Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Системотехніки
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

ГЮИК.503200.007 ПЗ
(позначення документа)

Розробка та дослідження методів інтелектуального пошуку крос-медійного контенту
(тема)

Виконав: здобувач групи ІТПм-20-1
спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)
освітньої програми ОПП Інформаційні технології проектування
(повна назва освітньої програми)
Мірошник С. М.
(прізвище, ініціали)

Керівник проф. Гребеннік І.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри системотехніки _____ Гребеннік І.В.
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
Кафедра Системотехніки
Рівень вищої освіти другий (магістерський)
Спеціальність 122 Комп'ютерні науки
(код і повна назва)
Освітня програма ОПП Інформаційні технології проектування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Мірошнику Станіславу Миколайовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Розробка та дослідження методів інтелектуального пошуку крос-медійного контенту
затверджена наказом по університету від 8 листопада 2021 р. № 1663 Ст
2. Термін подання студентом роботи до екзаменаційної комісії 10 грудня 2021 р.
3. Вихідні дані до роботи Функції системи: виконання розумного пошуку крос-медійного контенту на основі виявлення неявних семантичних зв'язків, формування персональних рекомендацій та добірок медіа-контенту на основі розумного пошуку. Вибірки даних для навчання та перевірки розумного пошуку. Інтегроване середовище розробки PHPStorm.
4. Перелік питань, що потрібно опрацювати в роботі 4.1 Вступ, 4.2 Аналіз предметної області, 4.2.1 Аналіз сучасного стану галузі, 4.2.2 Обґрунтування актуальності дослідження та розробки методів інтелектуального пошуку крос-медійного контенту, 4.2.3 Постановка задачі, 4.3 Розробка інтелектуального методу пошуку крос-медійного контенту, 4.3.1 Дослідження методів інтелектуального аналізу даних, 4.3.2 Дослідження технологій семантичного анотування, 4.3.3 Обґрунтування необхідності використання семантичного анотування та методів інтелектуального аналізу даних, 4.3.4 Розробка інтелектуального методу пошуку крос-медійного контенту, 4.3.5 Висновки за розділом 2, 4.4 Експериментальні дослідження, 4.4.1 Розробка вимог до програмних засобів, 4.4.2 Обґрунтування вибору програмних засобів для розробки програмних засобів, 4.4.3 Розробка програмних засобів, 4.4.3.1 Модель бази даних, 4.4.3.2 Розробка інтерфейсу клієнтської частини, 4.4.3.3 Імплементація алгоритму пошуку, 4.4.4 Проведення експериментальних досліджень, 4.4.4.1 Характеристика вихідних даних, 4.4.4.2 Аналіз результатів дослідження

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) Слайди презентації: титул, мета роботи, аналоги алгоритму, опис об'єкта і предмета досліджень, постановка задачі, етапи дослідження, технологія семантичного анотування, методи інтелектуального аналізу даних, алгоритму пошуку крос-медійного контенту, вимоги до консольного застосунку, вимоги до веб-застосунку, експериментальні дослідження, представлення результатів, висновки.

6. Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	проф. Гребеннік І.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Отримання завдання кваліфікаційної роботи	8.11.2021	
2.	Аналіз завдання та аналогів сервісу розумного пошуку крос-медійного контенту	9.11.2021-10.11.2021	
3.	Аналіз літератури	11.05.2021 – 14.11.2021	
4.	Дослідження методів інтелектуального аналізу даних	15.11.2021-17.11.2021	
5.	Дослідження технології семантичного анотування	19.11.2021-20.11.2021	
6.	Розробка методу пошуку крос-медійного контенту	21.11.2021-26.11.2021	
7.	Розробка програмного засобу	27.11.2021 – 30.11.2021	
8.	Проведення експериментальних досліджень	01.12.2021	
9.	Оформлення пояснювальної записки та документації	02.12.2021 – 04.12.2021	
10.	Оформлення презентаційних матеріалів	05.12.2021	
11.	Представлення на рецензування	07.12.2021	
	Представлення кваліфікаційної роботи в ДЕК	10.12.2021	

Дата видачі завдання 8 листопада 2021 р.

Студент  Мірошник С.М.
(підпис)

Керівник роботи _____ професор Гребеннік І. В.
(посада, прізвище, ініціали)

РЕФЕРАТ

Записка пояснювальна: 64 с., 18 рис. 4 табл., 5 додатків, 27 джерел інформації.

НЕЯВНІ СЕМАНТИЧНІ ВІДНОШЕННЯ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, МЕДІА-КОНТЕНТ, МАШИННЕ НАВЧАННЯ, ВИДЕДЕННЯ НОВИХ ЗНАНЬ, ПОВ'ЯЗАНІ ДАНІ, РЕКОМЕНДАЦІЙНІ СИСТЕМИ, СЕМАНТИЧНА АНОТАЦІЯ

Об'єктом дослідження є цифровий медіа-контент.

Предметом дослідження є інтелектуальне виведення нових знань за медіа-контентом з використанням машинного навчання.

Мета досліджень: розробка та реалізація інтелектуального методу пошуку медіа-контенту з використанням семантичної анотації і проведення аналізу медіа контенту для створення нових зв'язків з іншими об'єктами.

Методи дослідження – аналіз теоретичного матеріалу, технічної літератури, методів машинного навчання, практична реалізація додатку і експериментальне дослідження ефективності його роботи.

У роботі результаті проведених досліджень вирішено задачу побудови неочевидних зв'язків між різними об'єктами медіа-контенту та розроблено програмні засоби для демонстрації виведення результатів.

Отримані результати використовуються для генерування рекомендацій для користувачів, виокремлення найбільш релевантного контенту для результатів пошуку, формування добірок контенту за ознаками схожості, виявлених під час дослідження.

В якості інтегрованого середовища розробки використано PHPStorm, мови програмування PHP та JavaScript. Пропонована розробка корисна для вирішення задачі створення розумного сервісу для пошуку медіа-контенту.

Галузь застосування – розповсюдження медіа-контенту.

ABSTRACT

Explanatory note: 64 p., 18 pic., 4 tables, 5 ann., 27 sources.

DERIVATION OF NEW KNOWLEDGE, IMPLICIT SEMANTIC RELATIONS, INTELLIGENT ANALYSIS OF DATA, LINKED DATA, MEDIA CONTENT, MACHINE LEARNING, RECOMMENDER SYSTEMS, SEMANTIC ANNOTATION

The object of development – digital media content.

The subject of development – development and software implementation of an intelligent method of searching for media content using semantic annotation and analysis of media content to create new connections with other objects.

The purpose of the work – software development for smart searching for media content using semantic annotation and analyze media content to create new relations between other objects.

Methods of working – analyze of theoretical material and literature, machine learning methods, application implementation and experimental research of the effectiveness of it work.

As a result of the research, the task if constructing non-obvious connections between a different object of media content was resolved and software was developed to demonstrate the results.

The results are used to generate recommendations for users, highlight the most relevant content for the search result, and configure selections of content on similarity signs which were selected during the investigation.

PHPStorm was used as an integrated development environment, programming languages PHP and JavaScript. The proposed development is useful for solving the problem of creating a smart service for distributing media content.

Scope – distribution of media content.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних;

ІАД – інтелектуальний аналіз даних;

МН – машинне навчання – machine learning;

ПЗ – програмний засіб;

ПО – предметна область;

СА – семантична анотація;

СКБД – система керування базами даних;

ШНС – штучна нейронна мережа;

IDE – інтегроване середовище розробки;

MVC - Model-View-Controller;

PHP – Hypertext Preprocessor (гіпертекстовий препроцесор);

SQL – Structured Query Language (мова структурованих запитів).

ВСТУП

Комп’ютерні технології, та процеси, що виконуються з їх допомогою вже давно мають глибоку інтеграцію у повсякденне життя людини, оскільки навколо себе ми маємо величезну купу комп’ютеризованих об’єктів та гаджетів, що використовуються щодня.

Комп’ютери з габаритних та важких в використанні машин, для опановування котрих необхідно було витрачати неймовірно велику купу часу та ресурсів перетворилися на кишенькові пристрої, що здатні виконувати мільярди операцій за лічені секунди. Наразі ці пристрої суттєво спрощують і покращують життя кожної людини.

Як наслідок, ми живемо у епоху безперервного споживання різнотипної цифрової інформації, в тому числі і медіа контенту.

З поширенням цифрового контенту у різних галузях обсяг споживання медіа-контенту зростає. Наразі є величезна купа сервісів, що містять зображення, графіки, відео та інші типи контенту. Лише відеозаписів на ресурсі YouTube в день переглядається більше ніж мільярд годин. В ході суттєвого зростання, що продовжується з кожним роком, з’явилася проблема пошуку необхідного контенту, оскільки він є розподіленим між мільярдами сервісів та серверів, і користувачам доводиться витрачати великі обсяги часу на їх дослідження та формування коректних, з їх точки зору, результатів.

Саме з цим пов’язаний активний розвиток та дослідження методів пошуку контенту, що використовуються надалі у сервісах розповсюдження, пошуку та його рекомендації, зокрема на ресурсах, що мають великі бази інформації.

Основною проблемою наразі є те, що активна фаза розробки більшості таких сервісів, а як наслідок і дослідження відповідних методів, завершилася багато років тому і для формування рекомендацій та виконання

пошуку ці сервіси використовують не всі можливості, які сформувалися в ході розвитку штучного інтелекту за останні роки.

В ході виконання даної роботи пропонується виконати дослідження поточних методів, що використовують для пошуку серед крос-медійного контенту на основі ознак схожості та запропонувати розв’язання класу задач інтелектуального пошуку цифрового крос-медійного контенту в різних предметних галузях в наслідок розробки алгоритму розумного пошуку, що дозволить більш точно пропонувати контент, орієнтуючись на потреби користувачів за допомогою виявлення неявних семантичних зв’язків між різноманітними ресурсами.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Аналіз сучасного стану галузі

Контент – це певний вміст чи інформація, що зазвичай створений людиною [1]. Виділяють кілька основних видів контенту:

а) відео-контент – це будь який формат контенту, що містить ознаки відео. Поширеними формами такого типу контенту є влоги, кінематографічні твори, анімовані GIF-файли, відео трансляції тощо;

б) текстовий контент – є інформацією, що сформовано у текстовому виді, як то сценарії, книги, дописи у соціальних мережах тощо;

в) аудіо-контент – це будь-який матеріал чи інформація, що споживається в наслідок прослуховування. До даного типу відносяться пісні, музичні композиції, аудіо-книги, подкасти, голосові повідомлення тощо;

г) ігровий контент – включає у собі ігрові локації, предмети, персонажів тощо;

ґ) зображення. Сюди можна віднести картини, зображення, інфографіки, презентації тощо.

Контент був присутній у житті людини майже завжди. Одним з перших прикладів контенту можна назвати наскальні малюнки, які є прикладом зображень. За допомогою цього люди навчилися зображати певні події та демонструвати їх іншим.

З часом почали з’являтися пісні, люди почали писати, для того щоб можна було контактувати між собою не тільки звуками, але й передавати певні знання через писемність – і це можна було вже вважати першим серйозним прикладом текстового контенту.

Крім того, людство почало малювати картини, ставити вистави. З часом було винайдено фотоапарат і можна було робити фотографічні

знімки. В подальшому почала з’являтися можливість запису рухомих кадрів і світ побачив перший приклад відео-контенту.

У 20 столітті у світі почали з’являтися перші цифрові комп’ютери [2], а з ним текстовий, а далі і аудіо-візуальний контент і таким чином фактично було закладено основу для розвитку цифрового контенту, яка є основою поточної роботи.

В ході даного дослідження переважно проводиться аналіз відео та аудіо контенту, проте, опосередковано, використовується і текстовий контент.

Існує велика кількість сервісів, які наразі використовують методи для інтелектуального пошуку крос-медійного контенту. До найбільш розповсюджених можна віднести:

а) сервіси для розповсюдження відео та аудіо контенту, які містять у собі в тому числі рекомендаційні системи. Для таких сервісів це актуально, оскільки використання подібних методів дозволяє суттєво скорочувати ресурсні витрати на пошук оптимальних рішень, та використовувати отримані результати в подальшому в рекомендаційних системах, при формування плейлистів тощо;

б) різноманітні ресурси, що займаються розповсюдженням текстового та візуального контенту, такі як блоги, різноманітні ресурси з новинами тощо. Для цих сервісів є актуальним активний пошук за вмістом контенту, формування його тематики для більш коректних результатів пошуку тощо;

в) сервіси, що займаються розповсюдженням та продажами зображень та іншого контенту, оскільки використання даного класу методів може суттєво підняти рівень продажів в наслідок покращення якості результатів;

г) програмні засоби, що орієнтуються на пошук розпізнавання зображень, використовуються для ідентифікації людей (як то FaceID), знаходження музичних творів тощо.

В ході виконання даної роботи експериментальна частина дослідження методів виконується на основі сервісів, що займаються

розповсюдженням аудіовізуального контенту, отже є сенс розглянути актуальний стан справ у даній галузі.

Зародження сучасних сервісів розповсюдження медіа контенту почали розвиватися в той час, коли певні великі компанії в області IT та розповсюдження контенту (Apple, Google, Netflix та інші), проаналізувавши потенціал даного ринку інвестували кошти у розробку та покращення власних ресурсів, з використанням яких вони могли б в подальшому вчасно зайняти своє місце на ринку.

Подібна концепція призвела до появи одного з найвідоміших сервісів у світі довгий час, що був розроблений компанією Apple – iTunes [4]. Фактично він є медіа плеєром, що застосовується для організації та відтворення аудіовізуального контенту. В рамках даного сервісу клієнту надається можливість придбати музику, серіали, фільми тощо.

В рамках даного сервісу вперше було розроблено своєрідну рекомендаційну систему, що, з певними модифікаціями використовується і наразі, а саме – Genius. Основною ідеєю даної системи є аналіз поточної медіатеки користувачів та вибудовування рекомендацій на основі порівняння її с медіатеками інших користувачів. Цей алгоритм має декілька цікавих особливостей.

