Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра _____Інформаційних управляючих систем_____
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____другий (магістерський)_____

_____Дослідження методів побудови рекомендацій_____
_____для нових користувачів онлайн-кінотеатру_____
(тема)

Виконав:
студент _2_ курсу, групи _ІУСТм-21-1____
_____Ігор МАКСИМЕНКО _____
(власне ім'я, прізвище)

Спеціальність ____122 Комп'ютерні_____
_____науки_____
(код і повна назва спеціальності)
Тип програми ____освітньо-професійна_____
(освітньо-професійна або освітньо-наукова)
Освітня програма _Інформаційні управляючі системи та технології_____
(повна назва освітньої програми)

Керівник __проф. каф. ІУС Сергій ЧАЛИЙ __
(посада, власне ім'я, прізвище)

Допускається до захисту

Зав. кафедри _____ ___Костянтин ПЕТРОВ__
(підпис) (власне ім'я, прізвище)

2022 р.

Харківський національний університет радіоелектроніки

Факультет_____Комп'ютерних наук_____

Кафедра_____Інформаційних управляючих систем_____

Рівень вищої освіти_____другий (магістерський)_____

Спеціальність_____122 Комп'ютерні науки_____
(код і повна назва)

Тип програми _____освітньо-професійна_____
(освітньо-професійна або освітньо-наукова)

Освітня програма___Інформаційні управляючі системи та технології_____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри_____
(підпис)

«_21_»__листопада__ 20 22_ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові_____Максименко Ігорю Сергійовичу_____
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів побудови рекомендацій для нових користувачів онлайн кінотеатру_____
затверджена наказом університету від _14_ __листопада__2022_ р. № _1490Ст_

2. Термін подання студентом роботи до екзаменаційної комісії ____13__грудня__ 2022р.

3. Вихідні дані до роботи науково-технічна література, публікації, інформація з інтернет-ресурсів._____

4. Перелік питань, що потрібно опрацювати в роботі аналіз рекомендаційних систем, дослідження методів побудови рекомендацій, дослідження підходів до вирішення проблеми нових користувачів рекомендаційної системи, постановка задачі дослідження, дослідження комбінованих підходів для вирішення проблеми холодного старту в рекомендаційній системі, удосконалення гібридного методу колаборативної для нових користувачів з використанням даних анкетування, створення технології використання удосконаленого методу, програмна реалізація удосконаленого методу, проведення експериментальної перевірки удосконаленого методу.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	10.10.2022	Виконано
2	Аналіз методів побудови рекомендацій	15.10.2022 − 25.10.2022	Виконано
3	Дослідження методів вирішення проблеми холодного старту	28.10.2022 − 13.11.2022	Виконано
4	Створення інформаційної технології побудови рекомендацій з використанням даних анкетування	14.11.2022 − 18.11.2022	Виконано
5	практична реалізація отриманих результатів	19.11.2022 − 29.11.2022	Виконано
6	Підготовка пояснювальної записки та графічного матеріалу	22.11.2022 − 12.12.2022	Виконано

Дата видачі завдання 21 листопада 2022 р.

Студент _____
(підпис)

Керівник роботи _____ проф. каф. ІУС Сергій ЧАЛИЙ
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка містить: 86 сторінок, 24 рисунків, 29 джерел. 7 таблиць

ДЕМОГРАФІЧНА ФІЛЬТРАЦІЯ, МЕТОД КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ, ПОКАЗНИКИ ПОДІБНОСТІ, ПРОГНОЗ РЕКОМЕНДАЦІЙ, РЕКОМЕНДАЦІЙНА СИСТЕМА, ХОЛОДНИЙ СТАРТ.

Об'єктом дослідження є процеси побудови рекомендацій для нових користувачів.

Предметом дослідження – методи формування рекомендацій при холодному старті нових користувачів.

Мета роботи – дослідження методів побудови рекомендацій для нових користувачів онлайн кінотеатру.

В результаті виконання роботи проаналізовано рекомендаційні системи в цілому, проведено аналіз методів побудови рекомендаційних систем, досліджено проблему нових користувачів, розглянуто підходи для вирішення проблеми нових користувачів, запропоновано удосконалений метод побудови рекомендацій для нових користувачів.

В роботі приведено технологію використання удосконаленого методу, проведено програмну реалізацію та експериментальну перевірку удосконаленого методу.

ABSTRACT

The thesis note contains: 86 pages, 24 figures, 29 references, 7 tables.

COLD START, COLLABORATIVE FILTERING METHOD, DEMOGRAPHIC FILTERING, PERSONALITY TRAITS, RECOMMENDATION FORECAST, RECOMMENDATION SYSTEM, SIMILARITY INDICATORS

The object of research is the processes of building recommendations for new users.

The subject of the research is methods of forming recommendations during the cold start of new users.

The aim of the work is to study methods for building recommendations for new users of an online cinema.

As a result of the work, recommendation systems as a whole are analyzed, methods for building recommendation systems are analyzed, the problem of new users is investigated, approaches to solving the problem of new users are considered, and an improved method for building recommendations for new users is proposed.

The paper presents the technology of using the improved method, conducts software implementation and experimental verification of the improved method.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

РС – рекомендаційна система;

КФ – колаборативна фільтрація;

MAB – multi-armed bandit;

UCB – upper confidence bound;

SVD – singular value decomposition;

VBPR – visual Bayesian personalized ranking;

BPR – Bayesian personalized ranking;

TIPI – ten item personality inventory;

MAE – mean absolute error;

ВСТУП

У світі, ринок онлайн-кінотеатрів є сектором економіки, який виразно показує тенденцію до постійного позитивного зростання. Цей процес на світовому ринку проходить завдяки залученню сучасних цифрових технологій, розробки програмного забезпечення, збільшенням рівня платоспроможності населення, а також зменшенням популярності багатозальних кінотеатрів.

Сьогодні, як і раніше, можна не сидіти вдома, а вирушити в кінотеатр, щоб подивитися фільм або мультфільм на великому екрані. Часом такий варіант перегляду фільмів може бути кращим, але не кожного влаштовує подібний спосіб.

Сеанси у звичайних кінотеатрах проводяться у визначений час, через це потрібно заздалегідь виділити час на похід до кінотеатру. Щоб подивитися фільм у кінотеатрі, необхідно купити квитки і нерідко для цього доводиться стояти в довгій черзі. Прибути до кінотеатру слід до певного часу, інакше можливо не встигнути.

У сучасних умовах не виникає таких проблем, оскільки розповсюдженість швидкісного інтернету, надала можливість ринку кіноіндустрії надати глядачам можливість зручнішого перегляду фільмів не виходячи з дому.

Онлайн-кінотеатри працюють за простим принципом, користувач оформлює підписку, яка відкриває доступ до великих бібліотек з ліцензійними фільмами. На сьогоднішній час такими можливостями не здивувати нікого, хоч зовсім недавно про це можна було лише мріяти.

Оскільки фільмів в онлайн-кінотеатрах з годом тільки збільшується, необхідним стає використання рекомендаційних систем.

Зростання важливості інтернету як середовища для електронних і ділових операцій стало рушійною силою для розвитку технології рекомендаційних систем. Поштовхом для розвитку цього стала легкість з якою інтернет надає

можливість користувачам залишати відгук про те що їм сподобалось чи не сподобалось [2].

Рекомендаційні системи онлайн-кінотеатрів використовують систему рейтингу, який формується за рахунок зворотного зв'язку. Користувачі обирають числові значення(наприклад система оцінки якості від 1 до 5), які вказують на їхні вподобання до різних фільмів. На виході користувачу надають список фільмів які йому мають сподобатися, тим самим допомагають вирішити проблему перевантаження вибором, пропонуючи користувачеві самі ті елементи, що становлять для них інтерес.

При розробці онлайн-кінотеатрів в основному використовують два методи при побудові систем рекомендації – це фільтрація вмісту та колаборативна фільтрація.

Фільтрація вмісту зіставляє користувачів з контентом або товарами ґрунтуючись на основі минулих оцінок вподобаних фільмів.

Колаборативна фільтрація система зіставляє фільмі, які зазвичай однаково оцінені, або користувачів зі схожою історією вподобань, для формування нових зв'язків між фільмом та користувачем, на основі минулої поведінки користувача.

Так як притік нових користувачів є постійним в онлайн-кінотеатрах, необхідно, щоб вони з самого початку отримували актуальні і точні рекомендації. Однак через відсутність або неповноту представленої інформації про нових користувачів, інформаційна система не може побудувати точні рекомендації фільмів. Така проблема отримала назву – проблеми холодного старту для нових користувачів.

Для подолання проблеми холодного старту необхідно проаналізувати та позбутися недоліків вже існуючих систем рекомендації, тим самим створити удосконалений метод який зможе вирішити проблему холодного старту та покращити користувацький досвід для нових користувачів.

1 АНАЛІЗ МЕТОДІВ ПОБУДОВИ РЕКОМЕНДАЦІЇ

1.1 Аналіз рекомендаційних систем

З розвитком комерційних сайтів, поширилась проблема прийняття рішення при виборі товару. Ця проблема стає ще більшою, через зростання обсягу різноманітності пропозицій та відсутність особистого досвіду при оцінки потенційних видів товару що може запропонувати наприклад веб-сайт. Хоча й масове збільшення кількості варіантів вибору, надало можливість обрати найбільш цікаву пропозицію, це призвело до проблеми перевантаження вибору, тобто коли користувач має необмежену кількість варіантів вибору. Вирішенням цієї проблеми стало створення та залучення рекомендаційних систем.

Рекомендаційні системи (РС) – це програмні інструменти та методи, які пропонують можливі списки пропозицій, стосовно саме тих товарів, що скоріш за все зацікавлять конкретного користувача. Списки пропозицій стосуються різних процесів прийняття рішень, стосовно того наприклад який фільм обрати для перегляду, яку музику обрати для прослуховування, чи який телефон придбати [2].

«Предмет» — це загальний термін, який використовується для позначення того, що система рекомендує користувачам. РС зазвичай зосереджується на певному типі елемента (наприклад, компакт-диски чи новини) і зокрема його дизайн, графічний інтерфейс користувача та основну техніку рекомендацій які використовуються для створення рекомендацій, усі налаштовані, щоб забезпечити корисні та ефективні пропозиції для цього конкретного типу предмета [2].

РС в першу чергу спрямовані на осіб, які не мають достатнього особистого досвіду або компетентності, щоб оцінити потенційно переважну кількість альтернативних елементів, які може запропонувати, наприклад, веб-сайт. Яскравим прикладом є рекомендаційна система YouTube, яка допомагає користувачам вибрати відео для перегляду.

Оскільки рекомендації зазвичай персоналізовані, різні користувачі або групи користувачів отримують різноманітні пропозиції. Крім того, бувають і неперсоніфіковані рекомендації. Їх набагато простіше генерувати, і вони зазвичай містяться в них журнали чи газети. Типові приклади включають в себе добірки з десяти найкращих фільмів, композицій цього місяця тощо. Хоча вони можуть бути корисними та ефективними в певних ситуаціях, ці типи неперсоналізовані рекомендації зазвичай не розглядаються дослідженнями РС [2].

Зазвичай, персоналізовані рекомендації надаються користувачеві у вигляді упорядкованих списків предметів. При побудові рекомендацій, РС намагаються передбачити, які продукти чи послуги є найбільш підходящими, виходячи з уподобань і обмежень користувача. Щоб завершите дане обчислювання, РС збирає від користувачів їхні переваги, які виражаються явно, наприклад, як оцінки продуктів, або випливають з інтерпретації дій користувача. Наприклад, РС може розглядати перехід на сторінку конкретного продукту, як неявний знак переваги товару на цій сторінці.

Розробка РС почалася з дослідження, що люди часто покладаються на рекомендації, надані іншими при прийнятті рутинних щоденних рішень. Наприклад, прийнято покладатися на те, що рекомендують однолітки при виборі книги для читання; роботодавці розраховують на рекомендаційні листи у своїх повторних кадрові рішення; а коли вибирають фільм для перегляду, люди схильні читати і покладатися на рецензії на фільми, написані іншими користувачами. Прагнучі імітувати таку поведінку, перші РС застосували алгоритми для використання рекомендації, створені спільнотою користувачів для надання рекомендацій для активного користувача, тобто користувача, який шукає пропозиції. Рекомендації були для товарів, які сподобалися схожим користувачам (зі схожими смаками) [2].

Цей підхід називається колаборативна фільтрація і його основна ідея полягає в тому, що, якщо активний користувач погоджується в минулого з

деякими користувачами, то інші рекомендації, що надходять від цих схожих користувачів також має бути актуальним і цікавим для активного користувача.

РС використовуються для надання користувачам елементів, які мають відповідати їх потребам в даний момент часу.

Процес побудови рекомендацій можна поділити на три етапи. Першим етапом виступає збір інформації, на якому система враховує дані профілю користувача такі як: оцінки предметів, демографічні дані та інші. На другому етапі виконується навчання. На цьому етапі система навчається на вибірці профілів даних, порівнюють однотипні дані різних людей, для формування груп користувачів. На третьому етапі виконується прогнозування/побудова рекомендацій. На цьому етапі рекомендаційна система намагається спрогнозувати які саме предмети будуть цікаві користувачам. Зворотній зв'язок - це процес, що призводить до того, що результат функціонування будь-якої системи впливає на параметри, від яких функціонування цієї системи залежить. Інакше кажучи, на вхід системи подається сигнал, пропорційний її вихідного сигналу [2]. Процес побудови рекомендації наведено на рисунку 1.1.

