Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет *Інфокомунікаціїі*
(повна назва)

Кафедра *Інформаційно-мережної інженерії*
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Блокування сайтів з використанням методів інтелектуального аналізу даних
(тема)

Виконав:
студент 2 курсу, групи *ІМІм-19-2*
Семенченко О. А.

Спеціальності *172 Телекомунікації та радіотехніка*
(код і повна назва спеціальності)
Тип програми Освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма *Інформаційно-мережна інженерія*
(повна назва освітньої програми)
Керівник *доц. Омельченко А.В.*
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ *Безрук В.М.*
(підпис) (прізвище, ініціали)

2021 р.

Не містить відомостей, заборонених до відкритого публікування

Студент _____ *Семенченко О. А.*
(підпис) (прізвище та ініціали)

Керівник _____ *Омельченко А.В.*
(підпис) (прізвище та ініціали)

Харківський національний університет радіоелектроніки

Факультет *Інфокомунікацій*
(повна назва)

Кафедра *Інформаційно-мережної інженерії*
(повна назва)

Рівень вищої освіти *другий (магістерський)*

Спеціальність *172 Телекомунікації та радіотехніка*
(код і повна назва)

Тип програми Освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма *Інформаційно-мережна інженерія*
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри ІМІ_____
(підпис)
"____" _____2021 року

ЗАВДАННЯ
НА КВАЛІФЛКАЦІЙНУ РОБОТУ

Студентові *Семенченку Олександру Андрійовичу*
(прізвище, ім'я, по батькові)

1. Тема роботи *Блокування сайтів з використанням методів інтелектуального аналізу даних*

затверджені наказом університету від 12 березня 2021 року № 350*Ст*

2. Термін подання студентом роботи до екзаменаційної комісії *12 травня 2021 р.*

3. Вихідні дані до роботи *Об'єкт дослідження – методи виявлення небезпечного матеріалу у мережі Internet та засоби блокування вебсайтів.*
Необхідно проаналізовати способи виявлення небезпечного матеріалу у мережі Internet з застосуванням методів Data Mining.
Проаналізувати основні способи блокування вебсайтів з небезпечним контентом і виконати порівняння їх ефективності за кількісним критерієм.

4. Перелік питань, що потрібно опрацювати в роботі
Вступ
1. Загальні відомості про вебсайти
2. Аналіз особливостей задачі блокування вебсайтів
3. Використання методів Data Mining для виявлення небезпечного матеріалу у мережі Internet
4. Аналіз способів блокування вебсайтів та вибір кращого з них
Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) *Слайди у форматі Power Point (назва, мета роботи, актуальність проблеми, розподіл матералів по важкості порушення, процеси в Text Mining, хмара слів, дендограма слів, пошук слів по тексту, схема різновиду типів блоування, загальна схема пошуку і блокування інформації в Інтернеті, схема різновиду критеріїв блокування/розблокування сайтів, формули розрахунку ймовірності найкращого блокування, результат розрахунку ефективності блокування, результат розрахунку ефективності обходу блокування, висновки)*

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів атестаційної роботи	Строк виконання етапів роботи	Примітка
1	*Ознайомлення із завданням. Уточнення ТЗ*	*14.03.21*	
2	*Підбір літератури за темою роботи*	*15.03-25.03.21*	
3	*Виконання розділу 1*	*18.03-25.03.21*	
4	*Виконання розділу 2*	*26.03-04.04.21*	
5	*Виконання розділу 3*	*05.04-27.04.21*	
6	*Виконання розділу 4*	*28.04-7.05.21*	
7	*Оформлення пояснювальної записки*	*06.05-08.05.21*	
8	*Оформлення презентаційного матеріалу,*	*09.05-12.05.21*	
	підготовка до захисту у ЕК		

Дата видачі завдання _____*13.03.2021 р.*_____

Студент _____ *(Семенченко О. А.)*
(підпис) (прізвище та ініціали)

Керівник роботи _____ *(Омельченко А.В.)*
(підпис) (прізвище та ініціали)

РЕФЕРАТ

Пояснювальна записка: 94 с., 22 рис., 8 табл., 17 джерел, 5 додатки.

Об'єкт дослідження – методи виявлення небезпечного матеріалу у мережі Internet та засоби блокування вебсайтів.

Мета роботи – знаходження ефективних методів виявлення та блокування вебсайтів.

Досліджені методи виявлення небезпечного матеріалу та основні види блокування вебсайтів.

Розглянуті основні етапи пошуку небезпечного контенту по ключовим словам та блокування вебсайтів, а також основні варіанти обходу цих блокувань.

Проаналізовані способи виявлення небезпечного матеріалу у мережі Internet з застосуванням методів Data Mining. З використанням можливостей мови програмування r створено програму виявлення небезпечного контенту.

DATA MINING, WEB MINING TEXT MINING, ВЕБСАЙТ, ЦЕНЗУРА, DNS, БЛОКУВАННЯ САЙТУ, VPN

THE ABSTRACT

Explanatory slip 94 p., 22 fig., 8 tab., 17 sources,5 app.

The object of work is methods of detecting dangerous material on the Internet and means of blocking websites.

The purpose of the work is finding effective methods to detect and blocking websites.

Methods of detecting dangerous material and the main types of website blocking are researched. The main stages of searching for dangerous content by keywords and blocking websites are considered, as well as the main options for bypassing these blockages. Methods of detecting dangerous material on the Internet using Data Mining methods are analysed. Using the capabilities of the programming language R, a program for detecting dangerous content has been created.

DATA MINING, WEB MINING TEXT MINING, WEB SITE, CENSOR, DNS, BLOCKING SITE, VPN

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

DNS – Система доменних імен;

HTTP – (HyperText Transfer Protocol) протокол передачі гіпертекста;

ОС – операційна система;

ПЗ – програмне забезпечення;

ПК – персональний комп’ютер;

URL – Uniform Resource Locator;

VPN – Віртуальна приватна мережа;

DNS – Система доменних імен.

ВСТУП

Сучасні інформаційні технології постійно вдосконалюються, що приводить і до змін у світі. За останні 40 років швидкість передачі даних збільшилась у десятки разів. А з появою Інтернету з'являються тисячі нових сайтів, а кількість статей, відео і інших різноманітних матеріалів важко навіть уявити. На сьогоднішній день будь-яка подія освітлюється не тільки державними каналами, а і випадковими прохожими. Тому фальсифікувати дані становиться важче.

Із-за стрімкого розвитку сучасного цифрового світу виникли нові загрози як звичайному користувачу, так і державам установам. В кіберпросторі з'явилися як природні (ненавмисні), так і потужні кібератаки обумовлені інтересами окремих груп, злодіїв або держав.

Все більше поширюються випадки незакононного добування, накопичення, використання, видалення, розповсюдження, приватних даних, незаконних фінансових операцій, крадіжок та махінацій у мережі Інтернет.

Не потрібно також забувати про сайти і матеріали, які можуть нанести шкоду людині. Тому необхідно відстежувати такі ресурси і, у разі необхідності, блокувати їх. Широкі можливості з автоматизації цих процесів виникають при використанні засобів інтелектуального аналізу даних (Data Mining), зокрема Text Mining та Web Mining.

Використовуючи засоби Text Mining та Web Mining можна проаналізувати матеріал на наявність шкідливого або небезпечного матеріалу. До такого матеріалу відноситься: ненормативну лексику, заклики до суїциду, ущемлення прав віруючих, екстремістські матеріали; дитяча порнографія, використання образів, що сіють ворожнечу за расовою, національною, релігійною або статевою ознакою.

На сьогоднішній день блокуванням ресурсів користуються не тільки держави. Ними користуються компанії для більшої безпеки підприємства, а

також, щоб працівники менше відволікались на робочому місці. Також сам користувач мережі Інтернет може для себе заблокувати ті ресурси, які йому не довподоби.

Мережа Інтернет почала все більше грати важливу роль в політиці. Різні країни світу почали все більше приділяти увагу інформаційній безпеці. Багато сайтів мають незаконий характер щодо законодавства деяких держав. Яскравим прикладом є запобігання поширенню дитячої порнографії, і протидія нелегальної діяльності в Інтернеті блокуваня іноземних онлайн-казино. Це здійснено задля захисту прав інтелектуальної власності.

Розглянувши усі приклади блокування ресурсів, можно прийти до висновку – блокування сайтів Інтернету не є кращи рішенням проблеми, кібератак, забороненого контенту і нелегальної діяльності. Вона скоріше завдасть збиток як користувачам Інтернету, так і державам с національними компаніями.

Метою магістерської атестаційної роботи є аналіз сильних і слабких сторін різних видів блокування ресурсів в мережі Інтернет, а також спрощення роботи експертів шляхом застосування методів інтелектуального аналізу даних.

В даній роботі розглянуті такі аспекти, як:

- Основні закони і питання, що порушують питання кіберпезпеки користувачів, а також основні скорочення, які використовуються в даній сфері.
- Розробка на основі мови програмування R програми, яка допоможе експеру в пошуку небезпечних матеріалів.
- Різновиди сфер блокування. Під цим розуміється те, хто використовує блокування(користувач, держава чи підприємство). Також показані методи здійснення блокування для кожного виду користувача і різні особливості та характеристики для кожного виду даних.
- Також розглянуті базові (для звичайного користувача) методи протидії блокування. Приведено аналіз і розрахунки, при яких виявлено найпростіший метод, а також розглянуті наслідки обходу блокувань.

1 ЗАГАЛЬНІ ВІДОМОСТІ ПРО ВЕБСАЙТИ

Правильність роміщення інформації на сторінках сайту грає немалозначну роль. Це спричинено тим що користувач лише бігло продивляється конент. Розуміючи це, адмінітратори домену стараються уникати умовних «сліпих зон» при розміщені ключової інформації.

Згідно цієї моделі (рис 1.1), погляд користувача при вивченні сторінки послідовно проходить точки 1,2,3 і 4. Якщо ми подивимося на результати досліджень Якоба Нільсена (відомого UI/UX фахівця, автора 10 принципів упішного інтерфейсу), то видно, що у рамках першого екрану поведінка користувачів описується саме Z- патерном [3].

Сектори 1, 2 і 3 отримують найбільше уваги, тоді як 4 майже не видиме: відповідно до дослідження Нільсена далі користувач йде уздовж вертикальної осі F. І на тепловій карті це чітко видно. Недолік Z- патерну такий же - він описує обмежену кількість призначених для користувача сценаріїв : текстовий контент, монотонна сетка- це явно недостатньо для аналізу моделі призначеної для користувача поведінки. Також мають популярність такі патерни як зиг-заг та золотий патерн [3].

Незважаючи на все потібно не забувати про навігацію та розміщення контенту. На рис. 1.1 наведен приклад розміщення ресурсу для Інтернет магазину. Тобто потрапивши на сайт користувач спочатку побачить головну сторінку, або сторінку с товаром, де розміщена основна інформація. Далі починається переміщення по сторінкам, які несуть детальнішу або корисну інформацію, яка розміщена на головній сторінці.

Рисунок 1.1 – Z-патерн

Наприкінці поміщаються контакти. Так як багатьом ця інформація не цікава.

Також на веб ресурсах часто використовують реєстрацію користувачів.

Реєстрація – це процес надання сайту своїх данних для надання доступу до усього функціонала (рис 1.4) сайту, як додаткові можливості або доступ до ресурсів, які недоступні неавторизованим користувачам.

Нижче приведений алгоритм входу до різних сайтом різними способами:

Способи входу до сайту(на прикладі інтернет магазину olx).

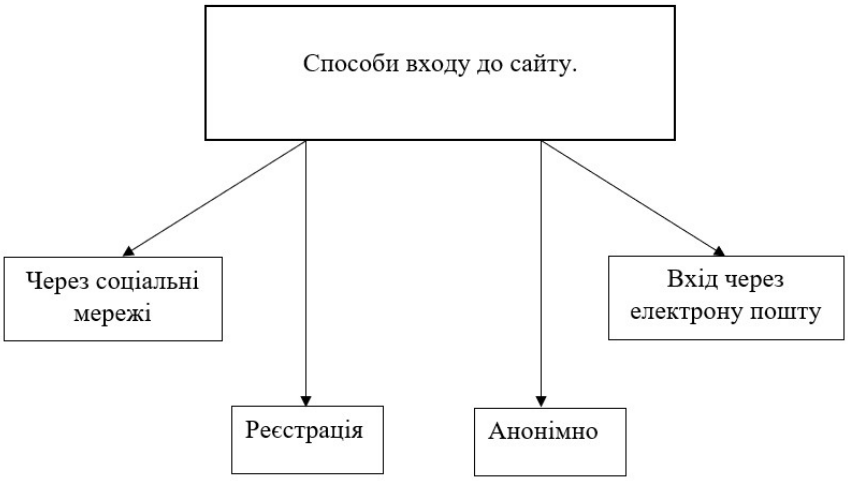

Рисунок 1.2 – Способи входу до сайту

Спосіб використання сайту інкогніто:

1. Заходимо до сайту.
2. Користуємося пошуком, або навігацією по сайту.

Спосіб входу через соцмережу:

1. Заходимо до сайту.
2. Натискаємо «Мій профіль».
3. Можемо одразу вибрати мережу «Facebook», якщо нам це не підходить натискаємо «Інші способи входу».
4. Вибираємо задовольняючу нас соц. Мережу.
5. Користуємося пошуком, або навігацією по сайту.
6. Замовляємо необхідний нам товар.

