийомів роботи з інформацією та її елементами. Інформація при цьому стає як об'єктом дослідження ринку, так і основою для створення його моделі.

Вище сформульовано завдання комп'ютерної конкурентної розвідки, розраховані на легітимну діяльність відповідних структур. Уся система конкурентної розвідки має дозволяти керівництву компанії не лише оперативно реагувати на зміни ситуації на ринках, а й оцінювати подальші можливості свого розвитку. Конкурентна розвідка забезпечує перехід від традиційного ухвалення рішень на основі недостатньої інформації до управління, заснованого на знаннях. При цьому вона також забезпечує зниження ризиків, безпеку бізнесу, а також набуття конкурентних переваг. Сучасна система конкурентної розвідки дозволяє не лише здійснювати моніторинг інформації, а й моделювати стратегію конкурентів, виявляти їхніх партнерів, постачальників, розуміти умови співробітництва.

Основні завдання систем конкурентної розвідки полягають у знаходженні та узагальненні інформації про конкурентів, ринки, товари, бізнес-тенденції та операції за такими основними об'єктами:

- партнери, акціонери, суміжники, союзники, контрагенти, клієнти, конкуренти;
- об'єднання компаній, злиття, поглинання, кризові ситуації тощо;

- кадровий склад компанії, партнерів, конкурентів тощо, а також кадрові зміни, їх динаміка;
- товарообіг, бюджет та його розподіл за статтями;
- укладені договори, угоди або домовленості.

Під час проведення конкурентної розвідки інтерес становить сфера діяльності компаній, сфери їхнього впливу та інтересів. Ці знання можуть застосовуватися, наприклад, для впливу на позиції партнерів та опонентів. Велике значення має інформація, що стосується політики конкурентів, їхніх намірів, сильних і слабких сторін, продукції та послуг, цін, рекламних кампаній, інших параметрів ринку.

1.2. Особливості комп'ютерної конкурентної розвідки

Сучасні відкриті мережеві ресурси, вебсайти, соціальні мережі, відеосервіси, месенджери перетворюються нині на основне джерело інформації та ефективний інструмент для конкурентної розвідки. Вони дають змогу не лише у режимі реального часу відстежувати дії компаній-конкурентів, а й виявляти останні тенденції за відповідною тематикою. Наведемо лише деякі способи використання інтернет-ресурсів для розв'язання завдань конкурентної розвідки[10]:

1. Отримання новин за цільовою тематикою

Сучасні мережеві новинні сервіси, такі як Google News, Yahoo News, UAPort.net, соціальні мережі типу Twitter, Facebook, Reddit, дозволяють отримувати новини, відібрані відповідно до інформаційних потреб користувачів. Наприклад, під час використання соціальної мережі Twitter можна скористатися пошуковим режимом і ввести запит, наприклад «банкрутство». Після цього користувач отримає список повідомлень, у деяких випадках супроводжуваних акаунтами користувачів, чиї повідомлення релевантні введеному запиту. Таким чином, можна ідентифікувати експертів, яких можна згрупувати відповідно до власних інформаційних потреб. Згодом, слідкуючи за думкою групи експертів, можна отримати

[10] Dmytro Lande, Ellina Shnurko-Tabakova. OSINT as a part of cyber defense system. Theoretical and Applied Cybersecurity, 2019. – №. 1. - pp. 103-108.

достатньо широке охоплення проблеми, кілька точок зору, нові інформаційні ресурси.

2. Виявлення тенденцій

За обраними за допомогою пошукових можливостей інформаційними ресурсами (веб-сайтами, соціальними мережами тощо) можна вручну або з використанням спеціальних аналітичних інструментів виявляти тенденції в обраній сфері.

3. Отримання розсилки цільових документів за підпискою (повідомлення месенджерів, електронна пошта, СМС)

Багато новинних агрегаторів і соціальних мереж (зокрема, Twitter) надають можливість якісних персоніфікованих періодичних розсилок, що охоплюють повідомлення, коментарі, експертні канали.

4. Побудова мереж інформаційних зв'язків, когнітивних карт

Для завдань конкурентної розвідки важливо не лише отримання цільової інформації (повідомлень), а й розуміння зв'язків, які виявляються під час аналізу інформації. Важливий не лише об'єкт аналізу, а й пов'язані з ним інформаційні ресурси, профілі в соціальних мережах, «друзі», групи обговорень тощо. У деяких випадках можна переглянути, хто є підписником даних профілів, хто цікавиться тією ж тематикою і, отже, може стати новим джерелом для отримання цільової інформації.

5. Отримання відповідей на запитання

Соціальні мережі, форуми, блоги можна використовувати як спосіб отримання відповідей на конкретні запитання, у тому числі з питань методології конкурентної розвідки. Якщо запитання сформульовано коректно, то з великою ймовірністю можна отримати на нього відповідь від інших користувачів.

6. Фільтрація «інформаційного шуму»

Для конкурентної розвідки не завжди становлять інтерес загальновідомі, часто хибні дані та інформація (Fake News), цікаві більшості, а адже саме на такі дані орієнтовані соціальні

мережі. Під час використання мережевих ресурсів як потужної бази для конкурентної інформації особливу увагу слід приділяти опрацюванню запитів, вибору джерел, експертів, установленню зв'язків.

1.3. Проблеми комп'ютерної конкурентної розвідки

Варто зазначити низку проблем, пов'язаних із комп'ютерною конкурентною розвідкою.

Першою і найбільш суттєвою проблемою є те, що колосальні обсяги інформації в Інтернеті, зокрема в соціальних мережах, ускладнюють пошук і вибір дійсно необхідних відомостей. Самі по собі необроблені, неузагальнені та непроверені дані не можуть забезпечити якісну підтримку під час ухвалення рішень у сфері конкурентної розвідки.

Нині пошуковими системами, зокрема системою Google, індексується понад трильйон документів, обсяги постійно зростають. Разом із цим, за словами Еріка Шмідта (Eric Emerson Schmidt) – голови ради директорів Google у 2001–2011 роках, навіть така потужна пошукова система, як Google, зможе проіндексувати всю наявну сьогодні інформацію лише приблизно через 300 років.

Традиційні пошукові системи в Інтернеті чудово справляються з простими одноразовими запитами, однак, як правило, слабо придатні для потреб конкурентної розвідки. За деякими оцінками, більш як 97 % критичної для конкурентної розвідки відкритої інформації неможливо знайти за допомогою традиційних інформаційно-пошукових систем.

Другою проблемою комп'ютерної конкурентної розвідки є те, що інформація в Інтернеті має динамічний характер: вона розміщується, модифікується та видаляється. Часткове вирішення цих проблем можливе застосуванням систем контент-моніторингу інформаційних потоків у Інтернеті.

Третя проблема, яку необхідно вирішити в цілях конкурентної розвідки, – автоматичне вилучення понять із формалізованих масивів інформації (таблиць, баз даних), а також неструктурованих текстів. Перспективним напрямом вирішення цієї проблеми в системах конкурентної розвідки є використання технологій генеративного штучного інтелекту та Text Mining

Четвертою проблемою є можливість автоматичного виявлення неочевидних закономірностей і зв'язків, зафіксованих у документах. Нині відомо кілька шляхів вирішення проблем вилучення понять із текстів і виявлення їхніх взаємозв'язків, як практичних, так і теоретичних. Одним із цих шляхів є побудова матриць і графів зв'язків понять, моделей предметних областей, когнітивних карт, до яких можна застосовувати відповідні математичні методи. Як правило, вузли цих графів – коефіцієнти, які пропорційні кількості документів, що відповідають досліджуваним поняттям.

П'ятою проблемою є пошук інформації в «прихованому» Інтернеті, де міститься непорівнянно більша кількість даних, потенційно цікавих для конкурентної розвідки, ніж у відкритій частині мережі. Не вся потенційно відкрита «несекретна» інформація є добре доступною, скоріше – навпаки. Добування необхідної в кожному конкретному випадку інформації є складним завданням. На думку експертів, лише порядка 10–15 % необхідної інформації наявне в Інтернеті в готовому вигляді, решту 85–90 % можна отримати в результаті порівняння, агрегування та аналізу численних розрізнених даних.

Отже, в Інтернеті міститься більша частина інформації, необхідної для проведення конкурентної розвідки, однак залишається відкритим питання її знаходження та ефективного використання. Причина – притаманні мережі Інтернет недоліки:

- непропорційне зростання рівня інформаційного шуму;
- засилля паразитної інформації;
- слабка структурованість і зв'язність інформації;
- динамічність інформації;
- відсутність цілісності інформації;
- багаторазове дублювання інформації;
- відсутність можливості семантичного пошуку;
- обмеженість доступу до «прихованого» веб.

Попри це можливості Інтернету оцінюються експертами в галузі конкурентної розвідки досить високо.

1.4. OSINT – розвідка по відкритим джерелам

1.4.1. OSINT як дисципліна розвідки

Як один із синонімів поняття конкурентної розвідки, що часто використовується в силових відомствах різних держав, вживається поняття «розвідка за відкритими джерелами» (OSINT). Це один із напрямів розвідки, який включає пошук, відбір і добування розвідувальної інформації, отриманої із загальнодоступних джерел (не обов'язково комп'ютерних або мережевих), а також аналіз цієї інформації.

OSINT базується на двох основних поняттях:

- відкрите джерело – це джерело інформації, яке надає її без вимоги збереження її конфіденційності, тобто надає інформацію, не захищену від публічного розголошення. Відкриті джерела належать до середовища загальнодоступної інформації і не мають обмежень у доступі для фізичних осіб;
- загальнодоступна інформація – це інформація, опублікована або розміщена для широкого використання; доступна для громадськості.
- За твердженнями аналітика ЦРУ Шермана Кента, станом на 1947 рік політики отримують із відкритих джерел до 80 відсотків інформації, необхідної їм для ухвалення рішень у мирний час. Пізніше генерал-лейтенант Самуель Вілсон, керівник Розвідувального управління Міністерства оборони США у 1976-1977 роках, зазначав, що «90 відсотків розвідданих надходить із відкритих джерел і лише 10 – за рахунок роботи агентури».
- Американський дослідник з питань безпеки Марк М. Ловенталь визначає відкриту інформацію як «будь-яку інформацію, яка може бути отримана із відкритих збірок: усі типи ЗМІ, урядові звіти та інші документи, наукові дослідження та звіти, комерційні постачальники інформації, Інтернет тощо. Основна характеристика відкритої інформації – це те, що для її отримання не вимагаються нелегальні методи збору і що вона може бути отримана за допомогою засобів, повністю відповідних авторським правам і комерційним умовам постачальників».

- Світова спільнота дедалі більше використовує інформацію з відкритих джерел з метою вирішення широкого спектра завдань. Матеріали OSINT слугують базою для всіх методів ведення розвідки як накопичувач розвідувальних даних, їх аналізатор і поширювач.
- Розвідка у відкритих джерелах OSINT є одним із способів ведення розвідки[11], який вносить значний внесок при плануванні бойових дій, а також надає всю необхідну інформацію під час їх проведення. Також визначається:
- OSINT є одним із методів ведення розвідки шляхом збору інформації з відкритих джерел, її аналізу, підготовки і своєчасного надання кінцевого продукту вищому керівництву з метою вирішення певних розвідувальних завдань.
- OSINT є методом ведення розвідки, розробленим на основі збору та аналізу загальнодоступної інформації, і не перебуває під безпосереднім контролем уряду США. OSINT є результатом систематизованого збору, обробки та аналізу необхідної загальнодоступної інформації.

Зокрема, роль OSINT під час проведення розвідки визначається низкою аспектів, серед яких оперативність надходження, обсяг, якість, ясність, легкість подальшого використання, вартість отримання тощо. Наступні фактори впливають на процес планування і підготовки ведення OSINT:

- Ефективне інформаційне забезпечення. Більша частина необхідних довідкових матеріалів про об'єкти інформаційних операцій добувається з відкритих джерел. Це здебільшого досягається шляхом збору інформації із ЗМІ. Накопичення даних з відкритих джерел є основною функцією OSINT.
- Релевантність. Доступність, глибина і масштаби публічно доступної інформації дозволяють знаходити необхідну інформацію без залучення спеціалізованих людських і технічних засобів розвідки.

[11] FMI 2-22.9, 5 December 2006. 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

- Спрощення процесів добування даних. OSINT надає необхідну інформацію, виключаючи потребу в залученні зайвих технічних і людських методів ведення розвідки.
- Глибина аналізу даних. Будучи офіційною частиною розвідувального процесу, OSINT дозволяє керівництву здійснювати глибокий аналіз загальнодоступної інформації з метою ухвалення відповідних рішень.
- Оперативність. Різке скорочення часу доступу до інформації в мережі Інтернет. Скорочення людино-годин, пов'язаних з пошуком інформації, людей та їхніх взаємовідносин на основі відкритих джерел. Швидке отримання цінної оперативної інформації. Стрімко мінлива обстановка під час криз повніше всього відображається у поточних репортажах CNN з місця подій.
- Обсяг. Можливість масового моніторингу певних джерел інформації з метою пошуку контенту, людей і подій, що цікавлять. Як показує досвід, грамотно зібрані фрагменти інформації з відкритих джерел у сукупності можуть бути еквівалентні або навіть більш значущі, ніж професійні розвідувальні звіти.
- Якість. Порівняно зі звітами спеціальних агентів інформація з відкритих джерел виявляється переважнішою вже тому, що позбавлена суб'єктивізму, не розбавлена брехнею.
- Ясність. Якщо у випадку використання OSINT надійність відкритих джерел може бути як ясною, так і неясною, то у випадку таємно здобутих даних ступінь їхньої надійності завжди викликає сумніви.
- Легкість використання. Будь-які таємниці прийнято оточувати бар'єрами з грифів секретності, особливих режимів доступу. Що ж стосується даних OSINT, то їх можна легко передавати будь-яким зацікавленим інстанціям. Можливе проведення комплексного розслідування на основі даних з Інтернету.
- Вартість. Вартість добування даних в OSINT мінімальна, визначається лише вартістю використовуваного сервісу.

Сьогодні пропоновані для OSINT програмно-технологічні рішення забезпечують:

- збір даних із соціальних мереж, таких як Facebook, Twitter або Youtube, аналіз зібраних даних;
- екстрагування із зібраного контенту суті подій;
- агрегування інформації, отриманої з мережі Інтернет;
- інформаційний вплив у мережі Інтернет;
- оцінку достовірності інформації;
- моніторинг і розпізнавання ідентичності в мережі Інтернет, у тому числі за допомогою геолокації;
- роботу з інформацією, отриманою з невидимих за допомогою традиційних мережевих пошукових систем сегментів веб-простору (dark web, hidden web, deep web).

1.4.2. Області застосування OSINT

Існує багато застосувань OSINT, серед яких можна назвати такі.

Розвідка

Відкриті джерела містять величезну кількість інформації, необхідної та такої, що задовольняє потреби розвідувальних органів – як державних, так і приватних, комерційних. Вони забезпечують розуміння об'єктивних і суб'єктивних факторів, пов'язаних, наприклад, із діяльністю конкурентів. При цьому, безумовно, для підвищення ефективності розвідувальної діяльності відкрита інформація використовується у комплексі з іншими ресурсами, зокрема агентурними.

Ініціатива Розвідувального співтовариства США щодо відкритих джерел (відома як National Open Source Enterprise) викладена у Директиві розвідувального співтовариства 301, оприлюдненій директором Національної розвідки. Директива встановлює повноваження та обов'язки помічника заступника директора Національної розвідки з питань відкритих джерел (ADDNI/OS), Центру відкритих джерел DNI та Національного комітету з роботи з відкритими джерелами.

OSINT у збройних силах

Нижче, як приклад, наведено підрозділи збройних сил США, які беруть участь у діяльності OSINT:

- Unified Combatant Command;
- Defense Intelligence Agency;

- National Geospatial-Intelligence Agency;
- US Army Foreign Military Studies Office;
- EUCOM JAC Molesworth;
- Foreign Media Monitoring in Support of Information Operations, U.S. Strategic Command.

Національна безпека

Open Source Enterprise (OSE) – це організація уряду США, що спеціалізується на розвідці на основі відкритих джерел. Спочатку вона входила до складу Office of the Director of National Intelligence, а нині є частиною Директорату цифрових інновацій у Central Intelligence Agency (CIA).

Попередниками цієї організації були Open Source Center (OSC) та Foreign Broadcast Information Service (FBIS).

1 жовтня 2015 року Open Source Center (OSC) змінив назву на Open Source Enterprise і був включений до складу Директорату цифрових інновацій ЦРУ. 21 грудня 2022 року директором Open Source Enterprise (OSE) було призначено Ренді Ніксона, який раніше обіймав посаду директора з цифрового майбутнього в ЦРУ. Ніксон обіймав цю посаду до вересня 2025 року.

Юстиція

Правоохоронне співтовариство застосовує OSINT для прогнозування, запобігання та розслідування злочинів, а також для переслідування злочинців, включно з терористами. Крім того, центри обробки та обміну інформацією (Fusion Centers) у США дедалі частіше використовують OSINT для підтримки розвідувальної діяльності та розслідувань. Такі центри спочатку створювалися під егідою Міністерства внутрішньої безпеки (DHS) та Міністерства юстиції і забезпечували обмін стратегічною інформацією між ЦРУ, ФБР, Міністерством оборони, службами надзвичайних ситуацій, а також місцевими адміністраціями тощо.

Прикладами успішного використання OSINT у правоохоронній діяльності є Scotland Yard OSINT та OSINT-підрозділ Королівської канадської кінної поліції (RCMP).

Поліцейський департамент Нью-Йорка (NYPD) має підрозділ OSINT, так само як і офіс шерифа округу Лос-Анджелес,

розташований у Бюро з надзвичайних операцій і пов'язаний із Об'єднаним регіональним розвідувальним центром Лос-Анджелеса.

У правоохоронній діяльності OSINT може застосовуватися для боротьби з такими явищами, як:

- організована злочинність і банди;
- педофілія;
- викрадення персональних даних і вимагання;
- відмивання грошей;
- злочини у сфері порушення інтелектуальної власності;
- діяльність екстремістських організацій.

При цьому за допомогою OSINT забезпечується виявлення залученості та посилення впливу в Інтернеті:

- ідентифікація ключових фігур і активістів;
- моніторинг конкурентів у режимі реального часу;
- обмеження поширення інформації;
- формування громадської думки;
- виявлення екстремістських організацій;
- ризики для громадського транспорту;
- санкції та правові вимоги;
- аналіз баз даних противників (HME, IED, TTPs);
- геолокація цілей;
- підтримка військових операцій.

Кібербезпека

У межах OSINT забезпечується підтримка процесів кібербезпеки. Зокрема, можна отримати відповіді на такі запитання щодо захисту телекомунікаційних мереж шляхом збору інформації:

- Хто атакує вашу організацію?
- Які їхні мотиви?
- Як вони організовані?
- Які інструменти вони використовують?

Бізнес

OSINT у бізнесі включає комерційну розвідку, інтелектуальну аналітику та бізнес-аналіз і часто є основною сферою діяльності приватних розвідувальних агентств.

Підприємства можуть використовувати інформаційних брокерів і приватних детективів для збирання та аналізу відповідної інформації з діловою метою. Така інформація може походити із засобів масової інформації, глибинної мережі (deep web), вебу нового покоління та комерційного контенту.

1.4.3. Технології OSINT

OSINT є вельми різноманітною формою збирання та аналізування інформації. Під час здійснення діяльності у сфері OSINT часто необхідно вживати запобіжних заходів під час збирання інформації з мережі Інтернет. Це може реалізовуватися шляхом використання VPN для забезпечення анонімності та прихованого збирання інформації, а також проксі-серверів у розподіленому мережевому середовищі. Оцінювання джерел стає важливим елементом загального процесу збирання та аналізування в рамках OSINT. Аналітик OSINT потребує застосування інтелектуального аналізу для виявлення істинних або хибних процесів, що впливатимуть на прогнозування майбутнього розвитку ситуації. Зрештою, аналітики мають забезпечити застосування оцінного аналізу таким чином, щоб його результати могли бути включені до готового класифікованого, некласифікованого або запатентованого інтелектуального продукту.

Збирання інформації в OSINT, як правило, відрізняється від збирання даних в інших розвідувальних дисциплінах, де отримання необробленої інформації, що підлягає аналізу, може становити основну труднощі. В OSINT основною труднощію є виокремлення релевантних та надійних джерел із величезного масиву загальнодоступної інформації.

Етапи OSINT

Процес OSINT складається з чотирьох етапів: планування, підготовки, збору та виробництва кінцевого матеріалу – аналітики, а також чотирьох основних процесів: аналізу, добування й накопичення розвідувальних даних, оцінювання та їх

розподілу за напрямами. Процес ведення розвідки, так само як і процеси підготовки відповідних інформаційних операцій (планування, підготовка, виконання та підбиття підсумків), перетинаються і повторюються відповідно до вимог практики.

Як зазначено в «Інструкції з ведення розвідки в польових умовах», OSINT підвищує ефективність і надає підтримку процесу ведення розвідки та інших операцій.

На рис. 2 подано типову схему процесу здійснення OSINT.

Збір розвідувальних даних синхронізує та інтегрує процеси планування, використання сил і засобів, оброблення та розподілу елементів системи для підтримки бойових операцій, що є об'єднаною розвідувальною та оперативною функцією.

Після аналізу інформація, отримана з різних джерел, перетворюється на розвідувальні дані, які містять необхідні відомості про противника, загрози, клімат, погодні умови, рельєф місцевості тощо.

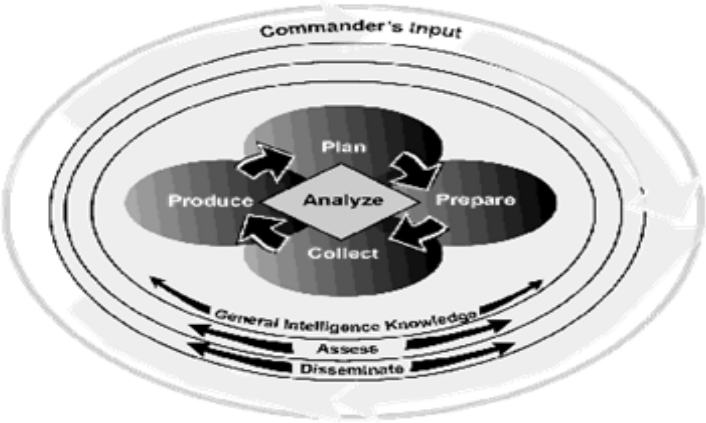


Рисунок 2 – Типова схема процесу ведення OSINT: План ⇒ Підготовка ⇒ Збирання ⇒ Виробництво. Загальні знання розвідки. Оцінка. Розповсюдження[12]

Установлено, що такі елементи структури OSINT, як постійний потік інформації, технічні засоби, програмне забез-

[12] Army open-source intelligence (ATP 2-22-9, 2012)

печення, безпека засобів комунікації та бази даних охоплюють засоби:

- забезпечення доступності розвідувальних даних. Забезпечення доступності розвідувальних даних є процесом, завдяки якому розвідувальні організації активно та оперативно отримують доступ до розвідувальної інформації;
- розроблення та підтримання автоматизованої розвідувальної мережі. Головним завданням є надання інформаційних систем, які забезпечують зв'язок, спільний аналіз і оброблення інформації, поширення матеріалів та створення умов доступності розвідувальних даних;
- створення та підтримання доступу. Це завдання передбачає встановлення, забезпечення та підтримання доступу до секретних і несекретних програм, баз даних, мереж, систем та інших інтернет-ресурсів для військ союзних держав, об'єднаних сил, національних агентств і міжнародних організацій;
- створення та ведення баз даних. Це завдання передбачає створення та підтримання як відкритих, так і секретних баз даних. Формування та ведення баз даних сприяє швидкому аналізу інформації, підготовці звітів, обробленню та поширенню даних, а також підтримці тривалих бойових операцій.

1.4.4. Міжнародний досвід

Ведення розвідки у відкритих джерелах підвищує ефективність діяльності всього розвідувального співтовариства – починаючи з національного і закінчуючи тактичним рівнями. Нижче наведено перелік деяких організацій, які у США займаються здобуванням, накопиченням, використанням, аналізом і поширенням інформації з відкритих джерел.

- Рада із захисту відкритих джерел (DOSC);
- Командування розвідки та безпеки Збройних сил США (INSCOM);
- Служба розвідувальної інформації Департаменту сухопутних військ (DA IIS);
- Центр відкритих джерел директора Національної розвідки (DNI OSC);

- Академія відкритих джерел;
- Департамент передових систем (ASD);
- ФБР;
- Федеральний науково-дослідний відділ (FRD), Бібліотека Конгресу.

Поряд із широким застосуванням OSINT у США наведемо також приклади використання цієї технології в інших країнах.

Служба зовнішньої розвідки Німеччини – Федеральна розвідувальна служба – також використовує переваги Open Source Intelligence у підрозділах Abteilung Gesamtlage/FIZ та Unterstützende Fachdienste (GU).

В Австралії експертом з відкритих джерел є Управління національних оцінок (Office of National Assessments), яке є однією з державних розвідувальних структур. У Великій Британії існує інформаційна служба BBC Monitoring, зосереджена на збиранні відкрито доступної інформації силами журналістів. Аналізом зібраних у BBC даних займаються підписники цього сервісу, зокрема й співробітники секретних британських спецслужб.

Висновки до розділу 1

У першому розділі монографії здійснено системний аналіз теоретико-методологічних засад комп'ютерної конкурентної розвідки та визначено її ключове місце в системі інформаційно-аналітичного забезпечення управління сучасними організаціями. Доведено, що комп'ютерна конкурентна розвідка являє собою легітимну діяльність зі збору, обробки та аналізу інформації виключно з відкритих джерел у комп'ютерних мережах, що принципово відрізняє її від промислового шпионажу суворим дотриманням правових та етичних норм. Основною метою цього процесу є підтримка прийняття управлінських рішень на стратегічному та тактичному рівнях, підвищення конкурентоспроможності суб'єктів господарювання та своєчасне виявлення загроз бізнес-середовищу.

Ключовими завданнями комп'ютерної конкурентної розвідки визначено інформаційне забезпечення процесу вироблення рішень, раннє попередження про потенційні ризики, прогнозування дій конкурентів, виявлення нових можливостей для розвитку та забезпечення інформаційної безпеки підпри-

ємства шляхом протидії розвідувальній діяльності опонентів у мережевій среде. Особлива увага приділяється трансформації підходів до збору даних у бік автоматизованих процедур моніторингу вебпростору, соціальних мереж, блогосфери та месенджерів, що дозволяє відстежувати динаміку інформаційних потоків у режимі реального часу та виявляти актуальні тенденції. Ідентифіковано термін OSINT як синонімічний поняттю інтернет-розвідки в контексті бізнесу, хоча в міжнародній практиці він традиційно асоціюється з державним сектором та силовими відомствами, де накопичено значний методичний досвід.

Проаналізовано основні проблеми галузі, серед яких домінують інформаційне перевантаження через колосальні обсяги даних в Інтернеті, динамічний характер мережевої інформації, складність автоматичного вилучення понять з неструктурованих текстів та обмеженість доступу до глибинного вебу. Доведено необхідність застосування спеціалізованих технологій контент-моніторингу, текстової аналітики та математичного моделювання для фільтрації інформаційного шуму, виявлення прихованих закономірностей та побудови мереж інформаційних зв'язків. Обґрунтовано важливість розмежування понять бізнес-розвідки та конкурентної розвідки, де остання зосереджена переважно на зовнішньому середовищі та безпосередніх конкурентах.

Розглянуто міжнародний досвід використання розвідки за відкритими джерелами у військовій сфері, національній безпеці, правозастосуванні та бізнесі, що підтверджує високу ефективність таких методів та їхню здатність забезпечувати значну частку необхідної для прийняття рішень інформації при мінімальних витратах. Загалом матеріали розділу формують базис для подальшого дослідження технологічних інструментів, математичних основ та програмних реалізацій систем конкурентної розвідки, які розглядаються в наступних розділах монографії, та підкреслюють актуальність розвитку цих напрямів в умовах глобалізації економіки та загострення інформаційного протистояння.

2. Технології конкурентної розвідки

Комп'ютерна конкурентна розвідка використовує у своєму арсеналі різноманітні засоби, найбільш розвиненими з яких є спеціалізовані інформаційно-аналітичні системи.

Інформаційно-аналітична система комп'ютерної конкурентної розвідки включає такі компоненти:

- комплекси контент-моніторингу інформації з відкритих мереж (вебпростору, соціальних, піринговых мереж тощо);
- засоби екстрагування понять (компаній, осіб, подій тощо) з повнотекстових документів;
- засоби виявлення та візуалізації інформаційних зв'язків, виявлення аномалій і неочевидних закономірностей;
- засоби формування аналітичних документів, що надаються особам, які приймають рішення (ОПР).

Змістовна частина, інформаційна база інформаційно-аналітичної системи конкурентної розвідки формується комплексом контент-моніторингу. Особливості сучасних комплексів контент-моніторингу полягають у тому, що вони повинні охоплювати величезні обсяги інформації з динамічно зростаючих інформаційних потоків у мережах за наявності шумової інформації, значної частини слабкодоступних ресурсів, так званого «прихованого Інтернету». У деяких випадках реалізація цього комплексу може бути передана так званим «процесорам збору даних» – компаніям, які займаються цілеспрямованим збором великих обсягів інформації із соціальних медіа відповідно до вимог замовників.

Формування бази даних (БД) ІАС ККР відбувається шляхом підключення до мережі Інтернет і збору (за певними критеріями та акаунтами) інформації з визначених інформаційних ресурсів (наведений нижче список може розширюватися):

Вебсайти

Блоги:

- X (Twitter);
- LiveJournal.

Соціальні мережі:

- Facebook;
- Instagram;
- LinkedIn;
- Reddit;
- Medium.

Відеосервіси:

- YouTube;
- TikTok.

Месенджери:

- Telegram;
- Viber.

Крім того, має бути передбачена можливість налаштування адміністратором комплексу контент-моніторингу ІАС ККР модулів автоматичного сканування та первинної обробки, а за потреби – створення службових акаунтів, через які організовується доступ до певних інформаційних ресурсів.

За допомогою комплексів контент-моніторингу в межах конкурентної розвідки, як правило, розв'язуються такі завдання:

- моніторинг діяльності партнерів, конкурентів, регуляторних органів;
- контроль медіаприсутності та медіаактивності учасників ринку;
- пошук інформації про учасників ринків;
- виявлення нових продуктів на ринках;
- виявлення нових гравців на ринках;
- організація ретроспективного інформаційного фонду документів для їх подальшого використання в аналітичній діяльності.

Процес перетворення «сирих» даних на знання та доведення їх до кінцевих споживачів прийнято називати розвідувальним циклом. У класичному розумінні розвідувальний цикл прийнято поділяти на п'ять основних етапів:

- визначення цілей, планування, визначення джерел інформації;
- збір, здобування даних;

- обробка розвідувальних даних – перетворення їх на розвідувальну інформацію;
- аналіз і синтез розвідувальної інформації – перетворення її на знання (висновки, рекомендації, рішення);
- доведення інформації до кінцевих споживачів.

Слід також зазначити деякі ключові особливості зазначених етапів:

- визначення цілей і планування доцільно поділяти на три рівні – стратегічний, тактичний і оперативний;
- на етапі збору інформації надзвичайно важливо залучити якомога більшу кількість незалежних і первинних джерел;
- процес обробки даних передбачає облік, класифікацію, відбір, верифікацію та оцінку здобутих відомостей;
- розвідувальний цикл у деяких випадках може не потребувати глибокого опрацювання: наприклад, за умов обмеженого часу він може бути неповним і завершуватися переданням споживачам не знань у вигляді остаточних висновків, рекомендацій чи проєктів рішень, а лише обробленої інформації у вигляді інформаційних довідок;
- у розвідувальному документі не повинно бути посилань на конфіденційні джерела інформації, оскільки це може призвести до їх розкриття;
- висновки та рекомендації повинні бути чіткими, короткими й однозначними, а прогнози – мати ймовірнісний характер;
- доведення інформації до кінцевих споживачів має здійснюватися у формі, адаптованій до сприйняття замовника та зручній для розуміння (цікаво зазначити, що, наприклад, ЦРУ надавало президенту США Р. Рейгану щоденну інформацію у вигляді відеофільму, який знімали щодня, оскільки колишній кіноактор сприймав таку форму подання інформації більш адекватно).

Отже, відкриті джерела є найбільш доступним каналом інформації. Під час їх використання зростає об'єктивність отримуваної інформації, проте різко збільшуються трудові витрати на вилучення потрібних відомостей. Тому в комп'ютерній кон-

курентній розвідці повинні застосовуватися спеціалізовані методики та системи. Такі методики й системи створювалися в інтересах спецслужб упродовж багатьох років як на Заході, так і в колишньому Радянському Союзі. Перехід за останні 10–20 років значної частини світової інформації з паперової форми в електронну, широке використання та зростання обсягів мережі Інтернет, сучасні інформаційні технології зробили конкурентну розвідку в Інтернеті одним із найперспективніших напрямів розвідувальної діяльності. Той факт, що так діють практично всі спецслужби світу, лише підтверджує перспективність цього напряму.

Для пошуку та збору інформації в комп'ютерних мережах в інтересах розвідки у всьому світі використовуються спеціальні моніторингові системи збору даних, процесори збору даних, які застосовують спеціальні програмні комплекси (на комп'ютерному сленгу їх називають «роботами» або «павуками»). Програма-робот самостійно обходить за заданим графіком зазначені адреси (URL) у мережі Інтернет, завантажує з них дані, а потім вилучає з них потрібну інформацію, використовуючи цілий арсенал засобів лінгвістичного, семантичного та статистичного аналізу. Такі програмні комплекси автоматично перехоплюють будь-яку інформацію, поставлену на моніторинг, щойно вона з'являється у доступному сегменті мережі.

Під час організації комп'ютерної конкурентної розвідки широке застосування отримав напрям науки, що виник на стику штучного інтелекту, статистики та теорії баз даних, відомий як Knowledge Discovery (виявлення знань), який використовує концепції Data Mining (глибинний аналіз формалізованих даних) та Text Mining / Information Extracting (глибинний аналіз текстів / вилучення знань з інформації). Унікальними особливостями цих концепцій і технологій є те, що з їх допомогою можна отримувати з «сирих» даних раніше невідомі, неочевидні, практично корисні та придатні для інтерпретації знання, необхідні для прийняття рішень у різних сферах діяльності. Такі технології здебільшого застосовувалися спеціальними службами.

Одним із перших розсекречених подібних комплексів стала французька система TAIGA (Traitement Automatique de l'Information Géopolitique d'Actualité – автоматична система

обробки актуальної геополітичної інформації)[13]. Цей програмний комплекс протягом використовувався в інтересах французької розвідки, а на цей час дозволений до комерційного використання. Новий більш досконалий комплекс Noemic, поставлений на озброєння французької розвідки, здатний обробляти інформацію зі швидкістю понад 1 мільярд знаків за секунду. Американський аналог цих програмних комплексів Topic (у перекладі – «Тема») також уже розсекречений і переданий для комерційного використання.

Аналогічні аналітичні системи створювалися і в колишньому СРСР, зокрема в Росії. Достатньо згадати такі відомі системи ФАПСІ, як «Барометр» і «Ельбрус». Вони займалися обробкою російської та зарубіжної преси, статистичної й оперативної інформації.

Створення та використання подібних систем триває й нині. Наприклад, система Radian6 (www.radian6.com) призначена для відстеження в реальному часі згадувань брендів у соціальних мережах з урахуванням тональності повідомлень і для участі в обговореннях. Інша система – Alterian SM2 – також дає змогу відстежувати згадування брендів, локалізувати місця обговорень і визначати демографічні характеристики користувачів соціальних мереж. На цей час провідними системами є ActivTrak, ChartMogul, Cluvio, Databox, Matomo Analytics, Plausible Analytics, Metabase, Tableau. В Україні також створено й розвивається десятки інформаційно-аналітичних систем конкурентної розвідки, про які йтиметься далі.

На перший погляд може здатися, що всі наведені приклади – це системи, які або використовуються державними структурами, або є надто дорогими для використання «середньостатистичними» компаніями. Насправді це не зовсім так. На сучасному ринку представлено цілу низку як західних комерційних продуктів, так і вітчизняних розробок, здатних у тому чи іншому обсязі виконувати подібні завдання в інтересах конкурентної розвідки комерційних структур.

[13] Gras, C. and Oiry, A., 2022. Géographies narratives. Entretien avec Cédric Gras, écrivain-voyageur, réalisé par Anaïg Oiry, le 16 mai 2022 à Paris. EchoGéo, (60).

2.1. Пошук інформації в Інтернеті

Щоб отримати в мережі Інтернет крихти інформації, необхідної користувачеві, потрібно обробити величезні масиви сирих даних. Природно, що для полегшення цього завдання використовуються спеціальні пошукові інструменти.

Пошук інформації в Інтернеті лише шляхом перегляду окремих вебсайтів, по-перше, має вибірковий і/або випадковий характер (до того ж інформація на окремих сайтах може бути досить суб'єктивною або навіть замовною), по-друге, є малопродуктивним.

Усі наявні засоби пошуку інформації в Інтернеті можна умовно поділити на кілька підгруп, а саме:

- засоби пошуку інформації на окремих сайтах;
- добірки посилань, каталоги;
- пошукові системи;
- метапошукові системи;
- системи моніторингу та контент-аналізу;
- екстрактори об'єктів, подій і фактів;
- системи Knowledge Discovery, Data Mining, Text Mining;
- спеціалізовані системи конкурентної розвідки;
- інтегровані системи.

Усі каталоги, пошукові системи та метапошукові системи є вебсайтами зі спеціалізованими базами даних, у яких зберігається інформація про інші вебсайти та документи, що на них розміщені. За запитом до таких систем видається список гіперпосилань, а іноді й короткий опис документів (сніпети). Як правило, пошук може здійснюватися за ключовими словами та фразами. Активувавши гіперпосилання, знайдене в результаті запиту, користувач потрапляє на оригінал документа. Природно, що якщо документ з часом змінився або вебсайт припинив своє існування, то й первинно проіндексований пошуковою системою документ через деякий час може бути не знайдений.

Основна відмінність пошукових систем від каталогів полягає у наявності автоматичного «робота», який постійно сканує вебпростір і накопичує нову інформацію в індексних файлах бази даних. До каталогів інформація, як правило, заноситься вручну – або власниками сайтів, або персоналом самих ката-

логів. Користування такими системами зазвичай є безкоштовним.

Метапошукові системи – це системи, що інтегрують результати пошуку різних пошукових систем. Оскільки окремі пошукові системи по-різному індексують різні сегменти мережі, то, природно, і результат пошуку за допомогою метапошукової системи буде повнішим, ніж за допомогою однієї окремої пошукової системи. Другою пошуковою перевагою таких систем є те, що одним запитом забезпечується пошук у багатьох пошукових системах, без необхідності багаторазового повторення одного й того самого запиту.

Системи моніторингу забезпечують регулярний пошук і «завантаження» інформації за заданими темами та з визначених сайтів, а також аналіз змісту «завантажених» документів. Такі системи, як правило, мають розвинену мову запитів, що дає змогу суттєво деталізувати та конкретизувати запити порівняно зі звичайними пошуковими системами. Крім того, такі системи зберігають у своїх базах даних повні тексти вихідних документів, що забезпечує збереження цих документів у часі та можливість їх оброблення і контент-аналізу як у поточному режимі, так і ретроспективно. Істотною перевагою таких систем є також те, що складні запити, які складаються з десятків або сотень пошукових слів і виразів і одного разу створені аналітиком-експертом, можуть бути збережені у вигляді каталогізованого запиту або рубрики та надалі викликатися автоматично або вручну зі збереженого списку для проведення пошуку чи аналізу.

За допомогою контент-аналізу такі системи дають змогу встановлювати перехресні зв'язки між темами, поняттями та об'єктами, поставленими на моніторинг, виявляти емоційне забарвлення документів, аналізувати динаміку появи тих чи інших документів у часі, здійснювати порівняльний аналіз інформаційної активності за різними тематиками та багато іншого.

Якщо системи моніторингу як системи фільтрації можуть виділяти з інформаційного потоку відомі об'єкти, то екстрактори об'єктів, подій і фактів уміють виділяти з потоку інформації об'єкти, заздалегідь невідомі, події або факти, що лише відповідають певному наперед визначеному типу, наприклад географічні поняття, персони, структури й організації, події

(дорожньо-транспортні пригоди, катастрофи, міжнародні зустрічі). При цьому факти можуть класифікуватися як звичайні або незвичайні. Прикладом звичайного факту в цьому випадку можна вважати виїзд автомобілів за межі міста, а прикладом незвичайного факту – виїзд за ті самі межі міста автомобіля без номерних знаків.

Системи типу Knowledge Discovery, технології Data Mining і Text Mining здатні виявляти нові знання та закономірності. Така система, наприклад, може самостійно, без участі людини, зробити висновок про факт знайомства між людьми, спираючись на наявні в системі дані про те, що вони закінчили одну й ту саму школу і один і той самий клас в одному й тому самому населеному пункті. Щоправда, самі правила, за якими така система робить висновки, усе ж поки що створюються і задаються людьми.

Спеціалізовані системи для конкурентної розвідки можуть включати одне або кілька з перелічених вище пошукових засобів, адаптованих до цих специфічних завдань. Крім того, потреби конкурентної розвідки передбачають використання як джерел інформації, окрім повнотекстових документів, також доступних у мережі Інтернет баз даних, власних документів, таблиць і баз даних компанії, а також формалізованих і неформалізованих документів і баз даних, отриманих з інших джерел.

У країнах Європейського Союзу звичайна людина зареєстрована більш ніж у 300 базах даних, таких як реєстрація місця проживання, страхування, водійські права, банки, кредитні бюро, інформаційні, рейтингові та рекрутингові агентства, служби зайнятості, медичні та поліцейські реєстри, супермаркети, клуби, системи управління взаємовідносинами з клієнтами комерційних фірм (так звані CRM-системи) тощо. В інтересах конкурентної розвідки та маркетингу аналізуються не лише ринки товарів і послуг, а й смаки та уподобання окремих клієнтів. Інформація про юридичних осіб, що зберігається у різних базах даних, є ще більш обширною.

З метою бізнес-розвідки необхідно аналізувати дані з усіх доступних джерел інформації, однак у межах цієї роботи не розглядатимуться джерела інформації, не представлені в Інтернеті.

Інтегровані засоби конкурентної розвідки включають не лише всі доступні пошукові засоби, а й банк виявлених (здобутих) і логічно пов'язаних між собою даних, інформації та знань.

З точки зору створення інформаційно-аналітичних систем така система концептуально має передбачати реалізацію трьох принципів:

- єдиний інформаційний простір взаємопов'язаних концептів – об'єктів і фактів незалежно від типу їхніх джерел або контенту;
- збереження зв'язків концептів із релевантними даними та джерелами інформації;
- історико-просторову модель банку даних системи, яка передбачає наявність у всіх об'єктів обліку атрибутів часу та місця.

Заради справедливості слід зазначити, що, згідно зі звітами Fuld's Intelligence Software Report, відомих комерційних версій повноцінних інтегрованих систем, які дозволяли б розв'язувати весь комплекс завдань конкурентної розвідки, поки що не існує, принаймні на Заході.

Пошукові системи чудово справляються з простими одноразовими запитами. Коли ж предметна область є складною або надто широкою (наприклад, «політика», «економіка») або, навпаки, надзвичайно вузькою і віддаленою в часі (наприклад, умови угоди деяких компаній п'ятирічної давності), а необхідно узагальнити всі інформаційні теми та приводи за цією тематикою, оцінити їх у часовій динаміці, знайти взаємозв'язки з іншими об'єктами, скласти цілісну картину щодо об'єкта, що цікавить, або виділити нестандартну подію із загального масиву, то можна переконатися, що:

- результати видачі пошукових систем або перевантажені тисячами непотрібних посилань, або, навпаки, є недостатніми;
- інформація в мережі Інтернет не зберігається довго: необхідну інформацію, що була на цільовому сайті місяць тому, сьогодні можна там не знайти;
- пошукова система не зберігає переглянуті аналітиком посилання, і йому щоразу доводиться починати рутинну роботу з нуля після вимушеної перерви;

- пошукова система не завжди відрізняє справді важливу інформацію від інформаційного шуму;
- пошукова система не завжди здатна узагальнювати або порівнювати інформацію за змістом чи іншими змістовними критеріями;
- пошукові системи не охоплюють деякі вебресурси або окремі види інформації (наприклад, інформацію з баз даних), а деякі вебресурси, навпаки, завжди відображаються на перших сторінках результатів, хоча їхній зміст не цікавить авторів запитів;
- пошукові системи можуть виконувати пошук інформації лише за безпосередньо введеним запитом і не завжди можуть автоматично повторювати його у визначений час без участі користувача.

На перший погляд може здатися, що всі перелічені приклади – це системи, які або використовуються державними структурами, або є надто дорогими, щоб їх могли застосовувати «середньостатистичні» компанії. Насправді це не зовсім так. На сучасному ринку представлено цілу низку як західних комерційних продуктів, так і вітчизняних рішень, здатних тією чи іншою мірою виконувати подібні завдання в інтересах конкурентної розвідки комерційних структур.

Значну частину критично важливої для бізнесу інформації з мережі Інтернет неможливо знайти за допомогою традиційних інформаційно-пошукових систем. Мережеві інформаційно-пошукові системи не повною мірою справляються із завданнями конкурентної розвідки. Тому розробляються спеціалізовані системи, орієнтовані на завдання мережевої аналітики, конкурентної розвідки. Наведемо опис деяких із них:

Website-Finder (www.softpedia.com) – програма, яка дає змогу шукати вебсайти, що погано індексуються пошуковою системою Google. Для кожного запиту видається 30 результатів. Програма проста у використанні, є безкоштовна версія.

Global Supplier Directory by Solusource (www.worldindustrialreporter.com/solusource) – веб-інтерфейс для конкурентної розвідки від компанії Thomas. Дає змогу знайти інформацію, наявну в ретроспективних базах даних Thomas (охоплення – понад 100 років), про компанії, продукти та галузі.

dtSearch (www.dtsearch.com) – пошукова програма, яка дає змогу обробляти терабайти тексту як на локальному диску, так і в мережевому оточенні. Підтримує статичні та динамічні дані. Дає змогу шукати у всіх форматах MS Office.

InfoNgen (www.infongen.com) – агрегатор, що охоплює в режимі перегляду понад 35 тисяч онлайн-джерел, який легко налаштовується на унікальні теми. Поєднує моніторинг, фільтрацію та агрегацію інформації за запитами конкретного користувача. Надає інформацію вісьмома мовами, забезпечує переклад англійською мовою.

Sentinel Visualizer (https://www.softwareadvice.com/bi/sentinel-visualizer-profile/) – одна з найкращих у світі програм із візуалізації зв'язків і відносин Sentinel Visualizer.

Web Content Extractor (newprosoft.com) – "Web Content Extractor" є найпотужнішим, простим у використанні програмним забезпеченням для вилучення даних із вебсайтів.

Screen-Scraper (screen-scraper.com) – дає змогу автоматично вилучати всю інформацію з вебсторінок, завантажувати переважну більшість форматів файлів, автоматично вводити дані в різні форми. Працює під усіма основними платформами, має повнофункціональну безкоштовну та дуже потужні професійні версії.

Attackindex (attackindex.com) – система, що дає змогу отримати відповіді на запитання: чи ведеться проти користувача інформаційна атака чи стався природний сплеск інтересу до події; коли розпочалась інформаційна операція, наскільки вона інтенсивна та масштабна; які сайти та акаунти в соцмережах використовуються для атаки; хто став ініціатором інформаційної операції та як пов'язані її учасники (Рис. 3).

Photoinvestigator (photoinvestigator.co) – сервіс для вилучення метаданих та іншої інформації з фотографій.

Visual.ly (visual.ly) – система пошуку інфографіки у вебпросторі.

CIRadar (www.ciradar.com/Competitive-Analysis.aspx) – комерційна англомовна система пошуку інформації для конкурентної розвідки у «глибинному» вебі. Реалізована як вебсервіс.

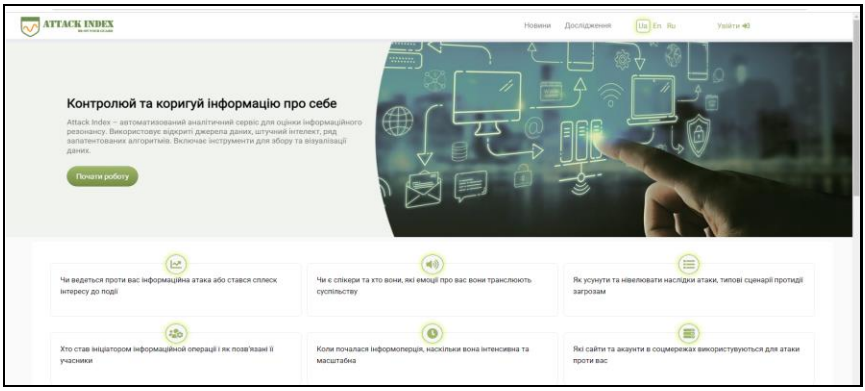


Рисунок 3 – Фрагмент сторінки вебсайту Attack Index (attackindex.com)

2.2. Моніторинг інформаційного простору

Сучасні методи контент-моніторингу – це адаптація класичних методів контент-аналізу та глибинного аналізу текстів (Text Mining) до умов формування й розвитку динамічних інформаційних масивів, наприклад, потоків інформації з мережі Інтернет. Перше типове завдання контент-моніторингу – побудова діаграм динаміки появи понять (відображення подій) у часі.

На прикладі ринку нафтопродуктів розглянемо, як із масивів текстової інформації з мережі Інтернет можуть бути виявлені документи, що містять максимальну кількість цінової інформації щодо цього ринку.

Розглянемо, як у системі контент-моніторингу InfoStream (www.infostream.ua) відстежуються публікації, що стосуються паливної кризи 2026 року. Для цього було складено запит «Oil crisis Iran», введений через вебінтерфейс системи.

На діаграмі, що з'явилася після виконання запиту, видно, що пік кризи припав на початок березня 2026 року (Рис. 4) і був пов'язаний зі збройною операцією США та Ізраїлю проти іранської тиранії. Якщо перейти в режим «Сюжети», у якому передбачено кластеризацію результатів пошуку з урахуванням вагових критеріїв, то можна видати користувачеві лише найвагоміші ланцюжки документів (Рис. 5). Після цього достатньо перейти в режим перегляду сюжету та проаналізувати документи, посилання на які видано системою (Рис. 6).

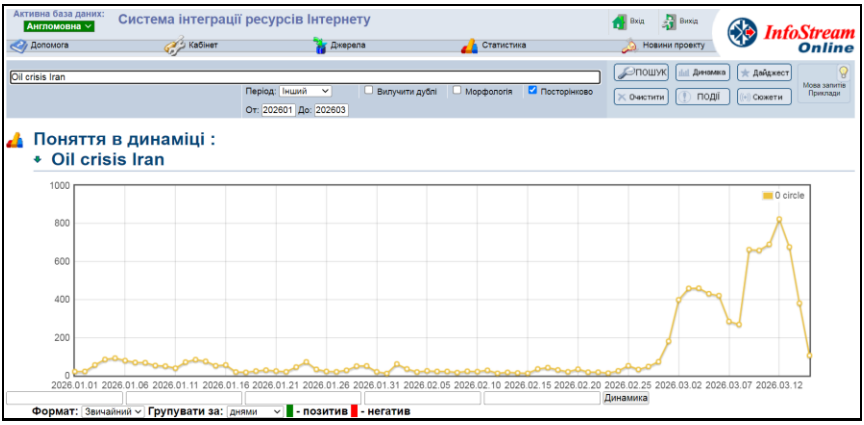


Рисунок 4 – Діаграма динаміки поняття у часі

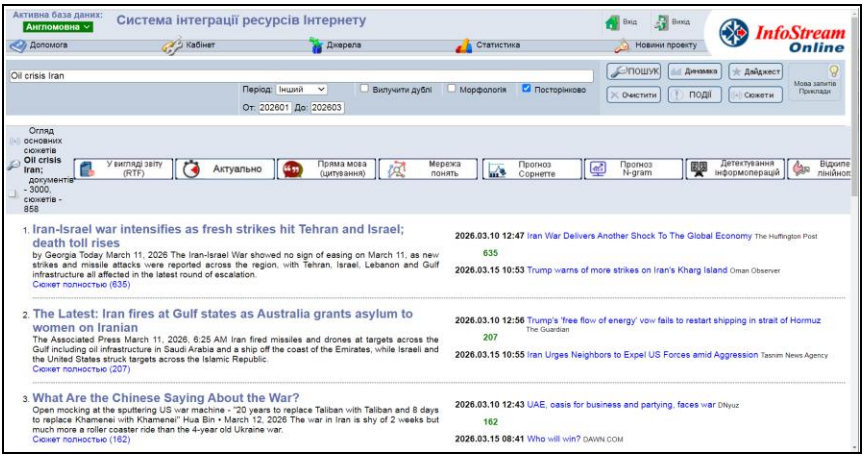


Рисунок 5 – Основні сюжетні ланцюжки за запитом

2.3. Text Mining, Information Extraction

Завдання, яке необхідно постійно розв'язувати під час проведення конкурентної розвідки, – автоматичне вилучення понять і фактів із формалізованих масивів інформації (таблиць, баз даних) та неструктурованих текстів, представлених у веб-просторі, а також виявлення глибинних зв'язків між окремими поняттями. Для цього передбачається використання в системах конкурентної розвідки технологій Knowledge

Discovery, концепції глибинного аналізу даних і текстів (Data Mining, Text Mining).

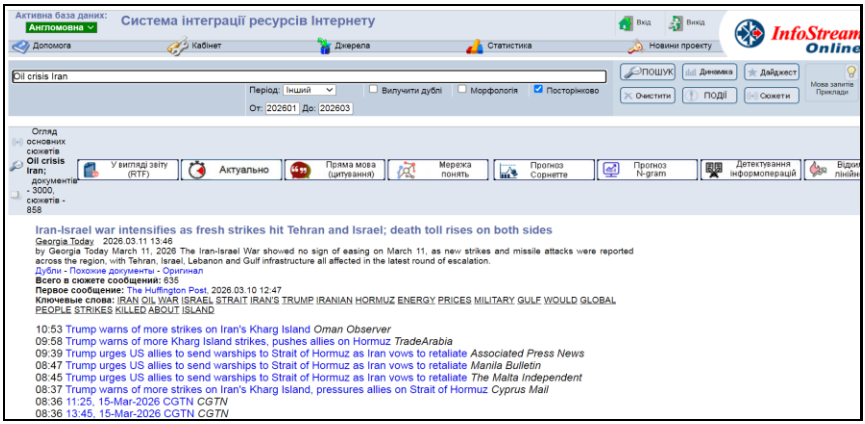


Рисунок 6 – Фрагмент ланцюжка основних сюжетів

Важливе завдання технології Text Mining пов'язане з вилученням із тексту його характерних елементів або властивостей, які можуть використовуватися як метадані документа, ключові слова, анотації. Інше завдання полягає у віднесенні документа до певних категорій із заздалегідь заданої схеми класифікації. Text Mining також забезпечує новий рівень семантичного пошуку документів.

Відповідно до методології, що склалася на цей час, до основних елементів Text Mining належать[14]: класифікація (Classification), кластеризація (Clustering), побудова семантичних мереж, вилучення фактів і понять (Feature Extraction), реферування (Summarization), відповіді на запити (Question Answering), тематичне індексування (Thematic Indexing) і пошук за ключовими словами (Keyword Searching). У деяких випадках цей набір доповнюється засобами підтримки та створення так-

[14] Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник / Д.В. Ланде, І.Ю. Субач, Ю.Є. Бояринова. -К.: ІСЗЗІ КПІ ім. Ігоря Сікрорського, 2018. - 300 с. ISBN 978-966-2577-12-9

сономій (Taxonomies), тезаурусів (Thesauri) та онтологій (Ontology).

Під час класифікації текстів використовуються статистичні кореляції для створення правил розміщення документів у певних категоріях. Завдання класифікації – це класичне завдання розпізнавання, коли на основі певної контрольної вибірки система відносить новий об'єкт до тієї чи іншої категорії. Особливість концепції Text Mining полягає в тому, що кількість об'єктів і їхніх атрибутів може бути дуже великою – передбачається застосування інтелектуальних механізмів оптимізації процесу класифікації.

Кластеризація ґрунтується на ознаках документів, застосуванні лінгвістичних і математичних методів без використання заздалегідь заданих категорій. Результатом кластеризації може бути таксономія або візуальна карта, яка забезпечує ефективне охоплення великих обсягів даних. Кластеризація в Text Mining розглядається як процес виділення компактних підгруп об'єктів із близькими властивостями. Засоби кластеризації дають змогу знаходити ознаки та розподіляти об'єкти на підгрупи на основі цих ознак. Кластеризація, як правило, передує класифікації, оскільки дає змогу визначити групи об'єктів.

Під час побудови семантичних мереж передбачається аналіз зв'язків між поняттями, екстрагованими з документів. Появі понять відповідає поява певних дескрипторів (ключових фраз) у документах. Зв'язки між поняттями можуть установлюватися в найпростішому випадку шляхом урахування статистики їхнього спільного згадування в різних документах.

Вилучення або екстрагування фактів (понять) призначене для отримання певних фактів із тексту з метою поліпшення класифікації, пошуку, кластеризації та побудови семантичних мереж.

Автоматичне реферування (Automatic Text Summarization) – це складання коротких викладів матеріалів, анотацій або дайджестів, тобто вилучення найважливіших відомостей з одного чи кількох документів і генерація на їхній основі лаконічних, зрозумілих і змістовних звітів.

На основі методів автоматичного реферування можливе формування пошукових образів документів. За автоматично побудованими анотаціями великих текстів – пошуковими образами документів – може здійснюватися пошук, що характери-

зується високою точністю (природно, за рахунок повноти). У деяких випадках замість пошуку в повних текстах масиву великих за обсягом документів доцільно здійснювати пошук у масиві спеціально створених анотацій. Хоча пошукові образи документів часто є утвореннями, що лише віддалено нагадують вихідний текст і не завжди сприймаються людиною, завдяки входженню найвагоміших ключових слів і фраз вони допомагають отримувати цілком адекватні результати під час проведення повнотекстового пошуку.

Унікальною особливістю концепції та технологій Text Mining є те, що з їхньою допомогою можна вилучати з «сирих» даних неочевидні, практично корисні та доступні для інтерпретації знання, необхідні для прийняття рішень у різних сферах діяльності, зокрема у сфері економічної конкуренції.

На сучасному ринку представлено цілу низку як західних продуктів, так і систем виробництва пострадянських країн, здатних у тому чи іншому обсязі здійснювати глибинний аналіз текстів.

Останнім часом усі основні західні бренди, що спеціалізуються на розробленні інформаційних сховищ і баз даних, корпоративних систем управління, розширили свої лінійки продуктів системами або модулями Text Mining. Про наявність таких модулів заявляють SAP, Oracle, SAS, IBM та інші компанії.

Процес конкурентної розвідки можна розглядати як побудову мережі з досліджуваних об'єктів і зв'язків між ними. Результати мають становити аналітичну інформацію, яка може бути використана для прийняття рішень. Аналітична інформація може бути представлена у вигляді наочних схем – семантичних мереж, дайджестів, наборів сюжетних ліній, взаємозв'язків ключових понять, компаній, осіб, технологій тощо.

Завдання конкурентної розвідки породили попит на спеціальні інформаційні технології, що забезпечують можливість вилучення й оброблення необхідної інформації, що, у свою чергу, спричинило потік пропозицій систем із боку розробників програмного забезпечення.

Сьогодні розв'язувати завдання конкурентної розвідки на основі інформації з мережі Інтернет допомагають загальнодоступні та спеціалізовані програми й сервіси. Зокрема, останнім часом набули популярності так звані «персоналізовані ро-

звідпортали», здатні відбирати інформацію з найвужчих, специфічних питань і тем та надавати її замовникам.

Нині декларуються технології та системи «комп'ютерної конкурентної розвідки», ідея яких полягає в автоматизації та прискоренні процесів вилучення з відкритих джерел інформації, необхідної для конкурентної боротьби, а також її аналітичного опрацювання.

Під час проведення конкурентної розвідки дедалі ширше застосовуються нові напрями науки і технологій, що отримали назви: «управління знаннями» (Knowledge Management) і «виявлення знань у базах даних» (Knowledge Discovery in Databases) або інакше Data і Text Mining – «глибинний аналіз даних або текстів».

Якщо системи управління знаннями реалізують ідею збирання та накопичення всієї доступної інформації як із внутрішніх, так і з зовнішніх джерел, то Data і Text Mining, як уже було показано, дають змогу виявляти неочевидні закономірності в даних або текстах – так звані латентні (приховані) знання. Загалом ці технології також визначають як процес виявлення у «сирих» даних раніше невідомих, але корисних знань, необхідних для прийняття рішень. Системи цього класу дають змогу здійснювати аналіз великих масивів документів і формувати предметні покажчики понять і тем, висвітлених у цих документах.

Характерним завданням конкурентної розвідки, яке зазвичай включається до систем Text Mining, є знаходження виняткiв, тобто пошук об'єктів, що за своїми характеристиками суттєво вирізняються із загальної маси.

Ще один клас важливих завдань, що розв'язуються в межах технології Text Mining, – це моделювання даних, ситуаційний і сценарний аналіз, а також прогнозування.

Для оброблення та інтерпретації результатів Text Mining велике значення має візуалізація. Часто керівник компанії не завжди адекватно сприймає запропоновану йому аналітичну інформацію, особливо якщо вона не цілком збігається з його розумінням ситуації. У зв'язку з цим служба конкурентної розвідки повинна прагнути подавати інформацію у вигляді, адаптованому до індивідуального сприйняття замовника.

Візуалізація зазвичай використовується як засіб представлення контенту всього масиву документів, а також для ре-

алізації навігації по семантичних мережах під час дослідження як окремих документів, так і їхніх класів.

2.4. Моделі предметних областей

Важливим завданням конкурентної розвідки є виявлення неочевидних закономірностей і зв'язків із текстів веб-сторінок та виявлення їхніх взаємозв'язків, побудова матриць і графів взаємозв'язків.

Наявні доступні фактографічні бази даних структурованої інформації не завжди можуть прийти на допомогу експерту-аналітику. Для оперативного визначення фактів і сутностей, моделювання інформаційних зв'язків між ними найбільш перспективним підходом виявляється врахування інформації, знань, що містяться в неструктурованих текстових документах, зокрема, в Інтернеті.

Сьогодні, коли практично в усіх зацікавлених користувачів уже накопичено великий досвід роботи з традиційними інформаційно-пошуковими системами, виявилося очевидним, що факти або поняття, які шукають за допомогою таких систем, самі по собі часто безглузді. Наприклад, якщо користувача цікавлять інформаційні зв'язки Ощадбанку з іншими банками або приватними особами, то він не знає, які банки або прізвища йому вказати в запиті, а всі документи, що містять слово «Ощадбанк», вказати фізично неможливо. У таких випадках інформаційні зв'язки, кількість яких виходить за межі статистичного фону, як правило, відображають реальність.

Інтерпретують зазвичай не самі поняття або факти, а взаємозв'язки між ними. Важливим виявляється не стільки дослідження самих понять, скільки дослідження їхнього взаємозв'язку. Відомо, що саме взаємозв'язок сприяє розумінню мотиваційно-цільових особливостей, тобто користувача цікавить не поняття саме по собі, а поняття в оточенні, щоб одразу мати уявлення про предметну область, за необхідності спрямувати уточнюючий пошук у потрібному напрямку. Подібні рішення, реалізовані у вигляді «інформаційних портретів», що містять опорні слова, використовуються в таких системах, як InfoStream (infostream.ua), CyberAggregator.

База даних практично будь-якої традиційної інформаційно-пошукової системи може розглядатися у вигляді графа, вершинами якого виступають об'єкти – терми, поняття,

дескриптори та ін., а ребрами – їхні зв'язки. Разом з тим, основа пошуку в цих випадках – пошук вершин, тобто пошук об'єктів. Пошук за взаємозв'язками, ребрами, здається на перший погляд менш ефективним. Дійсно, якщо припустити, що в графі N вершин, то число ребер теоретично може становити $N(N - 1)/2$, тобто, якщо припустити, що вершин всього 100 тис., то ребер може виявитися близько 5 млрд., що відповідає досить великій базі даних навіть за сучасними поняттями. Разом з тим, якщо як вершини графа використовувати такі поняття, як імена людей і назви компаній з новинних документів, то виявляється, що відповідна матриця інцидентності є дуже розрідженою. Вимірювання показали, що за кількості окремих понять, вилучених із 5 млн. новинних документів, що дорівнює приблизно $N = 1{,}5$ млн., кількість зв'язків становила лише $\upsilon = 4$ млн.

Крім того, як показали експерименти, розподіл ступенів вершин (ступінь вершини – кількість ребер, що з неї виходять) у подібних графах – степеневий, що свідчить про так звану безмасштабність, тобто про те, що багато характеристик (зокрема, співвідношення кількості вершин і ребер) мають залишатися на одному рівні. Тому як основа побудови бази даних зв'язків виявляється технічно можливим використання ребер розглянутого графа – зв'язків між окремими поняттями.

Як масиви документальної інформації для такої системи можуть використовуватися дані, що надходять від систем контент-моніторингу, таких як InfoStream, X-SCIF, Attack Index, а також результати моніторингу спеціалізованих веб-служб, таких як бази даних біографій людей, організацій, служб працевлаштування тощо.

Інформаційні взаємозв'язки між поняттями виявляються шляхом обробки документальних масивів і можуть зберігатися у спеціальній базі даних. Набір понять, що використовується при побудові бази даних зв'язків, формується шляхом екстрагування даних із доступного користувачеві текстового масиву, що надає системі цілісності.

У корпоративній інформаційній інфраструктурі база даних зв'язків може використовуватися різним чином, наприклад, окремо, або її можливості можуть бути доповнені можливостями наявних повнотекстових та/або фактографічних баз даних. При цьому основним результатом роботи є побудова так званих

«карт зв'язків», а як побічний ефект, що реалізує «режим доведення», може розглядатися вилучення самих документів як джерел зв'язків.

При проєктуванні баз даних зв'язків використовуються перспективні рішення у сфері створення інформаційно-аналітичних систем, зокрема, теорія та технології глибинного аналізу текстів – Text Mining, у тому числі методи екстрагування інформації (Information Extraction), технології баз даних надвеликих обсягів (Big Data), концепція «складних мереж» (Complex Networks).

У межах теорії складних мереж вивчаються характеристики, пов'язані з топологією мереж, а також статистичні феномени, розподіл ваг окремих вершин (як яких можна розглядати сутності, поняття, факти) і ребер, ефекти протікання та провідності в мережах тощо.

Схематично можливі технологічні етапи формування бази даних зв'язків можна подати таким чином.

За допомогою програми-робота здійснюється сканування вибраних веб-ресурсів, які містять інформацію, що належить до об'єктів досліджень. Після цього здійснюється екстрагування необхідних користувачам понять, наприклад, найменувань брендів, компаній, електронних адрес тощо. Відібрані поняття та відповідні відношення між ними завантажуються до бази даних зв'язків, яка також містить посилання на документи-першоджерела. Засоби екстрагування понять, як правило, орієнтовані на обробку документів, сканованих з мережі Інтернет, поданих різними мовами.

Запропонований підхід до пошуку, природно, тягне за собою деякі особливості в реалізації архітектури бази даних зв'язків понять. Нині найпопулярнішою платформою для такої бази даних є графова СУБД Neo4j. Крім того, архітектура бази даних зв'язків має бути орієнтована на такі можливі застосування, як виявлення неявних зв'язків (не виявлених явно комплексом екстрагування понять), пошук окремих об'єктів, а також взаємозв'язок з наявними фактографічними базами даних.

2.5. Концепція Big Data

2.5.1. Поняття Big Data

Термін Big Data (великі дані) з'явився як новий термін і логотип у редакційній статті Кліффорда Лінча, редактора журналу Nature, 3 вересня 2008 року, який присвятив цілий спеціальний випуск одного з найвідоміших журналів темі «що можуть означати для сучасної науки набори великих даних». Наразі цей термін уже прижився і досяг піку свого використання. Тут слово «великі» було пов'язане не стільки з якоюсь кількістю, скільки з якісною оцінкою. Час підтвердив слушність виділення великих даних як окремого феномену. Сьогодні, згідно з дослідженнями агенції Gartner, термін Big Data уже перетнув пік знаменитої гартнерівської кривої Hype Cycle.

На Рис. 7 наведено статистику запитів користувачів до системи Google за словосполученням «Big Data» (сервіс Google Trends, https://trends.google.com/).

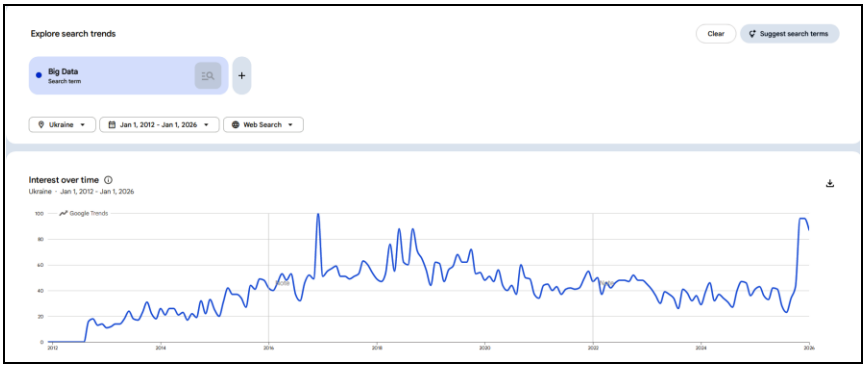


Рисунок 7 – Динаміка запиту “Big Data”

У 2012 році у статті[15] Дана Бойд та Кейт Кроуфорд сформулювали визначення Big Data як культурного, технологічного та наукового феномену, що включає в себе:

[15] Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". Information, Communication & Society. 15 (5): 662–679. doi:10.1080/1369118X.2012.67887

1. Технологію: максимізація обчислювальної потужності та складності алгоритмів для збору, аналізу, зв'язування та порівняння величезних наборів даних.
2. Аналіз: інтерпретація величезних наборів даних з метою виявлення патернів для формулювання економічних, соціальних, технічних та юридичних тверджень.
3. Міфологію: загальне переконання, що величезні набори даних репрезентують вищу форму знань та відомостей, які можуть генерувати інсайти, що раніше були неможливими, та з ореолом істинності, об'єктивності й точності.

Згідно з цим визначенням, великі дані – це термін, що позначає безліч наборів даних, настільки об'ємних і складних, що унеможливлює застосування наявних традиційних інструментів керування базами даних та додатків для їх обробки. Проблему становлять збір, очищення, зберігання, пошук, доступ, передача, аналіз та візуалізація таких наборів як цілісної сутності, а не локальних фрагментів. Як визначальні характеристики для великих даних відзначають «три V»: обсяг (англ. volume, у сенсі величини фізичного об'єму), швидкість (англ. velocity, що означає в даному контексті швидкість приросту та необхідність високошвидкісної обробки й отримання результатів), різноманітність (англ. variety, у сенсі можливості одночасної обробки різних типів структурованих і напівструктурованих даних). Провідною характеристикою тут є обсяг даних, який має бути розглянутий в аспекті застосунків.

- Чому обсяг даних перетворився на проблему? Зі збільшенням швидкодії комп'ютерів та обсягу пам'яті зростав і обсяг даних. Насправді, зростання даних навіть випереджало зростання швидкодії комп'ютерів, а лише деякі алгоритми лінійно масштабуються зі зростанням вхідних даних. Коротше кажучи, дані зростають швидше, ніж наша здатність їх обробляти. Таким чином, обсяг даних зростає швидше, ніж обчислювальні потужності. З цього випливає низка наслідків.
- Деякі методи та прийоми, що добре зарекомендували себе в минулому, тепер потребують перегляду або заміни, оскільки не масштабуються на сучасний обсяг даних.
- Алгоритми не можуть припускати, що всі вихідні дані вміщуються в оперативній пам'яті.

- Керування даними саме по собі стає нетривіальним завданням.
- Застосування кластерів або багатоядерних процесорів стає необхідністю, а не розкішшю.

Сучасність демонструє нам приклади жахливих розмірів генерованих сьогодні оцифрованих даних. Як стверджують гіганти IT-індустрії (EMC, Cisco, IBM, Google), у 2012 році у світі було згенеровано 2 зетабайти ($2 * 10^{21}$) або 2 тисячі екзабайтів, або 2 тисячі мільярдів гігабайтів інформації, а у 2025 році ця величина за даними Statista (www.statista.com) досягла 181 зетабайтів. Джерелами цієї лавини даних є численні цифрові пристрої, що концентрують і спрямовують у бездонні простори Інтернету продукцію людського розуму – твіти, пости у Facebook та ВКонтакті, запити до пошукових систем тощо, а також дані від сенсорів і контролерів мільйонів пристроїв, які вимірюють температуру та вологість, стан доріг і кондиціонерів та багато іншого, що сьогодні об'єднується терміном "Інтернет речей" IOT (Internet of Things).

Однак цьому перешкоджає не лише проблема кількості – обсяг даних, перша "V". Для великих даних, як уже було зазначено, важлива друга "V" – швидкість. Результати обробки великих даних мають бути отримані за час, що визначається проблемою, яку вирішують за їх допомогою. Це дасть можливість перетворити аналітику великих даних з інструменту, що відповідає на запитання "хто винен?", характерного для традиційних систем аналітики, на інструмент для отримання відповідей "що робити?". Аналітик у цьому випадку з лікаря-патологоанатома перетворюється на терапевта. Швидкість доступу до даних, швидкість їх обробки є важливим критерієм якості технологій, що входять у великі дані.

Нарешті, третя "V" – різноманітність даних говорить про те, що великі дані мають ефективно оброблятися незалежно від їхньої структурованості. Тут прийнято виділяти три основні види даних за ступенем їх структурованості.

Перший рівень – це звичні структуровані дані, які можуть бути представлені окремими й заздалегідь визначеними полями, у яких містяться біти, що мають різну семантику. Наприклад, усі таблиці мають у визначеному полі заданої довжини заголовки, в іншому заздалегідь заданому полі – один із фактів,

в іншому полі – інший із фактів, що визначають числові або текстові значення семантичних змінних, які містяться в заголовках. Структуровані дані добре зберігати в реляційних базах даних і керувати такими даними зручно, використовуючи спеціальну мову SQL – Structured Query Language. Незважаючи на свою поширеність, такі дані становлять лише 10% від усього обсягу згенерованих даних.

Другий рівень – це напівструктуровані (semistructured) дані. Дані такого типу мають структурні роздільники, але не можуть бути представлені у вигляді таблиці через відсутність частини атрибутів у різних даних. Прикладом таких даних можуть слугувати файли у форматі SGML – Standard Generalized Markup Language або BibTeX, у яких немає визначеної схеми зберігання даних, але семантичний зміст різних елементів даних може бути визначений за аналізом самого файлу. Іноді такі дані визначають як такі, що допускають самовпис. Багато даних, що зберігаються в Інтернеті, належать до напівструктурованих, дані бібліографічних описів публікацій, наукові дані.

Нарешті, неструктуровані дані, які за визначенням не можуть підпадати під раніше описані види. До них входять тексти, записані символами різних мов, записи звуків, нерухомі зображення, відеофайли, повідомлення електронної пошти, твіти, презентації та інша бізнес-інформація поза вивантаженнями баз даних. Вважається, що від 80 до 90 відсотків усіх даних в організаціях належать до неструктурованих даних. Нерідко до неструктурованих відносять і введені вище напівструктуровані дані. Іноді шкалу різноманітності розширюють, використовуючи цілу шкалу від структурованих даних до повністю неструктурованих. Вважатимемо показник варіативності даних нульовим для повністю неструктурованих даних і таким, що зростає до одиниці для добре структурованих із реляційних баз.

На думку учасників Всесвітнього економічного форуму 2012 року в Давосі, ті, хто осідлає тему інтелектуального аналізу великих даних, стануть господарями інформаційного простору. Цій темі було присвячено спеціальну доповідь на Форумі «Великі дані – великий вплив». Ключовий висновок доповіді – цифрові активи стають не менш значущим економічним активом, ніж золото чи валюта. Дослідження, проведені професо-

ром Бріньйолфсоном (E. Brynjolfsson) та двома його колегами у 2012 році, показали, що аналіз і прогнозування на основі великих даних береться на озброєння корпоративною Америкою. Вони вивчили 179 великих компаній і виявили, що ті з них, хто взяв на озброєння інтелектуальний аналіз великих даних за останні півтора року, отримали негайне покращення економічних показників на 5-6%.

Нині потреби суспільства роблять необхідною появу фахівців із великих даних як окремої професії. Назва цієї професії – Data Scientist – дослідник даних. Відомий журнал Harvard Business Review так озаглавив один зі своїх випусків: "Data Scientist: The Sexiest Job of the 21st Century" – дослідник даних – найпривабливіша робота 21 століття.

2.5.2. Техніки Big Data

Спочатку перелічимо ті функціональні операції над даними, методи їх зберігання та обробки. Звісно, цей перелік не вичерпує всієї різноманітності динамічно розвиваних технік, проте дозволяє побачити, що можна робити з великими даними для досягнення цілей, що стоять перед дослідником.

- Консолідація даних;
- Класифікація, кластеризація;
- Машинне навчання;
- Візуалізація.

Консолідація даних

Це цілий набір технік, спрямованих на вилучення даних з різних джерел, забезпечення їхньої якості, перетворення в єдиний формат та завантаження у сховище даних – «аналітичну пісочницю» (analytic sandbox) або «озеро даних» (data lake). Техніки консолідації даних різняться за видом аналітики, що виконується системою:

- Пакетна аналітика (batch oriented);
- Аналітика реального часу (real time oriented);
- Гібридна аналітика (hybrid).

При пакетній аналітиці періодично виконується вивантаження даних з різних джерел, дані аналізуються на наявність

збійних фрагментів, шуму та виконується їх фільтрація. При виконанні аналітики реального часу дані продукуються джерелами безперервно й утворюють набір потоків даних. Аналіз цих потоків і своєчасне отримання результатів у заданому темпі вимагають забезпечити асинхронне отримання даних у вигляді деяких повідомлень і маршрутизувати ці повідомлення в потрібні обробні вузли для обробки. Для гібридної аналітики, як правило, повідомлення даних мають бути не тільки промаршрутизовані на обробку, але й інтегровані в аналітичну пісочницю для подальшої обробки за результатами накопичення даних за значні інтервали часу.

Дані, отримані в результаті консолідації, мають відповідати певним критеріям якості. Якість даних - це критерій, що визначає повноту, точність, актуальність і можливість інтерпретації даних. Дані можуть бути високої та низької якості. Дані високої якості - це повні, точні, актуальні дані, які піддаються інтерпретації. Такі дані забезпечують отримання якісного результату: знань, які зможуть підтримувати процес прийняття рішень.

Сукупність процесів, що визначають консолідацію, називають ETL – Extraction-Transformation-Loading (Видобування-Перетворення-Завантаження). У застосунках бізнес-аналітики до процесів ETL включалися вельми складні перетворення даних, такі як квантування, що дозволяє знизити обсяг оброблюваних даних, нормалізація – процес приведення реляційних таблиць до канонічного вигляду або числових даних до єдиного масштабу, кодування даних – введення унікальних кодів для стиснення даних. У техніках великих даних зазвичай вважають, що необхідно працювати безпосередньо з брудними даними, оскільки нерідко саме характер збоїв може стати предметом аналізу, а стиснення даних є функцією власне аналітичних алгоритмів. Можливість же зберігання даних у вихідному вигляді мають надавати технічні засоби аналітичної системи. Якість великих даних нерідко важко оцінити методами формальних алгоритмів, і тоді вдаються до візуалізації на ранньому етапі дослідження. Крім оцінки якості та вибору методу препроцесингу, візуалізація може допомогти перейти до важливого етапу аналітики – вибору моделей, гіпотез для досягнення кінцевої мети – прийняття рішень.

Візуалізація

Техніка візуалізації є потужним методом інтелектуального аналізу даних. Як правило, її використовують для перегляду та верифікації даних перед створенням моделі, а також після генерації прогнозів. Візуалізація – це перетворення числових даних на певний візуальний образ, з метою спрощення сприйняття великих масивів інформації.

Для здійснення візуалізації слугують візуалізатори. Візуалізатори можуть бути або окремим застосунком, або плагіном, або частиною іншого застосунку. Можливості візуалізаторів дуже широкі. Наразі вони можуть представляти інформацію практично в усіх мислимих видах, аби лише аналітик міг сформулювати, що він хоче бачити.

Візуалізація текстів

Якщо дані є текстами природною мовою, то первинну допомогу в аналізі може надати візуалізація за допомогою розміченого тексту. Візуалізатор підраховує частоту згадування того чи іншого слова і присвоює словам умовну вагу, яка залежить від цієї частоти. Слова різної ваги при візуалізації мають різне розмічання, а отже, різне представлення на екрані. Одні слова виглядають більшими за інші. Цей тип візуалізації допомагає досліднику дуже швидко вхопити основні думки тексту.

Візуалізація кластерів

Однією з часто використовуваних візуалізацій є візуалізація кластерів. Кластерами називають групи в чомусь схожих або близьких за властивостями об'єктів. Алгоритми кластеризації, тобто розбиття множини об'єктів на групи, ми розглянемо нижче, а тут покажемо лише, як може бути візуалізована їхня робота. Більшість візуалізаторів підтримує алгоритми кластеризації і здатна розділяти дані на кластери. Зазвичай для візуального представлення кластерів для об'єктів з різних кластерів використовуються контрастні кольори.

Візуалізація асоціацій

Візуалізація асоціацій демонструє частоту, з якою ті чи інші елементи з'являються разом у наборі даних, за рахунок чого визначається структура організації даних (наприклад,

може йтися про те, які продукти часто продаються разом). Також можлива візуалізація інформації про силу асоціації даних.

Візуалізація гіпотез

Візуалізація гіпотез дозволяє показувати виявлені закономірності, що підтверджують висунуті гіпотези. Представлення інформації в різних візуалізаторах відрізняється. Наприклад, якщо рядки кругових 3D-діаграм відображають ознаки, використані класифікатором, то кожна кругова діаграма відображає ймовірність того, що величина ознаки або діапазон значень підходять для класифікації. На рисунку 2.10, поданому нижче, аналізується зарплата працюючого населення США. Візуалізатор відображає атрибути, які можуть впливати на класифікацію за зарплатою. Атрибути представлені рядами кругових тривимірних діаграм. Висота кругової діаграми (циліндра) показує кількість записів у даній категорії; колір показує, що зарплата більша або менша за 50 тис. дол. На кожен атрибут може бути кілька кругових діаграм, наприклад, для позначення статі (чоловіча/жіноча) є дві діаграми, а для віку – вісім діаграм. Їхня кількість залежить від кількості закономірностей, виявлених візуалізатором.

Візуалізація дерев рішень

Візуалізація дерев рішень дозволяє представити ієрархічно організовану інформацію у вигляді ландшафту та оглядати всю множину даних або їх частину у вигляді вузлів і гілок. Ландшафт може бути як двовимірним, так і тривимірним. Кількісні та реляційні характеристики даних стають видимими за допомогою ієрархічно з'єднаних вузлів.

Класифікація

Техніка класифікації є однією з базових методик інтелектуального аналізу великих даних. Її нерідко використовують при побудові моделі аналітичних систем поряд з ще однією технікою - кластеризацією. Класифікація - це розподіл об'єктів (спостережень, подій) дослідження за заздалегідь відомими класами на підставі подібності ознак. На відміну від класифікації, кластеризація виконує розподіл об'єктів (спостережень, подій) за невідомими заздалегідь класами.

Класифікація виконується відповідно до принципів машинного навчання з учителем (Supervised Machine Learning). Для проведення класифікації за допомогою математичних методів необхідно мати формальний опис об'єкта, яким можна оперувати, використовуючи математичний апарат класифікації. Кожен об'єкт (запис бази даних) повинен містити інформацію про деякі ознаки об'єкта.

Процес класифікації, як правило, зводиться до наступних кроків.

1. Набір вихідних даних (або вибірку даних) розбивають на дві множини: навчальну та тестову. Навчальна множина - множина, яка включає дані, що використовуються для конструювання моделі. Множина містить вхідні та вихідні (цільові) значення прикладів. Вихідні значення призначені для навчання моделі. Тестова множина також містить вхідні та вихідні значення прикладів. Тут вихідні значення використовуються для перевірки моделі.
2. Кожен об'єкт набору даних належить до одного з визначених класів. На цьому етапі використовується навчальна множина, на ній відбувається конструювання моделі. Отримана модель подається класифікаційними правилами, деревом рішень або математичними формулами.
3. Виконується оцінка правильності моделі. Відомі значення з тестової множини порівнюються з результатами використання отриманої моделі. Обчислюється рівень точності - відсоток правильно класифікованих об'єктів у тестовій множині.

Кластеризація

Техніка кластеризації є підходом до класифікації даних у випадку, коли заздалегідь невідомо, до якого класу має бути віднесений будь-який наявний об'єкт. Кластеризація здійснюється автоматичним знаходженням груп, на які мають бути розбиті аналізовані об'єкти. Такий процес може розглядатися як машинне навчання без учителя (Unsupervised Machine Learning). Відомо понад 100 різних алгоритмів.

Машинне навчання

Термін «машинне навчання», швидше за все, траплявся вам не раз. Хоча його нерідко використовують як синонім

штучного інтелекту, насправді машинне навчання – це один із його елементів. При цьому обидва поняття народилися в Массачусетському технологічному інституті наприкінці 1950-х років.

Машинне навчання (machine learning, ML) – клас методів штучного інтелекту, характерною рисою яких є не пряме вирішення задачі, а навчання в процесі застосування вирішень множини подібних задач. Для побудови таких методів використовуються засоби математичної статистики, чисельних методів, методів оптимізації, теорії ймовірностей, теорії графів, різні техніки роботи з даними в цифровій формі.

Розрізняють два типи навчання:

- Навчання за прецедентами, або індуктивне навчання, ґрунтується на виявленні емпіричних закономірностей у даних.
- Дедуктивне навчання передбачає формалізацію знань експертів та їх перенесення в комп'ютер у вигляді бази знань.

Дедуктивне навчання прийнято відносити до галузі експертних систем, тому терміни машинне навчання і навчання за прецедентами можна вважати синонімами.