Рисунок 1.1 – Процес побудови рекомендацій

Якість роботи РС залежить від багатої кількості властивостей вироблюваних нею рекомендацій, що найчастіше оцінюють за допомогою показників продуктивності, які показують вузьке уявлення про поведінку роботи системи. Для проведення оцінки точності рекомендацій, найчастіше за все використовують метрики, які вимірюють можливість РС точно представляти набір відомих переваг. Точність, у свою чергу, вимірюють за допомогою резервування частини існуючих даних рейтингів представлених у вигляді набору відомих уподобань і використання інших рейтингів для навчання РС та побудови рекомендацій. Для обчислення оцінки точності алгоритму рекомендації для набору даних, можна використати знайдені переваги які можуть використовуватися деякими метриками, такими як відгук чи точність [2].

При побудові РС використовують два основні підходи, в залежності від того для чого створюють РС змінюється й принцип за яким вони працюють:

– рекомендації на основі великих даних;

– рекомендації на основі принципу прогресивної персоналізації.

При побудові персоналізованих рекомендацій на основі великих даних. використовують історію поведінки цього користувача, а ще й історію поведінки усіх інших користувачів. Основною проблемою цього підходу є поява нових користувачів в системі про яких немає жодних даних в системі [3].

При побудові рекомендацій на основі принципу прогресивної персоналізації РС використовує дані профілю користувача.

Структуру рекомендаційної системи наведено на рисунку 1.2.

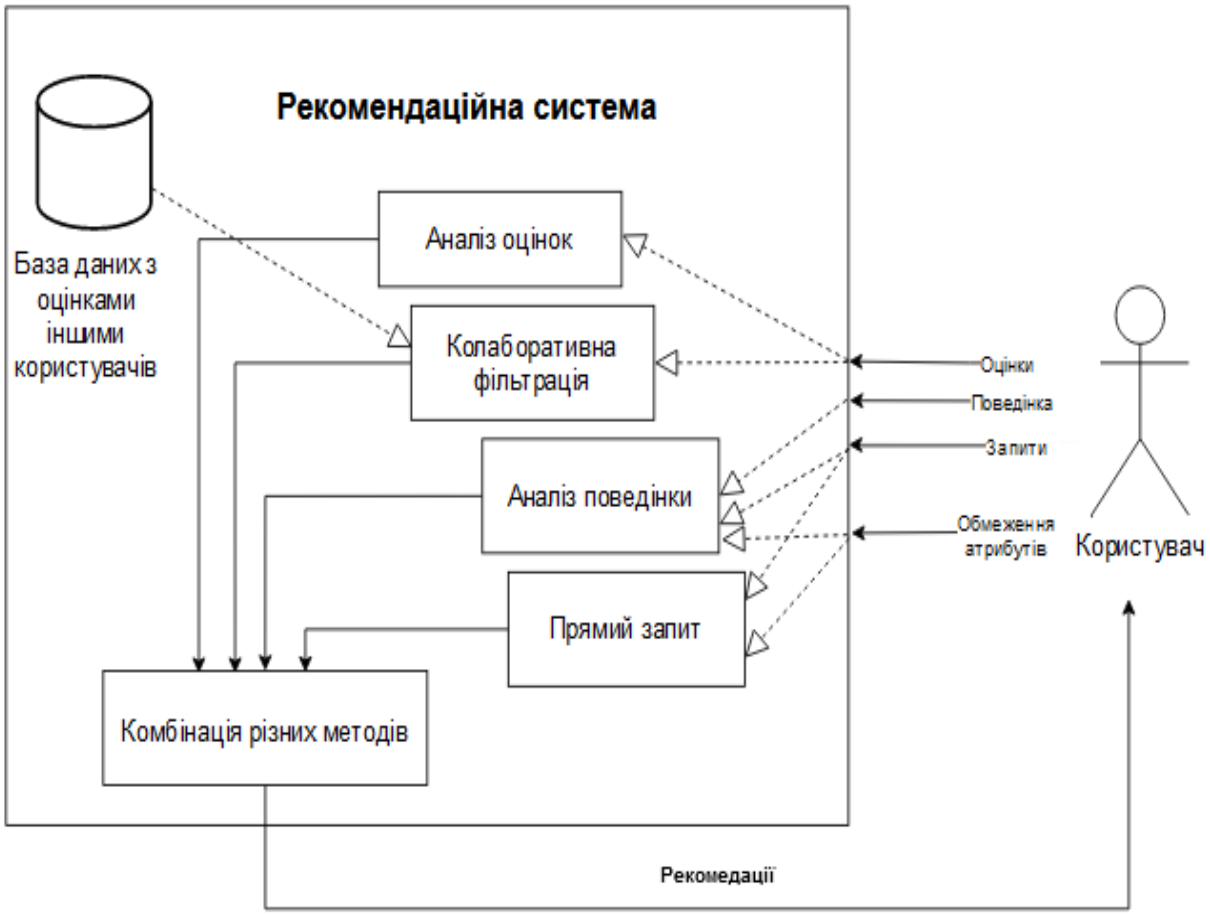

Рисунок 1.2 – Структура рекомендаційної системи

РС спочатку були розроблені та інтегровані для інтернет-магазинів електронної комерції. Оскільки користувачам було важко знайти саме необхідний товар серед купи варіантів, рекомендаційні системи в свою чергу показали свою ефективність з допомогою користувачам з вибором конкретного товару. Таким чином при взаємодії користувачів з системою, вони можуть отримувати випадкові пропозиції з недослідженої частини каталогу товарів, які б їх могли зацікавити. Їх застосування виявилось настільки успішним, що їх почали впроваджувати майже всюди, від соціальних мереж до інтелектуальних помічників [4].

Основна мета використання РС полягає в збільшенні продажів товарів. Привертаючи увагу користувачів до конкретних елементів, РС надають самі ті товари, які скоріш за все клієнт придбає. Це збільшить обсяг продажів, що в свою

чергу буде збільшувати прибуток власника. Крім того можна виділити загальні цілі РС:

– надання користувачу рекомендацій що є актуальними для нього;

– рекомендація нових товарів, які б могли зацікавити користувача;

– випадкові рекомендації, які могли б приємно здивувати чи відкрити щось нове для користувача;

– різноманітність рекомендованого товару, хоч користувач і отримує схожі рекомендації, через різноманітність користувачеві не набридає повторення схожих товарів.

Основний принцип роботи алгоритмів РС полягає в тому, що між діяльністю орієнтованою на користувача та товар-орієнтованою присутні значні залежності. Прикладом цього буде користувач, який зацікавлений в прослуховуванні класичної музики, скоріше зацікавиться в прослуховуванні іншого класичного композитора, а не в тяжкому року. Також різні категорії елементів можуть показувати значні кореляції, які необхідні для формування точнішого списку рекомендацій [4].

Загалом системи рекомендацій можуть служити двом різним цілям. На одному з боку, їх можна використовувати, щоб спонукати користувачів робити щось, наприклад купувати конкретну книгу або перегляд конкретного фільму. З іншого боку, рекомендаційні системи також можна розглядати як інструменти для боротьби з інформаційним перевантаженням, оскільки ці системи спрямовані на вибір найцікавіших предметів із більшого набору. Таким чином, дослідження систем рекомендацій також сильно вкорінені в області інформації пошук і фільтрація інформації.

Отже виходячи з цього, можна зазначити що використання і інтерес до рекомендаційних систем постійно зростає. РС вже давно є невід'ємною частиною багатьох веб-сайтів таких як Netlix, IMDb, Megogo, Amazon, Facebook, Youtube, та в багатьох інших. Розберемо на прикладі Netflix. У наш час мабуть нема людини, що б хоч раз не чула про Netflix. Netflix забезпечує користувачам по

всьому світі потокову доставку фільмів і телепередач за підпискою. Netflix активний за своєю природою, як тільки ви стаєте користувачем, він починає запитувати вас про ваші уподобання, жанри, заповнюючи інформацію. Це основний крок для надання рекомендацій. Рекомендації Netflix намагаються реалізувати рекомендацію, зводячи до мінімуму зусилля користувача, включені до пошуку [2].

Netflix – американська компанія, що надає платні послуги стримінгу фільмів та серіалів. З переходом до онлайн-послуг значення компанії щодо трендів методів рекомендаційних систем почало стрімко зростати. Власники Neflix визначили на основі аналізу поведінки користувача, що у рекомендаційної системи є близько 90 с для формування альтернатив на допомогу користувачу в пошуку цікавої стрічки, перш ніж він покине платформу та звернеться до іншого сервісу. Саме тому основна цінна пропозиція Netflix – надавати релевантні рекомендації своїм клієнтам Рекомендаційні системи охоплюють різні алгоритмічні підходи: навчання із підкріпленням, нейронні мережі, причинну модель, імовірнісні графічні моделі, факторизації матриць вподобань тощо [4].

За результатами досліджень проведених компанією McKinsey, більше 75% контенту, який дивились користувачі Netflix, було запропонований їхньою рекомендаційною системою. Система допомагає споживачам знайти цікавий контент, який вони хочуть подивитись, а також допомагає компанії економити на витратах на маркетинг [4].

Netflix використовує гібридний підхід у своїй РС. У своїй РС (відомій як Bellkor), Netflix використовує гібридний підхід на основі 27 алгоритмів. Серед усіх видів РС для онлайн кінотеатрів, система Bellkor показує найвищу точність. Однак подібну систему не кожен онлайн-кінотеатр може собі дозволити через складність реалізації та великі затрати [4].

Розвиток рекомендаційних систем полягає в тому, щоб постійно поліпшувати алгоритми побудови рекомендації. Основна ідея цього – надавати користувачам системи найбільш точні рекомендації, що відповідають їх

потребам в даний момент часу. Для досягнення цього, використовуються математичні алгоритми, що представлені в основі РС, повинні постійно покращуватися.

Для цього використовують інтелектуальний аналіз даних data mining. На прикладі це виглядатиме так: онлайн кінотеатр надає користувачеві набір рекомендацій, далі адміністратор системи отримує від нього оцінки чи вподобання що до представленого фільму, аналізує їх на предмет відповідності даних, раніше рекомендацій інтересам відвідувача, перенавчати математичну модель, потім знову пропонує рекомендації і так далі по колу [2].

Рекомендаційні системи незважаючи на усі свої переваги, залишаються з однією з головних проблем – проблемою холодного старту. Проблема холодного старту виникає тоді, коли в системі з’являється новий користувач, а в системі відсутня достатня кількість даних для формування рекомендації для нових користувачів. Проблема холодного старту є частою при побудові рекомендаційних систем для онлайн кінотеатрів, оскільки поява нових фільмів і користувачів є постійною.

В відомих онлайн-кінотеатрах нові користувачі з'являються щодня, вони реєструються на сервісі і стають холодними користувачами. Вони мало чи майже зовсім не взаємодіятимуть із каталогом в день їх реєстрації. Пряме включення нових користувачів до працюючої системи рекомендації неможливо за відсутністю чи неповнотою інформації. Чекати коли вони стануть теплими користувачами (за внутрішніми критеріями) також небажано, оскільки рекомендація відповідного для користувачів вмісту одразу має вирішальне значення, бо нові користувачі, які зіткнуться з не якісними рекомендаціями, можуть припинити подальше користування сервісом, як наслідок йде втрата клієнта..

За результатами огляду рекомендаційних систем та аналізу ринку, було виявлено, що рекомендаційні системи продовжують набирати популярність для використання в різних інформаційних системах, однак не можливо не зазначити

одну з головних проблем РС – холодний старт. Для вирішення проблеми відтоку клієнтів необхідно впровадження алгоритмів які будуть вирішувати проблему холодного старту.

1.2 Дослідження методів побудови рекомендацій

У наш час існують різноманітні підходи до побудови рекомендаційних систем. Дані підходи все ще мають спільні риси засновані на їх базових алгоритмах, що допомагає їх класифікувати. Рекомендаційні системи зазвичай класифікують за способами відбору даних для користувачів. При побудові РС використовують один з двох підходів це або колаборативна фільтрація або фільтрація на основі контенту. Однак на практиці зазвичай використовуються гібридні методи, що поєднують в собі переваги приведених нижче підходів (рисунок 1.3).

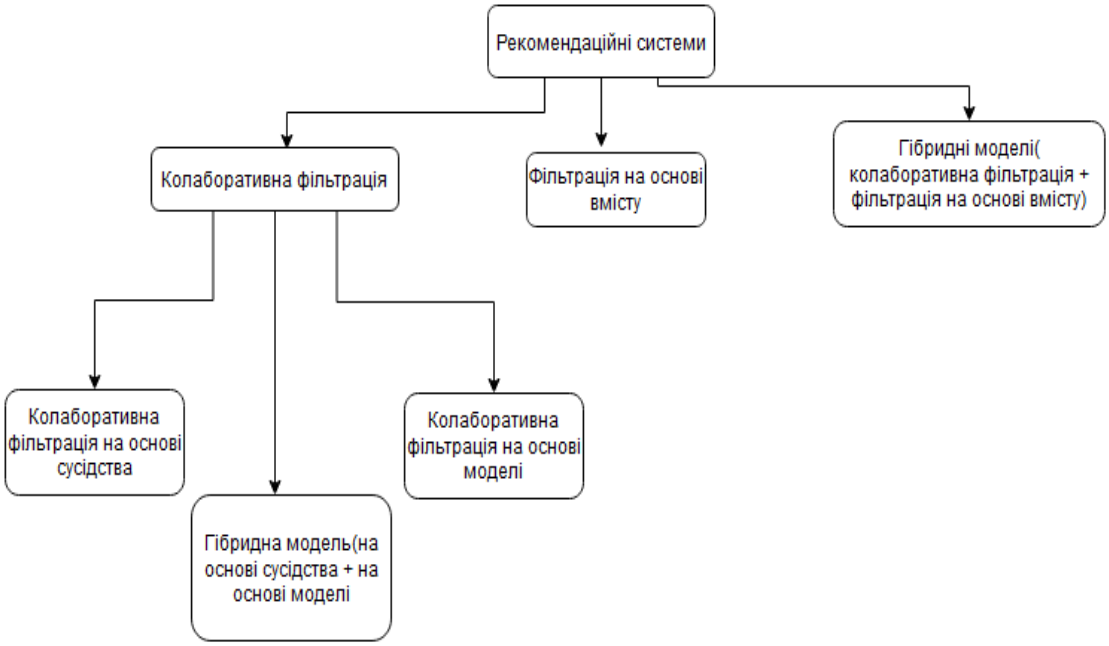

Рисунок 1.3 – Класифікація рекомендаційних систем

Таблиця 1.1 – Методи побудови рекомендаційних систем

Назва	Опис
1.Колаборативна фільтрація(КФ)	застосовують інформацію про оцінки, які користувачі виставляють об’єктам, та засновані на визначенні схожості користувачів чи об’єктів
1.1 КФ на основі сусідства	використовують коефіцієнти подібності між елементами системи (користувачами, об’єктами) для того, щоб формувати рекомендації на основі ступеню схожості елементів системи.
1.2 КФ на основі моделі	розробляються з використанням алгоритмів машинного навчання, щоб передбачити оцінку користувача для товарів без оцінки.
2. Фільтрація на основі вмісту	використовує функції елемента, щоб рекомендувати інші елементи, схожі на те, що подобається користувачеві, на основі попередніх дій або явних відгуків
3. Гібридні методи	поєднують у собі декілька методів та/або моделей роботи рекомендаційних систем.