Таблиця 1.1 – Переваги та недоліки типів входу

Спосіб входу до сайту	Плюси	Мінуси
Анонімно	Анонімність. Швидкість доступу.	Закриті деякі функції сайту
Вхід через соціальні мережі	Деякі дані про Вас заповнюються автоматично. Швидкий вхід.	Уязвімість (так як однаковий логін і пароль з соціальної мережі). Втрачається анонімність.
Реєстрація на сайті	Можливість вказати любий псевдонім на сайті. При взломі	Найдовша реєстрація Необхідність власноруч заповнювати дані

Продовження таблиці 1.1

Вхід через електронну	Більш захищенні данні	Довший вхід/реєстрація

пошту	ніж при вході через соціальні мережі. Залишаються усі покупки, а також варіанти пошуку. Часткова анонімність.	ніж при вході через соціальні мережі.

Плюси:

- Деякі дані про Вас заповнюються автоматично.
- Швидкий вхід.

Мінуси:

- Уязвімість (так як однаковий логін і пароль з соц. мережею).
- Втрачається анонімність.

Спосіб входу через електронну пошту:

1. Заходимо до сайту
2. Натискаємо «Мій профіль».
3. Натискаємо реєстрація.

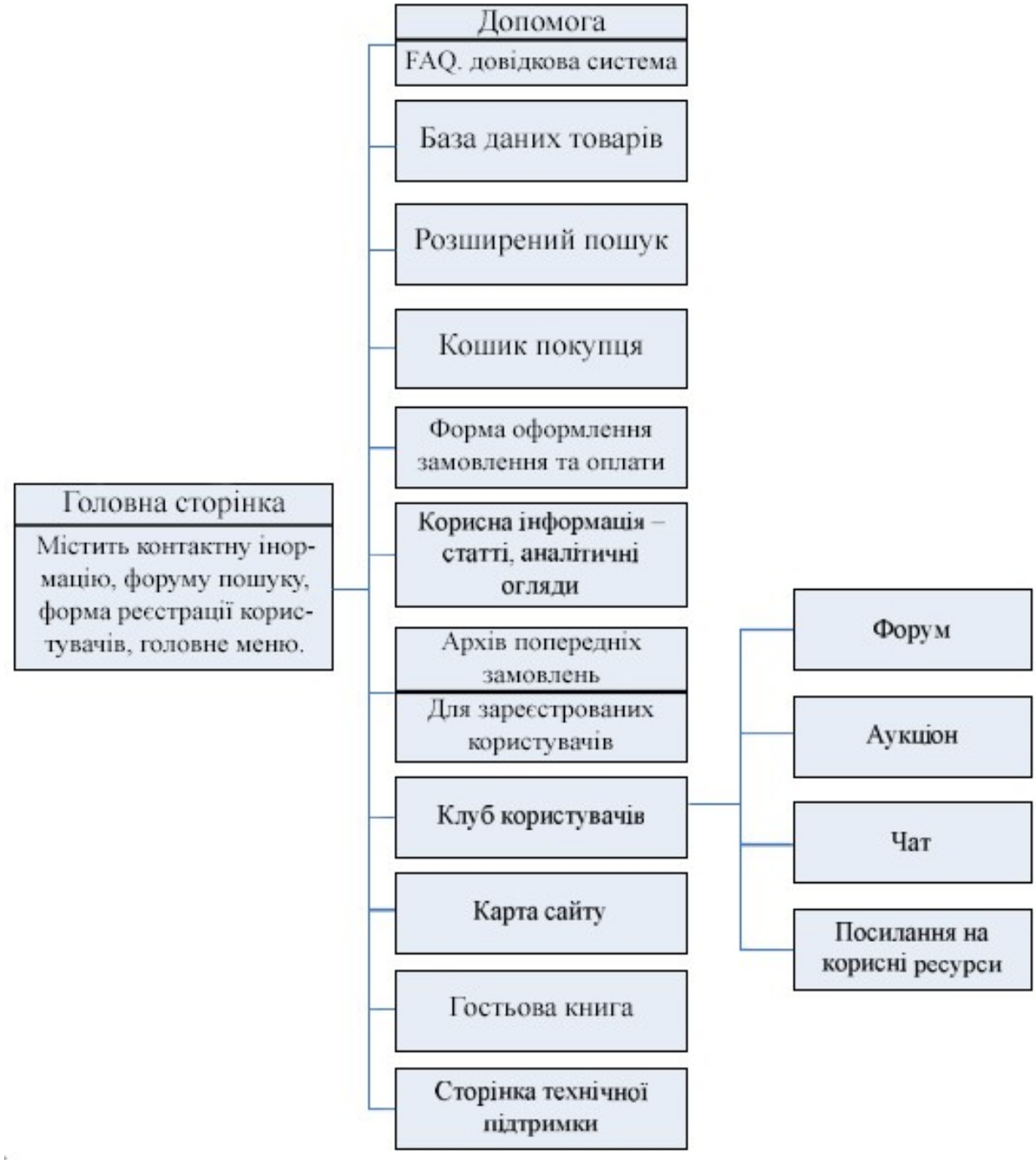

Рисунок 1.3 – Принцип розміщення інформації

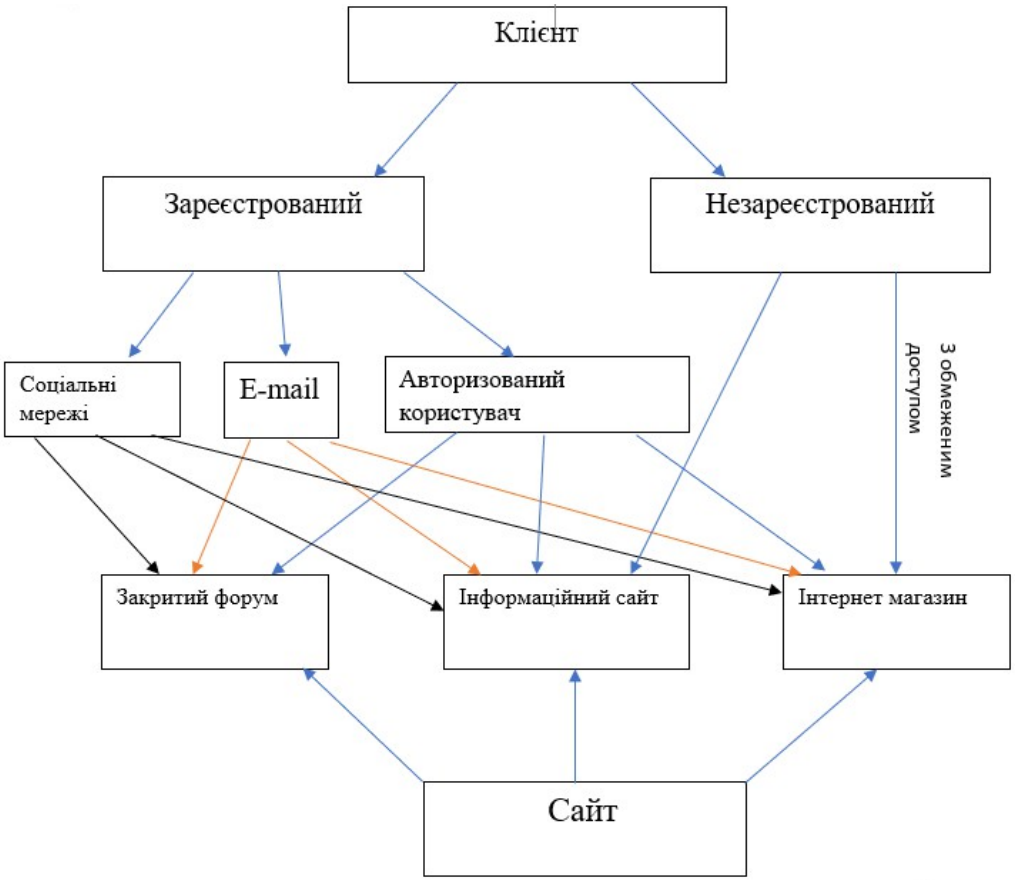

Рисунок 1.4 – Різновид доступу користувачів на сайті

4. Вводимо свій e-mail, вводимо пароль, погоджуємося з правилами користування сервісом (спочатку прочитавши їх), проходимо reCaptchy та натискаємо «Реєстрація».

5. Отримуємо на пошту посилання для закінчення реєстрації

6. Повторюємо вхід до системи.

7. Користуємося пошуком, або навігацією по сайту.

8. Замовляємо необхідний нам товар.

Плюси:

- Більш захищенні данні ніж при вході через соціальні мережі.
- Залишаються усі покупки, а також варіанти пошуку.
- Часткова анонімність.

Мінуси:

- Довший вхід/реєстрація ніж при вході через соціальні мережі.

2 АНАЛІЗ ОСБЛИВОСТЕЙ ЗАДАЧІ БЛОКУВАННЯ ВЕБ САЙТІВ

2.1 Змістовний опис і аналіз предметної області, структурних і функціональних особливостей блокування веб-ресурсів

З врахуванням того, що вже майже у кожної людини є доступ до мережі Інтернет, цензура контенту грає важливу роль для безпеки суспільства.

Принцип блокування веб-ресурсів можливо розділити на такі підрозділи як: запит на блокування, аналіз ресурсу, вибір блокування, блокування.

Далі буде представлена контекстна діаграма бізнес-процесу «Блокування сайтів» на основі стандарту IDEF3 з подальшою її декомпозицією.

На першому рівні описуємо те, що поступає в наш процес зі сторони користувача мережі, а саме запит на блокування веб-ресурсу. На виході отримаємо блокування ресурсу або ж відмову на блокування. Також не забуваємо про керуючі фактори (закони і технічні характеристики). Механізмами будуть працівники (адміністратор мережі, сервер), а також до механізмів запишемо користувача мережі, так як він грає значну роль в процесі (рис 2.1).

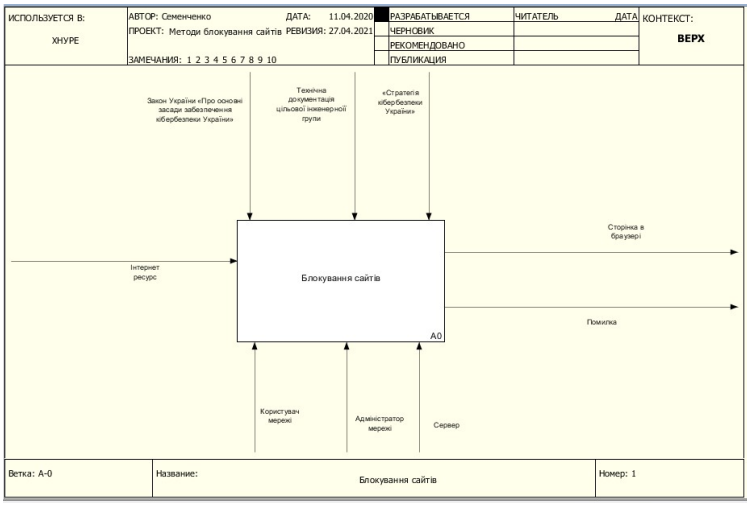

Рисунок 2.1 – Контекстна діаграма

Далі робимо декомпозицію і розглянемо більш детальніше хід роботи(рис 2.2).

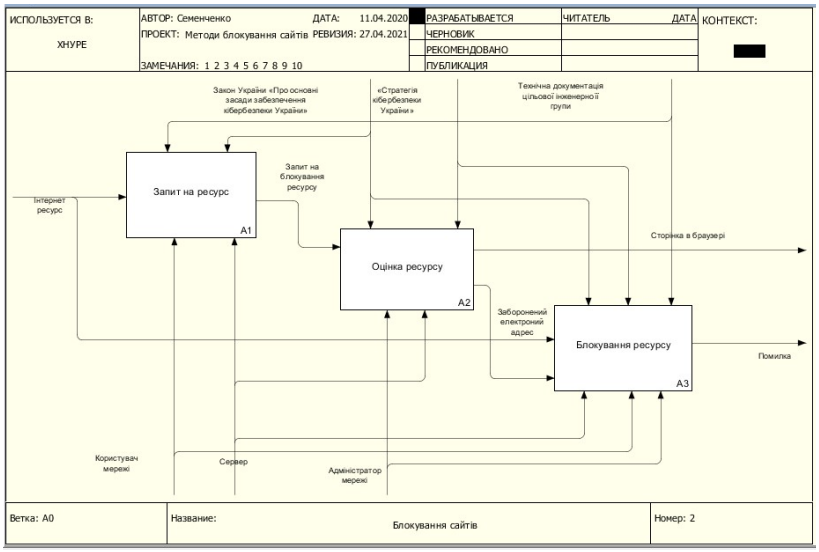

Рисунок 2.2 – Перший рівень декомпозиції контекстної діаграми

Тепер розглянемо більш детальніше блок A1, а саме «Запит на ресурс».

В цьому блоці користувач мережі знаходять веб-ресурс який має якийсь небезпечний матеріал. Далі іде запи на сервер де передають дані адмінісратору мережі длі подальшого аналізу на наявність небезпечного контенту(рис 2.3).

