МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Моделювання вибору користувача в умовах обмежень холодного старту рекомендаційної системи
(тема)

Виконав:
студент 2 курсу, групи ІПЗм-22-3

Сопун А.І.
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(код і повна назва спеціальності)

Керівник доц. Лещинськиий В.О.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри _____ З.В.Дудар
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова програма
Освітня програма Інженерія програмного забезпечення
(шифр і назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___» _____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Студентові Сопуну Артьому Івановичу
(прізвище, ім’я, по батькові)

1. Тема роботи: Моделювання вибору користувача в умовах обмежень холодного старту

2. Термін здачі студентом закінченої роботи: «31» грудня 2023р.

3. Вихідні дані до роботи: Науково-технічна література, академічні публікації та інформаційні ресурси в Інтернеті.

4. Перелік питань, що потрібно опрацювати в роботі: Аналіз рекомендаційних систем, дослідження підходів до вирішення нових проблем користувачів, формування програми досліджень, дослідження комплексного підходу до вирішення проблеми холодного старту, комплексний підхід до вирішення проблеми холодного старту, розробка гібридних методів спільної роботи з новими користувачами на основі даних опитувань, створення методик використання методів розробки, експериментальна перевірка вдосконалених методів

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналіз предметної галузі та постановка задачі	23.01 – 14.02.24	*виконано*
2	Аналіз та вибір API для дослідження	15.02 – 24.02.24	*виконано*
3	Аналіз та моделювання предметної області	17.02 – 28.02.24	*виконано*
4	Планування експериментів	25.02 – 28.02.24	*виконано*
5	Програмна реалізація кожного з обраних для дослідження API	25.02 – 01.04.24	*виконано*
6	Експериментальні дослідження	02.04 – 20.04.24	*виконано*
7	Аналіз результатів експериментальних досліджень та розробка рекомендацій	20.04 – 23.04.24	*виконано*
8	Написання та оформлення статті та тез доповіді	17.04 – 23.04.24	*виконано*
9	Підготовка пояснювальної записки	01.04 – 26.04.24	*виконано*
10	Підготовка презентації та доповіді	26.04 – 02.05.24	*виконано*
11	Нормоконтроль	02.06 – 04.06.24	*виконано*
12	Рецензування	02.05 – 04.05.23	*виконано*
13	Занесення диплома в електронний архів	05.06.2024	*виконано*
14	Попередній захист	06.06.2024	*виконано*
15	Допуск до захисту у зав. кафедри	08.06.2024	*виконано*

Дата видачі завдання 20 січня 2024р.

Студент _____ Сопун А.І.
(підпис)

Керівник роботи _____ доц. Лещинський В.О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 79 сторінок, 31 рисунок, 19 джерел. 7 таблиць.

ХОЛОДНИЙ СТАРТ, РЕКОМЕНДАЦІЙНІ СИСТЕМИ, ДЕМОГРАФІЧНА ФІЛЬТРАЦІЯ, МЕТОДИ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ, ПРОГНОЗ РЕКОМЕНДАЦІЙ, ПОКАЗНИКИ ПОДІБНОСТІ.

Об'єктом дослідження є процес генерації рекомендацій для нових користувачів.

Об'єктом дослідження є метод генерації рекомендацій при холодному старті для нових користувачів.

Мета дослідження - дослідити процес генерації рекомендацій для нових користувачів в умовах обмежень холодного старту.

В результаті дослідження запропоновано аналіз системи рекомендацій в цілому, аналіз методу генерації системи рекомендацій, дослідження проблеми нового користувача, дослідження підходів до вирішення проблеми нового користувача та вдосконалений метод генерації рекомендацій для нових користувачів в умовах обмежень холодного старту.

Представлено методику використання, програмну реалізацію та експериментальну перевірку розробленого методу.

COLD START, RECOMMENDATION SYSTEMS, DEMOGRAPHIC FILTERING, COLLABORATIVE FILTERING METHODS, RECOMMENDATION FORECAST, SIMILARITY INDICATORS.

The object of research is the process of generating recommendations for new users.

The object of the study is the method of generating recommendations during a cold start for new users.

The purpose of the study is to investigate the process of generating recommendations for new users under the constraints of a cold start.

As a result of the study, the analysis of the recommendation system as a whole, the analysis of the method of generating the recommendation system, the study of the problem of the new user, the study of approaches to solving the problem of the new user, and the improved method of generating recommendations for new users in the conditions of cold start restrictions are proposed.

The method of use, software implementation and experimental verification of the developed method are presented.

Я, *Сопун Артьом Іванович*, студент гр. *ІПЗм-22-3*, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота з дисципліни *«Теорія прогнозування»*, що буде представлена для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElAr KhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

BPR – Bayesian personalized ranking;

CF – колаборативна фільтрація;

MAB – multi-armed bandit;

MAE – mean absolute error;

РС – рекомендаційна система;

SVD – singular value decomposition;

UCB – upper confidence bound;

VBPR – visual Bayesian personalized ranking;

ВСТУП

Рекомендаційні системи у сучасному інформаційному віці відіграють важливу роль у вирішенні проблеми інформаційного перенасичення. Їхнє завдання полягає у наданні користувачам персоналізованих рекомендацій, спрямованих на задоволення їхніх індивідуальних потреб та уподобань.

Засновані на алгоритмах машинного навчання та штучного інтелекту, ці системи опираються на аналіз великого обсягу даних для здійснення прогнозів щодо вибору користувачів.

Однак, серед численних викликів, що виникають перед рекомендаційними системами, важливе місце посідає проблема холодного старту.

Це поняття відображає ситуацію, коли система має недостатньо даних про нових користувачів або новий контент, що призводить до складнощів у наданні рекомендацій. Недолік інформації про цих користувачів та новий контент ускладнює процес моделювання їхніх уподобань та робить прогнози менш точними.

Ця проблема особливо актуальна у сучасному цифровому середовищі, де темпи зростання нового контенту та залучення нових користувачів є нестримними.

Вирішення проблеми холодного старту є важливим завданням для забезпечення високої якості рекомендаційних систем та задоволення потреб користувачів.