По-перше, поточний алгоритм є достатньо добре оптимізованим за рахунок того фактору, що інформаціє про контент замінюється достатньо повільно, таким чином відсутня необхідність у безперервному аналізі контенту при кожному запиті зі сторони користувача. Цей фактор дозволяє зробити частину інформації про схожість контенту та його приналежність до певних груп статичною, вивільнивши таким чином чималу кількість ресурсів, що використовуються в ході аналізу.

По-друге, для виконання пошуку ознак схожості між об’єктами Apple використовує алгоритми пошуку інформації, що базуються на моделі векторного простору[5], виконуючи таким чином порівняння за ознакою схожості між поточним контентом з іншими його одиницями, розраховуючи

кут нахилу між ними (рис 1.1), де вектори d_1 та d_2 є документами, або ж прикладами контенту, що порівнюються між собою, а q – це пісня для якої відбувається пошук [3].

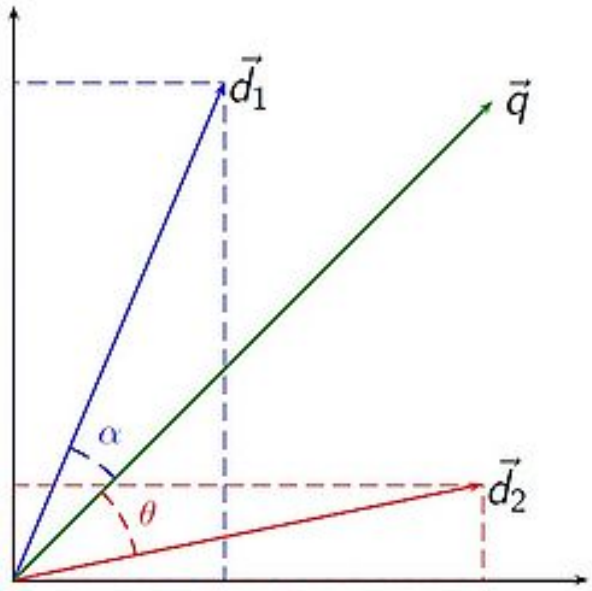

Рисунок 1.1 – Векторна модель

По-третє, в ході виконання процесу порівняння відбувається виокремлення факторів, для яких надаються вагові коефіцієнти. Для цього алгоритм виконує підрахунок частоти повторення контенту у користувача за специфічним набором факторів, як то жанр, ритм, виконавець та ін. для музичних творів та в подальшому використовує цю інформацію для порівняння її з медіатеками інших користувачів.

На сьогодні iTunes активно втрачає популярність, а прибутки компанії від нього скоротилися майже вдвічі порівняно з піковими показниками у 2012 році (рис. 1.2)[4]. Втрата популярності не пов’язана з поганою роботою алгоритму, оскільки Apple і надалі використовує його у своєму новому продукті. Причиною втрати є неактуальна для даного ринку модель розповсюдження контенту, у якій користувачі платять за кожну одиницю контенту.

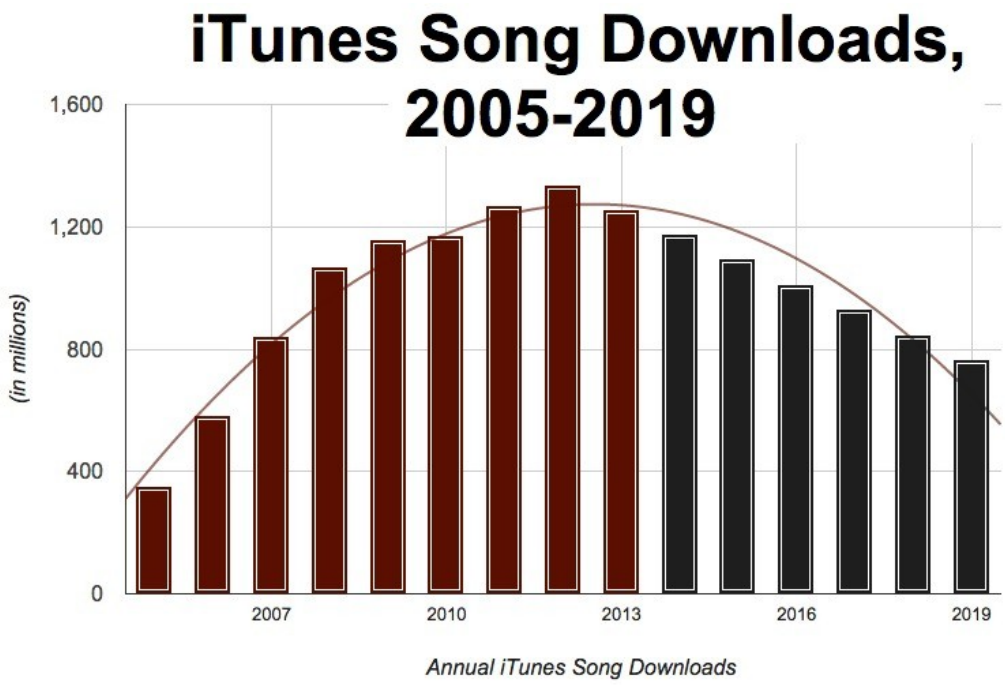

Рисунок 1.2 – Кількість завантажень пісень станом на 2013 рік

В якості аналога такій моделі продажів почала використовуватися інша модель, яка раніше вже використовувалася, наприклад, у друкарській сфері, коли читачі газет чи журналів виконували передплату за певний проміжок часу, в наслідок чого отримували паперові видання на час передплати. Цифровий контент почав розповсюджуватися за місячною підпискою, а користувач, який виконав передплату, отримав безлімітний доступ до всього наявного контенту.

До таких сервісів серед розповсюджувачів музики можна віднести Apple Music, Spotify, Youtube Music, серед розповсюджувачів відео контенту – Netflix. Саме після того, як з’явилися та отримали широке розповсюдження такі сервіси вперше за довгий час доходи музикантів почали зростати (рис. 1.3).

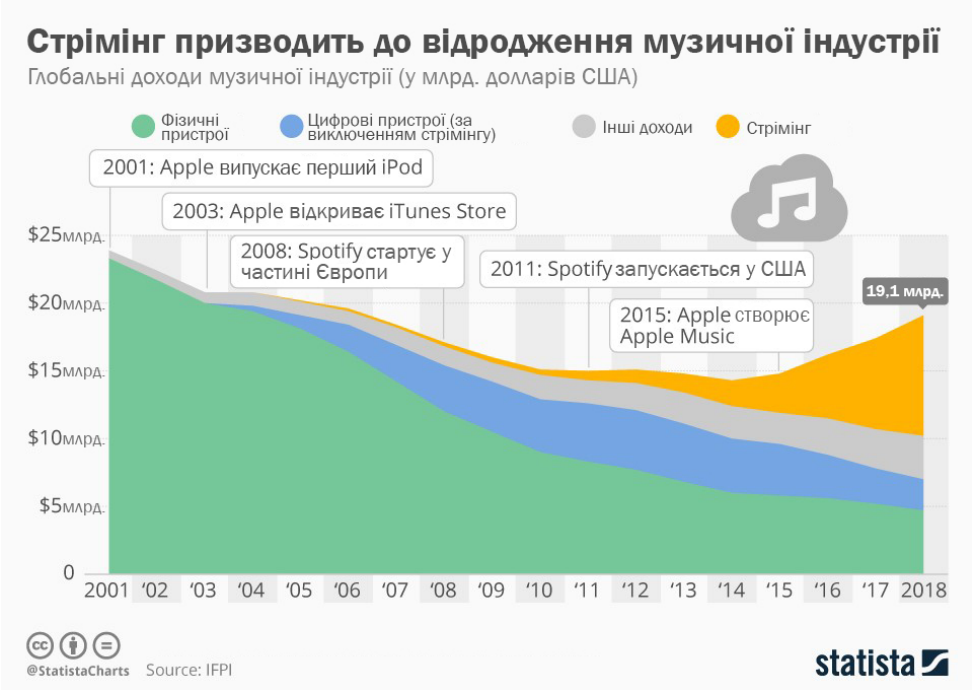

Рисунок 1.3 – Доходи музичної галузі

У кожного з даних гігантів IT індустрії є специфічні моменти, що пов’язані з рекомендаційною системою і кожен із них має свої позитивні і негативні нюанси. Розглянемо декілька прикладів.

Перший з них буде сервіс Youtube Music. Свого часу він пропагував ідею, що в переважній більшості випадків користувачам подобається новий контент, проте з часом сервіс почав більше спиратися на популярну музику, пропонувати контент, який вже прослуховувався велику кількість разів на відміну він спроб віднайти певний новий контент, який би міг потенційно зацікавити користувачів.

Негативною рисою даного сервісу також є принцип побудови систем рекомендації, що на початку генеруються певний «базовий рівень» для користувача, вирішаючи таким чином проблему холодного старту, а в подальшому цей рівень мінімально змінюється для позначення персоналізації сервісу [6].

Другим прикладом є Apple Music, і як вже було описано раніше, даний сервіс використовує векторну модель для пошуку схожого контенту для його подальшого використання в побудові музичних добірок.

Третім прикладом є Spotify, що також використовує векторну модель[5], пропонуючи користувачам твори певних виконавців, що мають ознаки схожості між собою. Для виконання пошуку даний ресурс додає додатковий фактор, що впливає на формування фінального результату, а саме алгоритм виконує аналіз наявності спільних слухачів. Як наслідок, у кількох груп, до прикладу, може бути 20000 спільних слухачів із загальної кількості у 50000, в такому випадку є велика вірогідність того, що іншим 30000 осіб система може порекомендувати групу [7].

Четвертим прикладом буде сервіс, що займається розповсюдженням відео контенту, а саме серіалів та фільмів – Netflix. Як і минулі застосунки даний сервіс використовує в своїй основі векторну модель, проте у нього є одна цікава особливість, яку варто розглянути детальніше [8].

Дана особливість полягає у спеціально розробленому нововведенні, що дозволило суттєво знизити обчислювальну інтенсивність алгоритмів і було розроблено працівниками компанії AT&T у 2009 році[3].

На початковому етапі обсяг обчислень, що виконувалися для порівняння між двома серіалами чи фільмами відповідав квадратичній функції, що масштабувалася в залежності від кількості порівнянь. Після імплементації розробленого алгоритму навантаження на обчислювальні центри знизилося на 10%.

Крім розглянутих вище прикладів можна перерахувати ще величезну кількість застосунків та сервісів, що виконують аналіз та пошук за медіа-контентом, визначають ступінь схожості та на його основі намагаються видавати результати, що найбільше відповідають встановленим на початку критеріям.

Проаналізувавши вищезазначене, можна зрозуміти загальний принцип побудови сучасних систем і підсумувати, що сучасні застосунки,

які мають за задачу виконання пошуку та визначення схожості цифровим контентом зазвичай звертаються до векторної моделі при формуванні рекомендацій, яка використовується для даних задач вже більш як 15 років.

1.2 Обґрунтування актуальності дослідження та розробки методів інтелектуального пошуку крос-медійного контенту

Наразі величезна кількість компаній змагаються між собою за право бути лідером у сферах, що, так чи інакше, зачіпають пошук різноманітного контенту. Такі фактори, як якість розпізнавання обличчя, швидкість пошуку продуктів за зображенням, актуальність згенерованих музичних добірок та багато іншого впливає на роль кожної великої IT компанії на ринку.

Незважаючи на сучасний розвиток методів машинного навчання, майже всі сучасні застосунки, що використовують методи для пошуку інтелектуального пошуку крос-медійного контенту, є заручниками векторної моделі. Дана модель обмежує розробників у намаганнях покращити якість і точність результатів пошуку. За останнє десятиліття в даній галузі не було помічено великих просувань уперед.

В наслідок цього компанії мають проблему з недосконалістю пошукових алгоритмів. Дану проблему можна було б вирішити розробкою нового алгоритму розумного пошуку в крос-медійному контенті.

В ході розробки алгоритму також варто враховувати і використовувати позитивні сторони сучасних сервісів, як то кешування, індексація тощо.

Тож, опираючись на дану інформацію, можна стверджувати, що застосування нових розумних методів дозволить вивести роботу застосунків на новий рівень шляхом розв’язання проблем покращення якості за допомогою покращення технології пошуку серед контенту. Ця задача наразі є актуальною і такою, що потребує детального вивчення.

1.3 Постановка задачі

Враховуючи інформацію, що була надана вище щодо поточного стану речей в контексті інтелектуальних методів пошуку крос-медійного контенту можна зазначити, що створення та розвиток продуктів у цій галузі є актуальним.

Наразі на ринку є велика кількість компаній, що займаються розробками та розповсюдженням сервісів, що мають в своїй основі один або кілька з існуючих методів для аналізу і пошуку контенту, проте ці сервіси виконують велику купу алгоритмів, що можуть бути покращені.

Виходячи з цього, можна сформувати задачу дослідження – розробити та здійснити програмну реалізацію інтелектуального методу пошуку медіа-контенту з використанням семантичної анотації, провести аналіз медіа контенту для створення нових зв’язків з іншими об’єктами.

Розроблений метод має відповідати наступним вимогам:

а) має отримувати набір текстових або числових параметрів, що описують медіа контент у певному форматі;

б) має виконувати обчислення семантичної відстані між екземплярами контенту;

в) має формувати на основі семантичної відстані вибірки відповідно до критеріїв пошуку;

г) має повертати результуючу вибірку.

Для досягнення зазначеної мети необхідно:

а) виконати аналіз існуючих методів анотування та методів обчислення семантичної відстані для їх подальшого використання при визначенні неявних семантичних відношень між одиницями контенту;

б) виконати аналіз існуючих методів інтелектуального аналізу даних в контексті можливості їх застосування для визначення результатів пошуку медіа-контенту;

в) обґрунтувати вибір визначених для розробки методів анотування та інтелектуального аналізу даних;

г) виконати розробку нового методу інтелектуального пошуку крос-медійного контенту;

ґ) виконати розробку програмного застосунку для проведення експериментальних досліджень;

ґ) провести експериментальні дослідження для визначення ефективності розробленого методу.