Багато методів індуктивного навчання розроблялися як альтернатива класичним статистичним підходам. Багато методів тісно пов'язані з вилученням інформації (information extraction, information retrieval), інтелектуальним аналізом даних (data mining).

На відміну від традиційного ПЗ, яке чудово справляється з виконанням інструкцій, але не здатне до імпровізації, системи машинного навчання по суті програмують себе самі, самостійно розробляючи інструкції шляхом узагальнення відомих відомостей.

Класичний приклад – розпізнавання образів. Покажіть системі машинного навчання достатню кількість знімків собак з позначкою «собака», а також котів, дерев та інших об'єктів, позначених «не собака», і вона з часом почне добре відрізняти собак. І для цього їй не потрібно буде пояснювати, як саме вони виглядають.

Навчання з учителем і без

Згаданий вид машинного навчання називається навчанням з учителем. Це означає, що хтось познайомив алгоритм з величезним обсягом навчальних даних, переглядаючи результати та коригуючи налаштування доти, доки не було досягнуто потрібної точності класифікації даних, які система ще «не бачила». Це те саме, що натискати кнопку «не спам» у поштовій програмі, коли фільтр випадково перехоплює потрібне вам повідомлення. Чим частіше ви це робите, тим точнішим стає фільтр.

Типові задачі навчання з учителем – класифікація та прогнозування (або регресійний аналіз). Розпізнавання спаму та образів – задачі класифікації, а прогнозування котирувань акцій – класичний приклад регресії.

При навчанні без учителя система переглядає гігантські обсяги даних, запам'ятовуючи, як виглядають «нормальні» дані, щоб отримати можливість розпізнавати аномалії та приховані закономірності. Навчання без учителя корисне, коли ви точно не знаєте, що саме шукаєте – у цьому випадку систему можна змусити вам допомогти.

Системи навчання без учителя можуть виявляти закономірності у величезних обсягах даних набагато швидше, ніж люди. Саме тому банки використовують їх для виявлення шахрайських операцій, маркетологи – для ідентифікації клієнтів зі схожими атрибутами, а ПЗ безпеки – для розпізнавання шкідливої активності в мережі.

Приклади задач навчання без учителя – кластеризація та пошук правил асоціації. Перша застосовується, зокрема, для сегментації клієнтів, а на пошуку правил асоціації ґрунтуються механізми видачі рекомендацій.

Способи машинного навчання

Розділ машинного навчання, з одного боку, утворився в результаті поділу науки про нейронні мережі на методи навчання мереж і види топологій їхньої архітектури, з іншого боку – увібрав у себе методи математичної статистики. Зазначені нижче способи машинного навчання виходять з випадку використання нейромереж, хоча існують й інші методи, що використовують поняття навчальної вибірки – наприклад, дискримінантний аналіз, який оперує узагальненою дисперсією та

коваріацією спостережуваної статистики, або байєсівські класифікатори. Базові види нейромереж, такі як перцептрон і багатошаровий перцептрон (а також їхні модифікації), можуть навчатися як з учителем, так і без учителя, з підкріпленням і самоорганізацією. Але деякі нейромережі та більшість статистичних методів можна віднести тільки до одного зі способів навчання. Тому, якщо потрібно класифікувати методи машинного навчання залежно від способу навчання, буде некоректним відносити нейромережі до певного виду, правильніше було б типізувати алгоритми навчання нейронних мереж.

- Навчання з учителем – для кожного прецеденту задається пара «ситуація, необхідне рішення».
- Навчання без учителя – для кожного прецеденту задається тільки «ситуація», потрібно згрупувати об'єкти в кластери, використовуючи дані про попарну подібність об'єктів, та/або знизити розмірність даних.
- Активне навчання – відрізняється тим, що алгоритм, який навчається, має можливість самостійно призначати наступну досліджувану ситуацію, на якій стане відома правильна відповідь.
- Навчання з частковим залученням учителя (semi-supervised learning) – для частини прецедентів задається пара «ситуація, необхідне рішення», а для частини – тільки «ситуація».
- Трансдуктивне навчання – навчання з частковим залученням учителя, коли прогноз передбачається робити тільки для прецедентів з тестової вибірки.
- Багатозадачне навчання (multi-task learning) – одночасне навчання групі взаємопов'язаних задач, для кожної з яких задаються свої пари «ситуація, необхідне рішення».
- Багатоваріантне навчання (multiple-instance learning) – навчання, коли прецеденти можуть бути об'єднані в групи, у кожній з яких для всіх прецедентів є «ситуація», але тільки для одного з них (причому, невідомо якого) є пара «ситуація, необхідне рішення».
- Бустинг (boosting – покращення) – це процедура послідовної побудови композиції алгоритмів машинного навчання, коли кожен наступний алгоритм прагне ком-

пенсувати недоліки композиції всіх попередніх алгоритмів.

- Байєсівська мережа.

Алгоритми машинного навчання потребують даних, якомога більшої кількості даних з якомога ширшого набору джерел. Чим більше вони «живляться» цими даними, тим «розумнішими» стають і тим більший їхній потенціал у прийнятті рішень. І хмари дають ці великі дані.

Великі дані обіцяють нам знайти багато цінного в процесі цифрової трансформації, в той час як хмара пропонує будівельні блоки для цього процесу. Машинне навчання, своєю чергою, стало першим по-справжньому промисловим інструментом для масштабного освоєння цих нових цінностей. Привабливість машинного навчання в тому, що можливості його використання практично безмежні. Воно може застосовуватися скрізь, де важливий швидкий аналіз даних, і справити простотаки революційний ефект там, де важливо виявляти тенденції або аномалії в обширних наборах даних – від клінічних досліджень до сфери безпеки та контролю за дотриманням стандартів.

Обмеження машинного навчання

Кожна система машинного навчання створює власну схему зв'язків, являючи собою щось на кшталт чорної скриньки. Ви не зможете шляхом інженерного аналізу з'ясувати, як саме виконується класифікація, але це й не має значення, головне, щоб працювало.

Однак система машинного навчання хороша настільки, наскільки точні навчальні дані: якщо подати їй на вхід «сміття», то й результат буде відповідним. При неправильному навчанні або надто малому розмірі навчальної вибірки алгоритм може видавати невірні результати.

2.5.3. Технології та інструменти Big Data

Розглянемо базові технології та інструменти, які на сьогодні набули найбільшого поширення у відомих проєктах. Цей перелік не вичерпує всіх уже апробованих технологій і, тим більше, тих, що перебувають у розробці, однак він дає змогу отримати достатньо цілісне уявлення про те, «чим» сьогодні користуються

дослідники даних і якими інструментами необхідно володіти, щоб розгорнути проєкт із використанням великих даних.

Технології великих даних мають забезпечувати рішеннями та інструментами, що дають змогу реалізовувати описані вище техніки на значних обсягах різнорідних даних із необхідною швидкістю. Досягається це високою паралелізацією обчислень і розподіленим зберіганням даних. Незважаючи на потребу у значній обчислювальній потужності та пам'яті, як правило, розгортання програмних продуктів великих даних здійснюється на кластерах із комп'ютерів середнього або навіть низького класу (commodity computers). Це дає змогу масштабувати системи великих даних без залучення суттєвих витрат. Останнім часом для розгортання систем великих даних дедалі ширше застосовуються хмарні сервіси (cloud computing services). У разі імплементації системи в хмарі вузли обчислювального кластера реалізуються на віртуальних машинах хмарної інфраструктури та гнучко адаптуються до завдання, знижуючи витрати на використання. Це слугує додатковим чинником, що приваблює багатьох розробників будувати системи великих даних на хмарних платформах.

Найпопулярнішою технологією великих даних, що вважається де-факто стандартом для побудови систем аналітики, які працюють у пакетному режимі, є сукупність рішень та програмних бібліотек, об'єднаних під назвою Hadoop. Якщо великі дані надходять у вигляді високошвидкісних потоків і реагування системи має відбуватися з малою затримкою, то замість пакетної аналітики застосовується аналітика реального часу. Тут поки що не виникло де-факто стандартних підходів, і з-поміж найпопулярніших ми розглянемо технологію під назвою Storm.

Apache Hadoop

Під назвою Hadoop спільнота Apache просуває технологію, що базується на використанні спеціальної інфраструктури для паралельної обробки великих обсягів даних. Hadoop забезпечує середовище для функціонального програмування задач, автоматичного розпаралелювання робіт, зміщення обчислювального навантаження до даних. Hadoop створив Даг Каттінг – творець Apache Lucene, бібліотеки текстового пошуку, що широко використовується. Hadoop походить від Apache Nutch – системи

вебпошуку з відкритим кодом, яка сама по собі була частиною проєкту Lucene.

Проєкт Nutch було запущено у 2002 році. Працездатний оглядач та пошукова система з'явилися дуже швидко. Однак розробники зрозуміли, що їхня архітектура не масштабуватиметься на мільярди веб-сторінок. Допомога прийшла у 2003 році, коли було опубліковано статтю з описом архітектури GFS (Google File System) – розподіленої файлової системи, яка використовувалася в реальних проєктах Google [16].

У 2004 році була опублікована стаття, в якій компанія Google представила світові технологію MapReduce [17]. На початку 2005 року у розробників Nutch з'явилася працездатна реалізація MapReduce на базі Nutch, а до середини року всі основні алгоритми Nutch були адаптовані для використання MapReduce та NDFS. Можливості застосування NDFS та реалізації MapReduce у Nutch виходили далеко за межі пошуку, і в лютому 2006 року було утворено незалежний підпроєкт Lucene, що отримав назву Hadoop. Приблизно в той самий час Даг Каттінг вступив до компанії Yahoo!, яка надала команду та ресурси для перетворення Hadoop на систему, що працює у вебмасштабах (див. далі врізку «Hadoop у Yahoo!»). Результати були продемонстровані у лютому 2008 року, коли компанія Yahoo! оголосила, що використовуваний нею пошуковий індекс був згенерований 10000-ядерним кластером Hadoop.

Історія Hadoop безпосередньо пов'язана з розробкою Google File System (2003 рік), а потім реалізацією технології MapReduce (2004 рік). На основі цих компонентів у 2005 році з'явився додаток пошуку інформації Apache Nutch, який наступного року дав дорогу проєкту Apache Hadoop. Хоча Hadoop найчастіше асоціюють із MapReduce та розподіленою файловою системою (HDFS, раніше званою NDFS), цим терміном часто позначають цілу родину взаємопов'язаних проєктів, об'єднаних інфраструктурою розподілених обчислень та великомасштабної обробки даних. Усі базові проєкти, що розг-

[16] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, «The Google File System», октябрь 2003 г., http://labs.google.com/papers/gfs.html.

[17] Jeffrey Dean, Sanjay Ghemawat, «MapReduce: Simplified Data Processing on Large Clusters», декабрь 2004 года, http://labs.google.com/papers/mapreduce.html.

лядаються в книзі, ведуться фондом Apache Software Foundation, який надає підтримку спільноті проєктів з відкритим кодом – включаючи вихідний HTTP-сервер, від якого походить назва. З розширенням екосистеми Hadoop з'являються нові проєкти, не обов'язково ті, що знаходяться під управлінням Apache, але які надають додаткові функції Hadoop або утворюють абстракції вищого рівня на основі базової функціональності.

Нижче коротко перераховані проєкти Hadoop.

Common – набір компонентів та інтерфейсів для розподілених файлових систем та загального введення/виведення (серіалізація, Java RPC, структури даних).

Avro – система серіалізації для виконання ефективних міжмовних викликів RPC та довгострокового зберігання даних.

MapReduce – модель розподіленої обробки даних та виконавче середовище, що працює на великих кластерах типових машин.

HDFS – розподілена файлова система, що працює на великих кластерах стандартних машин.

Pig – мова управління потоком даних та виконавче середовище для аналізу дуже великих наборів даних. Pig працює в HDFS та кластерах MapReduce.

Hive – розподілене сховище даних. Hive управляє даними, що зберігаються в HDFS, та надає мову запитів на базі SQL (які перетворюються ядром часу виконання на завдання MapReduce) для роботи з цими даними.

HBase – розподілена стовпцево-орієнтована база даних. HBase використовує HDFS для організації зберігання даних і підтримує як пакетні обчислення з використанням MapReduce, так і точкові запити (довільне читання даних).

ZooKeeper – розподілений координаційний сервіс високої доступності. ZooKeeper надає примітиви, які можуть використовуватися для побудови розподілених додатків (наприклад, розподілені блокування).

Sqoop – інструмент ефективного масового пересилання даних між структурованими сховищами (такими, як реляційні бази даних) та HDFS.

Oozie – сервіс запуску та планування завдань Hadoop (включаючи завдання MapReduce, Pig, Hive та Sqoop).

Hadoop складається з чотирьох функціональних частин:

- Hadoop Common;
- Hadoop HDFS;
- Hadoop MapReduce;
- Hadoop YARN.

Hadoop Common – це набір бібліотек та утиліт, необхідних для нормального функціонування технології. До його складу входить спеціалізований спрощений інтерпретатор командного рядка.

Коли набір даних переростає ємність однієї фізичної машини, його доводиться розподіляти по кількох різних машинах. Файлові системи, що керують зберіганням даних у мережі, називаються розподіленими файловими системами. Оскільки вони працюють у мережевому середовищі, проектувальнику доводиться враховувати всі складнощі мережевого програмування, тому розподілені файлові системи складніші за звичайні дискові файлові системи. Наприклад, одна з найсерйозніших проблем – зробити так, щоб файлова система переживала збої окремих вузлів без втрати даних. Hadoop постачається з розподіленою файловою системою, яка називається HDFS (Hadoop Distributed Filesystem). Іноді – у старій документації чи конфігураціях або у неформальному спілкуванні – також зустрічається скорочення «DFS»; воно означає те саме. HDFS – основна файлова система Hadoop, якій присвячено цей розділ, але в Hadoop також реалізовано абстракцію узагальненої файлової системи, і ми принагідно розглянемо інтеграцію Hadoop з іншими системами зберігання даних (наприклад, локальною файловою системою та Amazon S3).

HDFS (Hadoop Distributed File System) – це розподілена файлова система для зберігання даних на безлічі машин у великих обсягах. Проєктувалася так, щоб забезпечувати:

- Надійне зберігання даних на дешевому ненадійному обладнанні;
- Високу пропускну здатність читання-запису;
- Потоковий доступ до даних;
- Спрощену модель узгодженості;
- Архітектуру, аналогічну Google File System.

Файлова система HDFS спроєктована для зберігання дуже великих файлів зі потоковою схемою доступу до даних у клас-

терах звичайних машин[18]. Розглянемо це твердження більш детально.

Надвеликі файли

Під «надвеликими» в цьому контексті маються на увазі файли, розмір яких становить сотні мегабайт, гігабайт і терабайт. Зараз існують кластери Hadoop, в яких зберігаються петабайти даних[19].

Потоковий доступ до даних

В основу HDFS закладено концепцію одноразового запису/багаторазового читання як найефективнішу схему обробки даних. Набір даних зазвичай генерується або копіюється з джерела, після чого з ним виконуються різні аналітичні операції. У кожній операції задіяна велика частина набору даних (або весь набір), тому час читання всього набору даних важливіший за затримку читання першого запису.

Звичайне обладнання

Hadoop не вимагає дорогого обладнання високої надійності. Система спроєктована для роботи на стандартному обладнанні (загальнодоступне обладнання, яке може бути придбане у багатьох фірм) з досить високою ймовірністю відмови окремих вузлів у кластері (принаймні, для великих кластерів). Технологія HDFS спроєктована таким чином, щоб у разі відмови система продовжувала роботу без скільки-небудь помітного переривання. Також слід виділити області застосування, для яких наразі HDFS підходить не найкращим чином (при тому, що в майбутньому ситуація може змінитися):

Швидкий доступ до даних

Додатки, що вимагають доступу до даних з мінімальною затримкою (у діапазоні десятків мілісекунд), погано поєдну-

[18] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. The Hadoop Distributed File System. Proceedings of MSST2010, May 2010.

[19] «Scaling Hadoop to 4000 nodes at Yahoo!», http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html.

ються з HDFS. Нагадаємо, що система HDFS оптимізована для забезпечення високої пропускної здатності передачі даних, за яку доводиться розплачуватися уповільненням доступу. HBase наразі краще підходить для організації доступу до даних з мінімальною затримкою.

Багаточисельні дрібні файли

Оскільки вузол імен зберігає метадані файлової системи в пам'яті, межа кількості файлів у файловій системі визначається обсягом пам'яті вузла імен. Як показує досвід, кожен файл, каталог і блок займають близько 150 байт. Таким чином, наприклад, якщо у вас є мільйон файлів, кожен з яких займає один блок, для зберігання інформації потрібно не менше 300 Мбайт пам'яті. Зберігання мільйонів файлів ще прийнятне, але мільярди файлів уже виходять за межі можливостей сучасного обладнання.

Множинні джерела запису, довільні модифікації файлів

Запис у файли HDFS може виконуватися тільки одним джерелом. Запис завжди здійснюється в кінець файлу. Підтримка множинних джерел запису або модифікації з довільним зміщенням у файлі відсутня. (Можливо, ці можливості будуть підтримуватися в майбутньому, але, швидше за все, вони будуть відносно неефективними).

В основі архітектури HDFS лежать вузли зберігання – сервери стандартної архітектури, на внутрішніх дисках яких зберігаються дані. Для всіх даних використовується єдиний адресний простір. При цьому забезпечується паралельне введення-виведення інформації з різних вузлів. Таким чином, гарантується висока пропускна здатність системи.

HDFS оперує на двох рівнях: простору імен (Namespace) та зберігання блоків даних (Block Storage Service). Простір імен підтримується центральним вузлом імен (Namenode), що зберігає метадані файлової системи та метаінформацію про розподіл блоків файлів.

Багаточисельні вузли даних (Datanode) безпосередньо зберігають файли. Вузол імен відповідає за обробку операцій файлової системи – відкриття та закриття файлів, маніпуляція з каталогами тощо. Вузли даних відпрацьовують операції із

запису та читання даних. Вузол імен та вузли даних забезпечуються веб-серверами, що відображають поточний статус та дозволяють переглядати вміст файлової системи.

У HDFS немає POSIX-сумісності. Не працюють Unix-команди ls, cp тощо. Для монтування HDFS в ОС Linux необхідні спеціальні інструменти, наприклад, HDFS-Fuse. Файли поблочно розподіляються між вузлами. Всі блоки в HDFS (крім останнього блоку файлу) мають однаковий розмір – від 64 до 256 Мб.

Для забезпечення стійкості до відмов серверів, кожен блок може бути продубльований на кількох вузлах. Коефіцієнт реплікації (кількість вузлів, на яких має бути розміщений кожен блок) визначається в налаштуваннях файлу. Файли в HDFS можуть бути записані лише один раз (модифікація не підтримується), а запис у файл в один час може вести тільки один процес. Таким простим чином реалізується узгодженість даних.

Hadoop MapReduce

MapReduce – це модель програмування, орієнтована на обробку даних. Ця модель проста, але не настільки, щоб у її контексті не можна було реалізувати корисні програми. Hadoop дозволяє запускати програми MapReduce, написані різними мовами; у цьому розділі ми розглянемо одну й ту саму програму, написану мовами Java, Ruby, Python та C++. Але найважливіше полягає в тому, що програми MapReduce паралельні за своєю природою, а отже, великомасштабний аналіз даних стає доступним для всіх, у кого в розпорядженні є достатньо комп'ютерів. Переваги MapReduce повною мірою проявляються в роботі з великими наборами даних, тож почнемо з розгляду одного з таких наборів.

Hadoop MapReduce – це найбільш популярна програмна реалізація моделі паралельної обробки великих обсягів даних шляхом поділу на незалежні задачі, що вирішуються функціями Map і Reduce. Алгоритм MapReduce отримує на вхід 3 аргументи: вихідну колекцію даних, Map функцію, Reduce функцію, і повертає результуючу колекцію даних.

Вихідними колекціями даних є набори записів спеціального виду. Це структура даних типу Ключ, Значення (KEY, VALUE). Користувачеві необхідно задати функції обробки Map і Reduce. Алгоритм сам піклується про сортування даних, запуск

функцій обробки, повторне виконання транзакцій, що впали, та багато про що ще. Результуюча колекція складається з результатів аналізу в легкому для інтерпретації вигляді.

Робота алгоритму MapReduce складається з трьох основних етапів: Map, Group та Reduce. Як перший етап над кожним елементом вихідної колекції виконується Map функція. Як правило, вона приймає на вхід один запис виду (KEY, VALUE) і повертає по ній деяку кількість нових записів (KEY1, VALUE1), (KEY2, VALUE2), ..., тобто перетворює вхідну пару {ключ: значення} на набір проміжних пар. Також ця функція відіграє роль фільтра – якщо для даної пари ніяких проміжних значень повертати не потрібно, функція повертає порожній список.

Можна сказати, що обов'язок Map функції конвертувати елементи вихідної колекції в нуль або кілька екземплярів об'єктів {ключ: значення}.

На другому етапі (Group) алгоритм сортує всі пари {ключ: значення} і створює нові екземпляри об'єктів, згруповані за ключем. Операція групування виконується всередині алгоритму MapReduce і користувачем не задається. Функція Reduce повертає екземпляри об'єкта {ключ: згорнуте значення}, які включаються до результуючої колекції.

Для прикладу розглянемо спрощений варіант задачі, що стоїть перед пошуковими системами. Припустимо, у нас є база даних сторінок в Інтернеті, і ми хочемо знати, скільки разів посилаються на кожну сторінку. Нехай є сторінка first.com із посиланнями
на first.com, second.com, third.com, сторінка second.com з двома посиланнями на first.com і сторінка third.com, на якій немає посилань взагалі.

Щоб мати єдиний формат вихідної колекції даних, визначимо вид кожної збереженої сторінки як (KEY = URL, VALUE = TEXT). Результати легко інтерпретуються.

Як базова мова написання функцій використовується Java. Для програмування існує популярний Hadoop плагін в Eclipse. Але можна обійтися і без нього: утиліти Hadoop streaming дозволяють використовувати як Map і Reduce будь-який виконуваний файл, що працює зі стандартним введенням-виведенням операційної системи (наприклад, утиліти командного рядка UNIX, скрипти Python, Ruby тощо), є також SWIG-сумісний прикладний інтерфейс програмування Hadoop

pipes на C++. Крім того, до складу дистрибутивів Hadoop входять реалізації різних обробників, які найчастіше використовуються в розподіленій обробці.

Особливістю Hadoop є переміщення обчислень якомога ближче до даних. Тому користувацькі задачі запускаються на тому вузлі, який містить дані для обробки. Після закінчення фази Map відбувається переміщення проміжних списків даних для обробки функцією Reduce.

Зауважимо тут, що крім Hadoop існують різні імплементації MapReduce. Спочатку MapReduce був реалізований компанією Google. Пізніше з'явилися інші реалізації алгоритму. Розвитком MapReduce від Google став проєкт з відкритим кодом – MySpace Qizmt (MySpace's Open Source Mapreduce Framework). Іншою відомою версією алгоритму є та, що реалізована в системі MongoDB.

Hadoop YARN (Yet Another Resource Negotiator) – платформа управління ресурсами системи, відповідальна за розподіл обчислювальних ресурсів серверів та розклад виконання користувацьких задач.

У перших версіях Hadoop MapReduce включав планувальник завдань JobTracker, починаючи з версії 2.0 (2013 р.) цю функцію перенесено в YARN. У ній модуль Hadoop MapReduce реалізовано поверх YARN. Програмні інтерфейси здебільшого збережено, однак повної зворотної сумісності немає.

YARN іноді називають кластерною операційною системою. Це зумовлено тим, що платформа відає інтерфейсом між апаратними ресурсами та різними додатками, що використовують обчислювальні потужності. Основою YARN є логічно самостійний демон – планувальник ресурсів (ResourceManager), що абстрагує всі обчислювальні ресурси кластера та керує їх наданням додаткам розподіленої обробки. Йому підзвітні численні менеджери вузлів (Node Manager), відповідальні за відстеження поточного статусу та навантаження окремих серверів.

Працювати під управлінням YARN можуть як MapReduce-програми, так і будь-які інші розподілені додатки, що підтримують відповідні програмні інтерфейси. YARN забезпечує можливість паралельного виконання кількох різних задач у рамках системи серверів.

Розробнику розподіленого додатка необхідно реалізувати спеціальний клас управління додатком (AppMaster), який від-

повідає за координацію завдань у рамках тих ресурсів, які надасть планувальник ресурсів. Планувальник ресурсів відповідає за створення екземплярів класу управління додатком та взаємодії з ними через мережевий протокол.

На основі Hadoop створено цілий ряд продуктів для обробки даних. Ось список лише найбільш популярних з них:

- Pig – високорівнева мова потоків даних для паралельного програмування;
- HBase – розподілена база даних, яка забезпечує зберігання великих таблиць;
- Cassandra – стійка до помилок, децентралізована база даних;
- Hive – сховище даних з функціями об'єднання даних та швидкого пошуку;
- Mahout – бібліотека методів машинного навчання та видобування знань.

Hadoop є дуже динамічною технологією, що розвивається. Тому найсвіжішу інформацію рекомендується отримувати в Інтернеті на сайті http://hadoop.apache.org/.

Storm – система потокової обробки

Storm є безкоштовною технологією та програмною реалізацією розподіленої обчислювальної системи реального. Ця система дозволяє будувати надійну обробку необмежених потоків даних подібно до того, як Hadoop робить це з пакетною обробкою. Storm застосовується для аналітики реального часу, онлайнового машинного навчання, безперервних обчислень, розподілених ETL та інших операцій з потоками великих даних.

Storm може інтегруватися з технологіями черг і баз даних, які вже використовуються, і не залежить від мови програмування. Основою Storm є Storm топології та Storm кластер. Кластер є об'єктом, подібним до Hadoop кластеру, а замість запуску MapReduce job тут запускаються Storm topologies. Jobs і Topologies мають ключову відмінність – перші в нормальному режимі завершують роботу, а другі обробляють повідомлення завжди. У Storm кластері є два типи вузлів: master node та worker nodes (рисунок 2.28). На master node запускається де-

мон, який називається Nimbus, подібний до JobTracker в Hadoop. Nimbus відповідальний за розподіл коду по робочих вузлах кластера, розподіл задач по машинах та запуск і зупинку робочих процесів. Кожен робочий процес виконує підмножину топології. Працююча топологія складається з багатьох робочих процесів, розподілених по багатьох машинах. Кожен робочий вузол (worker node) має демона під назвою Supervisor. Цей модуль слухає всі процеси на своїй машині та запускає і зупиняє їх з ініціативи Nimbus. Координація між Nimbus та всіма Supervisor проводиться через спеціальний кластер, який називається Zookeeper. Цей кластер також зберігає на своєму дисковому просторі стан всіх процесів, що дозволяє відновлювати після збою окремо будь-яку машину робочого кластера.

Щоб виконати обчислення в реальному часі на Storm, потрібно створити топологію (topologies) – граф обчислень. Кожен вузол у топології містить логіку процесингу і лінк між вузлами, який показує, як дані мають бути передані між вузлами.

Основною абстракцією в Storm є потік (stream). Потоком називається необмежена послідовність кортежів (tuples). Джерела потоків даних для обробки представлені в топології абстракцією, яка називається spout, а обробники потоків, які можуть виконувати функції, фільтрувати потоки, агрегувати або об'єднувати потоки даних, взаємодіяти з базами даних, називаються bolt.

Стек Elastic

За останні кілька років з'явилися різні системи для зберігання та обробки великих масивів даних. Серед них можна виділити проєкти екосистеми Hadoop, деякі бази даних (БД) NoSQL, а також пошукові та аналітичні системи на кшталт Elasticsearch. Hadoop і будь-яка база даних NoSQL мають свої переваги та області застосування.

Elastic Stack – це велика екосистема компонентів, які служать для пошуку та обробки даних. Основні компоненти Elastic Stack – це Kibana, Logstash, Beats, X-Pack та 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 чудовим засобом візуалізації та користувацьким інтерфейсом для Elastic Stack. Компоненти Logstash та Beats

дозволяють передавати дані в Elastic Stack. X-Pack надає потужний функціонал: можна налаштовувати моніторинг, додавати різні сповіщення, встановлювати параметри безпеки для підготовки вашої системи до експлуатації. Оскільки Elasticsearch є ядром Elastic Stack...

Elasticsearch – це високомасштабована розподілена пошукова система повнотекстового пошуку та аналізу даних, що працює в режимі реального часу. Утиліта дозволяє зберігати, шукати та аналізувати великі обсяги даних. Зазвичай використовується як базовий механізм/технологія, допомагаючи додаткам зі складними функціями пошуку. Elasticsearch є основним компонентом Elastic Stack.

Elasticsearch як серце Elastic Stack відіграє основну роль у пошуку та аналізі даних. Вона побудована на унікальній технології – Apache Lucene. Завдяки цьому Elasticsearch докорінно відрізняється від традиційних рішень для реляційних баз даних чи NoSQL. Нижче перераховані основні переваги використання Elasticsearch як сховища даних:

- неструктурованість, документоорієнтованість;
- можливість пошуку;
- можливість аналізу даних;
- підтримка користувацьких бібліотек та REST API;
- легке управління та масштабування;
- робота в псевдореальному часі;
- висока швидкість роботи;
- стійкість до помилок та збоїв.

Огляд компонентів Elastic Stack

Деякі компоненти універсальні, їх можна застосовувати без Elastic Stack або інших інструментів.

Elasticsearch

Elasticsearch зберігає всі ваші дані, надає можливості пошуку та аналізу в масштабованому вигляді. Ми вже розглядали переваги та причини використання Elasticsearch. Ви можете працювати з Elasticsearch без будь-яких інших компонентів, щоб оснастити свій додаток інструментами для пошуку та аналізу даних.

Щоб працювати з реляційними базами даних, потрібно розбиратися в таких поняттях, як рядки, стовпці, таблиці та схе-

ми. Elasticsearch та інші сховища, орієнтовані на документи, працюють за іншим принципом. Система Elasticsearch має чітку орієнтацію на документи. Найкраще для неї підходять JSON-документи. Вони організовані за допомогою різних типів та індексів. Далі ми розглянемо ключові поняття Elasticsearch:

- індекс;
- тип;
- документ;
- кластер;
- вузол;
- шарди та копії;
- розмітку та типи даних;
- інвертований індекс.

Індекс

Індекс – це контейнер, який в Elasticsearch зберігає документи одного типу та керує ними. Індекс може містити документи одного типу. Індекси в Elasticsearch приблизно аналогічні за структурою базі даних у реляційних базах даних. Продовжуючи аналогію, тип в Elasticsearch відповідає таблиці, а документ – запису в ній.

Тип

Типи допомагають логічно групувати або організовувати однотипні документи за індексами.

Зазвичай документи з найбільш поширеним набором полів групуються під одним типом. Elasticsearch не вимагає наявності структури, дозволяючи вам зберігати будь-які документи JSON з будь-яким набором полів під одним типом. На практиці слід уникати змішування різних відомостей в одному типі, таких як «клієнти» та «продукти». Має сенс зберігати їх у різних типах та з різними індексами.

Документ

Як уже було сказано, JSON-документи найкраще підходять для використання в Elasticsearch. Документ складається з кількох полів і є базовою одиницею інформації, що зберігається в Elasticsearch. Наприклад, у вас може бути документ, що відпо-

відає одному продукту, одному клієнту або одній позиції замовлення.

Документи містять кілька полів. У документах JSON кожне поле має певний тип. У прикладі з каталогом продуктів, який ми бачили раніше, були поля sku, title, description, price та ін. Кожне поле та його значення можна побачити як пару «ключ – значення» в документі, де ключ – це ім'я поля, а значення – значення поля.

Вузол

Elasticsearch – це розподілена система. Вона складається з безлічі процесів, запущених на різних пристроях у мережі та взаємодіючих з іншими процесами. У розділі 1 ми завантажили, встановили та запустили Elasticsearch. Таким чином ми запустили так званий одиничний вузол кластера Elasticsearch. Вузол Elasticsearch – це одиничний сервер системи, який може бути частиною великого кластера вузлів. Він бере участь в індексуванні, пошуку та виконанні інших операцій, що підтримуються Elasticsearch. Кожному вузлу Elasticsearch у момент запуску присвоюються унікальний ідентифікатор та ім'я. Кожному вузлу Elasticsearch відповідає основний конфігураційний файл, який знаходиться в підкаталозі налаштувань. Формат файлу YML (YAML Ain't Markup Language). Ви можете використовувати цей файл для зміни значень за замовчуванням, таких як ім'я вузла, порти, ім'я кластера.

На базовому рівні вузол відповідає одному запущеному процесу Elasticsearch. Він відповідає за управління відповідною йому частиною даних.

Кластер

Кластер містить один або кілька індексів і відповідає за виконання таких операцій, як пошук, індексування та агрегації. Кластер формується одним або кількома вузлами. Будь-який вузол Elasticsearch завжди є частиною кластера, навіть якщо це кластер одиничного вузла. За замовчуванням кожен вузол намагається приєднатися до кластера з ім'ям Elasticsearch. Якщо ви запускаєте кілька вузлів всередині однієї мережі без зміни параметра cluster.name у файлі config/elasticsearch.yml, вони автоматично об'єднуються в кластер.

Кластер складається з кількох вузлів, кожен з яких відповідає за зберігання своєї частини даних та управління нею. Один кластер може зберігати один або кілька індексів. Індекс логічно групує різні типи документів.

Шарди та копії

Шарди допомагають розподілити індекс по кластеру. Вони розподіляють документи з одного індексу по різних вузлах. Обсяг інформації, який може зберігатися в одному вузлі, обмежується дисковим простором, оперативною пам'яттю та обчислювальними можливостями цього вузла. Шарди допомагають розподіляти дані одного індексу по всьому кластеру і тим самим оптимізувати ресурси кластера.

Процес поділу даних по шардах називається шардуванням. Це невід'ємна частина Elasticsearch, необхідна для масштабованої та паралельної роботи з виконанням оптимізації:

- дискового простору по різних вузлах кластера;
- обчислювальної потужності по різних вузлах кластера.

Розподілені системи на кшталт Elasticsearch пристосовані до роботи навіть при неполадках обладнання. Для цього передбачені репліки шардів, або копії. Кожен шард індексу може бути налаштований таким чином, щоб у нього було деяке число копій або не було жодної. Репліки шардів – це додаткові копії оригінального або первинного шарда для забезпечення високого рівня доступності даних.

Розмітка та типи даних

Elasticsearch – неструктурована система, завдяки чому в ній можна зберігати документи з будь-якою кількістю полів і типів полів. У реальності дані ніколи не бувають абсолютно безструктурними. Завжди є певний набір полів, спільний для всіх документів цього типу. Фактично типи всередині індексів повинні створюватися на основі спільних полів. Зазвичай один тип документів всередині індексу містить кілька спільних полів.

Типи даних

Elasticsearch підтримує широкий набір типів даних для різних сценаріїв зберігання текстових даних, чисел, булевих, бінарних об'єктів, масивів, об'єктів, вкладених типів, геоточок,

геоформ та багатьох інших спеціалізованих типів даних, наприклад, адрес IPv4 та IPv6. У документі кожне поле має асоційований тип даних.

Logstash

Утиліта Logstash допомагає централізувати дані, пов'язані з подіями, такі як відомості з файлів реєстрації (логів), різні показники (метрики) або будь-які інші дані в будь-якому форматі. Вона може виконати обробку даних до того, як сформувати потрібну вам вибірку. Це ключовий компонент Elastic Stack, який використовується для збору та обробки ваших контейнерів даних.

Logstash – це компонент на стороні сервера. Його мета – виконати збір даних з величезної кількості джерел введення в масштабованому вигляді, обробити інформацію та відправити її за місцем призначення. За замовчуванням перетворена інформація надходить до Elasticsearch, але ви можете вибрати один з багатьох інших варіантів виведення. Архітектура Logstash базується на плагінах і легко розширюється. Підтримуються три види плагінів: введення, фільтрації та виведення.

Kibana

Kibana – це інструмент візуалізації для Elastic Stack, який допоможе вам наочно представити дані в Elasticsearch. Його також часто називають вікном в Elastic Stack. У Kibana пропонується безліч варіантів візуалізацій, таких як гістограма, карта, лінійні графіки, часові ряди та ін. Ви можете створювати візуалізації буквально парою клацань миші та досліджувати свої дані в інтерактивному вигляді. Крім того, є можливість створювати красиві панелі управління, що складаються з різних візуалізацій, ділитися ними, а також отримувати високоякісні звіти.

У Kibana також передбачені інструменти для управління та розробки. Ви можете керувати налаштуваннями X-Pack для забезпечення безпеки в Elastic Stack, а за допомогою інструментів розробника створювати та тестувати запити REST API. Kibana Console являє собою зручний редактор, який підтримує функцію автозавершення та форматування запитів під час їх написання.

REST означає Representational State Transfer. Це архітектурний стиль для взаємодії систем одна з одною. REST розвивався разом із протоколом HTTP, і майже всі системи, засновані на REST, використовують HTTP як свій протокол. HTTP підтримує різні методи: GET, POST, PUT, DELETE, HEAD та ін. Наприклад, GET призначений для отримання або пошуку чого-небудь, POST використовується для створення нового ресурсу, PUT може застосовуватися для створення або оновлення існуючого ресурсу, а DELETE – для безповоротного видалення.

Elastic Cloud

Elastic Cloud – це хмарний сервіс з управління компонентами Elastic Stack, що надається компанією Elastic (https://www.elastic.co/) автором та розробником Elasticsearch та інших компонентів Elastic Stack. Всі компоненти продукту (окрім X-Pack та Elastic Cloud) створені на базі відкритого коду. Компанія Elastic обслуговує всі компоненти Elastic Stack, проводить тренінги, виконує розробку та надає хмарні сервіси.

Окрім Elastic Cloud, є й інші хмарні рішення, доступні для Elasticsearch, наприклад Amazon Web Services (AWS). Основна перевага Elastic Cloud в тому, що він створений і обслуговується авторами Elasticsearch та інших компонентів Elastic Stack.

Як ви можете бачити, Elasticsearch та Elastic Stack можна використовувати для широкого спектру задач. Elastic Stack – це платформа з розширеним набором інструментів для створення комплексних рішень пошуку та аналітики. Вона підходить для розробників, архітекторів, бізнес-аналітиків та системних адміністраторів. Цілком можливо створити рішення на базі Elastic Stack, майже не вдаючись до написання коду, виключно за рахунок зміни конфігурації. Водночас система Elasticsearch дуже гнучка, отже, розробники та програмісти можуть будувати потужні додатки завдяки широкій підтримці мов програмування та REST API.

Sphinxsearch

Ще одна повнотекстова пошукова система для великих даних – Sphinxsearch. Sphinxsearch (від SQL Phrase Index) поширюється за ліцензією GNU GPL або, для версій 3.0+ без вихідних кодів. Відмінною особливістю є висока швидкість індексації та пошуку, а також інтеграція з існуючими СУБД

(MySQL, PostgreSQL) та API для поширених мов веб-програмування (офіційно підтримуються PHP, Python, Java; існують реалізовані спільнотою API для Perl, Ruby, .NET та C++). Офіційний сайт системи – http://sphinxsearch.com/.

Система Sphinxsearch має такі особливості:

- Висока швидкість індексації (до 10-15 МБ/сек на кожне ядро процесора);
- Висока швидкість пошуку (до 150–250 запитів на секунду на кожне ядро процесора з 1 000 000 документів);
- Велика масштабованість (найбільший відомий кластер індексує до 3 000 000 000 документів і підтримує понад 50 мільйонів запитів на день);
- Підтримка розподіленого пошуку;
- Підтримка кількох полів повнотекстового пошуку в документі (до 32 за замовчуванням);
- Підтримка кількох додаткових атрибутів для кожного документа (тобто групи, часові мітки тощо);
- Підтримка однобайтових кодувань та UTF-8;
- Підтримка морфологічного пошуку – наявні вбудовані модулі для англійської, російської та чеської мов; доступні модулі для французької, іспанської, португальської, італійської, румунської, німецької, голландської, шведської, норвезької, датської, фінської, угорської мов;
- Нативна підтримка існуючих СУБД PostgreSQL та MySQL, підтримка ODBC сумісних баз даних (MS SQL, Oracle тощо).

У 2017 році команда Manticore Software зробила форк Sphinxsearch 2.3.2, який назвали Manticore Search. За словами розробників, менеджмент системи Sphinxsearch не справлявся із супроводом системи, а саме, не виправлялися виявлені помилки, не реалізовувалися оголошені можливості, діалог користувачів з розробниками Sphinxsearch був ускладнений. Sphinx версії 3 вже можна сприймати як пропрієтарне рішення для обмеженого кола користувачів. Фактично нова версія системи (Manticore Search) вирішила багато з цих проблем, в тому числі, забезпечена підтримка коду в цілому, організовано сучасну взаємодію з користувачами Sphinxsearch та Manticore, реалізовано такі можливості, притаманні, зокрема, Elasticsearch:

реплікація, auto id, JSON інтерфейс, можливість створити/видалити індекс на льоту, наявність сховища документів, розвинених real-time індексів.

Neo4j

Neo4j – це графова система управління базами даних з відкритим кодом, мовою Java, з підтримкою транзакцій (ACID). Станом на 2015 рік вважається найпоширенішою графовою СУБД[20]. Розробник – американська компанія Neo Technology, розробка ведеться з 2003 року.

Дані зберігає у власному форматі, спеціалізовано пристосованому для представлення графової інформації, такий підхід у порівнянні з моделюванням графової бази даних засобами реляційної СУБД дозволяє застосовувати додаткову оптимізацію у випадку даних з більш складною структурою. Також стверджується про наявність спеціальних оптимізацій для SSD-накопичувачів, при цьому для обробки графа не потрібне його поміщення цілком в оперативну пам'ять обчислювального вузла, таким чином, можлива обробка досить великих графів. Основні області застосування: соціальні мережі, системи надання рекомендацій, виявлення шахрайства, картографічні системи.

Термінологія графових баз даних

- graph database, графова база даних – це база даних побудована на графах – вузлах та зв'язках між ними
- Cypher – це мова для написання запитів до бази даних Neo4j (приблизно, як SQL в MySQL)
- node, нода – об'єкт у базі даних, вузол графа. Кількість вузлів обмежена 2 в степені 35 ~ 34 мільярди
- node label, мітка ноди – використовується як умовний «тип ноди». Наприклад, ноди типу movie можуть бути пов'язані з нодами типу actor. Мітки нод – реєстрозалежні, причому *Cypher не видає помилок, якщо набрати назву не в тому регістрі.

[20] Yuan, D., Zhou, K. and Yang, C., 2023. Architecture and application of traffic safety management knowledge graph based on Neo4j. Sustainability, 15(12), p.9786.

- relation – це зв'язок між двома нодами, ребро графа. Кількість зв'язків обмежена 2 в степені 35 ~ 34 мільярди
- relation identifier, тип зв'язку в Neo4j. Максимальна кількість типів зв'язків 32767
- properties, властивості ноди – це набір даних, які можна призначити ноді. Наприклад, якщо нода – це товар, то у властивостях ноди можна зберігати id товару з бази MySQL
- node ID, ID ноди – унікальний ідентифікатор ноди. За замовчуванням, під час перегляду результату відображається саме цей ID.

Зберігання даних у Neo4j

Файл nodestore.db містить записи певного розміру, що містять інформацію про ноду:

Мітка, яка показує, чи запис активна;

Вказівник на перший зв'язок, який містить дана нода;

Вказівник на першу властивість, яку містить дана нода. Нода не містить власного ідентифікатора. Оскільки кожен запис у nodestore.db займає однакову кількість місця, можна розрахувати вказівник на ноду.

Файл relationshipstore.db також містить записи однакового розміру, які описують зв'язки, але вони складаються з наступних елементів:

- Мітка, яка показує, чи запис активна;
- Вказівник на ноду, яка містить цей зв'язок;
- Вказівник на ноду, до якої цей зв'язок направлено;
- Вид зв'язку;
- Вказівник на зв'язок, який стоїть попереду (в межах даної ноди);
- Вказівник на зв'язок, який стоїть позаду (в межах даної ноди);
- Вказівник на зв'язок, який стоїть попереду (в межах Ноди, в якій цей зв'язок направлено);
- Вказівник на зв'язок, який стоїть позаду (в межах Ноди, в якій цей зв'язок направлено);

- Вказівник на першу властивість даного зв'язку.

Як модель даних вибрано орієнтований граф властивостей:

- Містить вузли (nodes) та зв'язки (relationships).
- Вузли мають властивості (properties). Вузли можна розглядати як документи, що містять властивості у вигляді пар ключ-значення.
- Вузли можуть бути позначені однією або кількома мітками (labels). Мітки групують вузли, вказуючи роль, яку вони відіграють у наборі даних.

Одному вузлу можна приписувати кілька міток (оскільки вузли можуть відігравати кілька різних ролей у різних доменах). Зв'язки зв'язують вузли та структурують граф. Зв'язки іменовані (завжди мають одне ім'я) і направлені (завжди мають напрямок, початковий та кінцевий вузли). Зв'язки також можуть містити властивості. Це дозволяє ввести додаткові метадані в графи, алгоритми, додати додаткову семантику зв'язкам, обмежувати запити в режимі реального часу.

Основні транзакційні можливості – підтримка ACID та відповідність специфікаціям JTA, JTS та XA. Інтерфейс програмування додатків для СУБД реалізований для багатьох мов програмування, включаючи Java, Python, Clojure, Ruby, PHP, також реалізовано API у стилі REST. Розширити програмний інтерфейс можна як за допомогою серверних плагінів, так і за допомогою некерованих розширень (unmanaged extensions); плагіни можуть додавати нові ресурси до REST-інтерфейсу для кінцевих користувачів, а розширення дозволяють отримати повний контроль над програмним інтерфейсом, і можуть містити довільний код, тому їх слід використовувати з обережністю.

У СУБД використовується Cypher – декларативна мова запитів до графів. Синтаксис цієї мови схожий на синтаксис SQL. Підтримуються операції зі створення, вибірки, оновлення, видалення даних. Cypher описує графи, використовуючи специфікацію за зразком – використовується проста форма ASCII-графіки, користувач малює частину графа, яка його цікавить, за допомогою ASCII символів; вершини беруться в дужки, їх мітки прописуються після «:»; для створення кількох вузлів їх слід перерахувати через »,«; зв'язки відображаються стрілками

(-> і <-), а назви зв'язків вказуються всередині квадратних дужок після «:»; властивості вузлів та зв'язків (пари ключ-значення) прописуються у фігурних дужках.

Мова запитів Cypher – найпоширеніша мова запитів до графових баз даних, що зумовлено її використанням у СУБД Neo4j. Cypher є декларативною мовою і дозволяє створювати, оновлювати та видаляти вершини, ребра, мітки та властивості, а також керувати індексами та обмеженнями. Для отримання даних зі сховища використовується запит, що містить шаблон фільтрації, який дозволяє отримувати:

- (n)->(m) – всі направлені ребра з вершини n у вершину m;
- (n:Person) – всі вершини з міткою Person;
- (n:Person:Russian) – всі вершини, що мають обидві мітки Person та Russian;
- (n:Person {name:{value}}) – всі вершини з міткою Person і відфільтровані за додатковою властивістю;
- (n:Person) -> (m) – ребра між вершинами n з міткою Person та m;
- (n)--(m) – всі ненаправлені ребра між вершинами n та m.

Запити в Neo4j можна робити й іншими способами, наприклад, безпосередньо через Java API та мовою Gremlin, створеній у проєкті з відкритим кодом TinkerPop. Cypher є не тільки мовою запитів, але й мовою маніпулювання даними, оскільки надає функції CRUD для графового сховища.

Gephi

Gephi – це пакет програмного забезпечення з відкритим кодом для аналізу та візуалізації графів (мереж). Gephi (https://gephi.org/) - це наразі найпопулярніша програма візуалізації та аналізу мереж і графів («мережевих графів»). Gephi забезпечує швидке компонування, ефективну фільтрацію та інтерактивне дослідження даних, а також є одним з найкращих варіантів для візуалізації великомасштабних мереж. Gephi - це мультиплатформне програмне забезпечення, яке поширюється з відкритим кодом згідно з ліцензіями CDDL 1.0 та GNU General Public License v3. За адресою https://gephi.org/ доступні версії для Mac OS X, Windows та Linux та вихідні коди.

Gephi активно використовується в цілій низці академічних дослідницьких проєктів, зокрема соціологічних; також швидко отримав популярність серед журналістів. Зараз його користувацьке середовище значно розширилося - за допомогою цього пакета можна займатися будь-якою темою мережевого аналізу. Gephi використовувався, серед іншого, для візуалізації глобальної зв'язності контенту New York Times і вивчення мережевого трафіку Twitter під час соціальних хвилювань; Gephi надихав створення LinkedIn InMaps і був використаний для візуалізації цілої мережі Truthy.

Gephi дозволяє обробляти графові структури досить великих обсягів (до 1 млн вузлів) на персональному комп'ютері за рахунок ефективних алгоритмів. Розробники Gephi описують цю програму як "щось на кшталт Photoshop, але для даних".

Програма включає безліч різних алгоритмів компонування (розміщення графіків на площині) і дозволяє налаштовувати кольори, розміри та мітки в графах. Gephi є інтерактивним програмним забезпеченням і надає засоби для виявлення спільнот, а також надається можливість розрахунку найкоротших шляхів або відносної відстані від будь-якого вузла до даного вузла. Плагіни від Gephi дозволяють розширювати її функціональність та додавати нові алгоритми, макети та інструменти вимірювань. Gephi має багатопотокову схему обробки даних і, таким чином, дозволяє виконувати кілька видів аналізу одночасно.

Інтерфейс користувача системи Gephi включає три основні розділи (вікна):

- «Лабораторія даних»: тут зберігаються всі вихідні дані про мережу, а також додаткові розрахункові значення;
- «Обробка даних»: тут відбувається більша частина операцій користувача, зокрема, ручне редагування мереж, тестування макетів, встановлення фільтрів;
- «Попередній перегляд»: тут уточнюється форма виведення графа, як правило, за допомогою набору інструментів граф доопрацьовується, в тому числі, і з естетичної точки зору. У цьому ж вікні реалізовано виклик експорту графа у формати PDF, PNG та SVG.

Програма включає безліч різних алгоритмів компонування (розміщення графіків на площині) і дозволяє налаштовувати кольори, розміри та мітки в графах. Gephi є інтерактивним програмним забезпеченням і надає засоби для виявлення спільнот, а також надається можливість розрахунку найкоротших шляхів або відносної відстані від будь-якого вузла до даного вузла. Плагіни від Gephi дозволяють розширювати її функціональність та додавати нові алгоритми, макети та інструменти вимірювань. Gephi має багатопотокову схему обробки даних і, таким чином, дозволяє виконувати кілька видів аналізу одночасно.

Ці три основні розділи охоплюють безліч вкладок, які дозволяють користувачеві реалізовувати окремі функції. Нижче розглядається кожне з основних та вторинних вікон - розділів та вкладок.

При аналізі великих і щільних мереж, швидке компонування (упорядкування вузлів графів) є вузьким місцем, оскільки більшість складних алгоритмів компонування є вимогливими до параметрів процесора, пам'яті та часу виконання. Водночас, Gephi постачається з ефективними алгоритмами компонування, такими як Yifan-Hu, Force-directed. Зокрема, алгоритм Yifan-Hu є ідеальним варіантом для застосування після інших, більш швидких і грубих алгоритмів. У той час, як більшість із запропонованих у Gephi методів можуть виконуватися протягом прийнятного часу, поєднання, наприклад, OpenOrd та Yifan-Hu, дає найбільш якісні візуальні представлення. Звісно, правильна параметризація будь-якого алгоритму компонування може впливати як на роботу, так і на результат візуалізації.

Gephi дозволяє завантажувати дані мереж у форматах GEXF, GDF, GML, GraphML, Pajek (NET), GraphViz (DOT), CSV, UCINET (DL), Tulip (TPL), Netdraw (VNA) та таблиць Excel. Крім того, Gephi дозволяє експортувати дані мереж у форматах JSON, CSV, Pajek (NET), GUESS (GDF), Gephi (GEFX), GML та GraphML. Завдяки цьому Gephi може взаємодіяти з іншими системами аналізу та візуалізації графів.

2.6. Математичні основи

У даній главі основну увагу приділено розпізнаванню інформаційних операцій на основі вивчення динамічних власти-

востей інформаційних потоків у глобальних комп'ютерних мережах, зокрема в мережі Інтернет.

Для дослідження інформаційних потоків в Інтернеті, тобто потоку повідомлень, які публікуються на сторінках веб-сайтів, у соціальних мережах, блогах тощо, має застосовуватися сучасний інструментарій. Так, відомі методи узагальнення інформаційних масивів (класифікація, фазове укрупнення, кластерний аналіз тощо) вже не завжди придатні навіть для адекватного кількісного відображення процесів, що відбуваються в інформаційному просторі[21].

Кількісний аналіз динаміки інформаційних потоків, які генеруються в Інтернеті, стає сьогодні одним із найбільш інформативних методів дослідження актуальності тих чи інших тематичних напрямів. Ця динаміка зумовлена різноманітними якісними чинниками, багато з яких не піддаються точному опису. Однак загальний характер часової залежності кількості тематичних публікацій в мережі Інтернет все ж таки допускає побудову математичних моделей, їх дослідження та прогнозування. Спостереження часових залежностей обсягів мережевих інформаційних потоків переконливо свідчать про те, що механізми їхньої генерації та поширення, очевидно, пов'язані зі складними нелінійними процесами.

Тут буде основну увагу приділено завданням OSINT з розпізнавання інформаційних операцій на основі вивчення динамічних властивостей інформаційних потоків у комп'ютерних мережах, зокрема, в Інтернеті.

Для вивчення інформаційних потоків у Інтернеті, тобто. потоку повідомлень, які публікуються на сторінках веб-сайтів, у соціальних мережах, блогах тощо, має застосовуватися сучасний інструментарій. Так відомі методи узагальнення інформаційних масивів (класифікація, фазове укрупнення, кластерний аналіз тощо) вже не завжди придатні навіть для адекват-

[21] Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах. - Київ: Наукова думка, 2009, - 295 с. ISBN 978-966-00-0973-9

ного кількісного відображення процесів, що відбуваються в інформаційному просторі[22].

Кількісний аналіз динаміки інформаційних потоків, що генеруються в Інтернеті, стає сьогодні одним із найбільш інформативних методів дослідження актуальності тих чи інших тематичних напрямків. Ця динаміка обумовлена різноманітними якісними факторами, багато з яких не піддаються точному опису. Однак загальний характер тимчасової залежності кількості тематичних публікацій у мережі Інтернет все ж таки допускає побудову математичних моделей, їх дослідження, прогнозування. Спостереження часових залежностей обсягів мережевих інформаційних потоків переконливо свідчать, що механізми їх генерації та поширення, очевидно, пов'язані зі складними нелінійними процесами. Саме цій темі присвячено цей розділ.

Для аналізу часових рядів, що відображають залежність обсягів інформаційних потоків від часу, використовують різноманітні методи та підходи. У цьому виявляється, що це підходи взаємопов'язані і більше, ключову роль грає поняття кореляції. Виклад побудовано навколо схеми, показаної на Рис. 5.1, причому особливу увагу приділено взаємозв'язкам.

2.6.1 Часові ряди

Часовий ряд – це набір значень, що спостерігаються, упорядкованих за часом. Далі розглядатимуться дискретні часові ряди, значення яких фіксувалися через рівні проміжки часу. Позначатимемо такий часовий ряд $x_1, x_2, \ldots, x_T$ або скорочено $\{x_t\}_{t=1}^{T}$. При цьому розуміємо, що, фіксація значень ряду відбувалося через рівний проміжок часу h: $t_0, t_0 + h, t_0 + 2h, \ldots, t_0 + (T-1)h.$.

[22] Information Operations Recognition. From Nonlinear Analysis to Decision-Making. A. Dodonov, D. Lande, V. Tsyganok, O. Andriichuk, S. Kadenko, A. Graivoronskaya. – LAP Lambert Academic Publishing, 2019. – 292 p.

Якщо значення часового ряду однозначно визначаються математичним відношенням (наприклад, $x_t = A \cdot \sin(vt)$), то мова йде про статистичний часовий ряд. Такі ряди будуть розглядатися далі. Аналізуючи часові ряди, ми будемо розглядати їх як реалізацію стохастичного процесу.

Далі як приклад використовуватимуться три часових ряди, які були отримані за допомогою популярного мережевого сервісу Google Trends. Ці часові ряди відображають рівень інтересу до Дональда Трампа, Хілларі Клінтон та "російських хакерів" з серпня 2016 року по квітень 2017 року. Часові ряди, отримані за допомогою Google Trends, показують динаміку популярності пошукового запиту. Максимальна точка на графіку дорівнює 100 і відповідає даті, коли запит був найбільш популярним, а інші точки на графіку визначаються у відсотковому співвідношенні до максимуму. Усі три часових ряди показані на Рис. 8. Для простоти посилань на дані ряди надалі позначимо їх Т (Д. Трамп), К (Х. Клінтон), Х («російські хакери»).

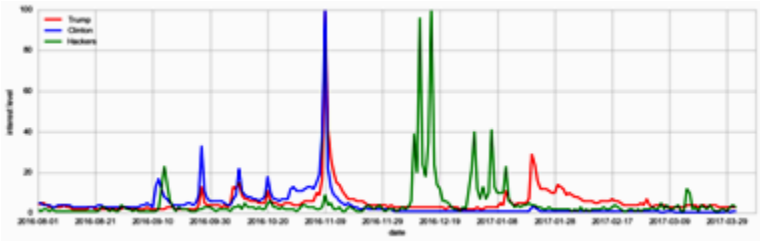


Рисунок 8 – Часові ряди, що відображають зацікавленість до Дональда Трампа (Т), Хілларі Клінтон (К) та "російських хакерів" (Х) з 1 серпня 2016 року по 1 квітня 2017 року з Google Trends[23].

У деяких випадках корисно розглянути гладкішу версію вихідного часового ряду. Згладжування допомагає виявити суттєві тенденції в динаміці ряду, приховавши при цьому шум

[23]Information Operations Recognition. From Nonlinear Analysis to Decision-Making A. Dodonov, D. Lande, V. Tsyganok, O. Andriichuk, S. Kadenko, A. Graivoronskaya - LAP Lambert Academic Publishing, 2019. – 292 p. ISBN-13: 978-620-0-27697-1,

та різні особливості, що виявляються при невеликих масштабах. Існують різноманітні методи згладжування. Найбільш простий спосіб згладжування – це обчислення ковзного середнього. Просте ковзне середнє дорівнює середньому арифметичному значенню елементів ряду з інтервалу заданої довжини, а саме:

$$SMA_t = \frac{1}{w}\sum_{i=0}^{w-1} x_{t-i},$$

де w – ширина інтервалу згладжування (кількість елементів, за якими розраховується середня), SMA_t – значення простого ковзного середнього у точці t. Отримане значення SMA_t відноситься до середини інтервалу згладжування, тому згладжений ряд y_t може бути визначений як $y_t = SMA_{t+\left[\frac{w}{2}\right]}$.

При використанні згладжування ковзним середнім, чим більше ширина інтервалу згладжування, тим більш гладкою вийде функція. Рис. 9 показано, як виглядає згладжений ряд T зі збільшенням значення w.

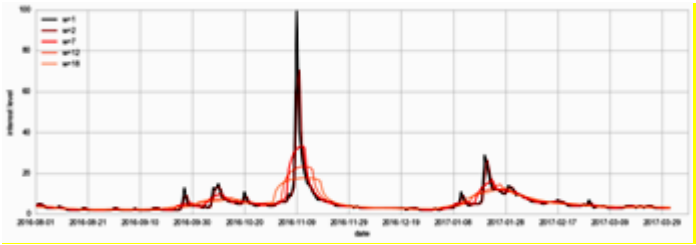


Рисунок 9 – Вихідний ряд T та згладжений простим ковзним середнім з шириною інтервалу згладжування 2, 7, 12, 18

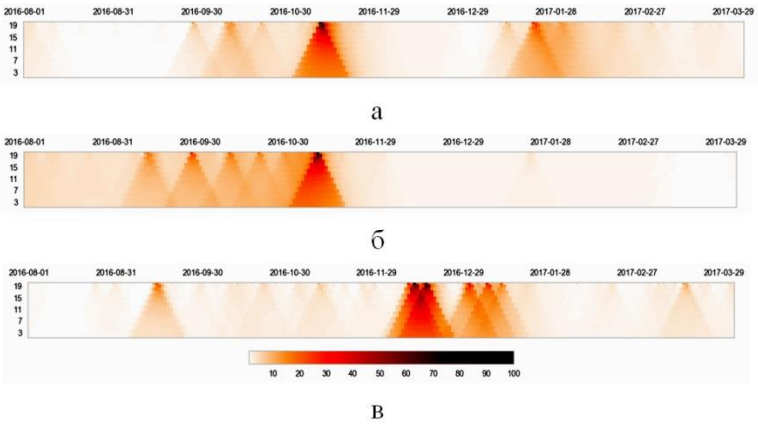
Результати згладжування ряду можна продемонструвати на графіку, де вісь абсцис відповідає часовій вісі, а вздовж вісі ординат відкладено ширину згладжуючого інтервалу. На графіку показані значення $y_t^{(w)}$ – тобто існують елементи згладженого ряду у точці t при використанні інтервалу шириною w (Рис. 10). При обчисленні простого ковзного середнього

всі точки, які потрапили в згладжуючий інтервал, мають однакову вагу.

Зрозуміло, що можна використовувати нерівні ваги. Отже, ми переходимо до визначення зваженого ковзного середнього значення:

$$WMA_t = \frac{1}{w}\sum_{i=0}^{w-1} a_i x_{t-i},$$

де $\sum_{i=0}^{w-1} a_i = 1.$



а

б

в

Рисунок 10 – Значення згладжених часових рядів *T* (а), *K* (б) та *X* (в) за допомогою простого ковзного середнього в залежності від ширини згладжуючого інтервалу

Як приклади розглядаються часові ряди *T*, *K* і *X*, які мають тижневу періодичність. Відомо, що публікація повідомлень новин часто відбувається з тижневою періодичністю, а також активність користувачів різниться в будні і вихідні дні. Це характерна властивість багатьох процесів в інформаційному просторі. Для того щоб виключити періодичну компоненту з рядів, згладимо їх за допомогою простого середнього ковзного з інтервалом шириною 7 (число днів у тижні) відповідно до формули:

$$x_t^{New} = \frac{x_{t-3} + x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2} + x_{t+3}}{7},$$

де x_t – вихідні значення ряду, x_t^{New} – нове значення ряду в момент часу t. На Рис. 11 наведені згладжені часові ряди.

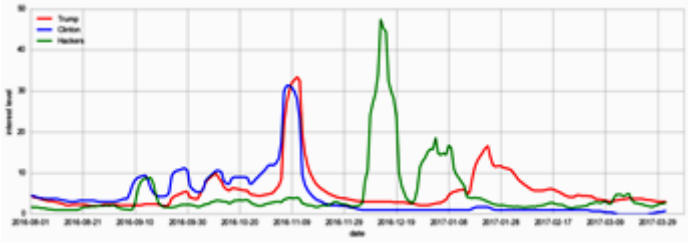


Рисунок 11 – часові ряди T, K і X, згладжені за допомогою простого середнього ковзного з інтервалом довжини 7

Ще один часто використовуваний метод згладжування рядів – це експоненціальне згладжування. Попередні значення ряду враховуються з вагою, що експоненціально зменшується. Позначимо елементи згладженого ряду y_t, і одразу визначимо $y_0 = x_0$. Наступні елементи ряду y_t отримуються за допомогою рекурсивної формули

$$y_t = \alpha x_t + (1-\alpha) y_{t-1},$$

де $0 < \alpha < 1$ – коефіцієнт згладжування. Очевидно, що при $\alpha = 1$ ряд, що отримується, y_t співпадає з вихідним x_t. Таким чином, якщо значення α близьке до 1, то найбільшиа вага при визначенні y_t призначається відповідному x_t, а попередня історія малозначуща. З іншого боку, якщо б α дорівнювалось 0, то весь ряд y_t згладився б до одного значення $y_t = y_0$. Тобто при α близькому до 0 попередня історія ряду враховується з більшою вагою, ніж поточне значення.

На Рис. 12 наведено ряд T, а також відповідні згладжені ряди при різних значеннях параметру α.

Як і у випадку з простим ковзним середнім, продемонструємо результати згладжування ряду на графіку. В даному випадку вздовж осі ординат відкладемо параметр α (Рис. 13). На графіках наведені значення $y_t^{(\alpha)}$ – значення у точці t згладженого з параметром α вихідного ряду.

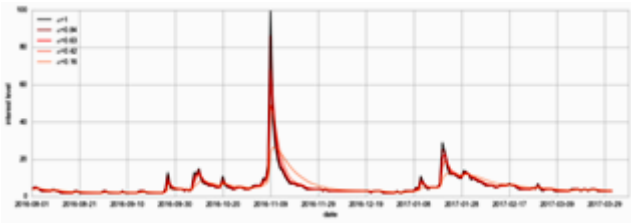


Рисунок 12 – Початковий ряд T та згладжений за допомогою експоненціального згладжування з параметром, рівним 0,84, 0,63, 0,42, 0,16

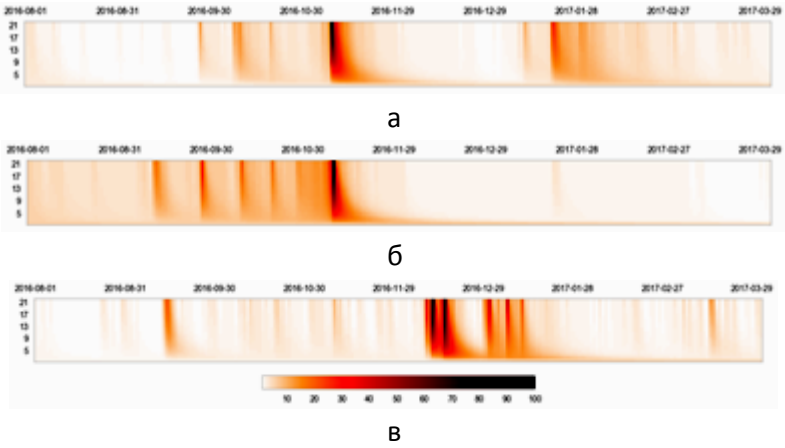


Рисунок 13 – Значення експоненціального згладжених часових рядів T(а), K(б) та X(в) залежно від параметра α. Вздовж осі абсцис відкладено час, а вздовж осі ординат – параметр α

Аналітика в Elasticsearch призначена для отримання повної картини даних. При пошуку детально розглядається декілька записів, аналітика ж дозволяє поглянути на дані ширше і згрупувати їх різними способами.

Реалізація агрегації у запитах

Агрегація в Elasticsearch, яка визначається елементом aggregations, дозволяє отримувати узагальнені за деякою ознакою дані. Всі запити на агрегацію мають вигляд:

```
GET /<index_name>/<type_name>/_search
{
"query": { ... тип запиту ... },
"aggregations": {
... тип агрегації ...
}
},
"size": 0
}
```

Елемент aggregations повинен містити фактичний запит агрегації. Тіло запиту залежить від бажаного типу агрегації. Необов'язковий елемент query вказує контекст агрегації, тобто якщо необхідно обмежити контекст агрегації, необхідно вказати елемент query. Агрегація буде враховувати всі документи даного індексу і типу, якщо не вказано елемент query (ми можемо вважати його рівним запитом match_all, якщо немає іншого запиту). Наприклад, цей параметр вказується, якщо необхідно, щоб агрегація працювала не за всіма даними, а тільки по певними документами, які відповідають конкретним умовам.

Запит фільтрує документи, які будуть оброблені елементом aggregations. Елемент size вказує, скільки відповідних пошуку документів повинно бути повернуто у відповіді. Значення за замовчуванням становить 10. Якщо значення size не вказано, відповідь буде містити не більше десяти релевантних документів. Зазвичай, якщо необхідно отримати тільки перші тестові результати агрегації, необхідно присвоювати елементу size значення 0, щоб не отримувати інших результатів.

Для отримання динаміки кількості публікацій за запитом в Elasticsearch, необхідно здійснити агрегацію даних, які відповідають тематиці, визначеній первинним запитом за полем, що відповідає значенню дати і часу. Конкретно, щоби здійснити агрегацію за датами всіх документів із індексу hb,

що відповідають запиту Samsung, за полем часу, вводимо запит у форматі JSON:

```
curl -XGET 'http://localhost:9200/hb/_search?pretty=true' –H 'Content-Type: application/json' -d '
{
"query":
     {"multi_match":
       {"query" : "Samsung",
       "fields":
          ["title","textBody"]
       }
     },
"aggregations":
     { "dates_with_holes":
       {  "date_histogram":
          {  "field": "pubDate",
             "interval": "day",
             "min_doc_count": 0
          }
       }
     },
"size": 0
}'
```

У результаті виконання цього запиту отримуємо вихідні дані для побудови діаграми – відповідь такого вигляду:

```
{
 "took" : 2,
 "timed_out" : false,
 "_shards" : {
  "total" : 1,
  "successful" : 1,
  "skipped" : 0,
  "failed" : 0
 },
 "hits" : {
```

```
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "dates_with_holes" : {
      "buckets" : [
        {
          "key_as_string" : "2019-07-27T00:00:00.000Z",
          "key" : 1564185600000,
          "doc_count" : 1
        },
        {
          "key_as_string" : "2019-07-28T00:00:00.000Z",
          "key" : 1564272000000,
          "doc_count" : 2
        },
. . .
 {
        "key_as_string" : "2021-03-04T00:00:00.000Z",
        "key" : 1614816000000,
        "doc_count" : 35
      },
      {
        "key_as_string" : "2021-03-05T00:00:00.000Z",
        "key" : 1614902400000,
        "doc_count" : 40
      }
    ]
  }
 }
}
```

Ці дані можна отримати із інструмента Console системи Kibana (Рис. 14).

Для подальшої обробки достатньо застосовувати дані, що відповідають полям key_as_string і doc_count.

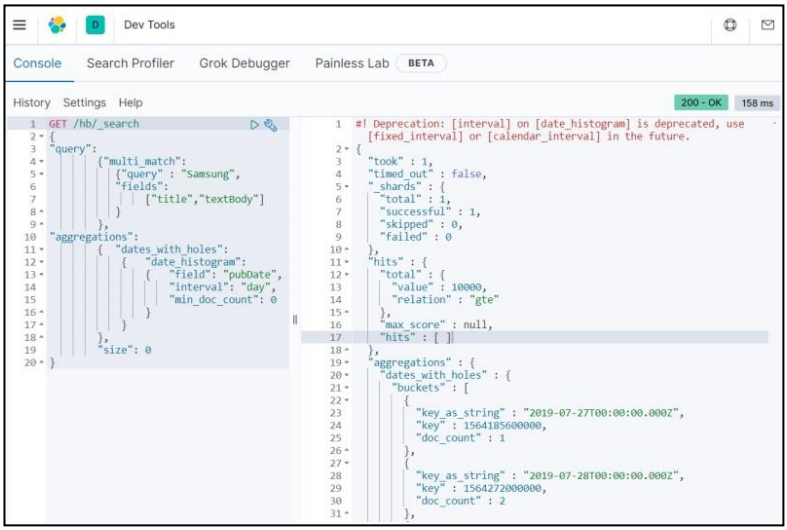


Рисунок 14 – Результати агрегації в консолі системи Kibana

Для застосування отриманих результатів у середовищі спеціалізованих систем обробки цифрових даних, можна перетворити їх до формату CSV за допомогою програми мовою Python.

У результаті виконання програми можна отримати дані у форматі CSV такого вигляду:

2019-07-27;1

2019-07-28;2

2019-07-29;0

. . .

2021-03-03;32

2021-03-04;35

2021-03-05;40

Для подальшої обробки даних в форматі завантажимо отриманий CSV-файл у середовище Excel і побудуємо графік динаміки повідомлень (Рис. 15).

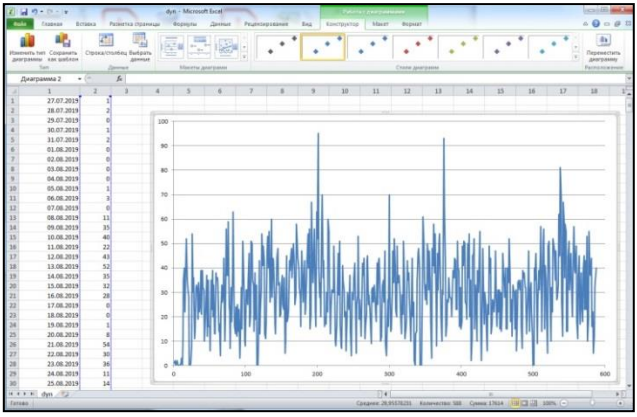


Рисунок 15 – Завантажені в систему Excel дані та побудований графік

Після завантаження CSV-файлу в систему цифрової обробки даних виникають прості можливості його статистичного оброблення. На Рис. 16 наведено приклад знаходження поліноміального тренду цього ряду.

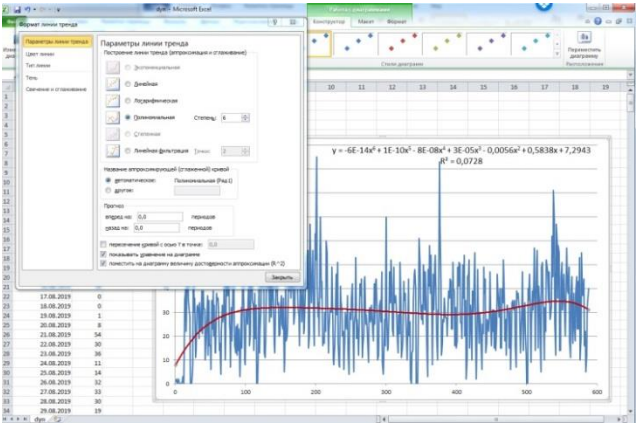


Рисунок 16 – Поліноміальний тренд часового ряду, що досліджується

2.6.2 Кореляційний аналіз

Багато методів дослідження часових рядів базуються на припущеннях про статистичну рівновагу або стан. Одним із таких корисних пропозицій є стаціонарність.

Часовий ряд називається строго стаціонарним або стаціонарним у вузькому сенсі, якщо його статистичні властивості не змінюються з часом. Формально, якщо спільне розподіл випадкових величин $x_t, x_{t+1}, ..., x_{t+n}$ співпадає з розподілом $x_{t+k}, x_{t+k+1}, ..., x_{t+k+n}$ при будь-яких цілих значеннях здвигу k, то часовий ряд $\{x_t\}_{t=1}^{T}$ називається строго стаціонарним. У стаціонарних часових рядів постійне математичне очікування

$$\mu = Ex_t$$

і дисперсія

$$\sigma^2 = Var(x_t) = E(x_t - Ex_t)^2.$$

При цьому значення μ і σ^2 можна оцінити як середнє вибіркове:

$$\hat{\mu} = \bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$$

та вибіркову дисперсію

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(x_t - \bar{x}\right)^2.$$

Властивість стаціонарності також має значення при порівнянні часових рядів. Лінійна залежність між двома випадковими величинами вимірюється коваріацією. Для часових рядів визначають кросковаріаційну функцію. За визначенням, кросковаріація з часовою затримкою k між випадковими процесами $\{x_t\}_{t=1}^{T}$ і $\{y_t\}_{t=1}^{T}$ дорівнює:

$$\gamma_{xy}(k,t) = Cov(x_t, y_{t+k}) = E\left[(x_t - \mu_x)(y_{t+k} - \mu_y)\right].$$

З припущення про стаціонарності у вузькому значенні випливає, що розподіл пар величин x_t, y_{t+k} однаково для довільного значення t. Отже, коваріація між величинами x_t і y_{t+k} не залежить від t, а залежить тільки від значення k, тобто $\gamma_{xy}(k,t) = \gamma_{xy}(k), \forall t.$ Набір значень $\{\gamma_{xy}(k)\}$ утворює кросковаріаційну функцію.

Після нормування отримуємо кроскореляційний коефіцієнт:

$$\rho_{xy}(k) = \frac{Cov(x_t, y_{t+k})}{\sigma_x \sigma_y} = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y}.$$

Кроскореляційна функція є мірою подібності між двома часовими рядами.

Найчастіше кросковаріаційні та кроскореляційні коефіцієнти оцінюють за формулами:

$$\hat{\gamma}_{xy}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{x})(y_{t+k} - \bar{y}), \quad \hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{\hat{\gamma}_{xy}(0)}.$$

Зауважимо, що такі оцінки справедливі для стаціонарних у вузькому значенні рядів, оскільки відповідні коефіцієнти не залежать від часу, а загальному випадку це може і не виконуватись. Часто використовують більш слабку вимогу, ніж стаціонарність у вузькому значенні, – стаціонарність у широкому значенні.

Часовий ряд $\{x_t\}_{t=1}^{T}$ стаціонарний у сенсі, якщо його математичне очікування змінюється з часом, тобто $\forall t \exists Ex_t = const$ і ковариационная функція залежить від різниці аргументів $Cov(x_t, x_s) = K(t-s).$

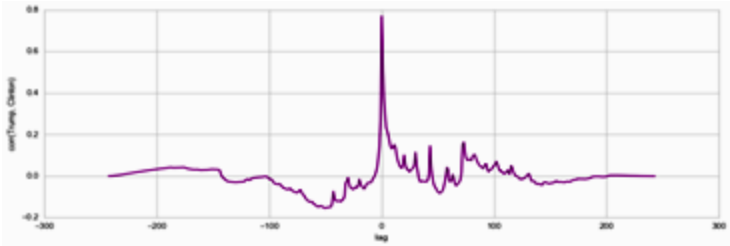
Так як у визначенні зазначено, що математичне очікування постійно і, легко помітити, що дисперсія також не змінюється з часом $Var(x_t) = Copv(x_t, x_t) = K(0) = const$, то в цьому випадку, як і для строго стаціонарних рядів, справедливі наведені раніш оцінки.

Для прикладу наведемо оцінку кроскореляційних функцій для рядів *T*, *K*, *X*. Рис. 17 а показана кореляційна функція для рядів *T* і *K*. Максимальне значення (приблизно рівне 0.8) функція досягає при тимчасовій затримці 0. Тобто два тимчасові ряди, пов'язані з інтересом до Дональда Трампа і Хіларі Клінтон, сильно корельовані.

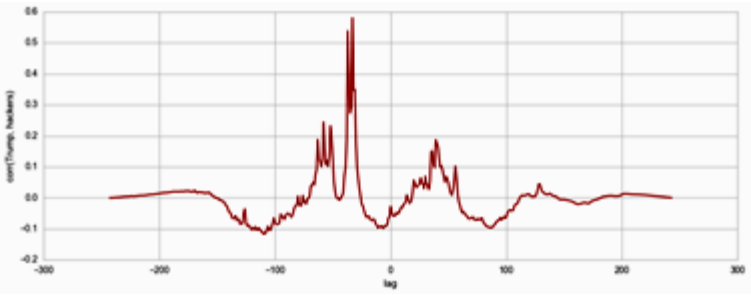
На Рис. 17 б показана кореляційна функція для рядів T і X. Максимальне значення (приблизно 0.7) функція досягає при тимчасовій затримці 34 дні. Це відповідає тому факту, що починаючи з 13 грудня 2016 року (34 дні після виборів у США 8 листопада) різко зросла кількість повідомлень новин про «російських хакерів».

Можна підрахувати коваріацію задля двох різних рядів, а одного ряду. Така коваріація називається автоковаріацією з тимчасовою затримкою або лагом *k*:

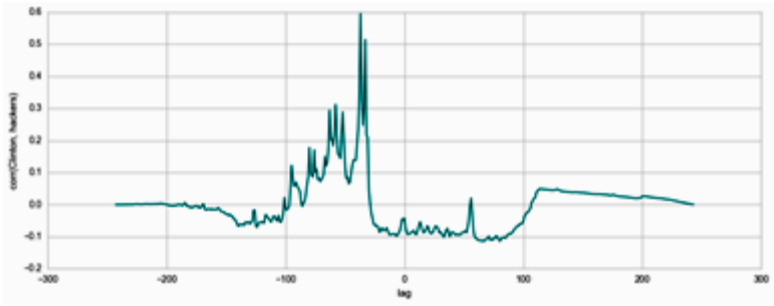
$$\gamma_k = Cov(x_t, x_{t+k}) = E\left[\left(x_t - \mu\right)\left(x_{t+k} - \mu\right)\right].$$



а



б



в

Рисунок 17 – Кореляційні функції для пар рядів T і K(а), T і X(б), K та X(в). Вздовж осі абсцис відкладено часову затримку (лаг), вздовж осі ординат – оцінка кореляційного коефіцієнта

Набір величин γ_k, $k = 0,1,2,...$ зветься автоковарійною функцією, а їхнє нормоване значення ρ_k, $k = 0,1,2,...$ – автокореляційною функцією:

$$\rho_k = \frac{E\left[\left(x_t - \mu\right)\left(x_{t+k} - \mu\right)\right]}{\sqrt{E\left(x_t - \mu\right)^2 E\left(x_{t+k} - \mu\right)^2}} = \frac{Cov(x_t, x_{t+k})}{Var(x_t)} = \frac{\gamma_k}{\gamma_0}.$$

Найчастіше автоковарійні та автокореляційні коефіцієнти оцінюють за формулами:

$$\hat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{x})(x_{t+k} - \bar{x}), \quad \hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}.$$

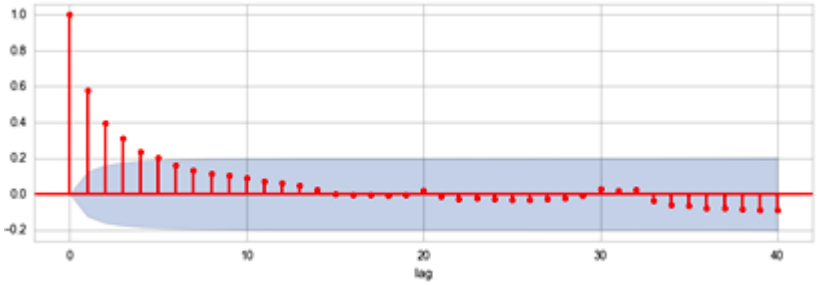
Після обчислення оцінок для кореляційних коефіцієнтів виникає питання: чи рівні коефіцієнти ρ_k нулю починаючи з деякого значення k? Для відповіді це питання потрібно порівняти значення оцінки $\hat{\rho}_k$ з його стандартним відхиленням. Якщо ми приймаємо припущення, що $\rho_k = 0$, то для стандартного відхилення оцінки $\hat{\rho}_k$ буде справедливо:

$$se(\hat{\rho}_k) \cong \frac{1}{\sqrt{T}}.$$

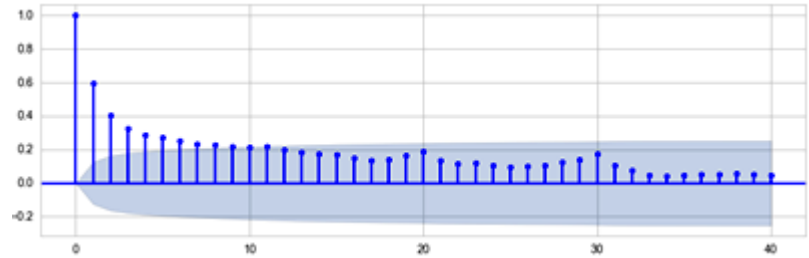
На Рис. 5.12 наведено автокореляційні функції часових рядів, що розглядались вище. На ньому уздовж осі абсцис

відкладено часову затримку (лаг), уздовж осі ординат – автокореляційний коефіцієнт. Потім темна область показує стандартне відхилення для оцінки автокореляційного коефіцієнта.

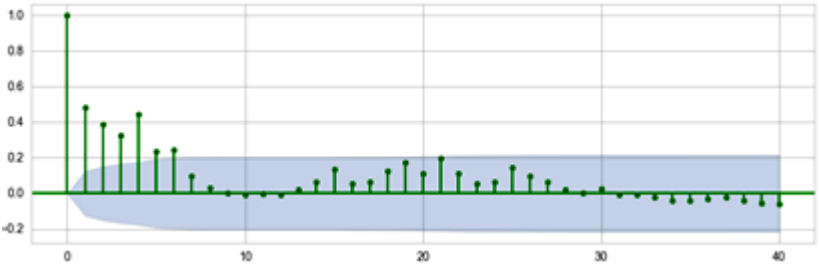
Насправді часто використовують емпіричне правило, за яким автокореляційні коефіцієнти оцінюють для часової затримки не більше як $T / 4$. На Рис. 18 наведені автокореляційні функції для рядів *T*, *K*, *X*.



а



б



в

Рисунок 18 – Автокореляційні функції для рядів *T*(*а*), *K*(*б*), *X*(*в*)

Визначення автокореляційної функції вводилося для стаціонарних часових рядів, але оцінити її значення можна

довільного часового ряду. Для нестаціонарних часових рядів така автокореляційна функція зменшується повільніше.

2.6.3. Аналіз Фур'є

Класичний аналіз Фур'є надає можливість досліджувати функцію у часовій та частотній області. Суть переходу в частотну область полягає в тому, що функція розкладається на складові, що є гармонійними коливаннями з різними частотами. При цьому кожній частоті відповідає коефіцієнт, який відображає амплітуду коливання цієї частоти. Якщо уявити функцію графічно у часовій області, отримаємо інформацію у тому, як функція змінюється з часом. Якщо зобразити функцію в частотній області, то отримаємо інформацію про частоти, коливання яких вона містить. Для цього використовують пряме та зворотне перетворення Фур'є

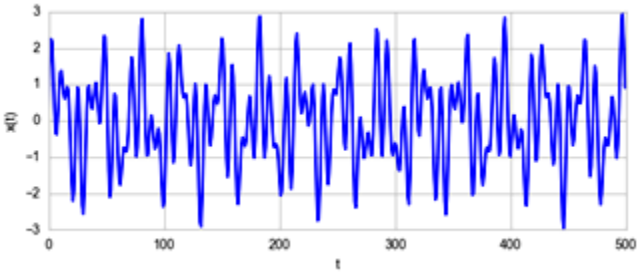
$$\hat{x}(v) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi vt} dt,$$

$$x(t) = \int_{-\infty}^{\infty} \hat{x}(v) e^{i2\pi vt} dv.$$

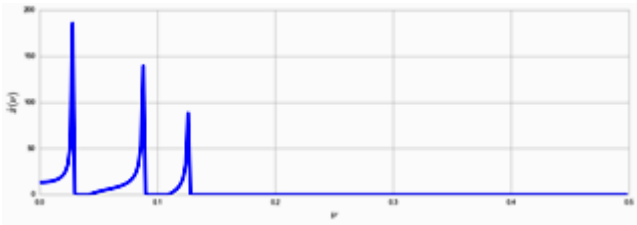
На Рис. 19 а показано приклад функції, яка насправді є сумою трьох синусоїд з різними періодами. Дивлячись лише на графік функції у часовій області досить важко зрозуміти, що вона складається з трьох гармонійних коливань та визначити їх періоди. На Рис. 19 б показано перетворення Фур'є для цієї функції. З графіка у частотній області наочно видно, що функція містить коливання на трьох різних частотах.