Ця кваліфікаційна робота спрямована на аналіз та дослідження моделей, призначених для подолання проблеми холодного старту у рекомендаційних системах.

Для подолання проблеми холодного старту у моделюванні вибору користувача потрібно провести дослідження, мета якого полягає у вдосконаленні підходів, що дозволить зменшити вплив цієї проблеми на точність та

релевантність рекомендацій для користувачів, що у свою чергу сприятиме покращенню їхнього користувацького досвіду та ефективності рекомендаційних систем.

1 ПОСТАНОВКА ЗАДАЧІ

1.1 Проблематика що розглядається

Проблема холодного старту є однією з найвикликаючих і ключових проблем у рекомендаційних системах, яка виникає при появі нових користувачів або нового контенту, для яких відсутні або обмежені дані. Вона має глибокі корені в обмеженості доступу до інформації про нові об'єкти чи користувачів. Відсутність історичних даних або обмежена кількість вхідних даних суттєво ускладнює можливість системи створювати релевантні та персоналізовані рекомендації.

Недостатність інформації про нових користувачів створює проблему при побудові їхніх профілів та визначенні індивідуальних уподобань. Рекомендаційні системи, базуючись на аналізі попередніх даних, стикаються з обмеженою здатністю робити точні та персоналізовані прогнози виборів нових користувачів. Це суттєво ускладнює встановлення взаємодії із цією аудиторією, знижуючи рівень задоволеності та відповідність рекомендацій їх очікуванням.

Крім того, при появі нових об'єктів контенту без історичних даних або з обмеженою інформацією, рекомендаційні системи не мають достатнього підґрунтя для здійснення точних рекомендацій. Це призводить до менш точних та неперсоналізованих рекомендацій для нових об'єктів, що може підірвати ефективність рекомендаційної системи в цілому.

Проблема холодного старту є суттєвим викликом, оскільки вона сильно обмежує здатність систем до адаптації та реагування на новий контент і нових користувачів. Розв'язання цієї проблеми має велике значення для покращення якості рекомендацій, задоволення потреб користувачів та підвищення ефективності систем.

1.2 Технічна задача курсового проєкту

Це технічне завдання націлене на розгляд, аналіз та порівняння різних моделей та підходів, які спрямовані на подолання проблеми холодного старту у рекомендаційних системах. Мета полягає в виявленні найбільш ефективних та точних підходів для рекомендацій користувачам з обмеженою або відсутньою історією взаємодії з системою.

Це технічне завдання представляє собою детальний план дослідження та розробки підходів для вирішення проблеми холодного старту у рекомендаційних системах.

Основні Вимоги:

1) Формулювання технічної задачі: розробити чітку та структуровану технічну задачу, охоплюючи всі аспекти дослідження, визначити обсяг та мету.
2) Збір та підготовка даних: вибрати відповідні набори даних для аналізу та провести попередню обробку для використання в експериментах.
3) Акумулювання теоретичного матеріалу: огляд літератури та існуючих досліджень з проблеми холодного старту, включаючи різноманітні підходи та методи моделювання вибору користувача в рекомендаційних системах.
4) Розробка та тестування моделей:

- Розробити та імплементувати різні методи для подолання проблеми холодного старту.
- Провести експериментальне тестування для оцінки їх ефективності та точності.

5) Аналіз та порівняльне обгрунтування: порівняти результати та ефективність розроблених моделей для вибору найефективнішої моделі.

6) Документування та презентація результатів: підготувати детальний звіт, включаючи опис всіх етапів дослідження та отримані висновки.

Етапи Реалізації:

1) Збір Даних: відбір та підготовка наборів даних для подальшого аналізу та тестування алгоритмів.
2) Аналіз існуючих методів: детальний огляд різних методів рекомендацій з фокусом на їх ефективність у вирішенні проблеми холодного старту.
3) Розробка комбінованих підходів: розробка та імплементація алгоритмів, що комбінують різні типи даних для створення рекомендацій.
4) Експериментальне тестування: проведення тестів для оцінки точності та ефективності розроблених алгоритмів.
5) Оптимізація та поліпшення: вдосконалення алгоритмів на основі результатів тестів та отриманих відгуків.
6) Документація та звіт: підготовка технічного звіту, опис методів, результатів та рекомендацій.
7) Валідація та порівняння: порівняння розроблених методів з існуючими на основі метрик для підтвердження їхньої працездатності.

Описані етапи та вимоги спрямовані на створення ефективних алгоритмів рекомендацій для нових користувачів, які не мають достатньої історії взаємодії з платформою.

Проведення аналізу існуючих методів, розробка підходів, їх експериментальне тестування та подальша оптимізація є ключовими етапами в цьому процесі. Кожен етап реалізації має чітко визначені кроки, що дозволить систематично та ефективно просуватися вперед у дослідженні.

Важливо враховувати, що впровадження комбінованих підходів вимагатиме не лише технічних знань, а й ретельного аналізу отриманих результатів для створення оптимальних рекомендаційних моделей.

Це технічне завдання має на меті створення та вдосконалення рекомендаційних алгоритмів, які забезпечать ефективні та персоналізовані

рекомендації для нових користувачів, спираючись на різноманітні джерела даних та методи аналізу.

1.3 Формування технічного завдання магістерського дослідження

Моє дослідження спрямоване на розуміння та подолання проблеми моделювання користувача в умовах холодного старту у рекомендаційних системах. Я прагну розкрити основні аспекти цієї проблеми, включаючи вхідні дані, які використовуються для нових користувачів, а також фактори, які призводять до обмеженості рекомендацій при старті користувача в системі.

Буде проведено експерименти з використанням різних моделей рекомендаційних систем, враховуючи контекст холодного старту, та порівняно їхню ефективність. Крім того, робота включатиме аналіз даних та розробку метрик для об'єктивної оцінки результатів експериментів.

Розкладемо задачу на пункти що необхідно буде виконати під час саме магістерського дослідження, без урахування теоретичного матеріалу:

1) Вибір та підготовка даних:
- Вибір набору даних, що відображає холодний старт у рекомендаційних системах.
- Попередня обробка даних, включаючи очищення та підготовку для подальшого використання.
2) Розробка укспериментального середовища:
- Створення середовища для проведення експериментів з визначенням параметрів аналізу та метрик оцінки ефективності.
- Налаштування середовища для точного відслідковування результатів.
3) Використання методів та інструментів:

- Використання статистичних та аналітичних методів для моделювання холодного старту та його впливу на точність рекомендацій.
- Використання програмного забезпечення для обробки даних, моделювання та аналізу результатів.