Колаборативна фільтрація (КФ). Основна ідея цих підходів полягає в тому, щоб використовувати інформацію про минулу поведінку або оцінки, чи ще якісь додаткові дані існуючої групи користувачів щоб передбачити, які предмети поточному користувачу системи найімовірніше вподобається або зацікавлять. Такі типи систем широко використовуються в різних інформаційних системах, зокрема, як інструмент на сайтах роздрібної торгівлі в інтернеті для налаштування вмісту потреби конкретного клієнта і таким чином просувати додаткові товари та збільшення продажів.

Колаборативна фільтрація використовує матрицю оцінок користувачів і предметів лише вхідні дані та зазвичай створюють такі типи виведення: (а) а

(числовий) прогноз, що вказує, наскільки поточному користувачеві сподобається чи не сподобається певне пункт і (b) список з n рекомендованих пунктів. Такий список перших N повинен, звичайно, не містити предметів, які поточний користувач уже купив.

Для ефективної роботи підходу потрібна велика кількість оцінок від користувача або оцінок елемента рекомендація, яка не працюватиме для нових користувачів, нових елементів або обох через відсутність оцінки в системі.

Формальну роботу РС з використання колаборативної фільтрації можна представити наступним чином (рисунок 1.4)

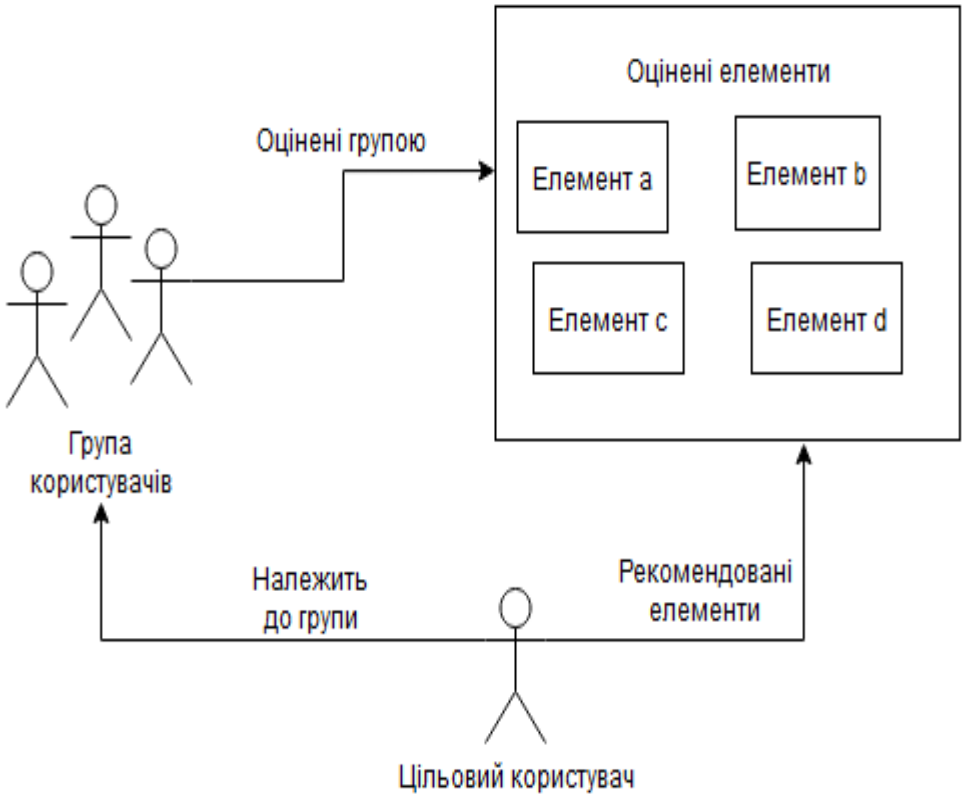

Рисунок 1.4 – Принцип роботи РС на основі КФ

Існує дві широкі категорії, та одна гібридна обох категорій, на які можна поділити колаборативну фільтрацію:

– колаборативна фільтрація на основі сусідства;

– колаборативна фільтрація на основі моделі;

– гібридна фільтрація колаборативної фільтрації на основі пам'яті та колаборативної фільтрації на основі моделі.

КФ на основі сусідства (або також називають методи на основі пам'яті), ґрунтується на побудові стосунків між елементами, чи між користувачами. Даний підхід моделює уподобання користувача к елементу на основі оцінок аналогічних елементів виставлених тим самим користувачем. Він обчислює подібність двох користувачів або предметів, та робить прогноз для користувача, приймаючи середнє зважене всіх оцінок. Методи на основі сусідства виділяються серед інших своєю простотою і ефективністю та можливістю виробляти точні та персоналізовані рекомендації. Також даний метод найчастіше використовується для впровадження в невеликих інформаційних системах, оскільки він є найдешевшим і найпростішим при реалізації [5].

КФ на основі моделі, ґрунтується на використанні інформації про користувачів і їх взаємодії з елементами. Даний підхід є більш комплексним та надає більш точніші рекомендації, бо допомагає розкрити приховані моделі поведінки яка мають пояснити взаємодії між користувачами та елементами.

Фільтрація на основі вмісту (на основі контенту), будують рекомендації товарів, ґрунтуючись на уподобань в минулому. Тобто створюється профіль користувача та профіль елемента. Параметри елементів повинні відповідати перевагам користувача. Даний підхід використовує дані о лише одного конкретного користувача і надає рекомендації лише для нього, не беручи до уваги інших користувачів [5]. Приклад фільтрації вмісту наведено на рисунку 1.5.

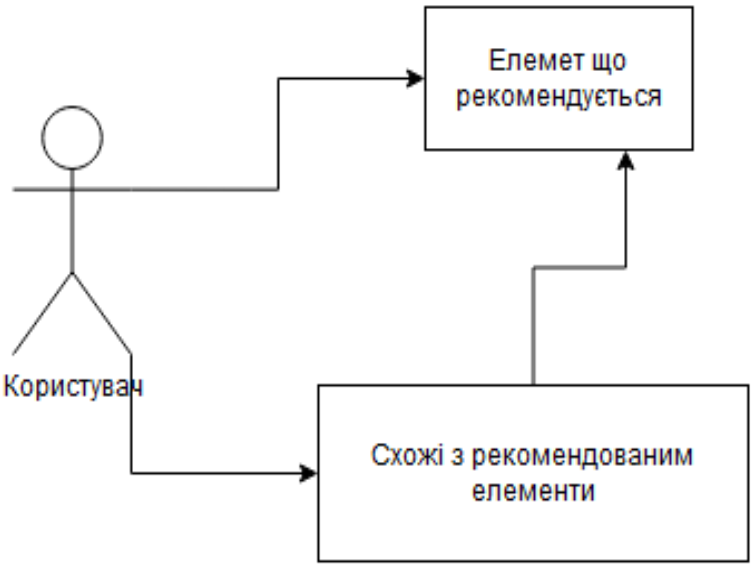

Рисунок 1.5 – Фільтрація на основі вмісту

Гібридний підхід містить у собі обидва підходи, заснований на моделі та заснований на сусідстві. Даний підхід виділяється серед інших, оскільки найчастіше застосовується при розробці рекомендаційних систем для комерційних веб-сайтів. Гібридний підхід допомагає подолати більшість недоліків, таких як: поліпшення якості передбачень уподобань користувачів, а також проблему втрати інформації. Однак незважаючи на всі переваги над іншими підходами, гібридний підхід більш складний і дорогий для реалізації і підтримки. Існує декілька основних типів комбінування при побудові гібридних систем:

– реалізація окремо колаборативних та контентних алгоритмів та поєднання їх припущень;

– включення деяких контентних правил до колаборативної методики;

– включення деяких колаборативних правил до контентної методики;

– побудова загальної моделі, що включає правила обох методик.

Основною характеристикою гібридних рекомендаційних систем є те, що вони поєднують оцінки з різних компонентів з точки зору презентації, а не з точки зору комбінування прогнозованих оцінок. У багатьох випадках

рекомендовані пункти представлені поруч з одним інший. Тому основною відмінною характеристикою таких систем є комбінація презентації, а не комбінація прогнозованих оцінок.

Хоча багато рекомендаційних систем насправді є гібридами, мало теоретичних робіт зосереджено на тому, як гібридизувати алгоритми та в яких ситуаціях можна очікувати вигоди від використання гібридного методу.

Прикладом створення такого методу був відкритий конкурс Netflix Prize, на якому сотні студентів і дослідників об’єдналися для покращення алгоритму рекомендаційної системи фільмів, для прогнозування оцінок користувачів фільмам на основі попередніх оцінок, без використання додаткової інформації. Було поєднано сотні різних методів та підходів колаборативної фільтрації для покращення загальної точності.

Використання рекомендаційних систем включає в себе багато переваг, такі як: притік нових користувачів, більш вигідні пропозиції для користувачів, конкуретно-спроможність, що призводить до збільшення прибутку. Однак як і в кожній системі РС не обійшлась без недоліків. Розберемо ретельніше переваги і недоліки найпопулярніших підходів при побудові рекомендаційних систем.

До переваг рекомендаційних систем на основі контента відносять:

– незалежність від даних інших користувачів при побудові рекомендації;

– відсутність проблеми холодного старту для нових товарів, оскільки використовуючи ознаки товарів, можливо легко знаходити подібні товари;

– результати побудови рекомендації – інтерпретовані.

До недоліків можна віднести:

– проблема холодного старту для нових користувачів в системі;

– формування груп схожих товарів, може обмежити рекомендації інших товарів. Користувачі будь отримувати одні і ті ж самі рекомендації, які входять до малої частини всіх товарів;

– при недостатності даних о товарах, трудно розрізняти товари та згруповувати їх, що призводить до зниження якості рекомендацій.

Серед переваг колаборaційних систем заснованих на сусідстві можна виділити:

– проста реалізація і інтерпретованість системи;

– очікуваний результат, коли користувач отримує подібні рекомендації за історією своїх вподобань;

– просте полегшення нових даних;

– масштабованість.

До переваг колаборaційних систем заснованих на моделі відносять:

– краща обробка розріджених матриць;

– краща робота з масштабністю даних.

Також для усіх підходів КФ властиві такі недоліки як:

– розрідженість даних, оскільки при великій кількості товарів в системі, в системі часто недостатньо оцінок від користувачів товарам, дана проблема властива системам колаборaційним підходом заснованих на сусідстві;

– шахрайство, часто трапляється коли люди можуть спеціально надавати високі оцінки одним товарам, а наприклад своїм конкурентам занижувати;

– відсутність різноманітності, коли користувачам постійно рекомендують одні й ті ж самі товари, не надаючи рекомендації нових цікавих пропозицій;

– проблема холодного старту, нові користувачі чи товари, що щойно з'явилися у системі, уявляють в собі проблему для рекомендаційних систем.

Гібридні системи допомагають подолати більшість проблем, беручи від обох підходів КФ найкраще, тим самим допомагаючи уникнути обмеження початкового оригінального підходу, що зустрічається підході заснованому на сусідстві, а також допомагає покращити якість результатів рекомендації. Однак незважаючи на усі свої переваги гібридний підхід страждає від складності реалізації та застосуванню системи, а також від дорожнечі.

При побудові РС частіше за усе виникають проблеми з відсутністю або неповнотою вхідних даних. Проблеми рекомендаційних систем наведено в таблиці 1.2.

Таблиця 1.2 – Проблеми побудови рекомендацій

Назва проблеми	Опис
1. Проблема рейтингів	Велика кількість користувачів не оцінюють товари, тим самим створюючи проблему, коли неможливо дізнатися чи сподобався товар чи ні.
2. Проблема холодного старту **2.1 Проблема нового користувача** 2.2 Проблема нового товару	Коли в системі з'являється новий користувач чи товар, у нього може бути або повністю відсутня інформація, або неповнота даних. Тому система не може надати рекомендації порівнюючи користувачів або товари.
3. Проблема розрідженості вхідних даних	Розрідженість рейтингової матриці, через те що користувачі не виставляють оцінки товарам приводить до неточних рекомендацій при використанні колаборативної фільтрації.
4.Проблема масштабованості	Через збільшення користувачів в системі, сповільнюється робота рекомендаційної системи
5. Проблема зміни даних з часом	Застаріння даних щодо товарів в рекомендаційній системі приводить до формування неточних рекомендацій
6. Проблема зміни уподобань користувача	У деяких користувачів з часом може змінитися уподобання, через це вони почнуть отримувати неперсоналізовані рекомендації
7. Проблема “синонімізації”	Виникає тоді коли товари що присутні в системі мають однакову назву чи ознаки, тому система може їх плутати між собою.
8. Проблема надмірної спеціалізації	Виникає якщо об'єкти що рекомендуються користувачеві занадто схожі один на одний.