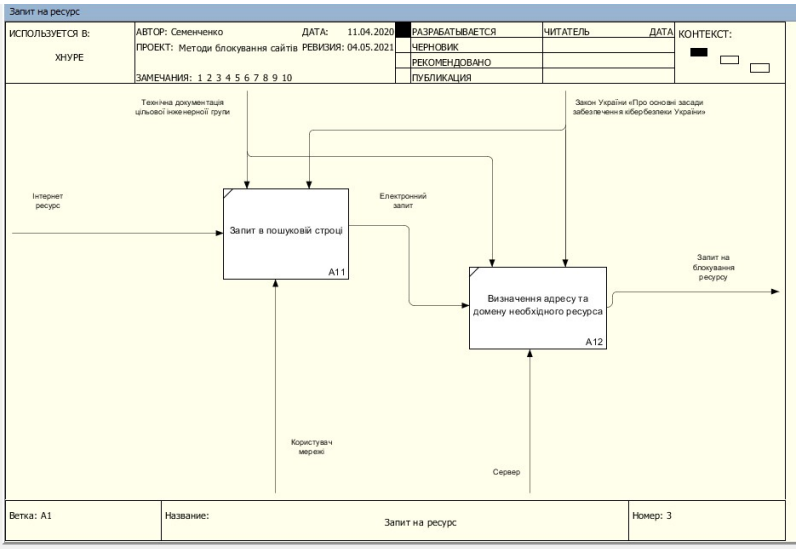

Рисунок 2.3 – Декомпозиція роботи «Запит на ресурс»

Тепер розглянемо «Оцінка ресурсу». В цій частині аналітик(в нашій схемі ми залишили цю роботу адміністратору мережі) користуючись Text Mining та Web Mining знаходить потенціально небезпечний матеріал. При знаходжені небезпечного матеріалу аналітик починає продовжувати роботу і підбирати принцип блокування. Якщо при аналізі нічого підозрілого не було знайдено то адміністратор просто закінчує перевірку веб-сайту і на цьому робота закінчується (рис 2.4).

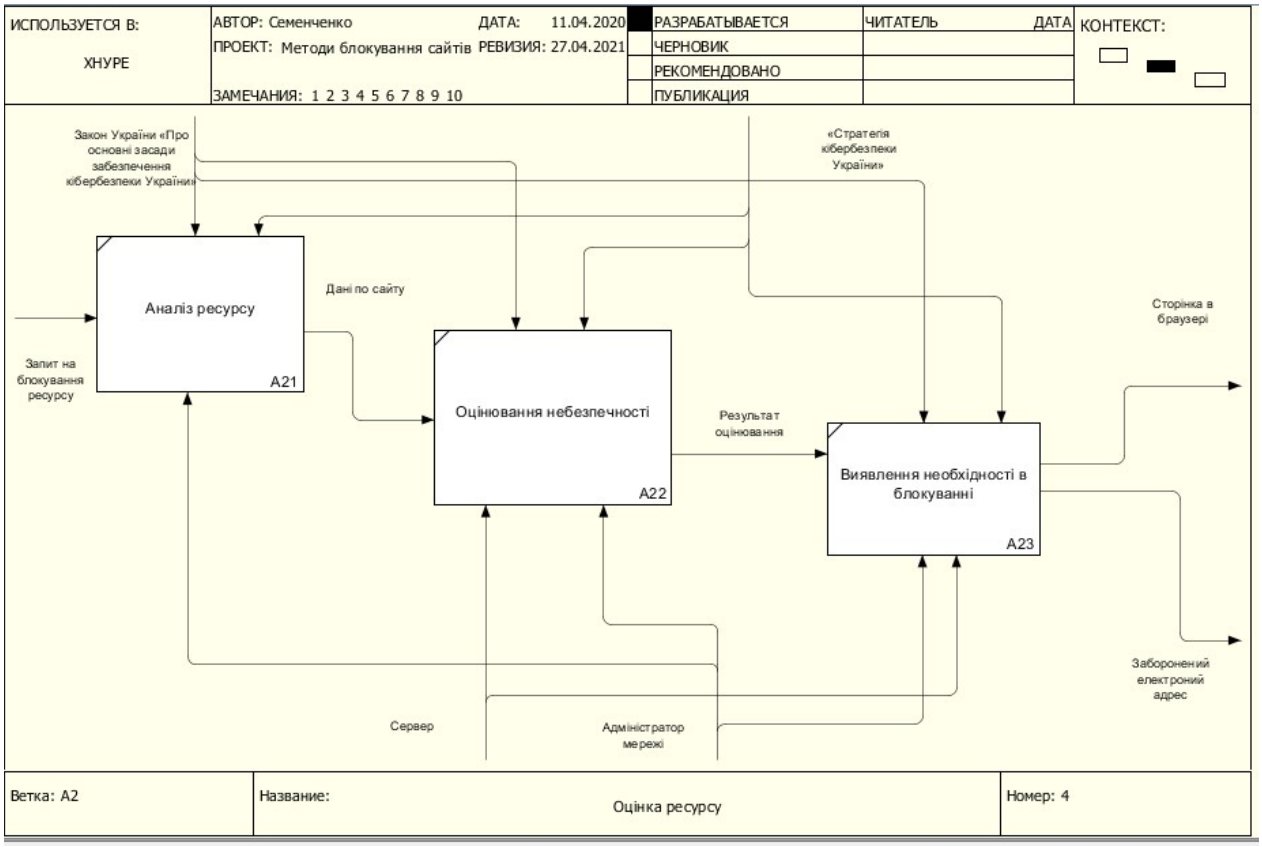

Рисунок 2.4 – Декомпозиція роботи «Оцінка ресурсу»

В «Блокування ресурсу» ми розглядаємо ресурс який потрібно заблокувати. Адміністратор мережі вибирає тип блокування, а також проводить данне блокування. В кінці данної процедури звичайний користувач не зможе отримати доступу до тих сайтів. Також вказаний пункт де сам користувач може заблокувати ресурс. Це визвано будь-якою неприязню самої людини і блокування буде разповсюджуватися лише на дану особу (рис 2.5).

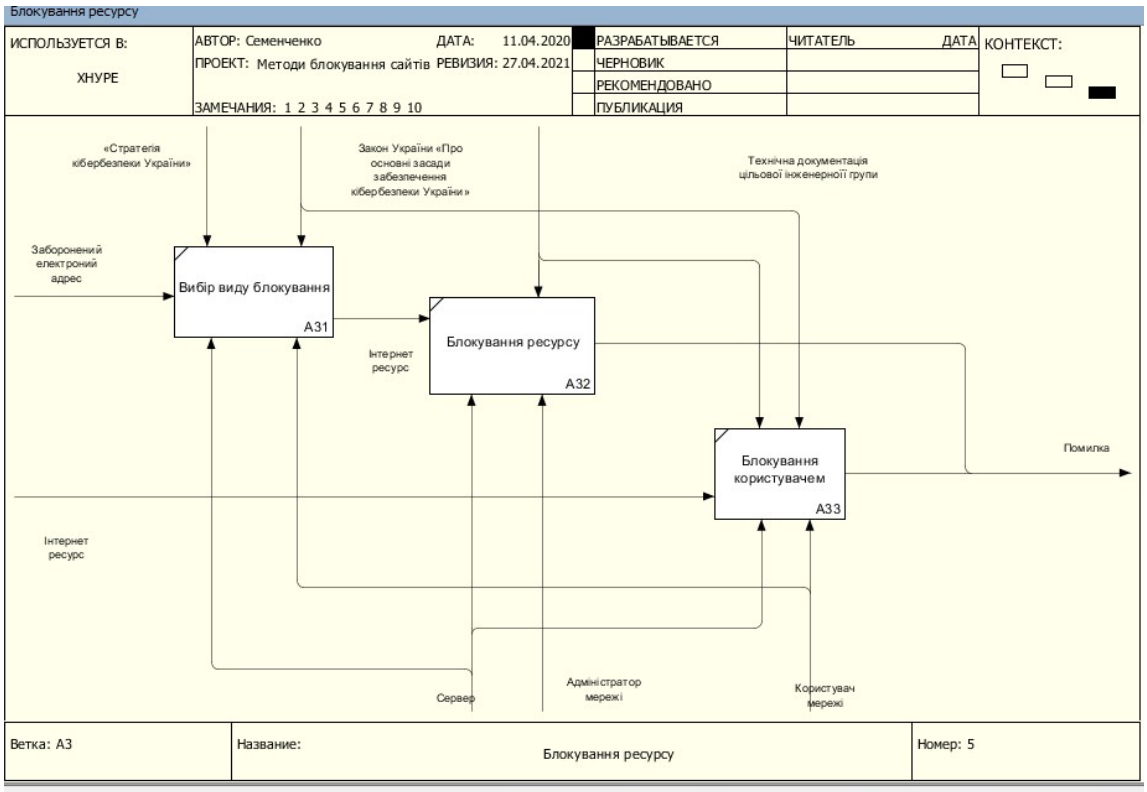

Рисунок 2.5 – Декомпозиція роботи «Блокування ресурсу»

Можна зазначити що дана схема являється узагальнена і для різних варіантів може мати невеликі зміни. Як прилад можно привести:

- у блоці «Запит на ресурс» може бути запит від держави про порушення законів, або зміна законів;
- у блоці «Оцінка ресурсу» може бути відправлена вимога сайту видалити матеріал. Після чого проводять переоцінку.

2.2 Аналіз контента веб-сайтів на небезпечність

Перед тим як блокувати сайт ми повині проаналізувати його на наявність незаконих або небезпечних матеріалів. До таких матеріалів можемо віднести:

- Заклик до єкстремізму або єкстремістські висловлювання;
- Заклики в участі в неузгоджених мітінгах;
- Розповсюдження дитячої порнографії;

- Розповсюдження інормації о способах здійснення самовбивства або схиляння до самогубства;
- Пропаганду вживання наркотиків;
- Оправдання тероризму;
- Поширення фейков;
- Неповага до держави та приниження влади;
- Розголошення персональних даних;
- Порушення авторських прав;
- Розповсюдження матеріалів небажаних організацій;
- Торгівля органами;
- Торгівля забороненими товарами (незарегестрована зброя, наркотики і т. д.);
- Розміщення нелегальних онлайн-казино та фінансових пірамід.

Проаналізувавши вище перерахований перелік небезпечнх матеріалів можливо їх розбити на такі підвиди(Таблиця 1).

Таблиця 2.1 – Розподіл матеріалів по важкості порушення

Одразу заблокувати	Обмежети доступ	Увідомити про необхідність видалення
Заклик до єкстремізму або єкстремістські висловлювання	Заклики в участі в неузгоджених мітінгах	Поширення фейков
Розповсюдження дитячої порнографії	Розповсюдження інормації о способах здійснення самовбивства або схиляння до самогубства	Неповага до держави та приниження влади

Продовження таблиці 2.1

Розголошення	Пропаганду вживання	Порушення авторських

персональних даних	наркотиків	прав
Розголошення персональних даних	Оправдання тероризму	Розповсюдження матеріалів небажаних організацій
Торгівля органами		
Торгівля забороненими товарами		
Розміщення нелегальних онлайн-казино та фінансових пірамід		

Матеріали в яких спочатку можливо сповістити власника сайта про те що його сайт буде заблокований. І якщо власник не реагує в заданий час то ресурс блокується.

Під обмеження доступу мається на увазі що спочатку блокується сайт, а тількі після чого приходить повідомлення про блокування. Власник має три дні с початку блокування на те щоб видалити увесь матеріал який визнали небезпечним.

2.3 Блокування на різних рівнях

2.3.1 Блокування сайтів державами

На сьогоднішній день мільярди людей часто стикаються з різними видами блокування. Це залежить країни і її законодавства, місця проживання та релігії. За таким блокуванням звичайно слідкує держава перекладаючи технічну частину на інтернет-провайдерів. За У сучасному світі мільярди людей стикаються з тією чи іншою формою блокувань при серфінгу в інтернеті. Доступ до інтернет-контенту в країнах блокується за допомогою обладнання інтернет-провайдерів.

Найбільша цензура для користувача мережі Інтернет вважається: Китай, Сирія, Іран, Ефіопія, Узбекистан [4].

Найяскравішим прикладом маштабного блокування можна назвати – «Великий китайський файрвол» або «Золотий щит». З його появи, він постійно аналізує ресурси по ключовим словам і блокує при наявності заборонених маеріалів. В той же час працівники кібербезпеки повторно перевіряють сайти. Як відомо на 2015 рік під цензуру в Китаї потрапилоо близько 3 тис. сайтів [5].

Розробку «Золотого щита» розпочалися в 1998 році на основі правил підготовлених Міністерством громадської безпеки. А саме:

- завдання шкоди нац. безпеці;
- розкриття державних таємниць;
- нанесення шкоди інтересам держави або громадянам.
- створювати і поширювати інформацію, яка спрямована проти Конституції КНР,
- заклики до революції,
- підривання національної єдності;
- наклеп,
- ширить чутки або шкодить громадському порядку;
- містить матеріали сексуального характеру або заохочує азартні ігри, насильство або вбивства.

Презентація Великого китайського файрвола відбулося в 2000 році в Пекіні. Він був представлений як інструмент , спрямований на "впровадження передових інформаційних і комунікаційних технологій в цілях зміцнення контролю, реагування і боротьби із злочинністю" [5].