4) Створення тестових сценаріїв:

- Розробка різних сценаріїв тестування для експериментальної перевірки моделей.
- Визначення параметрів обмежень даних та їх впливу на рекомендаційні системи.

5) Проведення укспериментів та аналіз результатів:

- Систематичне проведення тестів розроблених моделей відповідно до визначених сценаріїв.
- Проведення аналізу отриманих результатів та висновків з них щодо ефективності моделей.

6) Оцінка та висновки:

- Визначення ефективності розроблених моделей відповідно до метрик.
- Узагальнення результатів та висновки щодо працездатності моделей у контексті холодного старту.

Після створення технічного завдання магістерського дослідження, перейдемо до обирання засобів проведення даного дослідження, аналізу предметної галузі та підготовки до проведення самого магістерського дослідження.

2 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

2.1 Актуальність задачі надання рекомендацій

Системи рекомендацій дуже важливі для таких відомих компаній, як Amazon.com, YouTube, Netflix, Spotify, LinkedIn, Facebook, TripAdvisor, Last.fm, IMDb та Google.

Багато медіакомпаній зараз розробляють власні системи рекомендацій як частину користувацького досвіду.

Наприклад, Netflix пропонує високі винагороди тим, хто розробляє моделі рекомендацій.

Згідно з дослідженням McKinsey, понад 75% контенту, що переглядається на Netflix, обирається за допомогою системи рекомендацій, що полегшує споживачам пошук цікавого контенту, а компаніям - скорочення маркетингових витрат [1].

Гігант роздрібної торгівлі Amazon повідомляє, що 35% його продажів генеруються за допомогою рекомендаційних систем.

Система рекомендує товари, схожі на товари, які зазвичай купуються разом або вже є в кошику.

Ця стратегія також використовується в email-маркетингу і призводить до значного збільшення продажів.

Серія конференцій ACM, присвячених рекомендаційним системам, та щорічний конкурс RecSys Challenge [2] є важливими подіями для дослідників, на яких вони діляться новими розробками в цій галузі.

Рекомендаційні системи стали одним з найважливіших інструментів для вирішення проблеми інформаційного перевантаження, надаючи споживачам можливість автоматизованого та персоналізованого вибору продуктів [3].

2.2 Аналіз існуючих підходів щодо надання рекомендацій

Існує два типи товарної пропозиції: персоналізовані та неперсоналізовані. Приклади неперсоніфікованих рекомендацій включають рекомендації для найпопулярніших продуктів і рекомендації на основі бізнес-цілей.

Підходи до персоналізованих рекомендацій в рекомендаційних системах включають в себе контент-орієнтовані, CF та гібридні методи, що поєднують два попередні підходи.

Загальний принцип методів на основі контенту полягає в тому, щоб знайти спільні риси в продуктах, які позитивно оцінюються користувачами, і рекомендувати нові продукти з цими рисами цим користувачам. Системи, які лише рекомендують контент, часто страждають від обмеженого контент-аналізу та надмірної спеціалізації. Обмежений контент-аналіз виникає тоді, коли система має обмежену кількість інформації про користувача і продукт. Наприклад, персональні дані не можуть бути використані для аналізу через вимоги конфіденційності, або детальна інформація про продукт, така як фотографії чи музика, недоступна, коштує дорого або її важко зібрати. Інша проблема полягає в тому, що якість продукту не може бути оцінена за його змістом. З іншого боку, надмірна спеціалізація є побічним ефектом методів рекомендації нових продуктів на основі контенту, які мають вищі показники прогнозування, якщо вони схожі на продукти, яким користувач надає перевагу. Наприклад, у системі рекомендацій може рекомендуватися елемент тієї ж групи або з тим же описом, що й елементи, які користувач вже бачили. Однак вона не може рекомендувати інші елементи, які можуть зацікавити користувача.

Замість того, щоб покладатися на інформацію про контент, методи колаборативної фільтрації використовують інформацію про оцінки, надані продукту іншими користувачами системи. Ідея полягає в тому, що оцінка, яку користувач дає новому продукту, подібна до оцінок інших користувачів. CF може подолати деякі обмеження фільтрації на основі вмісту. Наприклад,

продукти, яким бракує контенту або які важко знайти, можуть бути рекомендовані на основі оцінок інших користувачів. Крім того, спільні рекомендації ґрунтуються на якості продукту за оцінкою інших користувачів, а не на контенті, який може бути не найкращим показником якості продукту. Колективна фільтрація також може рекомендувати різні продукти зі значно різним вмістом, оскільки інші користувачі виявляють інтерес до цих різних продуктів.

Підходи до колаборативної фільтрації можна розділити на два основні класи: виявлення сусідів і навчання на основі моделей. У підходах виявлення сусідів рейтинги продуктів, що зберігаються в системі, безпосередньо використовуються для прогнозування рейтингів нових продуктів. Це можна зробити двома способами: на основі рекомендацій користувачів і на основі рекомендацій продуктів. Системи на основі користувачів використовують рейтинги інших користувачів зі схожою поведінкою, відомих як "сусідні користувачі", щоб передбачити зацікавленість користувача в продукті. Сусідні користувачі, як правило, є користувачами, які найбільш тісно пов'язані з рейтингом користувача. У підході, заснованому на продуктах, користувачі оцінюють продукти на основі їхніх оцінок схожих продуктів. У цьому підході два продукти вважаються схожими, якщо кілька користувачів у системі дають їм однакову оцінку [4].