1.3 Дослідження підходів до вирішення проблеми нових користувачів рекомендаційної системи

Важливим при побудові будь-якої рекомендаційної системи є наявність даних які б вказували на потреби та вподобання користувачів. При виборі алгоритму побудови рекомендаційної системи, продуктивність відіграє важливу роль, однак точність рекомендацій заснованих на будь-якому класі рекомендаційних систем може погіршитися, якщо дані о користувачі відсутні, або присутні дані низької якості [4].

Холодний старт користувача або відвідувача трапляється тоді, коли в системі новий користувач з'являється вперше. Оскільки у нього немає історії користувача, система не знає особистих уподобань користувача, тому неможливо надати йому рекомендації.

Проблема з новим користувачем ускладнює для системи вивчення нових уподобань користувачів, ця проблема також відома як проблема холодного старту.

Основна стратегія роботи з новими користувачами полягає в тому, щоб попросити їх надати деякі параметри для створення початкового профілю користувача. Необхідно встановити поріг між тривалістю процесу реєстрації користувача, яка, якщо вона буде занадто довгою, може призвести до того, що занадто багато користувачів відмовляться від нього, та кількістю вихідних даних, необхідних для належної роботи рекомендаційних систем [2].

Проблема холодного старту привернула широку увагу в області досліджень систем рекомендацій. Було виявлено, що проблема холодного старту є крайньою стадією проблеми розрідженості даних. Для вирішення цієї проблеми було запропоновано різні підходи.

Одна з головних проблем РС – проблема холодного старту (Cold-start Problem, CSP). Вона виникає тоді, коли в системі з'являються нові предмети чи

нові користувачі (User Cold-Start), історія вподобань яких порожня, або нові предмети (Item Cold-Start), у яких ще немає оцінок або набору атрибутів [5].

У багатьох реальних системах CSP може набувати характеру циклічної проблеми для вже відомих користувачів або об'єктів. Наприклад, якщо частина користувачів змінює свої інтереси. Дана проблема отримала назву проблеми постійного холодного старту (Continuous CSP, CoCoS) [5].

Як і CSP, проблема CoCoS може виникати з користувачами (User Continuous CSP) та з об'єктами (Item Continuous CSP).

User Continuous Cold-start Problem виникає для користувачів, що змінюють свої вподобання, або рідко з’являються у системі та рідко оцінюють нові об’єкти [5].

Item Continuous Cold-start Problem виникає при наявності об'єктів, властивості яких можуть змінитися з часом.

Для вирішення проблеми холодного старту, як правило, застосовують наступні підходи:

– гібридизація РС з поєднанням контентної та колаборативної фільтрації.

– використання контексту, в якому створюються та надаються рекомендації (демографічні дані, час та дата, тощо).

Однак всі ці способи не підходять в разі проблеми Continuous Cold-start Problem, оскільки припускають, що після того, як користувач став «відомим», він залишається таким необмежену кількість часу, а об’єкти рекомендацій не можуть змінювати свої властивості. Для рішення даної проблеми необхідно не тільки прогнозувати вподобання користувачів, а й відслідковувати та прогнозувати зміну їх вподобань, а також враховувати можливість зміни властивостей об’єктів рекомендацій [6].

Дану проблему на сьогоднішній день намагаються вирішувати методами машинного навчання, що підвищують адаптивність системи до постійних змін.

Для вирішення проблеми холодного старту використовують різні підходи. Зосередимося на гібридних системах колаборативної фільтрації, оскільки вони є

найбільш популярними серед існуючих систем рекомендації. Розберемо різні сценарії при яких в системах може статися холодний старт, і приведемо можливі вирішення даної проблеми.

У сучасних системах рекомендації для подолання проблеми холодного старту нових користувачів найчастіше використовують такі підходи.

Загальну таблицю методів вирішення проблеми холодного старту наведено в таблиці 1.3.

Таблиця 1.3 – Методи вирішення проблеми холодного старту

Назва методу	Характеристика
1. Демографічна фільтрація	генерує рекомендації на основі демографічних атрибутів користувача. ДФ класифікує користувачів на основі їх атрибутів та рекомендує товари, використовуючи їх демографічні дані.
2. Contextual “bandit”.	Контекстна інформація може використовуватися для групування користувачів за загальними ознаками за допомогою методів класифікації або кластеризації з припущенням, що користувачі, що належать до одного і того ж кластера, схильні поводитися однаково, в той час як користувачі, що належать до різних кластерів, поводяться значно по-різному.

Кінець таблиці 1.3

3. Matrix factorization	Якщо кожен користувач пов'язаний з певними атрибутами (наприклад віком, статтю), то може бути визначена функція впровадження, яка, враховуючи дані користувача, оцінює його відповідні приховані фактори. Функція вбудовування навчається на даних теплих користувачів. У порівнянні ознак, можемо отримати обґрунтовану рекомендацію
4. Візуальний байєсівський персоналізований рейтинг (VBPR)	ідея полягає в створенні моделей, які включають візуальні функції до завдання персоналізованого рейтингу на основі даних неявного зворотного зв’язку.
5. Domain Recommender Systems	веб-додатки для електронної комерції зазвичай працюють у кількох доменах і використовують механізми агрегації кількох типів даних із кількох доменів.
6. Метод на основі тем.	Прогноз користувача моделюють за допомогою імовірнісного розподілу тем стосовно різних об’єктів. Під час взаємодії користувача з РС його наміри передбачається через марковский процес

РС, засновані на демографічній фільтрації (ДФ), генерує рекомендації на основі демографічних атрибутів користувача. ДФ класифікує користувачів на основі їх атрибутів та рекомендує товари, використовуючи їх демографічні дані. Припускається, що повинні бути надані різні рекомендації, Багато веб-сайтів використовують прості та ефективні рішення персоналізації на основі демографічних показників. Наприклад, користувачі надсилаються на певні веб-сайти залежно від їхньої мови чи країни. Або пропозиції можуть бути

налаштовані відповідно до віку користувача для різних демографічних груп На відміну від спільної фільтрації і системи рекомендацій на основі контенту, вона проста в реалізації і не вимагає оцінок користувачів [2].

Метод “Multi-Armed Bandit”(далі MAB) (або метод contextual bandit) використовує різні алгоритми для вирішення проблеми холодного старту. Холодний старт можна переосмислити і вирішити як проблему “бандита”. Що таке “бандитська” проблема? Класичний приклад такий: кожного дня в системі з'являються нові користувачі, кожен з них має свої уподобання. Але оскільки користувачі нові, невідомо що саме їм сподобається, тому неможливо надати відповідні рекомендації.

Існує кілька алгоритмів MAB, які використовуються для вирішення вищезазначеної проблеми “бандита”, кожен з яких різною мірою має свої переваги. Найпопулярнішими з них є Epsilon Greedy, Thompson Sampling і Upper Confidence Bound 1 (UCB-1) [6].

Epsilon Greedy, як випливає з назви, є найбільш жадібним з трьох алгоритмів MAB. В експериментах з епсилон-жадібністю константа ε (значення від 0 до 1) вибирається користувачем перед початком експерименту. При розподілі контактів за різними варіантами кампанії в більшості випадків вибирається випадково обраний варіант. В інших 1-ε випадках вибирається варіант з найбільшим відомим виграшем. Чим вище ε, тим більше цей алгоритм сприяє дослідженню.

Для кожного варіанту кампанії ми визначимо верхню межу довіри (UCB), яка представляє наше найвище припущення про можливий виграш для цього варіанту. Алгоритм призначить варіанти з найбільшим UCB.

Вибірка Томпсона, є підходом, який може дати більш збалансовані результати в несподіваних випадках. Для кожного варіанту будуємо розподіл ймовірностей (найчастіше бета-розподіл, з обчислювальних міркувань) справжнього показника успіху, використовуючи спостережувані результати. Для кожного нового контакту вибираємо один можливий показник успішності з бета-

дистрибутива, що відповідає кожному варіанту, і призначаємо контакт варіанту з найбільшим показником успішності вибірки. Чим більше точок даних ми спостерігали, тим більше будемо впевнені в істинному показнику успіху, і тому в міру збору більшої кількості даних показники успіху вибірки будуть все більш і більш близькі до істинного показника [6].

Одним з найбільш успішних алгоритмів рекомендаційної системи є matrix factorization (розкладання матриць). На цю тему було опубліковано велику кількість дослідницьких публікацій, і алгоритмічна парадигма набула широкого поширення у всіх видах інтернет-компаній. Найбільш загальною основою матричного розкладання є Функція SVD, яка моделює матричне розкладання, як засновані на характеристиках точкові твори збагачених функцій користувача і елемента вектору [7]. Майже всі варіанти розкладання матриці можуть бути змодельовані як конкретний приклад функції SVD. В основі методу МФ існують припущення, за якими він працює. По-перше це існування деяких факторів, що унікально описують послуги, які надаються інформаційною системою, а по-друге що ці фактори також використовуються при описі уподобань користувачів. Найчастіше система сама визначає зв'язки, які не мають жодної логічної інтерпретації , однак які обумовлені математично. Результатом МФ є розклад матриці user-item на дві окремі матриці item-factor і user-factor, матриці об'єктів та користувачів. Розмірність же вихідних матриць надає змогу відрегулювати ступінь персоналізації. Таким чином розкладаючи матрицю на вектор і вектор-стовпець, значення, що будуть набувати зв'язки, й будуть відповідати найбільш популярним елементам.

Візуальний байєсівський персоналізований рейтинг (VBPR). Основна ідея полягає в створенні моделей, які включають візуальні функції до завдання персоналізованого рейтингу на основі даних неявного зворотного зв’язку для подолання проблеми нових користувачів та підвищення якості рекомендацій. Методологічно візуальні характеристики продуктів виділяються за допомогою глибинної нейронної мережі, поверх якої встановлено додатковий шар, який

розкриває як візуальні, так і приховані розміри, які відповідають вподобанням користувачів. Для отримання рекомендацій використовується комбінація матричної факторизації, візуальних характеристик та процедури навчання на основі байєсівського персоналізованого рейтингу (BPR) [8].

Domain Recommender Systems – веб-додатки для електронної комерції зазвичай працюють у кількох доменах і використовують механізми для поєднання кількох типів даних із кількох доменів в одну одиницю. Наявність таких даних може принести користь у систему рекомендацій і дозволяє їй здійснювати, наприклад, перехресні продажі або боротися з проблемою холодного старту в його цільовому домені [9].

Метод на основі тем полягає в тому, що прогноз користувача моделюють за допомогою імовірнісного розподілу тем стосовно сайтів або інших об’єктів (наприклад, статей). Під час взаємодії користувача з РС його наміри передбачається через марковский процес [10].

Проблема холодного старту, яка описує труднощі надання рекомендацій, коли користувач є новим, залишається суттєвою проблемою для колаборативної фільтрації. Традиційно цю проблему вирішують, вдаючись до додаткового процесу співбесіди для встановлення профілю користувача перед наданням будь-яких рекомендацій.

За результатами дослідження було обрано саме методи демографічної фільтрації, як найефективніші серед розглянутих, для побудови удосконаленого гібридного алгоритму.

1.4 Постановка задачі дослідження

За результатами проведення аналізу існуючих підходів до побудови рекомендаційних систем було виявлено ряд проблем. Основною з них було

виділено проблему холодного старту для нових користувачів. Проблема холодного старту виникає тоді, коли в системі не вистачає даних, оскільки користувач скоріш за все не почне одразу ставити оцінки, а може й деякий проміжок часу не виявляти жодних дій, рекомендаційні системи не зможуть надати точних рекомендацій. Проблема холодного старту нових користувачів є дуже актуальною для онлайн-кінотеатрів, оскільки коли новий користувач починає використовувати сервіс для перегляду фільмів, через відсутність актуальних рекомендацій, користувач може припинити користуватися системою, що призведе до втрати клієнта.

Розглянуті методи можуть слугувати вирішенням проблеми холодного старту, однак проблемою є що більшість з них спираються на демографічну фільтрацію [11], не враховуючи особисті якості користувачів.

Було вирішено поєднати дані користувачів за результатами анкетування з демографічними характеристиками, оскільки користувачі з однаковими демографічними характеристиками можуть мати різні уподобання або поведінку, тому додаткові дані є необхідною умовою фільтрації для користувачів, щоб знайти кращих довідників для рекомендацій.

Вирішенням проблеми холодного старту шляхом об'єднання особистісних рис користувача і демографічних атрибутів, щоб отримати “найближчих сусідів” для нового користувача, що дозволяє нам знаходити точних схожих користувачів, навіть якщо у нового користувача немає історії рейтингів. Найбільш важливим завданням в рекомендаційної системі є пошук схожих користувачів з використанням відповідної міри подібності, оскільки різні вимірювання призводять до різних сусідніх користувачів, що, в свою чергу, призводить до різних рекомендацій.

У цьому дослідженні запропонуємо комбінацію з трьох типів обчислень подібності, роблячи це, ми можемо отримати кращу міру подібності користувача для призначеної для користувача рекомендаційної системи, щоб отримати найточніших користувачів для нового користувача.

У результаті виконання роботи буде отримано удосконалений метод що підвищуватиме точність надання рекомендацій для нових користувачів онлайн-кінотеатру.

Об'єктом дослідження є процеси побудови рекомендацій для нових користувачів рекомендаційної системи онлайн-кінотеатру.

Метою дослідження стало дослідження методів побудови рекомендацій для нових користувачів.

Для вирішення цієї задачі було сформовано наступні етапи дослідження:

– провести аналіз систем рекомендацій;

– провести дослідження методів побудови рекомендаційних систем;

– провести дослідження проблеми нових користувачів;

– розглянути підходи для вирішення проблеми холодного старту при побудові рекомендаційних систем;

– побудувати удосконалений метод формування рекомендацій для нових користувачів;

– сформувати технологію використання удосконаленого методу;

– провести програмну реалізацію використання удосконаленого методу;

– провести експериментальну перевірку удосконаленого методу.