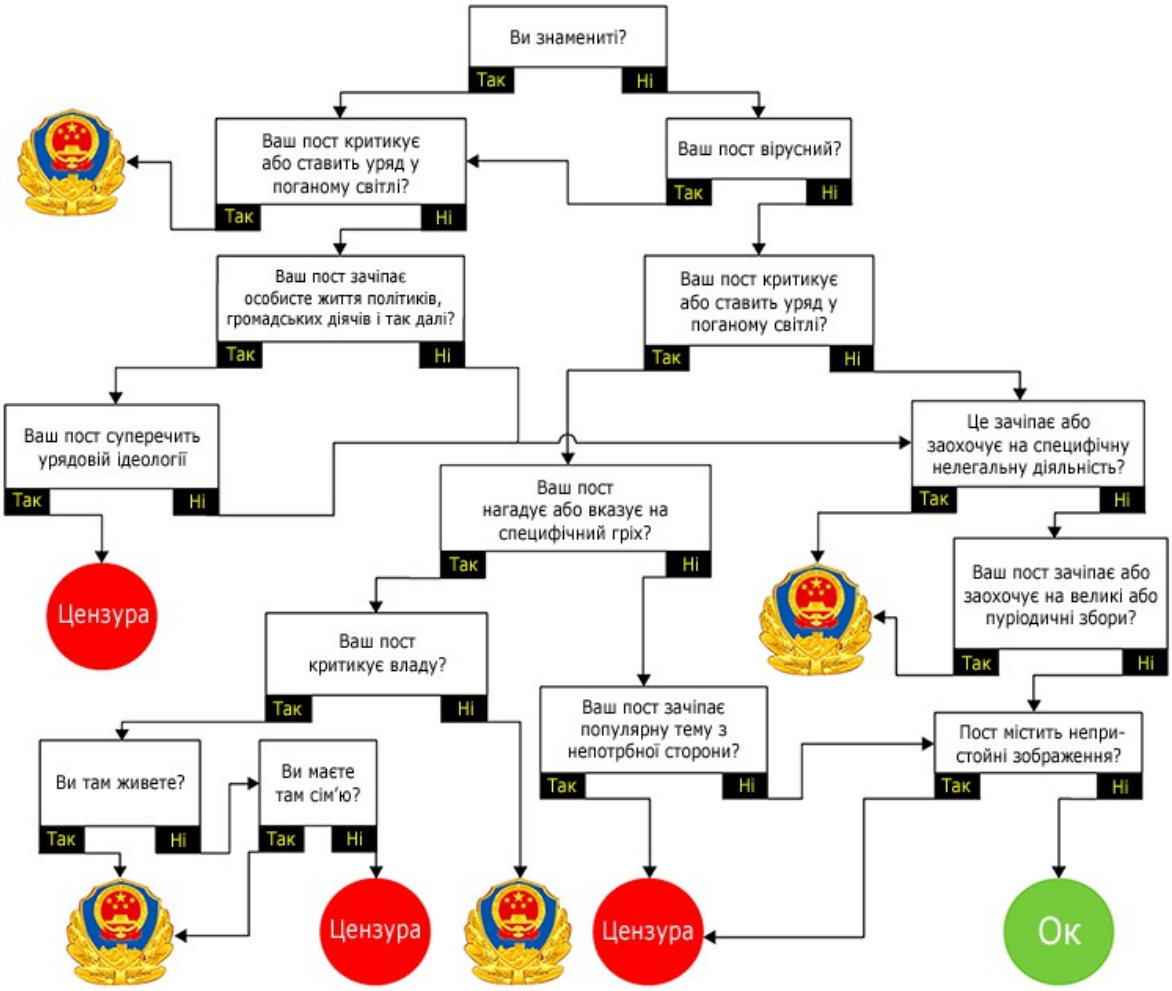

Рисунок 2.6 – Алгоритм блокування файрвола «Золотий щит»

Спочатку "Золотий щит" був багаторівневою системою баз даних. Її використовували для збору та зберігання інформації про більшу частину жителів Китаю. З часом одна із підсистем отримала назву «Великий файрвол». Її застосовують для аналізу контенту на відповідність встановленим правилам (рис 2.1).

Розвиток цього фаєрвола можливо поділити на декілька єтапів. Спочатку фільтр навчили блокувати тільки доменні імена IP-адреса. На сьогоднішній день цей метод досі існує в «Великому файрволі» досі, але тепер не грає таку важливу роль. Також слід зазначити що з'явився так званий «Чорний список».

На другому етапі розвитку файрвол навчили відсіювати неправомірний контент через пошук по ключовим словам. Програма аналізує увесь трафік на відповідність "національному чорному списку" ключових слів. У разі

знаходження заборонених зв’язок з'єднання приривається. Під цей вид блокування тоді потрапили компанії соцільних-мереж, такі як Facebook, Youtube, Twitter, Blogspot, Vimeo.

Однако через деякий час користувачі навчилися їх обходити за допомогою VPN і Shadowsocks. Тому наступний кроком було навчитися виявляти використання VPN і схожих методів. Це вдалося за допомогою виявлення протоколів IPSec, L2TP / TPSec і PPTP.

Окрім вищеперерахованих функцій "Золотий щит" здатний:

- Перехоплювати DNS.
- Фільтрувати контент на стороні користувача.
- Фільтрувати контент вибірково.
- Покладатися на самоцензуру.

Ще одна країна яка не поступається Китаю в цензурі – це Північна Корея. Однако важко провести аналіз так як їх система зберігається в великій таємниці. Проте в 2016 році в їх системі з’явилися збої в сервері імен верхнього рівня сталі відомі імена декількох доменів. Список 28 сайтів Північної Кореї опублікував користувач Reddit [6].

А доступ до «великої мережі» звичайному користувачу взагалі немає. Виключення лише північнокорейці, які займають високі посади або займаються питаннями моніторингу чи надають дозволи [7].

Вперше Північно Корейський домен з'явився лише в 2007 році - .kp. До цього державні сайти хостились в Китаї, Японії, Німеччині і США.

Також потрібно зазначити що усі ПК, які потрапили або зібрані в Північній Кореї ставляться на облік. Мобільний інтернет, а також продаж ноутбуків дозволений лише для послів та членів іноземних компаній. Це обумовлено захистом безпеки держави.

Однако в Північній Кореї розроблена своя власна внутрішня національна мережа, а саме Кванмен. В ній можлие листування через єлектрону пошту, а також є внутрішні сайти які пройшли жорстоку цензуру. На цих сайтах

зазвичай показують восхваляння влади або розмщенні статьї на науково-технічні теми.

На усіх офіційних сайтах використовується спеціальний скрипт, який аналізує текст на наявність в ньому імені Ким Чен Ина і збільшує його кеглю.

На даний час експерти почали відмічати що з приходом до влади Ким Чен Ина жителі Північної Кореї отримують все більше послаблень.Можливо настане час коли жителі Північної Кореї зможуть отримати доступ до глобальної мережі.

2.3.2 Блокування сайтів підприємствами

В останій час набуло розповсюдження надання робітникам доступ до мережі Інтернет з робочих комп’ютерів. Державні організації, як зазвичай відстають у цьому, але навіть у них є окремі мережеві сегменти або робочі станції, підключені до Інтернет.

Опираючись на звіт Symantec's Internet Security Threat Report всього 2.4% сайтів з дорослим контентом поширюють віруси, що набагато менше ніж в блогах, новинних порталах чи і інтернет-магазинах.

Яскравим прикладом цього є зломи сайтів New York Times, NBC, сайтів грузинського уряду, TechCrunch Europe та інших, з головних сторінок яких поширювалися віруси та здійснювалися фішингові атаки.

Така атака може привести до великих проблем. В якості прикладу можна навести:

- втрата або витік даних з ПК юриста або головного бухгалтера;
- вихід з ладу окремих вузлів інфраструктури.

І для того щоб мінімізувати такі проблеми підприємства звертаються до різних видів блокування.

Найрозповсюджена на сьогоднішній день це фільтрація за репутацією. В ньому свідомо блокують деякі сайти або розділи. Інформація про шкідливість знаходиться у хмарі і поновлюється кожні декілько хвилин. Якщо з’явився недавно то його рейтингшвидко складається за допомогою адміністраторами

які присвоють сайту оцінку і ми можемо отримати середній рейтинг цього сайту. Також потрібно пам’ятати про те що блокувати увесь сайт не є дуже якісним, тому кращим рішеням буде блокувати сумнівні розділи.

Також можно зробити фільтрацію по категоріям (URL-фільтрація). Завдяки цьому ми можемо одразу відсіяти усі непотрібні категорії(онлайн-казино, хакерські форуми і т. д.). Інформація для даної фільтрації також краще отримувати з хмари.

Однако не слід забувати про те що більша частина таких систем не здатна аналізувати російськомовні сайти. Тому потрібно ретельно підбирати систему. Також необхідно щоб працівник сканував на наявність вірусів все що він скачає.

Найкращим рішеням також буде при аналізі використовувати Advanced Malware Protection (AMP) або його аналоги. Центри аналітики VRT Sourcefire і SIO Cisco перевіряють чи зустрічався раніше саме цей файл в ході атаки в іншій організації, і якщо немає, то тестує його в пісочниці, аналізуючи їх дії.

Буде корисним обмежити доступ до соціальних мереж або месенджерів. Наприклад дозволити користуватися Skype і Facebook, але заборонити пересилку файлів і відео-дзвінки. Також треба заборонити усі програми для p2p обміну файлами, анонімайзери і утиліти для віддаленого управління.

Не слід забувати про заражені хости. Це один з найчастіших видів зараження. Розповсюджені випадки коли ПК становляться ботнетами.Нажаль антивірус не може розпізнати усі види вірусів, а також враховуючи що шахрайські програми працюють на низькому рівні, приховують свої процеси, з'єднання та існування в цілому від антивіруса. Цього можно уникнути якщо початианалізуваи трафік і підключити даний пошук до серверів з так званими «Чорними списками».

Але якщо це атака новим невідомим ботнетом то його можливо вирахувати за допомогою аналізу. А саме по шифруванню вмісту, малому обсягом переданих даних і тривалого часу з'єднання.

Розшифровка трафіку можливо здійснювати як на тому ж сайті, де і проводиться аналіз, так і на заздалегіть виділеному пристрої. Але треба пам’ятати те що при використовувати лише один хост то ми можемо втратити продуктивність самої мережі.

2.3.3 Блокування сайтів через персональний комп’ютер

Бувають випадки коли користувач сам вирішує блокувати ресурс. Батьки бажають захистити своїх дітей від травмуючих матеріалів, чи небажання бачити контент який тебе дратує. Найпростіший приклад таких блокувань є блокування рекламних банерів. Зазвичай мета блокування сайтів - це заборонений контент, матеріали для дорослих, або сайти шахраї.

Нижче перераховані методи дадуть змогу заблокувати доступ до інтернет ресурсів не тільки на одному ПК, але і розширити цю заборону для цілої мережі за допомогою WI-FI-роутера. Використовуючи маршрутизатори ми зможемо обмежити доступ не тільки комп'ютера, але і всіх інших пристроїв.

Можна виділити такі найпопулярнши види блокування користувачем:

- Блокування сайтів шляхом редагування файлу hosts.
- Блокування сайту в брандмауері Windows.
- Блокування за допомогою розширень в браузерах.
- За допомогою сторонніх програм.

Далі буде розглянуто кожен варіант більш детальніше(Рисунки послідовності виконання наведені у Додатку А)

Блокування сайтів шляхом редагування файлу hosts

Це блокування відбувається шляхом переадресації на іншу веб-сторінку.

Для того щоб провести даний вид блокування нам необхідно відкрити блокнот з правами Адміністратора. Далі вибираємо «Відкрити» в меню «Файл».

В відкритому вікні задаємо шлях, а саме: Диск C: -> Windows -> System32 -> drivers -> etc. Не забуваємо вибрати відображати «Усі файли» і відкриваємо файл «hosts». У цьому файлі ми будемо працювати.

Не чіпаємо все те що там було записано, а пишемо в кінці. Записи повині мати вигляд: адреса IP 127.0.0.1 і через пробіл вводимо IP-адресу або домене ім'я який ми бажаємо заблокувати.

Блокування сайту в брандмауері Windows

Даний спосіб є надійним тому що брандмауер грає роль фаєрвола і в ньому більш краще реалізовані принципи блокування.

Вразливість цього методу являється тим що блокування іде через IP-адресу сайту, а вона з часом може змінитися. Тому доведеться вносити цей сайт знову. Для того щоб провести таке блокування нам потрібно спочатку визначии IP-адресу сайту, який потрібно заблокувати. Відкриваємо командний рядок, для цього необхідно натиснути комбінацію клавіш *Win + R* і в відкрилася рядку «Виконати» та вводимо команду «cmd».

Тепер вводимо адресу сайту, який хочемо заблокувати. Після виконання команди ви побачите IP-адресу вашого сайту, яку ми будемо використовувати.

Далі запускаємо брандмауер, натиснувши на клавіші «*Win + R»* і в вводимо команду «control». Потрапивши до вікна «Панель управління» ми вибираємо «Брандмауер Захисника Windows». В цьому меню треба вибрати пункт «Додаткові параметри». Тепер переходимо до розділу «Правила для вихідних підключень». Після цього в розділі «Дії» вибираємо пункт «Створити правило». Відкриється вікно майстра створінь правил. В наступному вікні вибираємо пункт «Всі програми».

Тепер в меню «Протоколи і порти» нічого не чіпаємо і йдемо далі. В меню «Область», переходимо до «Вкажіть віддалений IP-адресу ...» виберіть пункт «Зазначені IP-адреси» та натисніть кнопку «додати». Якщо після виконання сайт продовжує відкриватися то повиноповторно дізнатися його IP-адресу та ввести усе повторно. Відомі сайти можуть мати велику кількість IP-адресів.

У вікні «Дія» обираємо «Блокувати підключення» і натискаємо Далі. У наступному вікні «Профіль» нічого не змінюємо. В останньому розділі «Ім'я» необхідний вказати ім'я нашого правила і натиснути кнопку «Готово».

Блокування за допомогою розширень в браузерах

Такі види блокування дуже популярні тому що їх дуже просто налаштовувати. А також є важливим бонусом є те що можливо виставити пароль для того щоб ніхто не міг внести зміни.