На відміну від підходу на основі виявлення сусідства, який використовує безпосередньо накопичені рейтинги для прогнозування, підхід на основі моделей використовує рейтинги для навчання моделі рекомендацій. Важливі характеристики користувача та продукту зберігаються в наборі параметрів моделі, які отримуються шляхом навчання моделі на навчальних даних і можуть бути використані для прогнозування нових рейтингів. Існують різні підходи до моделювання задачі рекомендацій. Деякі з них будуть розглянуті більш детально пізніше. Нарешті, для подолання обмежень фільтрації на основі контенту та CF використовуються гібридні методи рекомендацій, які поєднують переваги обох методів. Ці методи можна комбінувати різними способами, наприклад,

об'єднуючи окремі списки рекомендацій в один список або додаючи інформацію про контент до моделі колаборативної фільтрації. Кілька досліджень показали, що гібридні моделі можуть надавати точніші рекомендації, ніж суто контентні або колаборативні методи.

Оскільки популярність електронної комерції зростає, забезпечення того, щоб користувачі могли легко знайти те, що їм потрібно, серед широкого розмаїття продуктів, є значним викликом. Одним із рішень цієї проблеми є рекомендаційні системи, які активно розвиваються та вдосконалюються. Ці системи надають персоналізовані рекомендації щодо продуктів і послуг, які найкраще відповідають вподобанням і потребам користувача. Технологія, що використовується, базується на створенні профілів користувача і продукту та пошуку шляхів їхнього зв'язку між собою.

Рекомендаційні системи базуються на двох різних стратегіях (або їх комбінації). Підхід на основі контенту створює профілі на основі характеристик кожного користувача і продукту. Наприклад, профіль елемент включає такі атрибути, як група, відео і опис. Профілі користувачів можуть містити демографічну інформацію та відповіді на поширені запитання. Створені таким чином профілі дозволяють додаткам пов'язувати користувачів з відповідними продуктами. Однак цей підхід вимагає збору інформації, яка недоступна або яку важко зібрати.

Інша стратегія вимагає лише історії поведінки користувача без створення явного профілю. Цей підхід відомий як колаборативна фільтрація (CF): CF аналізує зв'язки між користувачами і залежності між продуктами, щоб знайти нові зв'язки між користувачами і продуктами. Наприклад, деякі системи CF включають схожі товари, які були оцінені або "вподобані". або включають користувачів зі схожою історією оцінок, які "вподобали" товар або оцінили чи "вподобали" схожі товари. або включають користувачів, які вподобали певний товар.

2.3 Аналіз методів побудови рекомендацій

Наразі існують різні підходи до побудови рекомендаційних систем.

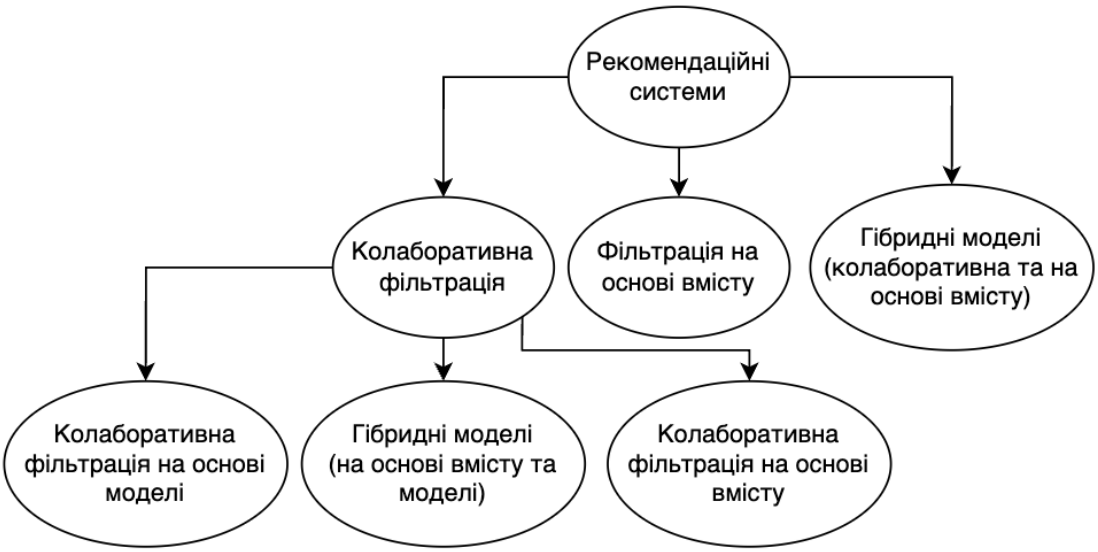

Рисунок 2.1 - Класифікація рекомендаційних систем

Ці підходи мають спільні риси, засновані на алгоритмах класифікації. Рекомендаційні системи зазвичай класифікують відповідно до того, як вони відбирають дані користувача, а для побудови РС використовують один з двох підходів - CF або фільтрацію на основі контенту. На практиці, однак, часто використовуються гібридні методи, які поєднують переваги цих підходів (рис. 2.1).

Таблиця 2.1 - Методи побудови рекомендаційних систем

Назва	Опис
1.Колаборативна фільтрація (CF)	Застосовує інформацію про оцінку, яку користувач дає об'єкту. Він базується на визначенні схожості користувача та об'єкта, що присвоюється об'єкту.

Продовження таблиці 2.1

1.1 CF на основі сусідства	Створювати рекомендації на основі схожості між елементами системи (користувачами, об'єктами), використовуючи коефіцієнти схожості між елементами системи.
1.2 CF на основі моделі	Розроблено з використанням алгоритмів машинного навчання для прогнозування користувацьких оцінок для продуктів без рейтингу.
2. Фільтрація на основі вмісту	Властивості елемента використовуються, щоб запропонувати інші елементи, схожі на вподобання користувача, на основі його минулої поведінки та явного зворотного зв'язку.
3. Гібридні методи	Поєднання різних методів і моделей система рекомендацій.

Колаборативна фільтрація (CF). Основна ідея цих підходів полягає у використанні інформації про минулу поведінку, рейтинги та інші додаткові дані від існуючих груп користувачів, щоб передбачити, які продукти, найімовірніше, сподобаються або зацікавлять існуючих користувачів системи. Цей тип систем широко застосовується в різних інформаційних системах і використовується інтернет-магазинами як інструмент для адаптації контенту до конкретних потреб клієнтів і, таким чином, для просування додаткових продуктів і збільшення продажів.