Оцінку ефективності розробленого методу варто розподілити на дві частини.

Першою частиною є оцінка повністю програмним шляхом. Для виконання даного порівняння треба розробити консольний програмний засіб. В ході даного оцінювання для розробленого методу треба:

а) сформувати навчальну та результуючі вибірки, що міститиме інформацію про медіа контент;

б) провести навчання;

в) виконати пошукові запити та оцінити ефективність, порівнявши отримані результати з існуючими даними про інші методів, визначивши час, за який виконано пошук для порівняння швидкості алгоритмів та відсоткове значення точності пошуку в порівнянні з результуючою вибіркою у вигляді числового значення від 0 до 100.

Другою частиною є суб’єктивна користувацька оцінка. Оцінка ефективності в ході такого дослідження проводиться шляхом отримання відгуків від користувачів застосунку, що реалізовує розроблений метод та матиме середнє числове значення від 1 до 5.

2 РОЗРОБКА ІНТЕЛЕКТУАЛЬНОГО МЕТОДУ ПОШУКУ КРОС-МЕДІЙНОГО КОНТЕНТУ

2.1 Дослідження методів інтелектуального аналізу даних

Інтелектуальний аналіз даних є процесом виявлення закономірностей та іншої цінної інформації на основі великих добірок даних з використанням методів машинного навчання, статистики та баз даних. Основна ідея інтелектуального аналізу даних полягає у використанні оптимізації, генетичних алгоритмів та ін. Результати даного аналізу в подальшому можуть бути використані для вирішення задач класифікації об’єктів, прогнозування певних подій та моделюванні [9].

Моделі інтелектуального аналізу даних використовуються для розв’язання цього ряду задач, до яких відносяться [10]:

а) задачі, що стосуються аналізу наявної інформації та формування на її основі очікуваних результатів, більш відомі як задачі прогнозування. Результати таких задач можуть бути використані для формування прогнозу рівня продажів;

б) задачі групування, до таких можна віднести вирішення проблеми групування комерційних пропозицій за певною тематикою для користувачів;

в) задачі знаходження послідовностей, що використовуються в тому числі для прогнозування дій користувачів у застосунках;

г) задачі обчислення ризиків. До таких можна віднести, наприклад, обчислення ймовірності повернення клієнтом банку боргових зобов’язань.

Використання методів інтелектуального аналізу даних є актуальним при вирішенні задачі розробки інтелектуального методу для пошуку крос-медійного контенту, оскільки присутня необхідність формування та вибору найбільш релевантного результату чи результатів пошуку.

Існує чимала кількість методів інтелектуального аналізу даних, що можуть бути використані в ході розробки нового методу інтелектуального пошуку. Розглянемо декілька таких методів.

Першим з таких методів є використання дерева рішень. Цей метод є найбільш зрозумілим для людей, що лише починають знайомитися з таким поняттям, як інтелектуальний аналіз даних. Крім того, даний метод достатньо часто використовуються в ході вирішення задач прогнозування або класифікації [11]. Виконаємо аналіз побудови дерева рішень, взявши за приклад формування результатів пошуку фільму за назвою «Once Upon a Time in Hollywood” (рис 2.1).

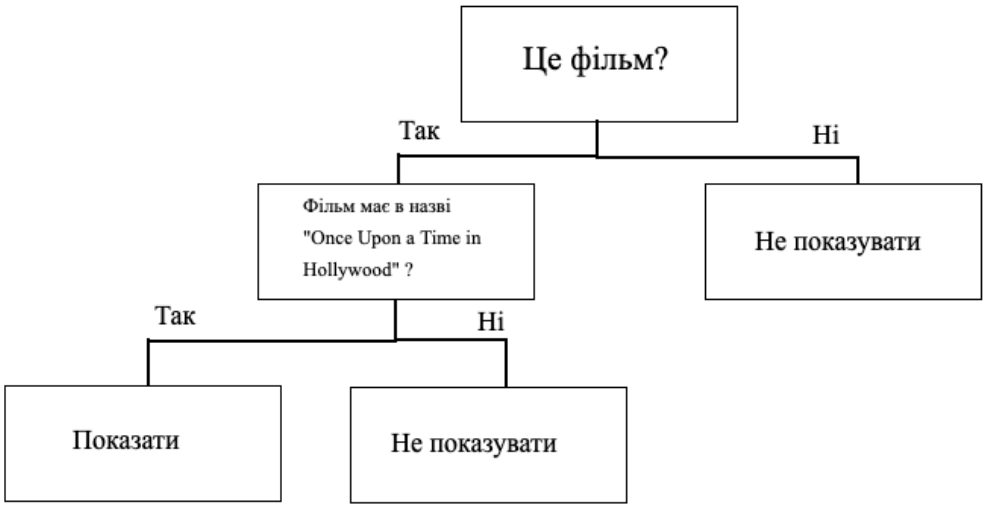

Рисунок 2.1 – Дерево рішень для пошуку фільму

Виходячи зі структури, що зображено (рис 2.1), в ході побудови дерева рішень використовуються лише два типи елементів:

а) вузли – включають у себе правила та виконують перевірку прикладів на відповідність певного атрибуту. Кожен із вузлів має своє правило і кілька варіантів вирішення. В результатів виконання перевірки у

вузлі відбувається перехід до одного з наступних вузлів. Даний процес продовжується до того часу, поки не буде досягнуто певного листка.

б) листки – визначають вирішення для прикладів, що дійшли до них, як до фінальної точки. Приклади потрапляють у лист, лише якщо відповідають усім правилам, що є на шляху до нього. До того ж, у кожен лист можна потрапити лише одним шляхом.

Даний алгоритм є так званим «жадібним» алгоритмом. Такий тип алгоритмів визначає, що при обранні оптимальних рішень по ходу обчислень в результаті буде сформовано ідеальне рішення [12].

Розглянемо формальний опис алгоритму дерева рішень.

Нехай є певна навчальна вибірка S, у якій є набір з різнотипного контенту і який має у собі m прикладів, кожен з яких має мітку класу $C_i(i = 1..k)$, та перелік із p атрибутів $A_j(j = 1..p)$, які відповідають за позначення приналежності прикладу до одного з класів, наприклад, тип контенту, автор, тематика тощо. В результаті можливі три варіанта розвитку подій:

а) у множині S існують приклади кожного з класів, що входять до множини C, таким чином вона включає у себе велику кількість прикладів різнотипного контенту (як то відео, фото, аудіо записи та ін.), які мають різні набори атрибутів. В такому випадку необхідно виконувати її розподілення на підмножини до тих під, поки кожен із прикладів у таких підмножинах не відповідатиме конкретному класу;

б) множина S пуста і не містить у собі ні одного прикладу контенту. За даних умов немає можливості виконувати навчання;

в) кожен з прикладів у множині S має ідентичну мітку класу C_i, таким чином поточна множина містить лише контент одного типу, що має ідентичний набір значень атрибутів. За таких умов відсутня необхідність проведення навчання через те, що кожен з прикладів, що міститься у множині в результаті матиме однаковий клас.

Серед позитивних сторін використання даного методу є:

а) немає необхідності виконувати нормалізацію вихідних даних, використовувати фіктивні змінні або видаляти пусті значення;

б) весь процес побудови фактично побудований на принципі бінарних дерев, таким чино для його опису можна використовувати булеву логіку;

в) простота методу і його зрозумілість для людей, які не пов’язані із галуззю машинного навчання.

Негативні сторони використання методу:

а) відсутність можливості застосування методу під час проектування та розробки великих застосунків, що містять великі обсяги даних, оскільки за таких умов побудова дерев буде дуже ресурсозатратною задачею, яка не матиме сенсу;

б) проблематичність формування оптимального дерева рішень.

Другим методом, що підлягає дослідженню, є використання кластеризації, а саме виконання розподілення вихідних об’єктів між кластерами за певним переліком ознак схожості [13].

Розглянемо детальніше принципи роботи даного методу та перелік кроків для його застосування[14]:

а) задати кількість кластерів x, до яких буде відбуватися розподілення об’єктів (у нашому випадків одиниць контенту);

б) рандомним чином вибрати із загального переліку t декілька об’єктів, що будуть використані в якості центрів кластерів. В ході виконання алгоритму кластеризації їх центри можуть змінюватися в залежності від обчислення центроїдів на подальших етапах;

в) провести послідовну вибірку кожного з об’єктів та виконати їх порівняння з обраними на попередньому кроці або підрахованими об’єктами, що є центрами кластерів з використанням обчислення відстані між ними, базуючись на значеннях їх атрибутів. Для обчислення відстані K використовується формула (2.1), міра схожості Чеканоського-Серенсена і Жаккара [13];

$$K = \frac{|t_1 \cap t_2|}{|t_1 \cup t_2|} \quad (2.1)$$

де t_1 і t_2 – множина параметрів, що описують об'єкти, а саме одиниці контенту, які порівнюються між собою, наприклад ритм, автор, жанр тощо.

г) провести розрахунок центроїдів кластерів за формулою (2.2), які обчислюються на основі середнього значення за кожною з ознак об'єктів у кластері. В результаті обчислення алгоритм повертається до кроку, що описано у пункті (в) та виконує повторення обчислень до того момент, поки точність результатів не досягне заданого рівня за формулою (2.3).

$$c_{ij} = \frac{\sum_1^{n_i} x_{ij}}{n_i}, j = 1,2, \ldots \quad (2.2)$$

де c_{ij} – множина центроїдів кластерів,

x_{ij} – множина об'єктів кластера,

n_i – кількість об'єктів, що належать до кластеру,

j – порядковий номер об'єкту у класетрі.

$$max_{ij}|K_{ij}(l+1) - K_{ij}(l)| < \varepsilon \quad (2.2)$$

де l – етапи інтерації ,

x_{ij} – множина об'єктів кластера,

ε – задана точність.

До позитивних сторін алгоритму кластеризації можна віднести точність результатів, так як він виконуватиме обчислення до того моменту, поки кордони сформованих кластерів не будуть незмінними.

До негативних сторін даного алгоритму варто віднести:

а) при його використанні при розв’язанні задачі класифікації медіа контенту необхідно на етапі початку обчислень мати вичерпний перелік параметрів, згідно з яким буде відбуватися розподілення та знати заздалегідь кількість кластерів;

б) ресурсомісткість поточного алгоритму через невідому заздалегідь кількість ітерацій, що необхідна для остаточного розподілення множини об’єктів.

Третім методом, або ж навіть групою методів для дослідження є штучні нейронні мережі.

Штучні нейронні мережі (ШНМ) представляють з себе обчислювальні системи, що натхнені звичайними біологічними нейронними мережами, які складають мозок тварин та людей. Отже, вони побудовані, базуючись на реальних біологічних мережах, так само складаються з нейронів, які є дуже спрощеною адаптацією біологічних нейронів.

Основна концепція роботи нейронів полягає у отриманні сигналів на вході, їх обробки згідно із певним порядком дій та видачі результатів на виході нейрону. Передача сигналів між нейронами відбувається з використанням синапсів, які пов’язують їх між собою [15].

Для початку проведемо дослідження простої ШНМ, використавши приклад оцінки чи буде дивитися фільм користувач (рис 2.2).

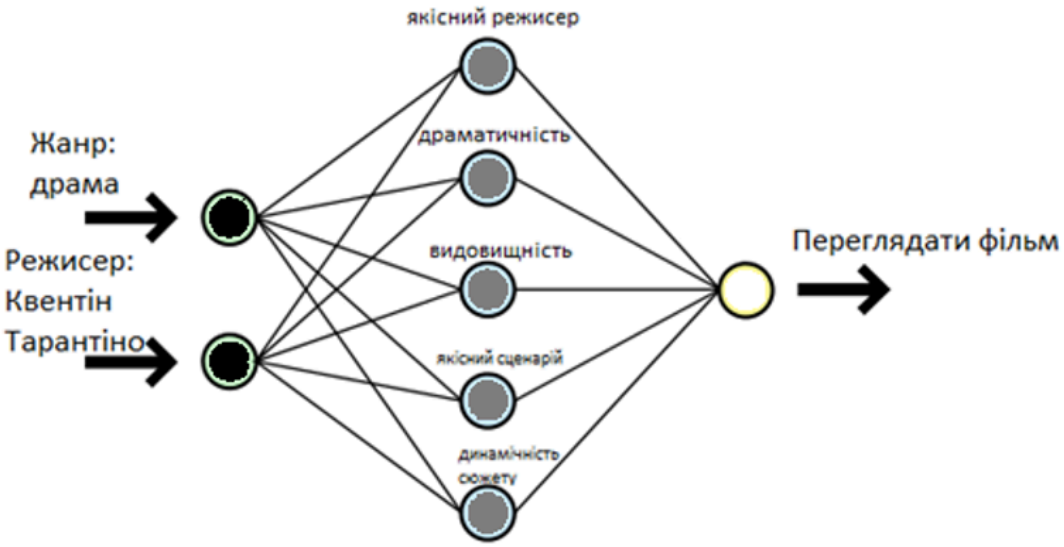

Рисунок 2.2 – Приклад штучної нейронної мережі

На зображенні (рис 2.2) зображено декілька типів нейронів. Перший з типів позначено чорним кольором, дані нейрони отримують сигнали з-за меж мережі, після чого виконують їх обробку та передають її результат за допомогою сигналів до нейронів, що знаходяться у прихованому шарі. Для нашого прикладу вхідними параметрами є жанр контенту та його режисер. Дані значення мають тестовий формат у тестовому прикладі для більш простого розуміння, при реалізації алгоритму такі значення зазвичай замінюються числовими.

В результаті аналізу і обробки вхідних значень режисера буде віднесено до певної групи, що матиме, наприклад, $v_1 = 5$ порядковий номер, множи, а жанр фільму в свою чергу матиме номер $v_2 = 3$.