Нижче приведемо список найпопулярніших розширень для блокування:

Adult Blocker - розширення дає змогу створити чорний список за допомогою якого будуть блокуватися сайти при завантажені

Block Site - це розширення працює так же як і попереднє. Встановлювання паролю для цього розширеня необов'язкове.

*За допомогою сторонніх програм*Під «стороніми» розуміюєть ті програми які не відносяться до стандартних в операційній системі. Однією з найпопулярніших програм можна виділити Adguard. Вона служить для захисту від вірусів і захис від нав'язливої реклами в інтернеті.

Використовуючи функцію «батьківського контролю» ми можемо ввімкнути блокування стандартних сайтів(завчасно вибраних програмою), а також внести адреса сайтів які хочемо власноручь заблокувати.

Також можна зазначити що функція «батьківського контролю» та блокування сайтів за *URL* адресами доступна у всіх популярних антивірусних програмах.

Child Control - являється повноцінним програмою яка збереже дієй від небезпечного контенту, а також надає можливість перевіряти їх дії в мережі.

Any Weblock дуже проста програма яка не може виділитися великою кількістю функцій, але там є цікава функція. Після внесення сайту в цю таблицю блокуваня буде працювати навіть тоді коли програма вимкнина. Це визвано тим що дані цієй програми зберігаються в кеші ПК.

3 ВИКОРИСТАННЯ МЕТОДІВ DATA MINING ДЛЯ ВИЯВЛЕННЯ НЕБЕЗПЕЧНОГО МАТЕРІАЛУ В МЕРЕЖІ INTERNET

3.1 Data Mining

Data Mining це дослідження збору, очищення, обробки, аналізу та отримання корисного аналіз даних. Існує широка варіація щодо проблемних доменів, програм, формулювання та подання даних, які зустрічаються в реальних додатках [6].

Отже, «Data Mining» - це широкий загальний термін, який використовується для опису цих різних аспектів обробка даних.

У сучасну епоху практично всі автоматизовані системи також генерують певну форму даних для діагностики або аналізу. Це призвело до потоку даних, який і був досягаючи порядку петабайт або екзабайт. Деякі приклади різних видів даних наступним чином [6].

- World Wide Web: Кількість документів у проіндексованому Інтернеті зараз на замовлення мільярдів, а невидима Мережа набагато більша. Доступ користувачів до таких документів створювати журнали веб-доступу на серверах та профілі поведінки клієнтів на комерційних сайтах. Крім того, пов'язана структура Інтернету називається веб-графіком, який сам по собі є різновидом даних [6].
- Financial interactions: Найпоширеніші транзакції в повсякденному житті, такі як використання картка автоматичного банкомата (банкомат) або кредитна картка, може створювати дані в автоматизованому режимі. Такі транзакції можуть бути видобуті для багатьох корисних даних, таких як шахрайство або інша незвична діяльність [6].
- User interactions: Багато форм взаємодії користувачів створюють великі обсяги даних. Наприклад, використання телефону зазвичай створює запис у телекомунікації компанія з подробицями про тривалість та пункт

призначення дзвінка. Багато телефонів компанії регулярно аналізують такі дані, щоб визначити відповідні моделі поведінки які можуть бути використані для прийняття рішень щодо пропускної спроможності мережі, рекламних акцій, цін або націлювання на клієнтів [6].

Для розв’язання цієї проблеми аналітики використовують конвеєр обробки, де знаходиться вихідне дані збираються, очищаються та перетворюються у стандартизований формат. Дані можуть бути зберігається в системі комерційних баз даних і остаточно обробляється для аналізу з використанням аналітичні методи. Насправді хоча аналіз даних часто викликає поняття аналітичних алгоритмів, реальність така, що переважна більшість робіт пов'язана з підготовкою даних частина процесу [6].

З аналітичної точки зору, видобуток даних є складною через велику різницю у проблемах та типах даних, які виникають. Наприклад, комерційний товар Проблема рекомендації сильно відрізняється від програми виявлення вторгнень, навіть у рівень формату вхідних даних або визначення проблеми. Навіть у межах суміжних класів проблем, відмінності досить значні. Наприклад, рекомендація щодо товару. Проблема багатовимірної бази даних сильно відрізняється від соціальної рекомендації проблема через різницю в базовому типі даних. Проте, незважаючи на це відмінності, додатки для інтелектуального аналізу даних часто тісно пов’язані з однією з чотирьох „надзадач” в інтелектуальному аналізі даних: видобуток моделей асоціацій, кластеризація, класифікація та винесення виявлення [6].

Дані можуть мати різні формати або типи. Тип може бути кількісним (наприклад, вік), категоріальні (наприклад, етнічна приналежність), текстові, просторові, часові або графічно орієнтовані. Хоча найбільше Звичайна форма даних багатовимірна, дедалі більша частка належить до більш складних типів даних. Хоча існує концептуальна переносимість алгоритмів між багатьма типами даних на дуже високому рівні це не так з практичної точки зору. Реальність така точний тип даних може суттєво вплинути на поведінку певного алгоритму. Як результат, можливо, доведеться розробити вишукані варіанти

базового підходу для багатовимірності дані, щоб їх можна було ефективно використовувати для іншого типу даних. Тому ця книга буде присвячуйте різні глави різним типам даних, щоб краще зрозуміти як впливає на методи обробки базовий тип даних [6].

3.2 Text Mining

Одним із різновидів Data Mining являється Text Mining. Під Text Mining підрозумляється процеси вискреблювання знань і високоякісної інформації з текстових масивів. Вискреблювання знань з текстів - це процес знаходження нових, потенційно корисних і простих шаблонів в неструктурованих текстових даних - в наборі документів, що представляють собою логічно об'єднаний текст без будь-яких обмежень на його структуру [6].

- веб-сторінки,
- електронна пошта,
- нормативні документи,
- мобільні текстові повідомлення і т.д.

Такий вид глибинного аналізу текстів здатна «просівати» великі обсяги неструктурованою інформації та виявляти з них тільки найважливіше, щоб людині не доводилося самому витрачати час на видобуток цінних знань «Вручну» [6].

По суті, Text Mining - це набір лінгвістичних, статистичних методів, а також алгоритмів машинного навчання, які здатні моделювати і структурувати інформаційний контент і текстові джерела з метою бізнес-аналітики, аналізу даних, досліджень [6].

Text Mining складається з наступних етапів:

1. Пошук інформації.
2. Попередня обробка документів.
3. Витяг інформації.
4. Застосування методів Text Mining.
5. Інтерпретація результатів

3.3 Web Mining

Мережа є унікальним явищем багато в чому, з точки зору свого масштабу, розповсюдженості і некоординований характер його створення, відкритість базової платформи та в результаті різноманітність застосунків, які він увімкнув Прикладами таких програм є електронна комерція, співпраця користувачів та аналіз соціальних мереж [6].

Окрім вмісту, доступного в веб-документах, використання Інтернету призводить до значного обсягу даних у вигляді журналів користувачів або веб-транзакцій. Там це два основних типи даних, доступних в Інтернеті, які використовуються алгоритмами майнінгу [6].

Web content information: Ця інформація відповідає веб-документам та посилання, створені користувачами. Документи пов’язані між собою за допомогою гіпертекстових посилань. Таким чином, інформація про вміст містить два компоненти, які можна видобути разом або ізольовано [6].

Document data: дані документа витягуються зі сторінок у Глобальній павутині.

Linkage data: Інтернет можна розглядати як масивний графік, на якому розміщені сторінки відповідають вузлам, а зв'язки відповідають краям між вузлами. Це Інформація про зв'язок може використовуватися різними способами, наприклад, для пошуку в Інтернеті або визначення подібності між вузлами [6].

Web usage data : ці дані відповідають шаблонам активності користувачів, які ввімкнено веб-додатками. Ці моделі можуть бути різних типів [6].

Web transactions, ratings, and user feedback: Інтернет-користувачі часто купують різні типи товарів в Інтернеті або виражають їх спорідненість із певними продуктами в форма рейтингу. У таких випадках поведінку покупців та / або рейтинги можна використовувати для висновків про уподобання різних користувачів. В деяких випадках, відгуки користувачів надаються у формі текстових відгуків користувачів, на які посилаються до як думки [6].

Web logs: поведінка перегляду користувачів фіксується у вигляді веб-журналів, які зазвичай підтримуються на більшості веб-сайтів. Ця інформація про перегляд може бути використовуючи для висновків про активність користувачів [6].

Ці різноманітні типи даних автоматично визначають типи загальних програм в Інтернеті. За погодженням з різними типами даних додатки також є орієнтований на вміст або використання [6].

Додатки, орієнтовані на вміст: документи та посилання в Інтернеті використовуються в різних додатках, таких як пошук, кластеризація та класифікація. Кілька прикладів таких програми такі:

- Data mining applications: веб-документи використовуються разом із різними типи програм обробки даних, такі як кластеризація та категоризація. Такий веб-портали часто використовують програми для упорядкування сторінок [6].
- Web crawling and resource discovery: Інтернет - це величезний ресурс знання про документи з різних предметів. Однак цей ресурс широко поширюється в Інтернеті, і його потрібно виявити та зберегти в одному місце для висновків [6].
- Web search: метою веб-пошуку є виявлення високоякісних, відповідних документів у відповідь на заданий користувачем набір ключових слів. Як буде видно пізніше, поняття якості та релевантності визначаються як зв'язком, так і змістом структура документів [6].
- Web linkage mining: у цих додатках фактичне або логічне представлення структури зв’язку в Інтернеті видобуваються для отримання корисної інформації. Приклади логічного представництва веб-структури включають соціальні та інформаційні мережі. Соціальні мережі - це пов'язані мережі користувачів, тоді як інформаційні мережі пов'язані мережі користувачів та об'єктів [6].

Usage-centric applications: Діяльність користувачів у Інтернеті видобувається для висновків. Різні способи видобутку активності користувачів такі:

Recommender systems: У цих випадках надайте інформацію про перевагу у формі будь-якої рейтинги товарів або поведінка покупців використовуються для надання рекомендацій іншим однодумцям [6].

Web log analysis: Веб-журнали є корисним ресурсом для визначення власників веб-сайтів відповідні моделі користувацького перегляду. Ці моделі можна використовувати для виготовлення умовиводи, такі як пошук аномальних шаблонів, інтереси користувачів та оптимальний Інтернет дизайн сайту [6].

Веб-сканери також називаються павуками або роботами. Основна мотивація для Інтернету сканування полягає в тому, що ресурси в Інтернеті розподіляються широко в усьому світі сайтів. Хоча веб-браузер надає графічний інтерфейс для доступу до цих сторінок у інтерактивним способом, повна потужність доступних ресурсів не може бути використана з використання лише браузера [6].

Веб-сканери мають численні програми. Найважливішим і найвідомішим додатком є пошук, в якому завантажені веб-сторінки індексуються, щоб надати відповіді запити ключових слів користувачів. У всіх відомих пошукових системах, таких як Google і Bing, працюють сканери, щоб періодично оновлювати завантажені веб-ресурси на своїх серверах. Такі сканери також називаються універсальними сканерами, оскільки вони призначені для сканування всіх сторінок на Інтернет незалежно від їх предметів або місцезнаходження [6].

3.4 Хід аналізу тексту

У даній роботі для розв’язання задач виявлення шкідливої інформації у текстах використано мову програмування R, яка є широко розповсюдженою, та

має у своєму розпорядженні прикладні пакети практично для будь-якого застосування, зокрема стосовно задач Data Mining. Додаткова зручність програмування мовою R забезпечується завдяки використанню середовища розробки програмного забезпечення RStudio. В ній можливо підключати різноманітні пакети функції яких спрощують роботу з кодом.

На початку роботи ми створимо файл з потрібною нам інформацією. Після чого підключили необхідні нам пакети:

- Tm пакет даних для пошуку частоти слів;
- ggplot2 пакет даних для візуалізації даних;
- biclust пакет даних для пошуку даних;
- cluster пакет даних для кластерного аналізу;
- igraph пакет даних для побудови графів;
- fpc пакет даних дляідентифікування кирилиці;
- pillar пакет даних для формування стовпців даних;
- SnowballC пакет даних для створення векторів з отриманих даних;
- Wordcloud пакет даних для створеня хмар слів;
- RColorBrewer пакет даних для додавання кольору в графіки;
- Rcampd пакет даних для розпізнавання різних форматів файлів.

Командою «inspect(docs[1])» ми отримуємо повну інформацію по нашому тексту, а саме кількість слів у блоці. А командою «writeLines(as.character(docs))» ми можемо прочитати наш текст. Як повний текст так і його фрагменти.

Після того як ми впевнилися, що документ правильно завантажився ми починаємо робити попередню обробку. В цій частині ми видаляємо цифри, великі літери, загальні слова та стоп-слова, розділові знаки, та підготовляємо текст до аналізу. Після чого видалимо деякі стоп-слова через команду:

«docs <- tm_map(docs, removeWords, c("случай", "понятие"))»

А також з’єднаємо деякі слова:

```
for (j in seq(docs))
{ docs[[j]] <- gsub("Информационная безопасность", "Информационная безопасность", docs[[j]])
  docs[[j]] <- gsub("компьютерные сети", "компьютерные сети", docs[[j]])
  docs[[j]] <- gsub("безопасность сети", "безопасность сети", docs[[j]])}
docs <- tm_map(docs, PlainTextDocument)
```

І в кінці видаляємо спільні закінчення в словах. Це необхідно для того щоб слова були зрозумілі для комп'ютера. Командою «docs <- tm_map(docs, PlainTextDocument)» ми закінчуємо попередню обробку даних і вказуємо програмі читати усю обробку як текстову частину.