Колаборативна фільтрація використовує лише матрицю оцінок користувачів і продуктів як вхідні дані і зазвичай дає такі результати, як (а) (числову) оцінку, яка показує, наскільки поточному користувачеві подобається або не подобається певний продукт, і (б) список з n рекомендованих продуктів. Звичайно, цей перший список з n не повинен включати продукти, які вже були придбані поточним користувачем.

Цей підхід вимагає великої кількості оцінок користувачів або оцінок рекомендованих товарів для ефективної роботи, і не буде працювати для нових користувачів, нових товарів або того й іншого, оскільки в системі немає оцінок.

Формальний механізм підходу колаборативної фільтрації показано на схемі нижче (рис. 2.2).

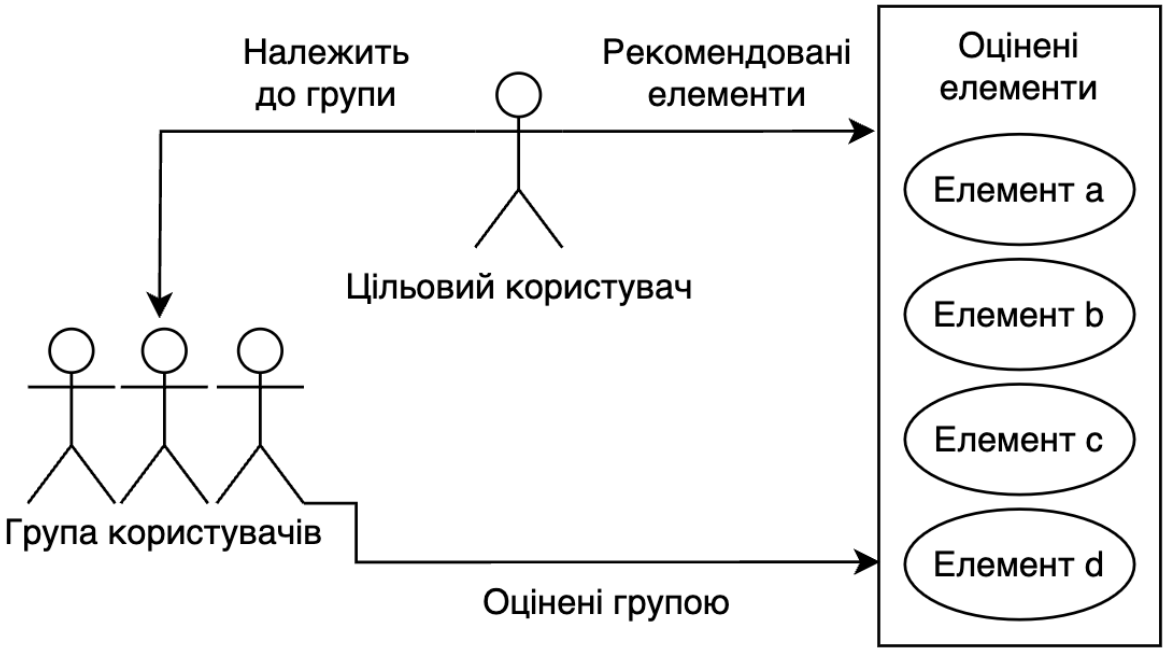

Рисунок 2.2 - Принцип роботи РС на основі CF

Існує дві основні категорії колаборативної фільтрації та гібриди обох категорій:

- Колаборативна фільтрація на основі найближчого сусіда;
- Колаборативна фільтрація на основі моделей;
- Гібрид CF на основі пам'яті та CF на основі моделей.

Колаборативна фільтрація на основі найближчого сусіда (також відома як колаборативна фільтрація на основі пам'яті) ґрунтується на встановленні зв'язку між продуктами та користувачами. Метод моделює вподобання користувача щодо продукту на основі оцінок схожих продуктів, наданих тим самим користувачем, обчислює схожість між двома користувачами або продуктами і бере середньозважене значення всіх оцінок, щоб зробити прогнози щодо

користувача. Методи на основі сусідства відрізняються від інших методів своєю простотою, ефективністю та здатністю надавати точні та персоналізовані рекомендації. Цей метод також є найдешевшим і найпростішим у реалізації, тому його часто використовують для невеликих інформаційних систем [4].

CF на основі моделей базується на використанні знань про взаємодію між користувачами та об'єктами. Цей підхід є більш комплексним і надає точніші рекомендації, оскільки допомагає виявити приховані поведінкові моделі, які пояснюють взаємодію користувача з об'єктом.

Фільтрація на основі вмісту (CBF) генерує рекомендації щодо продуктів на основі минулих уподобань. Це означає, що створюються профілі користувача і продукту. Параметри продукту повинні відповідати вподобанням користувача. Цей підхід використовує тільки дані про конкретного користувача і надає рекомендації тільки йому, а не іншим користувачам [4]. Приклад фільтрації контенту показано на рисунку 2.3.

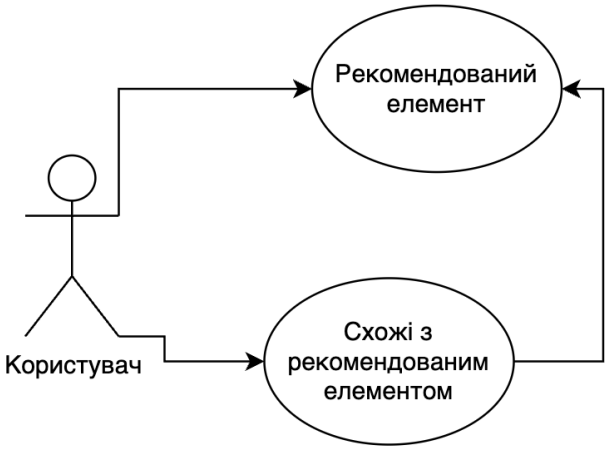

Рисунок 2.3 - Фільтрація на основі вмісту

Гібридний підхід поєднує в собі як підхід на основі моделей, так і підхід на основі околиць. На відміну від інших підходів, цей підхід широко використовується при розробці рекомендаційних систем для комерційних сайтів. Окрім покращення якості прогнозування вподобань користувачів, гібридний підхід допомагає подолати низку недоліків, таких як проблема втрати

інформації. Однак, незважаючи на свої переваги над іншими підходами, гібридні підходи є більш складними і дорогими у впровадженні та підтримці. Існує кілька основних комбінацій, які можна використовувати для створення гібридних систем:

- Реалізувати алгоритми співпраці та контенту окремо та об'єднати їхні передумови;
- Включити деякі правила координації в методологію контенту;
- Створити спільну модель з правилами обох методологій.