Сьогодні перетворення та спектри Фур'є знаходять різноманітні застосування у системах машинного навчання. Часто спектри Фур'є використовуються як навчальні параметри. Наприклад, існує модель прогнозування часового ряду, в якій

спектр Фур'є разом з деякими іншими параметрами подається на вхід нейронної мережі[24].



а



б

Рисунок 19 – Функція, яка є сумою трьох синусоїд з різними періодами (а) та оцінений спектр Фур'є для цієї функції (б)

Перетворення та спектри Фур'є часто використовуються при розпізнаванні мови з різними умовами навчання. Спектри Фур'є також використовуються як навчальні параметри для нейронних мереж в системах автоматичного детектування певних подій у мові або на тлі шуму. Іншим завданням у галузі розпізнавання є визначення емоційного забарвлення мови.

Зокрема, було запропоновано модель розпізнавання, в основі якої лежать певні параметри Фур'є, та демонструється ефективність використання таких параметрів для ідентифіка-

[24] Rodriguez N., Bravo G., Rodriguez N., Barba L. Haar Wavelet Neural Network for Multi-step-ahead Anchovy Catches Forecasting // Polibits. – 2014. – Issue 50. – P. 49-53.

ції різних емоційних станів у голосових сигналах25. Перетворення Фур'є можна як визначення кореляції між вихідним сигналом і гармонійними функціями з різними частотами коливання. Незважаючи на свої переваги та численні програми, перетворення Фур'є є поганим методом для дослідження функцій, які еволюціонують з часом. Для таких функцій потрібен певний спосіб оцінювання спектра не по всій довжині часового ряду, а по різних частинах.

2.6.4 Вейвлет-аналіз

На цей час успішно розвивається новий і важливий напрямок в теорії і технології обробки сигналів, зображень та часових рядів, що отримав назву вейвлет-перетворення, яке добре пристосоване для вивчення структури неоднорідних процесів.

Вейвлет-перетворення має кореляційну природу. У цьому разі розглядається кореляція вихідної функції з функцією вейвлетом різних масштабах. Для того щоб таку процедуру завжди можна було виконати і кореляційні коефіцієнти були інформативними, вейвлет повинен мати певні математичні властивості. Буквально слово вейвлет перекладається як «маленька хвиля» або «сплеск», і, як випливає з назви, вейвлет добре локалізований у часі.

Вейвлети представляють собою особливі функції у вигляді коротких хвиль (сплесків) з нульовим інтегральним значенням і з локалізацією по осі незалежної змінної (t або x), здатних до зсуву по цій осі і до масштабування (розтягування / стиснення). Будь-який з найбільш часто використовуваних типів вейвлетів породжує повну ортогональну систему функцій.

У разі вейвлет-аналізу (декомпозиції) процесу (сигналу) у зв'язку зі зміною масштабу вейвлети здатні виявити відмінність в характеристиках процесу на різних шкалах, а за

---

[25] W. Lu, X. Wang, C. Yang and T. Zhang, "A novel feature extraction method using deep neural network for rolling bearing fault diagnosis," The 27th Chinese Control and Decision Conference (2015 CCDC), Qingdao, China, 2015, pp. 2427-2431, DOI: 10.1109/CCDC.2015.7162328.

допомогою зсуву надають можливість аналізувати властивості процесу в різних точках на всьому досліджуваному інтервалі.

Узагальнений ряд Фур'є

Відомо, що довільна функція $f(t)$, для якої виконується умова:

$$\int_{t_2}^{t_2}\left[f(t)\right]^2 dt < \infty,$$

може бути представлена ортогональною системою функцій $\{\varphi_n(t)\}$, тобто

$$f(t) = C_0\varphi_0(t) + ... + C_n\varphi_n(t) + ... = \sum_{n=0}^{\infty} C_n\varphi_n(t),$$

де коефіцієнти визначаються із співвідношення:

$$C_n = \frac{1}{\|\varphi_n\|^2}\int_{t_1}^{t_2} f(t)\varphi_n(t)\,dt.$$

У цьому виразі $\|\varphi_n\|^2 = \int_{t_1}^{t_2}\varphi_n^2(t)\,dt$ – це квадрат норми, або енергія базисної функції $\varphi_n(t)$. При цьому передбачається, що ніяка з базисних функцій не дорівнює тотожно нулю і на інтервалі ортогональности $[t_1, t_2]$ виконується умова:

$$\int_{t_1}^{t_2}\varphi_n(t)\varphi_k(t)\,dt = \begin{cases}\|\varphi_n(t)\|^2, & k = n,\\ 0, & k \neq n.\end{cases}$$

Базисна функція $\varphi_n(t)$, для якої квадрат норми дорівнює одиниці $\left(\|\varphi_n(t)\|^2 = 1\right)$, називається нормованою (нормальною), а вся система функцій $\{\varphi_n(t)\}$ – ортонормованою або ортонормальною. У цьому випадку говорять, що задано ортонормований базис.

Ряд розкладення, у якому коефіцієнти C_n визначаються за наведеною вище формулою, називається узагальненим рядом Фур'є.

Добуток вигляду $C_n\varphi_n(t)$, що входить до цього ряду, являє собою спектральну складову сигналу $f(t)$, а сукупність коефіцієнтів $\{C_0,...,C_n,...\}$ називається спектром сигналу.

Суть спектрального аналізу сигналу $f(t)$ полягає у визначенні коефіцієнтів C_n (експериментально або аналітично). На основі ряду можливий синтез (апроксимація) сигналів при фіксованому числі N ряду

$$\hat{f}(t) = C_0\varphi_0(t) + ... + C_n\varphi_N(t) = \sum_{n=0}^{N} C_n\varphi_n(t),$$

При цьому узагальнений ряд Фур'є має наступну важливу властивість: при заданій системі базисних функцій $\{\varphi_n(t)\}$ і числі доданків N він забезпечує найкращий синтез (апроксимацію), даючи мінімум середньоквадратичної помилки ε, під якою розуміється величина:

$$\varepsilon = \int_{t_1}^{t_2} \left[f(t) - \hat{f}(t) \right]^2 dt = \int_{t_1}^{t_2} \left[f(t) - \sum_{n=0}^{N} C_n\varphi_n(t) \right]^2 dt.$$

Ортогональна система називається повною, якщо збільшенням N можна зробити ε як завгодно малим. Ряд розкладу у цьому випадку називається збіжним у середньоквадратичному.

Відносна помилка μ синтезу визначається за формулою:

$\mu = \varepsilon / E,$

де E – енергія сигналу, що чисельно рівна квадрату норми сигналу $f(t)$, тобто

$$E = \|F\|^2 = \int_{t_1}^{t_2} \left[f(t) \right]^2 dt$$

.

Ця формула із урахуванням розкладу в ряд може бути записана як:

$$E = \|F\|^2 = \sum_{n=0}^{\infty} C_n^2 \|\varphi_n\|^2,$$

а при використанні ортонормованій системи функцій $\{\varphi_n(t)\}$ як:

$$E = \sum_{n=0}^{\infty} C_n^2.$$

Питання щодо вибору раціональної системи ортогональних функцій залежить від поставленого завдання. Так при аналізі і синтезі сигналів, що впливають на лінійні ланцюги, найбільшого поширення набула система гармонійних функцій. По-перше, гармонійні коливання на відміну від інших зберігають свою форму при проходженні через ці ланцюги; змінюються лише амплітуда і початкова фаза. По-друге, широко використовується добре розроблений в теорії ланцюгів символічний метод. Із множини інших завдань найбільш важливою є задача наближеного розкладання складних сигналів, при якій необхідна точність забезпечується при мінімумі членів ряду.

Для представлення безперервних сигналів застосовуються поліноми і функції Лагерра, Лежандра, Чебишева, Ерміта і ін. Для подання сигналів з точками розриву використовуються шматково-постійні функції Уолша, Хаара, Радемахера. Для дискретизації безперервних сигналів у часі використовується ортогональний ряд Котельникова. В останні роки широко використовуються базисні функції типу вейвлетів.

5.5.2 Вейвлети

До кола сучасних інструментальних засобів оцінки рядів спостережень відноситься також вейвлет-аналіз[26]. Він особливо ефективний у тих випадках, коли крім загальних спектральних

26 Buckheit J., Donoho D. Wavelab and reproducible research // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. – 27 p.

характеристик потрібно виявляти локальні в часі особливості поведінки процесу, що досліджується. Основою вейвлет-аналізу є вейвлет-перетворення, яке є особливим типом лінійного перетворення, базисні функції якого (вейвлети) мають специфічні властивості. Аналіз даних з використанням вейвлет-перетворень є зручним, надійним і потужним інструментом дослідження часових рядів і дозволяє представити результати у наочному вигляді, зручному для інтерпретації.

Вейвлетом (малою хвилею) називається деяка функція, зосереджена в невеликій околиці деякої точки та різко убутна до нуля по мірі видалення від неї як у часовий, так і в частотній області. Існують найрізноманітніші вейвлети, що мають різні властивості. Разом з тим, усі вейвлети мають вигляд коротких хвильових пакетів з нульовим інтегральним значенням, локалізованих на часовій осі, які є інваріантними до зсуву і до масштабування.

До будь-якому вейвлету можна застосувати дві операції: зрушення, тобто переміщення області його локалізації в часі; масштабування (розтягання або стиск).

Головна ідея вейвлет-претворення полягає в тому, що нестаціонарний часовий ряд розподіляється на окремі проміжки (так званні «вікна спостереження»), і на кожному з них виконується обчислення скалярного добутку (величини, що показує ступінь близькості двох закономірностей) досліджуваних даних із різними зрушеннями деякого вейвлету на різних масштабах. Вейвлет-перетворення генерує набір коефіцієнтів, за допомогою яких представляється початковий ряд. Вони є функціями двох змінних: часу і частоти, і тому утворюють поверхню у трьохвимірному просторі. Ці коефіцієнти, що показують, наскільки поведінка процесу в даній точці аналогічно вейвлету на даному масштабі. Чим ближче вигляд аналізованої залежності в околі даної точки до вигляду вейвлету, тим більшу абсолютну величину має відповідний коефіцієнт. Негативні коефіцієнти показують, що залежність схожа на "дзеркальне відбиття" вейвлету.

Технологія використання вейвлетів дозволяє виявляти одиничні та нерегулярні «сплески», різкі зміни значень кількісних показників у різні періоди часу, зокрема, обсягів тематичних публікацій в Інтернет. При цьому можуть виявлятися моменти

виникнення циклів, а також моментів, коли за періодами регулярної динаміки настають хаотичні коливання.

Часовий ряд може апроксимуватися кривою, яка, у свою чергу, може бути представлена у вигляді суми гармонійних коливань різної частоти й амплітуди. При цьому коливання, що мають низьку частоту, відповідають за повільні, плавні, великомасштабні зміни значень часового ряду, а високочастотні – за короткі, дрібномасштабні зміни. Чим сильніше змінюється описувана даною закономірністю величина на даному масштабі, тим більшу амплітуду мають складові на відповідній частоті. Таким чином, досліджуваний часовий ряд можна розглядати в частотно-часовій області, у залежності як від часу, так і від частоти.

Вейвлет-перетворення одновимірного сигналу / функції – це його представлення у вигляді узагальненого ряду або інтеграла Фур'є по системі базисних функцій

$$\psi_{ab}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),$$

які сконструйовані з материнського (вихідного) вейвлету $\psi(t)$, що має певні властивості за рахунок операцій зсуву в часі (b) і зміни часового масштабу (a). Множник $1/\sqrt{a}$ забезпечує незалежність норми цих функцій від масштабуючого числа a.

Отже, для заданих значень параметрів a та b функція $\psi_{ab}(t)$ і є тим вейвлетом, що породжується материнським вейвлетом $\psi(t)$. У якості базисних функцій, що утворюють ортогональний базис, можна використовувати широкий набір вейвлетів.

Для практичного застосування важливо знати ознаки, якими неодмінно повинна володіти початкова функція, щоб стати вейвлетом. Наведемо тут основні з них.

Обмеженість. Квадрат норми функції повинен бути скінченним:

$$\|\psi\|^2 = \int_{-\infty}^{\infty} |\psi(t)|^2 \, dt < \infty.$$

Локалізація. Вейвлет-перетворення на відміну від перетворення Фур'є використовує локалізовану вихідну функцію і у часі, і у частоті. Для цього достатньо, щоб виконувалися умови:

$$|\psi(t)| \le C(1+|t|)^{-1-\varepsilon} \quad i \quad |f_\psi(\omega)| \le C(1+\omega)^{-1-\varepsilon}, \quad \text{при } \varepsilon > 0.$$

Наприклад, дельта-функція $\delta(t)$ і гармонійна функція не задовольняють необхідній умові одночасної локалізації у часовій і частотній областях.

Нульове середнє. Графік вихідної функції повинен осцілювати (бути знакозмінним) навколо нуля на осі часу і мати нульову площу:

$$\int_{-\infty}^{\infty} \psi(t)\,dt = 0.$$

З цієї умови стає зрозумілим вибір назви «вейвлет» – маленька хвиля.

Рівність нулю площі функції $\psi(t)$, тобто нульового моменту, призводить до того, що Фур'є-перетворення $f_\psi(\omega)$ цієї функції дорівнює нулю при ω = 0 і має вигляд смугового фільтра. При різних значеннях параметра a це буде набір смугових фільтрів. Часто для застосунків буває необхідно, щоб не тільки нульовий, а й всі перші n моментів дорівнювали нулю:

$$\int_{-\infty}^{\infty} t^n \psi(t)\,dt = 0.$$

Вейвлети n-го порядку дозволяють аналізувати більш тонку (високочастотну) структуру сигналу, пригнічуючи його складові, що змінюються повільно.

Автомодельність. Характерною ознакою вейвлет-перетворення є його самоподібність. Усі вейвлети конкретного сімейства $\psi_{ab}(t)$ мають те саме число осциляцій, що і материнський вейвлет $\psi(t)$, оскільки отримані з нього за допомогою масштабних перетворень (a) і зсуву (b).

Найбільш поширені дійсні базиси конструюються на основі похідних функції Гауса ($g_0(t) = \exp\left(-t^2/2\right)$). Це обумовлено тією обставиною, що функція Гауса має найкращі показники локалізації як у часовій, так і у частотній областях. При n = 1 отримуємо вейвлет першого порядку, так званий WAVE-вейвлет з рівним нулю нульовим моментом. При n = 2 отримуємо МНАТ-вейвлет, «мексиканський капелюх» (Mexican Hat). У нього нульовий і перший моменти дорівнюють нулю. Він має кращу розподільну здатність, ніж WAVE-вейвлет. Приклади графіків базових вейвлетів наведено на Рис. 20.

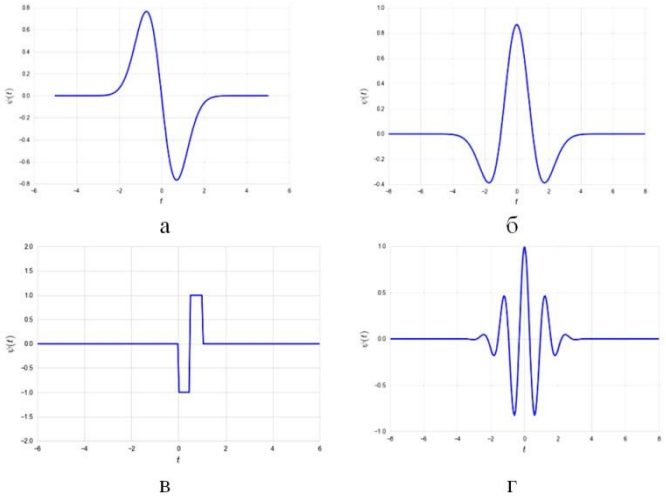


Рисунок 20 – Приклади вейвлетів, що часто використовуються у застосунках: (а) Гаусова хвиля (перша похідна функції Гауса), (б) мексиканський капелюх, (в) вейвлет Хаара, (г) вейвлет Морле (дійсна частина)

Неперервне вейвлет-перетворення

Неперервне пряме і зворотне вейвлет-перетворення для функції $f(t)$ будується за допомогою неперервних масштабних перетворень і переносів вибраного вейвлету $\psi(t)$ з довільними значеннями масштабного коефіцієнта a та параметра зсуву b:

$$W_f(a,b) = (f(t), \psi_{ab}(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-b}{a}\right) dt,$$

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_f(a,b) \psi_{ab}(t) \frac{dadb}{a^2},$$

де C_ψ – нормуючий коефіцієнт:

$$C_\psi = \int_{-\infty}^{\infty} |\Psi(\omega)|^2 |\omega|^{-1} d\omega < \infty,$$

де (., .) – скалярний добуток відповідних співмножників, $\Psi(\omega)$ – Фур'є-перетворення вейвлету $\psi(t)$. Для ортонормованих вейвлетів C_ψ.

Таким чином, вейвлет-спектр $W_f(a,b)$ (Wavelet Spectrum, або time-scale-spectrum – масштабно-часовий спектр) на відміну від Фур'є-спектру (single spectrum) є функцією двох аргументів: перший аргумент a (часової масштаб) аналогічний періоду осциляцій, тобто оберненій частоті, а другий b – аналогічний зміщенню сигналу по осі часу.

Слід зазначити, що $W_f(a_0,b)$ характеризує часову залежність (при $a = a_0$), тоді як залежності $W_f(a, b_0)$ можна поставити у відповідність частотну залежність (при $b = b_0$).

Якщо досліджуваний сигнал $f(t)$ являє собою одиночний імпульс тривалістю τ_u, зосереджений в околі $t = t_0$, то його вейвлет-спектр матиме найбільше значення в околі точки з координатами $a = \tau_u$, $b = t_0$.

Отримані коефіцієнти представляються у графічному вигляді картою коефіцієнтів перетворення, або скейлограмою. На скейлограмі по одній осі відкладаються зрушення вейвлету (вісь часу), а по іншій – масштаби (вісь масштабів), після чого точки схеми, що отримується, офарбовуються залежно від величини відповідних коефіцієнтів (чим більше коефіцієнт, тим яскравіше кольори зображення). На скейлограмі видні всі характерні риси вихідного ряду: масштаб та інтенсивність

періодичних змін, напрямок і величина трендів, наявність, розташування та тривалість локальних особливостей.

Наприклад, відомо, що комбінація декількох різних коливань може мати настільки складну форму, що виявити їх візуально не представляється можливим. Періодичні зміни, що відбуваються для значень коефіцієнтів вейвлет-перетворення на деякій неперервній множині частот виглядають як ланцюжок "пагорбів", що мають вершини, розташовані в точках (по осі часу), у яких ці зміни досягають найбільших значень.

Іншим важливим показником є виражена тенденція динаміки часового ряду (тренд) поза залежністю від періодичних коливань. Наявність тренда може бути неочевидною при простому розгляді часового ряду, наприклад, якщо тренд поєднується з періодичними коливаннями. Тренд відбивається на скейлограмі як плавна зміна яскравості уздовж осі часу одночасно на всіх масштабах. Якщо тренд наростаючий, то яскравість буде збільшуватися, якщо убутний – зменшуватися. Ще одним важливим фактором, якому необхідно враховувати при аналізі часових рядів, є локальні особливості, тобто можливі різкі, стрибкоподібні зміни характеристик вихідного ряду. Локальні особливості представлені на скейлограмі вейвлет-перетворення як лінії різкого перепаду яскравості, що виходять із точки, що відповідає часу виникнення стрибка. Аналіз локальних особливостей дозволяє відновити інформацію щодо динаміки вихідного процесу та у деяких випадках прогнозувати подібні ситуації.

Кожний з основних факторів динаміки часового ряду має свій, характерний відбиток на скейлограмі, при цьому вся аналітична інформація представляється в наочному та зручному для вивчення вигляді. Завдяки наочності подання результатів у вигляді скейлограми, іноді досить одного погляду, щоб побачити на ній найбільш суттєві фактори.

Вейвлет-спектр $W_f(a,b)$ (wavelet spectrum, або time-scale-spectrum – масштабно-часовий спектр) є функцією двох аргументів: перший аргумент a (часової масштаб) аналогічний періоду осциляцій, тобто обернений частоті, а другий b – аналогічний зміщенню сигналу по осі часу.

Слід зазначити, що $W_f(a_0,b)$ характеризує часову залежність (при $a=a_0$), тоді як залежності $W_f(a,b_0)$ можна поставити у відповідність частотну залежність (при $b=b_0$).

Якщо досліджуваний сигнал $f(t)$ являє собою одиночний імпульс тривалістю τ_u, зосереджений в околі $t=t_0$, то його вейвлет-спектр матиме найбільше значення в околі точки з координатами $a=\tau_u$, $b=t_0$.

Отримані коефіцієнти представляються у графічному вигляді картою коефіцієнтів перетворення, або скейлограмою (Рис. 21).

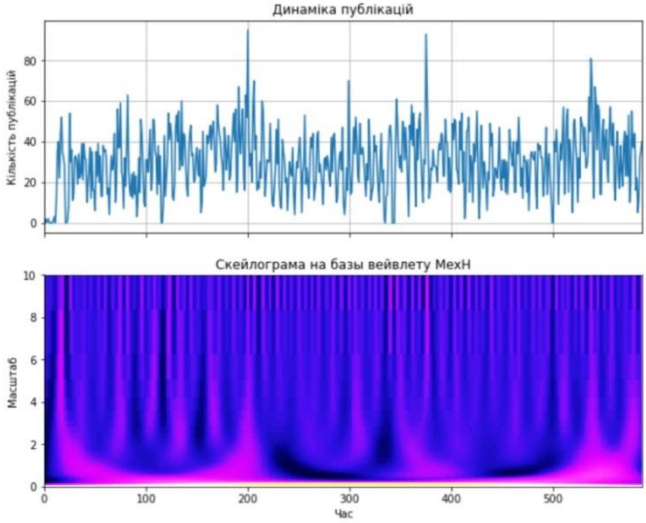


Рисунок 21 – Приклад вейвлет-скейлогами вихідного часового ряду

На скейлограмі по одній осі відкладаються зрушення вейвлету (вісь часу), а по іншій – масштаби (вісь масштабів), після чого точки схеми, що отримується, офарбовуються залежно від величини відповідних коефіцієнтів (чим більше коефіцієнт, тим яскравіше кольори зображення). На скейлограмі видні всі характерні риси вихідного ряду: масштаб та інтенсивність періодичних змін, напрямок і величина трендів,

наявність, розташування та тривалість локальних особливостей.

2.6.5 Кореляція з шаблоном

За допомогою безперервного вейвлет-перетворення виявляються ділянки досліджуваного ряду, які формою найбільш схожі на вейвлет (Рис. 22).

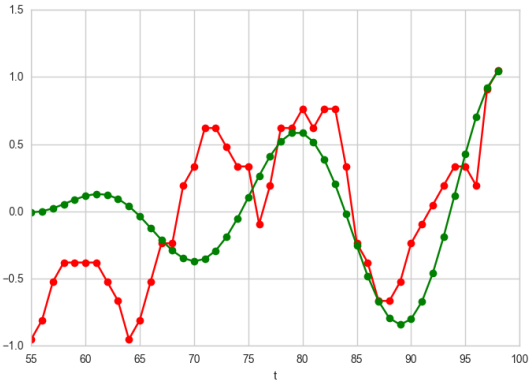


Рисунок 22 – Відрізок часового ряду з накладеним шаблоном

Ідея полягає в тому, щоб порівняти частини ряду з деяким шаблоном на різних масштабах (Рис. 23).

При цьому вейвлет як функція повинен мати певні математичні властивості, зокрема швидко зменшуватися до нуля на нескінченності. У деяких випадках корисно використовувати шаблон, який не відповідає вимогам до вейвлету.

Для цього замість вейвлет-перетворення обчислюватимемо кореляцію між частиною часового ряду та деяким шаблоном p:

$$C(l,k)=\frac{\sum_{i=1}^{k}(x_{l+i}-\bar{x})(p_i-\bar{p})}{\sum_{i=1}^{k}(x_{l+i}-\bar{x})^2\sum_{i=1}^{k}(p_i-\bar{p})^2},$$

Отриманий коефіцієнт $C(l,k)$ залежить від значень $x_{l+1},\ldots,x_{l=k}$. Тобто параметр 1 відповідає зсуву шаблону, а параметр k відповідає кількості точок у шаблоні та у розглянутому

відрізку ряду. Параметр k цьому випадку є аналогом масштабу s, який використовували під час вейвлет-перетворення.

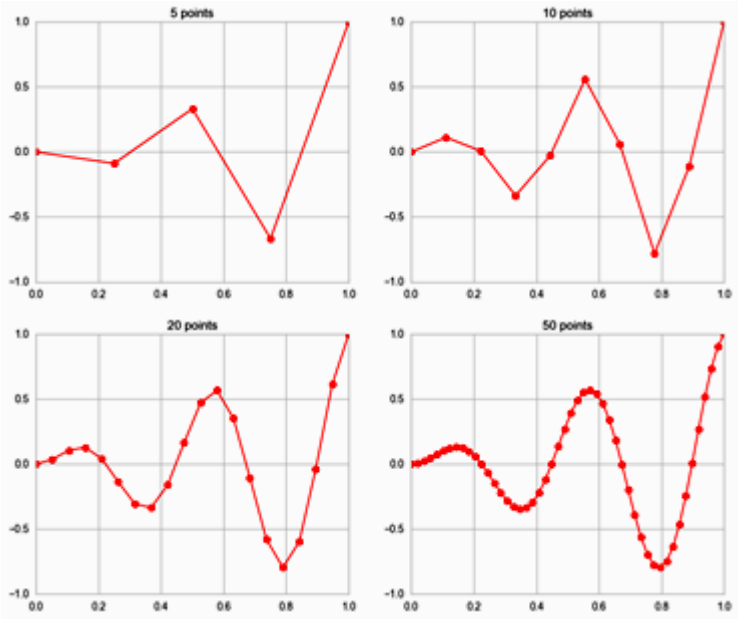


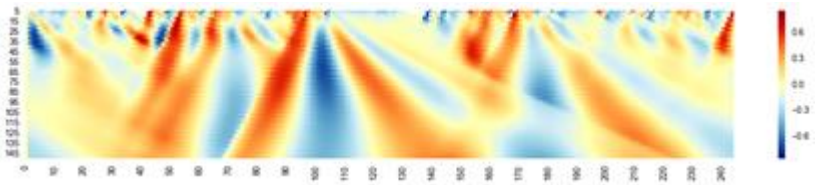
Рисунок 23 – Шаблон «змія» з різною кількістю точок

Якщо при обчисленні коефіцієнта вейвлет-перетворення завжди використовувався весь тимчасовий ряд, то в даному випадку для обчислення $C(l,k)$ використовуються k точок ряду і шаблон довжини k.

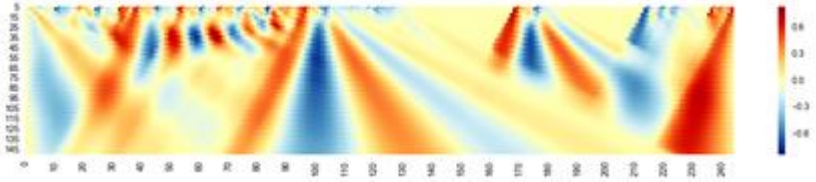
Отримані кореляційні коефіцієнти $C(l,k)$ представимо на графіку, що схожий на скейлограму (Рис. 24).

2.6.6 Фрактальний аналіз

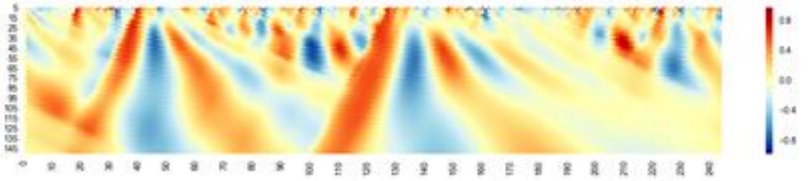
Термін фрактал, був запропонований Б. Мандельбротом у 1975 році для позначення нерегулярних самоподібних математичних структур.



а



б



в

Рисунок 24 – Кореляційні коефіцієнти C(1,k) обчислені для рядів T (а), K (б) та X (в) з використанням шаблону

Основне визначення фракталу, дане Мандельбротом, звучало так: «Фракталом називається структура, що складається із частин, які в якомусь змісті подібні до цілого»[27].

Головна особливість фракталів полягає у тому, що їх розмірність не укладається у звичні геометричні уявлення. Фракталам характерна геометрична «порізанність». Тому використовується спеціальне поняття фрактальної розмірності, введене Ф. Хаусдорфом та А. Безиковичем. Розмірність фракталів не є цілим числом, характерним для звичних геометричних об'єктів.

Алгоритм побудови фрактальної множини Мандельброта (Рис. 25) заснований на ітеративному обчисленні за формулою:

$$Z[i+1] = Z[i] \times Z[i] + C,$$

[27] B. Mandelbrot. “The Fractal Geometry of Nature”. Freeman and Co., San Francisco 1982. 460 pp.

де Z й C – комплексні змінні.

Ітерації виконуються для кожної стартової точки С прямокутної або квадратної області – підмножині комплексної площини. Ітераційний процес триває доти, поки $Z[i]$ не вийде за межі окружності заданого радіуса, центр якої лежить у точці (0,0), або після досить великої кількості ітерацій. Залежно від кількості ітерацій, протягом яких $Z[i]$ залишається усередині окружності, встановлюються кольори точок.

Завдяки тому, що кількість ітерацій відповідає номеру кольору, то точки, що перебувають ближче до множини Мандельброта, мають більш яскраве забарвлення.

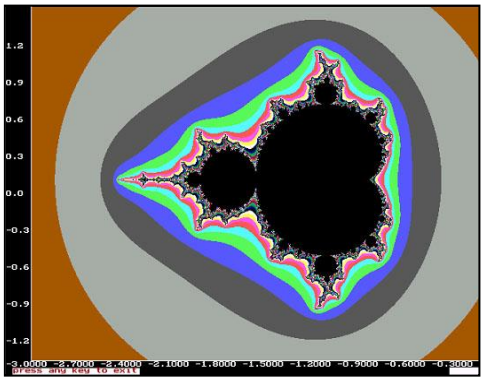


Рисунок 25 – Множина Мандельброта

Побудова іншої фрактальної множини, сніжинки Коха (Рис. 26), починається із правильного трикутника, довжина сторони якого дорівнює 1.

Сторона трикутника вважається базовою ланкою – вихідним положенням. Далі, на будь-якому кроці ітерації кожна ланка замінюється на утворюючий елемент – ламану, що складається по краях з відрізків довжиною 1/3 від довжини ланки, між якими розміщаються дві сторони правильного трикутника із стороною в 1/3 довжини ланки. Всі відрізки – сторони отриманої кривої вважаються базовими ланками для наступної ітерації. Крива, одержувана в результаті n-ї ітерації при будь-якому кінцевому n, називається предфракталом, і лише при n, що прямує до нескінченності, крива Коха стає фракталом. Одержана в результаті ітераційного процесу фрактальна мно-

жина являє собою лінію нескінченної довжини, що обмежує кінцеву площу. Дійсно, при кожному кроці число сторін результуючого багатокутника збільшується у 4 рази, а довжина кожної сторони зменшується тільки у 3 рази, тобто довжина багатокутника на n-й ітерації дорівнює $3*(4/3)n$ і прямує до нескінченності з ростом n.

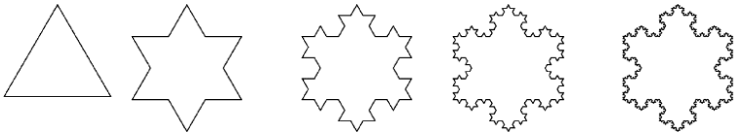


Рисунок 26 – Перші 5 поколінь сніжинки Коха

Площа під кривою, якщо прийняти площу утворюючого трикутника за 1, дорівнює:

$$S = 1 + 1/3 \sum_{k=0}^{\infty} (4/9)^k = 1{,}6.$$

У 80-х роках XX століття як простий метод одержання фрактальних структур з'явився метод "Систем Ітераційних Функцій" (Iterated Functions System – IFS). IFS являє собою систему функцій, що відображають одну багатомірну множина на іншу. Найбільш простою реалізацією IFS є афінні перетворення площини:

$$X' = A \times X + B \times Y + C;$$

$$Y' = D \times X + E \times Y + F.$$

У 80-х роках американські вчені М. Барнслі та А. Слоан запропонували ідею стиску та зберігання графічної інформації, засновану на міркуваннях теорії фракталів і динамічних систем. На підставі цієї ідеї був створений алгоритм фрактального стиску інформації, що дозволяє стискати деякі зразки графічної інформації у 500-1000 разів. При цьому кожне зображення кодується декількома простими афінними перетвореннями.

За алгоритмом Барнслі відбувається виділення в зображенні пар областей, менша з яких подібна більшій, і збереження декількох коефіцієнтів, які кодують перетворення, що пере-

водить більшу область у меншу. Потрібно, щоб множина таких областей покривало все зображення.

Як приклад використання IFS для побудови фрактальних структур, можна навести криву «дракону» Хартера-Хейтуея (Рис. 27). IFS застосовується для стиску зображень, наприклад, фотографій, що засновано на виявленні локальної самоподібності (на відміну від фракталів, де спостерігається глобальна самоподібність).

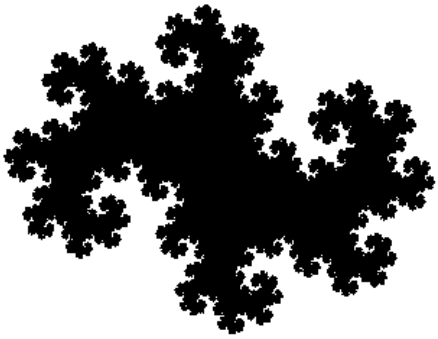


Рисунок 27 – «Дракон» Хартера-Хейтуея

Фрактали у природі

Один із кращих прикладів прояву фракталів у природі – структура берегових ліній. Дійсно, інколи на кілометровому відрізку узбережжя виглядає настільки ж порізаним, як і на стокілометровому.

Досвід показує, що довжина берегової лінії L залежить від масштабу l, яким проводяться виміри, і збільшується із зменшенням останнього за степеневим законом $L = \Lambda l^{1-\alpha}, \Lambda = const.$ Так, наприклад, для узбережжя Великобританії $\alpha \approx 1.24$, тобто фрактальна розмірність берегової лінії Великобританії дорівнює 1.24.

Інформаційний простір і фрактали

На цей час інформаційний простір прийнято розглядати як стохастичний. У багатьох моделях інформаційного простору вивчаються структурні зв'язки між тематичними множинами, що входять у цей простір. Самоподібність інформаційного простору виражається, насамперед у тому, що при його лавино-

подібному зростанні, частотні та рангові розподіли, одержувані в таких розрізах, як джерела, автори, тематика практично не міняють своєї форми. Застосування теорії фракталів при аналізі інформаційного простору дозволяє із загальної позиції глянути на закономірності, що становлять основи інформатики. Наприклад, тематичні інформаційні масиви сьогодні представляють самоподібні структури, що розвиваються і за своєю суттю є стохастичними фракталами, тому що їхня самоподібність справедлива лише на рівні математичних очікувань, як наприклад, розподілу кластерів за розмірами.

В інформаційному просторі виникають, формуються, ростуть і розмножуються кластери – групи взаємозалежних документів. Системи, засновані на кластерному аналізі, самостійно виявляють нові ознаки об'єктів і розподіляють об'єкти за новими групами.

Фрактальні властивості характерні для кластерів інформаційних веб-сайтів, на яких публікуються документи, що відповідають певним тематикам. Ці кластери, як набори тематичних документів, являють собою фрактальні структури, що мають низку унікальних властивостей.

Топологія та характеристики моделей веб-простору виявляються приблизно однаковими для різних підмножин, підтверджуючи тим самим спостереження про те, що "веб – це фрактал", тобто властивості структури всього веб-простору Bow Tie вірні і для його окремих підмножин.

З іншого боку, теорія фракталів розглядається як підхід до статистичного дослідження, що дозволяє одержувати важливі характеристики інформаційних потоків, не вдаючись у детальний аналіз їхньої внутрішньої структури та зв'язків. Для послідовності повідомлень тематичних інформаційних потоків у відповідності зі скейлинговим принципом, кількість повідомлень, резонансів на події реального миру, пропорційна деякому ступеню кількості джерел інформації (кластерів). Відомо, що всі основні закони наукової комунікації, такі як закони Парето, Лотки, Бредфорда, Ципфа, можуть бути узагальнені саме в рамках теорії стохастичних фракталів. Точно так само, як й у традиційних наукових комунікаціях, множина повідомлень в Інтернеті за однією тематикою в часі являє собою динамічну кластерну систему, що виникає в результаті ітераційних про-

цесів. Цей процес обумовлюється републікаціями, однобічним або взаємним цитуванням, різними публікаціями – відбиттями тих самих подій реального миру, прямими посиланнями тощо.

Фрактальна розмірність у кластерній системі, що відповідає тематичним інформаційним потокам, показує ступінь заповнення інформаційного простору повідомлень протягом певного часу:

$$N_{publ}(\varepsilon t) = \varepsilon^{\rho} N_k(t)^{\rho},$$

де N_{publ} – розмір кластерної системи (загальне число документів в інформаційному потоці); N_k – розмір – число кластерів (тематик або джерел); ρ – фрактальна розмірність інформаційного масиву; ε – коефіцієнт масштабування. У наведеному співвідношенні між кількістю документів і кластерів проявляється властивість збереження внутрішньої структури множини при зміні масштабів його зовнішнього розгляду.

Метод DFA

Для дослідження часових рядів сьогодні усе ширше використовується теорія фракталів. Часові ряди, породжувані тематичними інформаційними потоками, зокрема, мають фрактальні властивості та можуть розглядатися як стохастичні фрактали. Цей підхід розширює область застосування теорії фракталів на інформаційні потоки, динаміка яких описується засобами теорії випадкових процесів.

Метод DFA (Detrended Fluctuation Analysis) являє собою варіант дисперсійного аналізу, що дозволяє досліджувати ефекти тривалих кореляцій у нестаціонарних рядах. При цьому аналізується середньоквадратична помилка лінійної апроксимації в залежності від розміру відрізку апроксимації. У рамках цього методу спочатку здійснюється приведення даних до нульового середнього (вирахування середнього значення $\langle F \rangle$ з вихідного часового ряду F_n, $n = 1,..., N$) і будується випадкове блукання $y\ k$:

$$y(k) = \sum_{n=1}^{N} [F(n) - < F >_N].$$

Потім ряд значень $y\ k$, $k=1,...,N$ розбивається на відрізки, що не перекриваються довжини n, у межах кожного з яких методом найменших квадратів визначається рівняння прямої, що найкраще за критерієм χ^2 апроксимує послідовність $y\ k$.

Знайдена апроксимація $y_n\ k$ ($y_n\ k = ak+b$) розглядається як локальний тренд. При цьому коефіцієнти a та b обчислюються таким чином:

$$a = \frac{n\sum ky(k) - (\sum k)(\sum y(k))}{n\sum k^2 - (\sum k)^2};$$

$$b = \frac{(\sum y(k))(\sum k^2) - (\sum k)(\sum ky(k))}{n\sum k^2 - (\sum k)^2}.$$

Далі обчислюється середньоквадратична помилка лінійної апроксимації у широкому діапазоні значень n. Вважається, що залежність D(n) часто має ступеневий характер $D\ n \sim n^{\alpha}$, тобто наявність лінійної ділянки в подвійному логарифмічному масштабі $\lg D\ \lg n$ дозволяє говорити про існування скейлингу.

Як видно по Рис. 28, $D\ n$ для побудованого модельного ряду ступеневим чином залежить від n, тобто в подвійному логарифмічному масштабі ця залежність близька до лінійної.

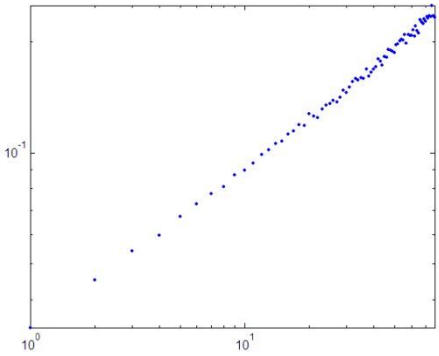


Рисунок 28 – Залежність $D\ n$ модельного ряду (вісь *Y*) від довжини відрізку апроксимації n (вісь *X*) у подвійній логарифмічній шкалі

Фактор Фано

Для вивчення поводження процесів і підтвердження їх самоподібності прийнято використати ще один показник – індекс розкиду дисперсії (IDC), так званий фактор Фано. Ця величина визначається як відношення дисперсії числа подій (у нашому випадку – точок ряду) на вікні спостережень заданої ширини k до відповідного математичного очікування:

$$F(k) = \sigma^2(k) / m(k).$$

Для самоподібних процесів виконується співвідношення:

$$F(k) = 1 + C \cdot k^{2H-1},$$

де C й H – константи. На Рис. 29. наведено графік значень $F(k)$ у подвійному логарифмічному масштабі.

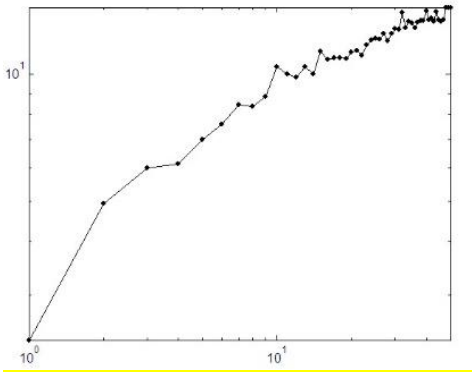


Рисунок 29 – Залежність фактора Фано від ширини вікна спостережень

Показник Херста

Показник Херста (H) зв'язують із коефіцієнтом нормованого розмаху (R/S), де R – «розмах» часового ряду, а S – стандартне відхилення.

Херст експериментально виявив, що для багатьох часових рядів справедливо: $R / S = (N / 2)^H$. Показник Херста пов'язується з традиційною фрактальною розмірністю (D) простим співвідношенням:

$D = 2 - H$.

Відомо, що показник Херста являє собою міру персистентності – схильності поведінки процесу до трендів (на відміну від звичайного броунівського руху). Значення $H > 1/2$ означає, що спрямована в певну сторону динаміка процесу в минулому, найімовірніше, спричинить продовження руху у тому ж напрямку. Якщо $H < 1/2$, то прогнозується, що процес змінить спрямованість. $H = 1/2$ означає невизначеність – броунівський рух.

Для вивчення фрактальних характеристик часових рядів $F(n)$, $n = 1,...,N$, досліджуються значення показника Херста, який визначається із співвідношення:

$$R(N)/S_N \cong \left(N/2\right)^H, \quad N >> 1.$$

Тут S_N – стандартне відхилення:

$$S_N = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(F(n) - \left\langle F\right\rangle_N\right)^2},$$

$$\left\langle F\right\rangle_N = \frac{1}{N}\sum_{n=1}^{N} F(n),$$

R – так званий розмах:

$$R(N) = \max_{1\le n\le N} X(n,N) - \min_{1\le n\le N} X(n,N),$$

де:

$$X(n,N) = \sum_{i=1}^{n}\left(F(i) - \left\langle F\right\rangle_N\right).$$

Якщо розмір ряду вимірів N досить великий, то розрахунок середнього значення і стандартного відхилення при зростанні N за наведеною формулою не є раціональним. Більш економними з обчислювальної точки зору є ітераційні методи. Обґрунтуємо формулу розрахунку середнього $\left\langle F\right\rangle_{N+1}$, якщо відомо N, $F(N+1)$, $\left\langle F\right\rangle_N$:

$$\left\langle F\right\rangle_{N+1} = \frac{1}{N+1}\sum_{n=1}^{N+1} F(n) = \frac{1}{N+1}\left(\sum_{n=1}^{N} F(n) + F(N+1)\right) = \frac{1}{N+1}\left(N\left\langle F\right\rangle_N + F(N+1)\right).$$

Формула для обчислення стандартного відхилення S_{N+1} на основі значень N, $F(N)$, $F(N+1)$, $\langle F\rangle_{N+1}$, S_N обґрунтовується дещо складніше:

$$S_{N+1}^2 = \frac{1}{N+1}\sum_{n=1}^{N+1}\left(F(n) - \langle F\rangle_{N+1}\right)^2,$$

$$(N+1)S_{N+1}^2 = \sum_{n=1}^{N+1}\left(F(n)^2 - 2F(n)\langle F\rangle_{N+1} + \langle F\rangle_{N+1}^2\right) =$$

$$= \sum_{n=1}^{N} F(n)^2 + F(N+1)^2 - 2\langle F\rangle_{N+1}\sum_{n=1}^{N} F(n) - 2\langle F\rangle_{N+1}F(N+1) + (N+1)\langle F\rangle_{N+1}^2.$$

Те ж саме для попереднього значення:

$$NS_N^2 = \sum_{n=1}^{N}\left(F(n)^2 - 2F(n)\langle F\rangle_N + \langle F\rangle_N^2\right) = \sum_{n=1}^{N} F(n)^2 - N\langle F\rangle_N^2,$$

Звідки:

$$\sum_{n=1}^{N} F(n)^2 = NS_N^2 + N\langle F\rangle_N^2.$$

Підставивши це значення у рівняння для S_{N+1}^2, отримуємо:

$$S_{N+1}^2 = \frac{1}{N+1}\left(NS_N^2 + N\langle F\rangle_N^2 + F(N+1)^2 - (N+1)\langle F\rangle_{N+1}^2\right).$$

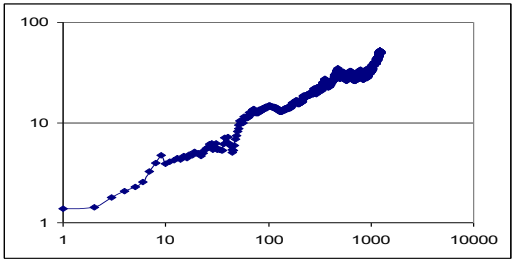


Рисунок 5.24 – Динаміка зміни показника значень *R/S* для модельного ряду у подвійній логарифмічній шкалі

Після виконання відповідної програми (для різних значень N) в інтерфейсі користувача буде відображено список значень *R/S*. Далі цей список через буфер обміну завантажується до інтерфейсу програми Microsoft Excel. Після цього засобами цієї програми будується графік, у якому вказується логарифмічна

шкала для обох осей (Рис. 5.30). Використовуючи засоби апроксимації можна переконатися у справедливості степеневого співвідношення для значень R/S.

2.6.7 ΔL-метод

Скейлограми, одержані за допомогою безперервного вейвлет-перетворення, використовують для візуалізації особливостей часового ряду. Запропоновано інший метод візуалізації, який також допомагає виявити тренди, періодичності та локальні особливості ряду[28]. Запропонований підхід значно простіше у реалізації, ніж вейвлет аналіз.

ΔL-метод базується на методі DFA (Detrended Fluctuation Analysis). Суть підходу полягає у визначенні та відображенні абсолютного відхилення точок ряду накопичених значень від відповідних значень лінійної апроксимації.

Опишемо ΔL-метод більш детально. Для початку зафіксуємо деяку ширину вікна s (масштаб, на якому розглядається ряд). Розглянемо точку x_l та виберемо для неї вікно ширини s так, щоб точка l була в центрі цього вікна (або зміщена на 1, якщо s парне). Побудуємо лінійну апроксимацію по точках вікна і позначимо $\Delta_{l,j,s}$ значення локальної апроксимації в точці j для відрізка з центром l. Далі обчислимо абсолютне відхилення x_j (Рис. 30) від лінії апроксимації $\Delta_{l,j,s} = \left| x_j - L_{t,j,s} \right|$.

Метод передбачає обчислення значень $\Delta_{l,j,s}$ для всіх $l = 1,\ldots,T$ і вікон шириною $s = 1,\ldots,[T/4]$. Для фіксованої ширини вікна обчислюється середнє квадратичне відхилення:

---

[28] Lande D.V., Snarskii A.A. Diagram of measurement series elements deviation from local linear approximations. Preprint ArXiv 0903.3328, 2009. DOI: https://doi.org/10.48550/arXiv.0903.3328

$$E(l,s)=\sqrt{\frac{1}{s}\sum_{j}\left|x_j-L_{l,j,s}\right|^2}=\sqrt{\frac{1}{s}\sum_{j}\Lambda_{l,j,s}{}^2}.$$

$$\mathrm{E}(\mathrm{l},\mathrm{s})=\sqrt{\frac{1}{\mathrm{s}}\sum_{\mathrm{j}}\left|\mathrm{x_j}-\mathrm{L_{l,j,s}}\right|^2}=\sqrt{\frac{1}{\mathrm{s}}\sum_{\mathrm{j}}\Delta_{\mathrm{l,j,s}}{}^2}.$$

Отримані значення демонструються на діаграмі, схожій скейлограму. Приклади таких діаграм показано на Рис. 31.

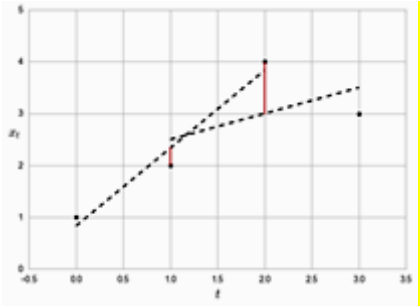


Рисунок 30 – Чотири точки часового ряду з лінійною апроксимацією для двох вікон із шириною три. Також показано відхилення $\Delta_{l,j,s}$ центральної точки вікна від відповідної лінійної апроксимації

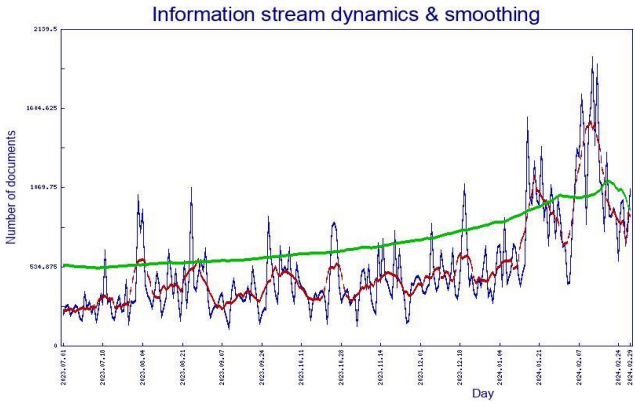


Рисунок 31 – Реальна та зглажена динаміка інформційного потоку

Запропонований метод візуалізації абсолютних відхилень ΔL, як і метод вейвлет-перетворень, дозволяє виявляти поодинокі та нерегулярні «сплески», різкі зміни значень кількісних показників у різні періоди часу, а також гармонійні складові у ряді.

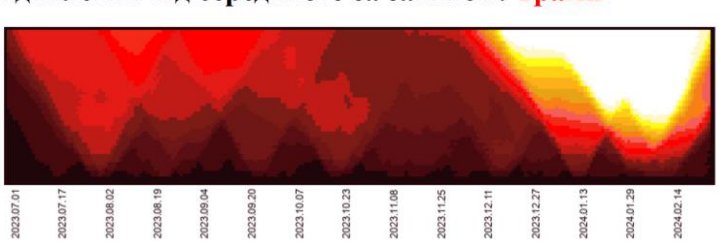


Рисунок 31 – Вихідний ряд динаміки повідомлень за запитом «Трамп» та коефіценти, отримані за допомогою ΔL-методу

2.6.8. Мережеві моделі

Останнім часом виокремився окремий науковий напрям – аналіз соціальних мереж (SNA, Social Networks Analysis), який базується, з одного боку, на соціології, а з іншого – на теорії складних мереж (Complex Networks)[29]. У межах теорії складних мереж вивчаються не лише топологічні характеристики мереж, а й статистичні феномени, розподіл ваг окремих вузлів та ребер, ефекти протікання та провідності. Незважаючи на те, що об'єктом розгляду теорії складних мереж є різноманітні мережі (електричні, транспортні, інформаційні), найбільший внесок у розвиток цієї теорії зробили дослідження соціальних мереж. У теорії складних мереж виділяють три основні напрями:

- дослідження статистичних властивостей, які характеризують поведінку мереж;
- створення моделей мереж;
- прогнозування поведінки мереж при зміні структурних властивостей.

Застосування мережевих моделей у сфері OSINT відноситись до аналізу: соціальних мереж, комунікацій, географічних ме-

[29] Newman M.E.J. The structure and function of complex networks // SIAM Review, 2003. – 45. – P. 167-256.

реж, економічних і фінансових мереж, технічних мереж, взаємодій між різними суспільними групами та структурами.

При аналізі соціальних мереж досліджуються взаємодії та зв'язки між особами або організаціями в соціальних мережах для виявлення ключових фігур, груп чи структур. При цьому можливе виявлення мережевих зв'язків між особами, визначення важливих учасників чи організацій.

При аналізі комунікацій здійснюється вивчення патернів комунікацій та обміну інформацією між об'єктами, такими як електронні повідомлення, телефонні дзвінки тощо. При цьому можливе виявлення ключових зв'язків та аналіз обміну інформацією для розуміння ситуації чи виявлення загроз.

Представлення географічної інформації у вигляді мережі для аналізу просторових взаємозв'язків здійснюється для мапування географічних зв'язків між об'єктами, виявлення географічних патернів та аналіз руху.

Економічні та фінансові мережі досліджуються для вивчення фінансових транзакцій, взаємодій між компаніями чи фінансовими інституціями, виявлення фінансових патернів, аналіз ринкових взаємодій та виявлення економічних тенденцій.

Технічні мережі вивчаються для моделювання технічних взаємодій, таких як комп'ютерні мережі, взаємодія пристроїв тощо, виявлення інформаційних загроз, аналіз взаємодій в комп'ютерних системах.

Вивчення взаємодій між різними суспільними групами та структурами здійснюється для аналіз соціокультурних зв'язків, виявлення групових динамік та ідентифікація потенційних напруг у суспільстві.

Мережевий аналіз у сфері конкурентної розвідки допомагає розкривати складні зв'язки та структури в інформації, отриманій з відкритих джерел. Це сприяє покращенню розуміння контексту подій, виявленню ключових фігур та формуванню стратегій аналізу для прийняття обґрунтованих рішень. Розглянемо деякі теоретичні поняття щодо застосування мережевого аналізу у цій сфері:

Мережевий аналіз спрямований на вивчення та розуміння взаємозв'язків між об'єктами (вузлами) в системі, де кожен вузол може представляти особу, організацію, ресурс чи інший

об'єкт. Це застосовується в OSINT для виявлення ключових акторів, визначення патернів комунікацій та інформаційних потоків між суб'єктами.

Аналіз характеристик окремих вузлів та властивостей мережі, таких як ступінь взаємодії, тип зв'язків, географічні координати тощо застосовується для визначення важливих вузлів, аналіз географічних аспектів взаємодій, визначення ключових властивостей об'єктів.

Визначення ступеня централізації в мережі шляхом аналізу ролі та важливості кожного вузла застосовується для виявлення центральних фігур або вузлів, які мають великий вплив у мережі, ідентифікація ключових посередників.

Кластерний аналіз мереж, розділення мережі на групи (кластери) вузлів, які взаємодіють між собою частіше, ніж з вузлами інших кластерів застосовується в OSINT для виявлення груп, субкультур або сфер взаємодії, визначення групових динамік та об'єднань.

Використання мережевого аналізу для розроблення сценаріїв діяльності, подій чи взаємодій на основі вивчення зв'язків у мережі застосовується в OSINT для створення прогнозів або аналітичних сценаріїв на основі зрозумілого розуміння взаємодій та властивостей об'єктів у мережі.

Мережі взаємозв'язку

Мережі взаємозв'язку об'єктів, сутностей, властивостей вузлів і мереж є ключовим поняттям у мережевому аналізі.

Розглянемо основні поняття у мережевому аналізі в OSINT, які дозволяють аналізувати та розуміти складні структури взаємозв'язків між різними об'єктами:

Мережа взаємозв'язку об'єктів представляє собою граф, де вузли (вузли мережі) представляють об'єкти чи сутності, а ребра (зв'язки) вказують на взаємодії або зв'язки між цими об'єктами. Об'єкти можуть бути різними, такими як особи, організації, ресурси, події тощо.

Сутність у мережі представляє об'єкт або елемент, який має свою унікальну ідентифікацію та може взаємодіяти з іншими сутностями у мережі. Це може бути, наприклад, конкретна особа, компанія, локація чи подія.

Кожен вузол мережі може мати свої властивості, які характеризують його особливості чи атрибути. Наприклад, у випадку соціальної мережі, властивості можуть включати ім'я, вік, місце проживання та інші характеристики для осіб.

Мережевий граф представляється у вигляді вузлів та ребер. Вузли представляють об'єкти, а ребра – взаємодії між цими об'єктами. Кожне ребро може мати свої властивості, які вказують на характер взаємодії.

Взаємодії між вузлами у мережі можуть бути різного типу. Це може включати фізичні зв'язки, соціальні взаємодії, обмін інформацією, фінансові транзакції та інше. Взаємодії можуть бути спрямованими або неспрямованими, важливими чи менш важливими.

У мережах можна виділити групи взаємодіючих вузлів, відомі як спільноти або кластери. Спільнота – це група вузлів, які взаємодіють частіше між собою, ніж з вузлами інших груп.

Для вимірювання та оцінки характеристик мережі використовуються різні метрики, такі як ступінь вузла, центральність, власне значення та інші. Ці метрики допомагають визначити важливість та роль кожного вузла у мережі.

Показники центральності

Показники центральності у мережевому аналізі визначають ступінь важливості або центральності вузлів у мережі. Ці показники дозволяють аналізувати різні аспекти центральності об'єктів у мережі, і вони можуть бути використані в OSINT для виявлення ключових фігур, груп чи об'єктів, які можуть мати важливий вплив на структуру та динаміку мережі. Наведемо деякі з них:

Ступінь вузла (Degree Centrality) визначає кількість зв'язків, які виходять чи входять у вузол. Визначається як загальна кількість зв'язків вузла. Зокрема, в системах OSINT визначає те, як часто об'єкт взаємодіє з іншими у мережі. Великий ступінь може свідчити про важливість об'єкта у мережі.

Близькість вузла (Closeness Centrality) – обернене значення середньої довжини найкоротшого шляху між вузлом та всіма іншими вузлами у мережі. Застосовується в OSINT для визначення того, як швидко об'єкт може отримати доступ до інших

об'єктів у мережі. Висока близькість може означати велику впливовість об'єкта.

Центральність по посередництву (Betweenness Centrality) – кількість найкоротших шляхів між парою вузлів, які проходять через даний вузол. Застосується для виявлення вузлів, які контролюють потік інформації між іншими вузлами. Велике значення може означати ключову роль в мережі.

Власне значення вузла (Eigenvector Centrality) – міра впливовості вузла, яка враховує інших важливих вузлів, з якими він пов'язаний. Визначення важливості вузла з урахуванням впливу його зв'язків. Об'єкти з високим власним значенням можуть мати велику впливовість в мережі.

Кластерний аналіз мереж

Кластерний аналіз мереж в задачах OSINT є методом групування вузлів мережі у класи або кластери на основі схожості їхніх взаємодій. Цей підхід дозволяє виявити структури та патерни у мережі, визначити групи взаємодіючих елементів і отримати краще розуміння глобальної організації мережі. Кластерний аналіз може забезпечувати додаткові засоби для розуміння групових зв'язків, організаційної структури об'єктів у великих мережах, що розглядаються.

Наведемо деякі напрямки кластерного аналізу мереж у контексті OSINT:

Для кластерного аналізу мереж використовують різні алгоритми, такі як це алгоритм на основі схожості, який визначає кластери за допомогою двох параметрів: радіуса є та мінімальної кількості точок, алгоритми агломеративної кластеризації, алгоритми кластеризації за модулярністю тощо. Кожен з них має свої переваги та обмеження.

Схожість між вузлами може визначатися на основі різних критеріїв, таких як частота взаємодій, схожість атрибутів, близькість у географічному просторі тощо. Важливо вибрати міру схожості, яка відображає специфіку ваших даних.

Однією з ключових характеристик кластерного аналізу є визначення оптимальної кількості кластерів. Існують різні методи для цього, зокрема, ті що базуються на застосуванні індексів кластеризації – статистичних показників, які оцінюють якість

кластеризації. Наприклад, індекс внутрішньої щільності, індекс зовнішньої щільності, індекс дисперсії.

Після кластеризації важливо проаналізувати структуру отриманих кластерів. Це включає виявлення внутрішньокластерних та міжкластерних зв'язків, визначення представників кожного кластеру та розуміння їхньої ролі в мережі.

Визначення ключових вузлів – центроїдів, ключових вузлів у кожному кластері, які можуть мати важливий вплив на структуру та функціонування групи.

Кластерний аналіз може допомагати виявляти спільноти або субкультури в мережі, що може бути корисним у визначенні групових динамік та взаємодій.

Наведемо декілька найбільш поширених алгоритмів кластеризації мереж:

Алгоритм k-середніх[30] – це алгоритм поділу та об'єднання, який починається з k випадково вибраних вершин. Потім вершини призначаються до кластера, який має найближчу середню. Ці кластери потім об'єднуються з іншими кластерами, поки не буде отримано бажану кількість кластерів.

Алгоритм DBSCAN[31] (density-based spatial clustering of applications with noise): це алгоритм на основі схожості, який визначає кластери за допомогою двох параметрів: радіуса ε та мінімальної кількості точок. Вершина вважається ядром, якщо вона має не менше minPts (мінімальне число точок, які повинні утворювати щільну область) сусідніх точок, які знаходяться в радіусі ε. Вершина вважається границею, якщо вона має менше minPts сусідніх точок, які знаходяться в радіусі ε, але при цьому є сусідом ядра. Вершина вважається шумом, якщо вона

[30] J. MacQueen (1967). Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (University of California Press): 281–297.

[31] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) / Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. – AAAI Press, 1996. – С. 226–231.

не є ядром і не є границею.

Алгоритм асоціаційної кластеризації (Association clustering algorithm) – це алгоритм поділу та об'єднання, який використовує асоціативні правила для визначення кластерів. Асоціативне правило визначає, що дві або більше вершин є частиною одного кластера, якщо між ними існує певна асоціація. Наприклад, правило «якщо вершина є членом спільноти A, то вона також є членом спільноти B».

Алгоритм на основі кластеризації за ієрархією[32] (HAC, Hierarchical Agglomerative Clustering) – це алгоритм агломерації, який створює ієрархічну структуру кластерів. Ієрархія створюється шляхом об'єднання кластерів один з одним, поки не буде отримано один кластер.

Вибір алгоритму кластеризації мереж залежить від конкретних завдань, які необхідно вирішити. Наприклад, алгоритм k-середніх може бути хорошим вибором для завдань, в яких необхідно отримати кластери з рівномірною кількістю вершин. Алгоритм DBSCAN може застосовуватись для завдань, в яких необхідно отримати кластери з певною формою. Алгоритм асоціаційної кластеризації може застосовуватись для завдань, в яких необхідно отримати кластери на основі асоціативних правил.

Формування сценаріїв на основі аналізу мереж

Формування сценаріїв на основі аналізу мереж в рамках OSINT – це процес генерування можливих сценаріїв подій на основі аналізу даних, отриманих з мереж. Цей процес може бути корисним для різних цілей, таких як:

- Прогнозування майбутніх подій
- Розслідування злочинів або інцидентів
- Аналіз конкурентів

При цьому важливо пам'ятати, що сценарії, сформовані на основі аналізу мереж, є лише гіпотезами. Вони можуть бути корисними для отримання нових знань і розуміння складних

[32] Nielsen, Frank (2016). "Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.

систем, але вони не можуть гарантувати точність.

Для формування сценаріїв на основі аналізу мереж використовуються різні методи, такі як:

Аналіз мережевих структур: цей метод використовується для виявлення зв'язків між різними вузлами мережі. Наприклад, можна використовувати цей метод для виявлення спільнот в соціальних мережах або мереж злочинних організацій.

Аналіз мережевих потоків: цей метод використовується для виявлення напрямку і частоти передачі інформації в мережі. Наприклад, можна використовувати цей метод для виявлення поширення дезінформації або пропаганди.

Аналіз мережевих взаємодій: цей метод використовується для виявлення взаємодій між різними вузлами мережі. Наприклад, можна використовувати цей метод для виявлення можливих ризиків для безпеки.

Ось деякі приклади використання формування сценаріїв на основі аналізу мереж:

Для прогнозування майбутніх подій можна використовувати аналіз мережевих структур для виявлення тенденцій і закономірностей. Наприклад, можна використовувати цей метод для прогнозування майбутніх політичних подій або економічних криз.

Для розслідування злочинів або інцидентів можна використовувати аналіз мережевих структур для виявлення зв'язків між різними учасниками. Наприклад, можна використовувати цей метод для розслідування хакерських атак або терористичних актів.

Для аналізу конкурентів можна використовувати аналіз мережевих структур для виявлення їхніх стратегій і тактик. Наприклад, можна використовувати цей метод для виявлення можливих загроз для бізнесу.

Формування сценаріїв на основі аналізу мереж є складним завданням, оскільки вимагає аналізу великої кількості даних і використання спеціальних методів. Однак цей процес може бути дуже корисним для отримання нових знань і розуміння складних систем.

2.7. Реалізовані технології конкурентної розвідки

2.7.1 Реалізація аналітичних складових в OSINT

Системи OSINT широко використовуються для збору, аналізу та вивчення великих обсягів відкритої інформації з різних джерел для розвідки та інтелектуального аналізу, створювання зв'язків та надання інтелектуальних висновків. Вони використовуються в різних галузях, таких як кібербезпека, правоохоронна діяльність, розвідка та бізнес-аналітика. Назвемо декілька прикладів таких систем:

Palantir Gotham[33] – аналітична система, що використовується для боротьби з тероризмом в офісах розвідувального співтовариства США (USIC) і Міністерства оборони США для обробки та аналізу великих обсягів різноманітної інформації для виявлення зв'язків та патернів. Palantir Gotham – це продукт, спрямований на використання державними спецслужбами. Початкова назва продукту – Government – відображала його специфічну орієнтацію на сектор державних служб. Основний механізм зберігання в Palantir Gotham використовує техніку онтологій, яка дозволяє різнорідні дані з різних джерел отримати смислову інформацію та уніфікувати їх для спільного аналізу. Засоби семантичного пошуку в платформі використовують можливості складання складних запитів до онтологій, включаючи пошук за близькістю значень і використання фонетичного пошуку. Користувачі-аналітики можуть зберігати встановлені факти і взаємозв'язки в семантичній формі, що далі беруть участь у подальшому аналізі. Платформа також підтримує підключення генетичних алгоритмів, які спеціально розробляються для конкретних сфер діяльності. Крім того, важливою складовою частиною платформи є підсистема групової роботи користувачів, яка дозволяє аналітикам обмінюватися повідомленнями та результатами аналізу.

Maltego (https://www.maltego.com) – це інструмент для інтерактивний інструмент для видобутку даних, аналізу великої кількості відкритої інформації для визначення зв'язків та залежностей між різними об'єктами, надає спрямовані графіки

[33] https://www.palantir.com/platforms/gotham/

для вивчення зв'язків. Пропонується як інструмент для графічного аналізу інтернет-посилань, пошукового інструменту у відкритих джерелах Інтернету в реальному часі та збору інформації, а також подання цієї інформації візуально на основі графів, завдяки чому шаблони та зв'язки між різними джерелами та відповідною інформацією легко ідентифікуються. Засіб використовується в онлайн-розслідуваннях для аналізу взаємозв'язків між різними частинами інформації з різних джерел в Інтернеті[34].

За допомогою Maltego ви можете видобувати дані з розподілених джерел, автоматично об'єднувати відповідну інформацію в одному графі та візуально наносити її на карту, щоб дослідити ваш ландшафт даних. Maltego пропонує можливість підключати дані та функції з різних джерел, використовуючи Transforms. Через Transform Hub ви можете підключити дані понад 30 партнерів, таких як Recorded Future, DomainTools, Crowdstrike, ThreatConnect, та різноманітні загальнодоступні джерела (OSINT), а також власні внутрішні дані. Однак для роботи з власними даними потрібні їх доволі важка конвертація та розроблення логістичної моделі та моделі асоціативних правил. Застосовується здебільшого для OSINT.

IBM i2 Analyst's Notebook (https://i2group.com/i2-analysts-notebook) є аналітичною системою, яка була започаткована на початку 2000 років, має кілька версій (Рис. 32). Систему орієнтовано на побудову різноманітних схем, але потрібна велика кількість операцій ручної обробки даних, бо вона не дуже пристосована для роботи з Big Data та Big Stream Data, має складну систему конвертації зовнішніх даних, доволі важку систему ГІС, потребує багато ресурсів і є коштовною.

IBM i2 Analyst's Notebook надає візуальне аналітичне середовище, яке дає змогу максимально ефективно використовувати величезні обсяги інформації, накопичені державними службами та підприємствами. Завдяки інтуїтивно зрозумілому ін-

[34] Узлов Д.Ю., Струков В.М. Сучасні інструментальні засоби кримінального аналізу. Проблеми застосування інформаційних технологій правоохоронними структурами України та вищими навчальними закладами зі специфічними умовами навчання : зб. наук. ст. за матеріалами доп. Між-нар. наук.-практ. конф. (м. Львів, 22 груд. 2017 р.) МВС України, Львів. держ. ун-т внутр. справ. Львів, 2017. С. 162–164.

терфейсу з урахуванням контексту воно дає можливість аналітикам швидко зіставляти, аналізувати і наочно уявляти дані з різних джерел, скорочуючи час на пошук важливої інформації у складних даних. IBM i2 Analyst's Notebook надає актуальні і дієві аналітичні засоби, що допомагають виявляти, передбачати, запобігати і припиняти злочинну, терористичну і шахрайську діяльність[35].

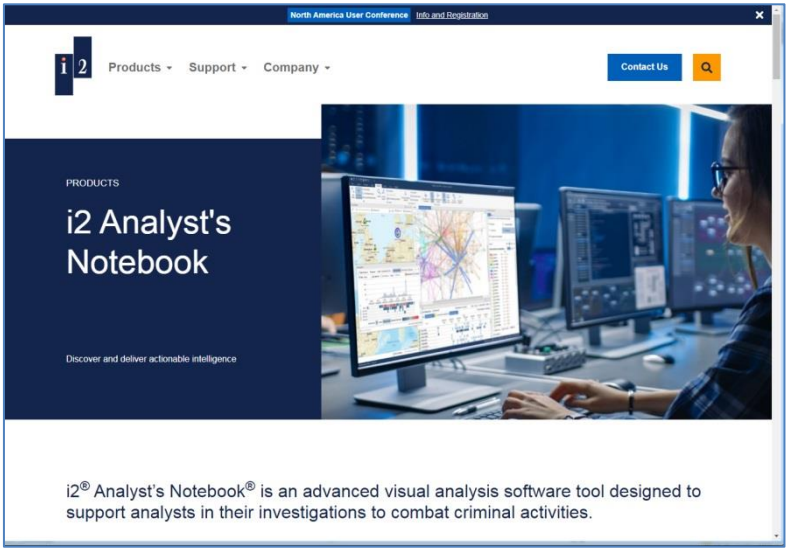


Рисунок 32 – стартова веб сторінка IBM i2 Analyst's Notebook

IBM i2 Analyst's Notebook допомагає вирішувати такі завдання:

- швидка систематизація розрізнених даних в єдиному узгодженому поданні;
- визначення ключових осіб, подій, зв'язків і закономірностей, які не завжди можна виявити іншими засобами;

[35] Корнейко О. В., Школьніков В. І., Овсянюк Д. І. Використання сучасних інформаційно-аналітичних технологій в діяльності центру кримінальної аналітики Національної академії внутрішніх справ // Інформаційні технології в освіті та практиці : матеріали Всеукр. наук.-практ. конф. (м. Львів, 18 груд. 2020 р.) / МВС України, Львів. держ. ун-т внутр. справ. Львів, 2020. С. 8–11.

- поліпшене розуміння структури, ієрархії і способів дій злочинних, терористичних і шахрайських організацій;
- спрощення обміну складними даними, що дає змогу ухвалювати своєчасні й точні оперативні рішення;
- можливість отримання вигоди завдяки швидкому впровадженню, яке забезпечує швидке зростання продуктивності, завдяки надійним рішенням для візуальної аналітики.

Recorded Future[36] – система від найбільшої розвідувально-аналітичної компанії у світі, що використовує штучний інтелект для аналізу і передбачення подій на основі великої кількості відкритих джерел. Зокрема, Recorded Future Intelligence Graph збирає, структурує та аналізує дані про загрози з усіх куточків Інтернету, перетворюючи великі обсяги даних на практичні висновки. Система збирає та структурує дані противника та жертви з тексту, зображень і технічних джерел, а також використовує обробку природної мови та машинне навчання для аналізу та відображення асоціацій між об'єктами у режимі реального часу.

IntelTechniques[37] – інструменти для пошуку і аналізу великої кількості відкритої інформації з метою виявлення та слідкування за особами та іншими об'єктами від Майкла Базелла. Ці інструменти були створені як додаток до книги Open Source Intelligence Techniques, 9th Edition та онлайн-навчання OSINT IntelTechniques[38].

Hunchly[39] – інструмент, який допомагає аналізувати та збирати інформацію з Інтернету, забезпечуючи можливість зберігання даних та навігації.

OSINT Combine[40] – інтегрована платформа, яка забезпечує інструменти для збору та аналізу великої кількості відкритої інформації.

Надалі наведемо детальний опис деяких систем.

[36] https://www.recordedfuture.com/
[37] https://www.inteltechniques.net/courses/open-source-intelligence
[38] https://inteltechniques.com/book1.html
[39] https://www.hunch.ly/
[40] https://www.osintcombine.com/

2.7.2 Palantir

Palantir Technologies (https://www.palantir.com/) – це американська компанія, розробник програмного забезпечення аналізу даних[41]. Ключові замовники компанії – американські силовики і спецслужби: ЦРУ, ФБР, Міністерство оборони США, Військово-повітряні сили США, Корпус морської піхоти США, Командування спеціальних операцій США, Військова академія США, міські поліцейські департаменти різних міст США . Типовий проект Palantir обходиться замовникові в суму від 5 до 100 млн. доларів.

Основна спрямованість продуктів фірми – візуалізація великих масивів даних з різнорідних джерел, що дозволяє користувачам знаходити взаємозв'язки між об'єктами, виявляти збіги між об'єктами і подіями навколо них, виявляти аномальні об'єкти, інтелектуальний аналіз даних (Data Mining) з упором на інтерактивний візуальний аналіз.

Palantir – наймасштабніший продукт з капіталізацією більш ніж 500 млн доларів США, який має багатий арсенал інструментів штучного інтелекту для роботи з Big Data і Big Stream Data. Працює з гетерогенними даними. Частково використовується Europol[42].

Як джерела, програмне забезпечення Palantir використовує як відкриті дані, традиційні бази даних та інші структуровані джерела, так і тексти, аудіо, відео. Вся робота ведеться в інтуїтивному графічному інтерфейсі, а запити до джерел формулюються природною мовою.

Усі типи даних Palantir називаються «об'єктами». Їх ділять на три категорії: сутності, події та документи. Всередині кожної категорії об'єкти групуються тематично. Наприклад, в сутність «особистість» (Person) входить не тільки ім'я конкретної людини, але і її адреси електронної пошти, банківські рахунки, номери карт соціального страхування, дані прав водія: інформація про зростання, вагу, колір очей і дата народження. Будь-який об'єкт має ряд властивостей (Property Types). Вони зале-

[41] Kevin Simler. What does Palantir Technologies do? URL: https://www.quora.com/What-specifically-does-Palantir-do

[42] Струков В. М., Узлов Д. Ю., Гнусов Ю. В. Інструментальні інтелектуальні платформи для кримінального аналізу. Право і безпека. 2021. № 4. С. 64–79. DOI: https://doi.org/10.32631/pb.2021.4.07.

жать від типу об'єкта – серед можливих варіантів є підлога, якщо це людина, або район та адреса, якщо це, наприклад, нерухомість.

Інформацію про об'єкти Palantir бере одночасно з кількох систем документообігу. Приклад з керівництва показує, що необхідні дані отримані з поліцейських баз Сан-Матео і Пало-Альто. Це приклад основної якості Palantir: система синтезує великі масиви даних із різних джерел. Palantir також може знаходити зв'язок між цими даними. Це дуже полегшує роботу, яка інакше виявилася б надзвичайно трудомісткою.

Компанія Palantir Technologies надає програмну платформу Palantir Gotham (https://www.palantir.com/platforms/gotham/), яку позиціонують як «операційну систему для прийняття рішень на глобальному рівні» (Рис. 33).

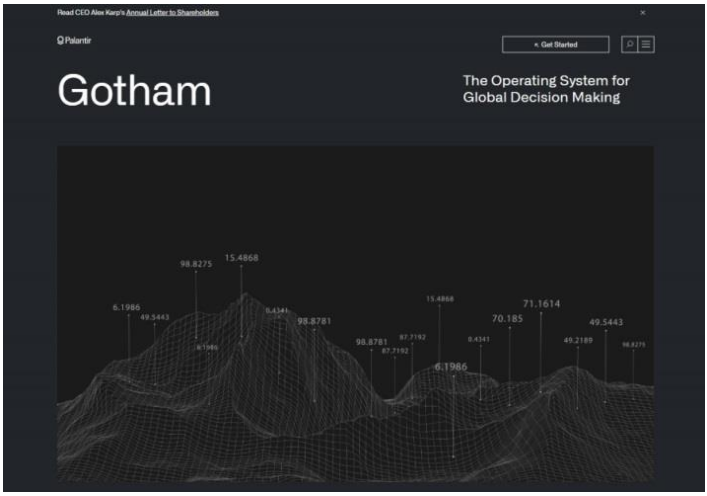


Рисунок 33 – Фрагмент стартової сторінки системи Gotham

Gotham дозволяє користувачам визначати закономірності, приховані глибоко в наборах даних, починаючи від джерел розвідки сигналів і закінчуючи звітами від конфіденційних інформаторів, а також полегшує передачу між аналітиками та оперативними користувачами, допомагаючи операторам планувати та виконувати реальні світові відповіді на загрози, які були виявлені на платформі.

Gotham – продукт, орієнтований для державних спецслужб, початкове найменування продукту – Government – відображало секторальну специфіку. Gotham використовує техніку онтологій, засобами яких різнорідні дані з багатьох джерел оснащуються смисловою інформацією і уніфікуються для спільного аналізу. Онтології в продуктах Palantir можуть бути одного з трьох типів:

- сутності – суб'єкти або об'єкти реального світу;
- події – дії над сутностями, що відбуваються в певний момент часу і в певній точці простору;
- документи – підтвердження відомостей про реальний світ, зведені в єдиний формат (на основі HTML).

Засоби семантичного пошуку в платформі використовують можливості складання складних запитів до онтологій, зокрема, підтримується пошук по близькості значення, є також засоби фонетичного пошуку. Встановлені користувачами-аналітиками факти і взаємозв'язки також зберігаються у формі з семантичною розміткою. Платформою підтримується підключення генетичних алгоритмів, спеціалізованим чином розробляються для тих чи інших сфер діяльності. Також складовою частиною платформи Gotham є підсистема групової роботи користувачів, що дозволяє аналітикам обмінюватися повідомленнями і результатами аналізу.

Для візуалізації даних Palantir використовує три інструменти: «Гістограма», «Карта» та «Оглядач об'єктів». Усі вони допомагають швидше відстежити зв'язок між різними типами даних. Гістограма допомагає відстежити збіги та повторювані дії об'єкта, що може розповісти про звички та поведінку людини. Режим картки дозволяє поліції зробити три речі: показує розташування об'єктів, шукає автоматичного визначника номерних знаків і будує «теплові карти» на основі концентрації певних об'єктів. «Оглядач об'єктів» – це комплексний інструмент аналізу. Він дозволяє поліції фільтрувати, сортувати та відображати десятки різних видів даних. Дані можна надати чотирма основними способами: числовими діаграмами, гістограмами, тимчасовими шкалами і круговими діаграмами. Інструкція Palantir пояснює, що тип відображення змінюється, залежно від того, яку інформацію аналізує поліція.

Palantir Technologies також пропонує Palantir Foundry, платформу, яка трансформує способи роботи організацій шляхом створення центральної операційної системи для їхніх даних; і дозволяє окремим користувачам інтегрувати та аналізувати потрібні дані в одному місці. Крім того, він надає Palantir Apollo, програмне забезпечення, яке забезпечує програмне забезпечення та оновлення для всього бізнесу, а також дозволяє клієнтам розгортати своє програмне забезпечення практично в будь-якому середовищі; і Palantir Artificial Intelligence Platform (AIP), яка забезпечує уніфікований доступ до відкритих, розміщених на власному хості та комерційних великих мовних моделей (LLM), які можуть перетворювати структуровані та неструктуровані дані в об'єкти, зрозумілі LLM, і перетворювати дії та процеси організацій на інструменти для людей і агентів, керованих LLM.

Засоби користувальницького завантаження даних в Palantir Technologies, включені в платформу – FEI (front-end import, засіб напівавтоматичного забору інформації з вебсайтів), Kite (засіб перетворення зовнішніх структурованих даних у внутрішній проміжний формат pXML на основі XML), Raptor (засіб напівавтоматичного обробки великих структурованих файлів), Phoenix (обробник журналів з вбудованим розпізнаванням атрибутивної інформації, наприклад, адрес, номерів телефонів). Крім того, надається API для створення додаткових інструментів для користувача забору даних для специфічних типів джерел. Окремо опрацьоване завдання тегування документів користувачами – розмітки текстів смисловою інформацією за допомогою вказівки прив'язок до нових або існуючих сутностей, позначки подій, збагачення наявних об'єктів інформацією; для пошуку відповідного об'єкта може бути використаний власний діалект OWL (NetOwl), або SAP Text Analytics. Основна техніка призначеного для користувача аналізу накопиченої інформації – візуалізація графів, в платформу включені різні засоби з їх перегрупування, укладання, в інтерфейсі роботи з графами можливо встановлювати, приховувати і видаляти зв'язки, доповнювати їх об'єктами .

З точки зору аналітика, є п'ять важливих особливостей аналітичної платформи:

- По-перше, і найважливіше, аналітик повинен контролювати ситуацію. Іншими словами, основним способом

взаємодії з інструментом аналізу мають бути запити, керовані людиною.
- Здатність узагальнювати великі масиви даних.
- Можливість візуалізації великих масивів даних.
- Здатність до швидкої ітерації (можливості аналітику поставити запитання, отримати відповідь, а потім швидко поставити наступне запитання, яке залежить від відповіді на початкове запитання).
- Здатність аналітика легко ділитися результатами свого аналізу зі своїми колегами.

Коли ці частини об'єднуються в узгоджену систему, в результаті виходить аналітична платформа, дуже загальна та дуже потужна. Зокрема, реалізація системи для поліції – Palantir Law Enforcement має інтуїтивно зрозумілий, зручний інтерфейс, який дає можливість будь-якому агенту, детективу чи слідчому швидко отримати доступ до всієї інформації в одному місці. Користувачі можуть здійснити пошук підозрюваного, цільового об'єкта або місця за допомогою єдиного порталу та отримати необхідні дані з усіх відповідних систем. Palantir підключається до національної системи обміну інформацією США (National Information Exchange Model), підтримує наявні системи управління справами, системи управління доказами, арештами, судовими даними, іншими даними про злочини й даними автоматизованої диспетчеризації (CAD), а також має підключення до федеральних сховищ США, оперативних баз і даних з державних сховищ. Уміє обробляти як структуровані та слабоструктуровані, так і неструктуровані дані, такі як сховища документів та електронні листи.

2.7.3 InfoStream

InfoStream (www.infostream.ua)[43] – сервіс контент-моніторингу веб-ресурсів, що надає доступ у пошуковому режимі до інформації з 10000 джерел, класифікацію інформації, сентимент-аналіз, екстрагування понять (персон, компаній, топонімів), формування сюжетних ланцюжків, оцінку тональності повідомлень, аналіз динаміки публікацій щодо певних

[43] http://infostream.ua/UKR/

об'єктів[44]. У базах даних системи зберігається близько 1 млрд. документів новин майже за 30 років.

InfoStream базується на скануванні інтернет-джерел, в першу чергу новинних, і наданні інструментів для аналізу отриманої інформації (Рис. 33).

InfoStream істотно відрізняється від глобальних світових пошукових систем. Не маючи їх охоплення, він сканує найважливіші, з точки зору цільової аудиторії – інформаційних аналітиків і маркетологів, ресурси. При цьому результати пошуку в InfoStream не спотворюються таргетуванням, ремаркетінгом, (само)цензурою та іншим чином, що властиво глобальним пошуковим системам.

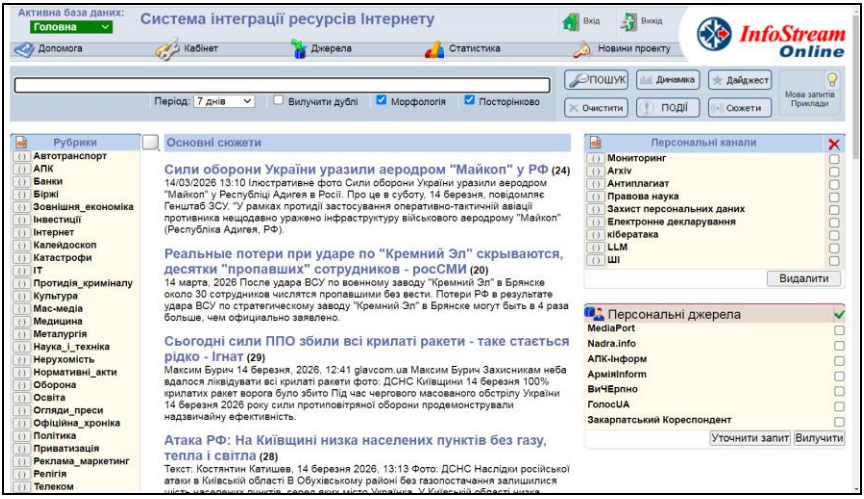


Рисунок 33 – Стартова веб-сторінка сервісу InfoStream Online

Головні переваги системи InfoStream у порівнянні з традиційними мережевими інформаційно-пошуковими системами:

- Оперативність – бази даних системи поповнюються

[44] Григор'єв О.М., Ланде Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.М. InfoStream. Моніторинг новин з Інтернету: технологія, система, сервіс: науково-методичний посібник. – Київ: ТОВ "Старт-98", 2007. – 40 с.