Тепер ми починаємо аналізувати свій текст. А саме знаходимо ті слова які найбільше та найменше зустрічаються.

```
freq
   1   2   3   4   5   6    7    8    9   10   11   12   13   14   15   16   17   19   20
1712  307  103   62   27   23   8   9   10    5   6   5   3   2   1   3   1   1    1   22    1
```

Рисунок 3.1 – Таблиця частоти слів

В цій таблиці ми бачимо в верхньому рядку кількість слів, а в нижній частоту появлення цих слів.

Далі виводимо 20 найбільш часто використовуючих слів.

```
                        word freq
безопасности     безопасности   97
информации         информации   69
информационной информационной   65
защиты                защиты   42
организации       организации   34
англ                     англ   23
```

Рисунок 3.2 – Таблиця найбільш використаних слів

За допомогою цієї таблиці ми можемо дізнатися усі найчастіше використані слова.

Далі ми створюємо графік найчастіше використаних слів

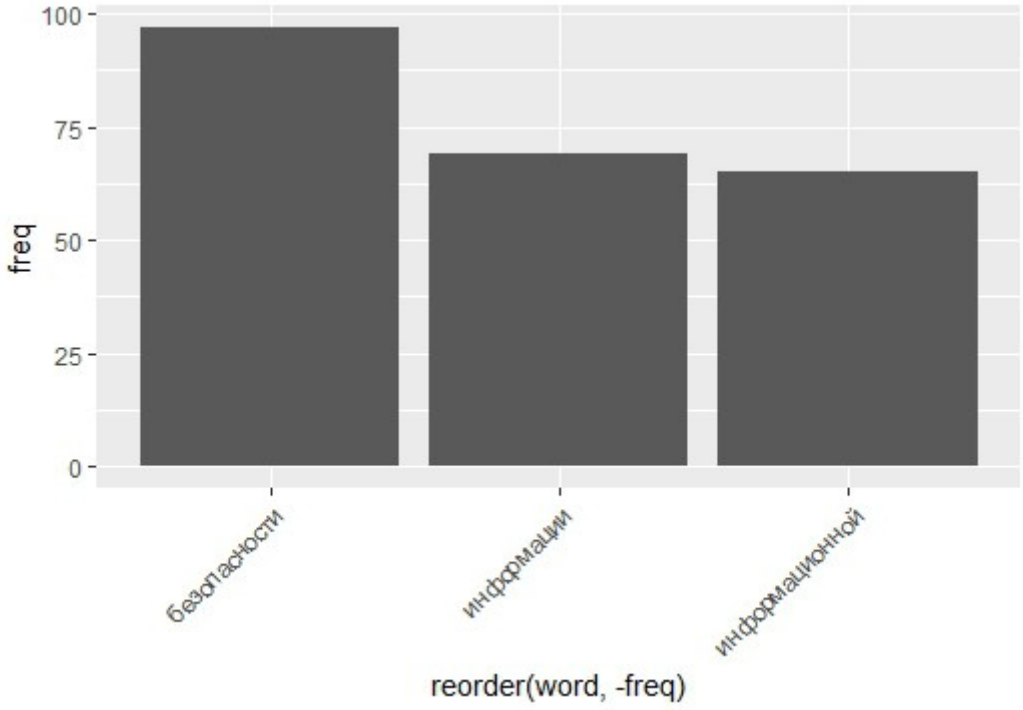

Рисунок 3.3 – Графік частоти слів

Тим самим ми можемо наглядно побачити найчастіші використані слова. Далі для більшої візуалізації ми створимо хмару слів.

```
set.seed(142)
dark2 <- brewer.pal(6, "Dark2")
wordcloud(names(freq), freq, max.words=100, rot.per=0.2, colors=dark2)
```

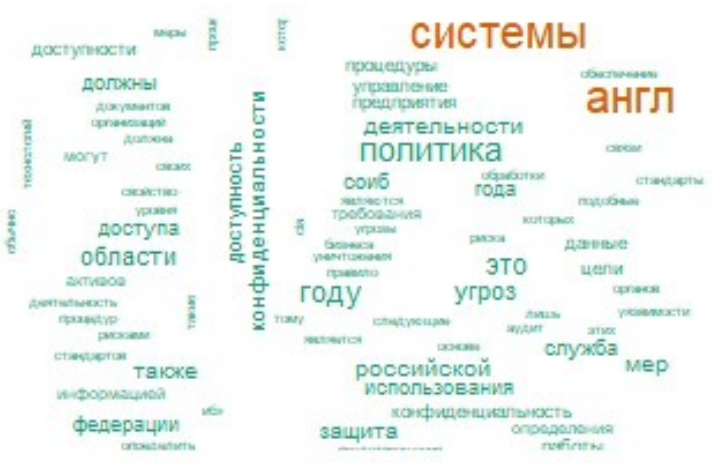

Рисунок 3.4 – Хмара слів

Тепер зробимо кластерізацію за схожістю термінів, а саме ієрархічну кластерізацію. Спочатку ми видаляємо усі слова які нас не цікавлять, а також ті які дуже рідко зустрічаються. А після чого вирахуємо відстань між словами та згрупуємо їх за подібністю.

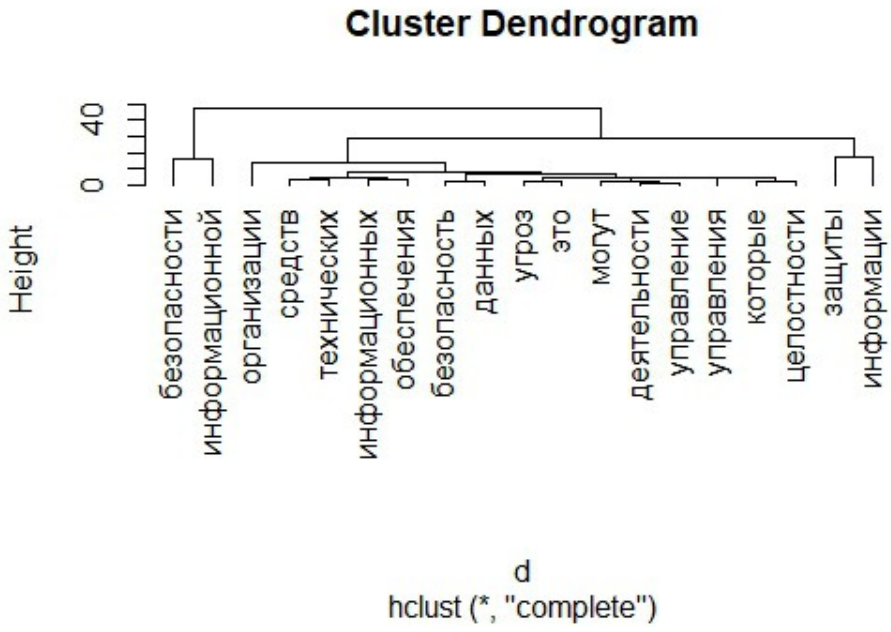

Рисунок 3.5 – Дейдаграма частоти слів

Так як у деяких людей може виникнути проблеми з читанням дейдаграм ми зробимо кластерізацію яка виділе слова по групам.

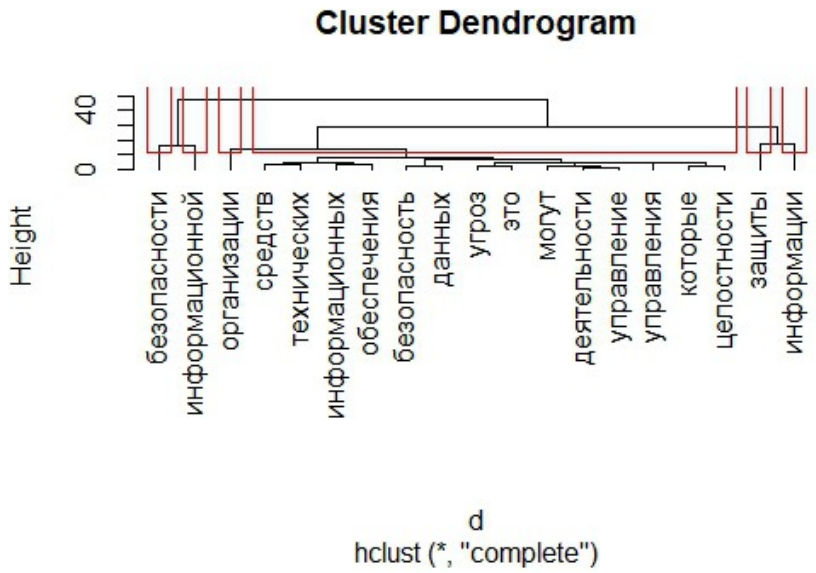

Рисунок 3.6 – Кластерізація дейдаграми частоти слів

В кінці можем проаналізувати текст та вивести К-кластеризаію. Цим ми згрупуємо слова в певну кількість груп (у нашому випадку 2), так що сума квадратів відстаней між окремими словами та одним із центрів групи.

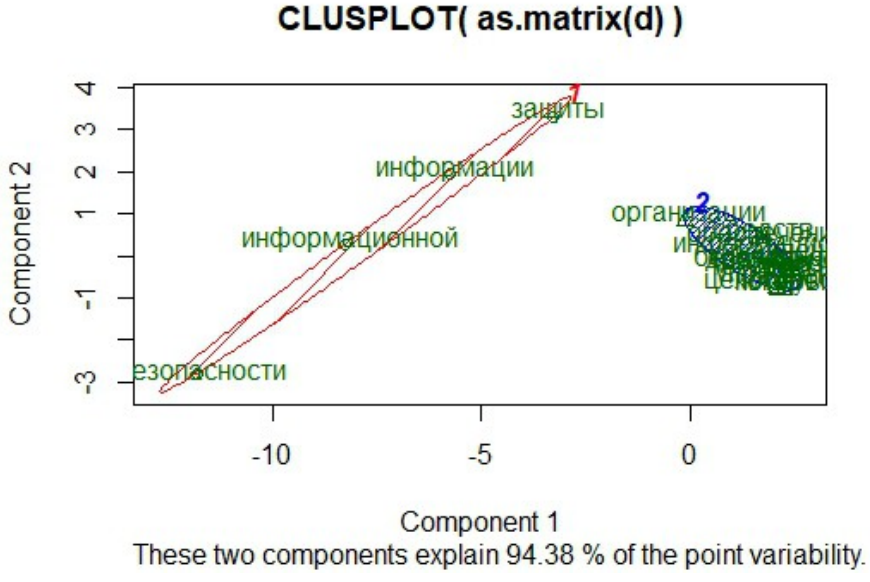

Рисунок 3.7 – К-кластеризація слів

Далі проведемо пошук по тексту нецензурної лексики(для даної роботи нецензурну лексику замінемо на сленгові слова). Для цього була підключен пакет tm після якого робимо запит:

```
 txta <- scan("Адреса тексту", "")
txta
```

Після чого отримуємо нумерацію кожного слова. В подальшому це допоможе для знаходження небезпечних слів.

Далі проводимо пошук по певнх небезпечних слів, а також більш детальніший пошук по частині слова. Для цього ми вводимо код:

```
for (j in seq(txta)) {  txta[[j]] <- gsub("/", " ", txta[[j]])
  txta[[j]] <- gsub("@", " ", txta[[j]])
  txta[[j]] <- gsub("\\|", " ", txta[[j]])
  txta[[j]] <- gsub("\u2028", " ", txta[[j]])}
grep("ништяк",txta)
grep("заточить",txta)
j<- grep("ништяк",txta)[1] j
gregexpr("захавал",txta)[j]
x<-txta
```

Проводимо аналіз аналіз двох текстів на наявність нецензурних слів. В першому тексті отримаємо такий результат.