Ключова особливість гібридних рекомендаційних систем полягає в тому, що замість того, щоб об'єднувати прогнозовані рейтинги, вони комбінують рейтинги з різних компонентів для забезпечення репрезентативності. У багатьох випадках.

У багатьох випадках рекомендовані елементи відображаються поруч. Тому основною особливістю таких систем є поєднання презентацій, а не прогнозованих рейтингів.

Багато рекомендаційних систем є гібридними на практиці, але існує мало теоретичних робіт про те, як алгоритми можуть бути гібридизовані і за яких умов можна очікувати, що гібридні методи будуть корисними.

Прикладом створення таких методів є відкритий конкурс Netflix Prize. Конкурс об'єднав сотні студентів і дослідників, які розробили алгоритм для системи рекомендацій фільмів, що прогнозує оцінки користувачів на основі попередніх оцінок без додаткової інформації. Сотні поширених методів фільтрації та підходів були об'єднані для підвищення загальної точності.

Використання рекомендаційної системи має багато переваг, зокрема збільшення прибутку завдяки притоку нових користувачів, покращенню користувацького досвіду та підвищенню конкурентоспроможності. Потрібно знати, що РС також має недоліки. Нижче приведені переваги та недоліки самих поширениих підходів побудови систем рекомендацій.

Переваги:

- При наданні рекомендацій не покладаються на інформацію про інших користувачів;
- Проблемп "холодного старту" відсутня для нових елементів, так як схожі елементи можна легко знайти за характеристиками.
- Інтерпретація результатів генерації рекомендацій.

Недоліки:

- Нові користувачі повстають перед проблемою "холодного старту";
- Додавання нових груп елементів схожих між собою може обмежувати рекомендації для інших елементів. Користувачі можуть отримати ті ж самі рекомендації для невеликої частини від всіх елементів;
- Відсутність достатньої кількості даних про товари ускладнює групування товарів, що призводить до низької якості рекомендацій.

Перевагами системи співпраці на рівні району є те, що

- Легко впроваджувати та інтерпретувати;
- Може надавати подібні рекомендації на основі історії вподобань користувача.
- Легко надавати нові дані;
- Є масштабованою.

Перевагами систем спільної роботи на основі моделей є:

- Добре працює з розрідженими матрицями;
- Дані легко розширювати;

Крім того, всі підходи ФМ мають такі недоліки:

- Недостатня кількість даних. ця проблема притаманна системам координації на районному рівні. Це пов'язано з тим, що кількість продуктів, включених до системи, часто є занадто великою, а користувачі часто не реєструють достатню кількість магазинів у системі;
- Шахрайство. Зазвичай це відбувається, коли користувачі навмисно завищують кількість балів за свій товар і занижують кількість балів за товар конкурента;

- Відсутність різноманітності. Замість нових і цікавих продуктів рекомендуються лише одні й ті ж продукти;
- Проблеми холодного запуску. Нові користувачі та нові продукти створюють виклики для рекомендаційних систем.

При створенні РС є багато поширених проблем, таких як недостатність або відсутність вхідних даних. Проблеми рекомендаційних систем наведені в таблиці 2.2.

Таблиця 2.2 Проблеми, пов'язані з генерацією рекомендацій

Назва проблеми	Опис
Проблеми з оцінюванням	Багатокористувачів не оцінюють елементи і це призводить до нерозуміння, чи подобається продукт, чи ні.
Проблема холодного старту	Коли в систему вводяться нові користувачі або продукти, інформація про них може бути відсутньою або не повною, тому система не може надати порівняльні рекомендації для користувачів або продуктів.
Проблема розрідженості вхідних даних	Розрідженість рейтингової матриці через те, що користувачі не оцінюють продукти, призводить до помилкових рекомендацій при використанні колаборативної фільтрації.
Проблема масштабованості	Зі збільшенням кількості користувачів системи зростала і робота над рекомендаційною системою.
Проблема зміни даних з часом	Застарілі дані про товари в рекомендаційних системах призводять до формування неточних рекомендацій.

Продовження таблиці 2.2

Проблема зміни даних з часом	Застарілі дані про товари в рекомендаційних системах призводять до формування неточних рекомендацій.
Проблема зміни уподобань користувача	Деякі користувачі можуть змінювати свої вподобання з часом і отримувати рекомендації, які не є персоналізованими.
Проблема “синонімізації”	Якщо товари, що існують в системі, мають однакову назву або характеристики, система буде плутати їх між собою.
Проблема надмірної спеціалізації	Виникає, коли рекомендовані об'єкти надто схожі між собою.

Гібридні системи можуть подолати багато з цих проблем, поєднуючи переваги обох підходів CF, допомагаючи уникнути обмежень оригінального підходу, з якими стикаються підходи на основі районів, і покращуючи якість результатів рекомендацій. Незважаючи на свої ж переваги, гібридні підходи досить часто страждають за рахунок складності та чималої вартості впровадження й використання однієї чи іншої системи.

2.4 Аналіз підходів рекомендаційних систем до вирішення нових проблем користувачів

Коли створюється рекомендаційна система потрібно мати інформацію про вподобання користувачів. Холодний старт користувача або відвідувача відбувається, коли новий користувач вперше з'являється в системі. Ці

користувачі не мають історії переглядів, тому система не знає їхніх особистих уподобань і не може надавати рекомендації.

Продуктивність має значну роль у виборі алгоритма побудови рекомендаційної системи, але недостатня кількість або низька якість даних про користувачів може поставити під загрозу точність рекомендацій, заснованих на будь-якому класі рекомендаційних систем [3].

Недолік з котрим одразу може стикнутись новий користувач, відомий як проблема "холодного старту", дуже ускладнює вивчення рекомендаційною системою вподобань нових користувачів.