Другим кроком сформований перелік атрибутів об’єкту буде передано у прихований шар ШНМ. Нейрони даного шару на рис. 2.2 позначено сірим. Даний шар складається з одного рівня, проте варто зазначити, що дуже часто кількість рівнів може бути суттєво більшою за 1. В ході формування ШНМ дуже часто у синапсів визначено значення їх ваги w, за допомогою

якої можна охарактеризувати рівень важливості кожного з сигналів, що буде подаватися для обчислення у нейрони. За відсутності ваги значення, що будуть отримувати на вхід нейрони матимуть для кожного з них однакове значення, що далеко не завжди є коректним при проведенні обчислень.

Розглянемо на прикладі застосування ваги w при обчисленнях у нейроні, що відповідає за драматичність фільму.

Для синапсу, що передає значення жанру встановимо вагу $w_1 = 0.87$ (87%). В свою чергу для синапсу, що передає значення про режисера використаємо значення $w_1 = 0.57$ (57%). В результаті ми визначили, що жанр фільму має дедалі більший вплив на його драматичність, ніж режисер, який його знімає. Виконає обчислення результат, що буде отримано у нейроні.

Першим кроком підрахуємо значення вхідних параметрів $H_{i\ input}$ (2.4), які будуть отримані в результаті підрахунку функції нейрону $H_{i\ output}$ (2.5). Для виконання даної операції нам необхідно задати функцію, згідно до якої буде проводитися підрахунок (2.6) та визначено значення, що буде використано як вихідний параметр та передано через синапс до наступного нейрона.

$$H_{i\ input} = (v_1 \times w_1) + (v_2 \times w_2) + \cdots + (v_n \times w_n) \tag{2.4}$$

де $v_1, v_2, \dots, v_n$ – множина значень із вхідних нейронів;

$w_1, w_2, \dots, w_n$ – значення ваги синапсів.

$$H_{i\ output} = f(H_{i\ input}) \tag{2.5}$$

$$f\left(H_{i\ input}\right) = \frac{H_{i\ input}}{n} \tag{2.6}$$

де n – це загальна кількість нейронів у поточному шарі.

Виконаємо обчислення вхідного значення:

$$H_{i\ input} = (5 \times 0.57) + (3 \times 0.87) = 2.85 + 2.61 = 5.46$$

Наступним кроком обчислимо значення функції у нейроні для отримання вихідного значення нейрона з використанням формули (2.5):

$$H_{i\ output} = \frac{5.46}{5} = 1.092$$

Таким чином вихідне значення нейрона, що відповідає за драматичність є 1.092.

Існує ще один тип нейронів, а саме вихідні нейрони. На схемі (рис 2.2) даний тип нейронів виділений білим кольором. Дані нейрони займаються обробкою результатів, що було отримано внаслідок обчислень у прихованому шарі та визначають результуюче значення.

Розглянемо приклад отримання результуючого значення роботи ШНМ. Відзначимо, що сума вагових коефіцієнтів вихідного шару не має перевищувати 1. Виконаємо обчислення результатів у прихованому шарі наступним чином:

а) якісний режисер – значення параметру $v_1 = 1.67$, при максимально можливому значенні 2, та ваговий коефіцієнт $w_1 = 0.21$;

б) драматичність – значення параметру $v_2 = 1.365$, при максимально можливому значенні 2, та ваговий коефіцієнти $w_2 = 0.14$;

в) видовищність – значення параметру $v_3 = 1.423$, при максимально можливому значенні 2, та ваговий коефіцієнт $w_3 = 0.25$;

г) якісний сценарій – значення параметру $v_4 = 1.782$, при максимально можливому значенні 2, та ваговий коефіцієнт $w_4 = 0.33$;

ґ) динамічність сюжету – значення параметру $v_5 = 1.252$, при максимально можливому значенні 2, та ваговий коефіцієнт $w_5 = 0.07$.

Виконаємо розрахунки вхідних значень нейрону, скориставшись формулою (2.4):

$$H_{i\ input} = (1.67 \times 0.21) + (1.365 \times 0.14) + (1.423 \times 0.25) + \\ +(1.782 \times 0.33) + (1.252 \times 0.07) \\ = 0.3507 + 0.1911 + 0.35575 + 0.58806 + 0.08764 = 1.57325$$

Таким чином вхідне значення нейрона дорівнює $H_{i\ input} = 1.57325$. Виконаємо розрахунок фінального значення. Для цього скористаємося формулою (2.7).

$$f(H_{i\ input}) = \frac{H_{i\ input}}{k} \tag{2.7}$$

де $H_{i\ input}$ – вхідне значення нейрона;

k – завчасно визначений коефіцієнт вихідного значення, який обчислюється із наступної формули (2.8)

$$k = \frac{(\sum_1^n \max(v_i))}{n} \tag{2.8}$$

де $\max(v_i)$ – максимально допустиме значення параметру;

n – кількість параметрів.

Обчислимо коефіцієнт для даного прикладу

$$k = \frac{(\sum_1^n 2 + 2 + 2 + 2 + 2)}{5} = 2$$

$$H_{i\ output} = \frac{1.57325}{2} = 0.786625$$

Результатом роботи ШНМ є значення 0.786625. Отриманий результат необхідно перевірити на відповідність умовам, відповідно до яких фільм може бути цікавим користувачеві. Якщо результуюче значення входить у чисельний діапазон або діапазони інтересів користувача – то цей контент може бути продемонстровано.

В результаті ми можемо виконати побудову ШНМ, яка буде видавати найбільш близькі результати.

До позитивних сторін нейронних мереж можна віднести:

а) ШНМ можуть працювати в умовах невизначеності;

б) ШНМ є достатньо стійкими до шуму, таким чином не треба виконувати обов’язковий аналіз даних перед початком роботи алгоритму;

в) ШНМ є в достатній мірі доступними для змін і адаптації під різні умови і системи;

г) структура ШНМ є дуже гнучкою, через те, що їх компоненти мають змогу до різноманітної взаємодії між собою.

До негативних сторін ШНМ відносяться:

а) ШНМ потребують великих проміжків часу та кількості об’єктів, що буде залучено до навчання для його успішного проведення [16];

б) результати, що отримуються в наслідок роботи ШНМ не є ідеально точними, адже базуються на близькому значенні.

2.2 Дослідження технологій семантичного анотування

Анотаціями ресурсів в більшості випадків є певний набір метаданих, що стосується даного ресурсу, тобто набір значень, через які відбувається його опис. Таким чином кожен ресурс має свій окремий набір значень, що його описує, відносить до певних груп, дозволяє ідентифікувати та виконати інші необхідні дії на основі даної інформації.

Загалом існує декілька основних видів анотацій [17]:

а) формальні. Даний тип анотацій має сформовану структуру, що в подальшому підлягає програмній обробці та є описаною за допомогою специфічних мов;

б) неформальні. Даний тип анотацій не може бути опрацьований програмно, оскільки не має завчасно визначеної чіткої структури, яку можна було б опрацювати, а його формування зазвичай відбувається за допомогою розмовної мови. З поточним розвитком технологій машинного навчання при використанні великого обсягу ресурсів можливо виконати його обробку, проте це потребуватиме великої кількості ресурсів;

в) онтологічні. Цей тип в основі має семантичну модель, що містить набір головних понять та формує відповідність об’єкту до певної ПО.

Враховуючи особливості інтелектуального методу пошуку та його потреби, найбільш ефективним із поточного перелік анотацій буде розгляд можливості використання онтологічної анотації. Даний вибір пов’язаний з тим, що крос-медійний контент містить велику базу метаданих, які можна буде оброблювати програмних шляхом, після чого, базуючись на них можна буде виконати аналіз кожного екземпляру.

Онтологічна модель фактично є знаковою системою, із застосуванням якої можливо виконувати оформлення семантичної анотації для досліджуваних об’єктів [18].

Одним із головних завдань, що має виконується в ході розробки інтелектуального методу пошуку є можливість класифікації контенту для отримання даних про нього з подальшим їх використанням в ході визначення вимог користувача та формування максимально коректного результату.

Суть семантичного анотування полягає у виконання формального опису кожного екземпляру об’єкту, що досліджується, для чого використовується множина кортежів. Таким чином формування метаданих виступає в ролі помічника при класифікації різних типів контенту, оскільки

даний тип анотування надає можливість виконати опис одиниці контенту з використанням лише набору кортежів.

Кортежі, що використовуються в ході виконання опису мають чітко відповідати наступній структурі (2.9) [19]:

$$< s, p, o > \tag{2.9}$$

де s – ідентифікатор суб'єкту;

p – ідентифікатор предикату;

o – ідентифікатор об'єкту.

Для повного розуміння суті кортежів виконаємо розгляд прикладу їх реалізації та порівняння між собою.

Наприклад, пісня "Blow" виконавців Ed Sheeran та Bruno Mars має жанри Hard Rock та Blues Rock, а пісня "Smells Like Teen Spirit" групи Nirvana має жанри Alternative Rock та Hard Rock. Використовуючи отриману інформацію ми можемо виконати формування кортежів для кожної з пісень:

а) <жанр Hard Rock, мати, пісня Ed Sheeran, Bruno Mars – "Blow">;

б) <жанр Blues Rock, мати, пісня Ed Sheeran, Bruno Mars – "Blow">;

в) <жанр Hard Rock, мати, пісня Nirvana – "Smells Like Teen Spirit">;

г) <жанр Alternative Rock, мати, пісня Nirvana – "Smells Like Teen Spirit">.

Виконаємо порівняння сформованих кортежів, для чого визначимо критерій схожості пісень базуючись на методі розрахунку семантичної відстані. Для виконання вимірів можна скористатися одним із наступних методів:

а) відстань Лемінга. Даний метод використовує принцип розрахунку мінімальної кількості дій для виконання видалення, вставлення або заміни символів у кортежах;

б) відстань Хеммінга. Даний метод виконує розрахунок загальної кількості позицій відповідно до яких можна відрізнити наявні концепти.

Враховуючи необхідність економії ресурсів раціональніше використати метод розрахунку відстані Хеммінга – L, що обчислюється за формулою (2.10).

$$L = \frac{\sum_{i=0}^{n} \frac{W_i}{maxW} \ (C_{i1}|C_{i2})}{N \sum_{i=0}^{n} \frac{W_i}{maxW}} \tag{2.10}$$

де W_i – ваговий коефіцієнт категорії в рамках домену;

maxW – вага категорії з максимальною вагою в рамках домену;

C_{i1} та C_{i2} – і-ті символи у кортежі;

N – кількість концептів у кортежі.

Домен в рамках крос-медійного контенту є параметром, за яким відбувається порівняння (як то жанр, автор, оцінка тощо). Концепт є значенням параметру. Категорія є формалізацією кортежу, та містить у собі суб’єкт, об’єкт та предикат.

Використавши формулу та визначивши, що усі концепти мають рівну вагу, виконаємо обчислення семантичної відстані між кортежами а та в, яке буде дорівнювати 0.667.

Застосування формули (2.10) є обґрунтованим у випадках розробки систем класифікації та пошуку крос-медійного контенту, через те, що за її допомогою є можливість, використавши вже існуючи об’єкти, що містять опис метаданих, виконати якісну класифікацію та чітко визначити до якої групи віднести поточний об’єкт для його подальшого використання.

Також важливо зазначити, що системне анотування може бути виконане декількома методами [20]:

а) автоматичне анотування. Даним метод виконується повністю автоматизовано з використанням програмних засобів, а анотація створюється лише на основі метаданих, що вже наявні у екземплярі;

б) ручне анотування. В ході використання даного методу усі метадані про об’єкт вносяться оператором або ж користувачем вручну, без застосування жодних засобів автоматизації;

в) напівавтоматичне анотування. Даний метод фактично є кооперуванням ручного і автоматичного анотування. При його використанні частина метаданих отримується автоматизовано з об’єкту, після чого оператор або ж користувач може додавати додаткові дані для поточної анотації.

2.3 Обґрунтування необхідності використання семантичного анотування та методів інтелектуального аналізу даних

В ході даного дослідження має бути розроблено новий інтелектуальний метод для пошуку крос-медійного контенту, що спростить вирішення наступного переліку задач:

а) формування результуючих добірок із контентом, що використовують для групування екземплярів контенту критерії схожості між ними;

б) покращення ефективності та якості роботи системи рекомендацій в ході отримання більш точних результатів на основі сформованих груп контенту та із врахуванням користувацьких вподобань;

в) виконання інтелектуального пошуку контенту, що в результатів повертатиме найбільш релевантні для користувача результати, які відповідатимуть критерію схожості екземплярів контенту до контенту, який користувач використовував раніше.

Для вирішення проблематики даних задач пропонується застосувати спільне використання методів інтелектуального аналізу даних та семантичного анотування.

Семантичне анотування надає можливість проаналізувати всі наявні ознаки контенту на основі його метаданих та використавши обчислення та аналіз семантичної відстані (2.10) і, базуючись на результатах даних обчислень, виконати групування екземплярів контенту у добірки, в рамках яких вони будуть пов’язані між собою за певним набором факторів.

Таким чином семантичне анотування забезпечить інтелектуальному методу можливість виконувати дослідження схожості контенту між собою та з пошуковим запитом для використання в подальшому отриманих даних в ході побудови методів пошуку або ж побудови роботи рекомендаційних систем.

Семантичне анотування може виконуватися трьома методами.

Перший з них – це ручне анотування. Даний метод є ефективним при роботі з невеликими обсягами даних та об’єктами, що не містять у собі метаданих. При реалізації методу інтелектуального пошуку він є неефективним, оскільки всю інформацію про кожну одиницю контенту необхідно вносити вручну.

Другим є автоматичне анотування, що дозволяє прибрати необхідність внесення інформації користувачами. Даний тип є ефективним за умови, що об’єкти гарантовано містять всю наявну інформацію, що необхідна в подальшому для обробки і їх ідентифікації. Поточний метод є більш актуальним за ручний, проте, враховуючи специфіку медіаконтенту є недостатньо ефективним, через те, що часто метадані контенту не містять всієї потрібної інформації.

Третім є напівавтоматичне анотування. Використання даного методу є найбільш релевантним в ході розробки інтелектуального методу пошуку, оскільки таким чином вирішується проблема недостатності метаданих та необхідності їх повністю ручного введення.