кожні 15 хвилин, джерела скануються в Мережі в міру їхнього оновлення, тоді як період індексації традиційних інформаційно-пошукових систем може вимірюватися тижнями.

- Доступність ретроспективного фонду – навіть якщо інформація видалена з веб-сайту джерела, вона збережена в інформаційному сховищі.
- Наявність аналітичного інструментарію – користувач може в режимі реального часу не тільки отримувати результати пошуку, а й формувати дайджести, будувати сюжетні ланцюжки, аналізувати зв'язок рубрик, динаміку понять і т.д.
- Можливість селекції дублікатів – система здійснює автоматичне маркування ідентичного за змістом інформації новин.
- Наявність інструментарію багаторівневого уточнення запиту.

У порівнянні зі звичайними новинними веб-сайтами, система InfoStream забезпечує такі переваги:

- Охоплення джерел – користувач має доступ до новин за тематикою, що його цікавить, одночасно з великої кількості веб-сайтів, включаючи і ті обрані, які він звик переглядати щодня.
- Принцип "одних рук" – користувач системи моніторингу має доступом до інформації з багатьох веб-сайтів з одного інтуїтивного інтерфейсу.
- Пошукові можливості – новинні веб-сайти, на відміну від системи InfoStream, не завжди мають розвинені пошукові можливості.
- Доступність ретроспективного фонду.

У порівнянні з відомими службами інтеграції новин сервіс InfoStream має такі переваги:

- Список джерел, що безперервно розвивається – системою InfoStream сканується понад 10000 джерел – усі основні інформаційні сайти України, а також провідні зарубіжні інтернет-ресурси.
- Облік місцевих ресурсів – навіть найбільші платні закордонні інтегратори новин лише частково охоплюють, наприклад, українські веб-ресурси. У систему InfoStream

можливе включення нових джерел за заявками користувачів.

- Розширений доступ до інформації – користувачам системи InfoStream доступні не тільки заголовки або анонси, але й повні тексти повідомлень новин, посилання на подібні документи та ін.
- Аналітичний інструментарій.
- Розвинена служба підтримки абонентів – користувачі системи за потреби можуть звертатися з питаннями до служби підтримки та отримувати вичерпні відповіді та консультації.

InfoStream надає широкий набір інструментів для професіоналів в області аналізу інформації та інформаційної боротьби:

- пошук першоджерел новин;
- обробка інформаційних потоків з метою відокремити природний життєвий цикл інформаційного події від штучного втручання;
- аналіз результатів інформаційних операцій;
- побудова онтологічних схем;
- автоматична побудова інформаційних сюжетів для тематичного аналізу;
- визначення і формування переліку цитат – прямої мови (Рис. 34) і багато іншого.

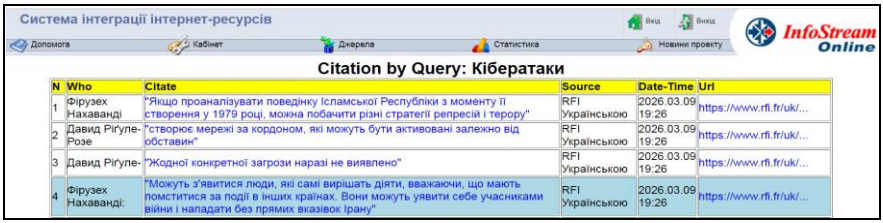


Рисунок 34 – виведення переліку цитат за запитом

Інформаційний сервіс та інформаційні продукти, засновані на базі системи автоматичного моніторингу інтернет-ЗМІ, дозволяють набагато спростити вирішення основних завдань інформаційного моніторингу ЗМІ:

- визначення та уточнення можливих основних інфор-

маційних приводів;

- максимально повне відстеження публікацій в електронних ЗМІ, що належать до цікавої події;
- визначення ступеня лояльності тих чи інших ЗМІ до компанії, бренду або персони;
- побудова медіа-репутації компанії, бренду або персони;
- оперативний і різноплановий моніторинг інформації про тенденції розвитку галузі або ринку;
- відстеження діяльності конкурентів, партнерів і регулюючих органів для оперативного реагування на події і своєчасного коригування власної стратегії;
- пошук потенційних клієнтів і партнерів;
- ретроспективний доступ до інформації про певну компанію, бренд, персону, подію або ринки, за допомогою архіву, що формувався понад 20 років;
- аналіз ефективності проведених заходів, акцій чи проектів для власних цілей та/або в інтересах замовника.

Як приклад застосування можливостей аналізу соціальних мереж наведемо фрагмент дослідження мережі зв'язків понять, які екстрагуються з корпусів неструктурованих текстів – масивів документів, що скануються з мережі Інтернет системою контент-моніторингу InfoStream.

При побудові мережі понять використовувалися алгоритми автоматичного вилучення понять із неструктурованих текстів. Слід зазначити, що підходи до вилучення різних типів понять із текстів суттєво відрізняються як у контексті їх уявлення, і за структурними ознаками. Так, для виявлення належності документа до рубрики тематичного класифікатора можуть використовуватися спеціальним чином складені запити, що включають логічні та контекстні оператори, дужки тощо.

Виявлення географічних назв передбачає використання таблиць, у яких крім шаблонів написання цих назв використовуються коди країн, назви регіонів та населених пунктів.

Ще один вид понять, такий як «персони», екстрагується з текстів на підставі правил, що враховують таблиці допустимих імен та прізвищ, шаблони ініціалів, можливі варіанти спільного написання ініціалів/імен та прізвищ.

Слід зазначити, що система InfoStream включає засоби отримання понять і, серед іншого, надає користувачам результа-

ти у вигляді «інформаційних портретів», які включають такі поняття, як ключові слова, географічні назви, прізвища персон, назви фірм і т.д. В рамках цієї системи аналізуються властивості мереж, утворених поняттями, пов'язаних один з одним згадками у тих самих документах.

Мережа, утворена поняттями, що витягуються з потоків текстів, не статична, а залежить від обсягів документів, з яких витягуються відповідні поняття. Отже, розуміння структури такий мережі необхідно враховувати її еволюцію.

Отримані емпіричні результати можуть бути корисними, наприклад, при моделюванні економіко-соціальних процесів, виявленні та візуалізації неявних зв'язків окремих об'єктів чи суб'єктів. Феномен стабілізації мережі зв'язків практично дозволяє шляхом аналізу невеликого масиву документів виявляти стійкі зв'язки, знижувати вплив шумових чинників. Водночас поки залишається відкритим питання оцінки кореляції отриманих інформаційних взаємозв'язків персон, розрахованих шляхом підрахунку частоти документів, у яких персони згадуються спільно, та взаємозв'язків реальних.

2.7.4 Attack Index

Attack Index (attackindex.com) – система, що дозволяє дати відповіді на запитання: чи ведеться проти користувача інформаційна атака, чи відбувся природний сплеск інтересу до події; коли почалася інформаційна операція, наскільки вона інтенсивна та масштабна; які сайти та облікові записи в соціальних мережах використовуються для атаки; хто став ініціатором інформаційної операції та як пов'язані її учасники.

Система Attack Index інтегрує окремі спеціальні модулі, що реалізують окремі інструментальні засоби дослідження інформаційних потоків, а саме візуального відображення динаміки інформаційних потоків, результатів згладжування відповідних часових рядів, графік розподілу повідомлень за тональністю, діаграми відхилення часового ряду від локальних лінійних трендів, діаграми кореляції часового ряду із шаблонами інформаційних операцій, перелік і мережу інформаційних джерел, динаміку показника стабільності часового ряду, результати прогнозування поведінки часових рядів тощо. В системі Attack Index реалізована методика розпізнавання інформаційних операцій і візуалізації аномалій в часових рядах динаміки тема-

тичних публікацій, зокрема, із Інтернет-простору. Система дозволяє виявляти основні тренди динаміки тематичних інформаційних потоків, виявляти аномалії, інформаційні операції, створює основу методології їх застосування до окремих предметних областей.

Систему Attack Index створила команда, яка складається з представників технічних та комп'ютерних наук, прикладної математики, медіа, соціології, IT та інформаційної безпеки. Саме тому усі сучасні наукові досягнення, у першу чергу власні, швидко імплементуються до Attack Index. Наукові дослідження команди, зазвичай перетворюються у робочі моделі з поетапним впровадженням до автоматизованого звіту.

Система Attack Index базується на низці математико-статистичних методах обробки інформації та інтерпретації інформаційних потоків, які були обґрунтовані в наукових роботах і опробовані на практиці. Attack Index використовує технології великих даних (Big Data), комплекс аналітичних алгоритмів, а також інструменти збору та візуалізації даних про атаку.

Перш за все, методи Attack Index базуються на тому, що сучасний інформаційний простір надає можливість отримання практично будь-якої інформації по обраному питанню при наявності відповідного інструментарію. Його використання дозволяє аналізувати взаємозв'язок минулих і поточних подій з інформаційної активністю обраного кола джерел. Такий підхід обумовлений, перш за все, статистикою.

Як інформаційний продукт Attack Index (індекс атак) – це інтегральний показник рівня інформаційної небезпеки, що враховує безліч факторів. У них входять: наявність інформаційної активності, активності можливих конкурентів, відхилення середнього фону, наявність інформаційних операцій і стадій їх розвитку, ретроспектива і динаміка негативної тональності публікацій, а також ступінь хаотичності процесів. Крім того, в розробці знаходиться інструмент прогнозування інформаційних подій.

Attack Index відповідає на такі запити користувачів:

- чи ведеться проти вас інформаційна атака або стався сплеск інтересу до події;
- чи є спікери та хто вони, які емоції про вас вони транслюють суспільству;

- як усунути та нівелювати наслідки атаки, типові сценарії протидії загрозам;
- хто став ініціатором інформаційної операції і як пов'язані її учасники;
- коли почалася інформоперція, наскільки вона інтенсивна та масштабна;
- які веб-сайти та акаунти в соціальних мережах використовуються для атаки.

Складові рішення Attack Index:

- пошук повідомлень за темами, що представляють інтерес в глобальних мережах;
- відстеження інформаційних потоків (історій), відповідних тем, подій і процесів;
- визначення динаміки інформаційних потоків;
- побудова динаміки тональності публікацій;
- визначення аномального і критичного в заданий момент у динаміці тематичних інформаційних потоків;
- визначення основних подій і об'єктів тематичного потоку інформації;
- візуалізація відносин об'єктів моніторингу;
- прогнозування розвитку ситуації.

Комплекс комп'ютерних програм реалізується як веб-застосунок. Для виходу на головний інтерфейс комплексу (Рис. 35) користувачу необхідно ввести адресу веб-сайту, логін і пароль.

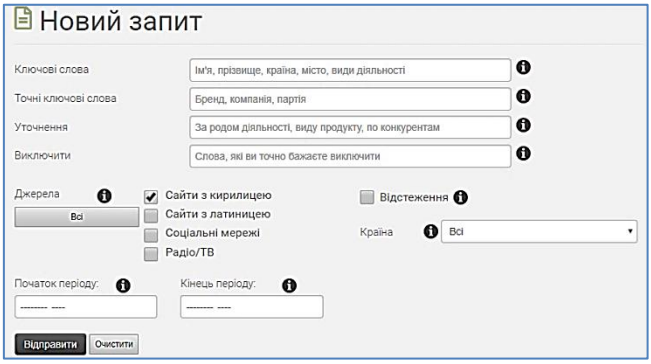


Рисунок 35 – Стартова сторінка комп'ютерного комплексу програми

У наведеному веб-застосунку з веб-форми запитуються такі дані: запит у вигляді ключових слів з параметрами включення або виключення, дати початку і кінця періоду, база даних джерел. У запиті можна вибирати мову публікацій, регіональну приналежність і тип бази даних з інформацією (бази даних моніторингу сайтів; блогів, форумів, соціальних мереж; теле/радіокомпаній).

В Attack Index реалізована система визначення тональності, яка базується на статистичному підході і навчанні нейронної мережі. Тональність визначається за допомогою машинного навчання. Система визначення тональності базується на статистичному підході і самонавчанні нейронної мережі. В основі статистики лежить виявлення найбільш часто вживаних слів в текстах з позитивною або нейтральною тональністю.

Тональність інформаційного масиву, що відповідає запиту, представляється на графіках, де зелений колір означає позитив, а червоний – негатив (Рис. 35).

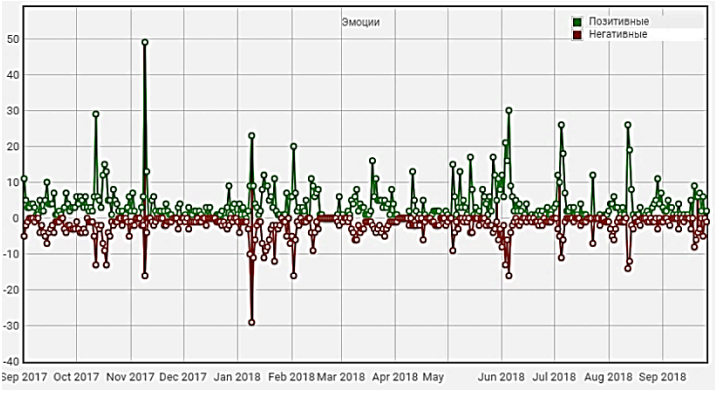


Рисунок 36 – Відображення тональності інформаційних потоків

Фактично, всі інші публікації визначаються як нейтральні. Чим більше симетрії спостерігається між цими двома кривими, тим більше можна вважати відповідний моменту матеріал в цілому нейтральним. Однак асиметрія в ту або іншу сторону, свідчить про переважання емоційного забарвлення відповідного типу.

Інформаційний простір завжди більш активно реагує на проблеми і негативні події. Як наслідок, в інформаційних пото-

ках, статистично, негатив зустрічається частіше. Навіть експерти не можуть дійти згоди, що може бути негативом, а що – позитивом, тому завдання системи правильно обробити знайдені текстові масиви і представити на розгляд оціночні значення.

Відмінна особливість системи Attack Index – визначення інформаційних операцій, виявлення подібності інформаційної динаміки шаблонам інформаційних операцій. Чим вище червона область хвильового графіка – тим більше ситуація навколо об'єкта дослідження у певний час відповідає інформаційній операції (Рис. 37).

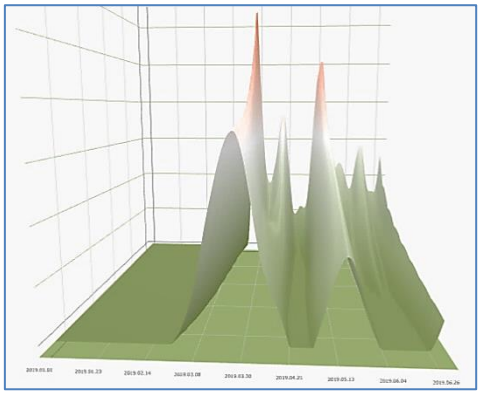


Рисунок 37 – об’ємне зображення близькості інформаційної динаміки ознакам інформаційних операцій

Контурна карта інформаційних операцій (Рис. 38) визначає дати початку і загасання виявлених хвиль, обумовлених інформаційними впливами. Чим більш насичений червоний колір – тим більше ознак інформаційної операції. Система аналізує наявність інформаційної операції тривалістю від 7 до 30 днів.

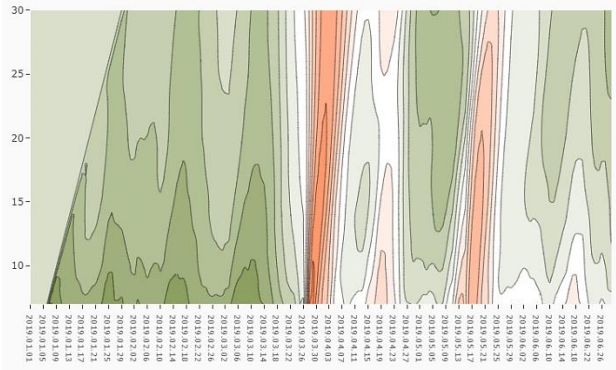


Рисунок 38 – двомірне зображення хвиль інформаційної активності з ознаками інформаційних операцій

При переміщенні курсору вертикально за датою можна побачити максимальне значення ймовірності збігу інформаційної динаміки з класичним шаблоном інформаційної операції. Можна відстежувати максимальне значення параметра інформаційної операції (приймає значення від 0 до 100). Це значення відповідає кількості днів К, протягом яких відбувалися основні фази операції. Для виявлених областей і дат інформаційних операцій представляються: топ-5 інформаційних сюжетів відповідних операцій та топ-10 ресурсів – розповсюджувачів інформації.

Граф взаємозв'язку джерел інформації надає топ-20 джерел за виявленими днями найбільших інформаційних сплесків або проведення інформаційної операції. Рейтинг будується за кількістю публікацій на одному джерелу. Когнітивна карта зв'язків (Рис. 39) показує взаємозалежність (наприклад, передрук чи наявність посилань) 20 найбільш значущих об'єктів інформаційного сплеску або операції в розраховану дату. Розмір кола джерела залежить від кількості публікацій. При натисканні миші на назві джерела синім кольором показується зв'язок для вибраного джерела з іншими. Суцільними лініями показуються найсильніші зв'язки між джерелами.

Стабільність інформаційної ситуації в системі серед іншого визначається значення коефіцієнта Херста, близькість якого до значення 0,5 говорять про хаотичну природу процесу. Чим ближче це значення до 1 – тим стабільніший процес, майбутня

поведінка повторює минулу, процес дотримується свого тренду. Різкі стрибки коефіцієнта Херста свідчать про наявність дуже різних неоднорідних процесів. Можливо, слід збільшувати досліджувані часові проміжки – для розуміння складових інформаційної активності досліджуваного об'єкта.

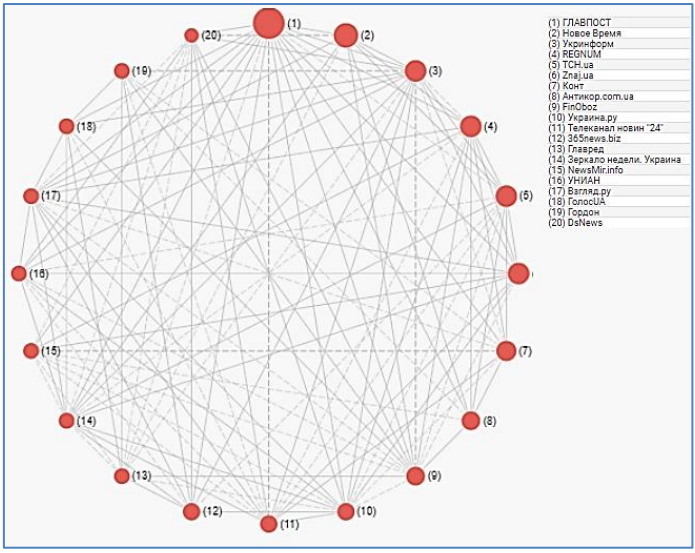


Рисунок 39 – графічне зображення кількості публікацій по джерелах, що брали участь в інформаційній операції

За допомогою методів виділення даних з текстів можна сформувати мережі взаємозв'язків понять. Їх вузли є ключовими словами, іменами персоналій, компаніями і т.д. Аналіз цих мереж дозволяє виявити явні і неявні зв'язки між окремими поняттями, оцінити вагу тих чи інших понять, уточнити критерії формування інформаційного потоку і побачити взаємозалежності в досліджуваних мережах.

Когнітивні карти можуть використовуватися для створення сценаріїв інформаційної підтримки. Вершини когнітивної карти відповідають поняттям і причинно-наслідкових зв'язків. При аналізі когнітивних карт вузли та посилання оцінюються щодо обраної концепції, після чого між цими вузлами утворюються узгоджені ланцюга.

Вузли можуть бути пов'язані між собою, якщо відповідні їм слова знаходяться поруч у тексті, належать одному реченню, з'єднані синтаксично або семантично.

Як підкреслюється на веб-сайті сервісу Attack Index, за 2022–2023 роки кількість джерел моніторингу збільшилася на 45%, у тому числі за рахунок кількості Телеграм каналів – на 50%, джерел в Україні та росії – до 20%. А також додались показники аналізу інформаційної ситуації та до восьми розділів, які були доступні раніше, додались іще чотири, які дозволили ширше характеризувати інформаційну активність пов'язану із запитом.

Сервіс дозволяє визначати чи ведеться інформаційна атака чи стався сплеск інтересу до події, які публічні персони-спікери залучені, які емоції стосовно запитаної теми, персон, компаній чи подій вони транслюють суспільству, хто став ініціатором інформаційної операції і як пов'язані її учасники. У випадку виявленої інформаційної операції, можна дізнатися наскільки вона інтенсивна та масштабна, які сайти чи акаунти в соціальних мережах задіяні. Новою характеристикою стало математичне прогнозування[45] подальшого розвитку інформаційної ситуації. Методика прогнозування подій на основі кореляційного відображення і Фурье-розкладу релевантних часових рядів призначена для прогностичного обчислення ознак того, що відбудеться деяка подія. Як вхідні данні для обчислення крім ряду, який необхідно продовжити прогнозними значеннями, застосовуються релевантні ряди близьких за змістом подій. Обчислення базується на моделі, в якій розраховуються кореляції сумарних значень розкладу Фур'є для всіх релевантних часових рядів із історичним рядом. При цьому певна кількість постійних параметрів розкладу Фур'є обчислюється як результат рішення оптимізаційної задачі. За визначеною моделлю і значеннями постійних параметрів проводиться розрахунок подальших значень історичного ряду, тобто визначення прогнозованої події.

[45] Lande, Dmitry and Shchutsky, V. and Shnurko-Tabakova, Ellina and Strashnoy, Leonard. Event Prediction Technique Based on Correlation Mapping and Fourier Decomposition. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4372251 (February 27, 2023). – 8 p.

Протягом свого існування від 2018 року аналітично-моніторинговий сервіс Attack Index (attackindex.com) розвивався як у напрямку розширення баз моніторингу за рахунок під'єднання нових веб-сайтів чи інших каналів розповсюдження інформації (наприклад, Телеграм), так і додаванням нових аналітичних функцій, у тому числі – математичних методів прогнозування розвитку інформаційної ситуації. І сьогоднішні нововведення теж стали результатами продовження наукової роботи, досвіду підготовки аналітичних звітів для клієнтів та регулярного зворотного зв'язку і консультацій з користувачами системи.

Лінгвостатистичний підхід проведення аналітичних досліджень на базі застосування соціальних мереж має об'єктивний характер, що є суттєвим компонентом методологічної основи аналізу та прогнозування.

Використання запропонованих засобів візуалізації дозволяє «розкласти» вихідні часові ряди за складом і особливостями фразеології і понять, виявити активність публікацій, яким відповідають певні наративи, виявити зв'язки фразеологізмів, особливості динаміки виникнення в інформаційному потоці нових фразеологізмів.

Розглянутий підхід може бути використаний для аналізу та візуалізації розподілу концептів, наративів для будь-яких вибраних наборів інформації з точки зору питань, що цікавлять дослідника та охоплюють значний часовий проміжок.

2.7.5 Cyber Aggregator

Актуальним підходимо вирішення проблеми створення такої корпоративної системи, що пропонується, 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 певний період часу пов'язаних із тематикою «кібербезпека».

Збирання інформації здійснюється у режимі пошуку в соціальних медіа (веб-сайтах, соціальних мережах, месенджерах, блогах, тощо). Запит (ключова фраза для пошуку у відповідній соціальній мережі, якщо це можливо, інакше – акаунт) зчитується програмою із спеціальних конфігураційних

таблиць. Далі йде здійснення пошуку і виведення записів, що відповідають запитам. Після цього здійснюється запис унікальних записів до БД сервера.

Аналіз існуючих підходів до агрегації тематичних новин привів до необхідності і можливості створення комплексу інструментальних засобів контент-моніторингу соціальних мереж з вибраних питань, зокрема, кібербезпеки[46].

Макет, що описується нижче, включає сучасні засоби персоналізації, надання доступу до баз даних в режимі онлайн, у тому числі з мобільних пристроїв, для чого широко застосовуються можливості форматів 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.

Для прикладу, розглянемо детально методологію добування інформації із вибраних каналів месенджера Telegram. Відомо три шляхи доступу засобів автоматизованого добування даних до цього ресурсу. Перший – доступ напряму до месенджера за адресою типу https://uk.tgstat.com/channel/@kpilive. У цьому випадку kpilive – це назва каналу (конкретно у цьому випадку – каналу НТУУ «КПІ ім. Ігоря Сікорського» КПІ live). Другий, в іншому форматі, за адресою редиректу типу: https://t.me/s/kpilive. Третій – доступ до інформації каналу у форматі Atom через зовнішній агрегатор RSSHub: https://rsshub.app/telegram/channel/kpilive.

1. Для створення текстового корпусу на основі контенту каналів месенджера Telegram першим кроком створюється список каналів, що можуть цікавити дослідника. Для цього можна звернутися до чисельних каталогів цих каналів, розміщених в мережі. Обираємо, наприклад, розміщений за адресою https://ru.telegram-store.com/catalog/product-category/channels/ (понад 50 тис. каналів). Вводимо в режимі пошуку слово, що нас цікавить, наприклад «Україна», отримуємо дві сторінки посилань за адресою: https://ru.telegram-

[46] Lande, D. Information Streams Analysis in the Global Computer Networks // Visnyk NAS of Ukraine, 2017. – N 3. – p. 46-54.

store.com/?s=Україна. Формуємо список каналів у форматі:

```
@nowasteukraine
@onlineukraine
@sos_ua
@ua_rozvynuta
@ukraine_novosti
…
```

2. Скануємо ресурси вибраних каналів із допомогою вільно доступної програми отримання інформації з мережевих ресурсів. Відомо, що існує декілька програмних агентів (програм-роботів), що сканують контент із ресурсів мережі за протоколами HTTP/HTTPS. Обираємо програму wget, яка входить до складу багатьох систем. Застосовуємо вибраний у п. 1 список каналів, формуємо переліки адрес для трьох шляхів доступу до ресурсу:

```
а:
https://uk.tgstat.com/channel/@nowasteukraine
https://uk.tgstat.com/channel/@onlineukraine
https://uk.tgstat.com/channel/@sos_ua
…
б:
https://t.me/s/nowasteukraine
https://t.me/s/onlineukraine
https://t.me/s/sos_ua
…
в:
https://rsshub.app/telegram/channel/nowasteukraine
https://rsshub.app/telegram/channel/onlineukraine
https://rsshub.app/telegram/channel/sos_ua
…
```

Здавалося б, найбільшу перевагу дає підхід в), його формат Atom найбільш стандартний, але, враховуючи те, що він відповідає зовнішньому сервісу по відношенню до Telegram, до реалізації рекомендується підхід б.

3. На третьому, останньому етапі здійснюється перетворення зібраних найповніших даних до формату текстового корпусу, необхідного для дослідження (до формату XML 1.0), який у подальшому може завантажуватися в інформаційно-пошукову систему Manticore, фрагмент якого такий:

```
<?xml version="1.0" encoding="utf-8"?>
<sphinx:docset>
<sphinx:schema>
<sphinx:field name="subject"/>
<sphinx:field name="content"/>
```

```
<sphinx:field name="source"/>
<sphinx:field name="datatime"/>
<sphinx:attr name="url"/>
</sphinx:schema>
 <sphinx:document id="1_tg">
<subject>Фестивалі, концерти та літературні читання</subject>
<content>Фестивалі, концерти та літературні читання - ці події зовсім не поставлені на паузу через карантин...</content>
<source>Telegram: novoe_vremya</source>
<datatime>20200425 14:20</datatime>
<url>https://tlgrm.ru/channels/@novoe_vremya/11263</url>
</sphinx:document>
…
```

Наведена методика може застосовуватися для добування даних в інформаційно-пошукових системах, аналізу текстів, що публікуються користувачами месенджера Telegram. Подібна методика із деякими адаптаціями використана і для формування текстових масивів із інших соціальних медіа.

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

1) Формування баз даних шляхом підключення до мережі Інтернет та збору за певними критеріями і 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. При необхідності створення службових акаунтів, через які буде організований доступ ПТК 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 (UTF-8).

4) Виявлення дублікатів, схожих за змістом інформаційних повідомлень (у т.ч. на різних мовах), групування дублікатів та

близьких за змістом інформаційних повідомлень у видачі пошукової системи.

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

6) Первинний аналіз текстових повідомлень, що зберігається в базах даних системи: автоматичне виявлення іменованих сутностей (особи, назви компаній, брендів, географічні назви тощо), визначення тональності, виявлення опорних слів за статистичними алгоритмами в інформаційних матеріалах, наведених різними мовами.

7) Формування аналітичних звітів, у тому числі інформаційних портретів і сюжетних ланцюжків, що ґрунтуються на використанні опорних слів, наведених різними мовами, тематична рубрикація документів.

8) Інтеграція із геоінформаційною системою.

9) Аналіз та візуалізація даних; візуалізація статистичних даних: за визначеними джерелам, кількість 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 в соціальних мережах. При цьому можуть виявитися моменти виникнення циклів, а також моменти, коли за періодами регулярної динаміки настають хаотичні коливання. Визначено, що динаміку інформаційних операцій найточніше відображують такі відомі вейвлети, як «мексиканський капелюх» та вейвлет Морле[47].

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,

[47] Information Operations Recognition. From Nonlinear Analysis to Decision-Making / A. Dodonov, D. Lande, V. Tsyganok, O. Andriichuk, S. Kadenko, A. Graivoronskaya. – LAP Lambert Academic Publishing, 2019. – 292 p.

методи нелінійного аналізу[48]. Проте часові ряди, породжувані тематичними інформаційними потоками, також мають фрактальні властивості і 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[49].

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 таки 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: – витяг понять; – геоінформаційна підтримка; – визначення тональності повідомлень; – формування зведень; – аналіз

[48] Dodonov O., Lande D., Nesterenko O., Berezin B. Approach to forecasting the effectiveness of public administration using OSINT technologies // Information technology and security. Proceedings of the XIX International Scientific and Practical Conference ITS-2019. – Kyiv: Engineering, 2019. – pp. 230-233.

[49] Sornette, D. How to predict the collapse of financial markets. Critical events in complex financial systems. – Litres, 2017. – 394 p.

динаміки повідомлень; – прогнозування; аналіз масиву джерел інформації тощо);

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

Реалізований макет системи аналізу великих обсягів даних із соціальних медіа Cyber Aggregator [50,51] надає користувачеві веб-інтерфейс, з якого йому доступні функції пошуку і аналізу інформації в соціальних медіа.

Користувачу системи надаються можливості пошуку (як у ретроспективній базі даних повідомлень із соціальних медіа (Search), так і у поточній інформації (Current), а також і аналізу даних (Analysis). Центральне місце інтерфейсу займає дайджест із найбільш релевантних потребам користувача повідомлень. У 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 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.

Якщо запит видає відповідні інформаційним потребам документи, то його можна зберегти для подальшого застосування (Add Query). Можливий подальше виведення знайдених повідомлень у форматі RSS (із подальшим завантаженням цих результатів в так званні RSS-агрегатори на постійній основі), а також виведення результатів пошуку із деталізацією на географічній карті, що масштабується як в автоматичному режимі, так і шляхом налаштувань (Рис. 40).

В аналітичному режимі (Analysis) користувачеві надається низка інструментів, перший з котрих – це графік (Graph), що

---

[50] Computer program "Computer program of social networks content monitoring on cybersecurity "CyberAggregator" ("CyberAggregator"). Lande D., Subach I., Sobolyev A. Ukraine. Certificate of registration of copyright to the work N 91831 from 31.07.2019.

[51] D. Lande; I. Subach; A. 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

відповідає часовому ряду кількості релевантних запиту повідомлень на добу.

Користувачеві також надається можливість перегляду головних сюжетів (Digest) за темою, кластерів, згрупованих за відповідністю заздалегідь визначених опорних слів.

В системі передбачені режими формування мереж із понять, що відповідають окремим повідомленням (персон, брендів), інформаційних джерел (Рис. 41). Ці режими дозволяють рейтингувати поняття, досліджувати взаємозв'язки між ними.

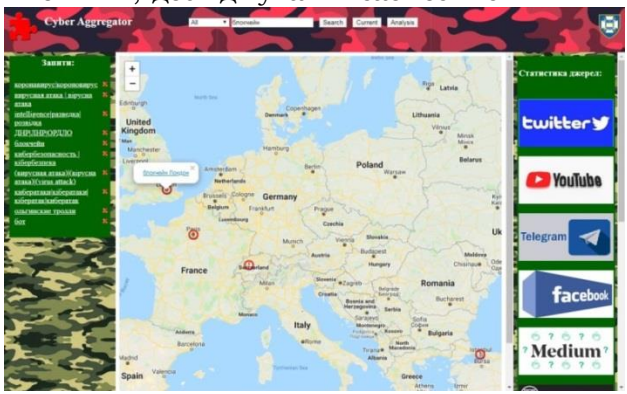


Рисунок 40 – Фрагмент інтерфейсу із застосуванням геоінформаційної системи

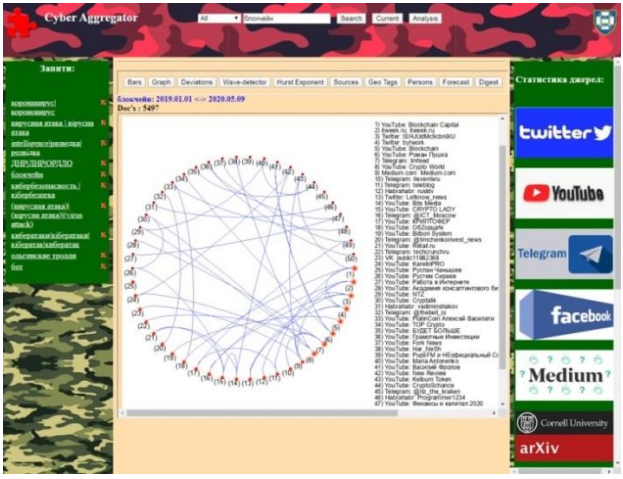


Рисунок 41 – Мережа взаємозв'язку джерел інформації, на яких публікувалися повідомлення за запитом «блокчейн»

У режимі «Аналітика» передбачено можливість прогнозування (Forecast) методом, запропонованим Д. Сорнетте[52], який заснований на аналізі закономірності руху ринкових цін на товарних і фондових ринках перед крахом. В роботі відзначається, 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:

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

У цій моделі t_c – критичний час (час кризи). Коефіцієнти моделі A, B, ω, φ визначаються за допомогою процедури підбору. За допомогою використання моделі Сорнетте (клавіша Forecast, Рис. 42) на основі даних моніторингу можна отримати прогнозні значення логарифму кількості відповідних публікацій.

Розглянемо кейси застосування системи Cyber Aggregator як інструмента забезпечення інформаційної та кібернетичної безпеки. Методика розслідуванні кіберінцидентів, що описується у першому кейсі, базується на аналізі змістовної складової інтернет-простору і дозволяє визначати об'єкти кібербезпеки та їх зв'язки, завдяки чому допомагає вирішувати різноманітні завдання, включаючи цільовий збір та обробку інформації, виділення потрібних сутностей, встановлення зв'язків між ними та створення мережі об'єктів. Також за допомогою цієї методики можна проводити кластерний аналіз мережі об'єктів, визначати центри кластерів та виконувати інші подібні завдання.

[52] Urentsov, O. Testing the possibility of predicting crises in the financial market using the method of D. Sornette. Proceedings of the Institute of System Analysis of the Russian Academy of Sciences, 2008. – N 40. – pp. 174-191.

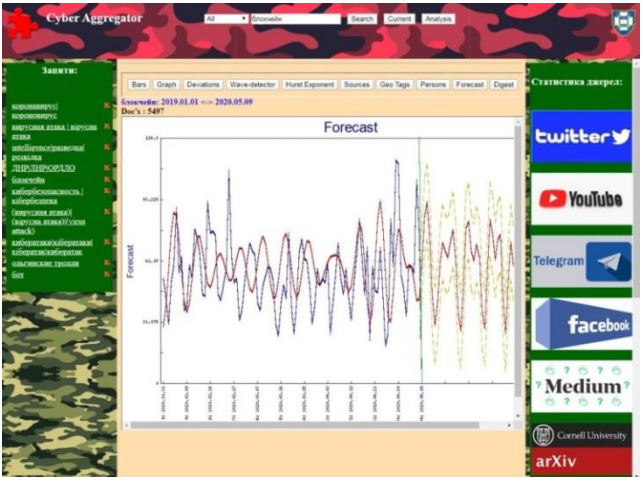


Рисунок 42 – Прогнозна лінія за алгоритмом Сорнетте для часового ряду, що відповідає певному запиту

Ця методика розслідування кібернетичних інцидентів, що вже відбулися передбачає виконання наступних кроків[53]:

Крок 1. Встановлення часового проміжку, під час якого відбувся кібернетичний інцидент.

Крок 2. Збір найбільш можливого обсягу посилань на повідомлення з Інтернету, що стосується вказаного кіберінциденту.

Крок 3. Отримання повних текстів всіх повідомлень, що відповідають зібраним посиланням.

Крок 4. Групування отриманих повідомлень за часом їхньої публікації в мережі Інтернет та побудувати часову послідовність динаміки публікацій.

Крок 5. Дослідження часового ряду, виявлення аномалій, періодичності тощо, порівняння з відомими шаблонами кібернетичних інцидентів.

[53] D. Lande, O. Puchkov, I. Subach, M. Boliukh, D. Nahornyi OSINT investigation to detect and prevent cyber attacks and cyber security incidents // Information Technology and Security, 2021. Vol 9 (2). – pp. 209-218. DOI: doi.org/10.20535/2411-1031.2021.9.2.249921.

Крок 6. Виділення з текстів концепти – іменовані сутності. Відображення на географічній картi топонімів, що відповідають вибраному кібернетичному інциденту.

Крок 7. Формування мережі цих сутностей та знаходження групи найбільш зв'язаних.

Крок 8. Формування дайджесту з повідомлень, що відображають найважливіші аспекти визначеного кіберінциденту.

Як приклад розслідування за цією методикою можна навести розслідування кібератаки Colonial Pipeline (7 травня 2021 року), яка за оцінками преси була найуспішнішою кібератакою на нафтову інфраструктуру США. У відповідності із представленою методикою було зібрано декілька тисяч документів за визначений проміжок часу, побудовано динаміку їх публікацій (Рис. 43), для виявлення аномалій і подібності фрагментів досліджуваного часового ряду у різних масштабах використовувався вейвлет-аналіз.

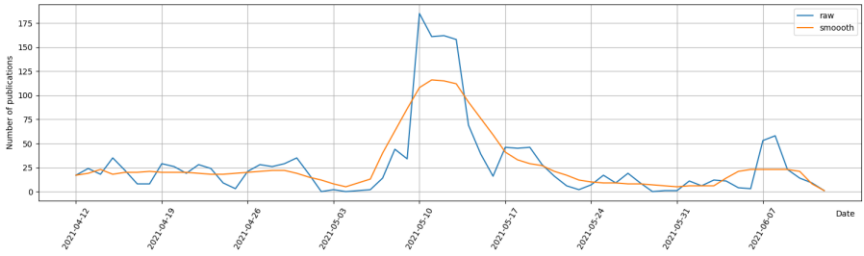


Рисунок 43 – Вихідний та згладжений ряди

Після цього були сформовані мережі сутностей, для відображення яких був використаний спеціальний програмний модуль, результат роботи якого – набір даних у форматі CSV, що відповідає матриці суміжності. Відображення і кластеризація здійснювалось за допомогою системи аналізу графів Gephi. У програмі Gephi відкривається створений CSV-файл. На основі зібраних даних було сформовано дайджест – список релевантних документів, що відображають різні найбільш важливі різні аспекти визначеного кіберінциденту.

Після цього були сформовані мережі сутностей, Відображення і кластеризація якої здійснювалось за допомогою системи аналізу графів Gephi. На основі зібраних даних було сформо-

вано дайджест – список документів, що відображають найбільш важливі різні аспекти кіберінциденту.

Виявлення злочинних кібернетичних угрупувань

Другий кейс присвячено проблемі екстрагування понять із текстів документів із мережевих джерел, а саме злочинних російських і білоруських хакерських угрупувань, що є учасниками сучасної кібервійни, а також відповідного шкідливого програмного забезпечення. Для вирішення цієї проблеми пропонується методика, сутність якої полягає у виконанні таких технологічних операцій, як:

1) добування інформації;
2) екстрагування понять – об'єктів кібербезпеки;
3) фільтрація понять із залученням експертів або засобів штучного інтелекту;
4) формування мережі об'єктів кібербезпеки;
5) аналіз (у тому числі кластеризація) і візуалізація цієї мережі;
6) візуалізація динаміки появи понять у часі.

Проводився аналіз активності російських/білоруських хакерських угрупувань впродовж 2022 і початку 2023 року. Для цього на 1-му етапі формується тематичний інформаційний масив, для чого мають використовуватись наявні інформаційно-пошукові системи, як загальнодоступні, так і корпоративні системи контент-моніторингу, зокрема, система Cyber Aggregator, яка дозволяє збирати інформацію із веб-сайтів і 12 соціальних мереж. Для отримання інформаційного масиву публікацій щодо кібербезпеки необхідно визначити необхідний період опрацювати тематичний запит, наприклад такий:

хакер|(вредоносн-программн)|(шкідл-програмн)|(кибер-атак)|кібератак|(кібер-атак)|кібератак

Після цього застосовувалась методологія визначення іменованих сутностей в області кібербезпеки[54], а на останньому

[54] Д.В. Ланде, О.О. Пучков, І.Ю. Субач. Методика виявлення об'єктів кібербезпеки на базі технології OSINT // Інформаційні технології і безпека. Матеріали XXII Міжнародної науково-практичної конференції ІТБ-2022. – Київ: Інжиніринг. – С. 9-13. Режим доступу: http://dwl.kiev.ua/art/itb2022-1/2022itb1.pdf

етапі здійснюється кластерний аналіз відібраної мережі та знаходження об'єктів – центрів кластерів (Рис. 44).

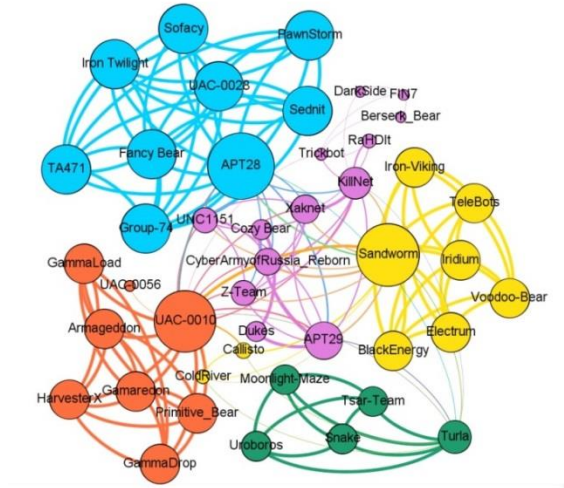


Рисунок 44 – Мережа основних хакерських угрупувань із росії/білорусії

Результати контент-моніторингу інтернет-ресурсів і подальшої кластеризації вказують на переважну приналежність розглянутих хакерських угрупувань до спецслужб рф і білорусії, а саме ФСБ рф; ГУ ГШ ЗС рф (ГРУ); СЗР рф; Міністерство оборони рб; інші проросійські угрупування.

Наведені методики враховують приховані знання, внесені експертним мережевим середовищем. Разом з цим, публікація детальної інформації з полів кібервійни у відкритих джерелах, безумовно, може бути застосована і ворогом, тому, вочевидь, вимагає деяких обмежень, які мають законодавчо регулюватись у сучасних умовах.

На українському ринку в сегменті інформаційно-аналітичних систем конкурентної розвідки представлені такі системи, як X-SCIF, The-hound (браузер, розроблений для розширення можливостей українських правозахисників, журналістів, правоохоронців та OSINT-дослідників під час збору ними даних з відкритих джерел), Bellingcat's map on Ukraine (Карта із геолокованими пошкодженнями цивільної інфра-

структури під час російського вторгнення в Україну.), scanbe.io (Сервіс шукає за відкритими державними реєстрами, бази даних доповнюються та оновлюються)), Телеграм-боти накшталт OpenDataUABot (для пошуку та моніторингу відкритих державних даних в Україні), DataLeakBot (Телеграм-бот, який здійснює пошук у великій кількості витоків даних – злитих баз) тощо.

Слід зазначити, що далеко не всі з названих систем мають повний функціонал та відповідні модулі, які забезпечують виконання всього спектра завдань конкурентної розвідки.

2.7.6 X-Scif

Як одну з найбільш повнофункціональних вітчизняних систем, обробка інформації в якій відповідає класичному інформаційному розвідувальному циклу, можна назвати систему X-SCIF.

Розглянемо, як реалізуються етапи розвідувального циклу за допомогою цієї системи; для цього зупинимося на описі можливостей системи X-SCIF дещо детальніше.

Онлайнова інструментальна корпоративна система моніторингу, агрегації та аналізу інформації X-SCIF являє собою програмно-технічний комплекс, призначений для розв'язання завдань автоматизованого збору, обробки, формування інтегрованого банку даних та аналізу різноманітної інформації.

Система X-SCIF забезпечує:

- моніторинг інформації з визначених користувачем вебсайтів (веб-сторінок) у мережі Інтернет (Інтранет) за заданими темами;
- пошук нових джерел інформації в мережі Інтернет за визначеними користувачем тематиками та їх подальше підключення до моніторингу;
- створення та збереження складних запитів за заданими темами у вигляді каталогізованого переліку або рубрики для подальшого автоматизованого моніторингу, пошуку чи контент-аналізу;
- приведення відібраної інформації до єдиного формату та її завантаження до сховища;
- фільтрацію, класифікацію, кластеризацію, рубрикацію та анонсування завантаженої повнотекстової інформації;

- автоматичне екстрагування (вилучення) з отриманої інформації сутностей (об'єктів і фактів);
- створення на основі завантаженої до системи неформалізованої повнотекстової та формалізованої фактографічної інформації інтегрованого банку даних (сховища) об'єктів, фактів, подій і документів, пов'язаних між собою різними видами та мотивами зв'язків, з урахуванням атрибутів достовірності, актуальності, а також вагових коефіцієнтів таких зв'язків;
- наскрізний пошук інформації за запитами або темами користувача, що охоплює як пошук у внутрішньому інтегрованому банку даних раніше завантаженої та накопиченої інформації, так і онлайновий метапошук у мережі Інтернет (пошукові системи, веб-сайти, блоги, соціальні мережі) та в інших підключених зовнішніх джерелах даних (офіційні реєстри державних органів, відкриті бази даних тощо);
- аналітичну обробку інформації (дає змогу аналізувати спільне згадування та виявляти неявні зв'язки між об'єктами, ідентифікувати об'єкти й групувати інформацію за сюжетами, будувати ланцюжки та графи зв'язків, аналізувати інформаційну активність, емоційне забарвлення документів, перетин заданих рубрик або тем, автоматично створювати інформаційний портрет відібраних за запитом документів, виокремлюючи згадувані в них об'єкти, джерела, регіони тощо, обчислювати індекс інформаційного сприяння та багато іншого);
- генерацію вихідних форм за визначеними користувачем параметрами (дає змогу автоматично створювати електронне досьє, схеми зв'язків, дайджести, огляди, порівняльні діаграми, інформаційні довідки та агреговані звіти);
- оперативну доставку результатів запитів різними каналами (до складу системи входить віртуальний офіс із власним віддаленим криптозахищеним сховищем документів та поштовою системою, що забезпечує як «онлайновий» безпечний доступ за шифрованим каналом до документів, що зберігаються в хмарі, так і «офлайнове» отримання підсумкових документів електронною поштою).

Структурно ІКС X-SCIF складається з кількох підсистем, орієнтованих на відповідні потреби корпоративних замовників, а саме:

- X-Stream – підсистема моніторингу веб-сайтів, створена на основі технології InfoStream, а також повнотекстова база даних (архів) неформалізованої інформації (статей, повідомлень тощо), яка автоматично поповнюється з 1996 року та є найповнішою серед наявних електронних архівів в Україні;
- X-Files – інтегрована база даних для накопичення різноманітної формалізованої довідково-фактографічної інформації, екстрагованої та агрегованої з усіх доступних системі джерел інформації, що перебувають на моніторингу, а також система наскрізного пошуку за внутрішніми та зовнішніми джерелами (веб-сайтами, блогами, онлайновими базами даних, соціальними мережами тощо);
- X-Office – система віртуального офісу, що забезпечує безпечний доступ до корпоративних ресурсів з будь-якої точки світу без встановлення додаткового програмного забезпечення. Система включає «хмарне» файлове сховище документів та захищену корпоративну веб-пошту. Додатково до віртуального офісу може бути інтегровано сервер VoIP-телефонії для ведення конфіденційних переговорів;
- X-Scoring – передскорингова система, яка дозволяє в автоматичному режимі здійснювати верифікацію даних та попередню перевірку надійності контрагентів (фізичних та юридичних осіб).

Зупинимося на розгляді кожної з підсистем детальніше.

Підсистема X-Stream, побудована на основі технології InfoStream компанії ElVisti, призначена для моніторингу інформації в мережі Інтернет за заданими користувачем параметрами, пошуку інформації за запитами або темами користувача, оперативної доставки результатів пошуку і, таким чином, мінімізації зусиль та економії часу, витраченого на пошук і обробку необхідної інформації. Підсистема X-Stream надає користувачеві доступ до інформації за цікавою для нього тематикою одночасно з великої кількості веб-сайтів, включно з тими

обраними, які він звик переглядати щодня (Рис. 45). Нині здійснюється автоматичний моніторинг понад 15000 джерел, потік інформації перевищує 100 000 документів на добу. Територіальне охоплення – російсько-, англомовні та україномовні видання України, Росії та інших країн ближнього і дальнього зарубіжжя (всього – 126 країн світу). За необхідності може бути охоплено будь-який веб-сайт, доступний у мережі Інтернет. Інформація з системи ніколи не видаляється, а переноситься до архіву. Архів публікацій ведеться безперервно з 1996 року і становить нині понад 85 мільйонів документів.

Відмінність цієї підсистеми від конкуруючих продуктів полягають у її обсягах та можливості індивідуального налаштування. Вона орієнтована не лише на швидку доставку загальних стрічок новин, яких багато у веб-просторі, а й на здійснення моніторингу за індивідуально заданими користувачем параметрами або архівного пошуку.

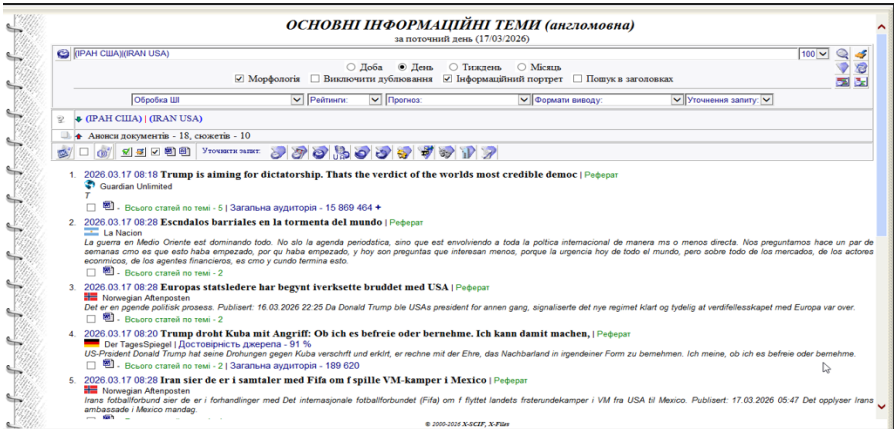


Рис. 45 – Виведення результатів пошуку

Перегляд інформації здійснюється через єдиний уніфікований інтерфейс. Користувач може в режимі реального часу не лише отримувати результати пошуку, а й формувати дайджести, інформаційні досьє, будувати сюжетні ланцюжки, аналізувати взаємозв'язок рубрик, інформаційну активність, інформаційні зв'язки та спільне згадування об'єктів тощо.

Нижче наведено приклади виведення результатів пошуку (рис. 45), перегляду окремого матеріалу (рис. 46) та аналітичної обробки за допомогою штучного інтелекту результатів пошуку інформації в мережі Інтернеті та соцмережах (рис. 47).

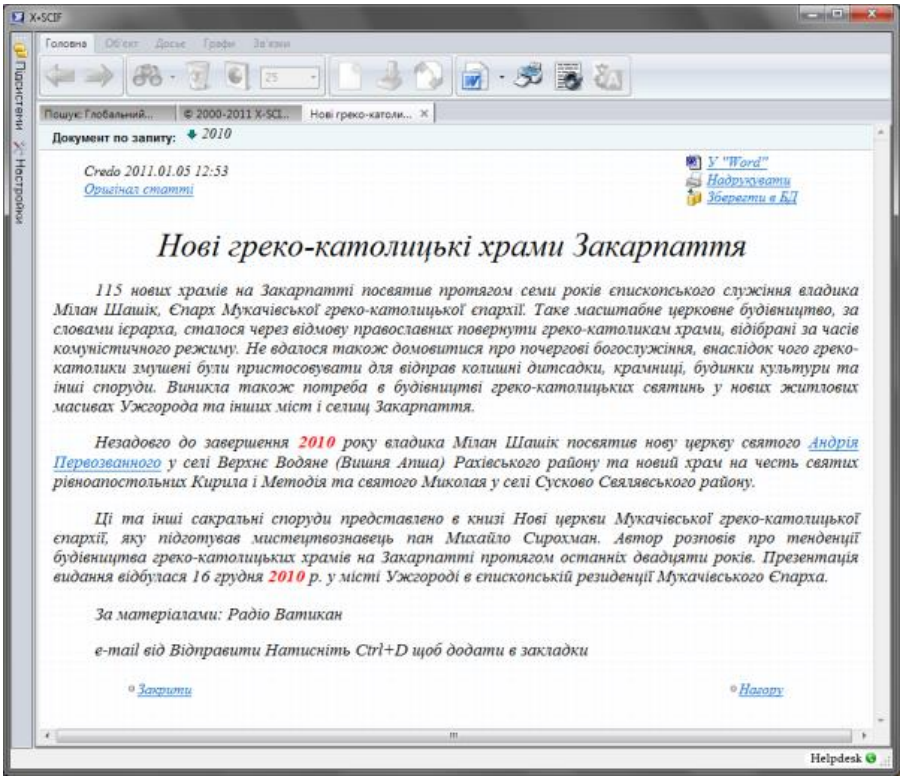


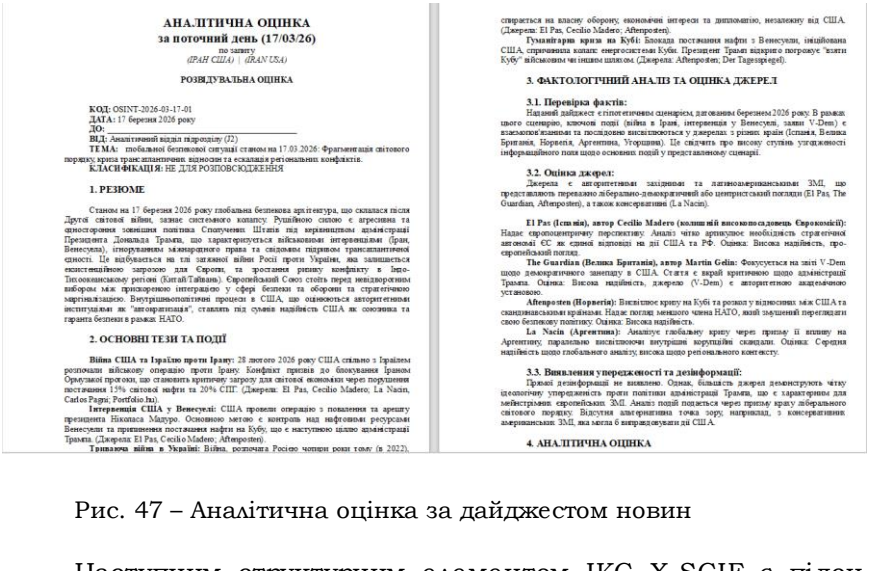
Рисунок 46 – Перегляд статті

Використання підсистеми X-Stream дозволяє:

- оперативно отримувати необхідну інформацію в міру її появи в Інтернеті, аналізувати події та своєчасно на них реагувати;
- формувати власні інформаційні канали, що визначаються запитами інформаційно-пошуковою мовою, ство-

рювати архіви для подальшої обробки та ретроспективного аналізу;

- аналізувати потік інформації, що надходить у режимі реального часу;
- своєчасно виявляти тенденції розвитку в політиці, економіці, військовій сфері, стан ринків товарів чи послуг;
- відстежувати інформацію про діяльність окремих організацій, партій, рухів, їхню PR-активність;
- оцінювати можливі сфери впливу конфліктних чи кризових ситуацій, здійснювати інформаційний контроль ймовірних джерел ризиків;
- проводити фактологічний аналіз отриманої інформації;
- виявляти упередженості та дезінформацію.



АНАЛІТИЧНА ОЦІНКА
за поточний день (17/03/26)
по запиту
(ІРАН США) | (IRAN USA)

РОЗВІДУВАЛЬНА ОЦІНКА

КОД: OSINT-2026-03-17-01
ДАТА: 17 березня 2026 року
ДО:
ВІД: Аналітичний відділ підрозділу (J2)
ТЕМА: глобальної безпекової ситуації станом на 17.03.2026: Фрагментація світового порядку, криза трансатлантичних відносин та ескалація регіональних конфліктів.
КЛАСИФІКАЦІЯ: НЕ ДЛЯ РОЗПОВСЮДЖЕННЯ

1. РЕЗЮМЕ

Станом на 17 березня 2026 року глобальна безпекова архітектура, що склалася після Другої світової війни, зазнає системного колапсу. Рушійною силою є агресивна та односторонньо зовнішня політика Сполучених Штатів під керівництвом адміністрації Президента Дональда Трампа, що характеризується військовими інтервенціями (Іран, Венесуела), ігноруванням міжнародного права та свідомим підривом трансатлантичної єдності. Це відбувається на тлі затяжної війни Росії проти України, яка залишається екзистенційною загрозою для Європи, та зростання ризику конфлікту в Індо-Тихоокеанському регіоні (Китай/Тайвань). Європейський Союз стоїть перед невідворотним вибором між прискореною інтеграцією у сфері безпеки та оборони та стратегічною маргіналізацією. Внутрішньополітичні процеси в США, що оцінюються авторитетними інституціями як "автократизація", ставлять під сумнів надійність США як союзника та гаранта безпеки в рамках НАТО.

2. ОСНОВНІ ТЕЗИ ТА ПОДІЇ

Війна США та Ізраїлю проти Ірану: 28 лютого 2026 року США спільно з Ізраїлем розпочали військову операцію проти Ірану. Конфлікт призвів до блокування Іраном Ормузької протоки, що становить критичну загрозу для світової економіки через порушення постачання 15% світової нафти та 20% СПГ. Джерела: El Pas, Cecilio Madero, La Nacin, Carlos Pagni; Portfolio Ini)

Інтервенція США у Венесуелі: США провели операцію з повалення та арешту президента Ніколаса Мадуро. Основною метою є контроль над нафтовими ресурсами Венесуели та припинення постачання нафти на Кубу, що є наступною ціллю адміністрації Трампа. (Джерела: El Pas, Cecilio Madero, Aftenposten).

Тривалість війни в Україні: Війна, розпочата Росією чотири роки тому (в 2022),

спирається на власну оборону, економічні інтереси та дипломатію, незалежну від США. (Джерела: El Pas, Cecilio Madero; Aftenposten).

Гуманітарна криза на Кубі: Блокада постачання нафти з Венесуели, ініційована США, спричинила колапс енергосистеми Куби. Президент Трамп відкрито погрожує "взяти Кубу" військовим чи іншим шляхом. (Джерела: Aftenposten, Der Tagesspiegel).

3. ФАКТОЛОГІЧНИЙ АНАЛІЗ ТА ОЦІНКА ДЖЕРЕЛ

3.1. Перевірка фактів:

Наданий дайджест є гіпотетичним сценарієм, датованим березнем 2026 року. В рамках цього сценарію, ключові події (війна в Ірані, інтервенція у Венесуелі, заяви V-Dem) є взаємопов'язаними та послідовно висвітлюються у джерелах з різних країн (Іспанія, Велика Британія, Норвегія, Аргентина, Угорщина). Це свідчить про високу ступінь узгодженості інформаційного поля щодо основних подій у представленому сценарії.

3.2. Оцінка джерел:

Джерела є авторитетними західними та латиноамериканськими ЗМІ, що представляють переважно ліберально-демократичний або центристський погляди (El Pas, The Guardian, Aftenposten), а також консервативні (La Nacin).

El Pas (Іспанія), автор Cecilio Madero (колишній високопосадовець Єврокомісії): Надає європоцентричну перспективу. Аналіз чітко артикулює необхідність стратегічної автономії ЄС як єдиної відповіді на дії США та РФ. Оцінка: Висока надійність, проєвропейський погляд.

The Guardian (Велика Британія), автор Martin Gelin: Фокусується на звіті V-Dem щодо демократичного занепаду в США. Стаття є вирій критичною щодо адміністрації Трампа. Оцінка: Висока надійність, джерело (V-Dem) є авторитетною академічною установою.

Aftenposten (Норвегія): Висвітлює кризу на Кубі та розкол у відносинах між США та скандинавськими країнами. Надає погляд західного члена НАТО, який зазнавий переглядати свою безпекову політику. Оцінка: Висока надійність.

La Nacin (Аргентина): Аналізує глобальну кризу через призму її впливу на Аргентину, паралельно висвітлюючи внутрішні корупційні скандали. Оцінка: Середня надійність щодо глобального аналізу, висока щодо регіонального контексту.

3.3. Виявлення упередженості та дезінформації:

Прямої дезінформації не виявлено. Однак, більшість джерел демонструють чітку ідеологічну упередженість проти політики адміністрації Трампа, що є характерним для мейнстримних європейських ЗМІ. Аналіз подій подається через призму кризу ліберального світового порядку. Відсутня альтернативна точка зору, наприклад, з консервативних американських ЗМІ, яка могла б виправдовувати дії США.

4. АНАЛІТИЧНА ОЦІНКА

Рис. 47 – Аналітична оцінка за дайджестом новин

Наступним структурним елементом ІКС X-SCIF є підсистема X-Files (не слід плутати з відомою російською системою). Дана підсистема призначена для накопичення та зберігання формалізованої інформації, отриманої з усіх доступних джерел, здійснення наскрізного пошуку та подальшої аналітичної обробки знайденої інформації.

Інформація, отримана з різних джерел, обробляється, формалізується, приводиться до єдиного вигляду та записується-

ся в інтегровану базу даних, яка структурно охоплює об'єкти та зв'язки між ними. Її структура, розроблена з урахуванням практичних потреб аналітиків, включає понад 40 типів об'єктів обліку та понад 1000 мотивів зв'язків між ними.

Пошук необхідної інформації здійснюється засобами глобального пошуку, який виконується за всіма доступними базами даних, а також передбачає автоматичне отримання інформації від онлайн-постачальників інформації.

На рис. 48 наведено інтерфейс введення запитів для глобального пошуку.

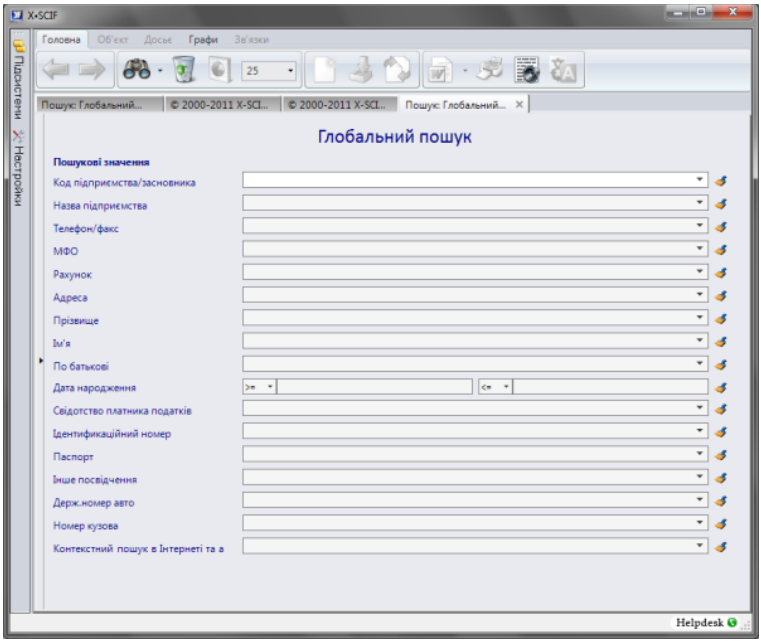


Рис. 48 – Інтерфейс введення запитів глобального пошуку

Підсистема дозволяє представляти результати пошуку у різній формі, найбільш зручній для вирішення поточного завдання.

Однією з найбільш поширених форм представлення відібраних даних є інформаційне досьє (Рис. 49). Дана форма дозволяє відображати інформацію про об'єкт обліку інтегрова-

ної бази даних у вигляді, в якому представлені всі реквізити цього об'єкта обліку, а також усі пов'язані з ним записи в інших базах даних.

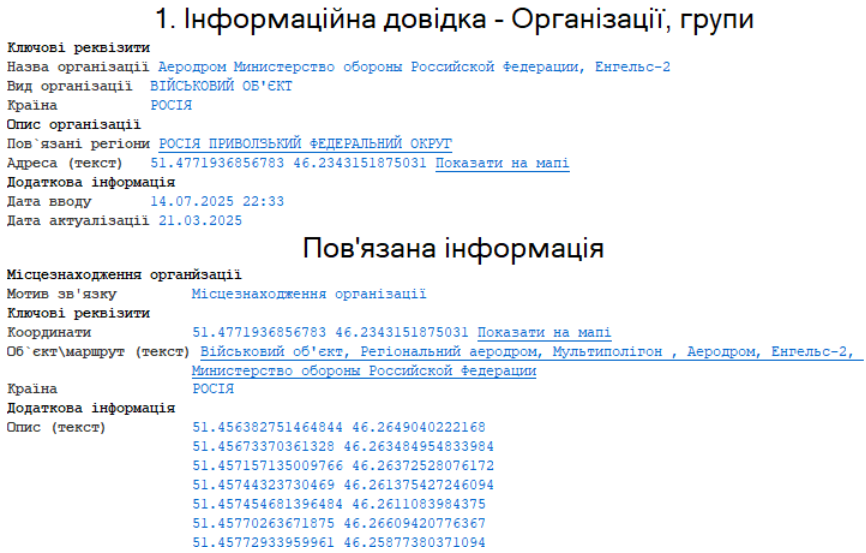


Рис. 49 – Інформаційне досьє

Будь-яка адресна інформація в системі автоматично переводиться у формат точних координат (14 знаків після градусів), що забезпечує прозору інтеграцію з геоінформаційними системами, такими як, OpenStreetMap, Google Maps, GoogleEarth та іншими. На мапі можливо проводити кластерізацію об'єктів, показувати зв'язки з іншими об'єктами з інформаційного досьє, переходити до візуального представлення схеми графу зв'язків (Рис. 50).

Ще одним прикладом форми подання даних є графічна схема досьє. Відібрані об'єкти разом із їхніми зв'язками відображаються у вигляді графа, у якому вершинами виступають об'єкти обліку, а ребрами – зв'язки між відповідними об'єктами (рис. 51). Така форма подання дозволяє здійснювати аналітичні дослідження як явних, так і неявних зв'язків об'єктів обліку, відображати їх на екрані у вигляді графів у різних масштабах, друкувати схеми цих графів тощо. Також у цьому режимі користувачеві доступні всі інструменти редагування, введення та

видалення інформації, що забезпечують інтуїтивне та швидке редагування формалізованих даних.

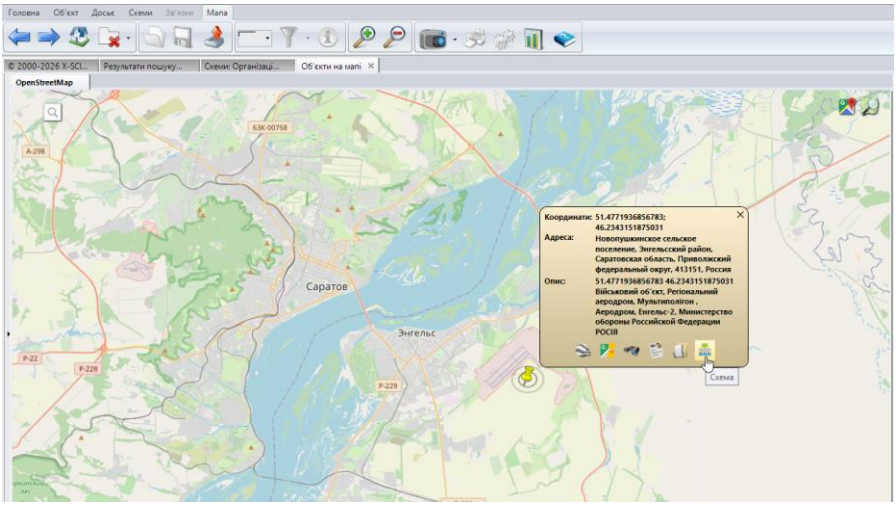


Рисунок 50 – Представлення об’єктів на мапі

Для аналітичної обробки великих обсягів однотипної інформації в системі передбачено аналіз за допомогою штучного інтелекту та механізм агрегованих форм. Він дозволяє на підставі вихідної інформації, яка важко піддається безпосередньому аналізу, будувати агреговані форми, графіки, проводити розрахунок інтегральних характеристик.

Однією з ключових можливостей підсистеми X-Files є модуль ETL (Extract-Transform-Loading) - модуль обробки і автоматизованого завантаження повнотекстових документів та розпізнавання сутностей в них за допомогою штучного інтелекту. Він дозволяє без участі оператора створювати різноманітні об'єкти обліку (особи, компанії, телефони, адреси, електронні адреси та інші) та встановлювати зв'язки між ними на основі неформалізованих документів (тексти, анкети, картки тощо).

В наслідок чого «інформаційна руда» – первинна вхідна розрізнена і неструктурована інформація, отримана в результаті моніторингу, може бути після обробки штучним інтелек-

том представлена у вигляді різних формалізованих документів - інформаційної довідки, аналітичного звіту, тощо (Рис. 52).

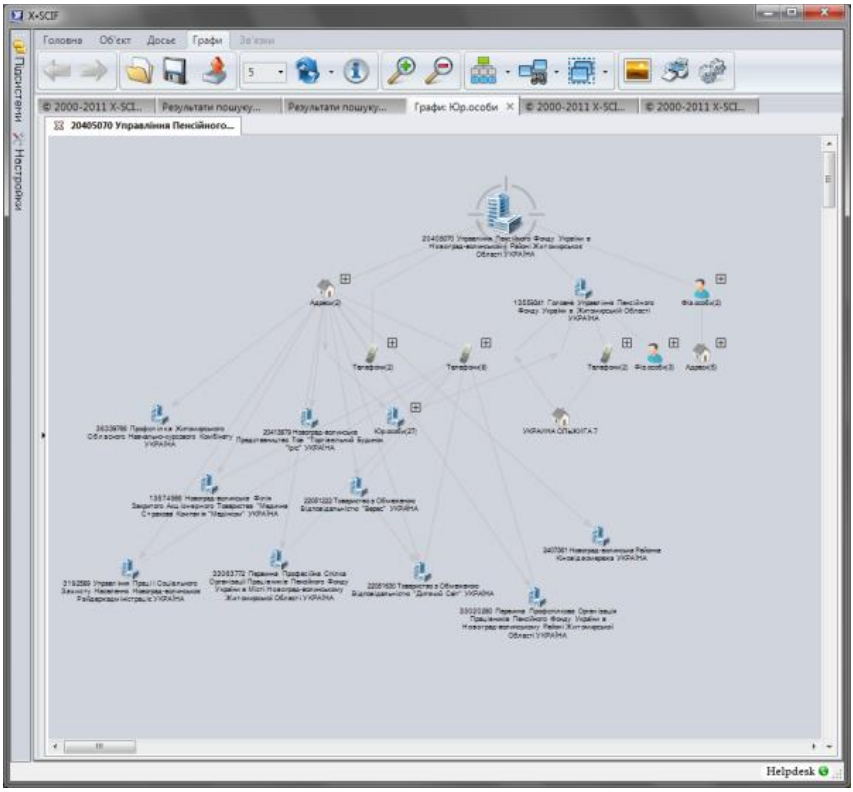


Рис. 51 – Візуалізація графів зв'язків об'єкту уваги

[redacted] УМАЛАТ МАГОМЕДРАСУЛОВИЧ, 09.12.198[redacted]
(інформаційна довідка)

Дата: 19.03.2026 №____ на вх. № ____

Фотографія, стисла характеристика

Коротка характеристика

[redacted] Умалат Магомедрасулович, [redacted] року народження, уродженець Дагестану, громадянин РФ. Імовірно, є діючим військовослужбовцем збройних сил РФ у званні капітана на посаді командира мотострілецької роти (МСР), на що вказують численні записи в телефонних книгах інших осіб. Паралельно з військовою службою, працював водієм. Наразі проживає в Ленінградській області. Володіє автомобілем KIA Cerato 2018 року. Активний користувач банківських послуг та державних онлайн-сервісів. Має відкриті виконавчі провадження щодо стягнення боргів. У соціальних мережах використовує псевдонім "Саид Омаров".

Основна інформація

ПІБ: [redacted] Умалат Магомедрасулович (рос. [redacted] Умалат Магомедрасулович).
Дата народження: 09.12.198[redacted] (в базі ФССП зустрічається помилкова дата 12.09.198[redacted]).
Місце народження: с. Мамедкала, Дербентський р-н, Республіка Дагестан, СРСР.
Ідентифікаційні коди:
ІПН (Індивідуальний податковий номер): [redacted]
СНІЛС (Страховий номер індивідуального особового рахунку): [redacted]

Контактні дані

Номери телефонів:
+790342[redacted] (основний)
+798180[redacted]
+792980[redacted]
+799962[redacted]
+793159[redacted]
+781242[redacted] (стаціонарний)
+796005[redacted] (пов'язаний з імовірним братом) .....

Рис. 52 – Приклад виведення у файл інформаційного OSINT-досье, сформованого за допомогою штучного інтелекту на основі автоматичного моніторингу мереж

Для забезпечення віддаленої взаємодії користувачів та забезпечення ефективної спільної роботи призначена підсистема X-Office. До її складу, своєю чергою, входять такі підсистеми:

- «Корпоративна веб-пошта». Забезпечує роботу з корпоративною електронною поштою з будь-якого місця через шифрований канал зв'язку, без необхідності попереднього налаштування та без залишення «слідів» на комп'ютері користувача;
- «Файлове сховище документів». Являє собою віддалене захищене файлове сховище, доступне з будь-якої точки мережі виключно для членів закритої групи. Надає доступ до особистих і корпоративних документів із можливістю спільної роботи кількох користувачів. Сховище документів забезпечує наскрізний пошук за змістом документів. Розмежування прав доступу до тексту документа здійснюється відповідно до профілю доступу користувача або за дозволом автора документа (Рис. 53);
- «Переговорна». Надає користувачам системи можливість спілкування всередині закритої групи в текстовому, голосовому та відеорежимах за захищеним протоколом зв'язку. Також доступна функція здійснення дзвінків на стаціонарні та мобільні телефони поза межами групи з неможливістю ідентифікації вихідного абонента.

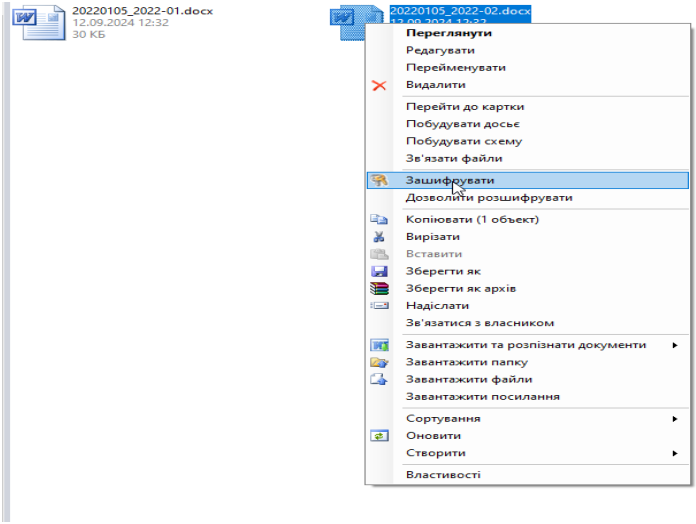


Рис. 53 – Операції, доступні в файловому сховищі

Підсистема X-Office має низку особливостей, які надають її користувачам суттєві переваги порівняно з аналогічними програмно-апаратними комплексами. Авторизація в підсистемі здійснюється за комбінацією відбитка пальця та пароля, а для комунікацій використовуються шифровані канали зв'язку та лише довірені сертифікати, що унеможливлює аналіз трафіку на рівні інтернет-провайдера або в будь-якій іншій точці перехоплення. Після завершення роботи з підсистемою на комп'ютері користувача не залишається жодних слідів її функціонування.

Файлове сховище корпоративних документів, яке входить до складу підсистеми X-Office, надає користувачеві можливість працювати з віддаленим хмарним сховищем так само просто, як із папкою на локальному диску його комп'ютера. Підсистема забезпечує прозору для користувача роботу з різними джерелами (локальний диск, корпоративні документи, файлові сховища). Редагування офісних документів можливе без встановлення та налаштування додаткового програмного забезпечення. Розмежування доступу до різних файлів здійснюється на основі шаблонів доступу, заданих адміністратором. У разі наявності особливо секретної інформації користувач має можливість додатково обмежити доступ до неї засобами шифрування безпосередньо з інтерфейсу програми.

Для перевірки надійності контрагентів в автоматичному режимі доступна підсистема X-Scoring. Підсистема реалізована на основі технології XML Web service, що забезпечує прозору інтеграцію з більшістю наявних систем на боці замовника. Попри великий обсяг інформації, що постійно поповнюється і на підставі якої ухвалюється рішення щодо надійності контрагента, система надає відповідь менш ніж за 3 секунди.

Алгоритм ухвалення рішення може гнучко налаштовуватися відповідно до потреб конкретного замовника. Типовий алгоритм проведення перевірки та ухвалення рішення щодо проблемнності фігуранта перевірки складається з таких етапів:

- Автоматичний пошук застережень передбачених нормативними актами України та відомчими актами відомств і організацій (Кабмін, РНБО, НБУ, МО, МВС, СБУ та інших), як самого фігуранта перевірки, так і його зв'язків до 3-го рівня .

- оцінка економічної платоспроможності клієнта за наданими ним відомостями;
- автоматична перевірка за банком даних, спрямована на перевірку відповідності відомостей, що надаються позичальником, та виявлення можливих спроб шахрайства з боку недобросовісних позичальників;
- детальна перевірка контрагента, заявка якого пройшла всі попередні етапи.

Слід зазначити, що повнофункціональні системи конкурентної розвідки не завжди є доступними чи навіть необхідними – з огляду на їхню вартість або інші чинники. Разом із тим, окремі завдання конкурентної розвідки можуть бути частково вирішені за допомогою цілком доступних інструментів. Використання нових підходів, а також відкритих, доступних і відносно недорогих інформаційних джерел дозволяє вже нині ефективно підтримувати процес ухвалення управлінських рішень у багатьох сферах бізнесу, зокрема й на стратегічному рівні.

Висновки до розділу 2

Проведений у другому розділі аналіз технологічного забезпечення комп'ютерної конкурентної розвідки засвідчує, що сучасний етап розвитку цієї діяльності неможливий без застосування спеціалізованих інформаційно-аналітичних систем та автоматизованих процедур обробки даних. Традиційні пошукові системи виявляються недостатньо ефективними для вирішення завдань розвідки через високий рівень інформаційного шуму, динамічність веб-простору та обмеженість доступу до прихованих сегментів мережі, що зумовлює необхідність використання комплексних рішень на основі контент-моніторингу та глибинного аналізу текстів. Ключовим елементом технологічного стеку виступають методи Text Mining та Data Mining, які дозволяють трансформувати неструктуровані масиви інформації у структуровані знання шляхом класифікації, кластеризації та екстрагування сутностей, що є критично важливим для виявлення неочевидних закономірностей та прихованих зв'язків між об'єктами моніторингу.

Інтеграція концепції великих даних Big Data у процеси конкурентної розвідки забезпечує можливість обробки надвеликих обсягів різнорідної інформації з необхідною швидкістю,

використовуючи розподілені обчислювальні середовища такі як Hadoop та Storm, а також спеціалізовані системи зберігання та візуалізації на кшталт Elastic Stack та графових баз даних Neo4j. Математичне підґрунтя аналізу, зокрема методи дослідження часових рядів, вейвлет-аналіз та теорія складних мереж, надає інструментарій для кількісної оцінки динаміки інформаційних потоків, виявлення аномалій та розпізнавання інформаційних операцій на ранніх етапах їх проведення. Практична реалізація розглянутих технологій представлена широким спектром існуючих програмних продуктів та систем, таких як X-SCIF, RCO, PolyAnalyst та інші, які демонструють можливість побудови єдиного інформаційного простору взаємопов'язаних об'єктів та фактів. Таким чином, технологічний арсенал конкурентної розвідки формується як багаторівнева система, що поєднує засоби збору, інтелектуальної обробки, математичного моделювання та візуалізації даних, що дозволяє перейти від інтуїтивного прийняття рішень до управління, заснованого на достовірних прогнозах та верифікованих знаннях, отриманих виключно з легитимних відкритих джерел.

3. Джерела інформації

В інформаційно-аналітичній роботі важливе значення має можливість доступу до джерел даних, інформації та знань. При цьому головною проблемою є знаходження змістовних і надійних джерел з-поміж усіх загальнодоступних. Коли такі джерела знайдено, включаються механізми перетворення даних на знання, для чого застосовуються відповідні технології. Під даними зазвичай розуміють «сирі», необроблені відомості, що ґрунтуються на фактах. Це можуть бути статистичні дані, факти з біографій ключових персон або, наприклад, відомості про звітність окремих компаній. Інформація являє собою вже певним чином оброблені та проаналізовані дані. Кінцевим же інформаційним продуктом будь-якої аналітичної роботи є знання – синтезовані висновки, рекомендації для прийняття рішень.

Інформація, як було зазначено вище, може бути отримана з офіційних, відкритих джерел, ЗМІ, оголошень, реклами, фірмових, банківських, урядових звітів, баз даних, від експертів шляхом аналізу або спеціальної обробки даних, текстів.

Нижче наведено докладний перелік видів інформаційних джерел, які найчастіше використовуються при конкурентній розвідці.

1. Прес-релізи компаній, офіційні заяви від імені компаній про нові технології, нові напрямки, угоди, перспективи. Такі прес-релізи створюються компаніями для власної популяризації, привернення уваги потенційних клієнтів, інвесторів, які шукають вигідні варіанти вкладення своїх коштів. Часто в таких заявах присутня інформація про наміри, заплановані події. Прес-релізи доступні на веб-сайтах компаній, у PR-службах, на загальних і профільних спеціалізованих майданчиках для розміщення прес-релізів.
2. Інтерв'ю співробітників компаній, відповідні матеріали в ЗМІ. В інтерв'ю інтерес становлять плани компаній. При цьому з боку служби конкурентної розвідки допускається ініціювання інтерв'ю когось із співробітників об'єкта інтересу.

3. Висловлювання співробітників компаній на форумах, у блогах, у приватних бесідах. При цьому можуть виявлятися плани компаній, кадрова політика, атмосфера в колективі тощо. Джерела інформації: 1) інтернет-ресурси (спеціалізовані форуми, блоги співробітників), блоги експертів, групи в соціальних мережах; 2) виставки, конференції, курси підвищення кваліфікації, професійні заходи.
4. Тендери, закупівлі. Предмети закупівель, обладнання, виконавці. Джерела інформації: 1) інтернет-ресурси (веб-сайти компаній, торговельні майданчики, профільні форуми); 2) партнери досліджуваної компанії, ті, хто брав участь у їхніх тендерах, у клієнтів і постачальників.
5. Патентні документи, авторські свідоцтва компанії та її співробітників. Для завдань конкурентної розвідки цікавий їхній зміст, спрямованість, списки співавторів. Інформація розміщується на відповідних сайтах. Для України: https://ukrpatent.org/; Google Patents: https://patents.google.com/; Євразійське патентне відомство: www.eapo.org. Патентування можливе в будь-якій країні, найкращі варіанти – країна реєстрації організації, країна ведення бізнесу, крім того США, Євросоюз, Росія, Японія та Китай.
6. Розробки компанії: ті, що ведуться, фінансуються, розробки, якими компанія цікавиться. Спостереженню підлягають спроби компанії проводити дослідження: закупівля специфічного обладнання, прийом на роботу спеціалістів, переговори, відвідування відповідних організацій тощо.
7. Активність компанії на ринку злиттів і поглинань (M&A). Інформація про те, які організації поглинаються, планують поглинути або ведуть переговори про поглинання. Інформацію можна отримати в Антимонопольному комітеті (АМК) України, аналогічних відомствах інших країн, за новинними повідомленнями на веб-ресурсах, присвячених M&A.
8. Вакансії компанії (відкриті, закриті), повідомлення про активний пошук співробітників, вимоги до вакансій, умови. Джерело інформації: веб-сайт компанії, сайти з

пошуку роботи та на сайти агенцій, з якими компанія співпрацює.