Головна стратегія для роботи з новоствореними користувачами – прохання вказати деякі параметри для створення початкового профілю користувача. Саме тому, в нашому випадку, тривалість процесу реєстрації користувача повинна бути встановлена як поріг поміж вхідними даними для потрібного функціонування системи рекомендацій [1].

Холодний старт в рекомендаційних системах отримав широку увагу в галузі досліджень. Доведено, що холодний старт є екстремальною стадією розрідженості даних. Для вирішення були запропоновані різні підходи.

Однією з головних проблем РС є проблема холодного старту (CSP). Ця проблема виникає при додаванні нових елементів або користувачів з порожньою історією вподобань (холодний старт користувача) або новий об'єкт, який ще не має набору оцінок або атрибутів (холодний старт об'єкта) [4].

У багатьох реальних системах CSP є повторюваною проблемою для вже відомих користувачів та об'єктів. Наприклад, коли користувач змінює інтереси. Ця проблема називається безперервною проблемою CSP (CoCoS) [4].

Проблема безперервного холодного старту користувача виникає, коли користувачі змінюють свої налаштування, рідко входять в систему або рідко оцінюють нові об'єкти [4].

Проблема безперервного холодного старту об'єкта виникає, коли є об'єкти, властивості яких змінюються з часом.

Для вирішення проблеми холодного запуску зазвичай використовують такі підходи: гібридизація, поєднання контентної та колаборативної фільтрації використовуючи контекст (наприклад, демографічну інформацію, дату і час), в якому генеруються і надаються рекомендації.

Однак ці методи не підходять для вирішення проблеми безперервної холодної ініціалізації, оскільки вони припускають, що після "розпізнавання" користувач зберігає цей стан на невизначений час і не може змінити властивості рекомендованого об'єкта. Для вирішення цієї проблеми необхідно не тільки прогнозувати вподобання користувача, але й відстежувати та прогнозувати зміни вподобань, а також враховувати можливість зміни властивостей запропонованого об'єкта [5].

Сьогодні ця проблема вирішується методами машинного навчання, які підвищують здатність системи адаптуватися до безперервних змін.

Для вирішення проблеми холодного старту використовувалися різні підходи. Тут ми зосередимося на гібридних системах колаборативної фільтрації.

Це пов'язано з тим, що гібридні системи колаборативної фільтрації є найпоширенішими серед існуючих рекомендаційних систем. Для того, щоб запропонувати можливі рішення цієї проблеми, проаналізовано різні сценарії перезапуску системи.

Методи вирішення холодного старту представлені в таблиці 2.3.

Таблиця 2.3 - Методи вирішення проблеми холодного старту

Назва методу	Характеристика
Демографічна фільтрація	Елементи рекомендацій генеруються на основі інформації про користувача; PF класифікує користувачів у відповідності до їх атрибутів і генерує рекомендації, використовуючи дані атрибутів.

Продовження таблиці 2.3

Contextual “bandit”.	Контекстна інформація може бути використана для групування користувачів за їхніми спільними характеристиками за методами кластеризації або класифікації, з припущенням, що користувачі в одному кластері поводяться однаково, а в різних кластерах дуже по-різному і дуже відрізняються.
Matrix factorization	Припустимо, що всі користувачі асоціюються з атрибутами – віком чи статтю, на основі цих даних можна визначити функцію вбудування, котра навчається на "гарячих" даних користувачів, для оцінки латентних факторів. Можна отримати обґрунтовану пораду при порівнянні атрибутів.
Візуальний байєсівський персоналізований рейтинг (VBPR)	Ідея заключається в побудові моделі, що включає візуальні характеристики персоналізованого завдання оцінки на основі неявних даних зворотного зв'язку.
Domain Recommender Systems	Веб-додатки з електронної комерції як правило працюють у кількох доменах і використовують механізми для агрегування різних типів даних з різних доменів.
Метод на основі тем.	Прогнози користувачів моделюються за допомогою імовірнісних розподілів тем, пов'язаних з різними об'єктами. Коли користувач взаємодіє з РС, його наміри прогнозуються за допомогою марковського процесу.

Демографічна фільтрація (ДФ) генерує рекомендації на основі демографічних даних користувача; ДФ класифікує користувачів на основі їхніх характеристик і використовує їхні демографічні дані для рекомендації продуктів. Якщо допустити, що рекомендації мають бути різними, тоді велика кількість веб-сайтів використовують прості та ефективні рішення для персоналізації на основі демографічних даних. Як приклад, користувачі переходять на певні веб-сайти на основі мови або країни. На відміну від систем CF та рекомендацій на основі контенту, вони прості у впровадженні та не потребують оцінки користувачів [1].

Методи багаторукого бандита (далі - MAB) (або методи контекстного бандита) використовують різні алгоритми для вирішення проблеми холодного старту. Проблему холодного старту можна переформулювати і розв'язати як проблему бандита. Що таке бандитська задача Типовим прикладом є те, що щодня до системи приєднуються нові користувачі, кожен з яких має свої власні вподобання. Так як це ми не знаємо уподобань нових користувачів, то і дати їм відповідні поради не можемо.

Існують різні алгоритми MAB, що використовуються для вирішення вищезгаданої "бандитської" проблеми, кожен з яких має свої переваги. До них відносяться Epsilon Greedy, Thompson Sampling та Upper Confidence Bound 1 (UCB-1) [5].

Epsilon Greedy найжадібніший з цих трьох алгоритмів MAB: В експериментах з використанням Epsilon Greedy константа ε (значення між 0 і 1) обирається користувачем перед початком експерименту. Коли людям призначають різні варіанти кампанії, у більшості випадків обирається випадково обраний варіант. В інших 1-ε випадках обирається варіант з найбільшим відомим виграшем; чим вище ε, тим легше запускати алгоритм.

Для кожного варіанту кампанії визначається верхній довірчий інтервал (ВДІ), який представляє найвищу оцінку можливого виграшу для цього варіанту. Алгоритм обирає лише один варіант з найвищим значенням ВДІ.

Thompson Sampling - підхід, що може забезпечити більш збалансовані результати в непередбачуваних ситуаціях. Для кожного варіанту спостережувані

результати використовуються для генерування розподілу ймовірності фактичних показників успішності (в основному бета-розподіл). Для кожного контакту вибираються можливі показники успішності.