Для розв’язання задачі отримання результатів пошуку за пошуковим запитом або певними вподобаннями користувача пропонується використати один з існуючих методів інтелектуального аналізу даних. Пропонується обрати в якості такого методу один з методів, що було розглянуто у розділі 2.1, а саме метод побудови дерев рішень, метод кластеризації або ж штучних нейронних мереж.

В ході використання інтелектуального методу планується, що його основним завданням буде робота з великими обсягами даних, що матимуть велику кількість параметрів, згідно до яких буде виконуватися формування результатів.

Метод дерева рішень є дуже неефективним при роботі з великою кількість даних, оскільки для його реалізації в такому випадку необхідна велика кількість ресурсів, а отже його використання в ході розробки методу є недоцільним і необґрунтованим.

Метод кластеризації є дуже точним, проте, як і метод використання дерева рішень, вимагає великої кількості обчислювальних ресурсів, а отже його використання теж є недоцільним.

В свою чергу використання нейронних мереж є достатньо ефективним при реалізації даного методу, оскільки вони достатньо добре і ефективно працюють з великими обсягами даних і є більш економними в таких задачах. Крім того, медіа контенту часто може не мати певних даних, що використовуються в ході проведення аналізу, а ШНМ є достатньо стійкими до відсутності даних або ж шумів.

Таким чином, для вирішення завдання доцільним є використання напівавтоматичного методу семантичного анотування для реалізації методу пошуку схожості разом з використанням штучних нейронних мереж для обробки отриманих результатів.

2.4 Розробка інтелектуального методу пошуку крос-медійного контенту

В ході проведення дослідження було проведено розробку інтелектуального методу, що відповідає за групування одиниць контенту за певними ознаками та використовує сформовані групи в ході отримання результатів для систем пошуку та рекомендацій. Розроблений метод консолідує у собі методи семантичного анотування та штучних нейронних мереж.

Розроблений алгоритм поділяється на дві частини.

Перша його частина використовує обчислення коефіцієнту схожості між об’єктами в ході виявлення неявних семантичних відношень. Її реалізація та використання відбувається в ході завантаження нової одиниці контенту, після внесення змін до існуючого, а також періодично виконується для всієї бази контенту за певним графіком. Необхідність виконувати постійне обчислення схожості відсутня, оскільки дані про схожість змінюються або при зміні даних про одиницю контенту або в ході зміни смаків користувачів, що відбувається нечасто. Дана частина алгоритму має виконуватися за наступною послідовністю:

а) для виконання обчислення коефіцієнтів схожості обирається один екземпляр медіа контенту, на основі метаданих про який відбувається формування набору кортежів, який в подальшому використовується для порівняння з іншими екземплярами;

б) формуємо набір кортежів з іншими екземплярами, у яких значення суб’єкту хоча б у одному з наявних кортежів співпадає із значенням суб’єкту хоча б у одному кортежі поточного екземпляру;

в) виконуємо розрахунок семантичної відстані між сформованими для нього кортежами поточного екземпляру крос-медійного контенту та інших екземплярів з використанням для поточних обчислень методу знаходження відстані Хеммінга (2.10);

г) отримуємо результат, в ході якого виконано формування залежності між екземплярами на основі коефіцієнту схожості, що додатково зберігається для його подальшого використання.

В подальшому отримання значення схожості та сформовані зв’язки між одиницями контенту будуть використанні у другій частині алгоритму.

Друга частина алгоритму є варіативною відносно використання інтелектуального пошуку в рамках, безпосередньо, пошуку та рекомендаційних систем.

Першим кроком розглянемо використання алгоритму для задачі пошуку медіа контенту. В ході розв’язання даної задачі нам також треба врахувати дві варіації алгоритму, оскільки пошук медіа контенту може відбуватися у декілька шляхів:

а) пошук за існуючим медіа контентом;

б) пошуку за певним текстовим запитом.

В рамках реалізації пошуку за існуючим контентом потенційний користувач системи має завантажити або ж обрати одиницю контенту відповідно до якої буде відбуватися обчислення.

Алгоритм побудови результатів пошуку на основі існуючого медіа контенту буде мати наступну послідовність:

а) виконати завантаження та обчислення коефіцієнту схожості або ж скористатися вже проіндексованими і збереженими значеннями;

б) виконати формування множини схожого контенту, використавши обчислений коефіцієнт схожості;

в) виконати формування результатів за кожним з атрибутів медіа контенту, обчислюючи для цього вхідні та вихідні значення атрибутів, що будуть передані до нейронів ШНМ (2.11), а також значення вагових коефіцієнтів синапсів (2.12).

$$v = \frac{n}{N} \tag{2.11}$$

де n – повторюваність атрибуту з поточним значенням;

N – повторюваність атрибуту з будь-яким значенням.

$$w = \frac{n_p - n_n}{2 \times N} + 0{,}5 \tag{2.12}$$

де n_p – повторюваність атрибуту у контенті у результуючій добірці;

n_n – відсутність атрибуту у контенті у результуючій вибірці;

N – загальна кількість результатів.

г) виконати формування формул для обчислення значень у нейронах мережі(2.4)(2.5)(2.13);

$$f\left(H_{i\ input}\right) = \sum_{i=0}^{n} \frac{H_{i\ input} \times w_i}{N} \tag{2.13}$$

де $H_{i\ input}$ – вхідні дані нейрону;

w_i – ваговий коефіцієнт;

n – кількість вхідних сигналів;

N – кількість нейронів у шарі.

ґ) виконати обчислення значень та передати їх на вихідний нейрон, у якому виконати обчислення відношення одиниці контенту до результуючої групи;

е) виконати виведення результату.

Алгоритм інтелектуального пошуку може бути використано і для пошуку результатів за текстовим запитом. У такому випадку значенню кожного з атрибутів у кортежах буде задано значення пошукового запиту, в подальшому іншому ж алгоритм буде ідентичним до алгоритму пошуку результатів за існуючим контентом.

Другою варіацією другої частини алгоритму є можливість його використання в рамках побудови системи рекомендацій, що фактично є пошуком найбільш актуальних результатів відповідно до наявних даних про користувача.

Використаємо для аналізу і пошуку результатів дані користувача, а саме історію перегляду або використання одиниць контенту. Таким чином ми зможемо отримати достатній обсяг метаданих з контенту для формування певних вподобань користувача і на їх основі виконати пошуку та формування результуючих добірок для рекомендаційних систем.

Варіація алгоритму для розробки системи рекомендацій матиме наступний вигляд:

а) вибрати одиниці контенту з бази даних, які були використані користувачем;

б) використавши отриману вибірку сформувати множину зі схожими екземплярами медіа контенту, використавши для цього обчислений раніше коефіцієнт схожості;

в) отримати з бази даних дані про активність користувача, проаналізувавши які будуємо перелік вподобань за кожним з атрибутів елементів контенту. Для цього обчислимо значення атрибутів (2.10), а також розрахуємо значення вагового коефіцієнту (2.14).

$$w = \frac{n_p - n_n}{2 \times N} + 0{,}5 \qquad (2.14)$$

де n_p – повторюваність атрибуту у контенті з позитивною оцінкою;

n_n – повторюваність атрибуту у контенті з негативною оцінкою;

N – повторюваність атрибуту з будь-яким значенням.

г) виконати формування формул для обчислення значень у нейронах мережі(2.4)(2.5)(2.13);

г) обчислимо значення у нейронах та передамо результати на вихідний нейрон, у якому визначаємо чи входить отриманий екземпляр контенту чи множина екземплярів у інтереси користувача;

д) для отриманих результатів, в разі необхідності, генеруємо окремі добірки використовуючи обчислені раніше коефіцієнти схожості;

е) вивести отриманий результат.

Варто зазначити, що в разі побудови системи рекомендації існує велика проблеми, що пов’язана з відсутністю даних про користувачів, які ще не зареєстровані у рамках системи.

Дана проблема називається проблемою холодного старту. Вона існує через те, що неможливо точно отримати розуміння про вподобання користувача в разі відсутності будь якої інформації про нього.

Існує два основних способи вирішення даної проблеми:

а) розробити певну «базову версію» користувача, що має узагальнені інтереси і в подальшому налаштовувати систему аналізуючи дії користувача з одиницями контенту впродовж певного періоду користування системою;

б) надати можливість користувачам обрати з наданого фіксованого переліку вподобань. Таким чином алгоритм одразу видаватиме найбільш релевантні для користувача результати.

При розробці сервісів з використанням поточного алгоритму рекомендується використання другого вирішення проблеми холодного старту, оскільки таким чином сервіс надасть можливість отримання максимально якісної системи рекомендацій.

2.5 Висновки за розділом 2

У другому розділі даної роботи було виконано детальне дослідження існуючих методів інтелектуального аналізу даних, розглянуто позитивні та негативні сторони їх використання.

Було виконано детальне дослідження методів семантичного анотування в контексті їх використання в рамках задачі знаходження неявних семантичних зв’язків.

На основі досліджених методів було обрано метод використання штучних нейронних мереж та напівавтоматичного анотування з використанням відстані Хемінга для розробки власного інтелектуального методі пошуку крос-медійного контенту.

Було виконано розробку власного методу для пошуку контенту, що має варіації як для, безпосередньо, пошуку так і пошуку найбільш оптимальних результатів для рекомендаційних систем.

Вихідними даними для даного методу виступають наявні метадані про крос-медійний контент.

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

3.1 Розробка вимог до програмних засобів

В ході проведення експериментальних досліджень буде виконано розробку двох програмних засобів для порівняння ефективності розробленого методу інтелектуального пошуку.

Першим програмним засобом (ПЗ) є консольний додаток, що матиме реалізацію розробленого методу інтелектуального пошуку. Розроблений програмний засіб має включати реалізацію розробленого методу для проведення експериментальних досліджень.

Другим таким засобом буде сервіс пошуку крос-медійного контенту, що спеціалізуватиметься на розповсюдженні фільмографічних та музичних творів та використовуватиме розроблений метод. Він має відповідати наступним вимогам:

а) має забезпечити можливість створення персональних акаунтів користувача при отриманні вичерпної інформації про нього;

б) має забезпечити можливість авторизації користувачів;

в) має забезпечити зміни інформації про користувача (як персональної, так і даних для авторизації);

в) має забезпечити можливість перегляду фільмографічних та прослуховування музичних творів;

г) має забезпечити можливість оцінювання контенту;

ґ) має мати зрозумілий і доступний користувацький інтерфейс, який забезпечить швидку орієнтацію користувача у застосунку;

д) має забезпечити формування рекомендаційних вибірок, орієнтуючись на вподобання клієнтів;

е) має забезпечити виконання пошуку крос-медійного контенту;

є) має забезпечити можливість авторизації адміністратора;

ж) має забезпечити можливість додавання чи маніпулювання музичними та фільмографічними творами чи вибірками.

3.2 Обґрунтування вибору програмних засобів для розробки програмних засобів

Для розробки програмних засобів обрано мову програмування PHP, оскільки вона є зручною і ефективною при розробці веб-застосунків та відкидає необхідність додаткової реалізації методу на інший мові програмування[21].

До переваг даної мови можна віднести:

а) інтерпритованість мови програмування, яка надає можливість миттєвого внесення змін до великих проектів, виключаючи тривалі компіляції, скорочуючи таким чином час на розробку програмного засобу;

б) висока продуктивність, завдяки якій дана мова має суттєву перевагу у швидкості в порівнянні з великою кількістю інших мов програмування;

в) інтегрованість з великою кількістю платформ, ОС та серверів. Крім того дана мова програмування підтримує роботу з великою кількістю баз даних (БД), надаючи таким чином варіативність вибору.

В якості системи керування базами даних (СКБД) обрано MySQL, що базується на мові запитів SQL. Дана СКБД має наступні переваги при проведенні експериментальних досліджень [22]:

а) вона є безкоштовною при її застовуванні для розробки навчальних проектів;

б) підтримується абсолютною більшістю серверів, що надає можливість її легкого перенесення між ними;

в) має якісний захист даних, що гарантується системою розподілення прав доступу між користувачами;

г) має механізм багатопоточної обробки даних, надаючи можливість швидкої взаємодії з БД при великих навантаженях.

В якості IDE було обрані ПЗ компанії JetBrains – PHPStorm, оскільки дана компанія надає велику кількість сервісів для зручної роботи з проектами, системами контролю версій та мають безкоштовну ліцензію для використання в ході розробки навчальних проектів [23].

При розробці веб-застосунку обрано архітектурний шаблон Model-View-Controller (MVC), оскільки він надає можливість формування простої і зрозумілої архітектури проекту. В результатів його використання проект матиме наступні частини [24]:

а) моделі, що виконують обробку даних та відповідають за роботу з ними;

б) вигляди, що надають можливість взаємодії з користувачем, відповідають за відображення сторінок застосунків;

в) контролер, який відповідає за обробку запитів, що передаються користувачами з використанням вигляду. Контролер може відправляти запити до моделей для отримання або модифікації певного набору даних.

3.3 Розробка програмних засобів

В ході розробки програмних засобів буде реалізовано веб-застосунок для проведення тестування з реальними користувачами та консольний застосунок для проведення дослідження виключно програмним шляхом.

3.3.1 Модель бази даних

Для консольного застосунку відсутня необхідність зберігання даних, а яка наслідок і проектування БД через вичерпність вхідних та вихідних даних.

В свою чергу для проведення тестування з використанням веб-застосунку спроектовано модель БД (рис. 3.3)(рис. 3.4).