2 ДОСЛІДЖЕННЯ МЕТОДІВ ВИРІШЕННЯ ПРОБЛЕМИ ХОЛОДНОГО СТАРТУ

2.1 Дослідження комбінованих підходів для вирішення проблеми холодного старту в рекомендаційної системи

Появу рекомендаційних технологій можна датувати з винайдення колаборативної (або спільної) фільтрації. З тих пір і до нашого часу з'явилося багато дослідницьких робіт з різних тем. Були винайдені такі важливі винаходи як матричне розкладання [12], навчання ранжуванню [13], навчання засновані на глибоких алгоритмах [14], які крок за кроком покращили показники технічної точності, та взаємодію користувачів з системою. Обладнання допоміжних алгоритмів побудови РС стає більш і більш складнішим.

У наш час майже у всіх онлайн кінотеатрах присутні системи персональних рекомендацій. Вони допомагають користувачам обрати саме ці фільмі та серіали, з більш ніж десятків тисяч одиниць контенту, які можуть сподобатися користувачеві.

Основним завданням рекомендаційної системи будь–якого онлайн-кінотеатру є знайти зв'язки, які фільми з якими разом зазвичай переглядають користувачі. За результатом отриманих даних можна сформувати сусідніх користувачів в групу за схожістю уподобань, оскільки зазвичай це дозволяє зробити висновки, що для сусідніх користувачів, які високо оцінили одній й ті самі фільми, можна рекомендувати одні й ті самі товари. Це може використовуватися як суто нова інформація, тобто передбачення вподобань користувача щодо нових елементів, так і для коригування роботи системи, за рахунок порівняння фактичної оцінки що надав користувач з прогнозованою.

Найбільш важливим завданням в системах рекомендацій на основі користувачів – це пошук подібних сусідніх користувачів, використовуючи відповідну міру подібності, оскільки різні вимірювання призводять до

формування різних груп користувачів, що, в свою чергу, призводить до різних рекомендацій.

Хоча рекомендаційні системи еволюціонують вже досить довгий час, все одно залишаються деякі внутрішні проблеми, які мають бути вирішними. Однією з найрозповсюджених проблем РС залишається проблема холодного старту. Вона виникає коли в системі з'являється новий користувач у якого відсутня історія відвідувань і ще не має сформованих персональних вподобань, це стає проблемою для роботи рекомендаційних систем.

Колаборативна фільтрація, один з найпопулярніших алгоритмів фільтрації традиційна система рекомендацій, заснована на рейтинговій структурі, зазвичай представленої у вигляді матриці оцінок користувачів і товарів. Значення кожної комірки представляє оцінки товару користувачем. Він прогнозує рейтинги на основі таких показників подібності, як коефіцієнт кореляції Пірсона, косинусну подібність, міру евклідової відстані та інші [15].

РС, в основі яких лежить колаборативна фільтрація для своєї роботи аналізують та використовують дані про користувачів. Кожному користувачу знаходиться деяка група користувачів, зі схожими з ним смаками. На основі цього створено гіпотезу, що користувачі, які однаково оцінили деякі об’єкти в минулому, швидше за все, однаково оцінюватимуть інші об’єкти в майбутньому [16].

Однак в класичному вигляді КФ стикається з проблемою холодного старту. Для вирішення цієї проблеми найчастіше використовують гібридні алгоритми на основі колаборативної фільтрації з використанням демографічних характеристик користувачів.

Розділимо метод колаборативної фільтрації на два окремі:

- фільтрація на основі сусідства;
- демографічна фільтрація.

В методі фільтрації на основі сусідства, рейтинг елемента e для користувача x прогнозується за рахунок існуючих оцінок інших користувачів,

які система визначила мають схожі до цільового користувача уподобання. Даний підхід залежить від обраної кількості сусідів користувача x, оцінки яких будуть враховуватися при прогнозуванні рейтингу. Власне ступінь сусідства, тобто наскільки користувачі мають однакові уподобання розраховується за допомогою міри схожості (кореляція Пірсона) [17].

Розглянемо три типи обчислення подібності.

Алгоритм заснований на сусідстві обчислює кореляцію між двома користувачами на основі їх рейтингу, який вони надають схожим товарам, формує прогноз для користувача, приймаюче середнє зважене значення всіх оцінок. Обчислення схожості між користувачами є важливими для цього підходу, використовується для цього кореляція Пірсона:

$$s(x,y) = \frac{\sum_{e=1}^{n}(R_{x,e}-\overline{R_x})(R_{y,e}-\overline{R_y})}{\sqrt{\sum_{e=1}^{n}(R_{x,e}-\overline{R_x})^2}\sqrt{\sum_{e=1}^{n}(R_{y,e}-\overline{R_y})^2}}, \quad (2.1)$$

де $R_{x,e}$ і $R_{y,e}$ – рейтинг користувачів x і y для елементів e;

$\overline{R_x}$ та R_y – їх відповідний середній рейтинг, n – кількість елементів.

Подібність, заснована на зміні інтересу користувача. Обчислення подібності за кореляцією Пірсону, для двох користувачів X і Y наведено нижче:

$$s(x,y) = \frac{\sum_{o=1}^{n} Esim(e,o)^2 \times t(x,o) \times (R_{x,e}-\overline{R_x})(R_{y,e}-\overline{R_y})}{\sqrt{\sum_{o=1}^{n} Esim(e,o) \times t(x,o) \times (R_{x,e}-\overline{R_x})^2}} \times \frac{\sum_{o=1}^{n} t(y,o) \times (R_{y,e}-\overline{R_y})}{\sqrt{\sum_{e=1}^{n}(Esim(e,o) \times t(y,o) \times (R_{y,e}-\overline{R_y})^2}}, \quad (2.2)$$

де $s(x,y)$ – подібність між користувачами x та y;

$Esim(e,o)$ – подібність між предметами e та o;

$t(y,o)$ та $t(x,o)$ – вага часу.

Чим ближче час до сьогодення тим більша ймовірність того що відображені дані будуть актуальні:

$$t(x,e) = (1-\varepsilon) - \varepsilon\frac{T_{xe}}{T_x}, \quad (2.3)$$

де ε – фактор для контролю ваги часу $\varepsilon \in (0,1)$.

Якщо ε більше то вага часу буде більше при розрахунку подібності.

Розглянемо більш детально метод демографічної фільтрації.

Демографічні дані користувачів відіграють дуже важливу роль при формуванні груп користувачів для надання рекомендацій. Так як РС часто стикаються з проблемою холодного старту при появі нових користувачів, через недостатню кількість даних про оцінки та уподобання нового користувача. В такій ситуації необхідно використовувати дані, які користувач надає з початку роботи з системою. Демографічні дані про користувача відносяться до такого типу даних. Використовуючи ці дані система зможе формувати рекомендації для користувача без історії більш точно, адже, знаючи вік, стать та країну проживання користувача значно легше віднести його до певної групи [17].

Метод демографічної фільтрації працює за принципом поділу користувачів на демографічні групи з точки зору їх особливих атрибутів. Саме ці групи виступають вхідними даними для побудови рекомендації. Цей процес необхідний для виявлення групи людей, яким подобається один товар. Наприклад користувачам з групи G подобається де-який товар P і є користувач h, з цієї ж групи, який ще не переглядав товар P, то цей товар можна рекомендувати користувачу h.

Подібність на основі демографічних атрибутів користувачів. Демографічні атрибути користувачів містять дані, які користувачі надають при реєстрації найчастіше це – ім'я, країна проживання, мова спілкування. Однак в залежності від вимог системи, ці елементи можуть змінюватися. Даний метод враховує лише демографічні атрибути користувачів. Даний метод широко використовується для полегшення проблеми холодного старту. Подібність на основі демографічних атрибутів користувачів наведено нижче:

$$D_{sim(x,y)} = \cos_{sim(\vec{x},\vec{y})} = \frac{\vec{x},\vec{y}}{|\vec{x}|*|\vec{y}|} = \frac{\sum_{e=1}^{n} x_e y_e}{\sqrt{\sum_{e=1}^{n} xe^2}\sqrt{\sum_{e=1}^{n} ye^2}}, \quad (2.4)$$

де D_{sim} – демографічна подібність;

$\cos_{sim}$ – косинусна подібність;

x_e і y_e компоненти векторів x та y відповідно.

В свою чергу, кожні користувачі x_e і y_e в системі мають такі демографічні характеристики:

$$x_e \; and \; y_e = \{Id, a, g, c, l\}, \quad (2.5)$$

де Id – унікальний ідентифікатор користувача;

a – вік користувача;

g – стать користувача;

c – країна проживання;

l – мова спілкування.

Однак, що робити у разі, коли уподобання нового користувач, якого система відносить до групи з схожими демографічними даними, не сходяться. Більшість досліджень використовують демографічну фільтрацію, щоб вирішити проблему холодного запуску, якщо користувачі не мають історії рейтингів подібність між ними буде розрахована на основі їхніх демографічних атрибутів, це означає, що якщо x1 схожий на x2 за віком, статтю тощо, це означає, що x2 може рекомендувати x1 і навпаки. Однак виникає проблема, коли два користувача однакового віку, статі та з однаковим місцем проживання можуть мати різні риси особистості. Тому що інтереси користувачів найчастіше відрізняються залежно від їх особистості.

Таким чином, рекомендація, заснована лише на демографічних характеристиках, може змусити нас прийняти неправильне рішення, тобто рекомендувати товари користувачам, які не відповідають їх вимогам. Якщо об'єднаємо демографічні атрибути та дані результатів анкетування користувачів

і обчислимо їхню схожість, результат буде кращим і точним, навіть якщо обидва користувачі не мають історії оцінок.

2.2 Удосконалення гібридного методу колаборативної фільтрації для нових користувачів з використанням даних анкетування

Розглянувши типи обчислення подібності: схожих користувачів на основі сусідства, на основі демографічних даних, було вирішено скомбінувати ці підходи, додавши обчислення подібності основі даних анкетування в один.

Для вирішення проблеми холодного старту, поєднаємо дані анкетування користувача і його демографічні атрибути, щоб отримати m-найближчих сусідів для нового користувача, що дозволить знаходити точних схожих користувачів, навіть якщо у нового користувача немає історії рейтингів.

За результатами роботи буде отримано удосконалений гібридний метод який буде враховувати демографічні дані користувачів, дані оцінок користувачів а також особистих даних користувачів в роботі рекомендаційної підсистеми інформаційної системи онлайн-кінотеатру (рисунок 2.2).

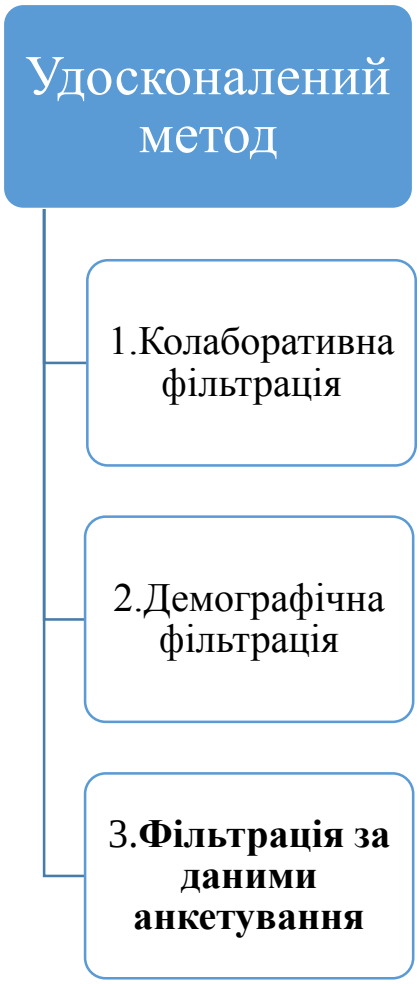

Рисунок 2.2 – Структура удосконаленого методу.

Ідея створеного удосконаленого методу полягає в об'єднанні кількох методів фільтрації в один. Таким чином це надає можливість поєднати переваги представлених методів, для більш точного рішення з побудови рекомендацій для нових користувачів.

Таким чином, рекомендація, заснована тільки на демографічних характеристиках, може привести нас до неправильного рішення, рекомендуючи товари невідповідним користувачам. Якщо додати демографічні характеристики та дані анкетування користувачів та розрахуємо їх схожість, результат буде краще і точніше, навіть якщо у двох користувачів немає історії оцінок.

Однак у удосконаленого методу все ще залишається недолік у вигляді високій обчислювальній складності при великій кількості даних.

Етапи виконання удосконаленого методу наведено нижче.

Етап 1. Збір даних. На цьому етапі формується профіль користувача. Під час початкової взаємодії із рекомендаційною системою, користувачу необхідно зареєструватися та пройти анкетування. Це дасть змогу РС отримати початкові дані для ініціювання побудови профілю для нового користувача. Ці дані необхідні в подальших розрахунках.

Етап 2. Обчислення подібності. На цьому етапі обчислюють вагові коефіцієнти подібності для кожного користувача з використанням демографічних даних та розрахунок подібності на основі оцінок, метою отримати кількість користувачів, які мають приналежність до цільового користувача, для формування груп.

Крок 1. Для кожного користувача *x* обчислюємо його схожість до користувача *y*. Для цього використаємо колаборативну фільтрацію засновану на сусідстві обчисленням подібності по Пірсону. це метод, який обчислює кореляцію між двома користувачами на основі їхніх оцінок, які вони дають подібним предметам, і враховує час, коли користувач оцінював товар, як один із факторів для визначення подібності, що це означає, оцінку, яку користувач дає елементу зараз може змінюватися з часом через зміну інтересів користувачів, оскільки рейтинг не є постійною величиною. Використаємо формулу (2.1).

За результатом цього кроку вибрали зі списку користувачів, які є найбільш наближеними до даного користувача, групу визначеного розміру.