Рисунок 3.8 – Результат аналізу першого тексту

Як видно по результату в цьому тексті відсутні шкідливі слова. Тому

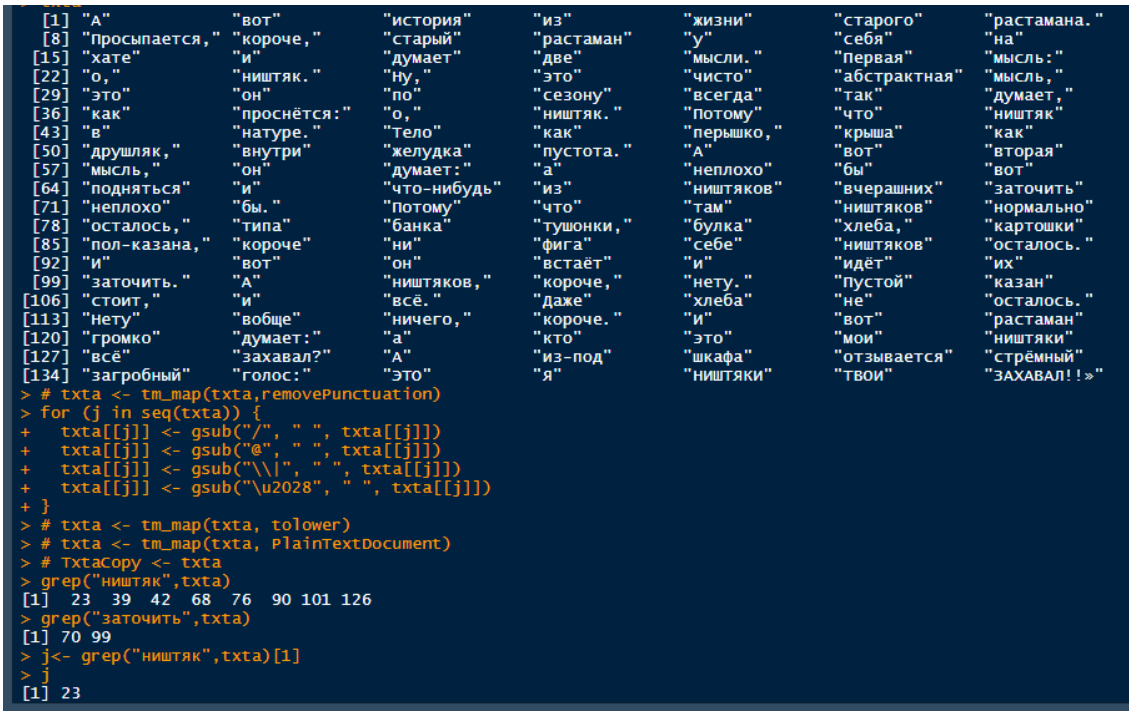

Рисунок 3.9 – Результат аналізу другого тексту

В даному тексті знайдені шкідливі слова які були представені в виді чисел(порядковий номер в тексті). Також після чого можемо провести пошук спільнокореневих небезпечних слів:

```
matches<-gregexpr('(?=((ниш)|зат))', x, perl=TRUE) set.match.length<-
  function(x) structure(x, match.length=as.vector(attr(x, 'capture.length')[,1]))
matches<-lapply(matches, set.match.length)
mapply(regmatches, x, lapply(matches, list))
write.csv(m, file="FindWords.csv")
```

```
character(0)

$вот
character(0)

$он
character(0)

$встаёт
character(0)

$и
character(0)

$идёт
character(0)

$их
character(0)

$заточить.
[1] "зат"

$А
character(0)

$`ништяков,`
[1] "ниш"

$`короче,`
character(0)

$нету.
character(0)

$пустой
character(0)
```

Рисунок 3.10 – Пошук по частині слова

Завдяки цій функції іде перебір усіх слів на наявність необхідної частини.

3.5 Оцінка небезпечності сайту відносно отриманих даних

Отримавши усі дані спеціаліст може проаналізувати матеріал на наявність шкідливого або небезпечного матеріалу орієнтуючись на таблицю 2.1.

Такий вид аналізу прискорить роботу правозахисників перевіряти ресурси, так як щоденно з’являються мільйони нових даних. Ми сводимо величезний текст до звичайної дейдаграми або хмари слів тим самим виводимо найпопулярніші слова в тексті, а також структуруючи їх по частоті проявляння. Якщо ж при такому аналізі ми знаходимо підозрілі слова і після чого вже більш детальніше вивчаємо дані які містить даний ресурс, перевіряючи код ресурсу, хост, власника, а також відправляючи запит на перевірку інформації яка проходить через нього.

Якщо при більш детальній перевірці ми знайшли небезпечний матеріал тоді ми можемо поступити таким чином:

- винести попередження, при якому ресурс повинні знищити матеріал;
- блокувати ресурс.

4 АНАЛІЗ СПОСОБІВ БЛОКУВАННЯ ТА ОБХОДУ БЛОКУВАНЬ ЗЛОВМИСНИКАМИ І ОЦІНКА ЇХ ЕФЕКТИВНОСТІ

4.1 Обхід блокувань зловмисниками

На сьогоднішній день винайдено багато способів обходу блокування сайтів.

Далі буде розглянуто найпопулярніші способи через які можливо отримати доступ до заблокованих сайтів.

Анонімайзер для відвідування заблокованих сайтів. Найкраще всього почати з найпоширеніших ресурсів. Корисування в них дуже просте. Потрібно лише на сайт CGI-ргоху і ввести необхідний веб-адресу. Також потрібно пам’ятати що користуючись такими ресурсами потрібно дуже обережно і не вводити ніяких паролів.

Браузерні розширення для обходу блокування сайту. Ще один дуже просий спосіб обійти блокування. Все що потрібно від користувача це лише встановити даний додаок і ввмкнути його.

Проксі для доступу на заблокований сайт. Такий спосіб обходу блокування може бути як анонімним, так і легко відстежуваним. Слід знати що при використані проксі-серверу може уповільнити швидкість інтернет-з'єднання.

VPN сервіси. VPN – віртуальна приватна мережа. За допомогою цієї технології можливо отримати доступ до мережі через захищений канал. Принцип VPN в тому що він як і Проксі-сервера, змінює ваш IP-адреса на власний і при потрапляні на сайт показує змінений IP-адрес користувача.

VPN-сервіси діляться на платні та безкоштовні. Для доступу на прості заблоковані сайти (приклад Вконтакті, GitHub) вистачить функцій безкоштовних сервісів. А для більш кращої анонімності в мережі необхідно ставити платну версію.

Доступ до заблокованого сайту за допомогою кеша Google. Якщо ціль тільки прочитати текст або проглянути картинки на сайті без яких було змін тоді вистачить і пошукової системи. Це здатнє за допомогою того що Google зберігаєкеш версії усіх сайтів.

IP-адреса для обходу блокування сайтів. Особливість даного способу заключається в що ми робимо запит через IP-адресу. Так як часто при блокування використовують ім'я сайту може все спрацювати. Але слід відмітити, що спосіб дуже ненадйний.

Перекладач для перегляду заблокованого сайту. Принцип роботи даного способу в тому що ми потрапивши на сайт перекладача вводимо та вводимо адрес сайту. Натискаєте перевести і через кілька секунд завантажується заблокований сайт.

Використання DNS для обходу блокування. За допомогою шлюзу DNS можливо зайти в інтернет через OpenDNS, в обхід DNS-серверів вашого провайдера.

TOR для обходу блокування сайтів. Зазвичай даний спосіб використовується для анонімного серфінгу в DarkNet. Також слід враховуввати що трафік TOR найчастіше помітніший ніж звичайний

Провівши аналіз усіх вищеперерахованих способів можемо привести оцінку ефективності методів обходу блокування. Для оцінку були приведені такі критерії як:

- Апаратні. Які пристрої потрібні для блокування.
- Програмні. Які програми необхідні для блокування.
- Умови можливості блокування. Наскільки можливо заблокувати ресурс.
- Кваліфікація користувача.

Також значення, які присвоєні в таблиці, являються критерієм оцінювання ймовірності блокування. Чим нижче це значення тим простіше користувачу обійти блокування.

Таблиця 4.1 – Методи обходу блокування

Метод обходу блокування	Кваліфікація користувача	Програми	Апаратура	Умови можливостей розблокування
Коєфіцієнт	0,2	0,2	0,3	0,3
CGI-proxy	1	2	1	2
VPN	1	2	1	1
Cash Google	2	1	1	2
IP-address	2	1	1	2
Translator	1	1	1	2
DNS	2	2	1	1
TOR	1	2	1	1

Для розрахунку ймовірності найкращого блокування була використана формула:

$$V = \frac{\sqrt[n]{\prod_{j=1}^{n} w_{ij}}}{\sum_{k=1}^{n} \sqrt[n]{\prod_{j=1}^{n} w_{kj}}}, \quad (4.1)$$

де $\prod_{i=1}^{n} w_{ij}$ – добуток всіх елементів рядка;

$\sum_{k=1}^{n} \sqrt[n]{\prod_{j=1}^{n} w_{kj}}$ – загальна сума всіх добудків кожної строчки матриці ймовірностей;

V_i – ймовірність обходу блокування.

Також ми можемо розрахувати враховуючи коефіцієнти важливості:

$$\overline{V_i} = \frac{\sqrt[n]{\prod_{j=1}^{n} w_{ij} \times k_j}}{\sum_{k=1}^{n} \sqrt[n]{\prod_{j=1}^{n} w_{kj} \times k_j}}, \quad (4.2)$$

де k_j – коефіцієнт важливості.

Розрахувавши усі значення приведемо в виді табл. 4.2.

Таблиця 4.2 Результат розрахунку ефективності обходу блокування

Метод обходу блокування	V	$\bar{V}$
Коэфіціент	0.1532	0.15328
CGI-proxy	0.1288	0.12889
VPN	0.1532	0.15329
Cash Google	0.1532	0.15328
IP-address	0.1532	0.15328
Translator	0.1288	0.12889
DNS	0.1532	0.15328
TOR	0.1288	0.12889

На основі виконаних розрахунків можемо зробити висновок, що із усіх перечислених варіантів слід виділити VPN, Translator а також TOR.

4.2. Аналіз способів блокування сайтів та вибір кращого з них

Майже кожна атака APT має наступні етапи (рис 4.1).

Рисунок 4.1 – Стадії APT

Особливості APT:

1. Розширення повноважень та подальша експлуатація найчастіше триває не один рік, що набагато ускладнює їх детектування;

2. Типові заходи та засоби із захисту телекомунікаційних мереж та інформаційно-телекомунікаційних систем не діють проти APT з достатньою ефективністю, оскільки:

а) об’єкти атак APT – це чітко визначені організації, телекомунікаційна інфраструктура, технічні рішення тощо. Таким чином, атаки APT мають невеликий розголос та рідко викликають суспільний резонанс;

б) суб’єктами атак APT використовуються найновітніші хакерські техніки та технології, ціна яких зазвичай – велика для більшості звичайних комп’ютерних зловмисників. Атакуюча сторона не зупиняється при здійсненні APT, зіткнувшись з ситуацією, коли об’єкт атаки має досить надійний захист;

в) типові пристрої та технології захисту, що використовуються у телекомунікаціях, можуть лише затримати час проведення етапу 2, проте не ефективні на етапах 3 та 4;

г) для захисту від APT ефективні проактивні та комплексні методики захисту, які дозволяють виявляти та попереджувати етап 2 та подальші етапи. APT може змінювати характеристики.

З кожним способом пов'язані обмеження технічного і політичного характеру, а також наслідки, які необхідно враховувати при розгляді окремих видів блокування контенту (рис 4.2).

Далі буде розглянуто 5 основних методів блокування (подробиці наведені в додатку В):

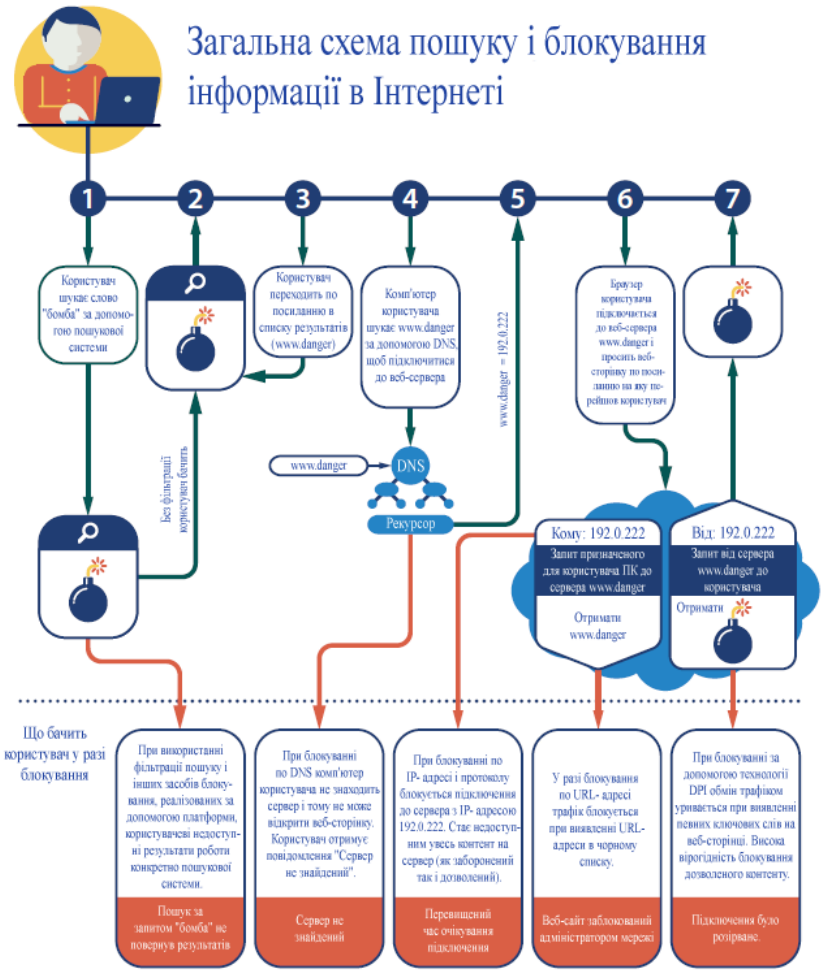

Рисунок 4.2 – Загальна схема пошуку і блокування інформації в Інтернеті

Блокування на основі IP встановлює бар’єри в мережі, такі як брандмауери, які блокують весь трафік до набору IP-адрес. Блокування на основі протоколу використовує інші ідентифікатори мережі низького рівня, такі як номер порту TCP/IP, який може ідентифікувати конкретну програму на сервері або тип протоколу програми. Ці найпростіші підходи до блокування вмісту насправді не блокують безпосередньо вміст, а блокують трафік до відомих IP-адрес або портів TCP / IP або протоколів, пов’язаних із певним вмістом або додатком. Блокування на основі IP та протоколу також може здійснюватися програмним забезпеченням на комп'ютерах користувача, як правило, з метою безпеки мережі [9].