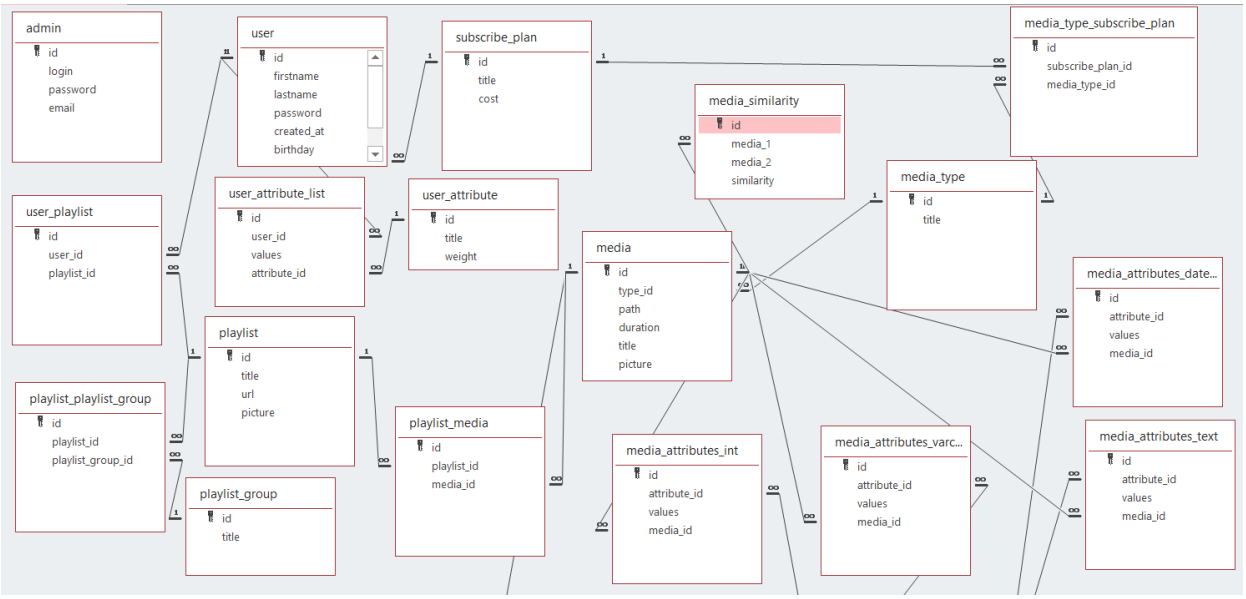

Рисунок 3.3 – Схема БД сервісу, частина 1

У БД буде зберігатися інформація про адміністраторів, користувачів, медіа контент, створенні та сформовані добірки (playlist).

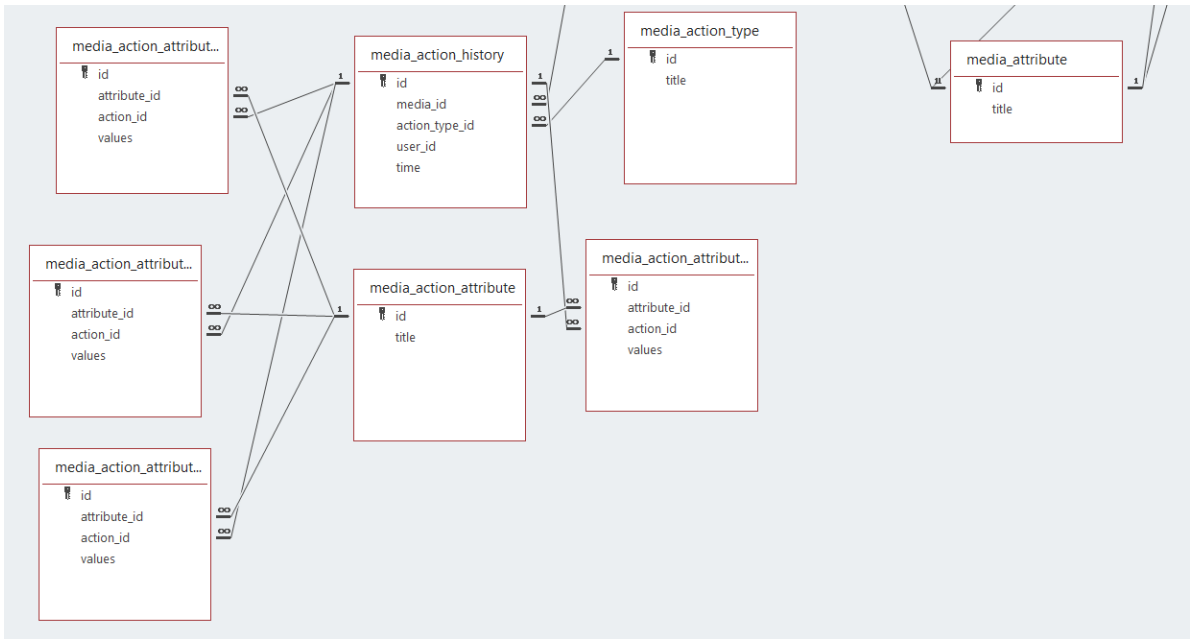

Рисунок 3.4 – Схема БД сервісу, частина 2

3.3.2 Розробка інтерфейсу клієнтської часини

Для консольного застосунку відсутня необхідність розробки інтерфейсу клієнтської частини. Для веб-застосунку проведено його розробку.

Однією з вимог до програмного засобу є простий і зрозумілий інтерфейс користувача [25].

Спочатку виконано розробку інтерфейсу для сторінок реєстрації та авторизації (рис 3.5). Дана сторінка містить меню, форму з полями для введення даних користувача та кнопку підтвердження дії.

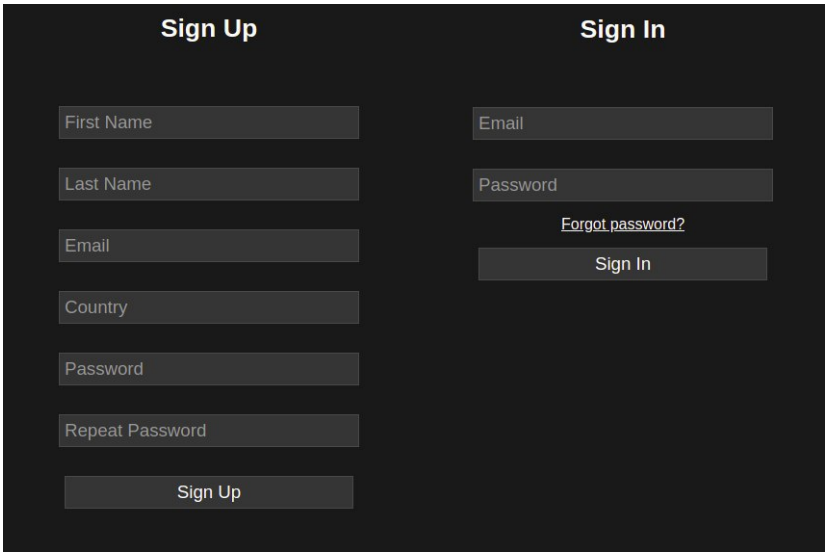

Рисунок 3.5 – Макет форм реєстрації та авторизації

Далі виконано розробку дизайну інтерфейсу головної сторінки (рис.3.6). Ця сторінка доступна для авторизованих користувачів і містить добірки з рекомендаціями музичних творів, меню та рядок для виконання пошуку.

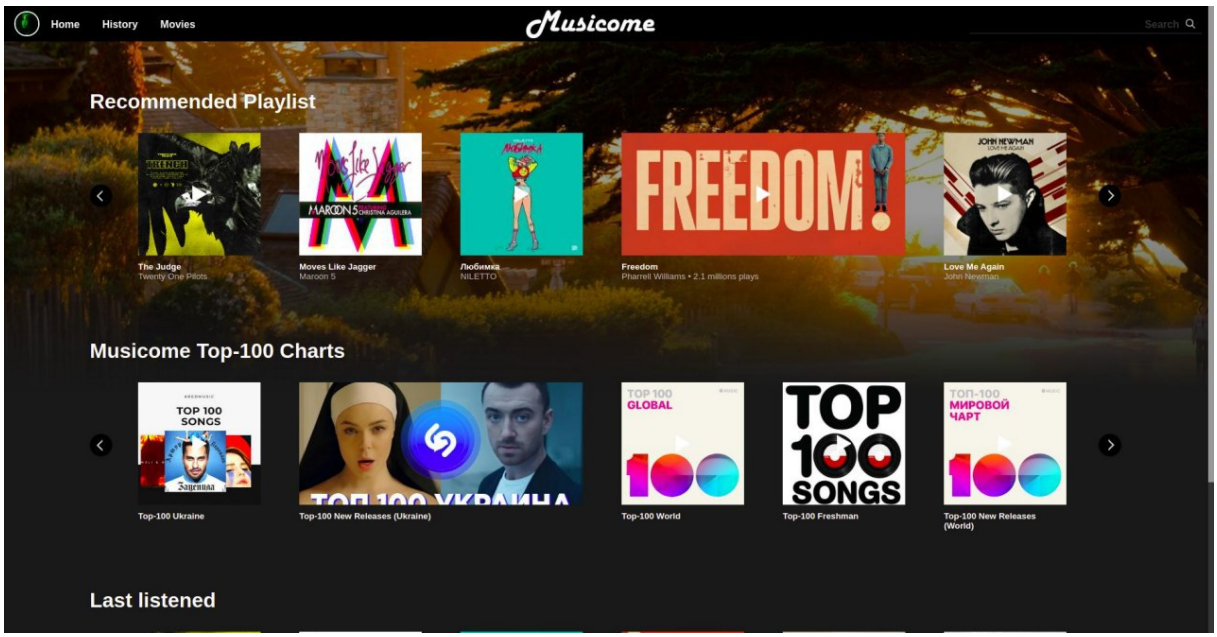

Рисунок 3.6 – Макет головної сторінки

Наступним кроком виконано розробку макету сторінки з добірками фільмографічних творів. (рис. 3.7).

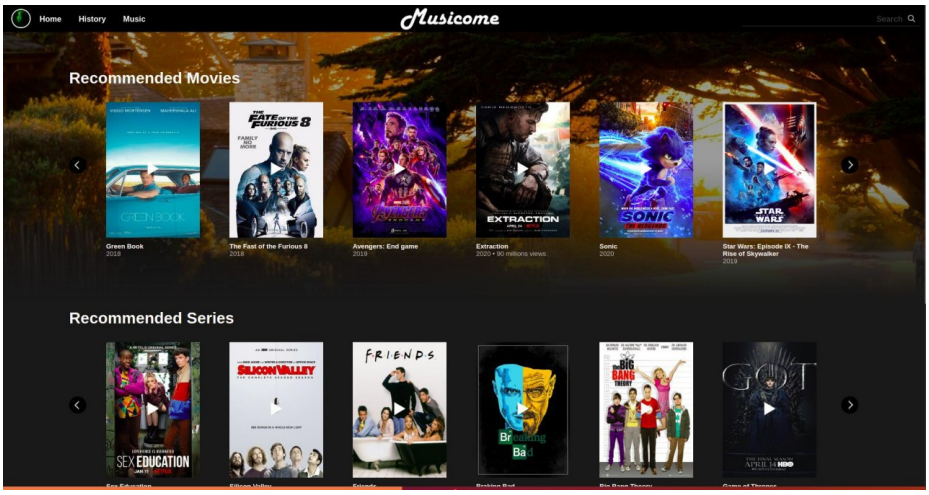

Рисунок 3.7 – Макет сторінки з добірками фільмів і серіалів

Також виконано розробку макетів для сторінок перегляду та прослуховування контенту.

Перший макет (рис. 3.8) відображає сторінку прослуховування музики і містить плейлист з переліком музчиних творів, область рекомендованих фільмів на основі добірки, область з інформацією про пісню, відтворюється та область керування прослуховуванням.

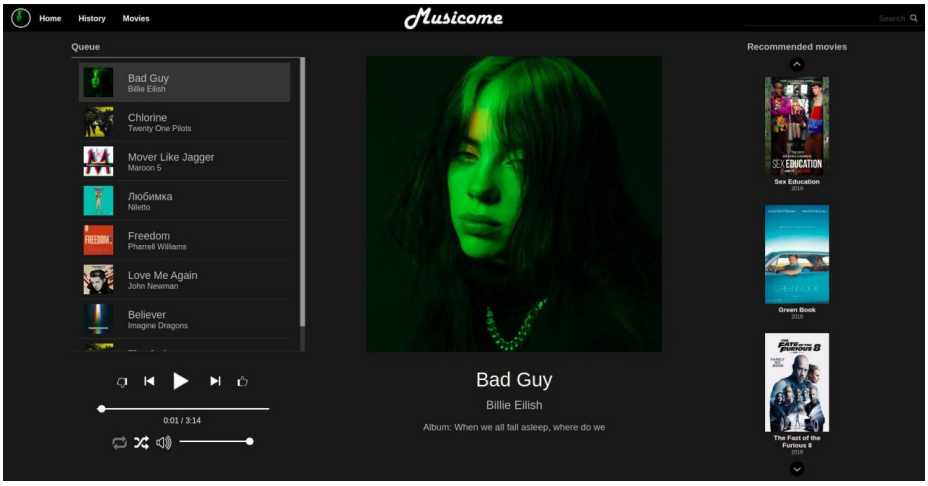

Рисунок 3.6 – Макет сторінки для прослуховування музики

Другий макет (рис. 3.9) відображає сторінку відео контенту і містить плеєр для програвання відео, добірку з рекомендованими фільмами, інформацію про поточний фільм та плейлист фільму.