Крок 2. Використаємо проаналізовані демографічні дані користувачів, отримані на етапі 1. Обчислюємо вагові коефіцієнти для кожного користувача подібність користувачів на основі демографічних атрибутів (вік, стать, країна проживання тощо).

Для цього використаємо формулу (2.4).

Результатом цього етапу буде отримано кількість користувачів, які мають подібні демографічні дані, з користувачам який формує сусідство (тобто належать до однієї групи).

Етап 3. Отримання прогнозів та побудова рекомендацій. Скомбінуємо демографічних атрибутів користувачів з їхніми результатами анкетування, що користувачі вказали під час опитування при реєстрації.

Крок 1. Розрахуємо вагові коефіцієнти для кожного користувача на основі подібності за даними анкетування (дані отримані від користувачів які окремо проходили особистісний тест для визначення особистісного бала Big 5) користувачів, використаємо міру схожості:

$$sp(x,y) = \frac{\sum_m (P_x - \overline{P_x})(P_y - \overline{P_y})}{\sqrt{\sum_m (P_x - \overline{P_x})^2}\sqrt{\sum_m (P_y - \overline{P_y})^2}}, \quad (2.6)$$

де $sp(x,y)$ – подібність, заснована на особистості користувачів x та y;

$\overline{P_x}$ та $\overline{P_y}$ є середніми особистостями користувачів x та y.

Крок 2**.** Скомбінуємо демографічні атрибути користувачів з їхніми даними отриманими при анкетуванні, що користувачі вказали під час опитування при реєстрації для отримання сусідніх користувачів групи. На даному етапі коефіцієнт подібності $Ed\_sim(x,y)$ розраховується як сума коефіцієнтів подібності за демографічними даними та даними анкетування:

$$Ed_{sim(x,y)} = D_{sim}(x,y) + sp(x,y), \quad (2.7)$$

$$Ed_{sim(x,y)} = \frac{\sum_{e=1}^{n} x_e y_e}{\sqrt{\sum_{e=1}^{n} xe^2}\sqrt{\sum_{e=1}^{n} ye^2}} + \frac{\sum_m (P_x - \overline{P_x})(P_y - \overline{P_y})}{\sqrt{\sum_m (P_x - \overline{P_x})^2}\sqrt{\sum_m (P_y - \overline{P_y})^2}}, \quad (2.8)$$

де $Ed\_sim(x,y)$ – розширена демографічна подібність між користувачами x та y;

$D_{sim}(x,y)$ – демографічна подібність між користувачами x та y;

$sp(x,y)$ – подібність, заснована на даних анкетування користувача.

У результаті отримали групу сусідніх користувачів, що є більш наближеними до даного користувача за демографічними та даними з анкетування.

Крок 3. На цьому кроці за прогнозом для цільового користувача побудуємо рекомендації. В алгоритмі КФ на основі користувачів, підмножина найближчих сусідів, з однієї групи, цільового користувача вибирається на основі їх схожості з ним, а зважена сукупність їхніх оцінок використовується для побудови рекомендацій для цільового користувача. Використовуємо удосконалений метод простого методу прогнозування на основі користувача для прогнозування оцінки для цільового користувача:

$$P_{x,e} = \bar{R}_x + \frac{\sum_{y=1}^{b}(R_{y,e} - \bar{R}_y) \times Ed\_sim(x,y)}{\sum_{y=1}^{b}|Ed\_sim(x,y)|}, \quad (2.9)$$

де $P_{x,e}$ – прогноз користувача x на елемент e;

$R_{y,e}$ – оцінка елементу e від існуючого користувача y;

$\bar{R}_x$ та $\bar{R}_y$ – середня оцінка існуючого користувача x та y;

b – кількість користувачів;

Ed_sim – розширена подібність з урахуванням демографічних і особистих рис користувачів.

Результатом цього етапу є впорядкований набір рекомендованих об'єктів, який враховує інтереси цільового користувача, і навіть найбільш популярні об'єкти схожих користувачів.

3 ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОБУДОВИ РЕКОМЕНДАЦІЙ З ВИКОРИСТАННЯМ ОСОБИСТИХ ДАНИХ

3.1 Технологія використання удосконаленого методу

Технологія використання удосконаленого методу побудови рекомендації, заснованому на демографії та особистості користувачів, виконується у три етапи:

1) збір даних;
2) розрахунок подібності;
3) прогноз рекомендації.

Етапи технології наведені на рисунку 3.1.

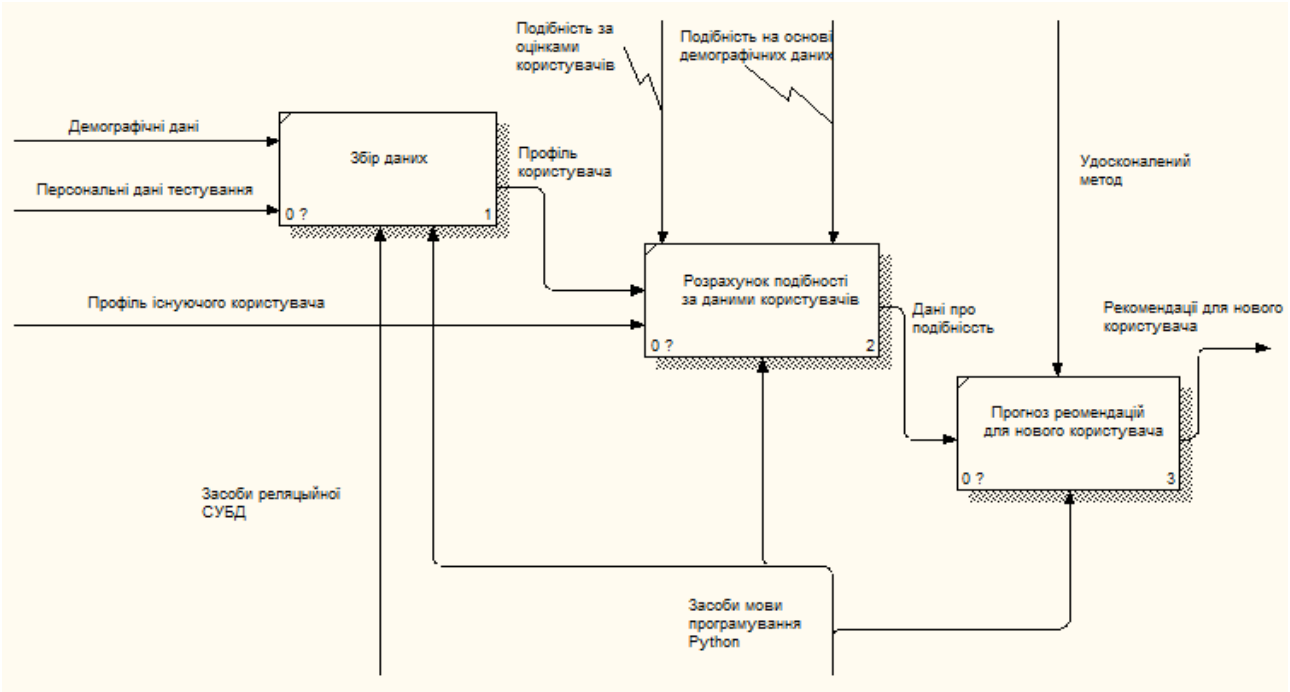

Рисунок 3.1 – Технологія удосконаленого методу

Збір даних – це етап, на якому збирається інформація про користувачів враховуючи демографічні та особистісні дані нових цільових користувачів. РС необхідно мати якомога більше інформації про користувача, щоб надати якісні рекомендації Під час початкової взаємодії користувача із рекомендаційною системою користувачу необхідно зареєструватися та пройти анкетування. Це

дасть змогу РС отримати початкові дані для ініціювання побудови профілю для нового користувача. Користувач повинен чітко визначити свою демографічну інформацію і пройти особисте тестування TIPI (Ten Item Personality Inventory) для оцінки особистостей користувачів (Велика п'ятірка Особистісні риси) і користувачі окремо проходили особистісний тест для визначення особистісного бала Big5 (відкритість, сумлінність, екстраверсія, доброзичливість, емоційна стабільність) [18].

Профіль користувача – це набір особистих даних кожного із користувачів, а саме: інтереси, уподобання користувача, демографічні дані, особисті риси, історію оцінок. Профіль користувача використовують для побудови моделі користувача.

Успіх будь-якої рекомендаційної системи багато в чому залежить від її здатності представляти поточні інтереси користувача. Точні моделі незамінні для отримання відповідних і надання точних рекомендацій із використанням будь-яких методів [2].

Етап розрахунку подібності використовує метод колаборативної фільтрації заснований на сусідстві, та розрахунок подібності за демографічними даними, щоб отримати кількість користувачів, що мають схожу демографію і схожі оцінки з цільовим користувачем, що формує сусідство.

На етапі прогнозування рекомендацій отримуємо пропозиції, що були позитивно оцінені сусідніми користувачами, які будуть запропоновані цільовому користувачу. Припускається, що користувачі зі схожими демографічними показниками та особистими рисами оцінюватимуть предмети однаково.

За результатом отриманого прогнозу, формується впорядкований набір об'єктів, який враховує інтереси цільового користувача.

4 ПРАКТИЧНЕ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

4.1 Програмна реалізація удосконаленого методу

Розроблювана рекомендаційна система буде представлена клієнт-серверним додатком. У серверній частині будуть реалізовані такі модулі: модуль обробки вхідних даних користувачів, модуль формування рекомендацій. У клієнтській частині, будуть запитуватися списки рекомендацій з серверної частини відповідно до вхідних даних користувача і у результаті виводити на екран список рекомендацій.

Модель, використана в даній роботі є універсальною та надає можливість для різних модифікацій та доробок з метою покращення та розширення роботи системи.

Запропонована структура містить наступні кроки:

– зчитування та збереження профілю користувача в основній пам'яті (для побудови моделі користувача). Представлені два типи даних в профілі користувача: демографічні атрибути, дані анкетування;

– застосування міри подібності на основі історії оцінок;

– застосування подібності на основі демографії та даних анкетування;

– виявлення сусідніх схожих користувачів;

– вибір пунктів для надання рекомендацій користувачеві;

– показ рекомендованих елементів користувачеві.

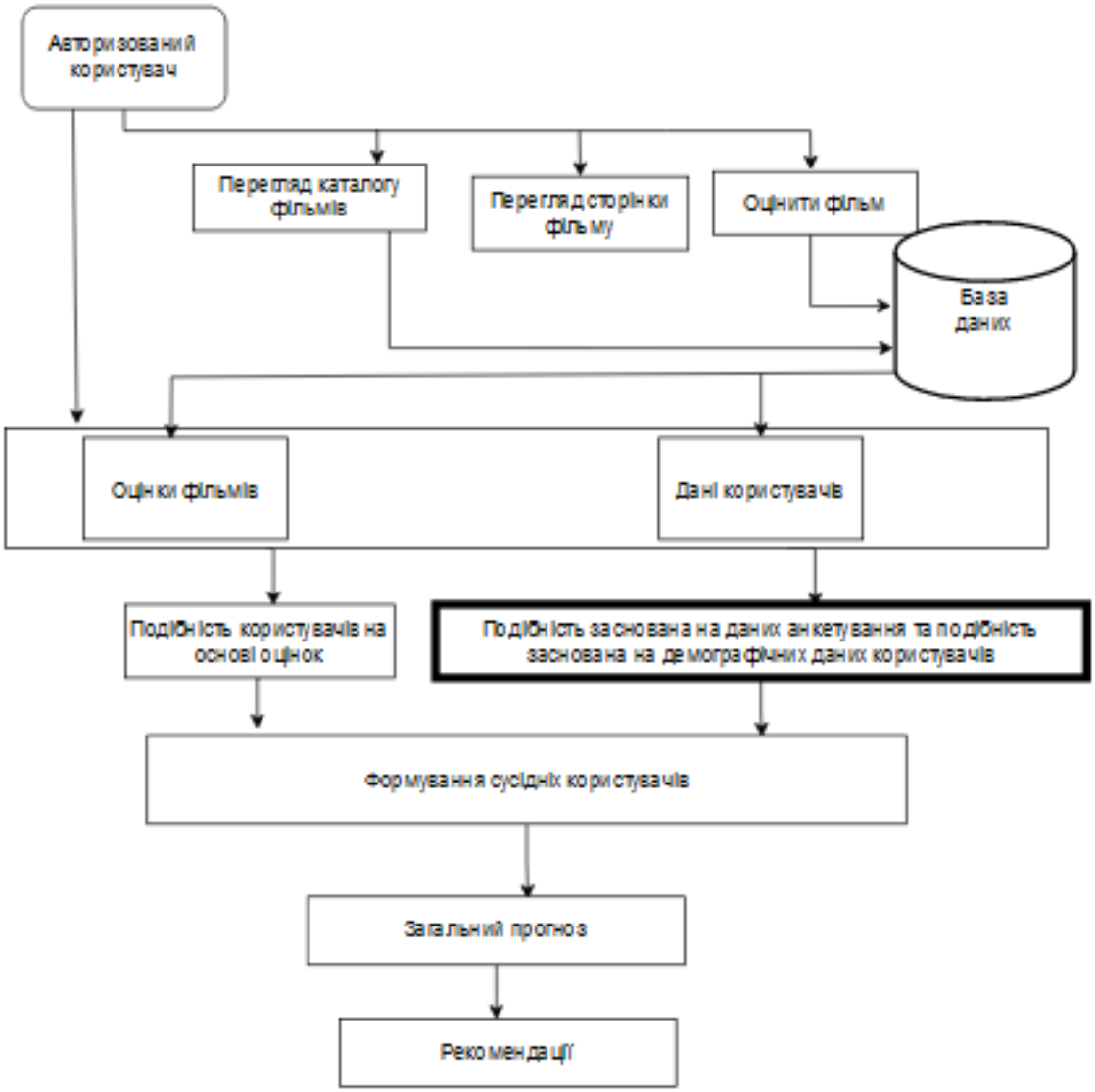

Рисунок 4.1 – Загальна архітектура системи

Вхідними даними виступає інформація, що користувачі вказували при реєстрації та дані додаткового опитування при реєстрації. Дані зберігаються таблицях баз даних. Приклад демографічних даних користувачів використовуваних при побудові рекомендацій наведено у таблиці 4.1.