Наприклад, якщо метою було заблокувати весь вміст, розміщений у міфічній країні Ельбонія, блокування IP можна було б використовувати, якби був відомий набір усіх IP-адрес, що розміщують вміст в Ельбонії. Так само, якби метою було заблокувати всі служби, блокування на основі протоколу могло б бути застосовано для зупинки служб VPN за допомогою відомих протоколів або номерів портів TCP/IP [9].

Різновидом блокування IP є зменшення трафіку. У цьому випадку не весь трафік заблокований, лише певний відсоток. Користувачі можуть сприймати послугу як дуже повільну або просто рухатись "вгору-вниз". Це може бути використано для того, щоб перешкодити користувачам користуватися послугою, зробивши її ненадійною, або заохотити використовувати альтернативні послуги, не виявляючи, що відбувається блокування [9].

Блокування на основі глибокої інспекції пакетів (DPI) використовує пристрої між кінцевим користувачем та рештою Інтернету, які фільтруються на основі конкретного вмісту, шаблонів або типів програм. Цей тип блокування мережі є обчислювально дуже інтенсивним і, отже, дорогим, оскільки весь вміст повинен оцінюватися на відповідність правилам блокування [9].

Для блокування DPI потрібен певний тип підпису чи інформація про вміст, щоб бути ефективними. Це можуть бути ключові слова, характеристики трафіку (наприклад, розміри пакетів або швидкість передачі), імена файлів або інша інформація, що стосується вмісту. Блокування DPI використовується дуже ефективно для блокування або регулювання певних [9].

Блокування DPI дуже часто використовується на підприємствах для систем захисту від витоку даних, продуктів, що захищають від спаму та зловмисного програмного забезпечення (антивірусів), а також пріоритетності трафіку (наприклад, підвищення пріоритету корпоративних відеоконференцій) для управління мережею [9].

Блокування на основі URL-адреси є дуже популярним методом блокування і може відбуватися як на окремому комп’ютері, так і в мережевому пристрої між комп’ютером та рештою Інтернету. Блокування URL-адрес

працює з веб-програмами і не використовується для блокування не-веб-програм (таких як VoIP). При блокуванні URL-адрес фільтр перехоплює потік веб-трафіку (HTTP) і перевіряє URL-адресу, яка відображається в запиті HTTP, щодо локальної бази даних або онлайнової служби. На основі відповіді фільтр URL-адреси дозволить або заблокує підключення до запитуваного веб-сервера [9].

Як правило, URL-адреси керуються категоріями (наприклад, «спортивні сайти»), а ціла категорія блокується, регулюється або дозволяється. У випадку, якщо національна політика вимагає блокування URL-адреси, політика он-лайн обслуговування та блокування, швидше за все, буде управлятися урядом. Фільтр URL-адрес може просто зупинити трафік або перенаправити користувача на іншу веб-сторінку, показавши заяву про політику або зазначивши, що трафік заблоковано. Блокування URL-адрес у мережі може застосовуватися як проксі-серверами, так і брандмауерами та маршрутизаторами [9].

Блокування URL-адреси вимагає, щоб сторона, що блокує (наприклад, Інтернет-провайдер користувача), мала можливість перехоплювати та контролювати трафік між кінцевим користувачем та Інтернетом. Блокування URL-адрес зазвичай дороге, оскільки фільтруючий пристрій, як правило, повинен знаходитись у строці між користувачем та Інтернетом, і, отже, вимагає високого рівня ресурсів для забезпечення прийнятної продуктивності [9].

У деяких випадках національні органи влади будуть співпрацювати з основними постачальниками інформаційних послуг, щоб заблокувати інформацію в межах свого географічного регіону, не блокуючи всю платформу. Найпоширеніші приклади фільтрування платформ - це провідні провайдери пошукових систем та платформи соціальних медіа [9].

Блокування на основі платформи – це техніка, яка вимагає допомоги власника платформи, наприклад оператора пошукової системи, такого як Google або Microsoft. У цій техніці запити від певного набору користувачів Інтернету до пошукової системи отримуватимуть різний набір результатів від

решти Інтернету - фільтрування покажчиків на вміст, який, певним чином, є несприятливим. У деяких випадках визначення того, що слід заблокувати, ґрунтується на місцевому регулюванні та державних вимогах, але це може також бути спричинене занепокоєнням оператора пошукової системи. Наприклад, пошукова система може блокувати вказівники на шкідливе програмне забезпечення або вміст, який вважається невідповідним відповідно до власних умов надання послуг [9].

Блокування пошукової системи впливає лише на користувачів, які вибирають конкретну пошукову систему, і лише тоді, коли користувачі визначаються як такі, що належать до певного набору з правилами фільтрації. Для блокування за віком, такого як Безпечний пошук (пропонується основними пошуковими системами та постачальниками вмісту), потрібно чітко визначити, чи потрібно це зробити [9].

Блокування вмісту на основі DNS дозволяє уникнути однієї з проблем, пов’язаних з іншими методами: вплив вартості та продуктивності фільтрації всього мережевого трафіку. Натомість блокування вмісту на основі DNS зосереджується на вивченні та контролі запитів DNS [9].

Завдяки блокуванню вмісту на основі DNS спеціалізований вирішувач DNS має дві функції: окрім виконання пошуку DNS, вирішувач перевіряє імена у списку блоків. Коли комп'ютер користувача намагається використовувати заблоковане ім'я, спеціальний сервер повертає неправильну інформацію, таку як IP-адреса сервера, що відображає повідомлення, що вміст заблоковано [9].

4.3 Аналіз та розрахунок найєфективнішого блокування

Після огляду усіх видів блокування сайтів можно зробити таблиці оцінки ефективності методів блокування з урахуванням різних рівнів. Для оцінку були приведені такі критерії як:

- Апаратні. Які пристрої потрібні для блокування.
- Програмні. Які програми необхідні для блокування.

- Умови можливості блокування.
- Кваліфікація користувача.
- Моніторинг. Як часто потрібно перевіряти заблокованість ресурсу.

Також значення, які присвоєні в таблиці являються показником оцінювання ймовірності блокування. Чим нижче це значення тим простіше користувачу обійти блокування.

Таблиця 4.3 – Оцінка різних видів блокування сайтів державою

Вид блокування	Кваліфікація користувача	Програми	Апаратура	Умови можливостей блокування	Моніторинг
Коєфіцієнт	0,1	0,3	0,3	0,2	0,1
Блокування по IP-адресу і протоколу	2	1	1	1	1
Блокування за допомогою технології DPI	2	2	3	2	2
Блокування по URL-адресу	2	2	3	1	2
Блокування за допомогою платформи	1	2	2	1	2
Блокування по DNS	1	1	1	1	1

Таблиця 4.4 – Оцінка різних видів блокування сайтів власником мережі

Вид блокування	Кваліфікація користувача	Програми	Апаратура	Умови можливостей блокування	Моніторинг
Коєфіцієнт	0,1	0,3	0,3	0,2	0,1
Блокування по IP-адресу і протоколу	1	1	1	1	1
Блокування за допомогою технології DPI	1	2	3	2	2
Блокування по URL-адресу	1	2	3	1	2
Блокування за	1	2	2	1	2

допомогою платформи					
Блокування по DNS	1	1	1	1	1

Таблиця 4.5 – Оцінка різних видів блокування сайтів користувачем

Вид блокування	Кваліфікація користувача	Програми	Апаратура	Умови можливостей блокування	Моніторинг
Коєфіціент	0,2	0,3	0,3	0,1	0,1
Блокування по IP-адресу і протоколу	2	1	1	1	1
Блокування за допомогою технології DPI	2	1	1	2	2
Блокування по URL-адресу	2	2	3	1	2
Блокування за допомогою платформи	1	1	1	1	2
Блокування по DNS	2	1	1	1	1

Після проведених розрахунків можемо створити таблицю в якій вкажемо ймовірність блокування V як з врахуванням коєфіціента важливості так і без нього.

Таблиця 4.6 Результат розрахунку ефективності блокування

Вид блокування	Блокування державами		Блокування підприємством		Блокування користувачем	
	V	$\bar{V}$	V	$\bar{V}$	V	$\bar{V}$
Блокування по IP-адресу і протоколу	0.1488	0,149	0,1418	0,141	0,1829	0,1182
Блокування за допомогою технології DPI	0,281	0,28	0,2678	0,268	0,2414	0,241
Блокування по URL-адресу	0,2446	0,245	0,2332	0,233	0,210	0,21
Блокування за допомогою платформи	0,1963	0,196	0,215	0,215	0,1829	0,182
Блокування по DNS	0,1295	0,1295	0,1418	0,141	0,1829	0,182

Опираючись на отримані дані можно виділити що DNS блокування підходить для використання державою, підприємством чи користувачем. Після проведених розрахунків отримуємо такі результати що із усіх перечислених варіантів можемо виділити блокування по DNS. Також можливо виділити що для підприємства також підходить блокування по IP-протоколу. А для користувача підійтуть блокування по IP-протоколу та за допомогою платформ.

ВИСНОВКИ

Стрімкий розвиток інформаційних технологій поступово трансформує світ. Відкритий та вільний кіберпростір розширює свободу і можливості людей, збагачує суспільство. Але на жаль не уся інформація може нести користь людині. Тому необхідно відстежувати такі ресурси і блокувати. Широкі можливості з автоматизації цих процесів виникають при використанні засобів інтелектуального аналізу даних (Data Mining), зокрема Text Mining та Web Mining.

Метою даної роботи є виявлення ефективних методів виявлення та блокування веб-сайтів.

В роботі розглянуті засоби виявлення шкідливого контенту в текстових даних в Інтернеті: використання ненормативної лексики.

Проаналізовані способи виявлення небезпечного матеріалу у мережі Internet з застосуванням методів Data Mining. З використанням можливостей мови програмування r створено програму виявлення небезпечного контенту.

Для знаходження необхідного нам рішення використана мову програмування R.

В ході виконання роботи було узято два коротких текста та проаналізовано їх на наявність ненормативної лексики (для даної роботи шукав сленгові слова).

Як результат в першому тексті немає заборонених слів, а у другому є. Відносно цього результату для другого тексту ми можемо поступити так:

- винести попередження, при якому ресурс повинні знищити матеріал;
- блокувати ресурс.

Проаналізовані основні способи блокування веб-сайтів і за кількісним критерієм виконано їх порівняння. Було виявлено, що найкращим видом блокування по DNS.

ПЕРЕЛІК ПОСИЛАНЬ

1. Закон України «Про основні засади забезпечення кібербезпеки України» [Електронний ресурс] Режим доступа: www/URL: https://zakon.rada.gov.ua/laws/show/2163-19
2. «Стратегія кібербезпеки України», затверджена Указом Президента України від 15 березня 2016 року № 96/2016. [Електронний ресурс] Режим доступа: www/URL: http://zakon.rada.gov.ua/laws/show/96/2016#n11
3. «Как пользователи видят сайты: F- и Z- паттерны, диаграмма Гутенберга» 14 жовтень 2016 Денис Нарижный [Електронний ресурс] Режим доступа: www/URL:/ https://netology.ru/blog/users-site-patterns
4. «Top 10 Countries With Highest Internet Censorship in 2016»[Електронний ресурс] Режим доступа: www/URL: https://fossbytes.com/countries-highest-internet-censorship
5. «Websites blocked in mainland China» [Електронний ресурс] Режим доступа: www/URL:/ https://en.wikipedia.org/wiki/Websites_blocked_in_mainland_China
6. «Data Mining The Textbook» Aggarwal, Charu C. [Електронний ресурс] Режим доступа: www/URL: https://doc.lagout.org/Others/Data%20Mining/Data%20Mining_%20The%20Textbook%20%5BAggarwal%202015-04-14%5D.pdf
7. «North Korea accidentally leaks DNS for .kp: only» [Електронний ресурс] Режим доступа: www/URL:https://www.reddit.com/r/technology/comments/53mr05/north_korea_accidentally_leaks_dns_for_kp_only_28
8. «What the North Korean internet really looks like» [Електронний ресурс] Режим доступа: www/URL: https://www.bbc.com/news/world-asia-37426725
9. «Symantec's Internet Security Threat Report» [Електронний ресурс] Режим доступа: www/URL :https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-cryptojacking-modern-cash-cow-en.pdf
10. «Как и зачем защищать доступ в Интернет на предприятии» [Електронний ресурс] Режим доступа: www/URL : https://habr.com/company/cisco/blog/230515/
11. «Как заблокировать доступ к сайту на компьютере?» [Електронний ресурс] Режим доступа: www/URL : https://public-pc.com/kak-zablokirovat-dostup-k-saytu-na-kompyutere/

12. Методичні вказівки з виконання атестаційної роботи бакалавра для студентів усіх форм навчання напряму 6.050903 «Телекомунікації» по кафедрі «Інформаційно-мережна інженерія» [Електронний документ] / Упоряд. А.І. Костромицький. – Харків: ХНУРЕ, 2019. – 37 с.
13. Про захист інформації в інформаційно-телекомунікаційних системах [Електронний ресурс]. – 1994. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/80/94-%D0%B2%D1%80#Text.
14. Закон України "Про інформацію" [Електронний ресурс]. – 1992. – Режим доступу до ресурсу: https://www.tax.gov.ua/diyalnist-/dpa-i-gromadskist/normativno-pravova-baza-u-sferi/arhiv-normativno-pravova-baza/53366.html.
15. Закон України “Про електронні документи та електронний документообіг” [Електронний ресурс]. – 2003. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/851-15#Text.
16. Закон України “Про основні засади забезпечення кібербезпеки України”"" [Електронний ресурс]. – 2017. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/2163-19#Text.
17. Закон України "Про електронні довірчі послуги" [Електронний ресурс] // 2017 – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/2155-19#Text.