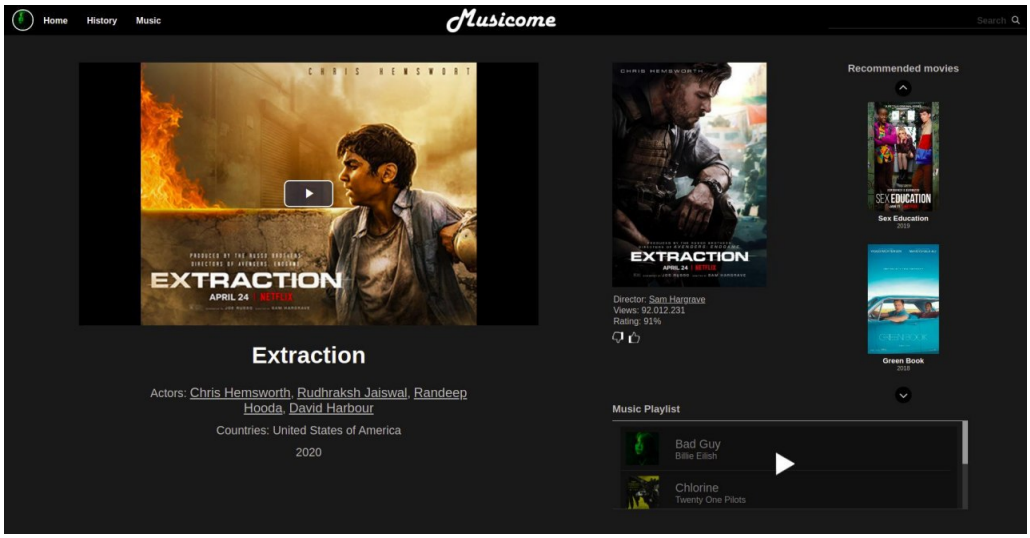

Рисунок 3.9 – Макет сторінки для перегляду відео контенту

Останнім макетом є сторінка пошуку (рис. 3.10), що містить пошуковий рядок та результати з піснями, фільмами, серіалами, альбомами тощо.

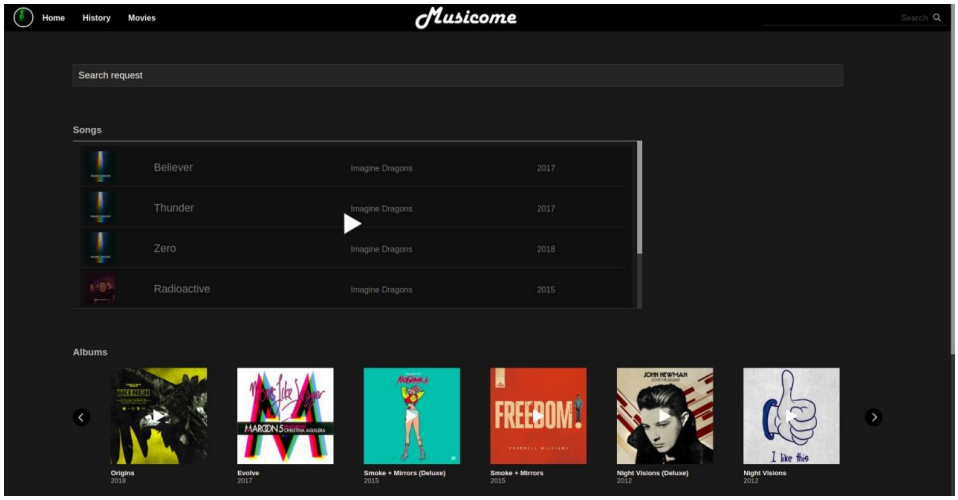

Рисунок 3.10 – Макет сторінки пошуку

3.3.3 Імплементація алгоритму пошуку

Відповідно до розділу 2.4 алгоритм інтелектуального пошуку складається з двох частин.

Перша частина алгоритму відповідає за обчислення коефіцієнтів схожості між екземплярами крос-медійного контенту.

Реалізацію даної частини у веб-застосунку виконано у моделі, що відповідає за опрацювання запитів на проведення маніпуляцій з контентом. Метод, що розроблено для визначення схожості викликається після отримання запитів для додавання чи редагування одиниці контенту.

Реалізацію для консольного застосунку виконано у стилістиці функціонального програмування, а запуск ПЗ виконується за допомогою виконання консольної команди.

Першим кроком реалізації методу є формування кортежів з даних про поточну одиницю контенту та отримання даних про кортежі існуючої множини контенту. В ході виконання програмної реалізації виконаємо групування даних про користувачів у масиви, в якості ключів для яких будуть використані кортежі, а в якості значень – вагові коефіцієнти параметрів контенту.

Далі використано вбудовану функцію мови програмування PHP для пошуку однакових значень серед множини ключів масиву і отримаємо таким чином результуючий маси із значеннями, що повторюються.

Отриману інформацію використовуємо для обчислення коефіцієнту схожості одиниць контенту, після чого виконаємо збереження інформацію про нього в таблицю у БД

Наступним кроком виконано реалізацію логіки формування результатів та добірок відповідно до обчислених раніше коефіцієнтів схожості та вподобань користувачів. Для реалізації даної частини алгоритму виконується формування масиву з даними про дії користувача, та отримуємо дані атрибутів контенту для їх подальшого використання в процесі порівняння.

В ході розробки програмного засобу виконано програмну реалізацію формул (2.11)(2.12)(2.13). В ході їх використання виконано обчислення вхідних даних та значення у нейронах, сформовано умову, відповідно до якої проводиться перевірка контенту на його відповідність до вподобань користувача. В результаті було отримано значення критерію відповідності для кожної з одиниць контенту.

Далі, в залежності від умов випадковим чином для рекомендаційних систем, або ж у відповідності до точності співпадіння для пошуку та консольного застосунку виконується вибірка обмеженої кількості одиниць контенту із отриманого переліку.

Для веб-застосунку виконано реалізацію виглядів для виведення результатів.

В випадку виведення аудіо контенту додатково відбувається формування добірок контенту, до яких виконується додатковий вибір контенту відповідно до значення коефіцієнту схожості.

В результатів виконання даних дій результуюча множина повертається в якості рекомендацій чи пошукових результатів.

Консольний засіб отримує перелік контенту у вигляді масиву, який порівнюється з результуючою вибіркою.

3.4 Проведення експериментальних досліджень

3.4.1 Характеристика вихідних даних

Для проведення тестування методу через консольний застосунок необхідно отримати навчальну вибірку, що міститиме в собі вичерпну інформацію про кожну одиницю контенту та заздалегіть сформовану результуючу вибірку, що має бути застосована для перевірки точності пошуку.

В ході підготовки даного етапу сформовано дві вхідні вибірки:

а) для перевірки точності роботи алгоритму вибірка містить невелику кількість записів, а саме дані про 96 пісень, які мають різний ступінь схожості. Дана вибірка міститься у додатку Б;

б) для формування відомостей про час роботи алгоритму сформовано велику вибірку даних із 60192 піснями, що містить 1444608 рядків з даними про дані пісні. Пісні у вибірці можуть повторюватися, оскільки поточна програмна реалізація не відкидає дублікати і потребує доопрацювання в разі використання у реальному проекті.

Для проведення тестування з використанням веб застосунку тестування необхідно отримати наступні вихідні дані:

а) ім’я користувача у вигляді текстового рядка;

б) прізвище користувача у вигляді текстового рядка;

в) пароль від акаунту у вигляді текстового рядка;

ґ) дані про вподобання щодо музичних та фільмографічних творів, що формуються виходячи з обраних одиниць контенту на етапі реєстрації;

ґ) навчальна вибірка із заповненими метаданими про контент.

В разі наявності даного набору даних можна виконати релевантних результатів пошуку та перших якісних рекомендацій для користувача.

Дослідження проведено із залученням певної кількості сторонніх користувачів для перевірки якості роботи сервісу. Користувачам надавалася можливість користування сервісом та форма у Google Forms для оцінки застосунку для його подальшого порівняння з іншими сервісами.

3.4.2 Аналіз результатів дослідження

По-перше, проаналізуємо результати програмного тестування алгоритмів, що застосовується для інтелектуального пошуку при знаходженні зв’язків між одиницями контенту.

В ході виконання тестування на якість пошуку було залучено вибірку із даними про 96 пісень.

В ході тестування було обрано за основу для пошуку пісню Billie Eilish – Bad Gay. При підготовці до проведення дослідження було сформовано результуючу вибірку із 8 пісень, до якої увійшли Billie Eilish – “burry a friend”, “I love you”, “COPYCAT”, “No time to die”, Imagine Dragons – “Natural”, Twenty One Pilots – “Chlorine”, Maroon 5 – “Girls Like You”, Lady Gaga & Bradley Cooper – “Shallow”.

При проведенні експериментального дослідження було отримано вибірку із 10 пісень, що мають найвищий коефіцієнт схожості із поточною піснею (табл 3.1). Можемо зазначити, що до фінальної вибірки входять усі 8 пісень, що було представлено у початковій умові. Таким чином на невеликій вибірці точність пошуку склала 100%.

Таблиця 3.1 – Результати експериментального дослідження

Музична композиція	Міра схожості
Billie Eilish – burry a friend	0.68528330449827
Billie Eilish – I love you	0.40549307958478
Billie Eilish - COPYCAT	0.19869160899654
Billie Eilish – No time to die	0.19869160899654
Twenty One Pilots – Chlorine	0.10137326989619
Imagine Dragons – Natural	0.10137326989619
Drake – God’s Plan	0.10137326989619
Maroon 5 - Girls Like You	0.10137326989619
Drake – In My Feelings	0.10137326989619
Lady Gaga & Bradley Cooper – “Shallow”	0.10137326989619

Варто зазначити, що при використанні пошуку із застосуванням семантичного анотування точність пошуку також залежіть і від повноти даних про одиниці медіа контенту, таким чином точність пошуку може варіюватися в залежності від обсягу та точності отриманих вхідних даних, проте у загальному випадку її значення буде перевищувати 95%.

Наступним кроком було проведення тестування навантаження на алгоритм, для якого було залучено вибірку із 60192 екземплярами медіа контенту (сумарно 1444608 атрибутів). В ході тестування перевірявся час виконання пошуку та було отримано результати, які продемонстровано нижче (рис. 3.11).

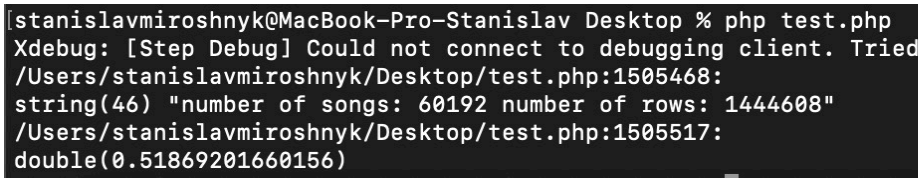

Рисунок 3.11 – Результати програмного тестування

Для проведення порівняння з існуючими методами використаємо відкриті дані про тестування алгоритмів з використанням векторної моделі, що було розроблено на мові програмування Python. Перед проведенням порівняння варто зазначити, що усі алгоритми, з якими відбувається порівняння суттєво оптимізовані і використовують індексацію для пришвидшення роботи.

Нижче представлено перелік методів, час виконання пошуку та час на побудову переліку індексів (табл. 3.2) [26]. Тестування методів проводилося з використанням вибірки «GoogleNews-vectors-negative300.bin», що містить близько 3 мільйонів слів і коротких фраз.

Таблиця 3.2 – Методи із застосуванням векторної моделі

Назва методу	Час виконання пошуку	Час на індексацію
Flat-CPU	9.1c	0c
NMSLIB (HNSW)	0.081c	173c
IVF16384, Flat	0.538c	240c
IVF16384,Flat (Titan X)	0.059c	5c
Flat-GPU (Titan X)	0.753c	0c

Найбільш коректним порівнянням буде порівняння з методом Flat-CPU, оскільки він мінімально застосовує індексацію в порівнянні з іншими методами та є лише частково оптимізованим, а саме його принцип полягає у відкиданні половини заздалегідь найбільш несхожих записів.

Час пошуку неявних семантичних зв’язків у розробленому методі прямо пропорційний до кількості атрибутів у вибірці то обчислимо його для 3 мільйонів записів наступним чином:

$$t = \frac{0.51869201660156 * 3000000}{1444608} = 1.07716145128$$

В результаті розроблений алгоритм має виконуватися для даної вибірки за 1.077с, що є суттєвим покращенням в порівнянні з методом Flat-CPU. При подальшому дослідженні та покращені алгоритму в разі застосування методів кластеризації для відкидання найбільш несхожих груп контенту та використанні індексації результати пошуку можуть бути суттєво покращені та конкурувати з найбільш швидкими алгоритмами, що застосовують векторну модель.

По-друге, проаналізуємо результати тестування із застосування веб-застосунку. Розглянемо основні кроки, які використовувалися користувачами в ході тестування з використанням веб-застосунку.

Першим кроком користувач створює новий акаунт (рис. 3.12).

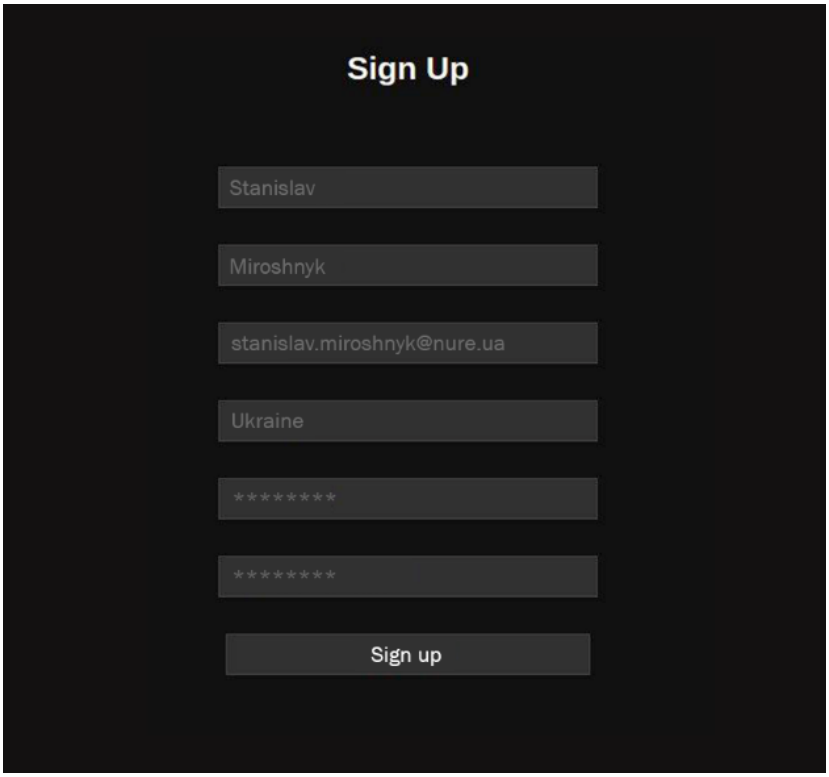

Рисунок 3.12 – Сторінка реєстрації

Далі користувач потрапляє на сторінку для вибору улюблених фільмів, серіалів та аудіо творів (рис. 3.13).