Таблиця 4.1 – Приклад демографічних даних користувачів

ID	Name	Age	Gender	Country	Language
37	Mark	26	M	England	English
38	Ivan	20	M	Ukraine	Ukrainian
39	Kate	35	F	USA	English

Кінець таблиці 4.1

40	Mike	22	M	Poland	Polish
41	Nina	49	F	Ukraine	English

Де Id – унікальний ідентифікатор користувача, Name - ім'я користувача, Age – вік користувача, Gender – стать користувача, Country – країна проживання користувача, Language – мова спілкування користувача.

Дані анкетування користувачів зберігаються у вигляді дескрипторів особистості, які формуються за результатами тесту на оцінку особистості Big5 (екстраверсія, доброзичливість, сумлінність, емоційна стабільність, відкритість). Однак слід враховувати час затрачуваний користувачами на анкетування, оскільки у разі надмірної втрати часу, це може відштовхнути користувача, від користування даною системою. Користувач відповідає на питання, за результатами яких Вибрані оціночні значення висловлювань перетворюються на бали (таблиця 4.2):

Таблиця 4.2 – Оціночна шкала

	-2	-1	0	1	2
Бали	5	4	3	2	1

Кожен із основних п'яти факторів складається з п'яти первинних факторів. Наприклад, основний фактор "екстраверсія - інтроверсія" складається з первинних факторів 1.1, 1.2, 1.3, 1.4, 1.5. У кількісному вираженні первинні фактори визначаються підсумовуванням трьох бальних оцінок.

Приклад даних користувачів за результатами тестування наведено у таблиці 4.3.

Таблиця 4.3 – Приклад даних користувачів за результатом тестування

Id	E	A	C	ES	O
37	4.3	4.7	5.4	5.3	3.5

Кінець таблиці 4.3

38	4.9	5.6	4.1	4.5	4.5
39	4.3	2.3	3.5	5.5	3
40	3.8	4	3.5	3.7	5.5
41	3.7	2.9	5	5.2	4.5

Де Id – це унікальний ідентифікатор користувача, E – екстраверсія, A – доброзичливість, C – сумлінність, ES – емоційна стабільність, O – відкритість.

Реалізацію алгоритму буде виконано використанням мови програмування Python.

Python – це інтерпретована об'єктно-орієнтована мова програмування високого рівня. Python була обрана через швидкість та легкість при розробці програм. В Python представлено багато бібліотек для роботи з даними, що і робить її привабливою для швидкої розробки програм.

Спочатку імпортуємо бібліотеки Python – pandas, numpy, scipy.stats. Ці бібліотеки призначені для обробки даних і обчислень. Також імпортуємо cosine_similarity для обчислення показників подібності.

```
import pandas as pd
import numpy as np
import scipy.stats


from sklearn.metrics.pairwise import cosine_similarity
```

Рисунок 4.2 – Скріншот програмного коду

В якості набору даних з якого береться інформація по фільмам що відображаються в списку рекомендації, було обрано базу даних «MovieLens».
В цьому наборі даних представлені актуальні оцінки фільмів від користувачів. В ньому представлена інформація по рейтингам більше ніж 100 тисяч фільмів. База даних постійно доповнюється, що допомагає досить точно відобразити інформацію для користувача додатку.

Для аналізу набору даних використаємо Google Colab. Google Colaboratory або Colab, Це ще один хмарний сервіс від Google Research. Це IDE, яка дозволяє будь-якому користувачеві писати вихідний код у своєму редакторі та запускати його з браузера. Зокрема, він підтримує мову програмування Python і орієнтований на завдання машинного навчання, аналіз даних, навчальні проекти тощо. Нижче наведено скріншот коду з його використанням.

```
from google.colab import drive
drive.mount('/content/drive')

import os
os.chdir("drive/My Drive/contents/recommendation_system")
```

Рисунок 4.3 – Скріншот програмного коду

Тепер зчитаємо дані о фільмах, для того щоб отримати назви фільмів.

```
movies = pd.read_csv('ml-latest-small/movies.csv')
movies.head()
```

Рисунок 4.4. – Скріншот програмного коду

Дані з набору даних MovieLens мають ідентифікатор фільму, назву та жанри.

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

Рисунок 4.5 – Фільми з набору даних MovieLens

Далі необхідно відфільтрувати фільми та залишити для аналізу лише ті, які мають понад 100 оцінок. Це робить обчислення доступними для керування пам’яттю Google Colab. Для цього спочатку групуємо фільми за назвою,

підраховуємо кількість оцінок і зберігаємо лише фільми з оцінками понад 100. Також розраховуються середні рейтинги фільмів.

```
agg_ratings = df.groupby('title').agg(mean_rating = ('rating', 'mean'),
                                      number_of_ratings = ('rating', 'count')).reset_in

agg_ratings_GT100 = agg_ratings[agg_ratings['number_of_ratings']>100]
agg_ratings_GT100.info()
```

Рисунок 4.6 – Скріншот програмного коду

Переглянемо отриманий список найпопулярніших фільмів та їх оцінки.

	title	mean_rating	number_of_ratings
3158	Forrest Gump (1994)	4.164134	329
7593	Shawshank Redemption, The (1994)	4.429022	317
6865	Pulp Fiction (1994)	4.197068	307
7680	Silence of the Lambs, The (1991)	4.161290	279
5512	Matrix, The (1999)	4.192446	278

Рисунок 4.7 – Найпопулярніші фільми з набору даних MovieLens

Наступним кроком перетворимо набір даних у матричний формат. Рядки матриці – це користувачі, а стовпці матриці – фільми. Значення матриці – користувацький рейтинг фільму, якщо рейтинг є. В іншому випадку відображається «NaN».

```
matrix = df_GT100.pivot_table(index='userId', columns='title', values='rating')
matrix.head()
```

Рисунок 4.8 – Скріншот коду програми

title	2001: A Space Odyssey (1968)	Ace Ventura: Pet Detective (1994)	Aladdin (1992)	Alien (1979)	Aliens (1986)	Amelie (Fabuleux destin d'Amélie Poulain, Le) (2001)	American Beauty (1999)	American History X (1998)	American Pie (1999)	Apocalypse Now (1979)	Apollo 13 (1995)
userId											
1	NaN	NaN	NaN	4.0	NaN	NaN	5.0	5.0	NaN	4.0	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	4.0	NaN	NaN	NaN	5.0	NaN	NaN	NaN	NaN
5	NaN	3.0	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0

Рисунок 4.9 – Матриця елемент-користувач

Визначимо схожих користувачів, розрахувавши матрицю подібності користувача за допомогою кореляції Пірсона.

```
user_similarity = matrix_norm.T.corr()
user_similarity.head()
```

Рисунок 4.10 – Скріншот програмного коду

userId	1	2	3	4	5
userId					
1	1.000000	NaN	NaN	0.391797	0.180151
2	NaN	1.0	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN
4	0.391797	NaN	NaN	1.000000	-0.394823
5	0.180151	NaN	NaN	-0.394823	1.000000

Рисунок 4.11 – Матриця подібності користувачів

Тепер можливо сформувати рекомендації для активних користувачів, однак неможливо буде визначити рекомендації для нових користувачів.

Для того щоб найти матрицю подібних користувачів з урахуванням нових користувачів використаємо для обчислення демографічні дані та особисті риси користувачів.

```
recommendation_df = pd.DataFrame()
recommendation_df['weighted_average_recommendation_score'] = temp['sum_weighted_rating'] / temp['sum_corr']
recommendation_df['movieId'] = temp.index
recommendation_df = recommendation_df.sort_values(by='weighted_average_recommendation_score',ascending=False)
recommendation_df.head(5)
recommendation_df

movie = pd.read_csv('datasets/movie.csv')
movies_from_user_based = movie.loc[movie['movieId'].isin(recommendation_df['movieId'].head(10))]['title']
movies_from_user_based.head(5)
movies_from_user_based[:5].values

user_similarity = matrix_norm.T.corr()
user_similarity.head()
```

Рисунок 4.12 – Скріншот програмного коду

Нарешті визначимо, які фільми рекомендувати новому користувачу. Отримаємо список фільмів, які зазвичай позитивно оцінюють сусідні користувачі, щоб запропонувати їх цільовому користувачеві. Припускається, що користувачі зі схожими демографічними характеристиками та рисами особистості оцінюватимуть елементи однаково, то й рекомендації отримуватимуть подібні.

```
item_score = {}
for i in similar_user_movies.columns:
  movie_rating = similar_user_movies[i]
  total = 0
  count = 0
  for u in similar_users.index:
    if pd.isna(movie_rating[u]) == False:
      score = similar_users[u] * movie_rating[u]
      total += score
      count +=1
  item_score[i] = total / count
item_score = pd.DataFrame(item_score.items(), columns=['movie', 'movie_score'])

ranked_item_score = item_score.sort_values(by='movie_score', ascending=False)
m = 10
ranked_item_score.head(m)

avg_rating = matrix[matrix.index == picked_userid].T.mean()[picked_userid]
```

Рисунок 4.13 – Скріншот програмного коду

У результаті виконання отримуємо список фільмів що можна порекомендувати новому користувачу ґрунтуючись на його демографічних атрибутах і особистих рисах.

	movie
16	Harry Potter and the Chamber of Secrets (2002)
13	Eternal Sunshine of the Spotless Mind (2004)
6	Bourne Identity, The (2002)
29	Ocean's Eleven (2001)
18	Inception (2010)

Рисунок 4.14 – Список рекомендованих фільмів

Таким чином опрацювання логіки побудови рекомендації для нових користувачів закінчено. Тепер впровадимо розроблену рекомендаційну систему до онлайн кінотеатру. На рисунках 4.15-4.16 представлено макет сайту.

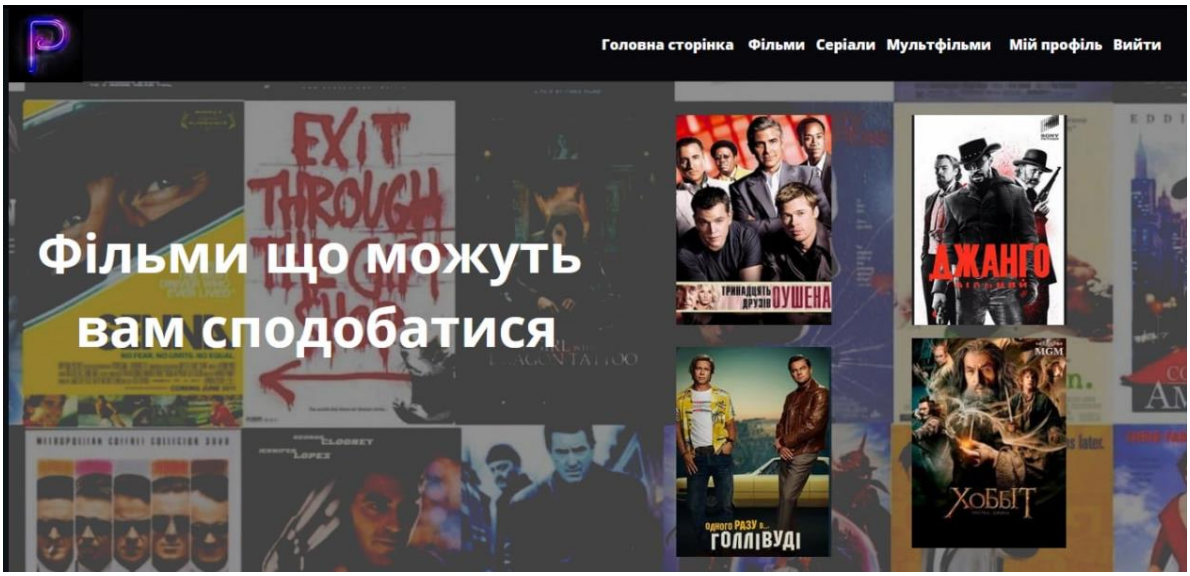

Рисунок 4.15 – Макет головної сторінка онлайн-кінотеатру з рекомендаціями для нового користувача

На головній сторінки відображаються 4 рекомендації фільмів відібрані з урахуванням демографічних даних та особистих рис користувачів. На початку користувач отримуватиме рекомендації лише за його особистими даними, однак вони можуть змінюватися, коли новий користувач почне оцінювати переглянуті фільми.

Після оновлення сторінки користувач отримує нові рекомендації.