9. Курси підвищення кваліфікації, навчання персоналу – вказівка на пріоритети в розвитку компанії. Інтерес становить те, чого навчають, яких спеціалістів запрошують для навчання, які вимоги висувають при залученні тих, хто навчає, які терміни навчання, яка кількість персоналу навчається.
10. Подяки та нагороди компанії та її співробітників.
11. Участь у заходах (виставки, конференції, круглі столи, презентації). З'ясування, в яких заходах беруть участь компанії, їхня спрямованість, коло учасників.
12. Участь в організаціях (спілки, асоціації, конфедерації тощо) – інформація про те, в яких об'єднаннях бере участь компанія, наскільки активно бере участь, що отримує від участі, на що розраховує, як використовує.

Інформація характеризується якісними, кількісними та ціннісними показниками. До якісних характеристик зазвичай відносять: достовірність, об'єктивність та однозначність інформації. До кількісних характеристик – її повноту (відсутність нез'ясованих прогалин) та релевантність (ступінь відповідності суті поставлених питань і завдань). Ціннісними характеристиками є вартість та актуальність інформації.

Діяльність конкурентної розвідки заснована на використанні лише легітимних джерел інформації, яких цілком достатньо для прийняття управлінських рішень у сфері бізнесу, необхідно лише провести деяку інформаційно-аналітичну обробку наявних відкритих даних. Серед таких джерел інформації можна назвати: дані статистики, матеріали з веб-сайтів, соціальних мереж, ЗМІ, галузевих звітів тощо.

Багато служб конкурентної розвідки не завжди можуть відокремити нелегітимну частину інформації від легальної, а замовник, як правило, цікавиться кінцевими результатами, джерела для нього виступають лише як підтвердження, проміжні дані. Водночас, солідні замовники самі зацікавлені в тому, щоб інформація добувалася законними засобами, щоб аналітичний звіт був легальним.

У конкурентної розвідки в останні десятиліття з'явилося і розвинулося до небачених раніше масштабів нове інформаційне джерело – веб-простір мережі Інтернет. Сьогодні, за

оцінками експертів, Інтернет за кількістю інформації знаходиться на першому місці, випереджаючи ЗМІ, галузеві видання та одержувані від колег новини, спеціальні огляди, закриті бази даних. При цьому у відкритих джерелах і спеціалізованих базах даних, доступних в Інтернеті, міститься більша частина інформації, необхідної для проведення конкурентної розвідки, проте залишається відкритим питання її знаходження та ефективного використання. Останні дослідження інформаційного веб-простору показали, що доступний через традиційні інформаційно-пошукові системи трильйон веб-сторінок – це лише «поверхнева видима частина айсберга». Близько 40 % усієї інформації в Інтернеті доступно безкоштовно. Навігацію по даному інформаційному простору забезпечують більше мільйона пошукових систем і каталогів, але й вони охоплюють лише малу частину інформаційних ресурсів. Прихованих і невидимих (deep, invisible) ресурсів мережі Інтернет значно більше – це, перш за все, динамічно генеровані сторінки, файли різноманітних форматів, інформація з численних баз даних. До «прихованого» веб можна віднести й такі мережі, як BitTorrent, DirectConnect, EMule, Napster та ін.

Сьогодні для конкурентної розвідки основними джерелами інформації слугують Інтернет, преса, а також відкриті бази даних. Дуже популярні серед спеціалістів з конкурентної розвідки бази даних державних і статистичних органів, торгово-промислових палат, органів приватизації тощо. Велику користь приносять і окремі доступні бази даних інших органів влади. Останнім часом дедалі популярнішими стають бази даних на основі архівів ЗМІ, у тому числі й мережевих. Традиційно конкурентна розвідка спирається на такі джерела інформації, як опубліковані документи відкритого доступу, що містять огляди товарного ринку, інформацію про нові технології, створення партнерств, злиття та придбання, оголошення про робочі вакансії, про виставки та конференції тощо. Широко використовуються відомості, що знаходяться в документах, які вже є в компаніях, що ведуть конкурентну розвідку, результати маркетингових досліджень, інформація, отримана на конференціях, при спілкуванні з клієнтами та колегами. Більша частина цих даних потрапляє в мережеву пресу, прес-релізи або публікується на корпоративних веб-сайтах. Тому останнім часом великої

популярності набувають бази даних на основі архівів мас-медіа, у тому числі (і переважно) мережевих.

3.1 Веб-ресурси

Веб-ресурси є важливим джерелом інформації для OSINT, які дозволяють отримати доступ до даних, що можуть бути використані для виявлення загроз та вразливостей у сфері кібербезпеки. Ці ресурси надають різноманітну інформацію, яку можна використовувати в аналітичній роботі, виявлення загроз та підвищення рівня кібербезпеки. Нижче наведено приклади веб-ресурсів, які використовуються у OSINT:

- WHOIS доменів, що застосовуються для визначення інформації про власника домену. Приклади ресурсів: WHOIS Lookup55 (whois.icann.org), WHOIS Domain Search (who.is).
- DNS бази даних, що застосовуються для отримання інформації про DNS-записи та конфігурації доменів. Приклади ресурсів: DNSstuff (https://www.dnsstuff.com/), MXToolbox (https://mxtoolbox.com/).
- Соціальні мережі, що дозволяють спостерігати за активністю їх користувачів. Приклади ресурсів: Facebook, Telegram, LinkedIn, Instagram, Reddit.
- Форуми та спільноти, що дозволяють виявляти обговорення та інформацію про кіберзагрози. Приклади ресурсів: Hack Forums (https://hackforums.net/), RaidForums (https://raidforums.com/), Nulled (www.nulled.to), Dark0de (http://darkode.cybercrime-tracker.net/), BlackHatWorld (BHW, www.blackhatworld. com).
- Архіви веб-сайтів та блогів дозволяють проводити аналіз контенту для виявлення нових трендів у кібербезпеці. Приклади ресурсів: Wayback Machine (archive.org), Google Blog Search (https://blog.google/ products/search/).
- Бази даних вразливостей та експлойтів дозволяють виявляти

[55] https://www.namecheap.com/domains/whois/

інформацію про вразливості та доступні експлойти, наприклад: Exploit Database (https://www.exploit-db.com/), National Vulnerability Database (https://nvd.nist.gov/), Vulnerability & Exploit Database (https://www.rapid7.com/db/).

- Бази даних і масиви, які містять інформацію щодо змін в інфраструктурі і дозволяють слідкувати за змінами, серед яких: SecurityTrails (https://securitytrails.com/), RiskIQ PassiveTotal (https://community.riskiq.com/), URL Classification Database (https://www.forcepoint.com/product/feature/forcepoint-url-database).

- Новинні джерела та дошки оголошень застосовуються для виявлення потенційних загроз. Приклади ресурсів: Google News, Bing News, Threatpost (https://threatpost.com/).

- Технічні форуми дозволяють зчитувати та аналізувати обговорення про технічні аспекти кібербезпеки. Приклади ресурсів: Stack Overflow (https://stackoverflow.com/), Information Security Stack Exchange (https://security.stackexchange.com/), Information Security Forum (https://www.securityforum.org).

3.2 RSS-фіди

RSS (Really Simple Syndication) є стандартом для форматування вмісту новин, блогів та інших інтернет-ресурсів в структуровану форму, яку легко можна використовувати для автоматизованого отримання оновлень[56].

Масиви документів, які надаються на ресурсах веб-серверів у форматі RSS мають назви «RSS-фіди» (від англ. Feed – годувати, постачати), канали інформації. Адреси цих фідів задаються у явному вигляді на сторінках веб сайтів (стандартне позначення – ), або їх можна знайти у вихідному коді HTML сторінок, наприклад, як у фрагменті першої сторінки веб-сайту Cyber Security Hub (https://www.cshub.com/):

```
<li class="list-inline-item">
   <a aria-label="RSS"
```

[56] Ben Hammersley. Content Syndication with RSS. O' Reilly & Associates, Inc., 2003. – 208 p. ISBN:978-0-596-00383-8

```
    href="https://www.cshub.com/rss-feeds">
    <i class="fa fa-rss-square fa-2x text-white">
    RSS Feed</i></a>
</li>
```

Для автоматичного завантаження RSS-фіди зазвичай можна використовувати стандартні утиліти завантаження інтернет-контенту, яких існує дуже багато, наприклад програми cURL (https://curl.se/) або wget (https://www.gnu.org/software/wget/).

cURL – це утиліта для організації вибірки даних з веб-сайтів, яка надає можливість оперувати з файлами на боці веб-сервера за допомогою параметрів, що можуть бути переданими в рядку URL. За допомогою cURL можна отримувати веб-сторінки, не використовуючи для цього браузера. Крім HTTP-запитів, cURL підтримує SMTP, IMAP, Telnet, FTP та інші мережні протоколи. Базове використання cURL полягає у простому наборі у командній консолі команди cURL, за якою іде URL для завантаження. cURL за замовчуванням відображає вивід отриманого у стандартний потік виводу системи, тобто виклик cURL покаже програмний код сторінки в вікні терміналу.

Програма cURL може записати вивід до файлу при завданні опції «-o»:

curl -o example.html www.example.com

Такий виклик забезпечує збереження коду титульної сторінки www.example.com у файлі example.html.

Wget – це консольна утиліта для завантаження файлів за протоколами HTTP, HTTPS та FTP. Wget дає змогу рекурсивно завантажувати мережеві файли, копіювати як окремі сторінки, так і цілі веб-сайти, конвертувати посилання тощо. Ця утиліта портована й запускається на багатьох UNIX-подібних системах, Microsoft Windows, MacOS X тощо. Wget – не інтерактивна програма, тобто, після того, як її запущено з певними параметрами, вона виконує всі необхідні дії і не потребує додаткового втручання у свою роботу. Wget може працювати як пошуковий робот, тобто отримувати ресурси, на які посилаються елементи HTML сторінки, й рекурсивно просуватися веб-деревом, доки всі необхідні файли не будуть завантаженні.

Процедури скачування RSS-фідів (каналів інформації) мають вигляд:

```
curl –o habr https://habr.com/ru/rss/all/all/
```

або

```
wget -O habr https://habr.com/ru/rss/all/all/
```

Нижче наведено список RSS-фідів, які можна використовувати для формування масиву даних для подальших практичних робіт з кібернетичної безпеки[57]:

```
Graham Cluley RSS Feed: grahamcluley.com/feed
Krebs on Security RSS Feed: krebsonsecurity.com/feed
Dark Reading RSS Feed: darkreading.com/rss/all.xml
CSO Online RSS Feed: csoonline.com/feed
Naked Security RSS Feed: nakedsecurity.sophos.com/feed
JISA Softech RSS Feed: jisasoftech.com/feed
Venture in Security RSS Feed: ventureinsecurity.net/feed
```

Приклад фіду:

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0" xmlns: content=
"http://purl.org/rss/1.0/modules/content/"
xmlns:wfw="http://wellformedweb.org/CommentAPI/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:atom="http://www.w3.org/2005/Atom"
xmlns:sy="http://purl.org/rss/1.0/modules/syndication/"
xmlns:slash=http://purl.org/rss/1.0/modules/slash/
>
<channel>
<title>Krebs on Security</title>
<atom:link href="https://krebsonsecurity.com/feed/" rel="self"
type="application/rss+xml" />
<link>https://krebsonsecurity.com</link>
<description>In-depth security news and investigation </description>
<lastBuildDate>Thu, 29 Feb 2024 22:24:46 +0000</lastBuildDate>
<language>en-US</language>
<sy:updatePeriod>    hourly  </sy:updatePeriod>
<sy:updateFrequency>        1       </sy:updateFrequency>
<generator>https://wordpress.org/?v=6.2.2</generator>
<item>
<title>Fulton County, Security Experts Call LockBit&#8217;s
Bluff</title>
<link>https://krebsonsecurity.com/2024/02/fulton-county-security-
experts-call-lockbits-bluff/</link>
<dc:creator><![CDATA[BrianKrebs]]></dc:creator>
<pubDate>Thu, 29 Feb 2024 22:18:54 +0000</pubDate>
<category><![CDATA[Ransomware]]></category>
<category><![CDATA[Brett Callow]]></category>
<category><![CDATA[Emsisoft]]></category>
```

[57] https://rss.feedspot.com/cyber_security_rss_feeds/

```
<guid isPermaLink="false"> https://krebsonsecurity.com/?p=66580
</guid>
<description><![CDATA[The ransomware group LockBit told officials
with Fulton County, Ga. they could expect to see their internal
documents published online this morning unless the county paid a
ransom demand. Instead, LockBit removed Fulton County's listing from
its victim shaming website this morning, claiming county officials
had paid. But county officials said they did not pay, nor did anyone
make payment on their behalf. Security experts say LockBit was
likely bluffing and probably lost most of the data when  the gang's
servers were seized this month by U.S. and U.K. law
enforcement.]]></description>
<slash:comments>22</slash:comments>
</item>
…
</channel>
```

Файли за тематикою, що визначається інформаційними джерелами, завантажені в форматі RSS, який можна вважати обмінним, комунікаційним форматом.

Багато сучасних браузерів, поштових клієнтів та засобів миттєвого обміну повідомленнями можуть працювати з RSS-стрічками, серед них Safari, Maxthon, Miranda, Mozilla Firefox (до Firefox 63), Mozilla Thunderbird, Opera, Opera Mini, Microsoft Internet Explorer (починаючи з 7-ої версії), Vivaldi (з версії 4.0). Крім того, існують спеціалізовані застосунки (RSS-агрегатори), які збирають та обробляють інформацію RSS-каналів.

У контексті OSINT у сфері кібербезпеки, використання RSS-фідів може бути корисним з точки зору моніторингу та аналізу інформації. Наведемо деякі аспекти використання RSS-фідів у сфері кібербезпеки:

- Слідкування за новинами і публікаціями організацій, вчених та експертів у сфері кібербезпеки, що публікують свої дослідження, аналіз та новини через RSS-фіди. Використання цих фідів дозволяє отримувати автоматичні оновлення про нові загрози, вразливості та інші події.

- Моніторинг блогів та форумів у яких експерти з кібербезпеки публікують свої думки та враження в блогах та форумах. Використання RSS-фідів дозволяє слідкувати за цим контентом, отримуючи сповіщення про нові записи.

- Стеження за безпековими бюлетенями, які регулярні випускають виробники програмного забезпечення та ор-

ганізації, що займаються кібербезпекою. Ці бюлетені містять інформацію про вразливості та патчі. Підписка на RSS-фіди цих бюлетенів дозволяє отримувати оновлення безпосередньо в реальному часі.

- За допомогою RSS-фідів можна отримувати інформацію про події, конференції, виставки та інші важливі події у сфері кібербезпеки можна слідкувати за подіями в індустрії. Це дозволяє бути в курсі останніх тенденцій та розвитку галузі.

- Використання RSS-агрегаторів дозволяє об'єднувати фіди з різних джерел, створюючи єдиний потік інформації. Це полегшує відслідковування і аналіз оновлень в кібербезпеці.

Використання RSS у контексті автоматизованого OSINT полегшує збір та аналіз відкритих даних, допомагає зробити збір інформації ефективнішим та швидшим, дозволяючи аналітикам швидко отримувати актуальну та важливу інформацію. Наведемо лише деякі процеси, в яких враховуються особливості представлення RSS-фідів.

- Для збору інформації з RSS-фідів можна використовувати техніки скрейпінгу або автоматизовані скрипти. Такі інструменти можуть регулярно перевіряти фіди, вилучати нові записи та зберігати їх для подальшого аналізу.

- RSS-фіди надають структуровану інформацію, таку як заголовок, опис, посилання тощо. Під час автоматизованого збору, парсингу та обробки цих даних їх можна використовувати для ефективного вилучення корисної інформації.

- Фільтрація та урахування ключових слів для розпізнавання подій та інцидентів. Технічно можна налаштовувати фільтри для відсіювання непотрібної інформації за допомогою ключових слів, які можуть вказувати на конкретні теми або типи подій. Автоматизовані системи можуть використовувати розпізнавання подій для автоматичного виявлення інцидентів або новин, пов'язаних з кібербезпекою.

- RSS може бути однією з джерел інформації у великій системі OSINT. Інтеграція з іншими джерелами, такими як соціальні мережі, форуми чи веб-сайти, дозволяє отримати більш повний цільовий контекст.

- Застосування технологій обробки природної мови (NLP) дозволяє аналізувати текстовий вміст статей або новин, щоб визначити важливість інформації та її контексту.

3.3 Соціальні мережі

Соціальні мережі грають важливу роль у сфері OSINT, зокрема, у в галузі кібербезпеки. При цьому важливо враховувати, що збір та використання інформації з соціальних мереж повинні відбуватися в межах етичних та правових стандартів. Наведемо деякі напрямки використання соціальних мереж як джерела інформації у сфері кібербезпеки:

- Інформація, яку користувачі викладають на соціальних мережах, може містити відомості про їх технічну інфраструктуру, використані програмні продукти, а також можливі вразливості. Аналітики можуть здійснювати пошук цієї інформації для виявлення потенційних точок вразливостей.

- Соціальні мережі можуть слугувати джерелом для виявлення загроз та атак. Кіберзлочинці можуть розміщувати інформацію про свої плани або методи на цих соціальних платформах. Моніторинг активності в мережах може допомогти заздалегідь виявити можливі загрози.

- Фішинг та соціальна інженерія: Аналіз профілів у соціальних мережах може надати інформації для проведення фішингових атак або атак соціальної інженерії, що потребує захисту. Кіберзлочинці можуть намагатися використовувати інформацію про користувачів для отримання несанкціонованого доступу.

- Соціальні мережі містять велику кількість фотографій та відео. Аналіз цього мультимедійного контенту може допомогти виявити можливі загрози або ресурси, пов'язані з кібербезпекою.

- Кіберзлочинці інколи організовуються в групи чи спільноти на соціальних мережах для обговорення та обміну інформацією. Моніторинг таких груп може розкривати нові тенденції та загрози у сфері кібербезпеки.

- Аналіз профілів користувачів у соціальних мережах дозволяє створювати профілі цільових об'єктів для подальшого вивчення, які можуть включати інформацію про освіту, досвід роботи, професійні контакти та інші деталі, які можуть бути корисними для аналізу загроз.

- Аналіз комунікації та поведінки користувачів у соціальних мережах може надати інформацію про їхні інтереси, взаємодії та можливі мотивації, що може бути корисно для аналізу загроз та атак.

Технічні аспекти добування інформації з соціальних мереж в рамках OSINT у сфері кібербезпеки включають в себе різні методи та інструменти для збору та обробки даних, назвемо їх:

- Використання API (прикладний програмний інтерфейс) доступу до своїх даних, які надають деякі соціальні мережі. Використання API дозволяє отримувати структуровані дані, серед яких профільні інформації, дописи, зображення та інше. Прикладами соціальних мереж з API є X API (раніше Twitter API), Facebook Graph API, LinkedIn API тощо.

- В тих випадках, коли API недоступне або має великі обмеження, можна використовувати техніку веб-скрейпінгу для отримання інформації з публічних сторінок соціальних мереж.

- Аналіз метаданих: Метадані соціальних мереж можуть містити цінну інформацію, таку як геолокаційні дані з дописів або інформація про пристрої, які використовуються для входу, можуть допомогти в обстеженні діяльності користувачів.

- Соціальні мережі можуть містити геотеги, які показують, де були опубліковані дописи. Це може бути корисно для визначення фізичного місця розташування осіб чи ор-

ганізацій.

- Аналіз хештегів та ключових слів у дописах може розкрити тенденції, специфічні теми чи навіть можливі загрози в обговореннях.

- Використання методів машинного навчання та аналітики текстового контенту допомагає проводити семантичну індексацію, екстрагувати ключові слова, теми, а також розпізнавати наміри та настрій користувачів.

- Моніторинг активності у соціальних мережах та сповіщення про нові події або зміни в поведінці може сприяти вчасному виявленню можливих загроз.

- Аналіз зв'язків користувачів соціальних мереж може розкрити інформацію про їхні професійні зв'язки, групи і спільноти.

- Використання спеціалізованих інструментів OSINT на базі соціальних мереж, які комбінують в собі різні технічні підходи для добування та аналізу інформації.

3.4 Спеціальні бази даних

Останні дослідження веб-простору показали, що доступні через традиційні інформаційно-пошукові системи понад трильйон веб-сторінок – це лише «поверхнева видима частина айсберга».

Важливою проблемою є пошук інформації в «прихованому» або «глибинному» веб-просторі, де, як було зазначено вище, міститься незрівнянно більша кількість даних, потенційно цікавих для конкурентної розвідки, ніж у відкритій частині Інтернету.

Це, передусім, динамічні веб-сторінки, інформація з численних баз даних, які можуть становити великий інтерес для аналітичної роботи. До розряду «прихованого» веб належать і повнотекстові інформаційні системи типу LexisNexis або Factiva.

До «прихованих» ресурсів мережі Інтернет можна також віднести пірингові мережі, такі як BitTorrent, EDonkey, EMule, Gnutella, Kazaa.

Як уже було зазначено раніше, необхідної (зокрема й для конкурентної розвідки) інформації в мережі Інтернет значно більше, ніж її охоплюють універсальні пошукові машини. Припускають, що на відміну від «пізнаваної» частини мережі Інтернет, «прихована» частина виявилася в сотні разів більшою за обсягом.

Бізнес-аналітик часто стикається з ситуацією, коли йому відомо про існування у веб-просторі якогось документа, але він не може знайти його за допомогою традиційних пошуковиків, якими сьогодні можна вважати такі системи, як Google, Yahoo!, Bing, Baidu, Рамблер або Мета. Однак, згадавши або знайшовши в закладках адресу (URL) цього документа, він без труднощів виходить на нього. Тобто у веб-просторі цей документ є, а знайти його звичним способом не можна. Користувач зіткнувся з невидимим (invisible) для пошукових систем ресурсом.

Збір інформації із спеціальних баз даних дозволяє аналітикам та спеціалістам з кібербезпеки отримувати більше інформації про потенційні загрози та вразливості для покращення оборонної стратегії та реагування на інциденти.

Отримання доступу до цих баз даних у межах OSINT дозволяє аналітикам та дослідникам відстежувати і аналізувати події в кіберпросторі. Наведемо деякі приклади спеціальних баз даних:

- National Vulnerability Database (NVD)[58]: NVD є базою даних, яка містить інформацію про вразливості програмного забезпечення. Вона є частиною системи NIST (National Institute of Standards and Technology) у США. NVD надає дані про вразливості, їх описи, засоби експлуатації та інші технічні деталі (Рис. 54).

- Common Vulnerabilities and Exposures (CVE)59: CVE – це словник вразливостей, який надає унікальні ідентифікатори (CVE-ідентифікатори) для визначення конкретних вразливостей. Інформація відображається у вигляді загальноприйнятих термінів, що полегшує обмін інфор-

[58] National Vulnerability Database (NVD). URL: https://nvd.nist.gov/

[59] CVE security vulnerability database. URL: https://www.cvedetails.com/

мацією про вразливості.

- Exploit Database (ExploitDB)[60]: ExploitDB – це база даних, яка містить експлойти та код для використання вразливостей програмного забезпечення. Вона може бути використана для дослідження та розуміння того, як атакувати вразливості та як їх уникати.

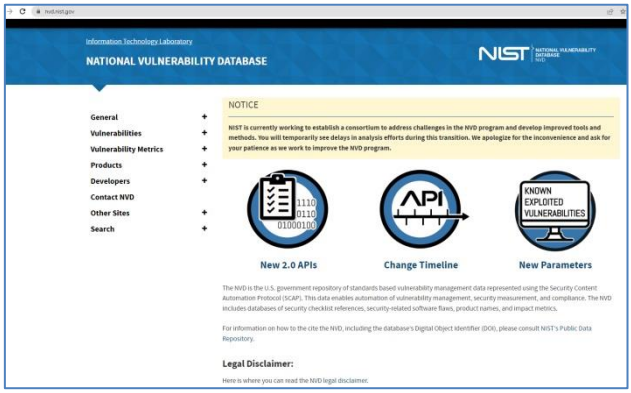


Рисунок 54 – Веб-ресурс системи NVD

- Shodan[61]: Shodan – це пошуковий двигун для Internet of Things (IoT) та підключених до Інтернету пристроїв. Він дозволяє аналізувати конкретні типи пристроїв та вразливості, пов'язані з ними, в реальному часі.

- National Institute of Standards and Technology Cybersecurity Framework[62]: NIST розробляє та публікує кілька стандартів та фреймворків з кібербезпеки, таких як NIST Cybersecurity Framework. Цей фреймворк містить рекомендації з кібербезпеки та включає в себе інформацію про кращі практики та стандарти.

- Dark Web Databases: В інформаційному просторі даркнету існують бази даних, які містять інформацію про зловмисні

[60] The Exploit Database. URL: https://www.exploit-db.com/
[61] Shodan. URL: https://www.shodan.io
[62] NIST Cybersecurity Framework. URL: https://www.nist.gov/cyberframework

дії, вартість хакерських інструментів, викрадені дані та інше. Зауважимо, доступ до даркнет-ресурсів потребує спеціальних заходів безпеки.

- Threat Intelligence Platforms (TIPs)[63]: Threat Intelligence Platforms надають можливість агрегувати та аналізувати різні джерела загроз, включаючи дані з різних спільнот з кібербезпеки, обігові та інші.

- Деякі OSINT-фреймворки (Open Source Intelligence Framework[64], Рис. 55), такі як Maltego (https://www.maltego.com/categories/osint/), Buscador OSINT Virtual Machine (https://inteltechniques.com/buscador/), або SpiderFoot (https://github.com/smicallef/spiderfoot), об'єднують різні джерела інформації, включаючи бази даних з кібербезпеки, для здійснення комплексного аналізу.

Технічні аспекти спеціальних баз даних як джерел інформації для OSINT включають в себе ряд особливих характеристик, форматів та інструментів, які забезпечують необхідну функціональність та зручність при роботі:

- Стандартизовані формати даних в спеціальних базах даних, що полегшує автоматизований аналіз і обробку. Наприклад, база CVE використовує формати JSON та XML для представлення даних щодо вразливостей.

- Деякі бази даних отримують дані через Threat Intelligence Feed[65], автоматизований механізм, який надає поточну інформацію про загрози в реальному часі.

[63] Top 7 Threat Intelligence Platforms & Tools for 2023. URL: https://www.esecurityplanet.com/products/threat-intelligence-platforms/

[64] OSINT framework. URL: https://osintframework.com/

[65] Daniella Balaban. 8 Best Threat Intelligence Feeds to Monitor in 2023 (https://cybeready.com/threat-intelligence/best-threat-intelligence-feeds-to-monitor-in-2023)

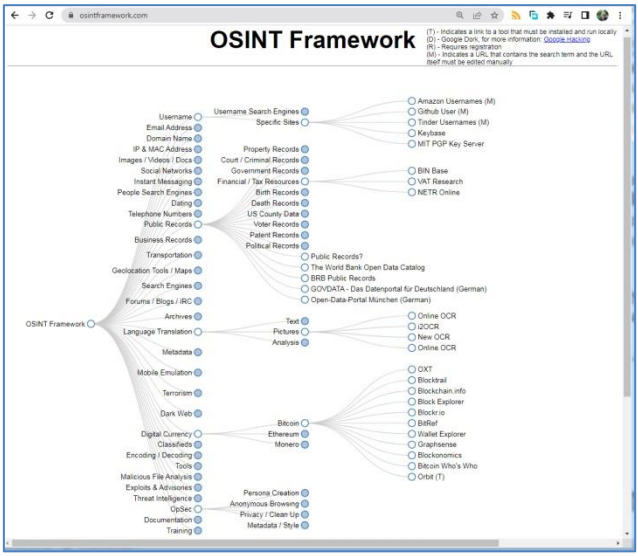


Рисунок 55 – Open Source Intelligence Framework (Фреймворк OSINT, зосереджений на зборі інформації з безкоштовних інструментів або ресурсів

Мета проекту полягає в тому, щоб допомогти людям знайти безкоштовні OSINT-ресурси)

- API для запитів і збору даних – деякі спеціальні бази даних пропонують API, використання яких дозволяє автоматизувати процеси збору та інтеграції інформації у системи безпеки.
- Використання класифікації для покращення доступу до даних дозволяє швидше фільтрувати та аналізувати інформацію.
- Взаємодія із наявними в корпорації інструментами безпеки, такими як SIEM (Security Information and Event Management), IDS (Intrusion Detection System), або іншими системами.
- Машинне навчання для аналізу даних: Застосування методів машинного навчання може допомогти впровадженню алгоритмів аналізу для автоматичного виявлення та класифікації загроз на основі накопичених даних.
- Обробка неперервних потоків даних: Оскільки інформація

в області кібербезпеки може надходити в режимі реального часу, спеціальні бази даних можуть використовувати системи обробки потоків даних для негайної аналізу та реакції на події.

- Розширені можливості пошуку та фільтрації: Враховуючи обсяг даних в кібербезпеці, бази даних часто мають розширені засоби пошуку та фільтрації для точного виділення необхідної інформації.
- Використання індексації даних: Індексація дозволяє швидко визначати та отримувати доступ до конкретних даних, що полегшує ефективний пошук та аналіз.

3.4.1. Глибинний веб

Сукупність джерел у веб-просторі, недоступних користувачам традиційних пошукових систем, утворює так званий «глибинний веб» – поняття, введене Джилл Еллсворт (Jill Ellsworth) у 1994 р. Тобто під глибинним веб (invisible web, deep web, hidden web) прийнято розуміти ту частину веб-простору, яка не індексується роботами (web crawlers) пошукових систем. Використовуючи аналогію, інформація, будучи недоступною для пошуку, знаходиться «в глибині» (англ. – deep). При цьому не варто плутати глибинний веб із ресурсами, взагалі недоступними з мережі Інтернет – це темний веб (dark web), і мова про нього тут не йтиме. Деякі ресурси, доступ до яких відкрито лише для зареєстрованих користувачів, також належать до глибинного веб.

У 2000 році американська компанія BrightPlanet (www.brightplanet.com) опублікувала сенсаційну доповідь, у якій стверджується, що у веб-просторі в сотні разів більше сторінок, ніж їх вдалося проіндексувати найпопулярнішим на той час пошуковим системам. Компанія розробила програму LexiBot, яка дозволяє сканувати деякі динамічні веб-сторінки, що формуються з баз даних, і, запустивши її, отримала несподівані дані. З'ясувалося, що в глибинному веб знаходиться у 500 разів більше документів, ніж доступно через пошукові системи. Звісно, ці цифри неточні. Крім того, стало відомо, що середня сторінка глибинного веб на 27 % компактніша за середню сторінку з видимої частини веб-простору.

Сьогодні ситуація змінилася, наприклад, провідні пошукові системи можуть індексувати документи, представлені у фор-

матах, що містять текст. Звісно, це, передусім, pdf, rtf та doc. У 2006 році Google запатентувала спосіб пошуку в глибинному веб: «Searching through content which is accessible through web-based forms» (Рис. 56). На думку різних авторів, до видимого веб належить лише 20–30 % веб-простору.

Рис. 56 – Фрагмент веб-ресурсу ВОІВ з описом патенту Google на пошук у глибинному інтернеті

У глибинному вебі знаходяться веб-ресурси, не пов'язані з іншими ресурсами гіперпосиланнями – наприклад, сторінки, що динамічно створюються за запитами до баз даних, документи з баз даних, доступні користувачам через пошукові веб-форми (але не за гіперпосиланнями). Такі документи залишаються недоступними для робота, нездатного в режимі реального часу правильно заповнити поля форми значеннями (формувати запити до баз даних).

Ось що говориться про глибинний веб у книзі[66]: «Більшість сторінок невидимого Інтернету можуть бути технічно проіндексовані, але не індексуються, тому що пошукові системи вирішили їх не індексувати... Більшість «невидимих» сайтів мають високоякісний контент. Просто ці ресурси не можуть

[66] Price G., Sherman C., Sullivan D. The Invisible Web: Uncovering Information Sources Search Engines Can't See. – Information Today, Inc., 2001. – 439 p.

бути знайдені за допомогою пошукових машин загального призначення...

... Деякі сайти використовують технологію баз даних, що дійсно складно для пошукової машини. Інші сайти, однак, використовують поєднання файлів, які містять текст і мультимедіа, а тому частина з них може бути проіндексована, а частина – ні.

... Деякі сайти можуть бути проіндексовані пошуковими машинами, але це не робиться тому, що пошукові машини вважають це непрактичним – наприклад, з причини вартості або тому, що дані настільки короткочасні, що індексувати їх просто безглуздо – наприклад, прогноз погоди, точний час прибуття конкретного літака, що здійснив посадку в аеропорту тощо».

Основні обмеження, пов'язані з роботами пошукових машин, можна пояснити такими основними причинами: для публічних пошукових служб важливіше забезпечити точність пошуку, ніж повноту, іноді важливіше забезпечити отримання відповіді на запит за прийнятний час, ніж точність. Звідси – обмеження на глибину сканування веб-ресурсів, спроби «фільтрації» контенту за змістом, відсіювання сторінок, що містять зайві вихідні гіперпосилання тощо. При цьому часто разом з водою вихлюпують і дитину.

Загальновизнано, що цінність ресурсів глибинного вебу іноді вища за цінність ресурсів видимої частини веб-простору.

Можна згадати ще одну причину поповнення глибинного вебу – власники свідомо не хочуть, щоб їхні веб-ресурси знаходили за допомогою пошукових систем. Найчастіше такі веб-ресурси являють собою щось не зовсім законне, хакерські форуми, архіви неавторизованого контенту тощо. Зрозуміло, що багато з таких ресурсів дуже цікаві для вивчення бізнес-аналітиками.

Багато компаній спочатку підключаються до загальної Мережі, і лише потім витрачають великі кошти на захист. Власники сайтів можуть спробувати заборонити індексацію тих чи інших сторінок своїх ресурсів, прописавши заборонну команду у файлі robots.txt, але пошукові системи можуть її проігнорувати. Тому такі ресурси або видаляють, або видаляють гіперпосилання, переводячи ресурси в глибинний веб. Наприклад, деякі бізнес-каталоги відмовляються віддавати свої оголошення ро-

ботам пошукових систем, тобто захищаючи свої інформаційні активи, компанії переводять свої ресурси в глибинний веб.

3.4.2. Ресурси глибинного веб

Існує кілька типів ресурсів глибинного вебу, наприклад, як було зазначено вище, це можуть бути веб-сторінки, що швидко застарівають. Крім того, до глибинного вебу належать веб-ресурси, що являють собою мультимедійну інформацію. Як відомо, на сьогодні ще не існує задовільних алгоритмів пошуку нетекстової інформації. Динамічно генеровані за запитом сторінки також часто потрапляють до глибинного вебу. Найчастіше без запиту таких сторінок не існує, вони генеруються при зверненні до баз даних. Виходить, що інформація начебто й присутня у веб-просторі, але виникає вона лише в момент обробки запиту, а універсального алгоритму заповнення ними пошукових форм роботами не існує. І, нарешті, якщо на веб-ресурс не ведуть жодні посилання, то роботи пошукових систем жодним чином не можуть дізнатися про його існування.

Засновник компанії BrightPlanet Майкл Бергман (Michael K. Bergman) зміг виділити 12 різновидів глибинних веб-ресурсів, що належать до класу онлайнових баз даних. До списку потрапили як традиційні бази даних (патенти, медицина та фінанси), так і публічні ресурси – оголошення про пошук роботи, чати, бібліотеки, довідники. Бергман зарахував до глибинних ресурсів і спеціалізовані пошукові системи, які обслуговують певні галузі чи ринки, бази даних яких не включаються до глобальних каталогів традиційних пошукових служб.

До глибинного вебу також належать численні системи інтерактивної взаємодії з користувачами – допомоги, консультування, навчання, що вимагають участі людей для формування динамічних відповідей від серверів. До них також можна віднести й закриту (повністю або частково) інформацію, доступну користувачам Мережі лише з певних адрес, груп адрес, іноді міст або країн. До «прихованої» частини Мережі багато хто зараховує й веб-сторінки, зареєстровані на безкоштовних серверах, які індексуються, в кращому випадку, лише частково – пошукові системи, щоб уникнути рекламного спаму, не прагнуть обходити їх у повному обсязі.

До глибинного вебу також належить категорія так званих «сірих» сайтів, що функціонують на основі динамічних систем

керування контентом (Dynamic Content Management Systems). У пошукових системах зазвичай обмежується глибина індексування таких сайтів, щоб уникнути можливого циклічного перегляду одних і тих самих сторінок.

Як же знайти веб-ресурси, розміщені в глибинному вебі? Якщо ресурси вимагають заповнення спеціальних форм, доповнених, наприклад, капчами, то необхідно вийти на базу даних, яка ймовірно містить потрібні документи. Знайти бази даних – джерела прихованого вебу – можна за допомогою звичайних пошукових систем, узагальнивши запит і ввівши уточнюючі слова, такі як «база даних», «банк даних», «database» тощо.

Наведемо загальновідомий приклад: користувачеві потрібна статистика катастроф літаків в Аргентині. Природний запит до традиційної пошукової системи видає величезний список газетних заголовків. За запитом «aviation database» можна одразу вийти на базу даних NTSB Aviation Accident Database (www.ntsb.gov/ntsb/query.asp).

Для пошуку в глибинному вебі, а саме в тому його сегменті, який складають бази даних, сьогодні вже існують деякі спеціалізовані ресурси. Лідером серед навігаторів у глибинному вебі є сайт CompletePlanet (www.completeplanet.com) компанії BrightPlanet. Цей сайт є найбільшим каталогом, що налічує понад 100 тисяч посилань. Компанія BrightPlanet також створила персональну утиліту для пошуку в онлайнових базах даних LexiBot, яка може забезпечувати пошук у кількох тисячах пошукових систем «глибинного» вебу. Метапошуковий пакет DeepQueryManager (DQM) цієї ж компанії забезпечує пошук більш ніж 70 тисячами «прихованих» веб-ресурсів.

Дослідження, проведене ще у 2006 р., показало, що глибинний веб охоплює понад 300 тис. сайтів, пов'язаних із понад 450 тис. баз даних, не охоплюваних традиційними пошуковими системами[67]. До найцікавіших для бізнес-аналітиків ресурсів глибинного вебу належать: бази даних юридичних та фізичних осіб; галузеві бази даних; репутаційні бази даних (чорні та білі списки); кримінологічні бази даних; бази даних то-

[67] He B., Patel M., Zhang Z., Chang K. C.-C. Accessing the Deep Web: A Survey. Communications of the ACM (CACM), 50(5):94-101, 2007.

варів та послуг; каталоги продукції тощо. До всесвітньо відомих бізнес-ресурсів, розміщених у глибинному вебі, належать: amazon.com, ebay.com, realtor.com, cars.com, imdb.com.

Наведемо ще кілька прикладів баз даних та каталогів глибинного вебу:

FindLaw (www.findlaw.com) – один із найпопулярніших у світі юридичних веб-сайтів – великий каталог правових ресурсів, що містить анотований список вільно доступних баз даних нормативно-правових документів, для яких даний ресурс є «точкою входу». Фрагмент веб-сайту сервісу FindLaw наведено на Рис. 57.

Рис. 57 – Фрагмент вебсайту сервису FindLaw

Politicalinformation.com (www.politicalinformation.com) – ресурс для журналістів, політиків, студентів та політичних діячів, сервіс, що забезпечує оперативний пошук на 5000 відібраних веб-сайтах політичного спрямування, надання новин із кількох десятків авторитетних джерел.

Академічна пошукова система Білефельда BASE (https://archive-it.org/) – одна з найбільших у світі пошукових систем академічних веб-ресурсів. Забезпечується доступ до

повних текстів близько 60% проіндексованих документів безкоштовно (відкритий доступ). BASE перебуває у віданні Бібліотеки Університету Білефельда.

CiteSeerX (https://citeseerx.ist.psu.edu/index) – це електронна бібліотека наукової літератури та пошукова система. Сервіс створений для поширення наукової літератури та покращення функціональності, зручності використання, доступності, вартості, повноти, ефективності та своєчасності доступу до наукових знань.

Data.gov (https://www.data.gov/) – «оселя даних». Згідно з умовами Федеральної політики відкритих даних 2013 р., новостворені урядові дані мають бути доступні у відкритих машиночитаних форматах, при цьому зберігається конфіденційність та безпека.

Mednar (https://mednar.com/mednar/desktop/en/search.html) – безкоштовна пошукова система в глибинній мережі, орієнтована на медицину. Оскільки Mednar є загальнодоступним пошуком, він не може отримати результати з особистої підписки або додаткових медичних ресурсів.

World Wide Science (https://worldwidescience.org/) – глобальний науковий портал, що складається з наукових баз даних та порталів.

Особливість більшості «прихованих» ресурсів полягає в їх вузькій спеціалізації. Для пошуку в них використовуються ті ж механізми, що й для «поверхневого» вебу, однак у більшості випадків роботи пошукових систем для глибинного вебу включають унікальні для кожного такого ресурсу модулі доступу до даних.

Традиційна пошукова система найчастіше може видати адресу бази даних, але не скаже, які документи конкретно містяться в ній. Типовий приклад – інформаційно-пошукові системи з українського (zakon.rada.gov.ua) або російського законодавства (www.kodeks.ru). Тисячі документів з баз даних стають доступними тільки після входу в систему, а роботи стандартних пошукових систем не в змозі проіндексувати контент баз даних.

Парадоксально, але як один із ресурсів глибинного вебу можна розглядати й архів ресурсів відкритого веб-простору. Такий архів – Internet Archive створюється з 1996 року

(www.archive.org). Сьогодні обсяг бази даних Alexa перевищує 538 млрд. веб-сторінок (Рис. 58), 28 млн. книжок та текстів, 14 млн. аудіозаписів, 6 млн. відео, 3,5 млн. зображень, 580 тисяч комп'ютерних програм.

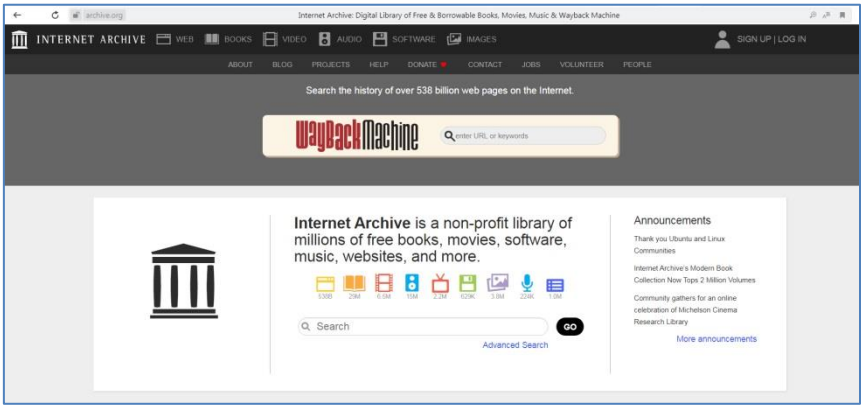


Рис. 58 – Стартова сторінка вебсайту www.archive.org

Технологія сховища Internet Archive включає низку сучасних засобів керування гігантським документальним сховищем. Наприклад, за допомогою цієї технології виконується кластеризація веб-ресурсів, тобто формування колекцій документів, близьких за тематикою. Особливий інтерес у користувачів сервісу Internet Archive викликає «Машина часу» (Wayback Machine), що відкриває доступ до часових зрізів веб-простору. Одне з найцікавіших практичних застосувань цієї технології – відновлення документів, колись опублікованих у веб-просторі, але згодом видалених. При цьому зростання глибинного вебу загрожує серйозними проблемами повноти у сховищі системи, пов'язаними зі збільшенням кількості сайтів, що експлуатують різні технології керування контентом, динамічної публікації документів із баз даних тощо.

3.4.3. Сервіси роботи з глибинним веб

Традиційні пошукові системи прагнуть звузити простір глибинного вебу, поступово захоплюючи такі ніші, як блоги, наукові сайти, інформаційні агентства. Так, як допоміжні сервіси для пошуку в глибинному вебі від Google можна рекомендувати: Google Book Search (books.google.com) – пошук книг, Google Scholar (scholar.google.com) – пошук наукових публікацій, Google Code Search (code.google.com) – пошук програмного коду.

Система Goldfire Research від компанії Invention Machine Corp. (inventionmachine.com) дає змогу обробляти контент глибинного вебу, розміщений на більш ніж 2000 сайтах урядових, академічних, дослідницьких та комерційних організацій США. Система Goldfire Research володіє інформацією про механізми доступу до баз даних глибинного вебу й автоматично генерує запити до них.

Наявні засоби аналізу та просування веб-ресурсів дають змогу по-новому підійти до оцінки співвідношення обсягів видимого та глибинного вебу. Так, на веб-сайті www.cypr.com наводиться інформація про реальну кількість документів на досліджуваному веб-сайті та про кількість документів, проіндексованих різними пошуковими системами, зокрема Google. Отримавши репрезентативну вибірку за сайтами, можна отримати оцінку співвідношення видимої та глибинної частини веб-простору.

Як показують розрахунки, обсяг інформації, що опинилася в глибинній частині веб-простору, перевищує обсяг інформації з видимої частини приблизно в 3-5 разів. Виявляється, за рідкісним винятком, що чим більший ресурс, тим більша його частина належить до глибинного вебу. У цьому сенсі невеликі веб-ресурси виграють у доступності. Оскільки велика частка новинних документів опиняється в глибинному вебі, то для завдань бізнес-аналітики потрібні спеціальні сервіси доступу до такої інформації. Саме такий сервіс надають служби інтеграції новинного контенту – архіви мережевих ЗМІ. Бізнес-аналітики активно використовують найбільші архіви інформації з відкритих джерел. Саме використання відкритих джерел дає змогу конкурентній розвідці діяти в межах правового поля, але при цьому мати високу ефективність.

Можна констатувати, що чим швидше зростає веб-простір, тим гірше він охоплюється традиційними каталогами та пошу-

ковими машинами. Через зростання кількості веб-сайтів і порталів, які використовують бази даних, динамічні системи керування контентом, появу нових версій форматів подання інформації, глибинний веб зростає дуже інтенсивно. З одного боку, Інтернет як величезне сховище збільшує обсяг інформації, доступної «в принципі», але з іншого боку – зростає інформаційний хаос, збільшується ентропія мережевого інформаційного простору. Дедалі менша частина інформаційних ресурсів стає реально доступною користувачам.

Провідні пошукові системи, як і раніше, намагаються знайти технічні можливості для індексації вмісту баз даних і доступу до закритих веб-сайтів, однак їхні завдання об'єктивно розходяться із завданнями бізнес-аналітиків – орієнтація традиційних пошукових служб на масовий сервіс у цьому випадку виправдана. Таким чином, ніша для систем пошуку в глибинному вебі стає дедалі ширшою.

3.4.4. Спеціальні бази даних

Як правило, для успішного ведення конкурентної розвідки має бути створений та підтримуватися банк даних, що включає такі основні бази даних:

- Конкуренти (діючі та потенційні);
- Інформація про ринок (тенденції, номенклатурна, цінова, адресна інформація);
- Технології (продукти, виставки, конференції, ГОСТи, якість);
- Ресурси (сировина, людські та інформаційні ресурси);
- Законодавство (міжнародні, центральні, регіональні та відомчі нормативно-правові акти);
- Загальні тенденції (політика, економіка, регіональні особливості, соціологія, демографія).

Сьогодні для конкурентної розвідки основними джерелами інформації слугують Інтернет, преса, а також відкриті бази даних. Але якщо доступ до звичайних інтернет-ресурсів можна вважати умовно безкоштовним, то в більшості випадків доступ до баз даних потребує не лише реєстрації, а й оплати таких послуг. Крім того, практично всі вони можуть бути віднесені до так званого «прихованого» веб-простору.

Дуже популярними серед фахівців з конкурентної розвідки є бази даних митних, податкових та статистичних органів, органів юстиції та судів, торгово-промислових палат, органів приватизації та фондових ринків, інформаційних, рейтингових, аналітичних та інших агентств тощо. Велику користь приносять і окремі доступні бази даних інших контролюючих органів та організацій.

Традиційно конкурентна розвідка спирається на такі джерела інформації, як опубліковані документи відкритого доступу, що містять огляди товарного ринку, інформацію про нові технології, створення партнерств, злиття та поглинання, оголошення про вакансії, про виставки та конференції тощо. Тому останнім часом дедалі популярнішими стають бази даних на основі архівів ЗМІ, у тому числі й мережевих.

До «Великої трійки» світових служб, що займаються наданням користувачам доступу до ділової та аналітичної інформації, входять LexisNexis, Factiva та Internet Securities. Найбільша у світі повнотекстова онлайнова інформаційна система LexisNexis (www.lexisnexis.com), яка містить понад 2 мільярди документів із 45 тисяч джерел з архівом глибиною понад 30 років з бізнес-інформації та понад 200 років з правової інформації, належить до розряду «прихованого» вебу (Рис. 59). Щотижня до архівів додається ще 14 млн документів. На відміну від неструктурованих масивів «поверхневого» вебу, користувачі LexisNexis можуть використовувати потужні інструменти пошуку для отримання достовірної та класифікованої інформації.

Служба Factiva (global.factiva.com), підрозділ компанії Dow Jones, займається наданням доступу до ділової та аналітичної інформації. В основі служби Factiva є понад 35 тис. первинних джерел зі 159 країн світу. У базі даних Factiva містяться матеріали більш ніж за 36,5 млн компаній, а також повна добірка інформації Investext.

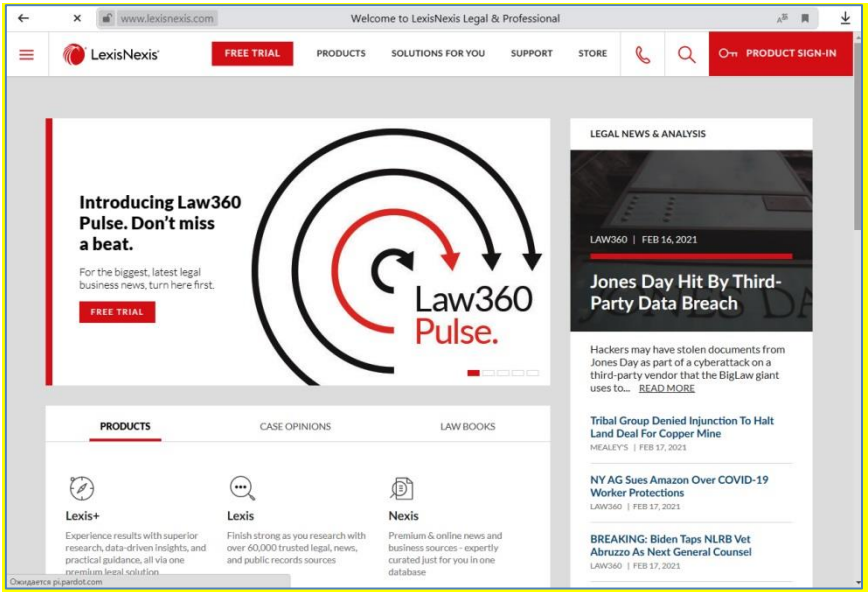


Рисунок 59 – Фрагмент веб-сайту служби LexisNexis

Компанія Internet Securities (www.internetsec.com), бренд ISI Emerging Markets, охоплює 80 тематичних інформаційних розділів, які формуються з 16 тис. джерел інформації – тексти статей, фінансові та аналітичні звіти, корпоративна інформація, макроекономічна статистика, дані по ринках (Рис. 60). Основні продукти ISI Emerging Markets: CEIC Data, Emerging Market Information Service (EMIS), Islamic Finance Information Service (IFIS), IntelliNews, ISI Compliance Edition, ISI DealWatch.

Europages (www.europages.eu) – Європейська бізнес-директорія – інформаційно-пошукова B2B-система, що охоплює понад 3 млн постачальників, виробників і дистриб'юторів у Європі та в усьому світі.

Завдання повного переліку всіх джерел інформації практично нездійсненне, оскільки цей ринок дуже динамічний, постійно з'являються нові бази даних, відбувається злиття наявних джерел, поглинання слабких сильними. Водночас одне з правил конкурентної розвідки формулюється так: «чим більшою кількістю незалежних джерел підтверджується інформація – тим вона достовірніша».

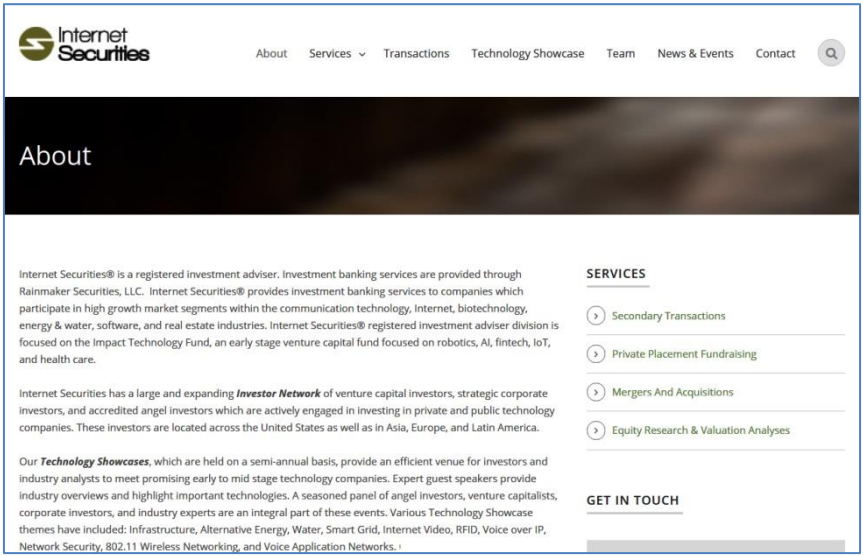


Рис. 60 – Фрагмент веб-сайту служби Internet Securities

Поряд із базами даних, одними з найефективніших джерел інформації можуть слугувати звіти та довідки аутсорсингових компаній, які професійно займаються конкурентною розвідкою та збором відомостей про комерційні структури та ринки. Їхня продукція, насправді, і є результатом конкурентної розвідки.

У світі існує безліч таких спеціальних компаній. Однією з таких найбільших компаній, якій належить близько 80 % західного ринку, є американська компанія Dun & Bradstreet (D&B), чию базу даних ми згадували вище. Довідка з будь-якої компанії в цій службі оцінюється з розрахунку в середньому 100 доларів і вище. Більш серйозний аналіз ринку чи конкурента може обійтися і в 10 тис. доларів. Терміни виконання – від кількох годин (інформація присутня в базі даних) – до кількох діб для довідок і до кількох місяців для аналітичної роботи.

На європейському ринку не менш відомі вищезгадана ірландська компанія Creditreform, німецька Schufa Holding AG (479 млн. документів у БД, зокрема, 66 млн. про фізичних осіб), австрійська Intercredit Infromation Holding, латвійська Coface IGK (відома IGK System – база даних боржників, що включає відомості про поточні борги, судові позови, а також процеси неплатоспроможності) та багато інших. Деякі з цих компаній

поєднують функції конкурентної розвідки з іншими видами діяльності, наприклад, обов'язками кредитних бюро.

Спільною проблемою під час звернення за інформаційними довідками до західних агентств, які мають представництва в Україні, є те, що, як правило, інформація, що надається щодо західних нерезидентів, набагато ширша та якісніша, ніж та, що надається щодо вітчизняних фірм. У зв'язку з чим, у таких випадках доцільно звертатися до місцевих інформаційних компаній, результати виявляються дешевшими та якіснішими.

В Україні існує ціла низка подібних компаній, серед яких можна назвати «Авеста-Україна», «Консалтингова компанія «СІДКОН», Міжбанківська служба безпеки «СКІФ» та багато інших.

Усі вітчизняні та зарубіжні інформаційні компанії мають свої представництва та приймають замовлення в Інтернеті, у зв'язку з чим їх можна віднести до специфічних інтернет-джерел.

Слід також зазначити, що, незважаючи на те, що в разі замовлення послуг аутсорсингової компанії вона виконує більшу частину інформаційної роботи за клієнта, остаточні висновки та рішення, рекомендації для прийняття управлінських рішень все ж таки залишаються за ним. Тільки клієнт може володіти всією необхідною повнотою зовнішньої та інсайдерської інформації.

3.5. Геопросторові джерела в OSINT

Картографічні ресурси можуть бути важливим джерелом інформації в області кібербезпеки, дозволяючи аналітикам та спеціалістам з кібербезпеки аналізувати географічні аспекти кіберзагроз та інцидентів, отримувати комплексні дані про географічні аспекти кіберзаходів, що дозволяє реагувати на загрози та виявляти слабкі місця в інфраструктурі. Такі ресурси можуть включати глобальні мапи, геодані з різних джерел, інструменти аналізу мереж тощо. Використання картографічних ресурсів у сфері кібербезпеки дозволяє:

- Проводити моніторинг географічних зон загроз, аналізувати географічний розподіл кіберзагроз та інцидентів, вказувати на конкретні географічні області, які піддаються атакам чи виявляють підвищені ризики.
- Використання картографічних інструментів дозволяє

візуалізувати географічні дані, такі як геолокаційні дані інцидентів, IP-адрес, атак чи аномалій мережевого трафіку.

- Аналіз Географічного Розташування IP-адрес: Деякі інструменти дозволяють виконувати аналіз географічного розташування IP-адрес, що може бути корисним для виявлення аномалій чи підозрілих активностей.
- Розуміти контекст кіберзагроз та інцидентів завдяки використанню геоданих, таких як інформація про об'єкти і спеціальні шари (наприклад, мережі, провайдери, географічні особливості).
- Виявляти певні закономірності та географічні залежності, що допомагають у формуванні стратегій кібербезпеки.
- Створювати карти для візуалізації інфраструктури та потенційних місць атак, включаючи розташування серверів, дата-центрів та інших ключових точок.
- Використовувати картографічні ресурси для отримання інтелектуальної інформації, зокрема, визначення стратегічних об'єктів, областей ризику та шляхів збройних конфліктів.
- Застосування карт для візуалізації загроз та ризиків може полегшити розуміння складних кіберзаходів та обов'язкових заходів безпеки.

Технічні аспекти роботи з картографічними ресурсами в якості джерел інформації в рамках OSINT можуть включати в себе різні технології та методи для збору, аналізу та використання географічних даних, які дозволяють ефективно використовувати картографічні ресурси для аналізу та моніторингу кіберзагроз у географічному контексті, полегшуючи виявлення, реагування та подальший аналіз:

1. Геодані та геоінформаційні системи (ГІС):
   - Використання стандартів геоданих, таких як Keyhole Markup Language (KML)[68] або GeoJSON (https://geojson.io/), для обміну та інтеграції гео-

[68] Keyhole Markup Language. URL: https://docs.fileformat.com/uk/gis/kml/

графічних даних.

- Використання ГІС для обробки та аналізу географічних інформацій, виявлення залежностей та відображення результатів на карті.

2. IP-Геолокація:

   - Використання сервісів геолокації для визначення фізичного розташування IP-адрес, що може бути важливим для виявлення географічно зосереджених загроз чи атак.

3. Карти загроз:

   - Створення та підтримка карт загроз, які дозволяють візуалізувати різноманітні кіберзагрози та інциденти на карті.

4. Використання API для мап:

   - Використання API картографічних сервісів (наприклад, Google Maps API[69], Mapbox API70) для інтеграції географічних даних у власні системи безпеки.

5. Техніка класифікації та кластеризації:

   - Застосування алгоритмів машинного навчання для класифікації та кластеризації географічних даних, що дозволяє виявляти закономірності та патерни.

6. Аналіз географічних зв'язків:

   - Використання інструментів для аналізу географічних зв'язків між різними об'єктами та подіями у кіберпросторі.

7. Збір та аналіз географічних аспектів соціальних мереж:

   - Використання інструментів для збору та аналізу гео-

---

[69] Geocoding Service | Maps JavaScript API. URL: https://developers.google.com/maps/documentation/javascript/geocoding

[70] Mapbox web services APIs. URL: https://docs.mapbox.com/api/overview/

графічних даних з соціальних мереж, що може допомагати в розумінні мережевих взаємозв'язків.

8. Застосування технік анонімізації та псевдонімізації:

   - Забезпечення захисту конфіденційності географічних даних, використовуючи техніки анонімізації чи псевдонімізації (управління даними та деідентифікації) при обробці та передачі інформації.

9. Моніторинг географічних змін:

   - Визначення та моніторинг змін у географічному контексті, таких як нові центри загроз або зміна географічного розташування інфраструктури.

У сучасному аналізі відкритих даних простір перестав бути лише метафорою – він став конкретним, вимірюваним, семантично насиченим виміром реальності. Будь-яка подія, незалежно від її природи: кібератака, соціальний протест, правова колізія, інформаційна кампанія, має географічну локацію або, принаймні, геопросторовий контекст. Це робить геопросторові дані не просто допоміжним шаром, а фундаментальним компонентом OSINT-аналітики, який дозволяє прив'язувати абстрактні факти до фізичної реальності, реконструювати логістику, виявляти інфраструктурні залежності та моделювати просторові взаємодії.

Серед усіх геопросторових джерел особливе місце займає OpenStreetMap (OSM) – колективний, відкритий, глобальний проект, який надає детальну картографічну інформацію про міста, дороги, будівлі, критичну інфраструктуру, військові об'єкти та багато іншого. На відміну від комерційних картографічних сервісів, OSM є повністю відкритим, має добре документовану модель даних (ноди, шляхи, відносини, теги) і підтримує потужні інструменти запиту (Overpass API), що робить його ідеальним джерелом для автоматизованого аналізу.

Проте OSM – лише частина ширшого екосистеми відкритих геопросторових ресурсів, до якої також входять Wikimapia, Natural Earth, NASA Worldview, Sentinel Hub, а також публічні шари Google Maps та Mapbox. Кожен із цих ресурсів має свою специфіку: одні надають векторні дані про інфраструктуру, інші – супутникові знімки, треті – геополітичні кордони чи де-

мографічні показники. Разом вони утворюють багатошарову геопросторову онтологію, яка дозволяє аналітику не просто визначити координати події, а зрозуміти її місце в системі фізичного, соціального та інфраструктурного середовища.

Цей підрозділ присвячений системному опису цих джерел, методам їхнього видобування, інтеграції в семантичні мережі та практичному застосуванню в задачах кібербезпеки, соціального моніторингу та стратегічного аналізу. Особлива увага приділяється технічним аспектам роботи з OSM через Python (osmnx, pyrosm), побудові геомереж, візуалізації через Neo4j та Gephi, а також етичним викликам, пов'язаним із використанням геоданих. У підсумку, геопросторові джерела перестають бути просто «картою», вони стають динамічною моделлю реальності, де кожен об'єкт має не лише координати, а й семантичні зв'язки, функціональне призначення та стратегічне значення.

3.5.1. Роль геопросторових даних у сучасному OSINT

У сучасному аналізі відкритих джерел геопросторові дані перестали бути допоміжним шаром інформації й перетворилися на фундаментальний контекст, що надає просторову прив'язку до будь-якої події. Якщо текст описує що сталося, а метадані – коли, то геолокація відповідає на питання де. Це триєдинство (що–коли–де) утворює просторово-часову онтологію події, яка є необхідною умовою для її достовірної реконструкції. У світі, де інформація часто фрагментарна, або навмисно спотворена, геопросторові дані надають об'єктивну опорну систему, оскільки координати, на відміну від тексту, важко сфабрикувати без технічних артефактів.

Геолокація в OSINT виконує подвійну роль: вона є одночасно атомарним фактом (наприклад, GPS-координати з EXIF-даного фото) і стратегічним індикатором. На мікрорівні координати дозволяють верифікувати достовірність зображення або відео: чи відповідає фон знімку реальному ландшафту у вказаній точці? На макрорівні геодані дозволяють аналізувати просторові патерни: концентрацію подій навколо критичної інфраструктури, переміщення військових колон, еволюцію міського середовища під час конфлікту.

Математично кожну подію є можна представити як елемент простору $E = T \times G \times S$, де T – це часовий компонент, $G \subset R^2$ – геопросторовий компонент (широта, довгота), S – семантичний зміст. Саме наявність G дозволяє застосовувати методи просторового аналізу: кластеризацію за відстанню (наприклад, DBSCAN), побудову теплових карт, аналіз близькості до стратегічних об'єктів. Без цього компонента подія залишається «плаваючим» фактом, який важко інтегрувати в загальну картину.

Геопросторові дані в OSINT моделюються в рамках векторної геометрії, де кожен об'єкт представляється як один із трьох базових типів:

Точка ($p \in R^2$) – це дискретна локація, наприклад, місце вибуху, база даних сервера, місце публікації в соцмережі;

Лінія ($l = \{p_1, p_2, ..., p_n\}$ – це послідовність точок, що утворюють маршрут, наприклад, траєкторія БПЛА, лінія фронту, маршрути доставки;

Полігон ($P \subset R^2$) – замкнена область, що визначає адміністративні кордони, зони відповідальності, межі міста, військові об'єкти.

Кожен з цих геометричних примітивів супроводжується атрибутами – семантичними властивостями, такими як назва, тип, статус, дата створення. Наприклад, полігон може мати атрибути {type: "military_base", name: "Object X", status: "active"}. Ця модель, відома як модель векторних даних, лежить в основі OpenStreetMap, GeoJSON, Shapefile та інших стандартів. Вона дозволяє не лише зберігати форму об'єкта, а й виконувати топологічні операції: перетин, об'єднання, відстань між об'єктами, що є критичним для аналізу зв'язків.

Відкриті геопросторові дані, зокрема OpenStreetMap (OSM), мають дві ключові переваги: безкоштовність та відкритість ліцензії (ODbL), що дозволяє вільно використовувати, модифікувати та поширювати дані. Це робить їх незамінними для державних установ, наукових досліджень та некомерційних OSINT-ініціатив. Крім того, OSM часто містить деталізацію, якої немає в комерційних картах: внутрішні дороги, промислові об'єкти, військові споруди – завдяки роботі місцевих маперів.

Проте ця ж відкритість є джерелом системних обмежень. По-перше, якість даних нерівномірна: у містах вони можуть бути надзвичайно деталізованими, тоді як у віддалених регіонах – відсутніми. По-друге, відсутність централізованого контролю робить OSM вразливим до вандалізму, помилок, навіть свідомої дезінформації (наприклад, додавання неіснуючих об'єктів). По-третє, відсутність історичних шарів у стандартному API ускладнює аналіз змін у часі (хоча це частково компенсується через Overpass API та OSMCha).

Навпаки, комерційні джерела (Google Maps, HERE, Maxar) забезпечують високу точність, регулярне оновлення, історичні знімки, але за умови платної ліцензії та обмежень на використання. У багатьох випадках, особливо в умовах конфлікту, доступ до таких даних може бути обмежений або цензурований.

Таким чином, відкриті геодані є найкращим компромісом між доступністю та корисністю. Вони не завжди надають максимальну точність, але забезпечують незалежність, прозорість і громадську валідацію – цінності, які є фундаментальними для етичної OSINT-аналітики. У поєднанні з іншими джерелами (соцмережі, супутникові знімки, технічні реєстри) вони утворюють геопросторовий каркас, на якому будується вся решта аналітичної моделі.

3.5.2. OpenStreetMap як ключове джерело OSINT

Серед усіх відкритих геопросторових джерел OpenStreetMap (OSM) займає особливе місце завдяки своїй повній відкритості, спільнотному характеру та глибокій семантичній структурі. На відміну від комерційних картографічних сервісів, OSM надає не лише візуальну карту, а й повний граф інфраструктурних об'єктів, який може бути використаний для реконструкції фізичного середовища, аналізу логістики, моніторингу змін у міському просторі та верифікації геолокаційних даних з інших джерел. У контексті OSINT OSM перетворюється з пасивного фону на активний компонент семантичної мережі, де кожен об'єкт – будівля, дорога, електропідстанція – стає вузлом із властивостями, який може бути пов'язаний з подіями, акторами, технічними артефактами.

Архітектура OSM базується на трьох основних примітивах, які утворюють планарний граф $G=(V,E,R)$, де V – множина нод (точок), E – множина шляхів (ліній або полігонів), R – множина відносин (складних структур).

Нода (node) є найпростішим елементом – це точка з географічними координатами(lat,lon). Вона може представляти окремий об'єкт (наприклад, поштову скриньку, дерево) або бути частиною більш складної структури.

Шлях (way) – це упорядкований список нод, який утворює лінію (дорога, річка) або замкнений полігон (будівля, парк). Якщо перша та остання ноди співпадають, шлях інтерпретується як полігон.

Відношення (relation) – це набір об'єктів (нод, шляхів, інших відносин), пов'язаних між собою через ролі. Наприклад, маршрут автобуса може бути представлений як відношення, де шляхи – це ділянки маршруту, а ноди – зупинки, кожна з яких має роль stop.

Кожен із цих примітивів може бути доповнений тегами – парами «ключ–значення» (key=value), які надають семантичний зміст. Наприклад, тег building=hospital вказує, що полігон є лікарнею, а power=substation – що нода або полігон є електропідстанцією. Ця система тегів є відкритою онтологією, яка постійно розширюється спільнотою.

Дані OSM зберігаються в двох основних форматах: XML (офіційний формат експорту) та JSON (частіше використовується в API). У XML-моделі кожен примітив має унікальний ідентифікатор (id), версію (version), час останньої зміни (timestamp), ідентифікатор автора (uid), а також список тегів у вигляді елементів <tag k="..." v="..."/>.

Наприклад, фрагмент XML для електропідстанції може виглядати так:

```
<node id="123456789" lat="50.4501" lon="30.5234">
  <tag k="power" v="substation"/>
  <tag k="name" v="Підстанція №5"/>
</node>
```

Семантика тегів визначається спільнотними угодами, зафіксованими у вікі OSM. Ключі організовані ієрархічно: power=*, military=*, landuse=*, building=* тощо. Значення можуть бути контрольованими (наприклад, power=substation, power=plant) або вільними (наприклад, name=...). Ця гнучкість

дозволяє моделювати надзвичайно різноманітні об'єкти, від військових баз до велосипедних доріжок.

З точки зору OSINT, найціннішими є теги, що описують критичну інфраструктуру:

power=substation, power=plant, power=line – енергетика;

military=airfield, military=barracks, military=naval_base – військові об'єкти;

highway=motorway, railway=station – транспорт;

amenity=hospital, emergency=fire_station – соціальна інфраструктура.

Ці теги дозволяють автоматично видобувати об'єкти певного типу для подальшого аналізу.

Для практичного використання OSM-даних існують кілька платформ доступу, кожна з яких орієнтована на певний сценарій.

Overpass API є найпотужнішим інструментом для вибіркового запиту. Він використовує спеціалізовану мову запитів Overpass QL, яка дозволяє фільтрувати дані за тегами, геозонами, часом оновлення. Наприклад, запит:

```
[out:json];
area["name"="Київ"]->.searchArea;
(
  node["power"="substation"](area.searchArea);
  way["power"="line"](area.searchArea);
);
out body;
>;
out skel qt;
```

повертає всі підстанції та лінії електропередач у Києві. Overpass API ідеально підходить для OSINT-завдань, де потрібно отримати лише релевантні об'єкти, а не весь регіон.

OSMnx – це бібліотека Python, призначена для аналізу транспортних мереж, але також підтримує загальний доступ до OSM. Вона автоматично завантажує дані через Overpass API, перетворює їх на граф NetworkX або GeoDataFrame, що дозволяє проводити топологічний аналіз, розраховувати найкоротші шляхи, виявляти критичні вузли.

Geofabrik та BBBike надають повні експорти регіонів у форматі .osm.pbf – стиснутому двійковому форматі, оптимізо-

ваному для зберігання великих обсягів даних. Ці файли використовуються, коли потрібен повний дубль OSM для певної території (наприклад, вся Україна), що корисно для масштабованих проектів або офлайн-аналізу.

Разом ці інструменти утворюють повноцінний стек для роботи з OSM у OSINT: від точкового запиту через Overpass API до масового аналізу через OSMnx і Geofabrik. Саме завдяки цій екосистемі OpenStreetMap перестає бути просто картою й перетворюється на динамічну, семантично збагачену модель реального світу, придатну для інтеграції в граф знань, побудови сценаріїв, прогнозування наслідків подій.

3.5.3. Методи екстрагування об'єктів із OSM

Ефективне використання OpenStreetMap у рамках OSINT-аналітики передбачає не просто візуальне спостереження, а систематичне видобування структурованих об'єктів із загального масиву геопросторових даних. Цей процес, відомий як екстрагування, є перехідною ланкою між сирим OSM-датасетом і аналітично придатною формою – будь то таблиця, граф або JSON-об'єкт. Він базується на трьох взаємопов'язаних етапах: формулюванні семантично точного запиту, автоматизованому завантаженні даних та трансформації у стандартні формати, придатні для подальшої обробки в Big Data-інфраструктурі.

Основою цілеспрямованого екстрагування є Overpass QL – декларативна мова запитів, розроблена спеціально для фільтрації даних OpenStreetMap. На відміну від SQL, Overpass QL оперує поняттями геопросторової топології та семантичних тегів. Запит формується як послідовність операторів, кожен з яких визначає підмножину примітивів $P \subseteq G$, де $G = (V, E, R)$ – повний граф OSM.

Ключовим механізмом є фільтрація за тегами. Наприклад, вираз виділяє всі ноди, що мають тег power=substation. Аналогічно,

way["building"="hospital"];

повертає всі полігони, що представляють лікарні.

Для обмеження простору пошуку використовуються геозони. Найчастіше – через визначення адміністративної одиниці:

area["name"="Львів"]["admin_level"="8"]->.city;

(

```
  node["military"="barracks"](area.city);
  way["landuse"="industrial"](area.city);
);
```

Цей запит видобуває військові казарми та промзони лише в межах міста Львів. Альтернативно можна задати bounding box у форматі (south, west, north, east), що корисно для нетипових регіонів.

Синтаксис Overpass QL також підтримує рекурсивні операції: >, <, >>, << – для отримання всіх залежних примітивів (наприклад, нод, що складають шлях). Це критично важливо для повного опису складних об'єктів, таких як транспортні мережі чи військові бази.

Після формулювання запиту наступним етапом є його автоматизоване виконання. У середовищі Python це реалізується через кілька підходів. Найпростіший – це використання бібліотеки requests для прямого HTTP-запиту до Overpass API:

```
import requests
que-
ry = "[out:json]; node['power'='substation'](50.4,30.5,50.5,30.6); ou
t;"
response = requests.post("https://overpass-
api.de/api/interpreter", data=query)
data = response.json()
```

Цей підхід дає повний контроль над запитом, але вимагає ручної обробки відповіді.

Більш високорівневим інструментом є OSMnx – спеціалізована бібліотека для роботи з OSM. Вона абстрагує деталі Overpass QL, надаючи інтуїтивні функції:

```
import osmnx as ox
gdf = ox.geometries_from_place("Київ, Україна", tags={"power"
: "substation"})
```

Результат повертається у форматі GeoDataFrame – розширення pandas DataFrame, що містить геометрію кожного об'єкта. Це дозволяє відразу застосовувати просторові операції: буферизацію, перетин, відстань.

Для складних pipeline'ів, де потрібна інтеграція з іншими системами (наприклад, Elasticsearch або Neo4j), часто використовується комбінація requests для завантаження та geopandas для нормалізації геометрії, що забезпечує гнучкість і масштабованість.

Останнім, але ключовим етапом є трансформація екстрагованих даних у формати, сумісні з аналітичними інструментами.

Найпростішим виходом є CSV – таблиця, де кожен рядок відповідає одному об'єкту, а стовпці містять його властивості: id, lat, lon, name, type, tags. Такий формат ідеально підходить для імпорту в MongoDB, Elasticsearch або для первинного аналізу в pandas.

Для побудови семантичних мереж дані перетворюються у графові формати, зокрема GEXF (Graph Exchange XML Format). У цьому випадку кожен OSM-об'єкт стає вузлом, а зв'язки (наприклад, «розташований у», «належить до інфраструктури») – ребрами. Це дозволяє імпортувати дані в Gephi або Neo4j для подальшого аналізу центральності, кластеризації, шляхів.

Нарешті, JSON (частіше GeoJSON) є універсальним форматом для веб-додатків та API. Він зберігає як геометрію (Point, Polygon), так і властивості (properties), що робить його ідеальним для інтеграції з картографічними бібліотеками (Leaflet, MapLibre) або для передачі в LLM-системи, які аналізують просторові зв'язки.

Таким чином, методи екстрагування з OSM утворюють замкнутий аналітичний цикл: від семантичного запиту – через автоматизоване завантаження – до структурованого виводу. Саме цей цикл перетворює OpenStreetMap із пасивної карти на активний компонент інтелектуальної інфраструктури OSINT, де кожен геооб'єкт стає частиною більшої моделі знань.

3.5.4. Інші відкриті геопросторові ресурси

Хоча OpenStreetMap є центральним джерелом відкритих геопросторових даних, повна картина фізичного середовища вимагає інтеграції додаткових ресурсів, кожен з яких надає унікальний шар інформації: від топоніміки та адміністративних меж до супутникових знімків та комерційних картографічних платформ з відкритим доступом. Ці джерела не замінюють OSM, а доповнюють його, формуючи багатошарову геопросторову модель реальності, де кожен шар має свою семантику, точність, частоту оновлення та обмеження. У цьому підрозділі розглядаються три ключові категорії: енциклопедичні та базові географічні бази, супутникові платформи та картографічні сервіси з відкритим API.

Wikimapia є колективним проєктом, що поєднує елементи вікіпедії та картографії. На відміну від OSM, де акцент робиться на топологічній точності, Wikimapia фокусується на описовій семантиці: кожен полігон супроводжується коротким текстом, історичною довідкою, фотографіями. Це робить його цінним джерелом для контекстуалізації об'єктів – наприклад, для розуміння, чи є певна будівля колишньою військовою частиною або промисловим підприємством. Дані доступні через API та експорт у форматі KML, що дозволяє інтегрувати їх у аналітичні системи.

GeoNames – це глобальна база географічних назв, що містить понад 11 мільйонів топонімів із координатами, ієрархією (країна → регіон → місто), кодами країн (ISO 3166), часовими зонами. Кожен запис має унікальний ідентифікатор geonameId, що дозволяє однозначно пов'язувати назви з іншими джерелами. GeoNames особливо корисний для геолокації текстових згадувань: коли в новині згадується «Львів», система може автоматично зіставити це з geonameId=702508 та отримати точні координати, адміністративну приналежність, населення. Це критично важливо для побудови часових рядів за регіонами.

Natural Earth – це набір векторних та растрових шарів, призначених для картографії в малих масштабах. Він містить адміністративні межі, річки, дороги, рельєф, кліматичні зони – усе це узгоджено за рівнями деталізації (1:10m, 1:50m, 1:110m). Natural Earth не призначений для локального аналізу, але є незамінним для стратегічного OSINT: візуалізації подій на рівні країн, регіонів, континентів. Дані надаються у форматі Shapefile та легко імпортуються в geopandas, QGIS, або Gephi.

Супутникові дані надають об'єктивний, безпосередній погляд на фізичну реальність, незалежно від того, як вона описується в текстах. Серед відкритих платформ ключове місце займає Sentinel Hub – інтерфейс до архіву місій Copernicus (Sentinel-1, Sentinel-2), що надає мультиспектральні знімки з роздільною здатністю до 10 м. Через API можна отримувати не лише RGB-зображення, а й індекси (NDVI, NDWI), що дозволяє аналізувати стан рослинності, наявність води, зміни в ландшафті. Це особливо корисно для моніторингу знищення інфраструктури, переміщення техніки, змін у міському середовищі.

NASA Worldview надає доступ до десятків шарів даних з супутників NASA (MODIS, VIIRS, Landsat) у реальному часі. Він орієнтований на глобальні явища: пожежі, урагани, аерозольні шари, температура поверхні. Хоча просторова роздільна здатність нижча (від 250 м до 1 км), часове покриття дуже густе (до кількох разів на день), що дозволяє виявляти динаміку подій – наприклад, поширення задимлення після вибуху.

Google Earth Engine (GEE), хоча й є хмарною платформою, надає відкритий доступ до петабайтів супутникових даних (Landsat, Sentinel, MODIS) через JavaScript/Python API. Його потужність полягає в можливості виконувати складні обчислення на стороні сервера: класифікація земного покриву, виявлення змін, часові серії. Для OSINT це означає можливість аналізувати зміни на території протягом років без завантаження великих обсягів даних.

Комерційні картографічні платформи, такі як Mapbox, HERE та Thunderforest, надають відкриті шари (часто на умовах attribution) для некомерційного використання. Ці шари часто мають вищу візуальну якість, кращу деталізацію міських територій, актуальніші дорожні мережі порівняно з OSM.

Mapbox дозволяє використовувати базові шари (Streets, Satellite) через Tile API, а також запитувати геокодування та маршрутизацію. HERE надає доступ до базових карт через XYZ-тайли, а також до спеціалізованих шарів (трафік, паркування). Thunderforest – це провайдер стильованих OSM-шарів (Transport, Landscape, Outdoors), які корисні для тематичної візуалізації.

Хоча ці сервіси не надають прямого доступу до векторних даних, їхні тайли можуть бути використані як фон для верифікації OSM-об'єктів або для візуалізації результатів аналізу. У деяких випадках (наприклад, при відсутності даних у OSM для певного регіону) вони стають єдиним джерелом актуальної інформації.

Разом усі ці ресурси утворюють розширений геопросторовий ландшафт, де OSM виступає як основа, а інші джерела надають контекст, візуальну верифікацію, часову динаміку та глобальну перспективу. Саме через їхню інтеграцію OSINT-аналітик отримує можливість не просто локалізувати подію, а реконструювати її фізичний контекст з високою точністю та достовірністю.

3.5.5. Побудова мереж на основі геопросторових даних

Геопросторові дані, хоча й надають точну локацію об'єктів, набувають аналітичної цінності лише тоді, коли вони інтегровані в структурну модель взаємодій. Саме тому сучасний OSINT переходить від пасивного картографування до активного моделювання геопросторових мереж – графів, у яких вузли представляють фізичні або інституційні сутності, а ребра відображають просторові, функціональні або інфраструктурні зв'язки між ними. Такий підхід перетворює карту зі статичного зображення на динамічну систему знань, придатну для аналізу вразливостей, прогнозування наслідків подій та виявлення прихованих залежностей.

Першим етапом побудови мережі є семантична ідентифікація сутностей серед загального масиву геопросторових даних. У OpenStreetMap це досягається через фільтрацію за тегами, які визначають онтологічний тип об'єкта. Наприклад, атомна електростанція може бути ідентифікована як:

$v \in V$: tag(v)={power=plant,plant:source=nuclear}

Міст – як шлях із тегом bridge=yes та highway=*; військова база – як полігон із military=barracks або landuse=military; адміністративний кордон – як relation із boundary=administrative та admin_level=4 (область) або 8 (місто).

Кожен такий об'єкт перетворюється на вузол графа $v_i \in V$, який має не лише координати(xi,yi), а й властивості: тип, назва, статус, час останньої зміни. Ця операція є критичною, оскільки саме від точності ідентифікації залежить достовірність всієї подальшої мережі. Помилкова класифікація (наприклад, інтерпретація складу як військової бази) може призвести до хибних висновків про загрози.

Другим етапом є визначення ребер $e_{ij} \in E$, які відображають зв'язки між вузлами. На відміну від соціальних мереж, де зв'язки часто явно задані (наприклад, дружба), у геопросторових мережах ребра є похідними від геометричних або семантичних відношень.

Найпростіший тип – просторова близькість: два об'єкти вважаються пов'язаними, якщо відстань між ними менша за поріг $d_{\max}$:

$$e_{ij} \in E \Leftrightarrow \left\| p_i - p_j \right\|^2 < d_{\max}.$$

Це корисно, наприклад, для виявлення критичної інфраструктури, розташованої поблизу потенційної цілі.

Більш складний тип – функціональна залежність. Наприклад, електропідстанція функціонально залежить від електростанції, якщо вона входить до її мережі передачі. Такі зв'язки можуть бути встановлені через аналіз тегів (power=line, cables=*) або через топологічний аналіз OSM-графа.

Нарешті, інфраструктурна суміжність виникає, коли об'єкти є частиною однієї системи: дорога, що з'єднує два міста; залізнична колія, що проходить через станції; трубопровід, що з'єднує родовище з переробним заводом. У таких випадках ребро формується не за відстанню, а за топологічною зв'язністю у графі OSM.

Ці три типи зв'язків разом утворюють багатошарову мережу, де кожен шар відповідає певному типу взаємодії. Така модель дозволяє не просто бачити об'єкти, а розуміти системну архітектуру території.

Остаточна сила геопросторової мережі розкривається при її інтеграції з текстовими даними. Коли в новині або Telegram-каналі згадується «атака на енергосистему Києва», система може автоматично зв'язати цей текст із вузлом OSM, що представляє найближчу підстанцію або електростанцію. Це досягається через геолокаційну верифікацію:

Видобуття топоніму («Київ») з тексту;

Зіставлення з geonameId або OSM-полігоном міста;

Пошук критичної інфраструктури в межах цього полігону;

Створення семантичного зв'язку: (текст) –[відноситься до]→ (OSM-об'єкт).

Такий зв'язок перетворює абстрактне повідомлення на геопросторово прив'язану подію, яка може бути включена в загальну семантичну мережу. Наприклад, якщо декілька незалежних джерел згадують одну й ту саму підстанцію, це підвищує достовірність інформації. Навпаки, якщо згадується об'єкт, якого немає в OSM, це може бути сигналом про фейк або про те, що об'єкт ще не задокументований.

У результаті геопросторова мережа стає мостом між цифровим і фізичним світом: кожен текстовий сигнал отримує локацію, кожен фізичний об'єкт – контекст. Саме ця інтеграція робить OSINT не просто інструментом спостереження, а платформою для реконструкції реальності.

3.5.6. Візуалізація та аналіз геомереж

Геопросторові дані, отримані з OpenStreetMap та інших відкритих джерел, набувають аналітичної цінності лише тоді, коли вони інтегровані в семантичну мережу, де кожен об'єкт стає частиною складної системи зв'язків. Цей перехід від ізольованої точки на карті до вузла в графі знань вимагає не лише технічного імпорту, а й методологічно обґрунтованого моделювання, подальшого топологічного аналізу та інтерактивної візуалізації. У цьому підрозділі розглядаються три ключові етапи: імпорт у графову базу даних, структурний аналіз мережі та створення інтерактивних семантичних карт.

Імпорт у Neo4j

Перший крок у побудові геомережі – це її формальне представлення у вигляді властивісного графа $G = (V, E)$, де кожен вузол $v_i \in V$ відповідає геооб'єкту (будівлі, підстанції, дорозі), а ребра $e_{ij} \in E$ відображають семантичні або просторові зв'язки. У системі Neo4j це реалізується через присвоєння кожному вузлу мітки (наприклад, :PowerSubstation, :MilitaryBase) та набору властивостей:

id – унікальний ідентифікатор OSM;

lat, lon – географічні координати;

name – назва об'єкта;

tags – JSON-рядок з усіма тегами OSM ("power": "substation", "operator": "Ukrenergo").

Ребра можуть мати типи, що відображають природу зв'язку: :LOCATED_IN (об'єкт у місті), :PART_OF (підстанція належить енергосистемі), :NEAR (просторова близькість до іншого об'єкта). Така модель дозволяє не просто зберігати геодані, а інтегрувати їх у загальну онтологію знань, де геооб'єкти пов'язані з особами, подіями, технічними артефактами. Наприклад, запит Cypher:

```
MATCH (s:PowerSubstation)-[:TARGETED_BY]->(a:Attack)
```

WHERE s.name CONTAINS "Київ"
RETURN s.name, a.date

дозволяє виявити всі атаки на київські підстанції, демонструючи силу семантичного моделювання.

Аналіз у Gephi

Після імпорту графа в Neo4j його часто експортують у Gephi – спеціалізовану платформу для візуального аналізу мереж. Тут застосовуються алгоритми теорії графів для виявлення прихованих структур.

Кластеризація (наприклад, алгоритм Louvain) розбиває граф на спільноти – групи вузлів, що мають високу внутрішню зв'язність. У контексті геомережі це може виявити концентрації критичної інфраструктури (наприклад, промзони, військові об'єкти), що формують зони підвищеної вразливості.

Центральність (Betweenness, Closeness, Eigenvector) визначає вузли, що відіграють ключову роль у передачі інформації або ресурсів. Вузол із високою Betweenness Centralities може бути стратегічною точкою, наприклад, електропідстанція, через яку проходить основний потік енергії.

Модулярність $Q \in [0,1]$ кількісно оцінює якість розбиття на спільноти: чим ближче Q до 1, тим чіткіша структура. Це дозволяє порівнювати різні сценарії розвитку подій, наприклад, як змінюється структура інфраструктурної мережі після втрати окремих вузлів.

Такий аналіз перетворює геомережу з пасивної карти на динамічну модель системи, де можна оцінювати стійкість, виявляти слабкі ланки, прогнозувати наслідки подій.

Візуалізація через CSV2Graph

Для оперативної верифікації результатів використовується CSV2Graph – це інструмент, що перетворює CSV-файли з вузлами та ребрами на інтерактивні HTML-графи. Ключовою перевагою є автоматичне додавання гіперпосилань на пошукові системи (Google, Bing) для кожного вузла. Наприклад, клік на вузол «Підстанція №5» відкриває нове вікно з запитом "Підстанція №5 Київ site:*.gov.ua", що дозволяє миттєво перевірити інформацію в офіційних джерелах.

Цей підхід забезпечує замкнутий цикл аналізу: від видобування даних з OSM – через побудову мережі – до верифікації

через зовнішні джерела. Інтерактивність дозволяє аналітику не просто бачити зв'язки, а досліджувати їхню достовірність, що критично важливо в умовах дезінформації.

Таким чином, візуалізація та аналіз геомереж є не завершальним етапом, а інтерактивним інструментом пізнання, де граф стає живим середовищем для формулювання гіпотез, тестування сценаріїв та прийняття рішень. Саме ця інтеграція геопростору, семантики та інтерактивності визначає майбутнє OSINT-аналітики.

3.5.7. Практичні кейси використання OSM у OSINT

OpenStreetMap, як джерело структурованих геопросторових даних, набуває особливої цінності в умовах кризи, конфлікту або швидких соціальних змін, коли офіційна інформація запізнюється, цензурується або навмисно спотворюється. У таких ситуаціях OSM перетворюється не просто на карту, а на динамічну модель фізичної реальності, яка дозволяє реконструювати події, виявляти загрози та прогнозувати наслідки. У цьому підрозділі розглядаються три типові сценарії застосування: моніторинг критичної інфраструктури, геолокаційна верифікація подій та виявлення змін у міському середовищі.

Моніторинг критичної інфраструктури під час конфліктів

Під час збройних конфліктів або гібридних атак критична інфраструктура – енергетична, транспортна, комунікаційна – стає первинною ціллю. OSM дозволяє систематично ідентифікувати такі об'єкти через семантичні теги:

- power=substation, power=plant, power=line – енергетика;
- military=airfield, military=barracks, military=naval_base – військові об'єкти;
- highway=motorway, railway=station, aeroway=aerodrome – транспорт.

Використовуючи Overpass API, можна автоматично видобути всі такі об'єкти в заданій геозоні (наприклад, області України) і побудувати мережу вразливості $V=(O,C)$, де O – множина об'єктів, а $C \subseteq O \times O$ – зв'язки близькості або функціо-

нальної залежності. Наприклад, якщо дві підстанції розташовані на відстані менше 5 км, вони можуть бути уражені одним і тим самим ударом. Аналіз центральності (Betweenness, Eigenvector) у такій мережі дозволяє виявити ключові вузли, чия втрата призведе до каскадного колапсу. Ця модель не лише фіксує стан інфраструктури, а й прогнозує наслідки пошкоджень, що є критичним для планування оборони та відновлення.

Реконструкція подій на основі геозаявок із Telegram/X

Часто перші повідомлення про події (вибухи, обстріли, протести) з'являються в соціальних мережах із геолокацією або описом місця. Проте такі заявки часто містять неточності, дезінформацію або навмисні фейки. OSM виступає як референсний простір для верифікації.

Наприклад, повідомлення в Telegram: «Обстріл біля школи №12 у Харкові» – може бути автоматично оброблене так:

1. Видобуття топоніму «Харків» та об'єкта «школа №12»;
2. Запит до OSM:
   node["amenity"="school"]["name"~"12"](area.Harkiv);
3. Отримання координат реальної школи №12;
4. Порівняння з геолокацією зображення або відео (через EXIF або візуальну геолокацію).

Якщо координати збігаються з точністю до 100 м, заявка вважається підтвердженою. Якщо ж у радіусі 1 км немає школи №12 – це сигнал до поглибленої перевірки. Такий підхід перетворює OSM на онтологічний фільтр, що дозволяє відокремити достовірні сигнали від шуму в умовах інформаційної війни.

Виявлення змін у міському середовищі

OSM є живим документом, який постійно оновлюється спільнотою. Це дозволяє використовувати його не лише як статичну карту, а й як сенсор змін. Через Overpass API можна отримувати дані з міткою часу (timestamp), що дозволяє будувати часові ряди змін.

Наприклад, запит:

```
 [diff:"2025-01-01T00:00:00Z","2025-03-01T00:00:00Z"];
(
  node["military"="checkpoint"](area.Kyiv);
```

```
  way["barrier"="block"](area.Kyiv);
);
out meta;
```

повертає всі нові блокпости та бар'єри, додані в Києві за останні два місяці. Аналогічно можна виявляти:

знесення будівель (building → видалено);

зміну функціонального призначення (landuse=residential → landuse=industrial);

появу тимчасових об'єктів (emergency=field_hospital).

Такий аналіз дозволяє відстежувати еволюцію міського простору в реальному часі, що є незамінним для моніторингу воєнізування територій, реакції на надзвичайні події або соціальних змін. Більше того, порівняння з історичними супутниковими знімками (наприклад, з Sentinel Hub) дозволяє верифікувати сам факт зміни, усуваючи можливість помилок редакторів OSM.

Разом ці кейси демонструють, що OpenStreetMap у контексті OSINT є не просто картографічним інструментом, а динамічною, семантично збагаченою моделлю фізичного світу, яка забезпечує основу для верифікації, прогнозування та стратегічного планування в умовах невизначеності.

3.5.8. Етичні та технічні виклики

Незважаючи на значні переваги OpenStreetMap як джерела відкритих геопросторових даних, його застосування в OSINT-аналітиці супроводжується низкою фундаментальних обмежень, що походять як із технічної природи колективного редагування, так і з етичних наслідків публікації детальної інформації про фізичне середовище. Ці обмеження формують системну невизначеність, яка впливає на достовірність, повноту та безпечність аналітичних висновків. У цьому підрозділі розглядаються три ключові аспекти: залежність від людського фактора, проблеми верифікації та ризики порушення конфіденційності.

OpenStreetMap є продуктом колективного інтелекту, де кожен елемент карти додається, редагується або виправляється окремим користувачем – так званим мапером. Ця модель забезпечує високу адаптивність, але водночас призводить до нерівномірності покриття. У містах з активною спільнотою

(наприклад, Київ, Львів) дані можуть бути надзвичайно деталізованими: кожна будівля, тротуар, лавка має свій полігон і семантичні теги. Навпаки, у віддалених регіонах або зонах конфлікту покриття може бути фрагментарним або повністю відсутнім.

З точки зору теорії інформації, це означає, що ентропія просторового розподілу даних $H(S)$ є нерівномірною: в одних регіонах $H(S) \to 0$ (висока структурованість), в інших $H(S) \to \log N$ (хаос, відсутність інформації). Така асиметрія ускладнює побудову глобальних моделей, оскільки аналітик не може припускати, що відсутність об'єкта в OSM означає його фізичну відсутність. Навпаки, відсутність може бути наслідком інформаційної сліпоти – відсутності локального мапера. Це створює системну упередженість, яка особливо небезпечна в умовах кризи, коли найбільш вразливі території часто є найменш картографованими.

Проблеми верифікації

Відкритість OSM, хоча й є її головною перевагою, одночасно є джерелом вразливості до маніпуляцій. Будь-який зареєстрований користувач може додати, змінити або видалити об'єкт, що відкриває можливість для вандалізму, дезінформації або просто непрофесійних помилок. Відомі випадки, коли в OSM додавалися неіснуючі дороги, фіктивні військові бази, або навіть «пастки» у вигляді неіснуючих будівель, щоб зловити комерційні сервіси, що копіюють дані.

Хоча спільнота OSM має механізми модерації (патрулювання змін, історія редагувань, інструменти типу OSMCha), вони не гарантують повної безпеки. З точки зору теорії надійності, кожен примітив $p \in G$ має ймовірність помилки $\varepsilon(p)$, яка залежить від досвіду автора, складності об'єкта та рівня модерації регіону. У відсутності зовнішньої верифікації (наприклад, через супутникові знімки або офіційні реєстри) ця ймовірність залишається невизначеною. Тому будь-який OSINT-аналіз, що ґрунтується виключно на OSM, повинен включати процедуру крос-перевірки з незалежними джерелами, щоб уникнути поширення хибної інформації.

Конфіденційність

Найглибшим етичним викликом є конфлікт між відкритістю та приватністю. OSM містить детальну інформацію про фізичні об'єкти, включаючи приватні будинки, внутрішні подвір'я, особисті садиби. Хоча спільнота OSM дотримується правила «не картографувати те, що не видно з публічного простору», на практиці це правило часто порушується – навмисно або через незнання.

У контексті OSINT це створює серйозні ризики:

- розкриття місця проживання особи, якщо її будинок детально зображено в OSM;
- ідентифікація критичної інфраструктури приватного характеру (наприклад, резервні генератори, сховища);
- використання даних у шкідливих цілях – від крадіжок до цілеспрямованих атак.

З правової точки зору, такі дані можуть підпадати під регулювання GDPR або національного законодавства про захист персональних даних, оскільки координати будинку є прямою ідентифікуючою інформацією. Тому аналітик, який використовує OSM у професійних цілях, повинен дотримуватися принципу мінімізації даних: використовувати лише ту інформацію, що є необхідною для завдання, і уникати публікації деталей, що можуть загрожувати безпеці осіб чи об'єктів.

Разом усі ці виклики вказують на те, що OpenStreetMap, хоча й є могутнім інструментом, не є абсолютно надійним дзеркалом реальності. Його ефективне застосування в OSINT вимагає не лише технічної майстерності, а й глибокого методологічного рефлексу, етичної відповідальності та системного підходу до верифікації. Лише за таких умов OSM може стати не джерелом ризиків, а основою для побудови достовірної, відповідальної та ефективної аналітики.

Висновки до розділу 3

У третьому розділі монографії здійснено комплексний аналіз сукупності джерел інформації, що становлять фундаментальну основу для проведення комп'ютерної конкурентної розвідки та розвідки у відкритих джерелах. Доведено, що сучасне інформаційне середовище характеризується значною

гетерогенністю та динамізмом, де традиційні вебсайти виступають лише початковим шаром доступу до даних, тоді як основний масив цінної інформації часто міститься у соціальних мережах, блогах та глибинному вебпросторі. Особливу увагу приділено трансформації рольового значення геопросторових даних, які з допоміжного інструменту перетворилися на ключовий компонент OSINT-аналітики, дозволяючи прив'язувати абстрактні події до фізичної реальності через такі ресурси, як OpenStreetMap та супутникові знімки, що забезпечує реконструкцію логістики та виявлення інфраструктурних залежностей.

Важливим висновком є теза про необхідність використання спеціалізованих інструментів моніторингу та аналізу для ефективного опрацювання великих масивів неструктурованих даних, оскільки традиційні пошукові системи не забезпечують достатньої повноти та глибини проникнення у спеціалізовані бази даних і приховані ресурси мережі Інтернет. Розглянуто специфіку роботи з глибинним вебом, де зосереджено значний обсяг потенційно важливої бізнес-інформації, доступ до якої вимагає застосування спеціальних технік та сервісів, що дозволяють долати обмеження стандартних індексаторів. Окремо наголошено на правових та етичних аспектах використання відкритих джерел, зокрема щодо захисту персональних даних, дотримання авторських прав та законодавчих норм, що є критичною умовою легітимності діяльності служб конкурентної розвідки та уникнення криміналізації їхніх процесів.

Таким чином, успішність конкурентної розвідки залежить не стільки від обсягу зібраної інформації, скільки від здатності інтегрувати різнорідні джерела у єдину аналітичну модель, забезпечуючи верифікацію даних та їх своєчасну обробку. Сформована класифікація джерел та методів роботи з ними створює методологічну базу для побудови ефективних інформаційно-аналітичних систем, здатних трансформувати сирі дані з відкритих мережевих ресурсів у структуровані знання для підтримки управлінських рішень.

4. Семантичний нетворкінг: теорія, моделі, підходи до побудови знань

У сучасному інформаційному середовищі аналітик, що працює з відкритими джерелами, опиняється перед фундаментальною епістемологічною дилемою: як перетворити надлишок неструктурованих даних на цілісну, логічно узгоджену модель реальності? Традиційні підходи –пошук по ключовим словам, частотний аналіз, класична NER-екстракція – вичерпали свій потенціал у світі, де події часто фіксуються фрагментарно, з затримкою або навмисно спотворено. Відповідь на цей виклик лежить у семантичному нетворкінгу – новій парадигмі аналітики, що поєднує силу великих мовних моделей (LLM), принципи семантичного вебу та методи графового моделювання для побудови динамічних мереж знань.

На відміну від класичних семантичних мереж, які будувалися вручну або на основі жорстких онтологій, семантичний нетворкінг спирається на автоматизоване, ітеративне, LLM-кероване видобування знань. Його суть полягає в тому, що кожний документ – будь то повідомлення в Telegram, законодавчий акт або наукова публікація – перетворюється не на набір слів, а на набір семантичних трійок (суб'єкт – предикат – об'єкт), які потім об'єднуються в єдину мережу. Ця мережа не є статичною картою; вона живе, розширюється, адаптується, оскільки нові дані автоматично інтегруються, а існуючі зв'язки перевіряються через зовнішні джерела.

Центральним інноваційним елементом цього підходу є концепція «рою віртуальних експертів» (Swarm of Virtual Experts, SVE), де один і той самий текст аналізується кількома LLM, кожна з яких діє в рамках окремої професійної ролі – кібераналітика, юриста, соціолога, психолога масової комунікації. Такий колективний інтелект не лише підвищує точність видобування, а й забезпечує багатовимірну інтерпретацію, що наближається до людського розуміння складних подій.

Ще однією ключовою особливістю семантичного нетворкінгу є відмова від хронологічного принципу на користь логічного. У реальності події часто пов'язані не часом, а причинно-наслідковими або асоціативними зв'язками. Семантична мережа дозволяє реконструювати такі події навіть тоді, коли часові мітки відсутні, суперечливі або навмисно сфальсифіко-

вані. Це робить підхід особливо цінним у умовах гібридних конфліктів, де інформаційна війна супроводжується системним знищенням часових орієнтирів.

Цей розділ присвячений системному викладу теоретичних основ, математичних моделей та практичних методів семантичного нетворкінгу. У ньому розглядаються формальні моделі семантичних мереж, механізми екстрагування знань за допомогою LLM, підходи до верифікації зв'язків через зовнішні пошукові системи, методи побудови каузальних та асоціативних мереж, а також інтеграція результатів у графові бази даних (Neo4j) та інструменти візуалізації (Gephi, CSV2Graph). Особлива увага приділяється авторській методиці, розробленій у рамках досліджень, – зокрема, моделі «мережі переважного семантичного приєднання», де нові поняття приєднуються не до випадкових, а до семантично найближчих вузлів, що забезпечує органічний ріст мережі знань.

Семантичний нетворкінг не є просто технічним інструментом – він є новою формою пізнання, яка дозволяє не просто збирати факти, а будувати картографію знань, де кожен елемент має своє місце, функцію та зв'язки. Саме ця здатність до синтезу, верифікації та прогнозування робить його центральним компонентом сучасної OSINT-аналітики.

4.1. Теоретичні основи семантичних мереж

Семантична мережа як форма представлення знань має глибокі коріння в когнітивній науці, лінгвістиці та штучному інтелекті, але лише в епоху великих даних і генеративного ШІ вона перетворилася з теоретичної абстракції на практичний інструмент аналізу. У своїй суті семантична мережа є графовою моделлю знань, де кожен факт або поняття представляється як структурований елемент, а зв'язки між ними – як відношення, що несуть семантичне навантаження. Ця модель дозволяє не просто зберігати інформацію, а моделювати логіку реальності через топологію зв'язків.

4.1.1. Історія та еволюція концепції семантичних мереж

Перші ідеї семантичних мереж виникли ще в 1950–1960-х роках у рамках когнітивної психології. Алан Коллінз і М. Росс

Квілліан запропонували модель семантичної пам'яті, де знання організовані у вигляді ієрархічної мережі понять, пов'язаних відношеннями типу «є різновидом» (is-a) або «має властивість» (has-property). Ця модель пояснювала, чому людина швидше розпізнає, що «канарка – це птах», ніж що «канарка – це істота».

У 1970–1980-х роках семантичні мережі стали основою систем штучного інтелекту на основі знань (knowledge-based systems), зокрема експертних систем. Тут мережа використовувалася як онтологія предметної області – формальна специфікація класів, властивостей, обмежень. Проте такі системи були статичними, ручно побудованими, масштабно обмеженими.

Прорив стався з появою Семантичного Вебу (Semantic Web) наприкінці 1990-х – ініціативи Тіма Бернерса-Лі, спрямованої на структурування всесвітньої павутини через стандарти RDF (Resource Description Framework), RDFS, OWL. RDF формалізував знання як трійки (суб'єкт, предикат, об'єкт), що стало основою для сучасних графів знань.

Сьогодні, у епоху LLM, семантичні мережі еволюціонували далі: вони більше не будуються вручну, а автоматично екстрагуються з неструктурованих текстів, ітеративно розширюються, верифікуються через зовнішні джерела, і інтегруються в динамічні аналітичні системи. Це перетворило їх із пасивного сховища знань на активний інтелектуальний інструмент.

4.1.2. Граф як модель знань

Формально семантична мережа є орієнтованим, мультиграфом з властивостями:

$$G = (V, E, \varphi, \varphi)$$

де V – множина вузлів (сутностей), $E \subseteq V \times V$ – множина ребер (зв'язків), $\varphi : V \to C$ – функція, що присвоює кожному вузлу клас (наприклад, Person, Organization, Event), а $\psi : E \to R$ – функція, що присвоює кожному ребру тип відношення (наприклад, employs, located_in, causes).

Кожен вузол $v_i \in V$ може мати властивості – пари «атрибут–значення» (наприклад, name: "Іван", date: "2025-03-15"). Кожне

ребро $e_{ij} \in E$ також може мати властивості, зокрема вагу $w(e_{ij}) \in [0,1]$, що відображає достовірність або силу зв'язку.

Типи зв'язків можуть бути онтологічно визначеними (наприклад, згідно з schema.org або MITRE ATT&CK) або екстрагованими автоматично (наприклад, «використовує», «планує», «реагує на»). Саме ця гнучкість дозволяє моделювати складні реальні процеси – від кібератак до правових колізій.

4.1.3. Класифікація семантичних мереж

Залежно від природи зв'язків, семантичні мережі можна класифікувати на кілька типів.

Неорієнтовані мережі використовуються, коли зв'язок симетричний (наприклад, «співпрацює з»). Однак у більшості випадків зв'язки є асиметричними, тому переважають спрямовані мережі (directed graphs), де ребро $v_i \to v_j$ не еквівалентне $v_j \to v_i$. Наприклад, «APT29 атакує банк X» не те саме, що «банк X атакує APT29».

Зважені мережі додають до кожного ребра числову характеристику $w(e)$, яка може відображати ймовірність, частоту, довіру або семантичну близькість. Це критично важливо в умовах, коли зв'язки отримані з різних джерел із різним рівнем надійності.

Бімодальні мережі (двочасткові графи) містять два типи вузлів, наприклад, особи та події, і ребра існують лише між різними типами. Така модель корисна для аналізу участі акторів у подіях без прямої взаємодії між ними.

Нарешті, гіперграфи узагальнюють поняття ребра: замість зв'язку між двома вузлами, гіперребро може з'єднувати будь-яку кількість вузлів. Це дозволяє моделювати n-арні відношення, наприклад: «група A використала інструмент B проти цілі C у час D». Такі структури точніше відображають складні події, але вимагають спеціалізованих інструментів для обробки.

Разом усі ці моделі утворюють спектр формальних засобів, які дозволяють адекватно представляти знання про реальний світ – від простих асоціацій до складних каузальних механізмів. Саме ця теоретична основа забезпечує наукову строгость сучасного семантичного нетворкінгу.

4.2. Моделі семантичного нетворкінгу

Семантичний нетворкінг як методологія побудови знань не обмежується єдиною архітектурою графа. Навпаки, він передбачає множинність моделей, кожна з яких відображає певний аспект реальності: асоціативні зв'язки між поняттями, причинно-наслідкові механізми подій, формальні онтологічні структури або динаміку змін у часі. Ці моделі не є взаємовиключними; навпаки, вони доповнюють одна одну, формуючи багатошарову семантичну мережу, здатну до глибокого, багатовимірного аналізу. У цьому підрозділі розглядаються чотири ключові класи моделей: асоціативні, каузальні, онтологічні та динамічні.

4.2.1. Асоціативні мережі

Асоціативна мережа є найбільш інтуїтивною формою семантичного нетворкінгу. Вона базується на принципі контекстної коокурентності: два поняття вважаються пов'язаними, якщо вони часто з'являються в одному й тому самому текстовому контексті. Формально, нехай $D = \{d_1, d_2, ..., d_n\}$ – множина документів, а $T(d_i)$ – множина термів у документі d_i. Тоді ступінь асоціативності між термами ta та tb може бути визначена через коефіцієнт Жаккара:

$$J(t_a, t_b) = \frac{\left|\left\{d_i \in D : t_a \in T(d_i) \wedge t_b \in T(d_i)\right\}\right|}{\left|\left\{d_i \in D : t_a \in T(d_i) \vee t_b \in T(d_i)\right\}\right|}.$$

Якщо $J(t_a, t_b) > \theta$, де θ – порогове значення, між ta та tb створюється ребро. Така модель особливо ефективна для виявлення тематичних кластерів, наративів або інформаційних бульбашок у соціальних мережах. Наприклад, постійне співзгадування «енергетика», «хаос», «держава» у Telegram-каналах може сигналізувати про сформований дезінформаційний наратив.

4.2.2. Причинно-наслідкові (каузальні) мережі: моделювання логіки подій

На відміну від асоціативних мереж, каузальні мережі фокусуються на логічних залежностях між подіями, а не на їхній

співпояві. У такій моделі ребро eij означає, що подія vi є причиною події vj. Каузальність не випливає автоматично з тексту, її потрібно інферувати за допомогою LLM, яка аналізує семантику речень на наявність слів-маркерів причинності («через», «внаслідок», «спровокував» тощо).

Формально, каузальна мережа є орієнтованим ациклічним графом $(G_C = V, E_c)$, де V – це множина подій, а $E_c \subseteq V \times V$ – множина каузальних зв'язків. Ключовою перевагою цієї моделі є можливість реконструкції подій без хронології: навіть якщо часові мітки відсутні, логіка подій залишається збереженою. Наприклад, мережа може містити шлях: («група APT28» → «використала Zloader» → «атака на банк X»), що дозволяє зробити висновок про причетність групи, навіть якщо дата атаки невідома.

4.2.3. Онтологічні мережі: формалізація предметної області

Онтологічна мережа вводить формальну структуру у семантичний простір, визначаючи класи, властивості, обмеження та ієрархії. Вона базується на стандартах семантичного вебу (RDF, OWL), де кожен факт подається як трійка (s, p, o), але з додатковими аксіомами. Наприклад, у домені кібербезпеки можна визначити:

- клас Attack як підклас Event;
- властивість hasActor з доменом Attack і рангом Actor;
- обмеження: кожна Attack має принаймні одного hasActor.

Така модель забезпечує логічну цілісність мережі: система може виявляти суперечності (наприклад, атака без виконавця) або виводити нові знання (наприклад, якщо APT28 є Actor, а Zloader – Tool, то (APT28 uses Zloader) є валідним зв'язком). Онтологічні мережі особливо корисні для юридичної аналітики, де точність формулювань є критичною.

4.2.4. Динамічні та адаптивні мережі

Нарешті, сучасний OSINT вимагає живих, а не статичних моделей. Динамічна семантична мережа $G(t) = (V(t), E(t))$ є функцією часу, де множини вузлів і ребер змінюються з над-

ходженням нових даних. Оновлення може відбуватися за двома сценаріями:

- інкрементне додавання: нові трійки, видобуті з останніх документів, інтегруються в існуючу мережу;
- ретроспективна корекція: при виявленні помилки (наприклад, через верифікацію) попередні зв'язки видаляються або модифікуються.

Для підтримки динаміки використовується механізм довіри: кожен зв'язок e_{ij} має вагу $w_{ij} \in [0,1]$, яка залежить від джерела, кількості підтверджень, результатів верифікації. При надходженні нових даних ваги перераховуються, що дозволяє мережі адаптуватися до змін у реальності. Наприклад, якщо спочатку група X була пов'язана з інструментом Y, але нові дані показують, що це була помилка, вага зв'язку знижується, і він може бути видалений при досягненні порога.

Разом усі ці моделі утворюють універсальний каркас семантичного нетворкінгу, де аналітик може вибирати або комбінувати підходи залежно від завдання: асоціативні – для виявлення тем, каузальні – для реконструкції подій, онтологічні – для формальної верифікації, динамічні – для прогнозування. Саме ця гнучкість робить семантичний нетворкінг потужним інструментом сучасної OSINT-аналітики.

4.3. Підходи до автоматизованої побудови семантичних мереж

Автоматизована побудова семантичних мереж є центральним етапом семантичного нетворкінгу, де неструктуровані тексти перетворюються на структуровані знання. Цей процес вимагає не просто застосування великих мовних моделей, а системного підходу, що поєднує точну промпт-інженерію, множинні інтерпретації, механізми верифікації та статистичну агрегацію. У цьому підрозділі розглядаються чотири ключові компоненти: екстрагування сутностей, конструювання промптів, концепція «рою віртуальних експертів» та агрегація результатів.

4.3.1. Екстрагування сутностей та зв'язків за допомогою LLM

На відміну від класичних NER-систем, які обмежені фіксованими класами (PERSON, ORG, LOC), сучасні LLM дозволяють видобувати довільні сутності та зв'язки без попередньої онтології. Процес формується як відображення

$$f : T \to K,$$

де T – множина текстових документів, а K – множина семантичних трійок (s, p, o), що складають базу знань. Ключовою перевагою LLM є здатність розуміти контекст: одне й те саме слово може бути інтерпретоване як організація, інструмент або подія залежно від оточення. Наприклад, у реченні «Sandworm використала Zloader проти банку» модель правильно ідентифікує Sandworm як групу, Zloader – як шкідливе ПЗ, а банк – як ціль. Це забезпечує глибину аналізу, недосяжну для правилових систем.

4.3.2. Промпт-інженерія для видобутку знань

Ефективність LLM критично залежить від якості промпта. У рамках семантичного нетворкінгу використовуються формалізовані шаблони, які визначають:

- роль моделі («Ти – кібераналітик»);
- предметну область («Аналізуй кіберзагрози»);
- типи сутностей та зв'язків («Видобий акторів, інструменти, цілі»);
- формат виводу («Поверни у форматі CSV: суб'єкт;предикат;об'єкт»);
- обмеження («Не вигадуй. Якщо немає зв'язку – поверни порожній рядок»).

Такий підхід перетворює LLM з «чорної скриньки» на контрольований інструмент видобування знань. Валідація вбудовується прямо в промпт: модель зобов'язана дотримуватися формату, що спрощує подальшу обробку. Наприклад, промпт може включати інструкцію: «Якщо ти не впевнений у зв'язку, пропусти його. Не використовуй припущення.» Це значно зменшує кількість галюцинацій на етапі первинного екстрагування.

4.3.3. Концепція «рою віртуальних експертів»: множинні перспективи аналізу

Навіть найкращий промпт не гарантує повної достовірності, оскільки LLM схильна до внутрішніх упереджень та помилок. Для подолання цього обмеження запропоновано метод «рою віртуальних експертів» (SVE). Суть його полягає в тому, що один і той самий текст аналізується кількома незалежними запитами, кожен з яких задає LLM окрему професійну роль:

- «Ти – юрист, що аналізує правові колізії»;
- «Ти – кібераналітик, що виявляє тактики MITRE ATT&CK»;
- «Ти – соціолог, що досліджує наративи дезінформації».

Кожен «експерт» формує власну інтерпретацію, яка фіксується у структурованому форматі. Це імітує роботу колегії фахівців, де кожен оцінює подію через призму своєї компетенції. Такий підхід не лише підвищує глибину аналізу, а й створює багатовимірну картину події, що наближається до людського розуміння.

4.3.4. Агрегація результатів, уникнення галюцинацій, підвищення повноти

Останнім етапом є статистична агрегація результатів від усіх «експертів». Кожен зв'язок (s, p, o) отримує вагу

$$w = N_{nconf},$$

де *nconf* – кількість «експертів», що підтвердили зв'язок, а N – загальна кількість запитів. Зв'язки з вагою $w < \tau$ (наприклад, $\tau = 0.3$) вважаються ненадійними і відсікаються.

Додатково, для верифікації використовується зовнішній пошук: для кожного зв'язку формується запит до Google/Bing (наприклад, "Sandworm" "Zloader"), і якщо результати відсутні, вага знижується. Остаточна достовірність оцінюється як зважена сума:

$$p_{final} = \beta \cdot p_C + (1-\beta) \cdot w,$$

де p_C – ймовірність, отримана з пошукової системи, а $\beta \in [0,1]$ – параметр довіри до зовнішніх джерел.

Цей механізм дозволяє одночасно підвищувати повноту (завдяки множинному аналізу) та точність (завдяки верифіка-

ції), усуваючи основні недоліки LLM – галюцинації та поверхневість.

Разом усі ці підходи утворюють замкнутий цикл автоматизованої побудови семантичних мереж, де кожен етап контролюється, верифікується та оптимізується. Саме ця система забезпечує перехід від хаотичного потоку текстів до цілісної, достовірної, багатовимірної моделі знань.

4.4. Формалізація процесу побудови знань

Семантичний нетворкінг як методологія набуває наукової цінності лише тоді, коли його процеси піддаються формальній математичній моделі. Це дозволяє не лише описувати результати, а й аналізувати властивості мережі, оцінювати достовірність зв'язків, прогнозувати її еволюцію та забезпечувати відтворюваність аналізу. У цьому підрозділі пропонується формальна модель семантичної мережі знань, механізми оцінки достовірності та правила її динамічної трансформації.

4.4.1. Математична модель семантичної мережі: множини, функції ваг, метрики центральності

Як було сказано вище, семантичну мережу знань можна формально визначити як зважений орієнтований граф $G = (V, E, \varphi, \psi)$.

На основі цієї моделі можна визначити стандартні метрики центральності, адаптовані до семантичного контексту:

Ступінева центральність:

$$C_D(v) = \sum_{u \in V} \left(\psi(v,u) + \psi(u,v) \right) -$$

загальна вага зв'язків вузла.

Близькість:

$$C_C(v) = \left(\sum_{u \in V \setminus \{v\}} d_\psi(v,u) \right)^{-1},$$

де d_ψ – найкоротша шляхова відстань, обернена до суми вагових значень ребер.

Посередництво:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

де σ_{st} – кількість найкоротших шляхів між s і t, а $\sigma_{st}(v)$ – кількість таких шляхів, що проходять через v.

Ці метрики дозволяють виявляти ключові сутності, ті, що відіграють центральну роль у передачі знань або координації подій.

4.4.2. Оцінка достовірності зв'язків

Достовірність кожного зв'язку $e=(s,p,o)$ оцінюється через три незалежні компоненти.

По-перше, частотність $f(e)$ – кількість документів, у яких зв'язок був видобутий. Це дає базову оцінку, але не гарантує істинності (наприклад, дезінформація може поширюватися масово).

По-друге, контекстуальна узгодженість $c(e)$ – міра семантичної когерентності зв'язку в межах одного документа. Вона може бути оцінена через косинусну подібність векторних уявлень суб'єкта, предиката та об'єкта у просторі LLM:

$$c(e)=\cos(v_s,v_o)\cdot I\left[p\sim(s,o)\right]$$

де I – індикаторна функція, що перевіряє онтологічну коректність (що є вілідним предикатом для (s,o)).

По-третє, зовнішня верифікація $v(e)$ – результат пошуку в незалежних джерелах (Google, Bing, офіційні бази). Якщо запит «s p o» повертає релевантні результати, то $v(e)\to 1$, інакше $v(e)\to 0$.

Остаточна вага зв'язку обчислюється як зважена сума:

$$\psi(e)=\alpha\cdot f(e)+\beta\cdot c(e)+\gamma\cdot v(e),$$

де $\alpha+\beta+\gamma=1$, а коефіцієнти відображають пріоритет системи (наприклад, у кібербезпеці γ може бути вищим, ніж у соціальному моніторингу).

4.4.3. Моделі трансформації мереж

Семантична мережа є динамічною системою, що еволюціонує з надходженням нових даних. Її стан у момент часу t позначається як $G(t)$. Трансформація мережі відбувається через три типи операцій:

1. Додавання: при надходженні нового зв'язку enew з вагою $\psi(e_{new}) > \tau_{add}$ (поріг додавання), виконується:
   $V(t+1) = V(t) \cup \{s, o\}$,
   $E(t+1) = E(t) \cup \{e_{new}\}$.
2. Видалення: якщо вага існуючого зв'язку падає нижче порогу τ_{del} (наприклад, через спростування у нових джерелах), то:
   $E(t+1) = E(t) \setminus \{e\}$.
   Якщо вузол втрачає всі зв'язки, він також видаляється.
3. Модифікація: якщо нові дані змінюють властивості вузла (наприклад, уточнюють тип сутності), застосовується:
   $\varphi'(v) = merge(\varphi(v) < \varphi_{new}(v))$,

де *merge* – функція узгодження властивостей (наприклад, пріоритет надається джерелам із вищою авторитетністю).

Ці операції забезпечують адаптивність мережі до змін у реальності, зберігаючи одночасно її логічну цілісність. Таким чином, семантична мережа перестає бути статичним артефактом і перетворюється на живу модель знань, здатну до неперервного навчання та самоуточнення.

Разом ці формальні механізми перетворюють семантичний нетворкінг із евристичного підходу на строгу наукову дисципліну, здатну до кількісного аналізу, верифікації та прогнозування.

4.5. Інтеграція з іншими парадигмами

Семантичний нетворкінг, хоча й є самодостатньою методологією, набуває повної аналітичної потужності лише в умовах глибокої інтеграції з іншими інтелектуальними парадигмами: семантичним вебом, машинним навчанням, графовими базами даних та системами візуалізації. Ця інтеграція не є механічним об'єднанням інструментів, а формує уніфіковану екосистему знань, де кожен компонент виконує свою функцію: LLM генерує гіпотези, семантичний веб забезпечує формальну онтологію, машинне навчання виявляє приховані патерни, а гра-

фові системи надають інфраструктуру для зберігання та дослідження. У цьому підрозділі розглядаються три ключові напрями такої інтеграції.

4.5.1. Семантичний веб (RDF, OWL, SPARQL) та його зв'язок із LLM-мережами

Семантичний веб, заснований на стандартах RDF (Resource Description Framework), OWL (Web Ontology Language) та SPARQL (SPARQL Protocol and RDF Query Language), пропонує формальну модель подання знань у вигляді трійок суб'єкт–предикат–об'єкт. Ця модель є природним доповненням до LLM-генерованих мереж, оскільки забезпечує логічну цілісність та вивід знань.

Формально, LLM-мережа $G_{LLM} = (V, E)$ може бути відображена в RDF-граф $G_{RDF} = (S, P, O)$, де кожен вузол $v_i \in V$ стає URI-ресурсом, а кожне ребро $e_{ij} \in E$ – трійкою (s_i, p_k, o_j). При цьому OWL-аксіоми дозволяють визначити класи Attack ⊑ Event), властивості (hasActor: Attack→ActorhasActor: Attack→Actor) та обмеження (∀ *a*: Attack.∃ *x*: Actor.hasActor(*a*,*x*)).

Така інтеграція дає два ключові переваги. По-перше, SPARQL-запити дозволяють виконувати точні семантичні пошуки, які неможливі в неформалізованих LLM-мережах.

По-друге, машинний вивід (reasoning) дозволяє автоматично виявляти нові знання: якщо APT28 є Actor, а Zloader –Tool, і визначено правило uses(Actor,Tool), система може інферувати зв'язок навіть за відсутності прямої згадки. Таким чином, семантичний веб перетворює LLM-мережу з евристичної конструкції на формально верифіковану онтологію.

4.5.2. Поєднання семантичного нетворкінгу з машинним навчанням

Машинне навчання (ML) додає до семантичного нетворкінгу здатність до виявлення прихованих закономірностей та прогнозування. У той час як LLM видобуває явні зв'язки, ML-моделі аналізують топологічні властивості графа для виявлення неявних.

Наприклад, алгоритми вбудовування графів (Graph Embedding), такі як Node2Vec або GraphSAGE, перетворюють кожен вузол v_i у вектор $z_i \in \mathbb{R}^d$, що зберігає його структурний контекст. На основі цих векторів можна будувати класифікатори для прогнозування типу сутності, регресійні моделі для оцінки ризику або кластеризацію для виявлення спільнот.

Більше того, графові нейронні мережі (GNN) дозволяють поширювати інформацію через ребра:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \frac{1}{\sqrt{|N(i)||N(j)|}} W^{(l)} h_j^{(l)} \right),$$

де $N(i)$ – сусіди вузла i, $W(l)$ – навчальна матриця, σ – активаційна функція. Це дозволяє передбачати відсутні зв'язки (link prediction) – наприклад, імовірність того, що група X використає інструмент Y у майбутньому. Таким чином, ML перетворює семантичну мережу з описової моделі на прогностичну систему.

4.5.3. Взаємодія з графовими базами даних та системами візуалізації

Графові бази даних, зокрема Neo4j, забезпечують інфраструктурну основу для зберігання, запиту та масштабування семантичних мереж. Модель властивісного графа Neo4j ідеально відповідає формату $G = (V, E, \varphi, \varphi)$: вузли мають мітки та властивості, ребра – типи та вагові значення. Запити на мові Cypher дозволяють виконувати складні семантичні операції:

```
MATCH (a:Actor)-[:USES]->(t:Tool)-[:USED_IN]->
(e:Event {target: "BankX"})
RETURN a.name, count(e) AS attacks
ORDER BY attacks DESC
```

Це забезпечує оперативний доступ до знань у реальному часі.

Для глибокого аналізу та дослідження використовується Gephi – платформа для візуального аналізу мереж. Вона реалізує алгоритми компоновки (ForceAtlas2), кластеризації (Louvain), ранжування (Betweenness, PageRank), що дозволяє аналітику бачити структуру мережі, виявляти ключові вузли, ізольовані спільноти, аномальні зв'язки. Експорт з Neo4j у

Gephi (через формат GEXF) створює замкнутий цикл: від зберігання – через аналіз – до візуального дослідження.

Разом усі ці парадигми утворюють інтегровану аналітичну екосистему, де семантичний нетворкінг є центральним інтелектуальним ядром, а інші компоненти забезпечують формальну строгость, прогностичну здатність, інфраструктурну стійкість та інтерактивну дослідницьку середу. Саме така інтеграція робить сучасний OSINT не просто збором даних, а системою побудови знань.

4.6. Етапи життєвого циклу семантичної мережі

Семантична мережа знань не є статичним артефактом, а являє собою динамічну, еволюційну систему, що проходить через чітко визначені етапи життєвого циклу: від концептуального замислу до практичної експлуатації. Цей цикл можна формально описати як послідовність станів $G(t_0) \to G(t_1) \to ... \to G(t_n)$, де кожен перехід відповідає певній фазі аналітичного процесу. У цьому підрозділі розглядаються п'ять ключових етапів: ініціалізація, побудова, верифікація, модифікація та експлуатація.

4.6.1. Вибір джерел, визначення мети аналізу

Ініціалізація є фундаментальним етапом, що визначає всю подальшу траєкторію розвитку мережі. На цьому етапі формулюється аналітична гіпотеза *H*, яка визначає предметну область, типи сутностей та очікувані зв'язки. Наприклад, гіпотеза може мати вигляд:

«Група APT28 використовує інструмент Zloader для атак на фінансовий сектор України».

На основі *H* визначається множина джерел даних $D = \{d_1, d_2, ..., d_k\}$, які мають найвищу релевантність: Telegram-канали, GitHub, судові бази, новинні сайти. Також визначаються онтологічні рамки: які класи сутностей (Actor, Tool, Target) та предикати (uses, targets, located_in) будуть використовуватися. Цей етап забезпечує цілеспрямованість подальшого збору, усуваючи хаотичний аналіз.

4.6.2. Побудова: ітеративне формування через LLM

Побудова мережі є ітеративним процесом, де кожен документ $d_i \in D$ обробляється за допомогою великого мовного моделювання (LLM) для видобування трійок (s, p, o). Формально, це можна визначити як відображення:

$$f_{LLM} : D \to K,$$

де K – множина семантичних трійок. Ключовою особливістю є ітеративність: початкова мережа G_0 формується на основі перших документів, а потім поступово розширюється за рахунок нових даних.

Для підвищення якості застосовується метод «рою віртуальних експертів», де кожен документ аналізується кількома LLM у різних ролях (кібераналітик, юрист, соціолог). Результати агрегуються, і лише зв'язки з вагою w>τ додаються до мережі. Таким чином, побудова є не просто технічним етапом, а процесом колективного інтелектуального синтезу.

4.6.3. Верифікація

Верифікація є критичним етапом, що забезпечує надійність знань. Вона включає три рівні:

1. Автоматизована крос-перевірка – для кожного зв'язку $e = (s, p, o)$ формується пошуковий запит до зовнішніх систем (Google, Bing, офіційні бази), і результати використовуються для корекції ваги $\psi(e)$.
2. Виявлення аномалій – застосовуються алгоритми виявлення викидів (наприклад, на основі розподілу центральності), що дозволяє виявити малоймовірні зв'язки (наприклад, зв'язок між студентом і військовою базою).
3. Людський контроль – аналітик перевіряє критичні вузли та ребра, особливо ті, що мають високу центральність, але низьку зовнішню підтримку.

Цей етап перетворює мережу з евристичної конструкції на верифіковану модель реальності.

4.6.4. Модифікація

Модифікація є творчим етапом, де мережа використовується не лише для опису минулого, а й для моделювання майбутнього. На цьому етапі застосовуються механізми сценарного аналізу:

- Реконструкція подій. Навіть при відсутності хронології, каузальні шляхи в мережі дозволяють відновити логіку подій.
- Генерація сценаріїв – шляхом аналізу можливих продовжень шляхів (наприклад, Actor → uses → Tool → targets → ?) система пропонує ймовірні цілі.
- Прогнозування – за допомогою графових нейронних мереж (GNN) оцінюється ймовірність появи нових зв'язків.

Формально, це можна визначити як відображення:

$$g : G(t) \to S,$$

де S – множина можливих сценаріїв. Таким чином, мережа стає інструментом активного прогнозування, а не пасивного фіксування.

4.6.5. Експлуатація

Останнім етапом є експлуатація, де мережа інтегрується в аналітичні та оперативні процеси. Це включає:

- Семантичний пошук – запити на кшталт «Покажи всі атаки на банки, пов'язані з APT28» виконуються через Cypher або SPARQL.
- Підтримка прийняття рішень – мережа надає контекст для оцінки ризиків, планування заходів безпеки, розслідування інцидентів.
- Аналітика – використання метрик центральності, кластеризації, модулярності для виявлення ключових акторів, слабких ланок, стратегічних точок.

На цьому етапі мережа перестає бути внутрішнім артефактом і перетворюється на публічний інструмент знань, здатний впливати на реальні процеси.

Разом усі ці етапи утворюють замкнутий цикл інтелектуального аналізу, де семантична мережа постійно зростає, уточ-

нюється, адаптується та застосовується. Саме такий цикл забезпечує перехід від хаосу даних до порядку знань.

4.7 Семантичне індексування

Семантичне індексування – це процес, під час якого документи аналізуються для виділення їх сутності та змісту шляхом екстрагування ключових слів, імен власних, назв організацій, а також інших релевантних ознак і категорій, з метою полегшення подальшого пошуку, аналізу та виявлення кореляції між документами.

У багатьох сферах, наприклад, у кібербезпеці цей процес має важливе значення, оскільки текстові масиви можуть містити розмиту по документам критичну інформацію про шкідливе програмне забезпечення, хакерські угрупування, типи атак, вразливості та інші аспекти.

Автоматизовані методи семантичного індексування дозволяють швидко знаходити найважливіші дані у великих обсягах текстів (наприклад, у звітах про кіберзагрози, технічній документації, записах мережевих подій, системних журналах), тим самим забезпечуючи ефективний аналіз загроз та відповіді на інциденти.

Семантичне індексування дозволяє структурувати великі обсяги текстової інформації для швидкого пошуку та аналізу критично важливих даних, таких як назви шкідливих програм, хакерських груп та вразливостей у кібербезпеці.

4.7.1. Етапи семантичного індексування

Можна виділити чотири основні етапи семантичного індексування, а саме 1) попередню обробку текстових документів, 2) екстрагування сутностей, 3) виявлення зв'язків між сутностями, 4) створення індексів, їх фіксація в полях даних.

Попередня обробка текстових документів

На цьому етапі текстові документи проходять базову обробку для підготовки до подальшого аналізу. Основними завданнями попередньої обробки є, по-перше, очищення тексту, а саме видалення зайвих символів, пунктуації, а також стоп-слів (загальних слів, які не несуть специфічного значення, таких як "і", "або", "це"). Після цього здійснюється лематизація і стемінг

[Pant, 2024], а саме, приведення слів до їх базової форми для уніфікації обробки.

Це дозволяє скоротити різноманіття словоформ та покращити точність пошуку ключових термінів. При цьому стемінг – це процес зведення слова до його "стему" або кореня, який може не завжди бути словниковою формою. Стемінг працює за принципом відсікання суфіксів і префіксів від слова, щоб отримати корінь.

У свою чергу, лематизація – це процес зведення слова до його леми, тобто базової чи словникової форми, враховуючи контекст і граматичні особливості слова. Лематизація є більш точним методом, оскільки вона використовує лінгвістичні правила для розпізнавання правильної базової форми слова залежно від його частини мови та контексту. Іноді комбіноване застосування стемінгу та лематизації дозволяє підвищити загальну ефективність системи обробки тексту, отримуючи швидкий результат зі стемінгу та коригуючи його там, де це необхідно, за допомогою лематизації. І завершує етап попередньої обробки документів токенізація, тобто поділ тексту на окремі одиниці (токени), що можуть бути словами або фразами, що потім будуть аналізуватися в подальших етапах.

У сфері кібербезпеки, крім стандартних методів токенізації, можуть застосовуватися спеціалізовані підходи. Наприклад, токенізація може виділяти абревіатури типу CVE, назви версій програмного забезпечення або типи атак, оскільки вони мають важливе значення в контексті безпеки.

Токенізація тісно пов'язана із застосуванням великих мовних моделей, зокрема можливостями їх додаткового навчання (тюнингу). Розбиття на токени допомагає зосередитись на найважливіших елементах тексту, видаляючи непотрібну інформацію (таку як пробіли, незначущі символи). Завдяки токенізації сутності можна чітко виділяти й аналізувати, що значно полегшує роботу на етапах індексування та класифікації. Крім того, формування ефективних індексів базується на чіткому поділі тексту на сутності й взаємозв'язки між ними, що стає можливим завдяки токенізації.

Після токенізації текст можна перетворити в різні формати для збереження та подальшого аналізу. Одним із таких форматів є JSON. Наприклад:

{

```
  "original_text": "Шкідлива програма WannaCry – вразливість CVE-
2023-1234.",
  "tokens": [
    {
      "token": "Шкідлива",
      "type": "adjective",
      "position": [0, 9]
    },
    {
      "token": "програма",
      "type": "noun",
      "position": [10, 18]
    },
    {
      "token": "WannaCry",
      "type": "Malware",
      "position": [19, 27]
    },
    {
      "token": "використовує",
      "type": "verb",
      "position": [28, 40]
    },

    {
      "token": "CVE-2023-1234",
      "type": "Vulnerability",
      "position": [41, 55]
    }
  ]
}
```


Цей приклад показує, як можна зберігати інформацію про токени в структурованому вигляді разом із додатковою інформацією, такою як тип токену та його позиція у вихідному тексті.

Екстрагування сутностей (Named Entity Recognition, NER)

Після попередньої обробки тексту відбувається екстрагування важливих сутностей. Цей етап полягає у виокремленні з тексту фактографічної інформації, яка має значення для кон-

кретної області аналізу. У контексті кібербезпеки це можуть бути:

- Назви шкідливих програм (наприклад, "WannaCry", "Emotet").
- Назви хакерських угруповань (наприклад, "APT28", "DarkSide").
- Типи атак (наприклад, "DDoS", "SQL Injection").
- Вразливості (наприклад, "CVE-2023-1234").

Екстрагування сутностей може бути реалізована за допомогою методів автоматичного аналізу тексту, таких як машинне навчання або використання словників з наперед визначеними значеннями. Основна мета цього етапу – знайти ключові елементи інформації, які є критичними для аналізу кіберзагроз.

Витяг зв'язків між сутностями

Цей етап полягає у виявленні та фіксації зв'язків між сутностями, які раніше були екстраговані з тексту. Зв'язки можуть мати різні типи і контекстуальні значення. У сфері кібербезпеки це можуть бути такі зв'язки, як, наприклад, зв'язок між хакерськими угрупованнями та шкідливими програмами, які вони використовують, зв'язок між типами атак і вразливостями, які експлуатуються, Зв'язки між різними загрозами, що належать до одного типу або мають спільні характеристики. Вони є важливою частиною семантичної мережі, оскільки вони дозволяють не лише ідентифікувати окремі сутності, але і виявляти їх взаємозв'язки, що важливо для прогнозування поведінки загроз.

Зв'язки між сутностями можуть бути збережені у вигляді індексних структур, де кожна сутність пов'язана з іншими. Це дозволяє швидко знаходити не тільки сутності, але і контексти, в яких вони зустрічаються. Кожен зв'язок може також мати свою вагу, що відображає ступінь важливості, ймовірність застосування цього зв'язку, наприклад, як інформаційного впливу або шляху при проведенні хакерських атак.

На цьому етапі виділені сутності також структуруються та фіксуються у вигляді індексів. Процес індексування включає: асоціацію сутностей з документами, коли кожен документ отримує відповідний набір індексів на основі сутностей, знайдених у тексті; структуроване збереження, коли індекси фіксу-

ються у полях баз даних або інших структурах (наприклад, у вигляді таблиць або графів). Зазначимо, індекси можуть бути динамічно оновлюванні у випадку надходження нових документів або змін у базі даних.

Індекси можуть бути представлені у структурованих форматах, таких як JSON. Це забезпечує гнучкість і простоту в обробці та зберіганні даних для подальшого аналізу. В JSON можна зберігати як самі сутності, так і їх зв'язки та вагові значення. Наведемо приклад даних у форматі JSON:

```
{
  "document_id": "doc123",
  "entities": [
    {
      "type": "Malware",
      "name": "WannaCry",
      "id": "entity001",
      "start_position": 15,
      "end_position": 24
    },
    {
      "type": "HackerGroup",
      "name": "APT28",
      "id": "entity002",
      "start_position": 58,
      "end_position": 63
    },
    {
      "type": "Vulnerability",
      "name": "CVE-2023-1234",
      "id": "entity003",
      "start_position": 87,
      "end_position": 100
    }
  ],
  "relations": [
    {
      "source": "entity002",
      "target": "entity001",
      "relation_type": "uses",
      "weight": 0.9
    },
    {
      "source": "entity003",
```

```
      "target": "entity001",
      "relation_type": "exploits",
      "weight": 0.8
    }
  ]
}
```

У наведеному прикладі документ містить три сутності: шкідливу програму *WannaCry*, хакерську групу *APT28*, та вразливість *CVE-2023-1234*. Ці сутності мають певні зв'язки між собою, наприклад, *APT28* використовує *WannaCry* ("uses"), а *WannaCry* експлуатує вразливість *CVE-2023-1234* ("exploits"). Кожен зв'язок також має вагу, що може відображати ступінь впевненості або важливості зв'язку.

Формат JSON є зручним для зберігання та обробки даних про сутності та їх зв'язки, оскільки він дозволяє легко представляти складні структури у вигляді вкладених об'єктів. Крім того, JSON може бути легко інтегрований з багатьма мовами програмування та базами даних, а також великими мовними моделями, що робить його універсальним форматом для зберігання індексованої інформації.

Подальші етапи обробки документів

Після проведення семантичного індексування, можливі й подальші етапи змістової обробки. Ці етапи зосереджені на глибшому аналізі даних і екстрагуванні додаткової інформації.

Класифікація сутностей

На цьому етапі сутності, що були виділені під час індексування, можуть бути класифіковані за категоріями або типами. Це може бути важливо для подальшого аналізу, зокрема для групування шкідливих програм за типами атак або класифікації загроз за рівнем ризику.

Визначення зв'язків і побудова семантичних мереж

Після класифікації сутностей можливе визначення зв'язків між ними. Наприклад, можна встановити зв'язок між хакерським угрупуванням і шкідливим програмним забезпеченням, яке воно використовує, або між вразливістю і типом загрози, що її експлуатує. Це дозволяє побудувати семантичні мережі, в яких сутності виступають як вузли, а зв'язки між ними – як ребра. Такі мережі додатково дозволяють виявляти так званні «приховані зв'язки» між подіями та сутностями.

Кластерний аналіз і відображення семантичних мереж

Після побудови семантичних мереж можна застосовувати кластерний аналіз для виявлення груп сутностей, що мають схожі характеристики або поведінку. Наприклад, можна кластеризувати шкідливі програми за типами атак або створити групи вразливостей, які використовуються для певного типу загроз.

Визначення мережі зв'язків документів та її кластеризація

На цьому етапі, для кожного документу будується семантична мережа, що відображає концепти та зв'язки між ними. Далі можна порівнювати ці семантичні мережі попарно і формувати мережу зв'язків між документами, яка відображає схожість між ними. Потім ця мережа кластеризується, і отримані кластери документів відповідають спільним подіям за вибраною тематикою за певний період часу.

Пошук як результат індексування

Інформаційний пошук у системах кібербезпеки є кінцевим результатом семантичного індексування та класифікації даних. Після того як документи були індексовані, класифіковані та проаналізовані, користувачі можуть виконувати запити для отримання конкретної інформації. Пошукові системи, побудовані на основі індексів, дозволяють швидко знаходити окремі релевантні документи або масиви за допомогою ключових слів або складних запитів. Пошук є не лише частиною інтерфейсу користувача, але і важливою складовою для безпеки, оскільки він дозволяє швидко ідентифікувати критичні загрози або ризики у великих текстових масивах.

4.7.2. Практичне застосування семантичного індексування

Розглянемо можливості семантичного індексування на прикладі системи «КіберАгрегатор», яка призначена для збору та обробки контенту з соціальних мереж за вибраними темами, зокрема кібербезпеки[71]. «КіберАгрегатор» (CyberAggregator)

[71] D. Lande; I. Subach; A. 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

поєднує методи інформаційного пошуку, аналізу даних та агрегування інформаційних потоків, що робить його потужним інструментом для роботи з великими обсягами даних. Останній напрямок її розвитку – це інтеграція можливостей пошуку в соціальних мережах і штучного інтелекту, що дозволяє значно покращити аналітичні можливості системи.

Основною особливістю цієї системи є здатність автоматично обробляти повні тексти з соціальних мереж та відслідковувати динаміку інформаційних потоків у часовому розрізі. Це дозволяє відстежувати розвиток кіберзагроз, трендів та подій, пов'язаних із кібербезпекою в режимі реального часу.

Витяг важливих концептів і зв'язків в системі «КіберАгрегатор» здійснюється за допомогою промптів до LLM. Модель оцінює важливість концептів і зв'язків, використовуючи свою базу знань.

Формально здійснюється екстрагування із тексту T множини концептів: $C(T) = \{c_1, c_2, ..., c_{n_t}\}$ і зв'язків між ними: $R(T) = \{(c_1, c_2) \mid relationship(c_1, c_2)\}$, тут $relationship(c_1, c_2)$ – зв'язок між концептами c_1, c_2. Практично будь яка сучасна модель LLM (застосовується Llama-3.1) на основі промптів може одночасно визначати важливі концепти і зв'язки між ними. Зокрема, промпт до LLM, у тіло якого має агрегуватись текст документа наступний вигляд:

Вибери до 20 пар зв'язаних понять українською мовою (саме пари понять, а не окремі поняття) із тексту і виведи ці пари у вигляді нумерованого списку через ";", як "поняття;поняття". Кожне поняття може складатися з декількох слів. Ось текст: ГУР знову атакувало російські телеканали – джерела Головне управління розвідки (ГУР) Міноборони України провело масштабну кібератаку на російських провайдерів і заблокувало десятки ресурсів промислових об'єктів рф. …

На основі відповіді моделі отримуємо такі результати:

ГУР;кібератака Міноборони;розвідка

```
російські ресурси;війна
інтернет-провайдери;мобільні оператори
промислові об'єкти;спецтехніка
військово-промисловий комплекс;силові відомства
...
```

У наведеному прикладі як концепти розглядаються ключові слова, які разом із визначеними зв'язками між ними дозволяють формувати мережу концептів $G = (V, E)$, тут:

V – множина вузлів (концептів),

E – множина ребер (зв'язків між концептами).

Система типу LLM може надавати різні варіанти відповідей у різний час під час обробки тексту, причому більшість з них є логічно обґрунтованими з точки зору експерта-людини. Підхід із залученням рою віртуальних експертів забезпечує те, що система LLM генерує декілька відповідей на однакові промпти, що дозволяє аналізувати концепти та зв'язки з різних точок зору. Кожна відповідь віртуального експерта може відповідати окремій ролі або погляду на задачу. Після отримання множини відповідей від різних віртуальних експертів, їх узагальнюють, об'єднують в єдиний файл, наприклад у форматах CSV або JSON, наприклад:

```
{
  "concepts": [
    {"id": 1, "label": "Cybersecurity"},
    {"id": 2, "label": "Malware"},
    {"id": 3, "label": "Threats"}
  ],
  "connections": [
    {"source": 1, "target": 2, "weight": 0.85},
    {"source": 1, "target": 3, "weight": 0.75}
  ]
}
```

Зв'язки між концептами отримують вагові значення залежно від частоти їх згадування. Це дає змогу створювати багатий набір даних для подальшого аналізу та інтелектуального пошуку.

Для кожної пари концептів (c_i, c_j) система LLM (або рій віртуальних експертів) може або встановити зв'язок між ними, або не встановити. Кожен раз, коли система встановлює зв'язок між c_i та c_j, збільшується вага цього зв'язку. Позначимо вагу зв'язку між концептами c_i та c_j як w_{ij}. Ця вага залежить від кількості разів, коли рій віртуальних експертів підтверджує наявність зв'язку між відповідними концептами. Нехай $r_{ij}(k)$ – результат відповіді k-го віртуального експерта для пари концептів (c_i, c_j), і нехай $r_{ij}(k) = 1$, якщо експерт підтвердив зв'язок між концептами c_i та c_j та $r_{ij}(k) = 0$, якщо експерт не підтвердив зв'язок. Загальна вага зв'язку між c_i та c_j визначається як сума всіх підтверджень від експертів:

$$w_{ij} = \sum_{k=1}^{K} r_{ij}^k,$$

де K – загальна кількість запитів до рою віртуальних експертів.

Для того, щоб всі вагові значення були у межах від 0 до 1, їх можна нормалізувати:

$$w_{ij}^{norm} = \frac{w_{ij}}{K},$$

тут w_{ij}^{norm} – нормалізована вага зв'язку, яка відображає відсоток експертів, що підтвердили цей зв'язок.

Для того, щоб залишити лише найвагоміші, найзначущі зв'язки, використовуємо пороговий критерій. Якщо вага зв'язку w_{ij}^{norm} перевищує певний поріг θ, то зв'язок зберігається, інакше він відкидається. Тобто залишаємо зв’язок між c_i та c_j

, якщо $w_{ij}^{norm} \geq \theta$. Значення порога θ можна налаштувати в залежності від потрібної точності. Вищі значення θ залишатимуть лише ті зв'язки, які підтвердили більшість експертів, що збільшить точність мережі за рахунок повноти.

У результаті виконання наведених вище етапів за тематикою «кібератаки на російські ресурси» на базі аналізу 30 документів із мережевих ЗМІ побудовано мережу концептів, центральна частина яких у нашому прикладі має вигляд, наведений на рис. 61.

На етапі формування та кластеризації мережі документів для порівняння документів та визначення їхньої близькості, можна використовувати підхід, заснований на підрахунку спільних зв'язків понять у кожному документі. Формально, процес формування матриці близькості документів M можна визначити наступним чином: нехай $D = \{d_1, d_2, ..., d_m\}$ – це множина документів, $CN(d_k)$ – мережа концептів документа d_k. Для кожного документа d_k визначимо множину пар понять: $P(d_k) = \{(c_i, c_j) \mid c_i, c_j \in CN(d_k), w_{ij} > 0\}$, тут множини $P(d_k)$ включають всі пари зв'язаних концептів у документі d_k.

Для кожної пари документів (d_k, d_l) кількість спільних пар понять між документами визначається як:

$$S(d_k, d_l) = |P(d_k) \cap P(d_l)|,$$

тут $|P(d_k) \cap P(d_l)|$ – це потужність множини $P(d_k) \cap P(d_l)$, тобто кількість спільних пар понять між документами d_k та d_l.

Матриця близькості $M = \|m_{kl}\|$ розмірності $m \times m$, де m – кількість документів, визначається таким чином:

$$m_{kl} = \frac{S(d_k, d_l)}{\max(|P(d_k)|, |P(d_l)|)},$$

де $\max(|P(d_k)|,|P(d_l)|)$ – це нормувальний фактор, що дозволяє враховувати різну кількість пар у кожному документі та уникнути переоцінки подібності через розмір документів.

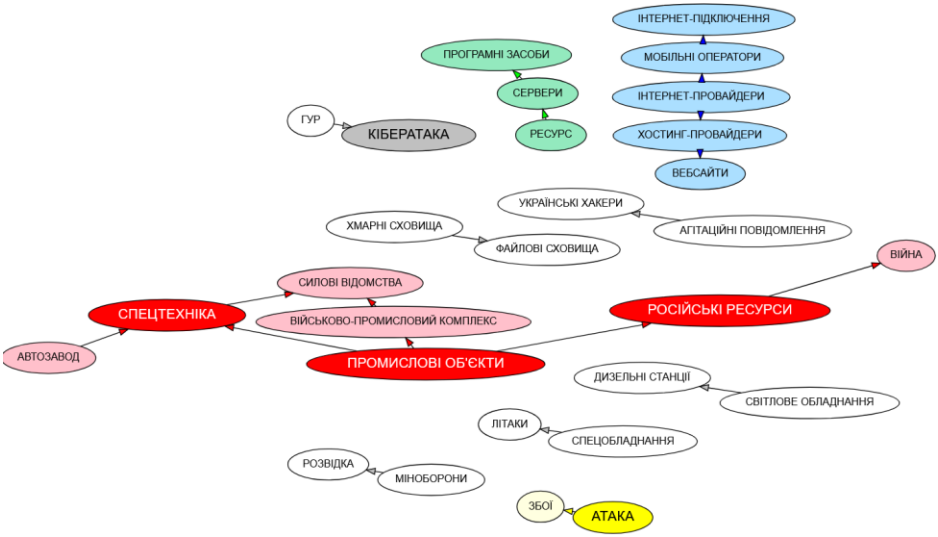


Рисунок 61 – Фрагмент мережі концептів, що відповідають тематиці «кібератаки на російські ресурси»

Елементи матриці близькості m_{kl} варіюються від 0 до 1, де 1 означає, що документи d_k та d_l мають максимальну схожість у парі понять, а 0 означає відсутність спільних пар понять. На рис. 40 наведено фрагмент матриці близькості

	Документ 1	Документ 2	Документ 3	Документ 4	Документ 5	…
Документ 1	0	0,01	0,2	0,01	0,01	
Документ 2	0,01	0	0,01	0,01	0,04	
Документ 3	0,2	0,01	0	0,2	0,01	
Документ 4	0,01	0,01	0,2	0	0,06	
Документ 5	0,01	0,04	0,01	0,06	0	
…						

Рисунок 40: Фрагмент матриці близькості документів

Для групування документів на основі матриці близькості документів нами застосовується метод кластеризації за класа-

ми модулярності. Існують різні види модулярності[72], але для матриць невеликого розміру зручно застосовувати модель Поттса[73], яка враховує так звану розподільну здатність. Перевага цієї моделі – застосування точного критерію якості кластеризації, недолік застосування методу «грубої сили» (у базовому варіанті) для досягнення оптимальності.

Цей підхід ґрунтується на фізичній моделі, яка використовує поняття енергії для оптимізації кластеризації в графах. Основна ідея полягає в мінімізації функціоналу енергії системи, що є аналогом задачі кластеризації. Формалізація цього підходу виглядає наступним чином: для графа з N вузлами та M ребрами, де a_{ij} – вага зв'язку між вузлами i та j, можна записати функціонал енергії, який потрібно мінімізувати:

$$E = -\frac{1}{2}\sum_{ij} J_{ij}\delta\left(c_i, c_j\right) + \sum_i h_i \delta\left(c_i\right),$$

де:

- J_{ij} – вага зв'язку між вузлами i та j,
- $\delta\left(c_i, c_j\right)$ – функція дельти, яка дорівнює 1, якщо вузли i та j належать до одного кластеру, і 0 в іншому випадку,
- h_i – зовнішній магнітний поляризаційний термін для вузла i.

Вагу зв'язку J_{ij} можна виразити через розподільну здатність γ як:

[72] Traag, V.A. and Šubelj, L., 2023. Large network community detection by fast label propagation. Scientific Reports, 13(1), p.2701.

[73] Inaba, K., Inagaki, T., Igarashi, K., Utsunomiya, S., Honjo, T., Ikuta, T., Enbutsu, K., Umeki, T., Kasahara, R., Inoue, K. and Yamamoto, Y., 2022. Potts model solver based on hybrid physical and digital architecture. Communications Physics, 5(1), p.137.

$$J_{ij} = \left(a_{ij} - \gamma \frac{k_i k_j}{2m} \right),$$

де:

- a_{ij} — вага зв'язку між вузлами i та j;
- k_i та k_j – степені вузлів i та j;
- m — загальна кількість зв'язків у графі;
- γ — розподільна здатність.

Функціонал енергії E можна переписати таким чином:

$$E = -\frac{1}{2} \sum_{ij} \left(a_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) + \sum_i h_i \delta(c_i),$$

де $\frac{1}{2} \sum_{ij} \left(a_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$ є частиною, яка визначає модулярність.

Розподільна здатність γ є константою, яка дозволяє масштабувати кількість очікуваних зв'язків між вузлами в межах одного кластеру. Вона впливає на те, як розраховується очікувана кількість зв'язків між вузлами всередині кластеру порівняно з випадковим розподілом, а саме висока γ зменшує вагу зв'язків всередині кластерів, що може знизити модулярність, а низька γ збільшує вагу зв'язків всередині кластерів, що може підвищити модулярність.

У практичних застосуваннях значення h_i (друга складова функціоналу енергії) можна визначити різними способами, залежно від контексту задачі. Ось кілька підходів:

Простий варіант визначення h_i – може бути встановлення його фіксованого значення, наприклад $h_i = 1$. Це підходить, коли немає конкретних зовнішніх факторів, які впливають на кластеризацію. Вибір h_i як ступеня вузла k_i може бути розум-

ним варіантом, особливо якщо важливість вузла у мережі визначається його ступенем. У цьому випадку: $h_i = k_i$.

Можна використовувати нормалізовані значення ступеня або інші метрики, які відображають важливість вузла, з метою забезпечення коректної масштабованості:

$$h_i = \frac{k_i}{\sum_j k_j}.$$

Цей підхід нормалізує ступінь вузла, роблячи його більш порівнянним із значеннями для інших вузлів.

Таким чином, у цьому пункті показано методику застосування LLM для витягу концептів і зв'язків з документів у сфері кібербезпеки, семантичного індексування, кластерного аналізу мереж концептів в мережі документів. Наведено приклад успішної інтеграції LLM з традиційними методами аналізу тексту (інформаційний пошук, кластерний аналіз).

Показано застосування великих мовних моделей для автоматичної обробки текстових даних, зокрема, у сфері кібербезпеки. Побудовані семантичні мережі та застосування кластерного аналізу дозволяють виявити складні взаємозв'язки між концептами та ефективно групувати документи. Отримані результати відкривають нові можливості для автоматизації аналітичних процесів та підтримки прийняття рішень.

4.7.3. Інтеграція інформаційного пошуку та LLM

Незважаючи на прогрес у технологіях інформаційного пошуку, традиційні методи часто стикаються з проблемами повноти та точності отриманої інформації. Ця проблема особливо гостра в сфері кібербезпеки, де своєчасний доступ до релевантних даних є критично важливим. Проблема полягає в необхідності створення більш розвинених систем, які можуть інтелектуально обробляти запити користувачів, визначати найбільш значущу інформацію та представляти її в стислому й зрозумілому форматі.

Зі зростанням обсягів інформації та складності кіберзагроз традиційні методи інформаційного пошуку та аналізу стають менш ефективними. В умовах, коли кіберзлочинці використовують все більш витончені методи атаки, необхідність швидкого та точного виявлення загроз є критичною. Для цього потріб-

но забезпечити автоматизацію обробки великих обсягів даних та підвищити якість аналітичної роботи.

Великі мовні моделі вже продемонстрували значний потенціал у покращенні якості обробки текстової інформації. Наявність програмного забезпечення з вільним доступом і відкритих моделей LLM, таких як LLama[74], відкриває нові можливості для їх інтеграції в закриті корпоративні системи. Це особливо актуально для систем, що займаються проблематикою кібербезпеки та кіберзахисту, де забезпечення надійного та оперативного моніторингу інформаційного простору є пріоритетним завданням.

Інтеграція таких технологій у існуючі системи, як CyberAggregator, дозволяє значно підвищити ефективність виявлення кіберзагроз, оптимізувати процеси інформаційного пошуку та аналізу, а також автоматизувати створення аналітичних зведень та побудову семантичних мереж[75]. Це забезпечує новий рівень захисту інформаційних систем та дозволяє більш ефективно протидіяти сучасним кіберзагрозами[76].

На цей час виникла наукова і практична проблема обґрунтування і створення методології, яка інтегрує LLM у систему моніторингу соціальних медіа, для покращення точності та релевантності інформаційного пошуку. Ця методологія включає семантичне індексування, модифікацію запитів і узагальнення результатів, що виконуються за допомогою LLM. Інтеграція інформаційного пошуку і великих мовних моделей передбачає:

---

[74] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[75] Dmytro Lande, Olexander Puchkov, Ihor Subach Method of Detecting Cybersecurity Objects Based on OSINT Technology // Selected Papers of the XXII International Scientific and Practical Conference "Information Technologies and Security" (ITS 2022) - Vol-3503. - pp 115-124. ISSN 1613-0073. [https://ceur-ws.org/Vol-3503/paper11.pdf]

[76] Ланде Д., Пучков О., Субач І., Рибак О. Інформаційна технологія визначення політичного спрямування джерел інформації для забезпечення інформаційної безпеки держави під час кризових ситуацій. Кібербезпека: освіта, наука, техніка, 2023. - N. 4 (20). - С. 142-152.

- створення методики, яка дозволить проводити попередню обробку текстових даних, визначаючи ключові поняття, зв'язки між ними та формуючи індекс для ефективного пошуку в базі даних.
- розробити підходи до динамічної модифікації запитів користувача на основі семантичного аналізу тексту, що забезпечить більш релевантні результати інформаційного пошуку.
- створення методології для автоматичного формування дайджестів, зведень та інших аналітичних продуктів на основі релевантних документів, отриманих в результаті пошуку.
- впровадження розроблених підходів у існуючу систему моніторингу соціальних медіа, проведення тестових досліджень та аналіз результатів.

Пропонується архітектура інтегрованої з LLM інформаційно-пошукової системи, яка орієнтована на моніторинг соціальних медіа з питань кібербезпеки. Основні компоненти архітектури включають:

- Модуль збору даних, що відповідає за агрегування даних з різних джерел, таких як соціальні медіа, форуми, блоги та інші публічні платформи. Цей модуль забезпечує регулярний і ефективний збір текстових даних, зокрема в реальному часі, з можливістю попередньої фільтрації та очищення даних.
- База даних та сховище даних на основі спеціалізованих СУБД, зокрема, таких як Elasticsearch, для зберігання зібраної інформації та забезпечення швидкого доступу до неї[77]. База даних структурується таким чином, щоб підтримувати ефективне семантичне індексування та пошук, а також можливість масштабування для обробки великих обсягів інформації.
- Модуль семантичного індексування, що виконує функції аналізу тексту та побудови семантичних індексів на ос-

[77] Ahir, D.D. and Shaikh, N.F., 2024. Evaluation of Elasticsearch Ecosystem Including Machine Learning Capabilities. International Journal of Safety & Security Engineering, 14(4).

нові ключових понять і зв'язків між ними[78]. Інтеграція з LLM дозволяє створювати складніші та точніші індекси, які враховують контекст та значення слів у різних доменах знань.

- Модуль пошукової оптимізації, в якому застосовується LLM для модифікації запитів користувачів, щоб покращити результати пошуку. Цей модуль автоматично аналізує вихідні запити, доповнюючи або уточнюючи їх для забезпечення максимальної релевантності та точності результатів.
- Модуль обробки результатів, який відповідає за узагальнення та аналітичну обробку результатів пошуку. Застосування LLM дозволяє автоматично створювати дайджести, аналітичні зведення, виявляти події та побудувати семантичні мапи для візуалізації взаємозв'язків між даними.
- Інтерфейс, який забезпечує взаємодію кінцевого користувача із системою. Інтерфейс включає панелі управління для налаштування запитів, перегляду результатів пошуку та отримання аналітичних продуктів у зручній для користувача формі. Іноді інтеграція з LLM може також дозволяти взаємодію через чатботи або інші інтерактивні інтерфейси.
- Модуль безпеки та управління доступом, який забезпечує захист даних і управління доступом до системи. Цей компонент особливо важливий в контексті інтеграції з корпоративними системами, де потрібно дотримуватися суворих вимог до кібербезпеки.

Запропонована методологія складається з трьох основних етапів:

1. Попередня обробка та семантичне індексування:

- Збір даних з різних відкритих джерел за допомогою системи CyberAggregator.

[78] Dimitri Busch, Dmytro Lande. Semantische Dokumentenindexierung mit generativer KI. Mitteilungen der Vereinigung Oesterreichischer Bibliothekarinnen und Bibliothekare, 2025, 75(1). DOI: 10.31263/voebm.v75i1.9251

- Очищення та нормалізація даних для підготовки їх до індексації.
- Використання LLM для проведення семантичного індексування, визначення ключових понять та відносин у даних. Індексовані дані зберігаються у базі даних Elasticsearch.

2. Обробка запитів:
- Аналіз і модифікація запитів користувачів для підвищення повноти та точності. LLM пропонує синоніми, пов'язані терміни та альтернативні структури запитів.
- Інформаційний пошук релевантних документів у базі даних Elasticsearch на основі модифікованого запиту.

3. Узагальнення та аналіз результатів:
- Узагальнення, тобто автоматична генерація дайджестів, зведень, семантичних карт і інших аналітичних матеріалів за допомогою LLM.
- Виявлення значущих подій та створення семантичних карт, які візуалізують зв'язки між ключовими поняттями.

4.7.4. Математична формалізація семантичне індексування

Нехай є набір документів $D = \{d_1, d_2, ..., d_n\}$ і $T = \{t_1, t_2, ..., t_m\}$ – множина термінів (або токенів), що використовуються для індексації. Процес індексації присвоює вагу w_{ij} кожному терміну t_j у документі d_i, що записується наступним чином:

$$I(d_i) = \{(t_j, w_{ij}) \mid t_j \in T, w_{ij} \geq 0\}.$$

Вага w_{ij} визначається за допомогою LLM, за допомогою яких оцінюється релевантність кожного терміну в контексті документа.

Велика мовна модель модифікує запит $q = \{t_1, t_2, ..., t_l\}$, розширюючи його додатковими релевантними термінами t_k', утворюючи розширений запит q':

$$q' = q \cup \{t_1', t_2', ..., t_p'\}.$$

Релевантність документа d_i до запиту q' оцінюється за допомогою функції подібності:

$$sim(q', d_i) = \sum_{t_j \in q'} w_{ij}.$$

Документ вважається релевантним, якщо його оцінка подібності перевищує визначений деякий поріг ε:

$$sim(q', d_i) > \varepsilon.$$

Процес узагальнення агрегує інформацію з релевантних документів $R = \{r_1, r_2, ..., r_k\}$, щоб створити набір узагальнень U:

$$U = \{u_1, u_2, \ldots, u_k\}.$$

Кожне узагальнення u_i генерується за допомогою LLM відповідного промпту (*prompt*), які виділяють найбільш значущі узагальнення з відповідного документа:

$$u_i = LLM(r_i, prompt).$$

Перейдемо до задачі виявлення зв'язків. Побудова матриці взаємозв'язків термінів: створюється матриця термінів A розміром $m \times m$, де n – кількість термінів у документі. Елемент a_{ij} цієї матриці визначає зв'язок між термінами t_i та t_j.

Обчислення значень в матриці:

- значення a_{ij} можна визначити як частоту спільної появи термінів t_i та t_j в документі d, наприклад, кількість разів, коли терміни t_i та t_j з'являються в одному контексті, або через значення взаємної інформації, що відповідає цим термінам.

Створення графа:

- нехай $G = (V, E)$ – граф, де V – множина вершин (термінів), а E – множина ребер (зв'язків між термінами).

Визначення вершин і ребер:

- вершини із V відповідають термінам t_i з набору $T(d)$;
- ребра E з'єднують пари термінів t_i та t_j, якщо a_{ij} перевищує певний поріг $_{\theta}$:

$$E = \left\{ \left(t_i, t_j \right) \mid a_{ij} > \theta \right\}.$$

де θ – це поріг значущості, який визначає, які зв'язки між термінами є суттєвими.

4.7.5. Нові можливості аналітичних можливостей систем

У результаті інтеграції системи CyberAggregator з великою мовною моделлю Llama-3.1 досягаються покращення аналітичних можливостей систем у сфері моніторингу соціальних медіа та інформаційного пошуку. Нові можливості Llama вдосконалюють різні режими роботи CyberAggregator, включаючи пошук інформації, аналіз динаміки, формування дайджестів та побудову мереж.

Інформаційні зведення (дайджести)

Поєднання технології пошуку з можливостями Llama забезпечує автоматичний аналіз новинних повідомлень і створення узагальнень. Модель Llama дозволяє системі CyberAggregator створювати детальні інформаційні дайджести, що включають:

1. Автоматичне формування дайджестів на основі обробки великих обсягів новинних повідомлень, екстрагування ключових слів, подій і фактів, і на їх основі генерування стислих оглядів основних подій в інформаційному просторі.
2. Лінгвістичні можливості Llama дозволяють глибше аналізувати зміст, пояснювати контекст та значення подій, що покращує точність і корисність дайджестів для кінцевих користувачів.

Мережі хакерських угрупувань

Система CyberAggregator, доповнена можливостями LLM, ефективно візуалізує зв'язки між хакерськими угрупуваннями:

1. Llama допомагає у виявленні та аналізі зв'язків між різними хакерськими групами, їхньою активністю, причетністю до кібератак та відношенням до силових відомств окремих держав.
2. Інтеграція з Llama дозволяє автоматично створювати візуалізації, які відображають зв'язки між угрупуваннями, що полегшує аналіз і виявлення патернів у їхній діяльності.

Мережі термінів

Функціонал LLM також забезпечує аналіз і побудову мереж термінів:

3. Модель допомагає автоматично ідентифікувати ключові терміни та їхні зв'язки, що дозволяє краще розуміти контекст інформації.
4. За допомогою Llama створюються семантичні мережі, які візуалізують зв'язки між термінами і поняттями, що полегшує розуміння складних концептів і їхніх взаємозв'язків.

Мережі персон

Лінгвістичні можливості систем типу LLM використовуються для створення та аналізу мереж персон – акторів кібервійни. Основні аспекти формування мережі персон полягають в аналізі активності осіб, тобто модель LLM дозволяє виявляти зв'язки між різними особами на основі їхньої активності в соціальних медіа та згадувань у різних джерелах. LLM допомагає виявляти не тільки прямі, але й непрямі зв'язки між особами, що дозволяє створювати більш точні і комплексні мережі.

Зокрема, для визначення акторів, які мають відношення до першої в світі кібервійни, запропоновано методику аналізу відібраних документів, доступних в електронних джерелах ме-

режі Інтернет, шляхом застосування системи генеративного штучного інтелекту[79].

На першому кроці методики формується запит до пошукової системи-агрегатора, наприклад, CyberAggregator з ключовими словами, які мають міститись в документі для подальшого аналізу. Даний запит повинен включати ключові слова, обов'язкові для наявності у документі для подальшого аналізу. Після знаходження достатньої кількості текстових повідомлень, дані документи фільтруються за допомогою згенерованого LLM програмного коду, наприклад, мовою програмування Python, пошуку пар понять, які мають формат "Ім'я Прізвище".

На наступному кроці здійснюється завдання фільтрації наданих словосполучень. Формується промпт до LLM з таким формулюванням:

Виділити імена та прізвища з даного файлу, ігноруючи власні назви та назви організацій

У нашому випадку, з понад 30 000 словосполучень було виділено близько 700 імен. Для оптимізації побудови мережі був розроблений програмний код мовою програмування Python за допомогою якого здійснюється підрахунок кількості повторень та вилучення всіх поява, окрім першої, а також вилучення слів, які згадуються менше заданої кількості разів (у нашому випадку – 3), оскільки вони не мають статистичної важливості і лише завантажують мережу зайвою інформацією.

За допомогою ChatGPT створюються зв'язки між акторами:

Знайди зв'язки між персонами, яких пов'язує їх діяльність, об можна було побудувати цільну мережу, і використай всі імена у зв'язках у форматі «персона1; персона2»

На третьому кроці, після встановлення зв'язків між учасниками в заданому форматі, отримана інформація записується у CSV-файл. На четвертому, заключному кроці, використовуючи спеціальний програмний застосунок, розроблений на базі

[79] Пучков О.О., Ланде Д.В., Субач І.Ю. Методики екстрагування об'єктів кібербезпеки з електронних джерел із застосуванням штучного інтелекту. Безпека інформаційних систем і технологій, 2024. N 2(8), 34-41. DOI: 10.17721/ISTS.2024.8.34-41

бібліотеки GraphViz[80], створюється графічне представлення мережі акторів кібервійни та їх зв'язків (Рис. 62).

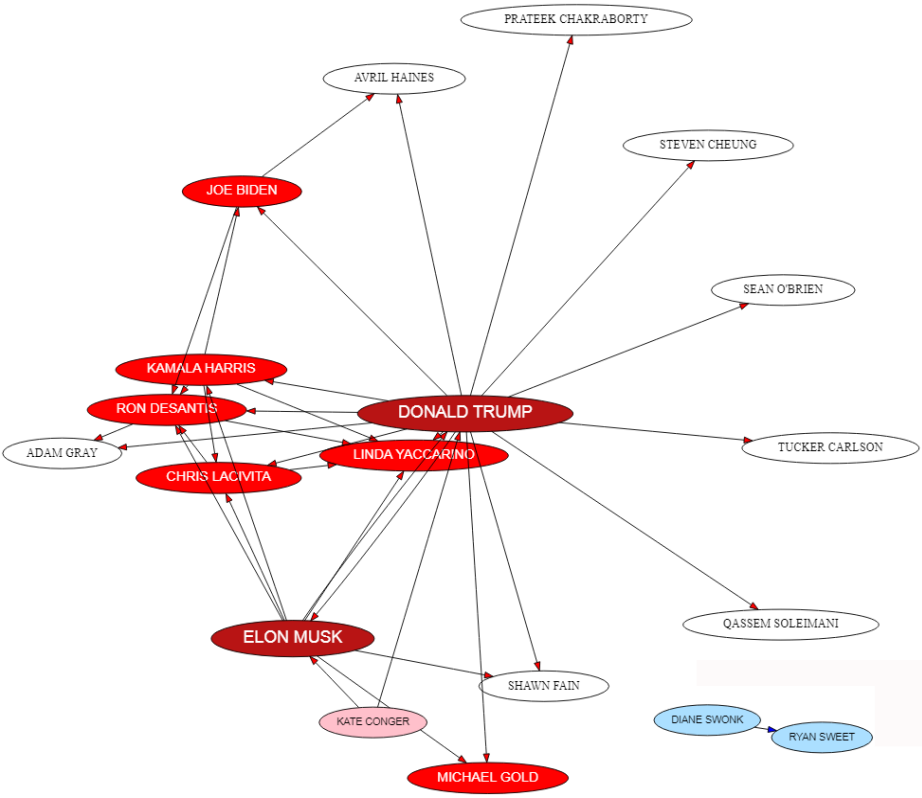


Рисунок 62 – Фрагмент мережі акторів кібервійни

Наведемо формалізацію методу виявлення суб'єктів кібербезпеки. Опішимо вихідні припущення:

Множина документів $D = \{d_1, d_2, ..., d_N\}$ – набір документів, отриманих за допомогою OSINT-систем на основі тематич-

[80] Tamilla Triantoro. Graph Viz: Exploring, Analyzing, and Visualizing Graphs and Networks with Gephi and ChatGPT (March 30, 2023). ODSC Community.

них запитів. Множина хакерських груп H – множина назв хакерських угрупувань, які потрібно виявити з текстів документів. Контекстуальні зв'язки C – множина зв'язків між хакерськими угрупуваннями, що витягуються з текстів документів. Розглянемо методику покроково:

Крок 1: Формування інформаційного масиву публікацій

Для кожного набору тематичних запитів Q (наприклад, запити на основі кібератак в Україні чи Ізраїлі), отримуємо множину документів D, що відповідають цим запитам.

$$D = \bigcup_{q \in Q} OSINT(q)$$

де $OSINT(q)$ – функція, що повертає множину документів за тематичним запитом q.

Крок 2: Екстрагування назв хакерських угрупувань

Для кожного документа $d \in D$ формуємо відповідний промпт до системи ChatGPT для екстракції назв хакерських угрупувань:

$$H(d) = ChatGPT(prompt, d)$$

де $H(d)$ – це множина хакерських груп, витягнутих з документа d, а prompt – це змістовний запит до LLM.

Крок 3: Побудова мережі зв'язків

На основі екстрагованих назв хакерських груп для кожного документа формуємо множину контекстуальних зв'язків:

$$C(d) = \left\{ (h_i, h_j) \mid h_i, h_j \in H(d) \right\},$$

де $C(d)$ – множина парних зв'язків між хакерськими групами з документа d.

Загальна множина зв'язків для всіх документів визначається як:

$$C = \bigcup_{d \in D} C(d).$$

Крок 4: Візуалізація та аналіз мережі

Мережа зв'язків хакерських угрупувань, побудована на основі множини H і множини зв'язків C, може бути представлена у вигляді графа $G=(H,C)$, де H – множина вершин (хакерських угрупувань), C – множина ребер (контекстуальних зв'язків між угрупуваннями).

Обчислювальна складність реалізації методики складається з декількох компонентів:

1. Складність процедури формування інформаційного масиву публікацій залежить від кількості запитів Q та кількості документів N. Складність формування множини документів можна оцінити як $O(|Q|\times N)$.
2. Складність екстрагування назв хакерських угрупувань розраховується з того, що для кожного документа d здійснюється звернення до системи ChatGPT. Нехай t_{GPT} – середній час обробки одного запиту до системи. Тоді загальна складність цього етапу: $O(N\times t_{GPT})$.
3. При побудові мережі зв'язків для кожного документа d екстрагуються зв'язки між угрупуваннями. Якщо в документі d знайдено $|H(d)|$ хакерських груп, то кількість зв'язків між ними можна оцінити як $O(|H(d)|^2)$. Загальна складність побудови мережі буде: $O\left(\sum_{d\in D}|H(d)|^2\right)$.
4. Складність візуалізації залежить від кількості вершин $|H|$. та ребер $|C|$ у графі $G=(H,C)$. У найгіршому випадку, складність візуалізації та аналізу можна оцінити як $O(|H|+|C|)$.

З урахуванням всіх етапів, загальна складність алгоритму:

$$O(|Q|\times N + N\times t_{GPT} + \sum_{d\in D}|H(d)|^2 + |H|+|C|).$$

Наведена методологія дозволяє ефективно екстрагувати та аналізувати взаємозв'язки між хакерськими угрупуваннями на основі даних з текстових джерел, використовуючи засоби генеративного штучного інтелекту. Інтеграція LLM забезпечує більш глибоке розуміння запитів користувачів та дозволяє отримувати більш релевантну та якісну інформацію. Це відкриває нові можливості для автоматизації процесів OSINT та підвищення ефективності роботи аналітиків кібербезпеки.

Система дозволила успішно виявляти та узагальнювати ключові події, що відбулися у сфері кібербезпеки, у автоматичному режимі створювати тематичні аналітичні дайджести з великої кількості документів, наданих інформаційно-пошуковою системою. Реалізовано побудову семантичних карт для візуалізації зв'язків між ключовими поняттями за визначеними напрямками у сфері кібербезпеки.

Інтеграція Llama в систему CyberAggregator дозволила значно підвищити якість і точність інформаційного пошуку та аналітичних процесів. Система тепер здатна автоматично генерувати більш детальні і корисні інформаційні дайджести, створювати точні мережі персон і угрупувань, а також проводити глибший семантичний аналіз термінів. Ці покращення сприяють підвищенню ефективності виявлення важливих подій і патернів у великому обсязі інформації, що є критично важливим для забезпечення кібербезпеки.

На практиці інтеграція системи LLM і інформаційного пошуку надало такі переваги:

Інтеграція Llama в систему CyberAggregator дозволила автоматизувати аналіз великих обсягів текстових даних. Модель автоматично обробляє новинні повідомлення, створює узагальнення та генерує зведення, що підвищує швидкість і точність аналітичних процесів.

Застосування Llama забезпечило значне покращення точності пошуку та релевантності результатів. Модель адаптує запити відповідно до контексту та специфіки запитуваної інформації, що дозволяє отримувати більш точні і корисні результати.

Розроблені математичні моделі забезпечують чітке розуміння та реалізацію процесів, що входять у систему, що дозволяє вдосконалити її функціонування та інтеграцію з Llama.

Таким чином, інтеграція технологій інформаційного пошуку та штучного інтелекту має великий потенціал, зокрема, у сфері кібербезпеки. Запропонована система демонструє, як LLM можуть бути використані для покращення точності та повноти інформаційного пошуку, а також для автоматичного узагальнення результатів. У майбутньому, розвиток цієї системи може призвести до створення більш розвинених інструментів для OSINT, що дозволять краще реагувати на сучасні загрози в кіберпросторі.

4.8. Обмеження та виклики сучасного семантичного нетворкінгу

Незважаючи на значні досягнення в галузі автоматизованої побудови знань, сучасний семантичний нетворкінг залишається обтяженим фундаментальними обмеженнями, що походять як із природи великих мовних моделей, так і з обчислювальних, епістемологічних та організаційних причин. Ці обмеження формують системну невизначеність, яка впливає на достовірність, масштабованість та автономність аналітичних процесів. У цьому підрозділі розглядаються три ключові аспекти: залежність від якості LLM, проблема масштабованості та необхідність людської експертизи.

4.8.1. Залежність від якості LLM

Семантичний нетворкінг, хоча й використовує LLM як інструмент видобування знань, залишається глибоко залежним від їхніх внутрішніх обмежень. По-перше, галюцинації – генерація правдоподібної, але фактично хибної інформації – залишаються нерозв'язаною проблемою. Навіть при точному промпті модель може «вигадати» зв'язок між неіснуючими сутностями, що призводить до поширення помилок у мережі.

По-друге, упередженість (bias) моделі, закладена на етапі тренування, проявляється у систематичному спотворенні певних тем, акторів або регіонів. Наприклад, модель, навчена переважно на англомовних джерелах, може недооцінювати події в україномовному інформаційному просторі.

По-третє, обмежена актуальність знань LLM (cut-off date) робить їх непридатними для аналізу подій, що відбулися після

дати останнього тренування. Це особливо критично в OSINT, де оперативність є ключовою.

Хоча метод «рою віртуальних експертів» та зовнішня верифікація частково компенсують ці недоліки, вони не усувають їх повністю. Таким чином, LLM залишається ненадійним джерелом первинної інформації, яке завжди потребує верифікації.

4.8.2. Проблема масштабованості

Масштабованість є ще одним серйозним викликом. Процес побудови семантичної мережі включає кілька ресурсоємних етапів:

- запуск LLM для кожного документа (часто – кілька разів для різних «експертів»);
- виконання зовнішніх пошукових запитів для верифікації;
- агрегація результатів, обчислення ваг, виявлення аномалій.

Формально, складність процесу можна оцінити як $O(N \cdot M \cdot K)$, де N – кількість документів, M – кількість «експертів», K – середня кількість зв'язків на документ. При великому N (наприклад, мільйони повідомлень з Telegram) ця складність стає неприйнятною для реального часу.

Крім того, часові затримки між надходженням даних та їх інтеграцією в мережу можуть знижувати оперативність аналізу. Навіть при використанні локальних LLM (Llama, Mistral), обробка великих колекцій займає години, що робить систему малопридатною для моніторингу в реальному часі. Тому сучасний семантичний нетворкінг часто вимагає розподілених обчислювальних кластерів або селективного аналізу лише найбільш релевантних джерел.

6.8.3. Роль ментора у керуванні мережею

Найглибшим обмеженням є те, що семантичний нетворкінг не може бути повністю автоматизованим. Навіть найскладніші алгоритми не здатні замінити людську інтуїцію, контекстне розуміння, етичне судження. Тому в архітектурі сучасних систем передбачається роль ментора – аналітика-експерта, який:

- формулює початкову гіпотезу та визначає онтологічні рамки;

- контролює критичні вузли мережі (наприклад, ті, що мають високу центральність, але низьку верифікацію);
- коригує помилки, виявлені в процесі аналізу;
- приймає остаточні рішення на основі результатів мережі.

Ця роль не є пасивною – ментор активно керує еволюцією мережі, визначаючи, які гіпотези розвивати, які зв'язки перевіряти, які сценарії моделювати. Таким чином, семантичний нетворкінг перетворюється не на автономну систему, а на гібридний інтелектуальний конвеєр, де людина залишається центральним елементом управління.

Разом усі ці обмеження вказують на те, що сучасний семантичний нетворкінг, хоча й є потужним інструментом, не є універсальним рішенням. Його ефективне застосування вимагає не лише технічної майстерності, а й глибокого методологічного рефлексу, критичного ставлення до результатів та відповідальної взаємодії між людиною та машиною. Лише за таких умов він може стати надійною основою для сучасної OSINT-аналітики.

Висновки до розділу 4

Розділ 4 продемонстрував, що семантичний нетворкінг є не просто технічним методом, а новою парадигмою інтелектуального аналізу, яка перетворює хаотичний потік неструктурованих даних на цілісну, верифіковану, динамічну модель знань. На відміну від класичних підходів, що спиралися на ключові слова або фіксовані онтології, семантичний нетворкінг використовує великі мовні моделі як інструмент автоматизованого видобування семантичних трійок, але в рамках строгого методологічного каркасу, що включає множинні інтерпретації, зовнішню верифікацію та статистичну агрегацію.

Центральною інновацією є концепція «рою віртуальних експертів», яка імітує колективне розуміння подій через призму різних професійних компетенцій. Цей підхід не лише знижує ризик галюцинацій, а й забезпечує багатовимірну інтерпретацію, наближену до людського аналізу. Разом із формалізацією процесу побудови знань – через математичну модель зваженого графа, функції довіри, метрики центральності – це перетворює

семантичну мережу з евристичної конструкції на кількісно оцінювану систему.

Особливо важливим є відхід від хронологічного принципу на користь логічного: мережа реконструює події на основі причинно-наслідкових або асоціативних зв'язків, навіть коли часові мітки відсутні або сфальсифіковані. Це робить підхід незамінним у умовах інформаційної війни, де деструктивні актори свідомо руйнують часові орієнтири.

Проте семантичний нетворкінг не є універсальним рішенням. Він залишається залежним від якості LLM, страждає від обчислювальної складності, потребує постійного людського контролю. Тому його ефективність визначається не стільки технологіями, скільки методологічною дисципліною: усвідомленням обмежень, системним підходом до верифікації, відповідальною взаємодією між людиною та машиною.

У підсумку, семантичний нетворкінг є мостом між штучним інтелектом та людським розумінням, між даними та знаннями, між минулим і майбутнім. Саме ця здатність до синтезу, верифікації та прогнозування робить його центральним компонентом сучасної OSINT-аналітики, здатної не просто реагувати на події, а формувати майбутнє.

5. Великі мовні моделі як інструмент автоматизованого аналізу OSINT

У епоху інформаційного надлишку, коли щодня генеруються петабайти неструктурованих текстів – від Telegram-каналів і новинних стрічок до законодавчих актів і наукових публікацій, – традиційні методи аналізу відкритих джерел вичерпали свій потенціал. Людина більше не здатна вручну обробляти такі обсяги, а класичні NLP-системи, засновані на правилах або статистичних моделях, не справляються з семантичною складністю сучасного контенту. Саме в цьому контексті великі мовні моделі (Large Language Models, LLM) перетворилися з технологічної новинки на центральний інтелектуальний інструмент сучасного OSINT.

LLM, навчені на масштабах людського знання, володіють здатністю не просто розпізнавати слова, а розуміти контекст, видобувати сутності, встановлювати зв'язки, генерувати гіпотези. Вони дозволяють автоматизувати найскладніші етапи аналізу: від первинного екстрагування фактів до побудови семантичних мереж, від верифікації через множинні інтерпретації до прогнозування подій на основі логічних шляхів. Проте ця потужність супроводжується фундаментальними обмеженнями: галюцинаціями, чорною скринькою, упередженістю, відсутністю актуальності. Тому ефективне застосування LLM у OSINT неможливе без строгого методологічного каркасу, який перетворює модель із джерела помилок на надійний інструмент знань.

Цей розділ присвячений системному дослідженню ролі великих мовних моделей у автоматизованому аналізі відкритих джерел. У ньому розглядаються архітектурні особливості сучасних LLM (GPT, Llama, Mistral, DeepSeek), механізми промпт-інженерії для структурованого видобування знань, концепція «рою віртуальних експертів» для верифікації та глибинного аналізу, а також підходи до інтеграції LLM у повноцінні аналітичні pipeline'и. Особлива увага приділяється безкодовим промпт-інтерфейсам, які дозволяють формалізувати запити, стандартизувати вивід та забезпечити відтворюваність результатів – критичні вимоги для наукової та професійної практики.

Великі мовні моделі не замінюють аналітика – вони розширюють його можливості, перетворюючи його з читача доку-

ментів на архітектора знань. Саме ця синергія людського інтелекту та машинного розуміння лежить в основі майбутнього OSINT – не реактивного, а проактивного, не описового, а прогностичного, не фрагментарного, а цілісного.

5.1. Роль LLM у сучасному OSINT-аналізі

Великі мовні моделі (Large Language Models, LLM) стали каталізатором якісного переходу в аналітиці відкритих джерел: від пасивного, людино-центричного моніторингу – до активного, масштабованого, генеративного інтелекту. Цей перехід не є просто технологічним оновленням; він відображає фундаментальну зміну в епістемології OSINT: знання більше не збираються, а конструюються через інтеракцію між машинним розумінням та людською експертизою. У цьому підрозділі розглядаються три ключові аспекти цієї трансформації: еволюція методів аналізу, переваги LLM як інструменту та їхні внутрішні обмеження.

5.1.1. Від ручного моніторингу до генеративного інтелекту

Традиційний OSINT базувався на ручному опрацюванні документів: аналітик читав новини, переглядав соцмережі, видобував факти, будував висновки. Цей підхід був точним, але масштабно обмеженим – він не міг охопити навіть частину сучасного інформаційного потоку. Навіть автоматизовані NER-системи, засновані на правилах або статистичних моделях (наприклад, spaCy, Stanza), були обмежені фіксованими класами сутностей і не володіли здатністю до семантичного узагальнення.

LLM, навпаки, вводять парадигму генеративного інтелекту: модель не просто класифікує, а інтерпретує, синтезує, моделює. Вона може перетворити неструктурований текст у структуровану трійку (суб'єкт, предикат, об'єкт), побудувати логічний ланцюг подій, запропонувати альтернативні інтерпретації. Це дозволяє переходити від фрагментарного збору фактів до цілісного моделювання реальності. Аналітик більше не шукає інформацію – він керує процесом її генерації, задаючи контекст, роль, формат виводу.

5.1.2. Переваги LLM: масштабованість, багатомовність, контекстне розуміння

Сучасні LLM володіють трьома ключовими перевагами, що роблять їх незамінними в OSINT.

По-перше, масштабованість: одна й та сама модель може обробляти мільйони документів паралельно, що неможливо для людини. Це забезпечує охоплення широкого спектру джерел – від Telegram-каналів до законодавчих баз – в єдиному аналітичному циклі.

По-друге, багатомовність: сучасні LLM навчені на корпусах, що охоплюють десятки мов, включаючи українську, російську, англійську, китайську. Це дозволяє аналізувати інформаційний простір глобально, не втрачаючи контексту через машинний переклад.

По-третє, контекстне розуміння: LLM враховує семантичне оточення слова, що дозволяє правильно інтерпретувати полісемічні терміни. Наприклад, слово «банк» у реченні «атака на банк» буде ідентифіковано як фінансова установа, а в «сидіти на березі» – як частина рельєфу. Ця здатність до дискримінативного розуміння є ключовою для точного видобування знань.

5.1.3. Обмеження: галюцинації, чорна скринька, затримка актуальності

Проте LLM залишаються ненадійними джерелами первинної інформації через три фундаментальні обмеження.

Галюцинації – генерація правдоподібної, але фактично хибної інформації – є наслідком того, що LLM оптимізовані на когерентність, а не на істинність. Модель може «вигадати» неіснуючу норму права, фейковий CVE або неіснуючий зв'язок між акторами, що призводить до поширення помилок у мережі знань.

Чорна скринька – відсутність прозорості у механізмах прийняття рішень – ускладнює верифікацію. Навіть при однаковому промпті різні запуски можуть давати різні результати, що порушує принцип відтворюваності, критичний для наукового аналізу.

Затримка актуальності (knowledge cutoff) – обмеження знань моделі датою її останнього тренування – робить її непри-

датною для аналізу подій, що відбулися після цієї дати. Наприклад, модель, навчена до січня 2024 року, не знатиме про події, що сталися у 2025–2026 роках, що є критичним недоліком у динамічному середовищі OSINT.

Ці обмеження вимагають системного підходу до верифікації: використання множинних інтерпретацій («рій віртуальних експертів»), крос-перевірки через зовнішні джерела, агрегації результатів. Без такого каркасу LLM перетворюються не на інструмент знань, а на джерело нових ризиків.

Таким чином, роль LLM у сучасному OSINT є подвійною: вони надають небувалу продуктивність і глибину аналізу, але одночасно вносять нові джерела невизначеності. Лише через строгу методологію можна перетворити цю двоїстість на силу, здатну будувати достовірні, масштабовані, прогностичні моделі знань.

5.2. Основні задачі OSINT, що вирішуються за допомогою LLM

Великі мовні моделі перетворилися з універсальних генераторів тексту на спеціалізовані інструменти для вирішення конкретних аналітичних завдань у сфері розвідки у відкритих джерелах. Ця трансформація стала можливою завдяки здатності LLM до контекстно-адаптивного видобування знань, семантичної класифікації, логічного висновування та структурованої генерації. У цьому підрозділі розглядаються чотири ключові задачі, які сьогодні ефективно вирішуються за допомогою LLM: екстрагування сутностей, тематична класифікація, виявлення дезінформації та генерація аналітичних продуктів.

5.2.1. Екстрагування іменованих сутностей

Традиційні системи NER обмежені фіксованими класами (PERSON, ORG, LOC), що робить їх непридатними для доменів із постійно змінною термінологією, таких як кібербезпека. Навпаки, LLM дозволяють динамічно визначати типи сутностей через промпт. Наприклад, запит може включати інструкцію: «Видобий усі сутності типу: хакерська група, malware, CVE, організація. Поверни у форматі JSON. Не вигадуй.»

Формально, це можна визначити як відображення:

f_{NER}: $T \rightarrow E$,

де T – множина текстових документів, а E – множина трійок (e,t,c), де e – сутність, t – її тип, c – контекст. Такий підхід дозволяє видобувати не лише відомі сутності (наприклад, APT28), а й нові (наприклад, CVE-2025-12345), що щойно з'явилися в джерелах. Це забезпечує адаптивність до швидкозмінних загроз, критичну для сучасного OSINT.

5.2.2. Класифікація документів за тематикою та рівнем загрози

Класифікація є ключовим етапом фільтрації інформаційного потоку. LLM дозволяють виконувати багаторівневу класифікацію: спочатку за тематикою (кібербезпека, соціальний моніторинг, правова аналітика), потім – за рівнем загрози (низький, середній, високий).

Наприклад, промпт може включати:

«Оціни рівень загрози документа за шкалою 1–5, де 5 – немедленна кібератака. Обґрунтуй оцінку.»

Модель аналізує семантичні маркери: наявність технічних деталей (IP, хеші), тональність (агресивна, погрозлива), конкретність («завтра атакуємо банк X» vs «можливо, хтось колись щось зробить»). Формально, це є функція:

$\text{fclass}: T \rightarrow [1,5] \times R^d$,

Такий підхід дозволяє пріоритезувати аналіз, зосереджуючись на найбільш критичних документах.

5.2.3. Виявлення дезінформації, фейків, deepfake-контенту

LLM можуть виступати як детектори семантичної непослідовності. Хоча вони не мають прямого доступу до візуальних даних, вони можуть аналізувати описи, метадані, контекстні протиріччя. Наприклад, якщо документ стверджує: «Відео з Києва від 15 січня 2026 року показує знищення будівлі», але в інших джерелах будівля ціла, модель може виявити розбіжність.

Більше того, LLM може аналізувати стиль мовлення: надмірна емоційність, використання клише, відсутність кон-

кретики – усе це є ознаками дезінформації. Формально, це можна подати як бінарну функцію:

$f_{disinfo}: T \rightarrow \{0,1\}$,

де 1 означає високу ймовірність дезінформації. Проте така оцінка завжди потребує зовнішньої верифікації, оскільки сама модель може помилятися.

5.2.4. Генерація аналітичних звітів, резюме, рекомендацій

Останнім, але ключовим застосуванням є автоматизована генерація аналітичних продуктів. На основі семантичної мережі LLM може формулювати:

- резюме події;
- звіти з описом акторів, інструментів, цілей;
- рекомендації щодо протидії.

Наприклад, промпт може включати:

«На основі наступних фактів напиши аналітичний звіт для керівництва: хто, що зробив, які наслідки, що робити.»

Це перетворює LLM на інтелектуального асистента аналітика, який не просто обробляє дані, а формує готові до використання інсайти. Формально, це є відображення:

$$f_{report}: K \rightarrow R,$$

де K – база знань (семантична мережа), а R – множина структурованих звітів.

Разом усі ці задачі демонструють, що LLM є не просто інструментом обробки тексту, а універсальною платформою для автоматизованого інтелектуального аналізу, яка охоплює весь цикл: від первинного видобування – до стратегічного висновку. Проте їхня ефективність повністю залежить від методологічного каркасу, що забезпечує верифікацію, уникнення галюцинацій та відтворюваність результатів.

5.3. Методологія автоматизованого аналізу текстів через LLM

Ефективне застосування великих мовних моделей (LLM) у OSINT-аналітиці вимагає не просто запуску промпта, а системної методології, що охоплює весь цикл перетворення неструктурованого тексту на структуровані знання. Цей цикл скла-

дається з трьох взаємопов'язаних етапів: формування цільового інформаційного масиву, застосування структурованих промптів для видобування семантики та забезпечення безкодового інтерфейсу для подальшої обробки. Разом вони утворюють замкнутий аналітичний конвеєр, де кожен компонент гарантує точність, відтворюваність та сумісність з іншими системами.

5.3.1. Формування інформаційного масиву

Першим етапом є формування тематично релевантного масиву документів D={d1,d2,...,dn}, який має бути достатньо повним, щоб охопити всі аспекти аналітичної гіпотези, але достатньо вузьким, щоб уникнути шуму. Це досягається шляхом застосування точних пошукових запитів, сформульованих з урахуванням предметної області. Наприклад, при аналізі кіберзагрози можна використовувати запит:

site:telegram.me "Zloader" intext:"bank" after:2025-01-01

Такі запити можуть генеруватися автоматично за допомогою LLM, яка аналізує контекст завдання та пропонує оптимальні комбінації ключових слів, доменів, часових рамок. Отриманий масив D стає вхідним сигналом для подальшої обробки, де кожен документ di розглядається як носій потенційних знань.

5.3.2. Структуровані промпти

Ключовим механізмом перетворення тексту на знання є структурний промпт – точно сформульована інструкція, яка визначає роль моделі, типи сутностей, формат виводу та обмеження. Такий промпт можна формально подати як кортеж:

$$P = (\rho, S, R, F, C),$$

де ρ – роль («Ти – кібераналітик»); S – множина типів сутностей («група, інструмент, ціль»); R – множина дозволених предикатів («використовує», «цілить»); F – формат виводу («JSON з полями subject, predicate, object»); C – обмеження («Не вигадуй. Якщо немає зв'язку – поверни порожній список»).

Наприклад, для видобування цитат промпт може включати:

«Видобий усі прямі цитати, що містять слова “атака” або “банк”. Поверни у форматі: {"quote": "...", "source": "...", "date": "..."}».

Такий підхід перетворює LLM з «чорної скриньки» на контрольований інструмент видобування, де кожен вихідний елемент має чітко визначену семантичну функцію.

5.3.3. Безкодовий інтерфейс

Останнім, але критичним етапом є стандартизація виводу. Результати роботи LLM повинні бути представлені у форматах, придатних для автоматичної інтеграції в аналітичний pipeline. Найпоширенішими є:

CSV – для імпорту в табличні системи (Elasticsearch, pandas);

JSON – для передачі в API, графові бази, веб-додатки;

GEXF – для візуалізації в Gephi або імпорту в Neo4j.

Ці формати реалізуються через безкодові інтерфейси – конфігураційні файли (YAML/JSON), у яких задаються параметри промпта, джерела даних, формат виводу. Наприклад, YAML-конфіг може містити:

- role: "Кібераналітик"
- entities: ["actor", "tool", "target"]
- output_format: "json"
- validation: true

Такий підхід забезпечує відтворюваність, масштабованість та інтеграцію – три ключові вимоги до сучасного OSINT-аналізу.

Разом усі ці етапи утворюють уніфіковану методологію автоматизованого аналізу, де LLM виступає не як ізольований інструмент, а як частина цілісної інтелектуальної інфраструктури. Саме ця інтеграція дозволяє переходити від хаотичного потоку текстів до структурованої, верифікованої, готової до використання моделі знань.

5.4. Концепція «рою віртуальних експертів» (SVE) у OSINT

У сучасному аналізі відкритих джерел однієї з найбільш гострих проблем залишається ненадійність великих мовних

моделей (LLM), які, незважаючи на свою семантичну потужність, схильні до галюцинацій, внутрішніх упереджень та поверхневих інтерпретацій. Для подолання цього обмеження було запропоновано концепцію «рою віртуальних експертів» (Swarm of Virtual Experts, SVE) – колективного підходу до аналізу, що імітує роботу групи фахівців з різних дисциплін, кожен з яких оцінює одну й ту саму інформацію через призму своєї професійної компетенції. Ця концепція не лише підвищує достовірність результатів, а й забезпечує багатовимірне розуміння складних подій, наближене до людського колективного інтелекту.

5.4.1. Принцип декомпозиції проблеми через множину ролей

Суть методу SVE полягає в декомпозиції аналітичного завдання на кілька підзавдань, кожне з яких виконується LLM у рамках заданої ролі. Наприклад, при аналізі повідомлення про кібератаку система одночасно активує:

- Кібераналітика – для ідентифікації тактик, інструментів, індикаторів компрометації;
- Юриста – для оцінки правових наслідків, перевірки посилань на законодавчі акти;
- Соціолога масової комунікації – для аналізу наративів, тональності, потенційного впливу на суспільну думку;
- Критика – для пошуку логічних суперечностей, семантичних аномалій, ознак дезінформації.

Формально, це можна визначити як розбиття первинної задачі T на підзадачі T_i, де кожна T_i вирішується функцією f_i: $D \rightarrow K_i$, що відповідає ролі r_i. Такий підхід перетворює монолітну інтерпретацію на багатогранну картину події, де кожен аспект аналізу отримує глибоку, спеціалізовану оцінку.

5.4.2. Паралельне опитування LLM з різними рольовими промптами

Для реалізації SVE кожен документ $d \in D$ обробляється паралельно кількома запитами до однієї або кількох LLM, де кожен запит містить рольовий промпт. Наприклад:

- Для кібераналітика: «Ти – експерт з кібербезпеки. Видобий усі індикатори загрози...»

- Для юриста: «Ти – юрист. Перевір, чи є посилання на неіснуючі норми...»

Ці запити можуть виконуватися як до однієї моделі (наприклад, GPT-4), так і до різних (Llama-3, Mistral, DeepSeek), що дозволяє врахувати архітектурні відмінності між моделями. Результат кожного запиту фіксується у структурованому форматі (JSON, CSV), що забезпечує подальшу агрегацію. Такий паралелізм не лише підвищує повноту аналізу, а й створює статистичну основу для оцінки достовірності.

5.4.3. Агрегація та верифікація результатів

Останнім етапом є агрегація результатів від усіх «віртуальних експертів». Кожен зв'язок e=(s,p,o) отримує вагу

$$w(e) = \frac{n_{conf}(e)}{N},$$

де $n_{conf}(e)$ – кількість експертів, що підтвердили зв'язок, а N – загальна кількість запитів. Зв'язки з $w(e)<\tau$ (наприклад, $\tau=0.3$) вважаються ненадійними.

Додатково, для верифікації використовується зовнішній пошук: для кожного зв'язку формується запит до Google/Bing, і якщо результати відсутні, вага знижується. Остаточна достовірність оцінюється як:

$$p_{\text{final}}(e)=\beta\cdot v(e)+(1-\beta)\cdot w(e),$$

де $v(e)$ – це результат зовнішньої верифікації, $\beta \in [0,1]$ – параметр довіри до зовнішніх джерел.

Цей механізм дозволяє одночасно уникати галюцинацій (оскільки випадкові помилки не підтверджуються більшістю експертів) та підвищувати повноту (оскільки різні ролі видобувають різні аспекти події). Таким чином, SVE перетворює LLM із джерела невизначеності на надійний інструмент колективного інтелектуального аналізу.

Разом усі ці елементи утворюють замкнутий цикл верифікованого знання, де кожен факт проходить через множину інтерпретацій, зовнішню перевірку та статистичну оцінку. Саме ця строгость робить концепцію «рою віртуальних експертів» центральним методологічним інноваційним елементом сучасного OSINT.

5.5. Побудова семантичних мереж на основі LLM-відповідей

Семантична мережа, побудована на основі відповідей великих мовних моделей (LLM), є центральним продуктом сучасного OSINT-аналізу. Вона перетворює неструктуровані тексти на формально описану модель знань, де кожен факт стає частиною складної системи зв'язків. Цей процес не є механічним перетворенням – він включає глибоку семантичну інтерпретацію, логічну верифікацію та інтеграцію в графові системи. У цьому підрозділі розглядаються три ключові етапи: формування графової структури, моделювання зв'язків без хронології та інтеграція з інструментами аналізу.

5.5.1. Від пар понять до графів: формування вузлів і ребер

Процес побудови мережі починається з екстрагування семантичних трійок (s,p,o) – суб'єкт, предикат, об'єкт – які LLM генерує на основі тексту. Кожна така трійка перетворюється на елемент графа:

- суб'єкт s та об'єкт o стають вузлами $v_i, v_j \in V$;
- предикат p стає орієнтованим ребром $e_{ij} \in E$.

Формально, семантична мережа визначається як зважений орієнтований граф

$$G = (V,E,\phi,\psi),$$

де ϕ: $V \to P$ – функція, що зіставляє кожному вузлу набір властивостей (наприклад, type, name, source), а ψ: $E \to [0,1]$ – функція ваги, що відображає достовірність зв'язку (визначена через агрегацію SVE та зовнішню верифікацію). Такий підхід забезпечує уніфіковане подання знань, придатне для подальшого аналізу.

5.5.2. Причинно-наслідкові та асоціативні зв'язки без прив'язки до хронології

Однією з ключових переваг семантичної мережі є її незалежність від хронології. На відміну від класичних методів, що

вимагають часових міток, мережа базується на логічних зв'язках:

- Каузальні зв'язки («спровокував», «призвів до», «використав») утворюють орієнтовані шляхи, що відображають причинно-наслідкову логіку подій;
- Асоціативні зв'язки («згадується разом з», «пов'язаний з») формують ненаправлені компоненти, що відображають тематичні кластери.

Наприклад, мережа може містити шлях:

(«група APT28» → «використала Zloader» → «атака на банк X»), навіть якщо дата атаки невідома. Це дозволяє реконструювати події в умовах інформаційного хаосу, де часові орієнтири свідомо спотворені або відсутні. Таким чином, мережа стає логічною, а не хронологічною моделлю реальності.

5.5.3. Інтеграція з Neo4j та Gephi

Для практичного застосування семантична мережа інтегрується в графові системи. У Neo4j вузли отримують мітки (:Actor, :Tool, :Event), а ребра – типи (:USES, :TARGETS). Дані імпортуються через Cypher-скрипти або CSV-файли, що дозволяє виконувати складні запити:

```
MATCH (a:Actor)-[:USES]->(t:Tool)-[:USED_IN]-
>(e:Event {target: "BankX"})
RETURN a.name, count(e) AS attacks
ORDER BY attacks DESC
```

Для глибокого аналізу мережа експортуються в Gephi у форматі GEXF. Тут застосовуються алгоритми:

- ForceAtlas2 – для компоновки;
- Louvain – для кластеризації;
- Betweenness Centrality – для виявлення ключових вузлів.

Метрика центральності CB(v) визначається як:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

де σ_{st} – це кількість найкоротших шляхів між s і t, а $\sigma_{st}(v)$ – це кількість таких шляхів, що проходять через v. Це дозволяє ви-

явити стратегічні точки у мережі – наприклад, інструмент, через який проходить більшість атак.

Разом усі ці етапи утворюють замкнутий цикл побудови знань: від LLM-видобування – через формалізацію – до інтерактивного аналізу. Саме ця інтеграція перетворює семантичну мережу з абстрактної конструкції на практичний інструмент підтримки прийняття рішень.

5.6. Інструменти та технології для LLM-аналізу OSINT

Ефективне застосування великих мовних моделей (LLM) у сфері розвідки у відкритих джерелах вимагає не лише методологічної дисципліни, а й надійної технологічної інфраструктури, що забезпечує доступ до моделей, їхню конфіденційність, масштабованість та інтеграцію в аналітичні pipeline'и. Сучасний ландшафт LLM-технологій охоплює три ключові шари: хмарні платформи доступу, локальні моделі з відкритим кодом та програмні бібліотеки для автоматизації. Разом вони утворюють гнучку, безпечну та продуктивну екосистему для OSINT-аналітики.

5.6.1. Платформи доступу

Хмарні платформи забезпечують найшвидший шлях до використання передових LLM без необхідності локального розгортання. OpenAI API надає доступ до серії моделей GPT, включаючи GPT-4 Turbo, з підтримкою структурованого JSON-виводу, що критично важливо для автоматизованого аналізу. Groq пропонує ультрашвидкий інференс завдяки спеціалізованому апаратному забезпеченню (LPU), що дозволяє обробляти тисячі запитів на секунду – ідеально для масштабованих OSINT-завдань.

Ollama – це локальна платформа, яка одночасно підтримує хмарні та on-premise моделі, забезпечуючи єдиний інтерфейс для роботи з Llama, Mistral, Qwen тощо. Hugging Face Inference API надає доступ до тисяч моделей з Model Hub, включаючи спеціалізовані версії для NER, класифікації, сумаризації. Ці платформи відрізняються за латентністю, вартістю, рівнем контролю та політикою конфіденційності, що вимагає вибору відповідно до специфіки завдання.

5.6.2. Локальні LLM

Для завдань, що вимагають максимальної безпеки, особливо в державному секторі, ключовим є використання локальних LLM з відкритим кодом. Моделі типу Llama-3 (Meta), Mistral (Mistral AI), Qwen (Alibaba) надають високу якість семантичного розуміння при повному контролі над даними. Оскільки вхідні тексти ніколи не покидають локальну інфраструктуру, такий підхід усуває ризики витоку конфіденційної інформації, що є критичним при аналізі чутливих OSINT-джерел.

Крім того, локальні моделі можна точно налаштовувати (fine-tune) на доменні корпуси – наприклад, на українське законодавство, термінологію кібербезпеки або військову тактику. Це значно підвищує точність видобування сутностей і зв'язків. Недоліком є потреба в потужному GPU-кластері, але з появою оптимізованих квантованих версій (GGUF, AWQ) навіть моделі з 70 млрд параметрів можуть працювати на одному сервері.

5.6.3. Автоматизація через мову Python

Для побудови повноцінних аналітичних систем використовується екосистема Python-бібліотек, що забезпечує інтеграцію, оркестрацію та розширення функціональності LLM.

LangChain надає універсальний каркас для побудови ланцюгів обробки (chains): від завантаження документів через загрузники (document loaders), розбиття на чанки, до застосування промптів, кешування, маршрутизації запитів. Він підтримує різні backend'и (OpenAI, Ollama, Hugging Face), що дозволяє легко перемикатися між моделями.

LlamaIndex спеціалізується на індексації та пошуку знань: він будує векторні індекси, графи знань, гібридні структури, що дозволяють ефективно отримувати релевантний контекст для LLM. Це особливо корисно при роботі з великими колекціями документів, де необхідно витягувати лише релевантні фрагменти.

Transformers (від Hugging Face) надає низькорівневий доступ до моделей: завантаження, токенізація, інференс, fine-tuning. Це дає максимальний контроль над процесом, що необхідно для наукових досліджень або нетипових завдань.

Разом ці інструменти дозволяють створювати повноцінні pipeline'и:

збір → очищення → індексація → промпт-інженерія → LLM-аналіз → верифікація → експорт у Neo4j/Gephi.

Таким чином, сучасний LLM-аналіз OSINT є результатом синтезу хмарних та локальних технологій, відкритих моделей та професійних бібліотек, що забезпечує баланс між продуктивністю, безпекою та гнучкістю. Саме ця інфраструктура лежить в основі надійного, масштабованого, етичного OSINT-аналізу.

5.7. Верифікація та оцінка якості LLM-результатів

Ефективність великих мовних моделей (LLM) у OSINT-аналітиці визначається не лише їхньою здатністю до семантичного видобування, а й надійністю отриманих результатів. Оскільки LLM схильні до галюцинацій, внутрішніх упереджень та поверхневих інтерпретацій, будь-який аналітичний висновок, заснований на їхніх відповідях, повинен пройти багаторівневу верифікацію. Цей процес базується на трьох взаємопов'язаних принципах: крос-перевірці через незалежні джерела, кількісній оцінці достовірності та активному людському контролі. Разом вони утворюють замкнутий цикл, що перетворює LLM із потенційного джерела помилок на надійний інструмент знань.

5.7.1. Крос-перевірка через різні моделі та джерела

Крос-перевірка є першим шаром захисту від хибних інтерпретацій. Вона реалізується через паралельне опитування кількох незалежних систем:

- Різні LLM (наприклад, GPT-4, Llama-3, Mistral) – кожна з яких має власну архітектуру, корпус тренування та внутрішні упередження;
- Зовнішні джерела – пошукові системи (Google, Bing), офіційні бази (zakon.rada.gov.ua, EUR-Lex), технічні реєстри (CVE, MITRE ATT&CK).

Для кожного зв'язку e=(s,p,o) формується запит типу "s p o" до зовнішніх систем. Якщо результати підтверджують існуван-

ня зв'язку, йому присвоюється додатковий бал достовірності. Формально, це можна визначити як функцію зовнішньої підтримки:

$$v(e) = \begin{cases} 1, & if \quad \exists \ R_{rel}; \\ 0, & otherwise. \end{cases}$$

Тут R_{rel} – релевантні результати. Такий підхід значно знижує ймовірність прийняття галюцинацій за факти.

5.7.2. Метрики достовірності

На основі крос-перевірки обчислюються кількісні метрики достовірності, які дозволяють ранжувати зв'язки за надійністю.

По-перше, частотність f(e) – кількість документів, у яких зв'язок був видобутий. Це дає базову оцінку, але не гарантує істинності.

По-друге, контекстуальна узгодженістьс(e) – міра семантичної когерентності зв'язку в межах одного документа. Вона може бути оцінена через косинусну подібність векторних уявлень суб'єкта та об'єкта у просторі LLM:

c(e)=cos(vs,vo)·I[p є валідним предикатом],

де I – індикаторна функція, що перевіряє онтологічну коректність.

Остаточна вага зв'язку обчислюється як зважена сума:

ψ(e)=αf(e)+βc(e)+γv(e),

де α+β+γ=1, а коефіцієнти відображають стратегічний пріоритет системи. Зв'язки з ψ(e)<τ (наприклад, τ=0.3) вважаються ненадійними і відсікаються.

5.7.3. Роль аналітика як ментора системи

Навіть найскладніші алгоритми не здатні повністю замінити людську експертизу. Тому в архітектурі сучасного OSINT передбачається роль ментора – аналітика, який:

- формулює первинну гіпотезу та визначає онтологічні рамки;
- контролює критичні вузли мережі (наприклад, ті, що мають високу центральність, але низьку верифікацію);

- коригує помилки, виявлені в процесі аналізу;
- приймає остаточні рішення на основі результатів мережі.

Ця роль не є пасивною – ментор активно керує еволюцією мережі, визначаючи, які гіпотези розвивати, які зв'язки перевіряти, які сценарії моделювати. Таким чином, LLM-аналіз перетворюється не на автономну систему, а на гібридний інтелектуальний конвеєр, де людина залишається центральним елементом управління.

Разом усі ці механізми забезпечують системну верифікацію, що дозволяє перетворити LLM із джерела невизначеності на надійний інструмент побудови знань. Саме ця строгость робить сучасний LLM-аналіз придатним не лише для дослідження, а й для підтримки стратегічних рішень у сфері безпеки, права та управління.

5.8. Етичні та правові аспекти використання LLM у OSINT

Впровадження великих мовних моделей (LLM) у практику розвідки у відкритих джерелах не лише підвищує аналітичну потужність, а й породжує серйозні етичні та правові виклики, які вимагають системного осмислення. Хоча дані, що аналізуються, формально є публічними, їхнє масове збирання, обробка та інтерпретація можуть порушувати фундаментальні права на приватність, недоторканність особистого життя та захист від автоматизованого профілювання. У цьому підрозділі розглядаються три ключові аспекти: конфіденційність у світлі GDPR, ризики поширення дезінформації через галюцинації LLM та необхідність етичної парадигми White Hat AI.

5.8.1. Конфіденційність, GDPR, право на забуття

Хоча OSINT оперує «відкритими» даними, це не означає, що їхнє використання є автоматично легітимним. Регламент ЄС GDPR (General Data Protection Regulation) встановлює, що будь-яка обробка персональних даних, навіть отриманих з публічних джерел, повинна відповідати принципам цільової обмеженості, мінімізації даних та законності. Особливо важливим є поняття «розумних очікувань приватності»: якщо особа

не могла передбачити, що її публікація в Telegram-каналі стане частиною масового аналізу, така обробка може вважатися незаконною.

Крім того, право на забуття (ст. 17 GDPR) надає особі право вимагати видалення її даних з індексів. Це створює прямий конфлікт із принципами OSINT, де повнота даних є критичною. Тому будь-яка LLM-система, призначена для роботи з європейськими даними, повинна включати механізми анонімізації, аудиту збору даних та процедури реагування на запити на видалення. У відсутності таких механізмів система ризикує не лише юридичними санкціями, а й втратою довіри.

5.8.2. Ризики поширення дезінформації через галюцинації

LLM, навіть найсучасніші, залишаються схильними до галюцинацій – генерації правдоподібної, але фактично хибної інформації. У контексті OSINT це набуває стратегічного характеру: помилкове посилання на неіснуючий закон, фейковий CVE або неіснуючий зв'язок між акторами може призвести до хибних аналітичних висновків, які, у свою чергу, можуть вплинути на рішення в сфері безпеки, права або державного управління.

Ще небезпечнішим є те, що галюцинації часто мають високу когерентність, що робить їх важкими для виявлення навіть для досвідченого аналітика. Тому використання LLM у OSINT вимагає обов'язкової верифікації кожного фактального твердження через незалежні джерела. Без такого каркасу LLM перетворюються не на інструмент знань, а на потенційне джерело дезінформації, що підриває саму основу розвідувальної діяльності.

5.8.3. Етичні рамки для автономних аналітичних систем

У світі, де штучний інтелект може бути використаний як для захисту, так і для атаки, виникає потреба в етичній парадигмі проектування. Концепція White Hat AI пропонує будувати інтелектуальні системи, які діють в інтересах людства, дотримуючись принципів прозорості, підзвітності, недискримінації та поваги до прав людини.

У контексті OSINT це означає:

- відмову від масового спостереження за окремими особами;
- використання анонімізованих, агрегованих даних;
- вбудовані механізми верифікації та аудиту;
- дотримання національного законодавства та міжнародних стандартів.

White Hat AI не є технічним стандартом – це філософія відповідальності, яка має лежати в основі будь-якої LLM-системи, призначеної для суспільного блага. Вона гарантує, що інтелектуальні технології будуть служити не контролю, а безпеці, не маніпуляції, а розумінню.

Разом усі ці аспекти вказують на те, що ефективний OSINT неможливий без глибокого етичного та правового рефлексу. Лише в рамках таких рамок великі мовні моделі можуть стати не просто інструментом аналізу, а надійним фундаментом для побудови відповідальної, прозорої та ефективної інформаційної безпеки.

Висновки до розділу 5

Проведене дослідження засвідчило, що великі мовні моделі трансформувалися з технологічного феномену на центральний інтелектуальний інструмент сучасного OSINT, здатний автоматизувати найскладніші етапи аналізу від видобування сутностей і побудови семантичних мереж до верифікації фактів і генерації аналітичних звітів. Ця трансформація стала можливою завдяки здатності моделей до контекстного розуміння, багатомовності, масштабованості та глибокої семантичної інтерпретації, що дозволяє переходити від фрагментарного збору фактів до цілісного моделювання реальності. Ключовим методологічним проривом виступає концепція рою віртуальних експертів, яка імітує колективне розуміння подій через призму різних професійних компетенцій і у поєднанні зі структурованими промптами та формалізованими форматами виводу перетворює модель із чорної скриньки на контрольований інструмент знань.

Разом з тим потужність великих мовних моделей супроводжується фундаментальними обмеженнями, такими як галюцинації, відсутність прозорості механізмів прийняття рішень та затримка актуальності знань, тому їхнє ефективне застосу-

вання неможливе без системного каркасу верифікації, що включає крос-перевірку через різні моделі та джерела, кількісну оцінку достовірності зв'язків, а також активний людський контроль. Аналітик у такій системі не зникає, а перетворюється на ментора, який керує еволюцією знань, формулює гіпотези, коригує помилки та приймає остаточні рішення, що забезпечує гібридний інтелектуальний конвеєр, де людина залишається центральним елементом управління.

Особливо важливим є те, що великі мовні моделі дозволяють відійти від хронології на користь логічного аналізу, реконструюючи події на основі причинно-наслідкових та асоціативних зв'язків навіть коли часові мітки відсутні або сфальсифіковані, що робить підхід незамінним в умовах інформаційної війни, де деструктивні актори свідомо руйнують часові орієнтири. Нарешті, використання великих мовних моделей у OSINT породжує серйозні етичні та правові виклики, пов'язані з ризиками порушення конфіденційності, поширення дезінформації та відсутності підзвітності, тому будь-яка система повинна будуватися в рамках парадигми White Hat AI, що гарантує служіння інтелектуальних технологій безпеці, а не контролю, та розумінню, а не маніпуляції. У підсумку великі мовні моделі є не просто інструментом, а мостом між людським інтелектом і машинним розумінням, між даними та знаннями, між реакцією та прогнозуванням, і саме ця синергія, підкріплена строгою методологією, етичною відповідальністю та технічною надійністю, визначає майбутнє OSINT як дисципліни, здатної не просто описувати світ, а формувати його безпечне майбутнє.

6. Репутаційний аналіз

6.1. Проблема керування репутацією

Репутація являє собою соціальну оцінку групи суб'єктів про людину, групу людей або компанію, сформовану на основі деяких критеріїв.

Репутація компанії – це комплекс оціночних уявлень цільової аудиторії про компанію, сформований на основі факторів репутації, що мають значення для цієї аудиторії. Згідно з інформаційним листом Вищого господарського суду України «Про деякі питання практики застосування господарськими судами законодавства про інформацію» від 28 березня 2007 року, ділову репутацію юридичної особи складає престиж її фірмового (комерційного) найменування, торговельних марок та інших належних їй нематеріальних активів серед кола споживачів її товарів та послуг.

Успіх компанії безпосередньо пов'язаний з її репутацією. Так, дослідження, проведене австралійськими вченими П. Робертсом і Г. Даулінгом, виявило, що чим вища репутація у компанії, то, по-перше, довший період, протягом якого вона отримує максимальний дохід від своєї діяльності, і, по-друге, тим менше часу компанії потрібно для досягнення середніх по галузі фінансових показників при впровадженні інновацій. Репутаційний капітал (Reputational Capital) – поняття не лише маркетингове, не менше відношення воно має і до фінансів. Грошовий еквівалент ділової репутації може бути виражений у формі гудвілу (goodwill). Відповідно до Міжнародних стандартів фінансової звітності (МСФЗ), гудвіл являє собою різницю між ціною, сплаченою за підприємство покупцями, та «справедливою вартістю» (ця величина часто значно відрізняється від простої вартості всіх активів компанії). Наприклад, у правилах бухгалтерського обліку під репутацією розуміється «різниця між покупною ціною організації та вартістю за балансом усіх її активів і зобов'язань».

Фінансова віддача компанії безпосередньо пов'язана з її репутацією. Так, дослідження, проведене австралійськими

вченими Г. Даулінгом і П. Робертсом[81], виявило дві переваги сприятливого іміджу компанії. Порівнявши дані рейтингу 500 найкращих і найбільш шанованих компаній США, щорічно складаного американським журналом Fortune, за 1984–1995 роки з фінансовими показниками компаній за цей же період, вчені виявили взаємозв'язок між репутацією фірми та її фінансовим рівнем. З'ясувалося, що чим вища репутація у компанії, тим, по-перше, довший період, протягом якого вона отримує максимальний дохід від своєї діяльності, і, по-друге, тим менше часу компанії потрібно для досягнення середніх по галузі фінансових показників при впровадженні інновацій.

Щоб мати можливість з'ясовувати нематеріальну ціну компанії, розробляються експертні оцінки репутації. Вартість репутації може визначатися експертами, наприклад, таким чином. Спочатку розраховується дохід, отриманий компанією за рахунок бренду (різниця між реальним прибутком і доходами, які можна отримати, продаючи небрендований товар), а потім отримана сума множиться на спеціально розрахований коефіцієнт (що залежить від становища компанії в галузі, стабільності фінансових показників тощо). Результат і є ціна бренду, що є важливою частиною репутації.

Існують і непрямі оцінки рівня репутації компаній, наприклад, засновані на результатах опитування керівників фірм та аналітиків, які оцінюють компанії за такими параметрами, як якість менеджменту та продукту, здатність залучити й утримати кваліфіковані кадри, фінансова стабільність, ефективне використання активів, інвестиційна привабливість, застосування нових технологій тощо.

Поняття «Управління репутацією в Інтернеті» (Online Reputation Management, ORM) по суті являє собою комплекс заходів щодо виявлення в мережі негативного контенту та зведення його до мінімуму в соціальних медіа та в результатах пошукової видачі. Це, свого роду, PR-кампанія у кіберпросторі. Гілкою ORM є SERM (Search Engine Reputation Management) – управління репутацією в пошукових системах. На Заході такі

[81] Roberts P.W., Dowling G. R. Corporate reputation and sustained superior financial performance // Strategic Management Journal, 2002. – 23. – № 12. – P. 1077–1093.

послуги практикуються дуже активно, і зростання ORM на рік становить близько 35–40 %.

Сьогодні за статистикою компанії Google 70 % користувачів шукають відгуки про товари та послуги, перш ніж купити їх. Історично першою компанією, яка стала практикувати двосторонній зв'язок із клієнтами в соціальних мережах, стала компанія eBay. На основі зворотного зв'язку було складено рейтинг продавців, на який могли спиратися покупці при прийнятті рішення про покупку. У Росії яскравим прикладом відображення репутації компанії, що базується на відгуках користувачів, можна назвати систему Яндекс.Маркет. Більше половини користувачів Інтернету при виборі того чи іншого продукту, компанії, замовника, виконавця тощо спираються на інформацію, надану іншими користувачами.

Роботи з управління репутацією проводять як спеціалізовані PR-агентства, що працюють на просторах веб-простору, так і підрозділи SEO-агентств, які запускають PR-кампанії, спрямовані на пошук та усунення негативного контенту. Крім того, такі послуги надають і приватні особи – фрілансери, спеціалісти в галузі інтернет-маркетингу та SEO. У великих компаніях існують свої власні відділи, робота яких спрямована на управління репутацією фірми, бренду, товару, послуги.

Поняття «Управління репутацією в Інтернеті» (ORM) вже стало усталеним терміном, і на ці цілі на Заході щорічно виділяється частина бюджету більшості великих компаній. Разом зі зростанням впливу соціальних медіа на погляди та вподобання людей зростає і необхідність великих компаній стежити за своїм іміджем у мережі. На цьому тлі не здається дивним зростання ринку ORM на 40 % щорічно.

Основне завдання управління репутацією – формування позитивного іміджу про компанію та її продукт. Оскільки складно охопити абсолютно всі користувацькі відгуки та прибрати весь негатив, зазвичай зусилля концентруються в трьох областях: пошуковій видачі, відгуках в електронних ЗМІ та згадках у соціальних медіа. Доводиться працювати як з контентом, створюваним редакторами різних видань, так і простими користувачами. Для створення цілісного позитивного образу інформація з цих трьох джерел має бути позитивною або нейтральною.

Управління репутацією в пошукових системах – Search Engine Reputation Management (SERM) – комплекс заходів, спрямованих на виключення негативних відгуків про компанію, товар або послугу в результатах видачі пошукової системи. Послуга управління репутацією в пошукових системах необхідна:

• компаніям, які бажають виключити або мінімізувати негативні відгуки про свою діяльність (продукцію);

• компаніям, які бажають сформувати позитивні відгуки або збільшити їх кількість та видимість для цільової аудиторії.

Негативна інформація, що завдає шкоди репутації в мережі, може бути різного походження[82]. Умовно виділяють три основні групи походження негативного контенту:

• ненавмисний негатив – це можуть бути як відгуки незадоволених клієнтів, які не мають наміру завдати шкоди репутації компанії, а просто не задоволені підсумками співпраці, так і необережно розміщені в Інтернеті фотографії з корпоративних свят, висловлювання співробітників на адресу клієнтів тощо. Зазвичай такий негатив не становить великої загрози, але ігнорувати його в жодному разі не можна;

• навмисний негатив з метою вдарити по репутації – у цьому випадку класичний приклад – негативні відгуки звільнених або тих, хто звільнився, співробітників, незадоволених концепцією компанії.

• чорна PR-кампанія – найнебезпечніший вид негативного контенту, що завдає серйозного удару по репутації. Такі PR-кампанії проводять спеціалісти, які ретельно вивчають бізнес конкурента і точно знають, де прихована ахіллесова п'ята. Організовуються великі рейдерські захоплення, здатні призвести до повного краху не лише репутацію, а й увесь бізнес загалом. Цю послугу у PR-спеціалістів замовляють великі серйозні конкуренти.

Найуразливішими тематиками в плані притягування негативних відгуків можна назвати:

• банки, фінансові інститути;

• діячі політики та шоу-бізнесу;

[82] Kuenzler, J., 2021. From zero to villain: Applying narrative analysis in research on organizational reputation. European Policy Analysis, 7, pp.405-424.

• туризм, подорожі (відгуки про готелі, курорти, туроператорів, авіаперевізників);

• мобільна техніка та зв'язок (оператори, телефони, електронні планшети);

• побутова техніка;

• заклади громадського харчування (ресторани, кафе, бари).

Як простір моніторингу для управління репутацією обирають мережеві ресурси, де розміщуються відгуки споживачів:

• соціальні мережі, месенджери;

• блоги та форуми;

• тематичні веб-сайти та портали;

• спеціальні сервіси відгуків.

Просуваються сторінки з позитивним контентом за допомогою стандартних інструментів пошукової оптимізації (Search Engine Optimization, SEO), таких як біржі посилань, купівля, обмін посиланнями на статті з тематичними ресурсами, розміщення анонсів, новин та ін. При цьому позитивний контент слід розміщувати регулярно, оскільки негативний контент здатен проявлятися знову і псувати репутацію.

Боротися з негативним контентом покликане пошукове управління репутацією - SERM. Завдання SERM полягає у витісненні з результатів пошуку веб-сторінок з небажаною інформацією, в результаті чого цільова аудиторія перестане бачити такі ресурси, оскільки користувачі не будуть виходити на них за допомогою пошукових систем. Для досягнення цієї мети створюються матеріали з позитивним контентом, припускаючи, що вони витіснять негативні небажані повідомлення.

Управління репутацією в мережі зазвичай починають з моніторингу пошукової видачі та соціальних медіа з метою виявлення інформації за заданим об'єктом. Існує декілька методів моніторингу:

• ручний моніторинг пошукових систем шляхом введення цільових пошукових запитів;

• використання систем оповіщення, інтегрованих з пошуковими системами, наприклад, Google Оповіщення (google.com/alerts). У цих випадках релевантна інформація надходить на електронну пошту підписника;

• використання спеціальних засобів моніторингу репутації компаній у соціальних мережах.

Як простір моніторингу для управління репутацією обирають мережеві ресурси, де розміщуються відгуки споживачів:

- соціальні мережі;
- блоги та форуми;
- тематичні веб-сайти та портали;
- спеціальні сервіси відгуків.

Одним із критеріїв якості послуги моніторингу репутації є повнота охоплення – частка інформації про об'єкт, досліджувана під час роботи, від загального обсягу інформації в мережі про об'єкт. Як і раніше, основним інструментом пошуку інформації є традиційні пошукові системи, вони охоплюють значну частину інтернет-контенту, а також деяку частину соціальних медіа.

Сьогодні у всьому світі існують сотні систем моніторингу репутації, серед яких можна назвати системи Babkee, Brandspotter, BuzzLook, Buzzware, IQBuzz, Крибрум, SemanticForce, Wobot, Youscan. У дослідженнях Кена Барбері (Ken Burbary) та Адама Коена (Adam Cohen)[83] наведено список із 230 систем моніторингу репутації, для багатьох з яких пропонуються безкоштовні тестові періоди для оцінки якості їхньої роботи.

6.2. Моделювання репутації в мережах

Останнім часом у рамках теорії аналізу соціальних мереж велика увага приділяється оцінці репутації окремих суб'єктів (агентів, вузлів соціальних мереж) та рівня довіри між ними.

Формально соціальна мережа являє собою граф, в якому множина вершин – це сукупність агентів, суб'єктів – індивідуальних або колективних, а множина ребер являє собою сукупність відносин, сукупності соціальних зв'язків між агентами.

При моделюванні соціальних мереж виникає необхідність врахування динаміки соціальних зв'язків – взаємного впливу агентів.

Вплив у даному випадку розглядається як процес і результат зміни індивідом (суб'єктом впливу) поведінки іншого суб'єкта – об'єкта впливу, його установок та оцінок у ході взаємодії . Таким чином, вплив – це здатність впливати на чиїнебудь уявлення або дії. Розрізняють спрямований і неспрямо-

[83] Burbary K., Cohen A. A Wiki of Social Media Monitoring Solutions // (on-line: http://wiki.kenburbary.com/

ваний вплив. Спрямований вплив використовує як механізми впливу на іншу людину переконання та навіювання. При цьому індивід – суб'єкт впливу – ставить перед собою завдання досягти певних результатів від об'єкта впливу. Неспрямований (нецілеспрямований, опосередкований) вплив – це вплив, за якого індивід не ставить перед собою завдання досягти певних результатів від об'єкта впливу.

Цілеспрямований вплив учасників соціальної мережі (або суб'єктів, які не входять до мережі, але використовують її як інструмент інформаційного впливу) є окремим випадком інформаційного управління, що полягає у формуванні в керованих суб'єктів такої поінформованості, щоб рішення, які вони ухвалюють на її основі, були найбільш вигідними для керуючого суб'єкта.

Можливості впливу одних учасників соціальної мережі на інших її учасників істотно залежать від репутації перших. Репутація – «думка, що склалася, про достоїнства або недоліки кого-небудь, чого-небудь, суспільна оцінка». Репутацію можна розглядати як «вагомість» думки спільноти про окремого агента або групу агентів, що визначається його поглядами та діяльністю (активністю). При цьому репутація може бути як індивідуальною, так і колективною.

Репутація зростає, якщо вибір агента (відповіді на деякі ключові питання) збігається з тим, чого від нього очікує спільнота, і знижується в іншому випадку.

Нехай $\{a_1, a_2, ..., a_n\}$ – множина агентів – вузлів соціальної мережі, які впливають один на одного. Матрицю впливу позначимо як $A = \left\| a_{ij} \right\|_{i=1,n}^{j=1,n}$ ($a_{ij} \geq 0$ позначає ступінь довіри i-го агента до *j*-го). При цьому очевидно, що якщо *i*-й агент впливає на *j*-го, а *j*-й впливає на *k*-го, то це означає наступне: *i*-й агент опосередковано впливає на *k*-го (транзитивність), що дозволяє будувати ланцюжки опосередкованих впливів.

Припустимо, що у кожного агента в початковий момент часу є думка з деякого ключового питання. Нехай думка спільноти агентів мережі відображає вектор початкових мок b^0 розмірності *n*. Думка кожного агента змінюється під впливом думок інших агентів соціальної мережі.

Вважатимемо, що думка і-го агента в момент часу t дорівнює:

$$b_i^t = \sum_{j=1}^{n} a_{ji} b_j^{t-1}$$

При багаторазовому обміні думками, думки агентів сходяться до результуючого вектора думок $B = \lim_{t\to\infty} b^t$. Таким чином, справедливе співвідношення $B = Ab$.

Позначимо r_i – параметр, що описує репутацію і-го агента в соціальній мережі (спільноті), яку можна визначити як нормовану суму його впливів на всіх інших агентів соціальної мережі (передбачається, що $a_{ij} \geq 0,\ \ i, j = 1,...,n$), тобто:

$$r_i = \frac{\sum_{i \neq j} a_{ij}}{R}, \quad j = 1,...,n.$$

Тут $R = \sum_{k} \sum_{j \neq k} a_{kj},\ \ k, j = 1,...,n$ – сумарний взаємний вплив один на одного всіх членів соціальної мережі.

Відповідно до наведеного виразу, агент має тим вищу репутацію, чим вищий його вплив на всіх інших членів соціальної мережі.

6.3. Рейтингування інтернет-ресурсів

З проблемою управління репутацією в мережі Інтернет тісно пов'язане поняття живучості інформації. Своєю чергою, для управління живучістю інформаційних об'єктів необхідне моделювання їхнього життєвого циклу: формування та розвитку, реакції на деструктивні впливи, відновлення, руйнування.

Під живучістю інформаційної системи розуміють здатність її (або її фрагмента) адаптуватися до нових непередбачених умов, протистояння небажаним впливам за одночасної реалізації основної функції – цільового інформування. Крім того, з живучістю інформаційних об'єктів сьогодні пов'язують таку

соціально важливу проблему, як забезпечення інформаційної безпеки.

Існує кілька механізмів, що забезпечують живучість інформаційних об'єктів в Інтернеті. Нижче розглядаються деякі найпоширеніші механізми забезпечення живучості, які на практиці застосовуються не в чистому вигляді, а, як правило, у комбінованому. Для вивчення проблем, пов'язаних із живучістю, необхідно чітко визначити як саме це поняття, так і навести формальну модель, на підставі якої можна розраховувати рівень живучості для таких важко формалізованих сутностей, як інформаційні об'єкти.

6.4. Цифрові сліди та тіні

Видалення інформаційного об'єкта з веб-ресурсу не може гарантувати його зникнення з Інтернету. Залишаються не лише «цифрові сліди» та «цифрові тіні».

Вираз «цифрові сліди» (Digital Footprint) стосується тієї інформації, яка залишається самим користувачем під час роботи в Мережі і за якою можна не тільки його ідентифікувати, а й «прив'язати» до певних дій, подій, відновити якісь фрагменти біографії.

Часто користувачі за власним бажанням зазначають свої П.І.Б., «прив'язуючи» подальшу інформацію до власної особи, дату народження, сімейний стан, освіту, професію, місця попередньої роботи та багато іншого, включаючи контактні телефони та адреси електронної пошти. Крім «цифрових слідів», які користувачі залишають самі, інформація про користувачів постійно тиражується і без жодної його участі.

Інформація про користувача, що створюється без його участі, отримала назву «цифрова тінь» (Digital Shadow), які виникають і накопичуються щоразу, коли хтось шукає користувача через пошукові системи, коли відбувається електронна розсилка за списками, в яких він фігурує, та в багатьох інших випадках. Індексація роботами пошукових машин сторінок з інформацією про користувача та їх подальше кешування – це теж створення «цифрової тіні», доступної кожному. Крім «цифрових тіней відкритого доступу», створюються і накопичуються «цифрові тіні обмеженого доступу» – записи камер спостере-

ження, банківські транзакції, білінги інтернет-магазинів, сервісів продажу квитків, телефонних дзвінків тощо.

За оцінкою аналітичної компанії International Data Corporation (IDC), що спеціалізується на дослідженнях ринку інформаційних технологій, обсяг «цифрової тіні», тобто інформації про користувача Інтернету, яка створюється без його участі, вже у 2007 р. перевищив обсяг інформації, яку створює сам користувач.

З проблемою репутації в Інтернеті щодня стикається дедалі більше користувачів. Про це свідчить і поява особливих сайтів (наприклад, www.suicidemachine.org), які дають змогу одночасно видалити реєстрацію та всі зроблені записи на різних форумах і в соціальних мережах. Така операція називається «накласти на себе руки в Інтернеті». Однак ця система поки що недосконала. Віднедавна цей клопіт беруть на себе спеціальні компанії, так звані «інтернет-чистильники», які налагоджують контакти з адміністрацією провідних пошукових систем і соціальних мереж, окремих веб-сайтів, використовують програмні інтерфейси взаємодії з кешами пошукових систем.

Як ілюстрацію можна навести дані адміністрації соціальної мережі (сервісу мікроблогів) Twitter про кількість запитів на видалення контенту. За даними аналітиків, за перше півріччя 2013 року уряди різних країн направили до Twitter 1157 запитів про надання інформації. За аналогічний період 2012 р. ця цифра становила 849. При цьому в 10 разів зросла кількість запитів на видалення контенту. За кількістю запитів на видалення інформації лідирує Росія. Крім того, відзначається різке зростання урядових запитів. 78 % усіх запитів про інформацію (902) припадають на частку США. На другому та третьому місці знаходяться Японія (87) та Великобританія (26).

Поняття живучості інформаційного об'єкта передбачає його здатність своєчасно виконувати свої функції (в даному випадку – інформування) в умовах дії дестабілізуючих факторів. Такими факторами можуть бути усунення окремих інформаційних об'єктів з інформаційного простору, втрата їхньої актуальності, доступності. Необхідно зазначити, що привернення уваги аудиторії до іншої теми, породження іншого інформаційного об'єкта також може знизити актуальність поточного інформаційного об'єкта.

При цьому слід враховувати, що найважливіша інформація, потрапивши до Інтернету, залишається там практично назавжди, і, як показує практика, розраховувати на її легке видалення або зміну не доводиться. Найкращим методом виявляється витіснення небажаної інформації новими сюжетами, проведення спеціальних заходів щодо змістовного виправлення помилок[84].

Враховуючи ефект надживучості інформації в мережі Інтернет, варто враховувати кілька важливих моментів під час боротьби з негативним контентом при управлінні репутацією в мережі:

не можна просто проігнорувати негативний контент; як відомо, інформаційні повідомлення, особливо негативного спрямування, багаторазово дублюються в мережі. Тому потрібні спростування, позитивний контент;

інтернет-чистильники – служби усунення негативу з мережі Інтернет можуть «механічно» лише частково вирішити проблему. Негативна інформація все одно десь залишиться і колись спливе. Тому слід витіснити негативний контент позитивним;

позитивний контент має бути правдивим, об'єктивним. Інтернет – чудовий детектор брехні;

необхідно розміщувати «вишитовхуючу негатив» позитивну інформацію в мережі на різних цільових ресурсах, дбаючи про гіперпосилання на цю інформацію.

Спостережуваний нині процес у сфері інтелектуалізації автоматизованих систем, переходу від простого оброблення даних до процесів підтримки прийняття рішень потребує нових підходів. Саме тому особливе місце посідають завдання, пов'язані із забезпеченням живучості як інформаційних систем, так і інформаційних об'єктів у мережевому середовищі.

Висновки до розділу 6

Проведене дослідження проблематики репутаційного аналізу в контексті комп'ютерної конкурентної розвідки дозволяє стверджувати, що управління репутацією виступає кри-

[84] Ланде Д.В. Керування репутацією в інформаційних мережах. // Правова інформатика, 2013. - N 3 (39). - С. 3-10.

тичним елементом забезпечення конкурентоспроможності суб'єкта господарювання. Репутація компанії розглядається не лише як маркетингова категорія, а й як фінансовий актив, що має грошовий еквівалент у формі гудвілу та безпосередньо впливає на період отримання максимального доходу та швидкість досягнення галузевих фінансових показників. Сучасні підходи до управління репутацією в Інтернеті, відомі як ORM та SERM, передбачають комплекс заходів щодо виявлення та мінімізації негативного контенту через пошукову видачу та соціальні медіа, проте ефективність цих заходів залежить від розуміння природи негативної інформації, яка може бути наслідком неумісних дій, цілеспрямованого впливу конкурентів або організованих чорних PR-кампаній.

Важливим науковим результатом розділу є формалізація процесів моделювання репутації в соціальних мережах, де агенти та їхні взаємозв'язки представляються у вигляді графів з урахуванням динаміки соціальних зв'язків та взаимного впливу. Використання матриць впливу та векторів початкових думок дозволяє кількісно оцінити репутацію агента як нормовану суму його впливів на інших учасників мережі, а застосування гіперкомплексних числових систем надає інструментарій для врахування позитивного та негативного ставлення суб'єктів до ключових питань з можливістю навчання системи та урахування семантичної синонімії. Такі моделі дають змогу перейти від якісних оцінок до кількісного вимірювання рівня довіри та лояльності суб'єкта щодо суспільства, що є важливим для прогнозування інформаційних загроз.

Окрему увагу в аналізі приділено проблемі інформаційної живучості об'єктів у мережі Інтернет, що тісно пов'язана з репутаційними ризиками, оскільки видалення інформації не гарантує її зникнення через наявність цифрових слідів та тіней, кешування пошуковими системами та архівування даних. Доведено, що інформаційні об'єкти мають власний життєвий цикл, а їхня живучість може бути оцінена через ймовірнісні, булеві або марковські моделі, які враховують розподіл копій на серверах та ступінь критичності функцій інформування. З огляду на ефект надживучості інформації, стратегія управління репутацією не може базуватися виключно на механічному видаленні негативу, а має передбачати витіснення небажаної інформації новими позитивними сюжетами та проведення за-

ходів щодо змістовного виправлення помилок, що вимагає постійного моніторингу пошукової видачі та соціальних медіа з використанням спеціалізованих систем аналізу тональності та охоплення аудиторії.

Таким чином, репутаційний аналіз в системі конкурентної розвідки інтегрує методи соціологічного моніторингу, математичного моделювання мережевих впливів та теорії інформаційної живучості, формуючи цілісну методику захисту ділової репутації. Ефективність протидії негативним інформаційним впливам забезпечується поєднанням технічних засобів моніторингу, правових механізмів захисту та стратегічного планування комунікацій, спрямованого на формування стійкого позитивного іміджу в цифровому просторі, де будь-яка оприлюднена інформація набуває властивостей довгострокового зберігання та потенційного впливу на прийняття управлінських рішень стейкхолдерами.

7. Правові питання конкурентної розвідки

7.1. Конкурентна розвідка в правовому полі

Безумовно, конкурентна розвідка як сфера діяльності має здійснюватися в рамках правового поля держави. Основою для цього є конституційні права на пошук, отримання, передавання та використання інформації у всіх цивілізованих державах. При цьому слід зауважити, що в ряді країн законодавство, яке обмежує діяльність зі збору та обробки інформації, фактично ставить конкурентну розвідку поза законом.

Водночас в Україні «кожен має право вільно збирати, зберігати, використовувати та поширювати інформацію усно, письмово або будь-яким іншим способом – на власний розсуд» (Конституція України, розділ 2, ст. 34).

Таким чином, в Україні правове регулювання в інформаційній сфері, до якої, безумовно, належить і конкурентна розвідка, ґрунтується на таких принципах:

1. свобода пошуку, отримання, передавання, виробництва та поширення інформації будь-яким законним способом;
2. встановлення обмежень доступу до інформації лише законами держави;
3. відкритість інформації про діяльність державних органів та органів місцевого самоврядування і вільний доступ до такої інформації, окрім випадків, передбачених законами держави;
4. за категорією доступу інформація поділяється на відкриту (загальнодоступну) та з обмеженим доступом. Водночас інформація з обмеженим доступом за своєю правовою природою також поділяється на два види: відомості, що становлять державну таємницю; конфіденційну інформацію.

Незважаючи на те, що конкурентна розвідка сьогодні є визнаною сферою діяльності, законодавчо закріпленого поняття «конкурентна розвідка» в Україні наразі не існує, хоча діяльність зі збору, зберігання, обробки та поширення інформації регулюється низкою законодавчих та нормативних актів:

Закон України «Про інформацію» від 02.10.1992 р. № 2657-XII (зі змінами від 13.01.2011 р.), ст. 5-7 .

Закон України «Про друковані засоби масової інформації (пресу) в Україні» від 16.11.1992 р № 2782-XII, ст. 6, 25.

Закон України «Про охоронну діяльність» № 4616-VI від 22.03.2012 р. ст. 9, 13, 19.

Закон України «Про захист персональних даних» № 2297-VI від 01.06.2010 р.

Цивільний кодекс України (ст. 505), Кримінальний кодекс України (ст. 231, 232), Кодекс України про адміністративні правопорушення (ст.163, ст.163);

Указ Президента України «Питання європейської та євроатлантичної інтеграції» від 20.04.2019 р. № 155/2019.

Указ Президента України «Про Національний Координаційній центр кібербезпеки» від 07.06.2016 р. № 242/2016.

Не можна забувати, що проведення заходів із забезпечення безпеки бізнесу навіть у рамках конкурентної розвідки інколи може сприйматися як здійснення оперативно-розшукової діяльності, проводити яку, згідно із Законом України «Про оперативно-розшукову діяльність» від 18.02.1992 р. № 2135-XII, мають право лише суб'єкти, зазначені в окремих статтях цього Закону. При цьому перелік суб'єктів є вичерпним, а здійснення оперативно-розшукової діяльності іншими юридичними та фізичними особами забороняється.

В затвердженій Указом Президента України №96/2016 від 27 січня 2016 року Стратегії кібербезпеки України декларуються основні завдання для силових органів, серед яких: «на розвідувальні органи України – реалізація розвідувальної діяльності з виявлення загроз національній безпеці України в кіберпросторі, інших подій та обставин, що стосуються сфери кібербезпеки», а також передбачено «створення системи своєчасного виявлення, протидії та нейтралізації кіберзагроз, зокрема за залучення волонтерських організацій», усе це, безумовно, відноситься до застосування засобів OSINT (або конкурентної розвідки) у цій сфері.

Водночас чинним Кримінальним кодексом України передбачена кримінальна відповідальність за незаконний збір з метою використання або використання відомостей, що становлять комерційну таємницю, а також за розголошення комерційної

таємниці. Очевидно, такі відомості виходять за межі конкурентної розвідки.

В Україні наразі триває поступова гармонізація законодавства з Регламентом ЄС 2016/679 (General Data Protection Regulation (EU) 2016/679) щодо захисту персональних даних. Основна мета: досягти рівня сумісності із GDPR, щоб українські компанії та органи влади могли легально обробляти дані громадян ЄС та уникати санкцій при співпраці з європейськими партнерами.

Основні моменти:

1. Закон України «Про захист персональних даних» (2010) зараз модернізується, щоб відповідати вимогам GDPR. Зокрема:
   - уточнюється правова основа обробки персональних даних;
   - вводяться нові права суб'єктів даних, зокрема право на перенесення даних, право на забуття;
   - посилюється обов'язок контролерів і процесорів даних щодо безпеки, ведення реєстрів і повідомлення про витоки даних;
   - запроваджуються чіткі штрафи та механізми контролю за порушеннями.
2. Уряд України та Мінцифри розробляють проект нового закону, який безпосередньо імплементує GDPR у національне законодавство, з урахуванням національних особливостей. Він передбачає:
   - узгодження термінів і визначень із GDPR;
   - розширення повноважень Уповноваженого із захисту персональних даних;
   - визначення чітких правил для трансферу персональних даних за кордон;
   - механізми для бізнесу та органів влади щодо відповідності GDPR.
3. Поступова імплементація через практику та рекомендації:
   - багато компаній уже адаптують внутрішні політики та документи під GDPR;

- державні органи впроваджують стандарти безпеки та процедури для захисту персональних даних, особливо в сфері електронного урядування та е-демократії.

При достатньо широкому тлумаченні норм законодавства будь-які процедури збору, обробки та зберігання інформації про конкурентів стають, з одного боку, легітимними, практично безкарними, а з іншого – ускладненими. В Україні фактично закрито доступ до великого обсягу бізнес-інформації, яка у більшості країн є вільнодоступною, наприклад, щодо нерухомості (наявної та закладеної), земельних ділянок, наявності банківських рахунків тощо. У цих країнах більшість відомостей можна отримати лише шляхом консультацій із відповідними експертами.

Сьогодні, як ніколи, гостро стоїть проблема криміналізації окремих служб конкурентної розвідки. Багато служб безпеки нині користуються базами даних із інформацією про осіб. Такі бази використовуються з цілком благими цілями, наприклад, для перевірки даних про співробітників, партнерів та конкурентів. Очевидно, такими базами даних вони користуватимуться і надалі, проте будуть змушені порушувати закон і «йти в підпілля». Технічно можливості використання та ведення подібних баз даних забезпечують численні системи типу Cronos (оболонки, що розповсюджуються цілком легально). За допомогою таких інструментальних засобів будь-якому зацікавленому користувачеві мережі Інтернет стають доступні численні бази даних, що працюють із цими оболонками.

В результаті діяльність компаній, які займаються конкурентною розвідкою, привертає підвищену увагу з боку державних контролюючих органів.

Це пов'язано з кількома групами правових проблем, які можна згрупувати, виділивши проблеми, пов'язані з:

- захистом комерційної таємниці;
- захистом персональних даних;
- дотриманням авторських прав;
- можливістю конкуренції на ринку самої конкурентної розвідки.

- Також можна виокремити три класи основних проблем авторського права, що стосуються конкурентної розвідки: це проблеми, пов'язані з такими аспектами:
- правомірністю використання вхідної інформації (джерел інформації), на підставі якої формуються звіти – результати конкурентної розвідки;
- авторськими правами на результати конкурентної розвідки;
- правами на застосування (використання) спеціалізованого програмного забезпечення, необхідного для проведення конкурентної розвідки.

Крім того, одна з проблем, що постає перед службами конкурентної розвідки в Україні – практично повна відсутність антидемпінгового законодавства. Незважаючи на те, що вихід на цей ринок великих міжнародних гравців ускладнений через відсутність необхідних зв'язків, баз даних, архівів і навіть лінгвістичної та правової підготовки, з їхнього боку можливий прояв демпінгу на послуги конкурентної розвідки.

Ситуація може змінитися, якщо буде створено чітку правову базу для діяльності, пов'язаної зі збором та аналітичною обробкою інформації, і, зокрема, для конкурентної розвідки.

7.2. Конкурентна розвідка та захист комерційної таємниці

Важливе значення для становлення конкурентної розвідки мав ряд статей Закону України «Про захист від недобросовісної конкуренції» № 236/96-ВР від 07.06.1996, де (ст. 15-1) забороняється «Неправомірний збір комерційної інформації», «Розголошення комерційної інформації», «Неправомірне використання комерційної інформації» (гл. 4, ст. 16, 17, 19, відповідно).

У постанові Кабінету Міністрів України від 9 серпня 1993 року № 611 «Про перелік відомостей, що не становлять комерційної таємниці» визначено цілий клас документів, що стосуються діяльності бізнес-структур, які є фактично відкритими, зокрема, установчі документи, форми звітності, інформація про участь засновників та посадових осіб в інших компаніях тощо.

Часто зусилля конкурентної розвідки спрямовані на отримання комерційної таємниці конкурентів. І хоча в різних законодавчих актах даються різні формулювання, можна пого-

дитися з тим[85], що комерційна таємниця характеризується такою сукупністю ознак: інформація є секретною, є невідомою та не є легкодоступною для осіб, які зазвичай мають справу з видом інформації, до якого вона належить; у зв'язку з тим, що є секретною, вона має комерційну цінність. Таким чином, комерційна таємниця – це інформація, яка є корисною і не є загальновідомою суспільству. Вона має дійсну або комерційну цінність, з якої можна мати прибуток і для захисту якої власник вживає заходів у всіх сферах життя та діяльності». Отже, можна сказати, що діяльність конкурентної розвідки іноді спрямована на добування інформації, яка не є загальнодоступною і охороняється законом. Ці діяння порушують величезну кількість статей Кримінального кодексу України, зокрема, статтю 231 «Незаконний збір з метою використання або використання відомостей, що становлять комерційну або банківську таємницю».

Таким чином, комерційна розвідка може легітимно використовувати лише ті методи та способи збору й обробки інформації, які не суперечать законодавству, тобто основні функції конкурентної розвідки – якісний збір, систематизація і, головне, аналіз інформації, а не стеження, підкупи та незаконні хакерські злами.

Вперше право на збереження комерційної таємниці було проголошено Законом СРСР від 4 червня 1990 р. «Про підприємства в СРСР». У ст. 33 зазначеного Закону розкривалося поняття комерційної таємниці як відомостей, що не є державними секретами, пов'язаних з виробництвом, технологічною інформацією, управлінням, фінансами та іншою діяльністю підприємств, розголошення (передача, витік) яких може завдати шкоди їхнім інтересам.

Наразі українське законодавство про охорону службової та комерційної таємниці являє собою сукупність статей, які містяться в різних правових актах, присвячених загалом регулюванню інших суспільних відносин.

Цивільний кодекс України, у свою чергу, визначає комерційну таємницю (ст. 505 п. 1) як інформацію, «яка є сек-

[85] Основи методики розслідування незаконного збирання та розголошення комерційної таємниці // Юридичний журнал, 2006. – № 8. – С. 48-66.

ретною в тому розумінні, що вона в цілому чи в певній формі та сукупності є невідомою та не є легкодоступною для осіб, які звичайно мають справу з видом інформації, до якого вона належить, у зв'язку з цим має комерційну цінність та була предметом адекватних наявним обставинам заходів щодо збереження її секретності, вжитих особою, яка законно контролює цю інформацію».

Відповідно до цих визначень, щойно інформація в результаті якихось дій потрапляє, наприклад, на сторінки будь-якого веб-сайту, вона перестає вважатися комерційною таємницею, оскільки стає легкодоступною.

Хоча в багатьох статтях Кримінального кодексу України (ст. 231, 232, 232-1, 361, 363) встановлено кримінальну відповідальність як за розголошення комерційної таємниці, так і за незаконний збір та використання відомостей, що до неї належать, однак, існуюча нормативно-правова база чітко не регламентує, які саме відомості про фінансово-господарську діяльність підприємства є комерційною таємницею (за винятком хіба що банківської таємниці, визначення якої надано в ст. 60 Закону України «Про банки і банківську діяльність»).

7.3. Конкурентна розвідка і захист персональних даних

Державні установи, банки, великі корпорації не завжди можуть забезпечити захист баз персональних даних, які в них зберігаються, в результаті чого величезний потік конфіденційної інформації надходить на ринок. Забезпечення безпеки персональних даних – об'єктивна потреба. Сьогодні персональні дані, інформація про людей перетворюється на найдорожчий товар. Така інформація в руках зловмисника – потужна зброя. Тобто персональні дані необхідно захищати.

Персональні дані – важлива складова більш широкого поняття – приватність. Тому захист персональних даних є складовою частиною забезпечення приватності. Приватність, поряд зі свободою слова та іншими правами, є однією з основних цінностей людства.

На сьогодні, основними європейськими документами у сфері захисту персональних даних є Конвенція Ради Європи «Про захист осіб у зв'язку з автоматизованою обробкою персональних даних» та Директива Європарламенту «Про захист

фізичних осіб при автоматизованій обробці персональних даних», ETS № 108, 1981 р., яка є обов'язковою для всіх держав-членів Європейського Союзу і яка є взірцем для наслідування у сфері законодавства, в тому числі, й нашою країною. Країни Євросоюзу послідовно приводять своє законодавство у відповідність до Директиви. У Великобританії ще в 1998 році був прийнятий «Закон про захист персональних даних» – «Data Protection Act 1998». Його технічна реалізація – проект стандарту «Specification for the management of personal information in compliance with the Data Protection Act 1998» (BS 10012, 2009). Паралельно з англійцями свою версію стандарту з безпеки персональних даних випустили в США. Проєкт документа із захисту персональних даних для американських державних структур – «Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)» (SP 800122) регламентує виконання Законів «The Privacy Act of 1974» та «Privacy Protection Act of 1980». Канада випустила «Privacy Code» – набір документів для реалізації законодавства із захисту відомостей про приватних осіб (The Privacy Act та PIPEDA).

У державах-членах Євросоюзу визначення персональних даних, як правило, максимально широкі, в результаті чого громадянами на практиці часто не виконується відповідне законодавство через надмірне «навантаження». Відповідні органи державної влади, як правило, не вживають жодних дій, крім особливих випадків. Важливими залишаються питання виникнення колізій між вимогами приватності та інтересами свободи слова. Сучасними європейськими законами, як правило, забороняється збір, зберігання, використання та поширення без згоди суб'єкта даних саме критичних персональних даних.

Право на приватність гарантується Конституцією України. Стаття 32 Конституції України гласить: «Ніхто не може зазнавати втручання в його особисте і сімейне життя, крім випадків, передбачених Конституцією України». Крім того, в Конституції України передбачено захист ще деяких аспектів приватності. Так, стаття 30 захищає недоторканність житла (територіальна приватність), стаття 31 – таємницю листування, телефонних розмов, телеграфної та іншої кореспонденції (комунікаційна приватність), стаття 32 передбачає заборону збору, зберігання, використання та поширення конфіденційної інформації про особу без її згоди (інформаційна приватність), а

стаття 28 передбачає заборону піддавати особу без її вільної згоди медичним, науковим чи іншим дослідженням (захищаючи деякі елементи фізичної приватності).

Конвенція про захист осіб у зв'язку з автоматизованою обробкою персональних даних Страсбург, 28 січня 1981 (ратифікація від 06.07.2010) визначає положення щодо передачі через національні кордони за допомогою будь-яких засобів персональних даних, що піддаються автоматизованій обробці або зібраних з метою їх автоматизованої обробки.

Наступні дані часто використовуються для ідентифікації конкретної особи, зазначені як особисті Управлінням США з менеджменту та бюджету:

- повне ім'я (мається на увазі ім'я разом із прізвищем)
- національний ідентифікаційний номер;
- IP-адреса (у деяких випадках);
- номерний знак транспортного засобу;
- номер водійських прав;
- обличчя, відбитки пальців або почерк;
- номери кредитних карток;
- цифрова ідентичність (цифровий підпис);
- дата народження;
- місце народження;
- генетична інформація.

Згідно із законодавством більшості європейських держав персональні дані поділяються за критерієм «чутливості» на дані загальні та «чутливі» (вразливі) особисті дані.

Загальні особисті дані:

- ідентифікаційні дані (прізвище, ім'я, по батькові, адреса, телефон тощо);
- паспортні дані;
- особисті відомості (вік, стать, сімейний стан тощо);
- склад сім'ї;
- освіта;
- професія;
- житлові умови;
- спосіб життя;
- життєві інтереси та захоплення;

- споживчі звички;
- фінансова інформація.

«Чутливі» особисті дані:

- інформація про расове, етнічне походження та національність;
- відомості, що стосуються політичних, світоглядних і релігійних переконань;
- відомості про членство в політичних партіях, профспілках, релігійних або громадських організаціях;
- відомості про стан здоров'я та статеве життя;
- генетичні та біометричні дані;
- місце знаходження та шляхи пересування особи;
- інформація про застосування до особи заходів у рамках трудового слідства;
- інформація про вчинення щодо особи різних видів насильства.

Конституційні норми визначають вичерпний перелік підстав для втручання в приватність та умов для такого втручання. Однак у пострадянських державах існує багато галузевих норм права, що суперечать вимогам їхніх Конституцій. Саме такі норми не відповідають міжнародним стандартам, практиці Європейського законодавства.

Відповідно до українського законодавства персональними даними в Україні є П.І.Б. у супроводі будь-яких інших ідентифікаційних даних, наприклад, адреси, телефону або освітнього статусу.

Для з'ясування, яке ж відношення має фізична особа або компанія до захисту персональних даних, велике значення має визначення суб'єктів відносин, пов'язаних із персональними даними (стаття 4 Закону України № 2297-VI): «Суб'єктами відносин, пов'язаних із персональними даними, є:

- суб'єкт персональних даних;
- володілець бази персональних даних;
- розпорядник бази персональних даних;
- третя особа;

- уповноважений державний орган з питань захисту персональних даних;
- інші органи державної влади та органи місцевого самоврядування, до повноважень яких належить здійснення захисту персональних даних.

В українському законодавстві передбачено повідомний характер обробки персональних даних. Володілець або розпорядник (оператор) до початку обробки персональних даних зобов'язаний повідомити уповноважений орган із захисту прав суб'єктів персональних даних про свій намір здійснювати обробку персональних даних. Потім дані про володільців або розпорядників (операторів) вносяться до спеціального реєстру операторів. Інформація, що міститься в реєстрі операторів, стає загальнодоступною.

Закони про персональні дані стосуються більшості населення як учасників процесу «обробки» даних. А оскільки суб'єктом персональних даних є кожна людина, то Закон має загальний характер і стосується кожного.

Цей законодавчий акт має пряме відношення до сфери інформаційних технологій та телекомунікацій, обидва містять спірні, такі, що суперечать усталеній практиці, здавалося б, нездійсненні положення. Вимоги закону поширюються на всі юридичні та фізичні особи, і інтернет-сфера не є винятком. Закон про захист персональних даних може змінити принципи роботи українських інтернет-ресурсів: сервісів електронної пошти, знайомств, онлайн-магазинів та соціальних мереж, хоча самі учасники ринку сподіваються, що сайти не підпадуть під дію закону. Власникам інтернет-ресурсів для дотримання всіх положень закону про персональні дані необхідно ретельно продумувати організацію своєї діяльності. В даний час існує чимало веб-служб, у рамках яких відбувається збір, зберігання, використання персональних даних. Дотримання вимог закону є непростим завданням для власників цих ресурсів, зокрема, чиновники мають можливість зобов'язати інтернет-компанії брати письмову згоду на використання анкетних даних у кожного користувача. Не секрет, що на багатьох сайтах розміщується інформація, що містить персональні дані людей (наприклад, тих, хто шукає роботу, знайомства), у тому числі й ті, що належать до спеціальних категорій, наприклад, націо-

нальність або віросповідання. Завдання тих, хто забезпечує подібні сервіси, легітимно обробляти подібну інформацію й одночасно захищати її згідно з вимогами законодавства.

Зокрема, персональні дані широко використовуються в соціальних мережах і сервісах електронної пошти. Наприклад, власникам веб-ресурсів досить складно дотриматися вимоги закону про отримання згоди кожного користувача на обробку його персональних даних. При цьому закон покладає саме на оператора обов'язок доведення факту отримання ним такої згоди.

Сучасна інтернет-компанія збирає та обробляє різні категорії персональних даних – своїх співробітників, своїх контрагентів за договорами та деякі дані користувачів своїх сервісів. Люди, які розміщують інформацію про себе в соціальних мережах або службах знайомств, свідомо роблять її відкритою для всіх користувачів ресурсу, і за законом її можна трактувати як «загальнодоступну», а отже, дотримання особливого режиму конфіденційності щодо неї не потрібно, але в соціальних мережах є й інформація, яку користувач приховує, роблячи її доступною тільки для окремої групи користувачів («друзів»). У цьому випадку інтернет-ресурс має передбачати для неї спеціальні засоби захисту.

У практиці конкурентної розвідки доводиться стикатися з численними суперечностями та казусами в чинному законодавстві, наприклад, в українському Законі «Про захист персональних даних» (частина 9 ст. 6) йдеться: «використання персональних даних в історичних, статистичних або наукових цілях може здійснюватися тільки в знеособленому вигляді». Тобто записи у звітах конкурентної розвідки мають виглядати приблизно так: «Особа А провела переговори з особою Б». У наукових звітах не можна робити посилань на інших колег, навіть за наявності їхньої письмової згоди. Викликає певні складнощі й необхідність сповіщати орган влади «про кожну зміну відомостей, необхідних для реєстрації відповідної бази», яка серед іншого включає інформацію про всіх розпорядників (користувачів) такої бази даних.

Крім того, багато служб конкурентної розвідки, які цілком на законних умовах створюють базу даних персональних даних

для вирішення поставленого ними завдання, зобов'язані знищити плоди своєї роботи, досягнувши мети. Адже, якщо основна мета, наприклад, при наданні послуг клієнтам – це виконання цих самих заявок, то супутня мета будь-якої організації, що себе поважає – це й напрацювання бази клієнтів. І ця база часто має власне комерційне значення. Відомі численні випадки легального перепродажу баз даних клієнтів, наприклад, при припиненні діяльності фірми-власника. В українському законодавстві суворої статті немає, проте передбачені умови знищення персональних даних, серед яких (ст. 15), «припинення правовідносин між суб'єктом персональних даних та володільцем або розпорядником бази...». А це означає, наприклад, що оператор – виконавець послуги повинен знищити всю напрацьовану за час виконання послуги базу даних.

Тому володільці та розпорядники подібних баз даних переформульовують свої цілі спеціальним чином, наприклад, як «надання послуги з можливістю зберігання персональних даних протягом гарантійного строку...». Таким чином, дотримуються норми законодавства та забезпечуються інтереси виконавця – володільця або розпорядника (оператора) бази персональних даних.

Підрозділи конкурентної розвідки займаються обробкою персональних даних, які знаходяться у відкритих джерелах в мережі Інтернет, тобто є загальнодоступними. Для їх обробки згоди суб'єкта персональних даних не потрібно. Однак при цьому обов'язок доведення, що оброблювані персональні дані є загальнодоступними, покладається на володільця або розпорядника. А це означає, що необхідно або накопичувати докази того, що дані взяті із загальнодоступних джерел, або отримувати згоду від суб'єкта персональних даних і потім зберігати цей документ. Крім того, потрібно мати документ, що підтверджує загальнодоступність джерела персональних даних. При цьому залишається без відповіді питання доказу того, що володілець інформаційного ресурсу (веб-сайту) має письмову згоду на обробку.

Як ніколи гострою стала проблема криміналізації окремих служб конкурентної розвідки. Багато служб безпеки сьогодні користуються базами даних з інформацією про осіб. Такі бази використовуються з цілком благими цілями, наприклад, для

перевірки даних про співробітників, партнерів і конкурентів. Очевидно, такими базами даних вони користуватимуться й надалі, однак будуть змушені порушувати закон і «йти в підпілля». Технічно можливості використання та ведення подібних баз даних надають численні системи типу Cronos (оболонки, що поширюються цілком легально). За допомогою подібних інструментальних засобів будь-якому зацікавленому користувачеві Інтернету стають доступні численні бази даних, які працюють під цими оболонками.

На державному рівні в США основним правовим механізмом ведення розвідки у відкритих джерелах міністерства оборони є Рада із захисту відкритих джерел (DOSC). Вона слугує форумом для координації та сприяння заходам і програмам ведення розвідки у відкритих джерелах для всіх служб і бойових командувань. Ця рада консультує та доповідає заступнику міністра оборони з розвідки про питання ведення розвідки у відкритих джерелах, про нові ініціативи щодо покращення ефективності роботи підрозділу OSINT та діяльності міністерства оборони в цілому. До обов'язків Ради входять:

координує діяльність підрозділу OSINT та затверджує його план ведення розвідки у відкритих джерелах;

визначає послідовність вимог до процесу ведення розвідки у відкритих джерелах.

Армійський стандарт США «ATP 2-22.9» встановлює загальні поняття, основні концепції та методи збору розвідувальних даних із відкритих джерел для Армії США. У цьому документі підкреслюється характеристика OSINT як розвідувальної дисципліни, його зв'язки з іншими розвідувальними дисциплінами та можливості його застосування під час об'єднаних операцій.

Використання загальнодоступної інформації є важливим аспектом технічної розвідки (TECHINT). Незважаючи на те, що наміри, можливості та фактори уразливості противників і потенційних загроз підлягають засекречуванню, результати OSINT (зокрема, відкритого сервісу «Google Earth») сприяють отриманню інформації про найпотаємніші держави та організації. Такі приклади свідчать про відповідальність діяльності в цій області.

Авторське право є однією з форм захисту опублікованих і неопублікованих робіт, передбачених главою 17 Кодексу США, що визначає авторів «оригінальних робіт авторів», у тому числі літературних, драматичних, музичних і художніх творів.

Національні закони про авторські права є обмеженнями конкурентної розвідки. Порушення прав, зокрема, передбачених главою 17 Кодексу США, законами про авторські права, все ж залишають можливість правомірного використання конкурентної розвідки, що визначається чотирма факторами:

- метою та характером використання;
- властивостями використовуваних авторських робіт;
- кількістю та частинами авторської роботи, які використовуються;

впливом використання авторських робіт на потенційний ринок або цінність цих робіт.

7.4. Конкурентна розвідка і захист авторського права

Можна виокремити три класи основних проблем авторського права, що стосуються конкурентної розвідки: це проблеми, пов'язані з такими аспектами:

- правомірністю використання вхідної інформації (джерел інформації), на підставі якої формуються звіти – результати конкурентної розвідки;
- проблеми з авторськими правами на результати конкурентної розвідки;
- права на застосування (використання) спеціалізованого програмного забезпечення, необхідного для проведення конкурентної розвідки.

Крім того, одна з проблем, що стоять перед службами конкурентної розвідки в Україні – практично повна відсутність антидемпінгового законодавства:

- ситуація може змінитися, якщо буде створено чітку правову базу для конкурентної розвідки;

- авторське право є однією з форм захисту опублікованих і неопублікованих робіт, передбачених розділом 17 Кодексу США, що визначає авторів «оригінальних авторських робіт», у тому числі літературних, драматичних, музичних та художніх творів. Національні закони про авторські права є обмеженнями конкурентної розвідки. Незважаючи на це, все ж таки залишаються можливості правомірного використання конкурентної розвідки, що визначається чотирма факторами:
- метою та характером використання;
- властивостями авторських робіт;
- обсягом та частинами авторської роботи;
- впливом використання авторських робіт на потенційний ринок або цінність цих робіт.

Висновки до розділу 7

Проведений у сьомому розділі комплексний аналіз правових засад функціонування конкурентної розвідки дозволяє зробити висновок, що дана сфера діяльності невіддільна від загального правового поля держави та має базуватися на конституційних гарантіях свободи пошуку, отримання та поширення інформації. Водночас виявлено суттєву прогалину в національному законодавстві, що полягає у відсутності законодавчо закріпленого поняття конкурентної розвідки, що призводить до правової невизначеності та створює передумови для потенційної криміналізації діяльності спеціалізованих служб, які оперують базами даних та відкритими джерелами. Ключовим критерієм легітимності виступає використання виключно відкритих джерел інформації та методів, що не суперечать чинним нормам, чим конкурентна розвідка принципово відрізняється від промислового шпионажу, спрямованого на незаконне заволодіння відомостями, що становлять комерційну таємницю. Особливого значення набувають питання дотримання законодавства про захист персональних даних, де імплементація європейських стандартів та вимоги національних законів накладають суворі обмеження на обробку інформації про фізичних осіб, вимагаючи від аналітиків ретельної верифікації статусу даних як общедоступних або отримання відповідної згоди суб'єктів. Паралельно існують ризики пору-

шення авторських прав при використанні вхідної інформації та інструментарію, що вимагає дотримання балансу між інтересами правовласників та правомірним використанням результатів інтелектуальної праці в аналітичних звітах. Загалом, подальший розвиток інституту конкурентної розвідки в Україні стримується відсутністю чіткого антидемпінгового законодавства та спеціалізованої нормативної бази, що регулює збір та аналітичну обробку бізнес-інформації, тому необхідним кроком є формування правового механізму, який би легалізував професійну діяльність у цій сфері, усунув суперечності між захистом конфіденційності та потребами бізнесу в інформаційній безпеці, а також запобіг тінізації ринку послуг конкурентної розвідки.

8. Протидія інформаційним операціям

Останніми роками завдяки численним документам та публікаціям Міністерства оборони США набув популярності термін «інформаційні операції», насамперед тому, що інформаційні технології відіграють постійно зростаючу роль у військових операціях. При цьому інформаційні операції визначаються як «дії, спрямовані на вплив на інформацію та інформаційні системи противника, а також на захист власної інформації та інформаційних систем»[86]. Інформаційні операції розглядаються як поєднання основних можливостей радіоелектронної боротьби, комп'ютерних мережевих операцій, психологічних операцій, військових дій та операцій із забезпечення безпеки з метою впливати, руйнувати, спотворювати інформацію, необхідну для прийняття противником рішень, а також захищати власну інформацію. Інформаційні операції охоплюють цілий комплекс процесів, що здійснюються в найрізноманітніших сферах. При цьому слід зазначити, що інформаційні операції є суттєвою та традиційною складовою бойових операцій. Незважаючи на те, що формальне визначення в документах Департаменту оборони США орієнтоване на військові аспекти інформаційних операцій, воно цілком застосовне практично до будь-якої сфери життя.

Нижче розглядатимуться такі інформаційні операції, які реалізуються за допомогою інформаційних систем (ІС). Живучість цих ІС значною мірою визначає живучість інформаційних операцій, які реалізуються у вигляді інформаційних впливів на свідомість людей.

Інформація є відображенням вкладеного в неї змісту, тому сьогодні інформація перетворилася з абстрактного терміна на об'єкт, мету та засіб інформаційних операцій, стала критичним поняттям у проблематиці безпеки. Колишній міністр оборони США Вільям Коен 18 березня 1999 року заявив, що «здатність армії використовувати інформацію для домінування у май-

[86] Information operations roadmap – DoD US, 30 october 2003. – 78 p.

бутніх битвах дасть США новий ключ до перемог протягом багатьох років, якщо не протягом кількох поколінь»[87].

Під час моделювання та проведення інформаційних операцій необхідно враховувати значення цінності інформації для осіб, що приймають рішення. Цінність інформації включає її своєчасність, точність та «аналітичність». З практичної точки зору цінність інформації також може бути визначена як її значущість або придатність до використання. Під придатністю інформації розуміється забезпечення доступу ОПР до готової до використання інформації. Стандарт ISO 9241 визначає придатність у термінах ефективності та задоволення потреб визначеного набору користувачів для вирішення визначеного набору завдань у специфічному середовищі. На практиці більша частина корисної інформації надходить до ОПР від інформаційно-аналітичних систем, що забезпечують орієнтацію в ситуації та підтримку при прийнятті рішень. Згідно з польовим статутом військового відомства США «Інформаційні операції» (FM 100-6)[88], «орієнтація в ситуації означає комбінацію чіткого уявлення про диспозицію власних та ворожих сил з оцінкою ситуації та намірів з боку командування».

Інформаційні операції здійснюються в певному соціальному середовищі, відповідно, для успішного їх проведення необхідно адаптуватися до цього середовища, подолати певний бар'єр недостатньої уваги до інформаційного впливу. Цей бар'єр виникає завдяки так званій імунній системі середовища, яка може не пропустити інформаційні впливи, якщо вона достатньо потужна та/або вже навчилася захищатися від подібних впливів. До підготовчих дій для проведення інформаційних операцій може належати створення «імунодефіциту» соціального середовища шляхом впливу через інформаційний простір, наприклад, за допомогою матеріалів у ЗМІ. Дуже часто інформаційні впливи використовують механізми «вірусного маркетингу», наприклад, у вигляді чуток, коли сенсаційно подана дезінформація поширюється з величезною швидкістю.

---

[87] Hill J.M.D., Surdu J.R., Ragsdale D.J., Schafer, J.H. Anticipatory planning in information operations // Systems, Man, and Cybernetics, 2000 IEEE International Conference, 2000. – 4. – P. 2350-2355.

[88] Army, U. S. "FM 100-6 information operations." (1996).

Саме імунна система чинить опір подібним інформаційним операціям. Дуже часто імунну систему суспільства ототожнюють з державою, покликаною забезпечувати безпеку цього суспільства, тобто за наявності сильного державного апарату ймовірність успіху антигромадських інформаційних операцій суттєво знижується. Читач чудово знає, як відбувалася протидія подібним інформаційним процесам у тоталітарних державах. У демократичному суспільстві, звичайно, тоталітарні методи не застосовні. У цьому випадку імунітет досягається за рахунок «навчання», тобто демократичне суспільство має пройти через багато інформаційних атак, впливів, впливу стереотипів, щоб виробити необхідний імунітет.

Рівень готовності до проведення інформаційних операцій сьогодні вважається ключовим фактором успішного проведення будь-якої соціальної процедури або кампанії. Особливою метою під час проведення інформаційних операцій є інформаційно-аналітичні системи об'єкта впливу. Впливаючи на такі системи, можна досягти того, що особи, які приймають рішення з табору противника, зроблять неадекватні висновки, і необхідний соціальний процес змінить траєкторію у напрямку, необхідному для сторони, що здійснює вплив[89] (рис. 63).

У цьому випадку до безпосередніх інформаційних впливів може бути віднесено розміщення в інформаційному просторі документів, що компрометують супротивника, реклама (у тому числі прихована) власних переваг, спотворені дані про зовнішнє середовище, спотворена інформація про наміри тощо.

Соціальні процедури та процеси, як правило, складно оцінювати та моделювати, оскільки їхні результати належать до психологічних і соціологічних, а не фізичних. Саме цей факт також визначає проблематичність прогнозування результатів моделювання інформаційних операцій. Крім того, експеримен-

---

[89] Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія / В.П. Горбулін, О.Г. Додонов, Д.В. Ланде. – Київ: Інтертехнологія, 2009. - 164 с. - Бібліогр.: с. 153-162. ISBN 978-966-1648-12-7

тування з інформаційними впливами в рамках інформаційних операцій є більш складним і небезпечним, ніж під час моделювання фізичних процесів.

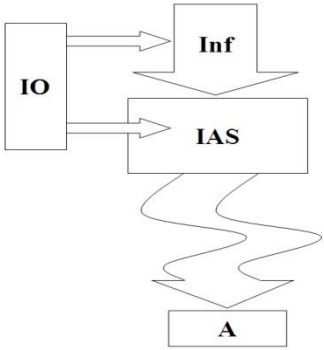


Рис. 63 – Вплив на інформаційно-аналітичну систему супротивника: Inf – інформаційний простір; IAS – інформаційно-аналітична система; A – абонент системи – ОПР; IO –інформаційні впливи

Дії для досягнення ефективності впливу на процеси прийняття рішень супротивником іноді необхідно здійснювати протягом тривалого часу, перш ніж вони набудуть чинності.

Одна з основних компонентів інформаційних операцій – соціальний вплив, що охоплює все різноманіття процесів впливу. Істотні зміни в переконаннях або ставленні людей до певної проблеми чи явища, як очікується, призведуть до змін у поведінці, пов'язаній з цією проблемою.

У 1948 році Гарольд Д. Лассвелл[90] розробив модель передачі комунікацій, що складається з п'яти компонентів:

- джерело – особа, яка впливає або переконує інші особи;
- повідомлення – засіб, за допомогою якого джерело намагається переконати цільову особу;
- ціль – особа, на яку джерело намагається впливати;
- канал – спосіб передачі повідомлень;
- ефект – реакція цільової особи на повідомлення.

[90] Lasswell H.D. The structure and function of communication in society // The Communication of Ideas. / Ed.: L. Bryson. – New York: Harper and Brothers, 1948.

Хоча Лассвелл насамперед цікавився масовою комунікацією, його модель передачі інформації може застосовуватися у міжособистісній комунікації типу циклічних моделей Шеннона – Вівера (Shannon–Weaver) та Осгуда – Шрамма (Osgood–Schramm), які включають петлі зворотного зв'язку в процесі комунікації, стверджуючи, що комунікація є циклічним, а не лінійним процесом[91,92].

Моделювання об'єктивних факторів соціального впливу вимагає міждисциплінарних підходів, що стосуються інформатики, маркетингу, політології та соціальної психології. Найвідоміші моделі формування громадської думки та соціального впливу ґрунтуються на теорії динамічного соціального впливу Латане[93,94], розвиненої багатьма іншими авторами, насамперед у працях[95,96,97,98].

Намагаючись обґрунтувати механізм соціального впливу повідомлень, Латане підкреслив важливість трьох ознак взаємовідносин джерела та об'єкта впливу:

- сила – соціальна сила, ймовірність або рівень впливу на індивідів;

---

[91] Schramm W., D.F.Roberts (eds.) The Process and Effects of Mass Communication. Univ. of Illinoise Press, 1974.

[92] Osgood Ch. E. Psycholinguistics. A Survey of Theory and Research Problems // Supplement to the International Journal of American Linguistics. Vol. 20. No 4. Oct. 1954, mem. 10. Baltimore: Waverly Press, 1954.

[93] Latane B. The psychology of social impact // American Psychologist, 1981. – 33. – P. 343-356.

[94] Latane B., Nowak A. Causes of polarization and clustering in social groups // Progress in communication sciences, 1997. – 13. – P. 43-75.

[95] Nowak A., Szamrej J., Latane B. From private attitude to public opinion: A dynamic theory of social impact // Psychological Review, 1990. – 97. – P. 367-376.

[96] Lewenstein M., Nowak A., Latane B. Statistical mechanics of social impact // Physical Review, 1993. – A, 45. – P. 763-776.

[97] Kacperski K., Holyst J.A. Physica A. Phase transitions as a persistent feature of groups with leaders in models of opinion formation // Statistical Mechanics and its Applications, 2000. – 287, Issues 3-4. – P 631-643.

[98] Sobkowicz P. Effect of leader's strategy on opinion formation in networked societies // Preprint Arxiv (on-line: http://arxiv.org/ pdf/cond-mat/0311566)

- безпосередність – фізична або психологічна відстань між індивідами;
- кількість джерел – кількість джерел, спрямованих на об'єкт.

Сучасний стан моделювання інформаційних операцій характеризується низкою невирішених проблем, основні з яких стосуються розуміння понять інформаційного впливу та ефекту.

8.1. Інформаційні впливи, атаки та операції

Універсальними характеристиками об'єктів є їхній стан та можливість впливу на інші об'єкти. Реалізація можливості впливу потребує певних умов, які зазвичай називають його впливом. При цьому об'єкт, який може здійснювати свою волю, називають суб'єктом, а управлінням прийнято називати вплив на об'єкт впливу, застосований з певною метою.

Коли індивід є ціллю впливу одного або кількох джерел, динамічна соціальна теорія впливу стверджує, що рівень соціального впливу на індивіда можна подати рівнянням, яке є основою так званої індивідуум-орієнтованої моделі:

$$I_i = -S_i\beta - \sum_{j=1, j\neq i}^{N} \frac{S_j O_j O_i}{d_{i,j}^{\alpha}},$$

де I_i – величина (кількість) соціального тиску, що чиниться на індивіда i, $(-\infty < I_i < \infty)$; O_i та O_j представляє думку індивіда (i або j, відповідно) щодо актуального питання – +1 або –1 – підтримку або заперечення щодо даного питання, відповідно. S_i (S_j) представляє силу індивіда i (j) або вплив $(S_i > 0,\ S_j > 0)$; β – опір індивіда до змін ($\beta > 0$); $d_{i,j}^{\alpha}$ – відстань між індивідами i та j ($d_{i,j}^{\alpha} \geq 1$); α – показник скорочення відстані ($\alpha \geq 2$); N – загальна кількість агентів (індивідів, що складають спільноту).

Значення β – тенденція зберігати власну думку або чинити опір змінам визначає те, що індивіди в межах моделі можуть

вимагати більших або менших обсягів соціального тиску для зміни своєї думки. Великі рівні значення α відповідають ефекту зростання відстані між джерелом і метою, що впливає на обсяг соціального тиску на ціль.

На основі введених термінів формулюється поняття «інформаційного поля об'єкта», описуються його характеристики. Це дає змогу визначити інформаційний вплив як вплив на інформаційне поле об'єкта. Досліджуючи інформаційні поля об'єктів і суб'єктів соціальних систем, можна визначати інформаційні впливи та управління. При цьому інформація може розглядатися як обєкт, так і як засіб впливу. Використання інформації як засобу впливу вимагає в процесі управління здійснити підготовку даних, виробництво відповідної інформації, а лише потім реалізовувати створену інформацію у вигляді впливу.

Одним із основних методів ведення інформаційних операцій є інформаційний вплив, що чиниться з метою інформаційного управління. Під інформаційним управлінням у даному випадку розуміється механізм керування, коли керівний вплив носить неявний, непрямий інформаційний характер, і об'єкту управління надається певна інформаційна картина, під впливом якої він формує лінію своєї поведінки. Таким чином, інформаційне управління – це спосіб впливу, що спонукає людей до впорядкованої поведінки та виконання необхідних дій.

Процес інформаційного впливу одного об'єкта на інші доцільно декомпозувати на такі етапи:

- генерація джерелом впливу даних, інформаційних елементів та інформаційних сукупностей;
- передача інформації джерелом впливу;
- прийом інформації реципієнтом;
- генерація сукупності даних, інформаційних елементів та нових сукупностей об'єкта впливу;
- відповідні активні дії об'єкта впливу.

Інформаційні впливи на елементи систем можна класифікувати за такими ознаками, як джерела виникнення, тривалість впливу, природа виникнення тощо.

Для вибору конкретних способів реалізації інформаційного управління необхідно конкретизувати завдання, що вирішуються за допомогою інформаційного впливу, провести аналіз

процесу формування інформаційних операцій і розробити критерії їх оцінки.

Інформаційне управління розглядають як процес, що охоплює три взаємопов'язані напрямки:

- управління обміном даними між реальним світом та віртуальним світом суб'єкта впливу;
- управління віртуальним світом суб'єктів впливу та механізмами прийняття рішень;
- управління процесом перетворення рішень на дії суб'єктом впливу у реальному світі.

Інформаційний вплив може мати два основні види:

- зміна даних у потрібному напрямку, які використовує інформаційно-аналітична система об'єкта впливу при прийнятті рішень;
- безпосередній вплив на процес прийняття рішень об'єкта впливу, наприклад, на процедури прийняття рішень або окремих осіб, що приймають рішення.

Важливе значення для проведення інформаційних операцій має навколишнє середовище, стан об'єктів інформаційного впливу та їх взаємний вплив. Зокрема, якщо в якості об'єктів інформаційних операцій обирається певне електоральне поле, важливо враховувати всі електоральні популяції, що входять у це поле, які представляють прихильників (або противників) тих чи інших політичних сил.

Незважаючи на те, що надалі розглядатимуться й деякі моделі, у яких явно постулюється однорідність середовища, у загальному випадку щодо інформаційних операцій навколишнє середовище може складатися з областей:

- домінуючого сприйняття;
- підвищеної чутливості;
- індиферентності до відповідних інформаційних впливів.

8.2. Етапи інформаційних операцій

Зупинимося окремо на етапності інформаційних операцій. Очевидно, не існує єдиного «стандартного» плану проведення як наступальних, так і оборонних інформаційних операцій. Можна лише розглянути приблизну послідовність дій, отриману

шляхом узагальнення деяких уже реалізованих інформаційних операцій.

На практиці інформаційна операція як процес інформаційного впливу на масову свідомість, як правило, реалізується наступним чином: в результаті попередньої розвідки формується план наступного етапу – оперативного управління, та намічаються відповідні заходи оперативної розвідки, які є наближеною моделлю рішення, після чого реалізується оперативне управління противником. На етапі оперативної розвідки визначається рівень відхилення початкової моделі від реальності, і якщо воно незначне, реалізується початковий план. В іншому випадку будується новий план оперативного управління та управління противником. Далі цикл повторюється доти, поки оперативна розвідка не підтвердить використану модель. При цьому остаточне рішення приймається з певним оперативним ризиком.

Таким чином, процес інформаційного впливу охоплює такі основні етапи[99] (Рис. 64):

- попередня розвідка (preliminary intelligence, PI);
- визначення поточної обстановки, стану противника (Ор);
- управління противником (management of enemy, M) – інформаційний вплив на противника з метою передавання йому відомостей, відповідних задуму керівника;
- оперативна розвідка (operational intelligence, OI) – перевірка результатів рефлексивного управління;
- оперативне управління (operational management, OM) – дії керівника для досягнення необхідної мети.

При плануванні або моделюванні соціальних процесів, зокрема інформаційних операцій, завжди необхідно враховувати, що загальну поведінку соціальних систем неможливо визначити, оперуючи виключно вишуканими математичними моделями. Це насамперед зумовлено тим, що такі процеси значною мірою залежать від соціально-психологічних факторів.

Виділяють два основні типи інформаційних операцій – наступальні та оборонні. Однак на практиці більшість інфор-

[99] Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.

маційних операцій є змішаними. Крім того, більшість процедур інформаційних операцій одночасно стосується як наступальних, так і оборонних дій. Кожен із типів інформаційних операцій, включно з наведеними вище основними етапами, передбачає певні особливості та уточнення.

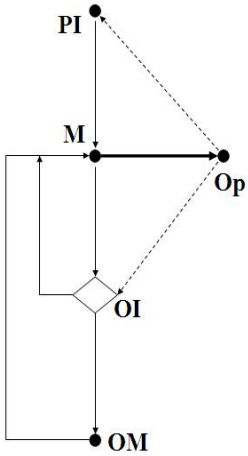


Рис. 64 – Основні етапи інформаційних операцій

Особливістю наступальних інформаційних операцій (інформаційних атак) є те, що об'єкти впливу таких операцій визначені, і планування базується на досить точній інформації про ці об'єкти. Інформаційна атака найчастіше вимагає знаходження або створення інформаційного приводу (для оборонних інформаційних операцій приводом може бути сама інформаційна атака противника), «розкрутки» цього приводу, тобто пропаганди (на відміну від заходів контрпропаганди при оборонних інформаційних операціях), а також необхідності вжиття заходів для протидії інформаційному спротиву.

Таким чином, план типової інформаційної операції включає на верхньому рівні для інформаційних операцій обох типів такі етапи, як оцінка, планування, виконання та завершальна аза. Наведемо більш детальний перелік компонентів інформаційних операцій.

У наступальних інформаційних операціях можна виділити такі основні фази:

1. Оцінка необхідності проведення операції:

- визначення мети, прогноз досяжності, ступеня впливу;
- збір інформації.

2. Планування.

3. Виконання інформаційного впливу:

- пошук або створення інформаційного приводу;
- розкрутка інформаційного приводу (пропаганда);
- оперативна розвідка;
- оцінка впливу;
- перешкоджання інформаційному протидіянню;
- коригування інформаційного впливу.

4. **Завершальна фаза:**

- аналіз ефективності;
- використання позитивних результатів інформаційного впливу;
- протидія негативним результатам.

Типова оборонна інформаційна операція охоплює такі основні етапи:

1. Оцінка:

- аналіз можливих вразливостей (цілей);
- збір інформації про можливі операції;
- визначення можливих «замовників» інформаційних впливів:
  - визначення сфер спільних інтересів об'єкта та потенційних «замовників»;
  - ранжування потенційних замовників за їхніми інтересами.

2. Планування:

- стратегічне планування оборонної операції (явне або неявне):
  - визначення критеріїв інформаційних впливів;
  - моделювання інформаційних впливів з урахуванням: зв'язків об'єкта; динаміки впливу; «особливих» (критичних) точок впливу;
  - прогнозування наступних кроків;
  - розрахунок наслідків.
- тактичне планування контроперацій.

3. Виконання – відбиття інформаційного впливу:

- виявлення та «згладжування» інформаційного приводу;
- контрпропаганда;
- оперативна розвідка;
- оцінка інформаційного середовища;
- коригування інформаційного протидіяння.

4. Завершальна фаза:

- аналіз ефективності;
- використання позитивних результатів інформаційного впливу;
- протидія негативним результатам.

Оперативне управління інформаційними операціями з використанням інформаційно-аналітичних систем можна проілюструвати за допомогою діаграми, представленої на Рис. 65.

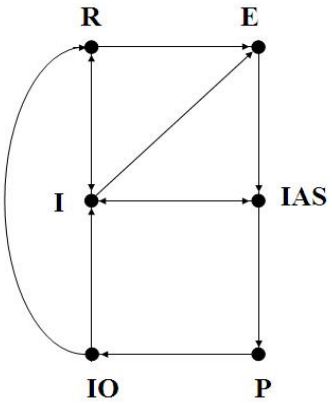


Рис. 65 – Діаграма оперативного управління з використанням інформаційно-аналітичних систем

Відповідно до наведеної діаграми, інформація з реального світу (R) надходить у інформаційний простір, зокрема до засобів масової інформації (I) або безпосередньо до експертів (E), також через засоби масової інформації.

Від експертів або безпосередньо з інформаційного простору (наприклад, за допомогою засобів контент-моніторингу) інформація надходить до інформаційно-аналітичної системи (IAC).

Інформаційно-аналітична система передає особам, які приймають рішення (Р), дані, що визначають заходи інформаційного впливу на інформаційний простір та безпосередньо на об'єкти реального світу (людей, довкілля, комп'ютерні системи тощо).

8.3. Моделювання інформаційних операцій

Моделювання можна розглядати як один зі способів розв'язання проблем, що виникають у реальному світі, зокрема під час планування та проведення інформаційних операцій. Найчастіше моделювання застосовується у випадках, коли експерименти з реальними об'єктами неможливі або є надто витратними. Моделювання охоплює відображення реальної проблеми у світ абстракції, вивчення, аналіз і оптимізацію моделі, а також перенесення оптимального рішення назад у реальний світ.

У моделюванні існують два альтернативні підходи – аналітичне та імітаційне моделювання. Ідеальні аналітичні моделі допускають суворе аналітичне розв'язання або, принаймні, постановку задачі, наприклад у вигляді систем диференціальних рівнянь. Однак аналітичні розв'язання не завжди досяжні. Тому, особливо останнім часом, і особливо під час розв'язання задач у сфері соціальної динаміки, дедалі частіше застосовуються методи імітаційного моделювання (Simulation Modeling). Імітаційне моделювання є більш потужним і практично незамінним засобом аналізу соціальних процедур. Імітаційну модель можна розглядати як множину правил, що визначають майбутній стан системи на основі її поточного стану. При цьому процес моделювання полягає у спостереженні еволюції системи в часі відповідно до цих правил і, відповідно, в оцінюванні адекватності моделі, коли це можливо.

Найперспективнішим напрямом моделювання інформаційних операцій є математичний опис самоорганізації середовища сприйняття та поширення інформації з урахуванням умов, що склалися на поточний момент. Самоорганізовані середовища, для яких відсутній центральний механізм управління, а розвиток відбувається за рахунок множини локальних взаємодій, вивчаються теорією складних систем. Ця теорія охоплює такі галузі знань, як нелінійна фізика, термодинаміка

нерівноважних процесів, теорія динамічних систем. Взаємодії між окремими елементами складних систем визначають виникнення складної поведінки за відсутності централізованого управління. Для дослідження подібної поведінки застосовуються найсучасніші методи, які охоплюються міждисциплінарною основою сучасної методології – концепцією складності. Нині до теоретичних і технологічних основ цієї концепції належать теорії детермінованого хаосу, фракталів і складних мереж, синергетика, хвильовий (вейвлет) аналіз, багатоагентне моделювання, теорія самоорганізованої критичності (яка вивчає динамічний розвиток до критичного стану, що характеризується сильними просторово-часовими флуктуаціями без зовнішнього управління [100], теорія перколяції (percolation – протікання) тощо.

Моделювання соціальних процедур (інформаційні операції, безумовно, належать до таких) передбачає проведення обчислювальних експериментів, оскільки найчастіше виникають суттєві обмеження, що ускладнюють проведення «польових» натурних експериментів.

Під час моделювання інформаційних операцій обчислювальний експеримент дає змогу скоротити операції з уточнення обмежень, добору вихідних даних, вибору правил функціонування компонентів моделі тощо. У цьому разі з'являється можливість врахування ситуацій, які складно реалізувати на практиці, використовуючи реальні дані лише для ідентифікації параметрів математичної моделі. Водночас математичне моделювання має свої обмеження: реальний світ виявляється надто складним для моделювання з достатнім рівнем деталізації та точності. Тобто більш-менш достовірні математичні моделі настільки складні й багатопараметричні, що не піддаються аналізу та оцінюванню точними методами.

Відпрацювати математичні моделі під час планування інформаційних операцій можна лише у процесі моделювання конкретних процедур, постійно співвідносячи їх із реальністю.

Чітко сформульована мета методології оцінювання інформаційних операцій полягає в тому, щоб забезпечити своєчасний і точний аналіз можливих невідповідностей між

[100] Bak P. How nature works: The science of self-organized criticality. – New York: Springer-Verlag Inc., 1996. – 212 p.

запланованою операцією та фактичним впливом. Коли виявляються суттєві відмінності, які впливають на ймовірність успіху операції, аналітична система повинна повідомляти про це особам, що приймають рішення, щоб скоригувати поточні плани та рішення. Водночас під час планування інформаційних операцій не можна діяти методом проб і помилок, тому необхідно розвивати методи, що дають змогу узагальнювати ретроспективні дані та на їхній основі перевіряти адекватність моделей.

В основу успішних моделей інформаційних операцій закладаються синергетичні підходи. Дійсно, суспільство є складною системою, кожний компонент якої характеризується множиною ознак і має багато ступенів свободи. При цьому важливою властивістю цієї системи є самоорганізація, що є результатом взаємодії таких компонентів, як випадковість, множинність, позитивний і негативний зворотний зв'язок.

Особливістю математичного моделювання інформаційних операцій слід вважати відносну простоту інтерпретації отриманих результатів. Такі поняття, як «чисельність електорату», «політична вага» тощо, сприймаються на інтуїтивному рівні навіть без ознайомлення з точними, наскільки вони тут можливі, визначеннями. Це дає змогу робити подібний аналіз актуальних ситуацій предметом широкого обговорення.

З огляду на те, що деякі рішення є нестійкими щодо своїх параметрів, значення таких параметрів необхідно визначати з високою точністю. Для цього потрібен комплекс методик, заснованих не лише на обробленні великих обсягів статистичних даних, а й на різнобічних соціологічних дослідженнях.

Нині реалістичною виглядає постановка задачі, що полягає у використанні математичних моделей для прогнозування можливих сценаріїв динаміки соціальних процесів на якісному рівні. У такому формулюванні моделювання динаміки займає ніби проміжний рівень між тим, що викладено тут, і точним прогнозуванням. І все ж необхідний вибір значень параметрів, які б у певному розумному наближенні відповідали досліджуваній ситуації, причому в більшості випадків продуктивним виявляється використання відносних величин. Звичайно, таким чином неможливо отримати достовірні дані про майбутній розвиток подій, але, найімовірніше, можна скласти

більш-менш адекватну картину того, що і як може відбутися. А це вже немало.

Для досягнення успіху окремі інформаційні впливи необхідно розглядати як частини єдиної інформаційної операції, так само як артилерійський обстріл або авіаційні атаки можна розглядати як узгоджені частини військової операції.

При цьому інформаційним операціям притаманні такі основні особливості:

- інформаційні операції – це міждисциплінарний набір методів і технологій у таких галузях, як інформатика, соціологія, психологія, міжнародні відносини, комунікації, військова наука;
- дотепер не існує стандартів проведення інформаційних операцій;
- у розвитку технологій інформаційних операцій зацікавлені не лише оборонні відомства, а й багато урядових і комерційних організацій;
- завдання формування наукового підходу до інформаційних операцій є нагальним і актуальним.

Під час проведення інформаційних операцій суттєвим є виявлення змісту (знань), вкладеного в інформацію, з урахуванням найрізноманітніших аспектів – соціальних, політичних, релігійних, історичних, економічних, психологічних, ментальних, культурних, притаманних різним верствам суспільства. Тому нині доцільно розглядати інформаційні операції ширше – як операції, що базуються на знаннях (Knowledge Operations) [101].

Звичайна мережева інформаційна атака у веб-середовищі сьогодні здійснюється таким чином: як правило, створюється веб-сайт (назвемо його «першоджерелом»), який протягом певного часу функціонує, публікуючи цілком коректну інформацію. У годину Х на його сторінці з'являється документ – зазвичай компромат на об'єкт атаки, достовірний або сфальсифікований. Після цього відбувається так зване «відмивання

[101] Burke M.M. Knowledge Operations: above and beyond Information Operations. 6th International Command and Control Research and Technology, June 19 – 21, 2001. – 16 p.

інформації». Документ передруковують інтернет-видання двох типів – ті, що зацікавлені в атаці, і ті, яким просто бракує матеріалів для заповнення власного інформаційного поля. У разі виникнення претензій усі видання, що передруковують матеріал, посилаються на «першоджерело» і, у крайньому випадку, на прохання або вимогу об'єкта атаки видаляють інформацію зі своїх веб-сайтів. Першоджерело за потреби також прибирає інформацію або взагалі ліквідується (після чого з'ясовується, що воно було зареєстроване в Інтернеті на неіснуючу особу). Водночас інформація вже поширилася, завдання першоджерела виконано, і атака розпочалася.

Сучасний інформаційний простір створює унікальну можливість отримання будь-якої інформації з обраного питання за умови наявності відповідного інструментарію. Використання такого інструментарію дає змогу аналізувати взаємозв'язки можливих подій або подій, що вже відбуваються, з інформаційною активністю певного кола джерел інформації. З іншого боку, під час ретроспективного аналізу будь-якого процесу або явища інтерес становлять певні характеристики його розвитку, а саме:

- кількісна динаміка, притаманна процесу або явищу, наприклад, кількість подій за одиницю часу або кількість повідомлень, що стосуються цього явища;
- визначення критичних, порогових точок, які відповідають кількісній динаміці явища;
- визначення проявів у критичних точках, наприклад, виявлення основних сюжетів публікацій у ЗМІ щодо обраного процесу або явища;
- після виявлення основних проявів явища в критичних точках ці прояви ранжуються, і досліджується динаміка розвитку окремих визначених проявів до та після певних критичних точок;
- здійснюється статистичний, кореляційний і фрактальний аналіз загальної динаміки та динаміки окремих проявів, на основі яких проводяться спроби прогнозування розвитку явища та окремих його проявів.

Для дослідження взаємозв'язку реальних подій і публікацій про них у мережі Інтернет авторами використовувалася система InfoStream, що забезпечує інтеграцію та моніторинг мережевих інформаційних ресурсів.

Кількість вебпублікацій на день з певної теми, а особливо зміни (динаміка) цієї величини, інколи дають змогу навіть нечисленним фахівцям у відповідній предметній галузі робити більш-менш точні висновки.

Отримати дані такої динаміки можна, наприклад, щоденно відвідуючи сайти агрегаторів новин. Звісно, у вигіднішому становищі перебувають користувачі професійних систем моніторингу, таких як X-SCIF або InfoStream.

8.4. Виявлення інформаційних операцій

Для оперативного аналізу інформаційної обстановки з метою виявлення інформаційних операцій застосовуються спеціалізовані системи моніторингу інформаційного простору (контент-моніторингу). Такі системи забезпечують, по-перше, оперативність, яку не можуть забезпечити традиційні пошукові системи (час індексації мережевого контенту навіть у найкращих із них становить від кількох діб до кількох тижнів). По-друге, повноту (як у плані джерел, так і представлення матеріалів джерел), яку не завжди забезпечують звичайні агрегатори новин. І, по-третє, необхідні аналітичні засоби, що дозволяють користувачеві створювати аналітичні звіти, які базуються на публікаціях за заданою тематикою у необхідний період часу.

У плані профілактики інформаційних операцій слід уважно стежити за динамікою публікацій про цільову компанію, якщо є можливість, з урахуванням тональності цих публікацій, користуватися доступними аналітичними засобами, наприклад, вейвлет-аналізом. При цьому слід орієнтуватися на можливі моделі інформаційних атак, наприклад, якщо така модель охоплює фази: «фонові публікації» – «затишшя» – «артпідготовка» – «затишшя» – «атака» (рис. 66), то вже за першими трьома компонентами можна з великою ймовірністю передбачити майбутні події.

Наведений вище план, очевидно, є ідеальним, орієнтованим виключно на дані контент-моніторингу веб-ресурсів.

Звичайно, у кращому становищі перебувають користувачі професійних систем контент-моніторингу. Багато сучасних інформаційно-аналітичних систем містять у своєму складі засоби відображення статистики входження до баз даних понять, що відповідають користувацьким запитам. Зокрема, авторами

використовувалася підсистема статистики в межах системи контент-моніторингу веб-простору InfoStream, яка реалізує цю функціональність.

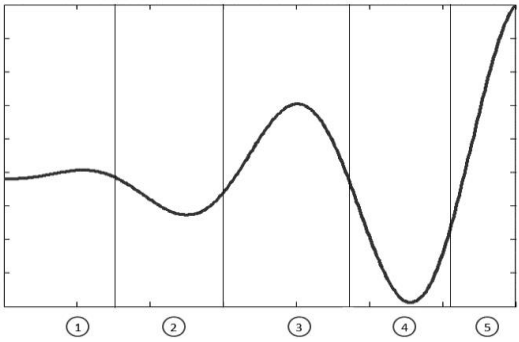


Рисунок 66 – Динаміка кількості тематичних повідомлень під час проведення інформаційної операції: 1 – фон; 2 – затишшя; 3 – «артпідготовка»; 4 – затишшя; 5 – атака / тригер зростання

Під час дослідження трендів інформаційних операцій як часові ряди розглядаються саме ряди за кількістю тематичних публікацій за певний проміжок часу (найчастіше – за добу), що відповідають цим інформаційним операціям. Тому для виявлення трендів досліджуються інформаційні потоки, що відповідають тематикам інформаційних операцій – тематичні інформаційні потоки.

Тренди повідомлень, що відповідають етапам інформаційної операції[102], подано на рис. 67. При цьому аналітикам слід орієнтуватися на такі моделі: наприклад, якщо моніторинг дозволяє визначити фази «фон» – «затишшя» – «артпідготовка» – «затишшя» – «атака», то вже за першими трьома компонентами можна з великою ймовірністю передбачити майбутні події.

Слід зазначити, що подібна динаміка кількості тематичних повідомлень під час проведення інформаційних операцій добре описується відомим рівнянням поширення електромагнітних хвиль:

[102] Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.

$$y = A + Bx\sin(x),$$

де x – час, A і B – константи, що визначаються емпірично.

Як відомо, нині інноваційна діяльність також опосередковано вимірюється кількістю публікацій, що стосуються інновацій. Існує кілька моделей інноваційних процесів, серед яких можна виділити модель дифузії інновацій [103]. Водночас упровадження інновацій також можна розглядати як інформаційні операції. Тому звернемося до результатів відповідних досліджень. На рис. 67 наведено обґрунтовану в діаграму кількості публікацій, що відповідає тренду інноваційної діяльності.

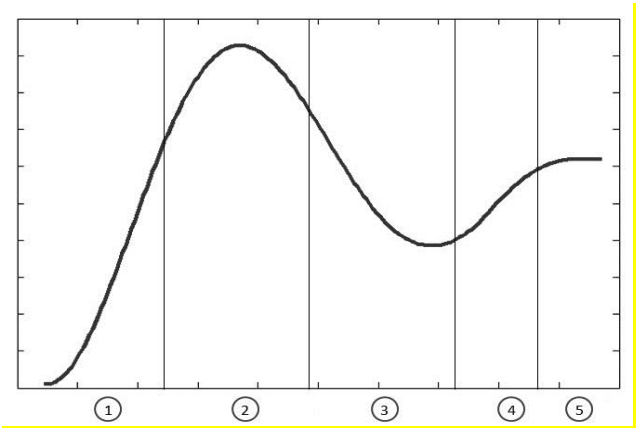


Рисунок 67 – Діаграма кількісті публікацій, що відповідають тренду інноваційної діяльності: 1 – атака/тригер зростання; 2 – пік завищених очікувань; 3 – втрата ілюзій; 4 – суспільне усвідомлення; 5 – продуктивність/фон

Об'єднавши графіки, що відповідають початку інформаційної операції (Рис. 66) та тренду інноваційної діяльності (Рис. 67), можна отримати повний графік, який відповідає

[103] Bhargava S.C., Kumar A., Mukherjee A. A stochastic cellular automata model of innovation diffusion. Technological forecasting and social change, 1993. – 44. – № 1. – P. 87-97.

відображенню інформаційних операцій в інформаційному просторі (Рис. 68).

Запропоновані моделі повністю відповідають реальним даним, які екстрагуються системами контент-моніторингу [104,105]. Тому наведені залежності можуть бути використані як шаблони для виявлення інформаційних операцій – як шляхом аналізу ретроспективного фонду мережевих публікацій, так і для оперативного моніторингу появи деяких їхніх ознак у реальному часі. Як відомо, для виявлення інформаційних операцій слід уважно стежити за динамікою публікацій щодо цільової теми і, за наявності можливості, користуватися доступними аналітичними засобами, засобами цифрової обробки даних та розпізнавання образів, наприклад, вейвлет-аналізом або поліномами Кунченка [106].

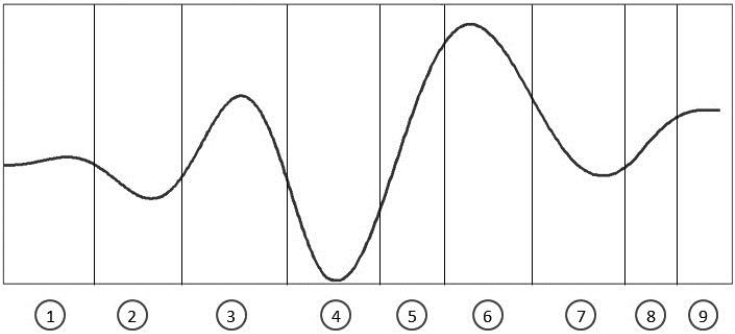


Рисунок 68 – Узагальнена діаграма, відповідна всім етапам життєвого циклу інформаційних операцій: 1 – фон; 2 – затишшя; 3 – «артпідготовка»; 4 – затишшя; 5 – атака/тригер зростання; 6 – пік завищених очікувань; 7 – втрата ілюзій; 8 – суспільне усвідомлення; 9 – продуктивність/фон

Як приклад, на рис. 69 показано динаміку публікацій в Інтернеті тематичних інформаційних потоків за запитами «Бан-

[104] Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах. – К: Наук. думка, 2009. – 295 с.

[105] Ланде Д. OSINT у кібербезпеці : навч. пос. - Київ: ТОВ "Інжиніринг", 2024. - 522 с. ISBN 978-966-2344-97-4

[106] Чертов О.Р. Поліноми Кунченка для розпізнавання образів. Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка, 2009. – № 50. – С. 105-110.

ки, Кіпр», «Офшор», «Вірджинські острови» за березень–квітень 2013 року, у період відомих кризових подій, отриману за допомогою системи InfoStream. Як видно з рис. 103, пік публікацій, пов'язаних із банківською кризою на Кіпрі, припадає на 17–18 березня 2013 року, тоді як більшість публікацій щодо Вірджинських островів припала на 4–5 квітня, коли там, зі значно меншими масштабами, почали проявлятися події, подібні до кіпрських. При цьому слід зазначити слабку кореляованість динаміки інформаційних потоків, пов'язаних із Кіпром та Вірджинськими островами. У цьому випадку коефіцієнт взаємної кореляції відповідних числових рядів становив лише 0,3. Водночас відзначається високий рівень взаємної кореляції рядів, що відповідають тематикам «Офшор» і «Банки Кіпру» (0,73), а також «Офшор» і «Вірджинські острови» (0,77).

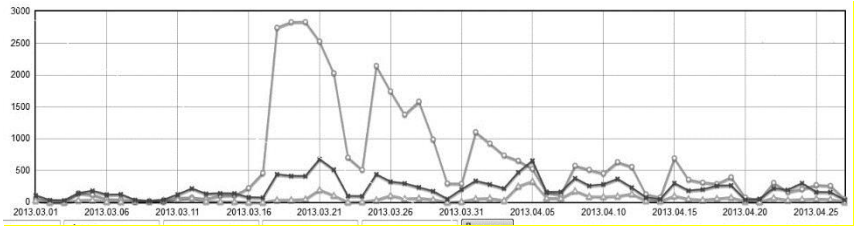
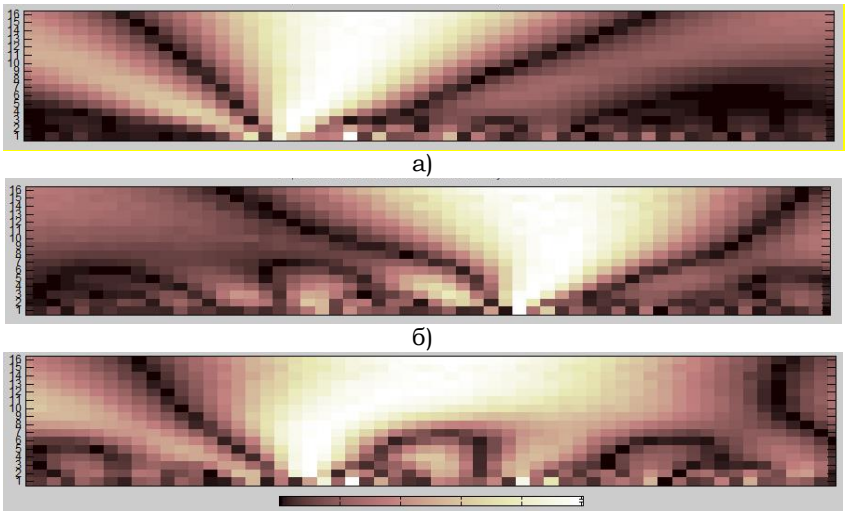


Рисунок 69 – Діаграма динаміки тематичних інформаційних потоків за запитами: o – «Банки Кіпру»; Δ – «Віргінські острови»; *x* – «Офшор»

Ймовірно, прояви інформаційних операцій у сфері офшорних банків у даному випадку найкраще видно під час аналізу більш загальної тематики – «Офшори». На графіку відповідного числового ряду чітко видно дві області локальних екстремумів, що відповідають кризовим ситуаціям на Кіпрі та на Вірджинських островах, а також фази, що відповідають «затишшю» та «артпідготовці».

Можна припустити, що якщо динаміка приватного інформаційного потоку в якийсь момент починає суттєво відрізнятися від динаміки потоку, що відповідає більш загальній тематиці (як у розглядуваному випадку, «Банки Кіпру» і «Офшор»), то можливе проявлення ознак початку інформаційної операції, що стосується вузької тематики.

Під час проведення вейвлет-аналізу[107,108] було ухвалено рішення про використання вейвлету «Мексиканський капелюх», як близького за формою до діаграми, наведеної на Рис. 70.



а)

б)

в)

Рисунок 70 – Вейвлет-спектрограми, що відповідають динаміці тематичних інформаційних потоків за запитами: а – «Банки Кіпру»; б – «Вірджинські острови»; в – «Офшор»

Розглянуті процеси чітко проглядаються як на вейвлет-спектрограмах, так і на відповідних їм скелетах (графіках ліній екстремумів).

Наведені моделі та методи придатні для описання загальних тенденцій динаміки інформаційних процесів, однак проблема прогнозування залишається відкритою. Ймовірно, більш реалістичні моделі можуть бути отримані з урахуванням додаткового набору факторів, більшість з яких не відтворюються у часі. Разом із тим, структура правил, що лежать в основі

[107] Buckheit J., Donoho D.Wavelab and reproducible research. Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. – 27 p.

[108] Torrence, C. and Compo, G.P., 1998. A practical guide to wavelet analysis. Bulletin of the American Meteorological society, 79(1), pp.61-78.

функціонування більшості з доступних моделей, дозволяє вносити відповідні корективи, наприклад, штучно моделювати випадкові відхилення.

Зауважимо, що відтворення результатів у часі є серйозною проблемою при моделюванні інформаційних процесів і становить основу наукової методології. Наразі лише ретроспективний аналіз вже реалізованих інформаційних операцій залишається відносно надійним способом їх верифікації.

Звісно, на практиці орієнтація лише на єдиний тип джерел може призвести до дефіциту інформації, необхідної для прийняття рішень, неточностей, а подекуди – до дезінформованості. Лише застосування комплексних систем, що базуються на використанні численних джерел і баз даних, разом із наведеними вище можливостями системи контент-моніторингу, може гарантувати ефективну інформаційну підтримку при протидії інформаційним операціям.
Виділені зразки поведінки рядів інтенсивностей тематичних публікацій можуть розглядатися як шаблони (зразки) функціональної залежності. Ці шаблони можна взяти як єдиний базисний елемент деякого лінійного простору, тобто як породжуючий елемент для моделювання за допомогою поліномів Кунченка[109].

Тоді як лінійну комбінацію лінійно-незалежних перетворень $f_1(e), f_2(e),..., f_n(e)$ відповідного породжуючого елемента можна побудувати поліном P_n наближення n-го порядку до частини вихідного сигналу $f_s(e)$:

$$P_n = \sum_{\substack{k=0,\\ k\neq s}}^{n} c_k f_k(e),$$

де коефіцієнти c_k визначаються з умови забезпечення мінімуму відстані між поліномом, що будується, і сигналом. Елемент c_0 визначається виразом:

[109] Чертов О.Р. Поліноми Кунченка для розпізнавання образів. Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка, 2009. – № 50. – С. 105-110.

$$c_0 = \frac{\langle f_s(e), f_0(e)\rangle - \sum\limits_{k=1,\ k\neq s}^{n} c_k \langle f_k(e), f_0(e)\rangle}{\langle f_0(e), f_0(e)\rangle},$$

а інші коефіцієнти c_k – як розв'язок системи лінійних рівнянь:

$$\sum_{k=1,\ k\neq s}^{n} c_k F_{i,k} = F_{i,s},\ i = 1,...,n,\ i \neq s,$$

де центровані кореляти $F_{i,k}$ також обчислюються за допомогою відповідних перетворень:

$$F_{i,k} = \langle f_i(e), f_k(e)\rangle - \frac{\langle f_i(e), f_0(e)\rangle \cdot \langle f_k(e), f_0(e)\rangle}{\langle f_0(e), f_0(e)\rangle}.$$

Числовою характеристикою, яку можна використовувати в критеріях якості зіставлення сигналу з виділеним шаблоном, тобто як міру наближення полінома Кунченка P_n до сигналу $f_s(e)$, можна вважати коефіцієнт ефективності d_n:

$$d_n = \frac{\sum\limits_{k=1,\ k\neq s}^{n} c_k \langle f_k(e), f_s(e)\rangle}{\langle f_s(e), f_s(e)\rangle}.$$

Розглянутий метод розпізнавання певних зразків за допомогою побудови простору з породжувальним елементом та пошуку коефіцієнтів відповідного полінома Кунченка може бути використаний у будь-якій предметній області, в якій можна апріорі у часовому ряді виділити певні характерні шаблони.

Таким чином, побудувавши типові моделі поведінки рядів інтенсивності тематичних публікацій під час проведення інформаційних операцій та зіставивши шаблони, отримані на їх основі, можна використовувати метод на основі поліномів Кунченка для визначення (та попередження) можливої інформаційної атаки.

Динаміка тематичних інформаційних потоків визначається комплексом як внутрішніх, так і зовнішніх нелінійних механізмів, які повинні бути відображені при моделюванні (мож-

ливо, у неявному вигляді). Часто задовільним виявляється спрощене розуміння тематичного інформаційного потоку як деякої залежної від часу величини, поведінка якої описується в аналітичному вигляді нелінійними рівняннями. Сьогодні при моделюванні інформаційних потоків використовуються переважно аналітичні нелінійні моделі, застосовуються методи нелінійної динаміки, теорії клітинних автоматів, перколяції, самоорганізованої критичності [110],[111].

Для аналізу динаміки реальних тематичних інформаційних потоків (ТІП) та, відповідно, оцінки їхніх моделей необхідно якимось чином отримати відповідну статистику, представлену у вигляді часових рядів.

Динаміку реальних тематичних інформаційних потоків (ТІП), наприклад, відображає мультиагентна модель, в рамках якої окремі документи ТІП асоціюються з агентами, життєвий цикл яких – з життєвим циклом документів в інформаційному просторі. Відповідно, весь простір мультиагентної моделі асоціюється з тематичним інформаційним потоком. Припускається, що протягом дискретних моментів часу відбувається еволюція популяції агентів. При цьому окремі агенти можуть:

1) «самозароджуватися» (народжуватися з причин, що виникають поза межами розглядуваного мультиагентного простору);
2) «породжувати» нових агентів;
3) «вмирати» – зникати з простору агентів (відповідає втраті актуальності документів);
4) отримувати посилання від інших агентів. Кожен агент володіє «потенціалом», що залежить від його віку (часу життя на поточний момент – t), від авторитетності (посилань, проставлених на нього – ns) та пло-

[110] Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах. - К: Наукова думка, 2009, - 295 с. ISBN 978-966-00-0973-9

[111] A. Dodonov, D. Lande, V. Tsyganok, O. Andriichuk, S. Kadenko, A. Graivoronskaya. Information Operations Recognition. From Nonlinear Analysis to Decision-Making. - LAP Lambert Academic Publishing, 2019. - 292 p. ISBN-13: 978-620-0-27697-1, ISBN-10: 6200276978, EAN: 9786200276971

дючості (кількості породжених безпосередньо ним агентів – k). Потенціал агента Pot визначається формулою:

$$Pot = \frac{1 + ns + k}{t}.$$

На Рис. 71 наведено приклад можливої динаміки мультиагентної системи: процеси народження нових агентів від існуючих позначені суцільними стрілками, процеси встановлення посилань на агентів зображені пунктирними стрілками, живі агенти – чорними колами, «мертві» агенти на момент $t = 5$, – порожніми колами.

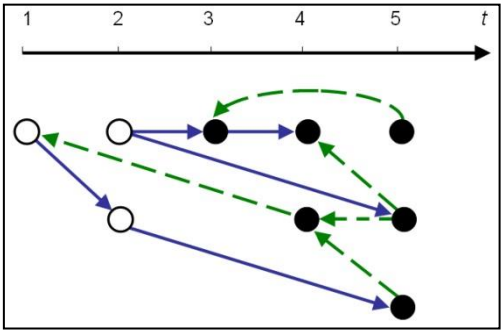


Рисунок 71 – Фрагмент мультиагентного простору

Отже, керуючі параметри моделі є такими:

- ймовірність «самозародження» P_1;
- ймовірність «народження» від існуючого агента: $P_2 \cdot Pot$;
- ймовірність «смерті» агента: P_3 / Pot;
- ймовірність посилання на агента: $P_4 \cdot Pot$.

Варіювання цими чотирма параметрами P_1, P_2, P_3 та P_4 уможливило моделювання типових профілів поведінки ТІП.

На Рис. 72 представлено результати чисельного моделювання кількості агентів (вісь ординат на графіку) у розглядуваній мультиагентній системі залежно від кількості тактів моделювання (вісь абсцис).

Розглянута модель еволюції простору агентів за різних значень керуючих параметрів узгоджується з динамікою реальних тематичних інформаційних потоків, ідентифікованих за допомогою системи InfoStream.

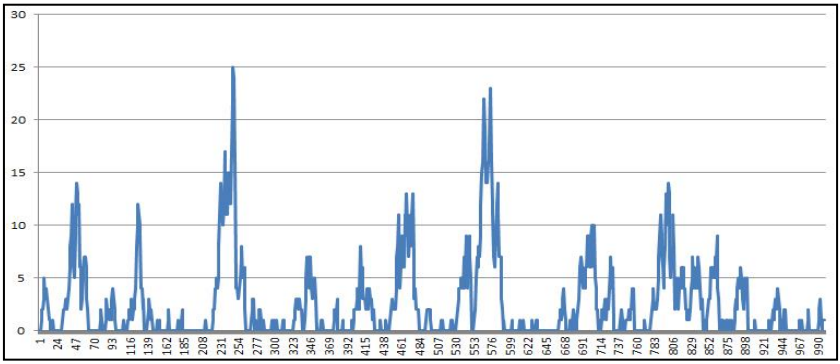


Рисунок 72 – Динаміка зміни кількості агентів в моделі

На практиці орієнтація лише на один тип джерел та математичних моделей може призвести до нестачі інформації, необхідної для прийняття рішень, неточностей, а іноді – до дезінформованості. Лише застосування комплексних систем, що ґрунтуються на численних джерелах, базах даних та математичних моделях, разом із зазначеними вище можливостями систем контент-моніторингу, може гарантувати ефективну інформаційну підтримку під час протидії інформаційним операціям.

8.5. Шляхи протидії інформаційним операціям

Розглянуті практичні приклади дозволили розробити певну загальну методику проведення оборонної інформаційної операції з використанням системи контент-моніторингу вебресурсів. Припустимо, об'єктом агресивної інформаційної операції є компанія «АБВ». Пропонується наступні 12 кроків протидії:

1) збір інформації з публікацій у «чужих» (не пов'язаних з «АБВ», неафілійованих) ЗМІ про компанію;

2) побудова графіка динаміки появи повідомлень про компанію «АБВ» в інтернет-ЗМІ;
3) аналіз динаміки з ретроспективою за 6–12 місяців за допомогою методів аналізу часових рядів. Після цього аналізується контент публікацій у порогових точках, визначаються моменти, тривалість, періодичність впливу, прив'язка моментів впливу до інших подій із сфери інтересів об'єкта;
4) визначення джерел, які публікують найбільшу кількість негативу (публікацій з негативною тональністю) про компанію «АБВ»;
5) визначення «першоджерел» публікацій у ЗМІ – тих джерел, які першими опублікували негативну інформацію;
6) визначення ймовірних «замовників» – власників або осіб, які впливають на видавничу політику окремих ЗМІ;
7) визначення сфер спільних інтересів компанії «АБВ» та потенційних «замовників» (шляхом виявлення спільних інформаційних характеристик – перетинів «інформаційних портретів» системи InfoStream, що будуються для об'єкта та «замовника»), ранжування потенційних «замовників» за їхніми інтересами;
8) визначення критеріїв інформаційних впливів на основі найбільш рейтингових інтересів;
9) моделювання інформаційних впливів, для чого визначаються зв'язки «замовника» – найбільш пов'язані з ним особи та організації, аналізується динаміка впливу з боку замовника та будується прогноз цієї динаміки, аналізується контент публікацій у порогових точках кривої динаміки – визначаються критичні точки впливу;
10) прогнозуються подальші кроки впливу шляхом аналізу аналогічної динаміки публікацій для інших компаній у ретроспективній базі даних системи InfoStream;
11) з урахуванням реалій та публікацій із ретроспективної бази даних оцінюються ймовірні наслідки;
12) організується інформаційна (і не лише) протидія. Приклади публікацій у контексті протидії містяться в ретроспективній базі даних.

8.6. Приклади інформаційних операцій

Антимонопольна діяльність та створення в державі конкурентного середовища передбачають боротьбу з проявами монополізму на ринках товарів і послуг, у тому числі відбиття відповідних інформаційних операцій, що проводяться монополістами, а також проведення наступальних інформаційних операцій. Для здійснення антимонопольної діяльності з боку держави та створення конкурентного середовища необхідно використовувати всі доступні та легальні інформаційні й програмні засоби. Однак сьогодні спостерігається реальний дефіцит оперативної ринкової інформації, зумовлений як слабкими комунікаціями між окремими органами влади, так і неповнотою й неточністю відповідних офіційних баз даних. З іншого боку, існує величезний інформаційний ресурс – вебпростір. Очевидно, що, попри такі переваги, як оперативність та широке охоплення інформації, цей ресурс не може бути доказовим джерелом, проте його не можна відкидати в окремих важливих застосуваннях. Оперативність, притаманна вебсередовищу, зокрема, має вирішальне значення під час реалізації концепції управління OODA, також відомого як цикл Бойда. У перекладі абревіатура OODA означає «Спостереження – Орієнтація – Рішення – Дія»[112]. Концепція OODA знаходить у всьому світі широке застосування в управлінні інформаційним протистоянням та запобіганні інформаційним операціям. Очевидно, що й в антимонопольній діяльності ця концепція може і повинна знайти застосування шляхом створення центрів швидкого реагування на монопольні прояви. Загальновідомо, що антимонопольна діяльність – це комплекс заходів, спрямованих на обмеження діяльності монополій у межах усієї держави, а також на створення відповідного законодавства, тоді як конкурентна розвідка спрямована на підвищення конкурентоспроможності лише окремих суб'єктів господарювання. Відповідно до цього окремі завдання конкурентної розвідки можуть бути узагальнені до рівня антимонопольної діяльності на державному рівні наступним чином:

1) збір інформації та своєчасне інформаційне забезпечення відповідних державних органів;

[112] Richards, C., 2020. Boyd's OODA loop. Necesse, 5(1), pp.142-165.

2) виявлення факторів ризику та загроз для конкурентного середовища держави;
3) виявлення факторів, що впливають на отримання окремими компаніями монопольних переваг;
4) розробка прогнозів та рекомендацій, що впливають на розвиток конкурентного середовища;
5) посилення сприятливих та локалізація несприятливих факторів для розвитку конкурентного середовища.

За допомогою методів конкурентної розвідки, яка стає сучасним напрямом дослідження поведінки конкурентів на ринку, створюються альтернативні моделі ринку для визначення характеристик його учасників та оптимізації тактики і стратегії розвитку суб'єктів господарювання на певних ринках. Досягнення таких цілей вимагає використання ефективних прийомів роботи з інформацією та її елементами. Інформація в цьому сенсі стає об'єктом у процесі дослідження ринку та створення його моделі.

Усі зазначені завдання реалізуються в рамках замкненої схеми взаємодії ринкового середовища та віртуального інформаційного простору.

Як відомо, ринкова реальність знаходить своє відображення у віртуальному інформаційному просторі; саме з ним працюють експерти-аналітики, які готують інформацію та прогнози для осіб, які приймають рішення (ОПР), котрі, у свою чергу, забезпечують цілеспрямований вплив на ринкове середовище.

Ймовірно, усі зазначені функціональні компоненти конкурентної розвідки можуть використовуватися й для загальних завдань, що стоять перед антимонопольними органами держави.

Можливості використання засобів конкурентної розвідки, зокрема засобів контент-моніторингу, в антимонопольній діяльності проілюструємо на прикладі колапсу цін на гречку на початку 2010 року в Україні. Антимонопольний комітет України лише у жовтні 2011 року (через півтора року!) виявив і покарав учасників змови на ринку гречки (Рис. 73), тоді як сотні користувачів системи контент-моніторингу InfoStream могли бачити фігурантів справи вже у лютому 2010 року в «інформаційному портреті» цієї системи (Рис. 74).

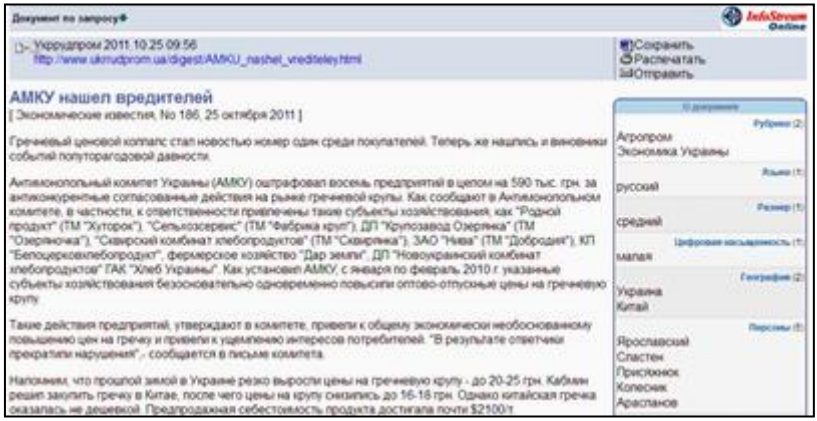


Рис. 73 – Винні у кризі знайдені (жовтень 2011 р.)

Безумовно, система підтримки антимонопольної діяльності, як і системи конкурентної розвідки, що використовують Інтернет як один із інформаційних ресурсів, має бути налаштована відповідно до специфіки конкретних ринків. Вона повинна включати відповідну класифікацію, гнучкі механізми пошуку, оперативної доставки даних, а також якісної оцінки інформації.

Одним із найважливіших завдань аналізу інформації при цьому є визначення її достовірності, тобто розв'язання завдання аналізу та фільтрації шуму й хибної інформації. Після аналізу достовірності інформації мають слідувати оцінки її точності та важливості. Головним критерієм достовірності даних на практиці є підтвердження інформації іншими джерелами, що заслуговують на довіру.

Умови дослідження стану ринку за допомогою електронних засобів, зокрема під час проведення антимонопольної діяльності, повинні відповідати сучасним умовам конкурентної розвідки:

По-перше, повинні застосовуватися методи та програмні засоби дослідження інформації, отриманої з відкритих джерел, з дотриманням вимог законодавства та етичних норм.

По-друге, успіх чи невдача у вирішенні практичного завдання моделювання стану ринку залежать від спрощення інтегрованої інформації, яку необхідно обробити.

По-третє, досягнення успіху під час дослідження ринків пов'язане з проблемою подолання складності доступу до інформаційних ресурсів з відкритих джерел, у тому числі з мережі Інтернет.

Методологія виявлення антиконкурентних дій учасників ринку за результатами аналізу моделі стану ринку повинна відповідати можливостям наявних комп'ютерних засобів та методів конкурентної розвідки.

Рисунок 74 – Відображення «гречаної кризи» 2010 р.

Наприклад, дані, інформація та знання, отримані в результаті антимонопольних досліджень, повинні подаватися у вигляді, що відповідає за структурою та формою розвідувальній інформації.

Сучасні засоби, що застосовуються в конкурентній розвідці в мережевому середовищі, забезпечують:

- доступність необхідної частини інформації;
- широке охоплення інформації;
- оперативність та врахування динаміки інформаційних потоків.

Водночас ці засоби не можуть замінити всі інструменти, необхідні для антимонопольної діяльності. Для прийняття рішень у цій сфері потрібно використовувати комплексні системи, які дозволяють збирати та узагальнювати інформацію про об'єкти дослідження з різних джерел.

Отже, можна зробити висновок, що конкурентна розвідка доповнює технологію пошуку даних та інформації в інтернет-просторі та цільове екстрагування корисних понять про стан і розвиток товарного ринку методами збору, зберігання, обробки та аналізу даних, створюючи середовище інтегрованої інформації для аналізу та формування конкурентної політики.

Цілі та засоби антимонопольної діяльності обумовлюють практичні вимоги до створення нових механізмів і технологій та вимагають об'єднання інструментів конкурентної розвідки різної природи відповідно до різних алгоритмів дослідження.

Висновки до розділу 8

У восьмому розділі монографії обґрунтовано фундаментальну роль протидії інформаційним операціям як невід'ємної складової забезпечення інформаційної безпеки та конкурентоспроможності суб'єктів господарювання в сучасному мережевому середовищі. Інформаційні операції розглядаються як комплексний процес впливу на інформаційні системи та свідомість осіб, що приймають рішення, з метою зміни траєкторії соціальних процесів у потрібному для ініціатора напрямку. Встановлено, що успішність проведення таких операцій значною мірою залежить від живучості інформаційних систем та здатності середовища протидіяти деструктивним впливам через вироблення своєрідного імунітету, що в демократичному суспільстві досягається шляхом навчання та адаптації до інформаційних атак.

Доведено, що моделювання інформаційних операцій вимагає застосування міждисциплінарних підходів, що поєднують методи нелінійної динаміки, теорії складних систем та агентного моделювання. Аналіз динаміки тематичних інформаційних потоків, зокрема кількісної динаміки публікацій, дозволяє виявляти характерні ознаки підготовки та реалізації інформаційних атак. Використання методів вейвлет-аналізу та поліномів Кунченка дає змогу ідентифікувати типові шаблони поведінки інформаційних потоків, такі як фази фонових публікацій, за-

тишшя, артпідготовки та безпосередньої атаки, що є критично важливим для завчасного попередження загроз. Практична реалізація заходів протидії базується на системному моніторингу інформаційного простору за допомогою спеціалізованих систем контент-аналізу, які забезпечують оперативність та повноту збору даних порівняно з традиційними пошуковими системами.

Розроблена методика оборонної інформаційної операції передбачає послідовне виконання етапів від збору інформації та аналізу динаміки публікацій до ідентифікації першоджерел і ймовірних замовників негативного впливу. Ефективність запропонованих підходів підтверджено ретроспективним аналізом реальних інформаційних кампаній, зокрема проти фінансових установ та страхових компаній, де своєчасне виявлення аномалій у інформаційних потоках могло б запобігти значним репутаційним та економічним втратам. Окремо наголошено на можливостях використання інструментарію конкурентної розвідки в антимонопольній діяльності держави, що дозволяє виявляти ознаки змови та маніпулювання ринком на ранніх стадіях, як це було продемонстровано на прикладі гречаного цінового колапсу.

Разом з тим, констатовано, що проблема прогнозування розвитку інформаційних операцій залишається відкритою через складність врахування всіх соціально-психологічних факторів та невідтворюваність умов у часі. Ретроспективний аналіз вже реалізованих операцій наразі є більш надійним способом верифікації моделей, ніж точне прогнозування майбутніх подій. Тому для забезпечення ефективної протидії необхідне використання комплексних систем, що інтегрують численні джерела даних, математичні моделі та інструменти конкурентної розвідки. Такий інтегрований підхід дозволяє не лише виявляти факти інформаційного впливу, але й формувати обґрунтовану доказову базу для антимонопольних органів та керівництва компаній, забезпечуючи перехід від інтуїтивного прийняття рішень до управління, заснованого на достовірних знаннях та аналітичних прогнозах, що є запорукою стійкості бізнесу та держави в умовах гібридних загроз.

Висновки

У монографії комплексно досліджено проблематику комп'ютерної конкурентної розвідки як стратегічного інструменту забезпечення безпеки та підвищення конкурентоспроможності суб'єктів господарювання та державних інституцій в умовах цифрової економіки середини двадцять першого століття. Обґрунтовано, що актуальність конкурентної розвідки в останній час значно зросла через такі процеси, як глобалізація економіки та конкуренції, віртуалізація бізнес-процесів, стрімкий розвиток інформаційних технологій та загострення гібридних загроз. Встановлено, що сучасна конкурентна розвідка еволюціонувала від епізодичного збору відомостей до системної діяльності, що базується на автоматизованій обробці великих масивів даних з відкритих джерел, інтеграції штучного інтелекту та застосуванні складних математичних моделей. Доведено, що інформаційне середовище стало основним полем боротьби за ринкові частки та вплив, де своєчасність, достовірність та глибина аналізу визначають успішність прийняття управлінських рішень.

Технологічне ядро сучасної конкурентної розвідки становить поєднання методів контент-моніторингу, текстової аналітики, семантичного нетворкінгу та використання великих мовних моделей. У роботі показано, що автоматизоване видобування знань з неструктурованих текстів за допомогою великих мовних моделей дозволяє трансформувати хаотичні інформаційні потоки у структуровані мережі знань, де кожен факт пов'язаний логічними та причинно-наслідковими зв'язками. Особливу увагу приділено концепції рою віртуальних експертів, яка імітує колективний аналіз фахівців різних профілів і дозволяє мінімізувати ризики галюцинацій штучного інтелекту та підвищити достовірність висновків шляхом кросверифікації результатів. Розроблені підходи до побудови семантичних мереж на основі відповідей мовних моделей та їх інтеграції з графовими базами даних забезпечують можливість виявлення прихованих залежностей між об'єктами, що є критично важливим для розкриття складних схем впливу та координації дій конкурентів. Окремо наголошено на ролі геопросторових даних у сучасному OSINT, де використання відкритих картографічних ресурсів дозволяє прив'язувати абстрактні

факти до фізичної реальності та моделювати просторові взаємодії.

Математичний апарат, викладений у монографії, включає методи нелінійної динаміки, вейвлет-аналізу, фрактального та мультифрактального аналізу часових рядів, що дозволяє виявляти аномалії в інформаційних потоках та ідентифікувати ознаки підготовки інформаційних операцій. Доведено ефективність використання цих методів для розпізнавання шаблонів поведінки тематичних публікацій, таких як фази фонових повідомлень, затишшя, артпідготовки та безпосередньої атаки, що надає можливість завчасного попередження загроз. Моделювання інформаційних операцій на основі теорії складних систем та агентного підходу дозволяє прогнозувати розвиток соціальних процесів та оцінювати стійкість інформаційного простору до деструктивних впливів. Разом з тим констатовано, що проблема точного прогнозування залишається відкритою через складність врахування всіх соціально-психологічних факторів, тому ретроспективний аналіз та верифікація моделей на реальних даних залишаються надійним способом підтвердження гіпотез.

Важливим компонентом дослідження є правове та етичне регулювання діяльності у сфері конкурентної розвідки. Проаналізовано межі між легитимною розвідкою та промисловим шпионажем, порушенням авторських прав та захистом персональних даних. Обґрунтовано необхідність дотримання принципів мінімізації даних, прозорості алгоритмів та відповідальності за наслідки використання автоматизованих систем аналізу. В умовах гармонізації законодавства з міжнародними стандартами, такими як GDPR, підкреслено важливість розробки механізмів анонімізації та аудиту збору даних, що дозволяє здійснювати розвідувальну діяльність виключно в правовому полі без порушення фундаментальних прав людини. Це створює передумови для формування довіри до інструментів конкурентної розвідки з боку суспільства та державних регуляторів.

Окремий розділ присвячено проблемам управління репутацією та протидії інформаційним операціям, що є критичними для виживання бізнесу та забезпечення національної безпеки. Показано, що живучість інформаційних об'єктів в мережі Інтернет вимагає постійного моніторингу та активних дій щодо

витіснення негативного контенту позитивними наративами, оскільки повне видалення інформації є технічно неможливим через феномен цифрових слідів та тіней. Розроблена методика оборонної інформаційної операції передбачає комплексний підхід до виявлення першоджерел загроз, аналізу динаміки публікацій та формування доказової бази для антимонопольних органів. Практичні кейси, наведені в роботі, демонструють ефективність використання засобів конкурентної розвідки для виявлення картельних змов та маніпулювання ринком на ранніх стадіях, що підтверджує потенціал цих технологій для державного регулювання економіки.

У підсумку монографія стверджує, що майбутнє конкурентної розвідки полягає у створенні гібридних інтелектуальних систем, де штучний інтелект виступає інструментом обробки даних, а людина залишається центральним елементом прийняття рішень та етичного контролю. Інтеграція різних парадигм аналізу, від семантичного вебу до машинного навчання, дозволяє перейти від реактивного моніторингу до проактивного прогнозування та формування бажаного майбутнього. Реалізація запропонованих підходів вимагає агрегування зусиль науковців, практиків та законодавців для створення цілісної екосистеми інформаційної безпеки, здатної протистояти гібридним загрозам та забезпечувати сталий розвиток економіки в умовах глобальної невизначеності. Таким чином, комп'ютерна конкурентна розвідка постає не лише як прикладна дисципліна, а як фундаментальна наука про управління знаннями в цифрову епоху, що визначає стратегічні переваги суб'єктів у боротьбі за ресурси та вплив.

Наукове видання

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ

ДОДОНОВ Олександр Георгійович
ЛАНДЕ Дмитро Володимирович
ПРИЩЕПА Віктор Володимирович

КОМП'ЮТЕРНА КОНКУРЕНТНА РОЗВІДКА

Монографія

Друге видання

Київ, ТОВ «Інжиніринг», 2026

Підп. до друку 04.02.2026. Формат 60×84/16.
Наклад 300 прим. Замовлення № 104