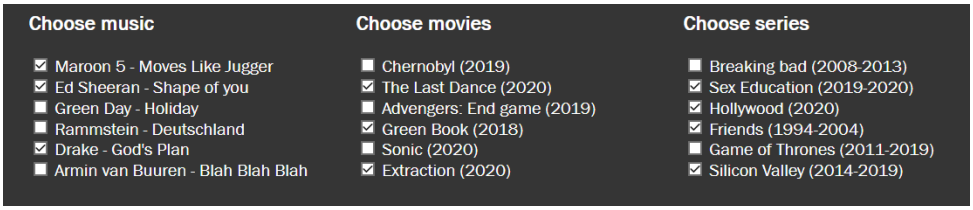

Рисунок 3.13 – Сторінка вибору вподобань

В ході успішної реєстрації користувач потрапляє на головну сторінку сервісу. В процесі переадресації відповідно до вподобань користувача буде сформовано результуючу множину з музичними творами (рис 3.6).

Також, користувач може перейти на сторінку з кінематографічними творами, на якій отримає рекомендації фільмів та серіалів (рис. 3.14).

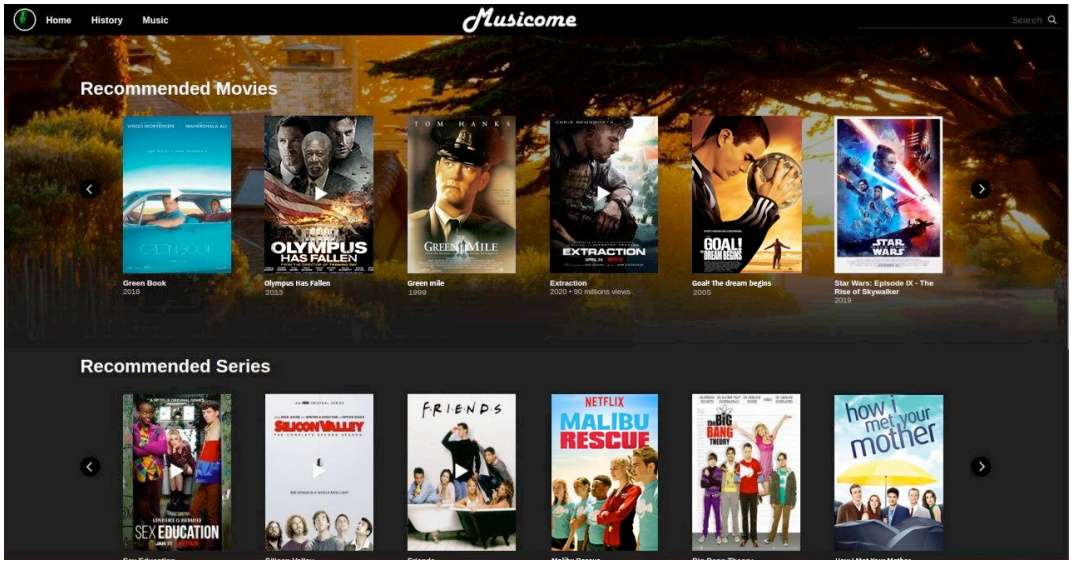

Рисунок 3.14 – Сторінка з добіркою фільмів і серіалів

Таким чином, виходячи з результатів, що були отримані при тестуванні можна побачити, що фільми, серіали та музичні добірки були сформовані за даними про користувача правильно і в цілому відповідають тематиці, яка обиралася на етапі реєстрації.

Далі для додаткової перевірки ефективності роботи алгоритму користувач має потрапити на сторінку однієї з музичних добірок та

переглянути, чи є отриманий набір творів релевантним до пісні, що була обрана на сторінці рекомендацій чи сторінці пошуку.

Для отримання користувацької оцінки з приводу роботи розробленого сервісу було створено форму із застосуванням сервісу Google Forms (рис. 3.15) та отримано наступні оцінки (табл 3.3).

Рисунок 3.15 – Форма для оцінки

Таблиця 3.3 – Оцінки користувачів

Ім’я	Якість рекомендацій	Якість формування добірок	Якість пошуку	Зручність сервісу	Середня оцінка
Антон	5	4	5	4	4.5
Костянтин	4	5	5	5	4.75
Ірина	5	5	5	5	5
Антон	5	5	5	3	4.5
Юлія	5	5	4	5	4.75
Єгор	5	5	5	5	5
Андрій	5	5	5	5	5
Галина	5	4	5	5	4.75
Богдан	5	5	5	3	4.5
Володимир	4	4	5	5	4.5
Вікторія	5	5	5	5	5
Загалом					4.75

Виконаємо порівняння із популярними сервісами з розповсюдження музичних творів. Оцінки сервісів отримано із застосунку App Store та представлено нижче (табл 3.4).

Таблиця 3.4 – Оцінки сервісів розповсюдження музики

Назва сервісу	Оцінка
Youtube Music	4.7
Spotify	4.8
Apple Music	3.5
Deezer	4.6

Таким чином розроблений сервіс має достатньо високу оцінку користувачів та конкурує за цим показником з іншими подібними сервісами.

Виходячи з результатів експериментальних досліджень можна сформувати висновок, що розроблений метод при певній доробці є конкурентоспроможним та може бути застосований в рамках задач пошуку крос-медійного контенту.

ВИСНОВКИ

Під час виконання кваліфікаційної роботи проведено детальний аналіз існуючих сервісів, що займаються розповсюдженням та використовують різноманітні методи для визначення схожості крос-медійного контенту, використовуючи результати роботи подібних методів в рекомендаційних та пошукових системах.

Виконано обґрунтування актуальності кваліфікаційної роботи, враховуючи поточні тенденцій розвитку сервісів.

Сформовано завдання на проведення дослідження та розробки нового методу інтелектуального пошуку крос-медійного контенту, в якому виконано виділення основних вимог, що використовуються в подальшому для оцінки ефективності методу.

В ході виконання проведено ряд теоретичних досліджень. Розглянуто ряд методів інтелектуального аналізу даних, до яких увійшли методи побудови дерев рішень, кластеризації та застосування можливості побудови штучних нейронних мереж в контенті можливості їх подальшого застосування в рамках розроблюваного методу інтелектуального пошуку. В ході аналізу даних методів приведено їх детальний опис, та сформовано переліки позитивних та негативних аспектів їх використання. Базуючись на даних аспектах виконано вибір одного з методів ІАД, що є найбільш ефективним в контексті його подальшого використання для роботи з великими обсягами даних.

Наступним кроком проведено дослідження можливості використання семантичного анотування в контексті розв’язання задачі організації наборів метаданих. Досліджено декілька видів анотацій, що можуть бути застосовані в ході даного процесу, розглянуто переваги та недоліки в контексті їх використання для анотування крос-медійного контенту та обрано онтологічну модель анотування для формалізації даних про екземпляри контенту з метою їх подальшої обробки.

Розглянуто існуючі методи для визначення семантичної відстані між одиницями контенту на основі отриманої інформації. Обрано метод пошуку відстані Хеммінга, на основі якої відбувається виокремлення неявних семантичних зв’язків.

В результаті дослідження методів інтелектуального аналізу даних та семантичного анотування виконано розробку власного інтелектуального методу пошуку крос-медійного контенту та описано можливості його адаптації для роботи при його використанні у рекомендаційних або пошукових системах.

Для проведення експериментальних досліджень розроблено два програмних засоби – консольний та веб-застосунки. В рамках даних засобів було реалізовано розроблений в ході теоретичних досліджень алгоритм.

Консольний додаток використовує розроблений алгоритм для проведення дослідження на точність та швидкість його роботи. Для проведення даних тестів було сформовано дві навчальні вибірки, що використовуються в якості вхідних даних та одну результуючу вибірку, що застосовується при тестуванні системи на точність отриманих результатів схожості.

В рамках веб-застосунку, який реалізовано в якості сервісу для розповсюдження музики, алгоритм використано для формування рекомендацій, виконання пошуку та генерацій музичних добірок.

В рамках тестування визначено, що розроблений метод добре показав себе в порівнянні з існуючими методами, а отримані результати повністю відповідають початковим вимогам, що було встановлено до методу.

Результати виконаного дослідження пройшли апробацію на 11-й Міжнародній конференції із застосування інформаційно-комунікаційних технологій та статистики в економіці та освіті «ICAICTSEE-2021» (м. Софія, Болгарія) [27].

Завдання на дипломну роботу було виконано у повному обсязі. Було реалізовано усі поставлені завдання.

ПЕРЕЛІК ДЖЕРЕЛ

1. Segura A. The Top 12 Types of Social Media Content to Create. Mailchimp. 2018. URL: https://mailchimp.com/resources/top-12-types-of-social-media-content-to-create/. (дата звернення 20.10.2021).

2. The Modern History of Computing [Електронний ресурс] // Stanford Encyclopedia of Philosophy. – 2017. – URL: https://plato.stanford.edu/entries/computing-history/. (дата звернення 20.10.2021).

3. Mimsarchive C. How iTunes Genius Really Works. MIT Technology Review. 2010. URL: https://www.technologyreview.com/2010/06/02/91325/ how-itunes-genius-really-works/.(дата звернення 20.10.2021).

4. Slavik N. Sources Say Apple Will Kill iTunes Downloads in 2-4 Years, Apple Denies. Djbooth. 2016. – URL: https://djbooth.net/features/2016-05-12-apple-kill-itunes-downloads. (дата звернення 20.10.2021).

5. Steinberg E. Semantic vector search: the new frontier in product discovery [Електронний ресурс] / Eugene Steinberg // Grid Dynamics. – 2020. – Режим доступу до ресурсу: https://blog.griddynamics.com/semantic-vector-search-the-new-frontier-in-product-discovery/. (дата звернення 20.10.2021).

6. Madrigal A. C. How YouTube's Algorithm Really Works. The Atlantic. 2018. URL: https://www.theatlantic.com/technology/archive/2018/11/how-youtubes-algorithm-really-works/575212/. (дата звернення 20.10.2021).

7. How "Fans Also Like" Works. Spotify for artists. 2019. – URL: https://artists.spotify.com/blog/how-fans-also-like-works. (дата звернення 20.10.2021).

8. Postmus S. Recommender system techniques applied to Netflix movie data. Vrije Universiteit Amsterdam. Amsterdam, 2018. №1. С. 31.

9. Ситник В. Ф., Краснюк М.Т. Інтелектуальний аналіз даних (дейтамайнінг). Київ: КНЕУ, 2007. 376 с.

10. Olson D. L. Data mining in business services / D. L. Olson. // Service Business. – 2007. – №3. – С. 181–193.

11. Breiman L., Friedman J.H., Olshen R.A,m Stone C.J. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. 368 с.

12. Quinlan J. R. Induction of Decision Trees, Machine Learning. 1986. №1. С. 81–106.

13. Tryon R. Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality / Robert C. Tryon. – Ann Arbor: Edwards Brother, 1939. – 122 с.

14. Strategy for One-to-One Pickup and Delivery Problem Using the Cyclic Transfer Approach / [R. Dupas, I. Grebennik, I. Litvinchev та ін.]. // EAI Endorsed Transactions on Energy Web, Special issue on Energy Conservation, Information Technologies and Large Scale Optimization.. – 2020. – №27.

15. Bhadeshia H. K. D. H. Neural Networks in Materials Science. ISIJ International. 1999. №39. С. 966–979.

16. Mackay D. J. C. Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press, 2003. 640 с.

17. What are semantic annotations? / E. Oren, et. al. National University of Ireland, Galway: Digital Enterprise Research Institute. 2006. №1. С. 15.

18. Ontologies and Semantic Annotation. Part 1: What Is an Ontology [Електронний ресурс] // Medium. – 2018. – URL: https://medium.com/sciforce/ontologies-and-semantic-annotation-part-1-what-is-an-ontology-1de10caf2c77.

19. Головянко М. В., Плиско Д.А. Построение распределенной системы онтологий на базе технологии Peer-To-Peer (P2P). Збірник наукових праць Харківського університету Повітряних сил. 2010. №3. С. 131–133.

20. Захарова О. В. Основні аспекти семантичного анотування великих даних / О. В. Захарова. // Проблеми програмування. – 2020. – №4. – С. 22–23.

21. PHP Advantages and Disadvantages | What is PHP Language? Merits and Demerits of PHP [Електронний ресурс] // AplusTopper. – 2021. – URL: https://www.aplustopper.com/php-advantages-and-disadvantages/. (дата звернення 4.10.2021).

22. MySQL Advantages And Disadvantages [Електронний ресурс] // W3spoint. – 2015. – URL: https://www.w3spoint.com/mysql-advantages-disadvantages. (дата звернення 4.11.2021).

23. PhpStorm Review: Pricing, Pros, Cons & Features [Електронний ресурс] // CompareCamp. – 2021. – URL: https://comparecamp.com/phpstorm-review-pricing-pros-cons-features/. (дата звернення 4.11.2021).

24. What is MVC? Advantages and Disadvantages of MVC [Електронний ресурс] // InterServer. – 2016. – URL: https://www.interserver.net/tips/kb/mvc-advantages-disadvantages-mvc/. (дата звернення 4.11.2021).

25. User Interface Design Basics [Електронний ресурс] // Usability. – 2014. – URL: https://www.usability.gov/what-and-why/user-interface-design.html. (дата звернення 4.11.2021).

26. Методи наближеного пошуку найближчих сусідів [Електронний ресурс] // Habr. – 2017. – URL: https://habr.com/en/company/vk/blog/338360/. (дата звернення 4.11.2021).

27. Miroshnyk S. Intelligent Search Of Cross-Media Content / S. Miroshnyk, I. Grebennik. – Sofia, Bulgaria: University of National and World Economy, 2021. – (ICAICTSEE. Conference proceedings.).