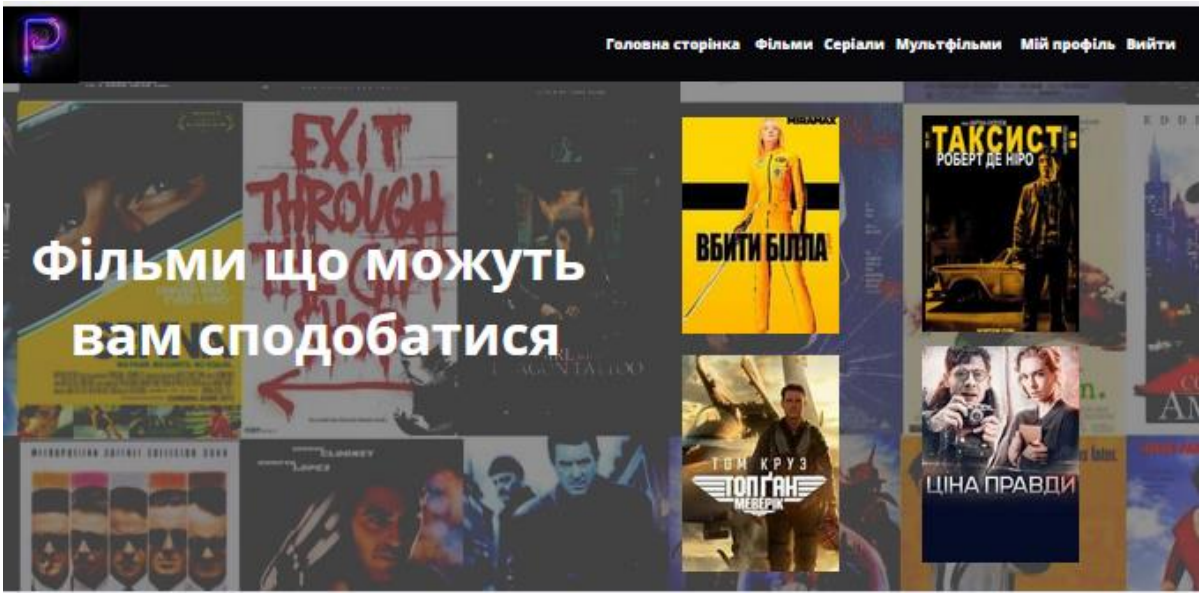

Рисунок 4.16 – Макет головної сторінка онлайн-кінотеатру з рекомендаціями для нового користувача

Подальший розвиток цієї розробки можна вести у таких напрямках як оптимізація структур даних, оптимізація продуктивності, підбір міри схожості, модифікація алгоритму фільтрації.

4.2 Експериментальна перевірка удосконаленого методу

Мета дослідження полягає в тому, щоб покращити продуктивність системи рекомендацій на основі користувачів шляхом розробки нової гібридної рекомендаційної архітектури, щоб вирішити проблему холодного старту шляхом поєднання схожості користувачів на основі їх особистості з демографічною фільтрацією.

Основні аспекти проекту можна розділити на чотири окремі етапи, які в цілому охоплюють процеси збору даних, навчання та тестування, прогнозування та рекомендації, а потім оцінювання ефективності.

Необхідно провести експериментальну оцінку запропонованого підходу, який враховує зміну інтересів користувачів, особистісні та демографічні характеристики користувача, а потім необхідно оцінити ефективність прогнозу.

Щоб оцінити створений метод, використаємо експериментальні дані з набору даних MovieLens. Цей набір даних складається з 1000 209 оцінок (1-5) від 6040 користувачів для 3900 фільмів і такі демографічні ознаки як вік, стать, країна проживання. Кожен користувач оцінив не менше 20 фільмів. Рівень розрідженості становить 1-1000209(3900*6040) = 0,958, набір даних дуже розріджений.

Випадковим чином обрали 70% даних для навчання, а решту – для тестування. Буде використано набір даних, зібраний з тестування при реєстрації, особистості користувачів для системи рекомендацій, тут користувачів просили відповісти на особистісні тести на основі моделі великої п’ятірки для створення профілю користувача. Використовуючи наведені вище два набори даних, порівняємо запропонований метод з іншими методами.

З набору даних 20% буде витягнуто випадковим чином, щоб служити тестовим набором, а решта буде використана як навчальний набір.

Якість рекомендаційної системи можна визначити за результатами оцінювання. Тип використовуваної метрики оцінювання залежить від типу КФ. Для оцінки ефективності запропонованого методу, використаємо середню абсолютну похибку (MAE). MAE метрика, яка повідомляє нам середню абсолютну різницю між прогнозованими значеннями та фактичними значеннями у наборі даних. Чим нижче MAE, краще модель відповідає набору даних.

Незважаючи на його обмеження при оцінюванні систем, зосереджених на рекомендації певної кількості елементів, простота його обчислення та його статистичні властивості зробили його одним із найпопулярніших показників при оцінюванні систем рекомендацій.

$$MAE = \frac{\sum_{e=1}|P_{x,e} - R_{x,e}|}{N}, \qquad (4.1)$$

де N – загальна кількість елементів у тестовому наборі;

$P_{x,e}$ – прогнозована оцінка користувача *x* до елемента *e*;

$R_{x,e}$ – фактична оцінка користувача *x* до елемента *e*.

Менше значення MAE означає вищу якість рекомендації або кращий прогноз або більшу точність рекомендації що надає алгоритм.

Після введення експериментальних налаштувань на основі обраного набору даних, були отримані результати оцінювання удосконаленого методу в порівнянні з іншими розглянутими методами заснованими на виявленні схожості користувачів. Було отримано такі результати (таблиця 4.4).

Таблиця 4.4 – Порівняння якості рекомендацій MAE

Кількість користувачів	S	Demg	SP	Hybrid
5	1.143531	0.956617	0.812321	0.797213
10	1.139181	0.831445	0.783252	0.751182
15	1.138663	0.820066	0.781158	0.668644
20	1.091772	0.808995	0.779451	0.659126
25	1.000312	0.801102	0. 779213	0.630811

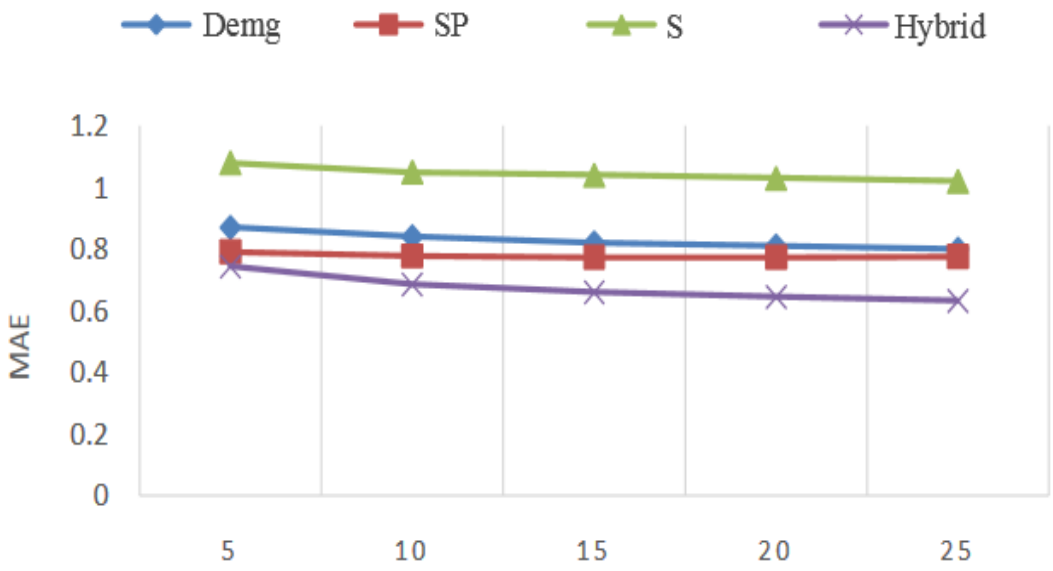

Рисунок 4.17 – Порівняння продуктивності удосконаленого методу

Де S – схожість на основі зміні інтересів користувачів, SP – схожість на основі особистих рис користувача, Demg – схожість на основі демографічних даних користувачів, Hybrid – удосконалений метод схожості користувачів в поєднанні демографічних даних з даними особистого анкетування.

Результат на рисунку 4.17 показує, що гібридна подібність може досягти найкращої якості рекомендацій на MAE. В MAE, чим менше значення, тим краще, як ми бачимо на графіку, запропонований метод (Hybrid) отримує значення MAE все менше і менше, зі зростом кількості схожих користувачів.

Суттєві відмінності у порівнянні з іншими методами виражаються в тому, що запропонований метод може зробити більш точний і якісний прогноз та надати рекомендації.

ВИСНОВКИ

Метою роботи є дослідження існуючих методів побудови рекомендацій для нових користувачів онлайн кінотеатру.

Для виконання поставленої мети в роботі було виконано:

– аналіз систем рекомендацій;

– дослідження методів побудови рекомендаційних систем;

– дослідження проблеми нових користувачів;

– розглянуто підходи для вирішення проблеми холодного старту при побудові рекомендаційних систем;

– запропоновано удосконалений метод формування рекомендацій для нових користувачів;

– надано технологію використання удосконаленого методу;

– програмна реалізація використання удосконаленого методу;

– експериментальна перевірка удосконаленого методу.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Методичні вказівки щодо розробки та оформлення кваліфікаційної роботи (для студентів усіх форм навчання другого (магістерського) рівня вищої освіти спеціальності 122 Комп'ютерні науки освітньо-професійної програми «Інформаційні управляючі системи та технології») / У поряд.: Петров К.Е., Левикін В.М., Чалий С.Ф., Євланов М.В., Саєнко В.І., Міхнов Д.К., Міхнова А.В., Чала О.В. – Харків: ХНУРЕ, 2021. – 30 с.

2. Ricci F., Rokach L., Shapira B. Recommender Systems Handbook. New York, 2015.

3. Moghaddam F., Elahi M. Cold-start solutions for recommendation systems – Bonn, Nordrhein-Westfalen, Germany. – 2019. 35 p.

4. Netflix's Recommendation Systems: Entertainment Made for You [Електронний ресурс] – режим доступу: https://illumin.usc.edu/netflixs-recommendation-systems-entertainment-made-for-you/ (дата звернення 05.11.2022).

5. Мелешко Є.В. Дослідження методів побудови рекомендаційних систем в мережі Інтернет / Є.В. Мелешко, Г.С. Семенов В.Д. Хох. // Збірник наукових праць "Системи управління, навігації та зв'язку". Випуск 1(47). – Полтава: ПНТУ ім. Ю. Кондратюка. – 2018. – С. 131-136.

6. Solving Cold User problem for Recommendation system using Multi-Armed Bandit [Електронний ресурс] – режим доступу: https://towardsdatascience.com/solving-cold-user-problem-for-recommendation-system-using-multi-armed-bandit-d36e42fe8d44 (дата звернення 10.11.2022).

7. T. Chen, W. Zhang, et. al. "SVDFeature: A Toolkit for Feature-based Collaborative Filtering", Journal of Machine Learning Research, 2012.

8. J. McAuley, R. He VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback – University of California, San Diego. – 2016. 7 p.

9. I. Cantador, P. Cremonesi Cross-Domain Recommender Systems – California. – 2014. 89 p.

10. М. Лобур, М. Шварц, Ю. Стех Моделі і методи прогнозування рекомендацій для колаборативних рекомендаційних систем. – 2018. 8с.

11. L. Safoury and A. Salah, “Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System,” Lect. Notes Softw. Eng., 2013. vol. 1, no. 3, pp. 303–307.

12. Y. Koren, R. Bell, C. Volinsky. Matrix factorization techniques for recommender systems. – 2009. 42 p.

13. M. Tavakol and U. Brefeld. Factored mdps for detecting topics of user sessions. In Proceedings of the 8th ACM Conference on Recommender Systems – 2014. 33–40p.

14. 10 найкращих алгоритмів глибокого навчання, які повинен знати кожен ентузіаст штучного інтелекту [Електронний ресурс] – режим доступу: https://ciksiti.com/uk/chapters/5902-top-10-deep-learning-algorithms-that-every-ai-enthusiast-sho (дата звернення 21.11.2022).

15. D. Asanov, “Algorithms and Methods in Recommender Systems,” Other Conf., 2011.

16. J. Basilico and T. Hofmann. Unifying collaborative and content-based filtering. In Proceedings of the 21th International Conference on Machine Learning, 2004 – P. 9-16.

17. M. Pazzani A Framework for Collaborative, Content-Based, and Demographic Filtering // Artificial Intelligence Rev. – 1999. – P. 393-408.

18. S. D. Gosling, P. J. Rentfrow, and W. B. Swann, “A very brief measure of the Big-Five personality domains,” J. Res. – 2003. 528 p.

19. J. L. Herlocker, J. A. Konstan, and J. Riedl. “Explaining collaborative filtering recommendations”, In Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, ACM Press, 2000. C. - 241-250

20. S. Tan, “Neighbor-weighted K-nearest neighbor for unbalanced text corpus,” Expert Systems with Applications. vol. 28, 2005. C. - 667-671.

21. Resnick P. GroupLens:An Open Architecture for Collaborative Filtering of Netnews / P. Resnick // Proceedings of CSCW ’94, Chapel Hill, NC, 1994. – P. 571–585.

22. Чалий С.Ф., Прібильнова І.Б. Ситуаційне представлення споживачів рекомендаційної системи. Матеріали дев’ятої міжнародної науково-технічної конференції. С.35.

23. T. Hofmann, “Latent Semantic Models for Collaborative Filtering”, ACM Transactions on Information Systems 22, 2004. C. - 89–115.

24. Recommender systems explained [Електронний ресурс] – режим доступу: https://medium.com/recombee-blog/recommender-systems-explained-d98e8221f468 (дата звернення 24.11.2022)

25. Z. Tilahun, H. Dong Jun, A. Oad Solving Cold-Start Problem by Combining Personality Traits and Demographic Attributes in a User Based Recommender System. China – 2017. – 9 p.

26. B. Chikhaoui, M. Chiazzaro, and S. Wang, “An improved hybrid recommender system by combining predictions,” Proc. - 25th IEEE Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA 2011, no. March, pp. 644–649, 2011.

27. S. Solanki, “Recommender System using Collaborative Filtering and Demographic Characteristics of Users.”

28. Chalyi S. Доповнення вхідних даних рекомендаційної системи в ситуації циклічного холодного старту з використанням темпоральних обмежень типу “next” / S. Chalyi, V. Leshchynskyi, I. Leshchynska // Системи управління, навігації та зв’язку. Збірник наукових праць. – Полтава: ПНТУ, 2019. – Т. 4 (56). – С. 105-109. – doi:https://doi.org/10.26906/SUNZ.2019.4.105.

29. Чалий С.Ф., Лещинський В.О., Лещинська І.О. Моделювання контексту в рекомендаційних системах. Науковий журнал «Проблеми інформаційних технологій», 2018, No. 1(023). С. 21-26.