З бета-розподілу вибирається можливий відсоток успішності, що відповідає кожному варіанту, і контакту присвоюється варіант з найвищим відсотком успішності у вибірці. Чим більше точок спостереження, тим вища довіра до істинного показника успішності, тому чим більше даних зібрано, тим ближче показник успішності вибірки до істинного показника успішності [5].

Одним з найуспішніших алгоритмів для рекомендаційних систем є матрична факторизація. На цю тему опубліковано багато наукових праць, і ця алгоритмічна парадигма широко поширена в усіх інтернет-компаніях. Найпоширенішою основою для матричної факторизації є функція SVD, яка моделює матричну факторизацію як точковий добуток елементів вектора на збагачену функцію користувача [6]. Майже всі операції матричної факторизації можуть бути змодельовані як окремі випадки функції SVD; метод MF базується на припущенні, що функція SVD працює. По-перше, існує декілька факторів, які однозначно описують послуги, що надаються інформаційною системою, а по-друге, ці фактори також використовуються для опису вподобань користувачів. У багатьох випадках система сама визначає математично обумовлені зв'язки, які не мають логічної інтерпретації; результатом МП є декомпозиція матриці користувач-об'єкт на дві окремі: матрицю об'єкта та матрицю користувача. Ступінь персоналізації можна регулювати відповідно до розмірності вихідної матриці. Розкладаючи матрицю на вектор і векторний стовпець, отримані значення відповідають найпопулярнішим продуктам.

Візуальне байєсівське персоналізоване ранжування (VBPR). Основна ідея полягає в тому, щоб побудувати модель, яка включає візуальні характеристики в персоналізовану задачу оцінки на основі даних неявного зворотного зв'язку для подолання нових проблем користувачів і поліпшення якості рекомендацій. Методологічно, глибокі нейронні мережі використовуються для вилучення візуальних характеристик продукту, а додаткові шари вбудовуються для

визначення як візуальних, так і латентних вимірів, що відповідають уподобанням користувача. Поєднання матричної факторизації, візуальних характеристик і процедур навчання Байєсівського персоналізованого ранжування (BPR) використовується для генерації рекомендацій [7].

Доменні рекомендаційні системи - є веб-сайтами з електронної комерції котрі в основному працюють з кількома доменами і використовують механізми для об'єднання різних типів даних з кількох доменів в одне ціле. Така доступність даних корисна для рекомендаційних систем, наприклад, для перехресних продажів і вирішення "холодного старту" в цільовому домені [8].

Тематичні методи використовують розподіл ймовірностей тем, пов'язаних з веб-сайтом або іншими об'єктами (наприклад, статтями), для моделювання прогнозів користувачів. Під час взаємодії користувача з PC наміри користувача прогнозуються за допомогою марковського процесу [9].

Холодного старт - складність надання рекомендацій, коли користувач є новим у системі і є основною проблемою спільної фільтрації. Традиційно ця проблема вирішується за допомогою додаткового процесу інтерв'ю для створення профілю користувача перед наданням рекомендацій.

За результатами аналізу метод демографічної фільтрації був обраний як найбільш ефективний серед оцінених методів для створення покращеного гібридного алгоритму.

2.5 Результати аналізу методів та підходів до вирішення проблеми

При аналізі існуючих підходів до створення системи рекомендацій виявив декілька проблем. Проблема холодного старту для нових користувачів виявилась найважливішою і виникає тоді, коли в системі недостатньо даних. Оскільки користувач не одразу починає оцінювання або деякий час не активний, система рекомендацій не може надати коректні рекомендації. Проблема холодного

старту нових користувачів має велике значення для моделювання вибору користувача в рекомендаційних системах. Коли нові користувачі починають користуватися рекомендаційною системою, вони можуть перестати користуватися системою і втратити клієнтів через відсутність актуальних рекомендацій.

Описані вище методи можуть бути вирішенням проблеми "холодного старту", але насамперед вона полягає в тому, що більшість з них базуються на демографічній фільтрації [10] без урахування характеристик індивідуальних користувачів.

Так як користувачі з одними й тими ж демографічними характеристиками можуть мати різні вподобання та поведінку, додаткові дані є необхідною умовою для пошуку оптимальних рекомендацій.

Для вирішення даної проблеми можна використовувати комбінацію персональних і демографічних характеристик, що дасть змогу отримати "найближчого сусіда" нового користувача і знайти такого ж користувача, навіть у випадку коли він не має історії оцінок. Найважливішим завданням рекомендаційної системи є пошук схожих користувачів за допомогою відповідної міри схожостіи

У цій статті пропонується комбінація трьох різних мір схожості. Таким чином, можна отримати найкращу міру схожості користувачів для системи рекомендацій та найточнішу міру схожості для нових користувачів.

Результати дослідження надають вдосконалений метод підвищення точності надання рекомендацій новим користувачам онлайн-кінотеатру.

Об'єктом дослідження є процес створення рекомендацій для нових користувачів рекомендаційних систем онлайн-кінотеатрів.

Мета дослідження - дослідити, як генеруються рекомендації для нових користувачів. Для розкриття цієї теми були визначені наступні етапи дослідження

- Аналіз рекомендаційних систем;
- Аналіз того, як побудовані системи рекомендацій;

- Дослідження нових проблем користувачів;
- Дослідження підходів до вирішення проблеми холодного старту при створенні рекомендаційних систем;
- Розробка вдосконалених методів створення рекомендацій для нових користувачів;
- Розробка методик використання розроблених методів;
- Експериментальна перевірка розробленої методики;

3 ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОБУДОВИ РЕКОМЕНДАЦІЙ З ВИКОРИСТАННЯМ ОСОБИСТИХ ДАНИХ

3.1 Технологія використання удосконаленого методу

На основі демографічної інформації та інформації про самого користувача, технологія, що використовується для вдосконаленого методу генерації рекомендацій, виконується наступними кроками:

1) Збір, зберігання та обробка даних;

2) Розрахунок подібності даних;

3) Висновок самої рекомендації, прогнозування, її оцінка.

Всі 3 етапи технології показані на малюнку 3.1

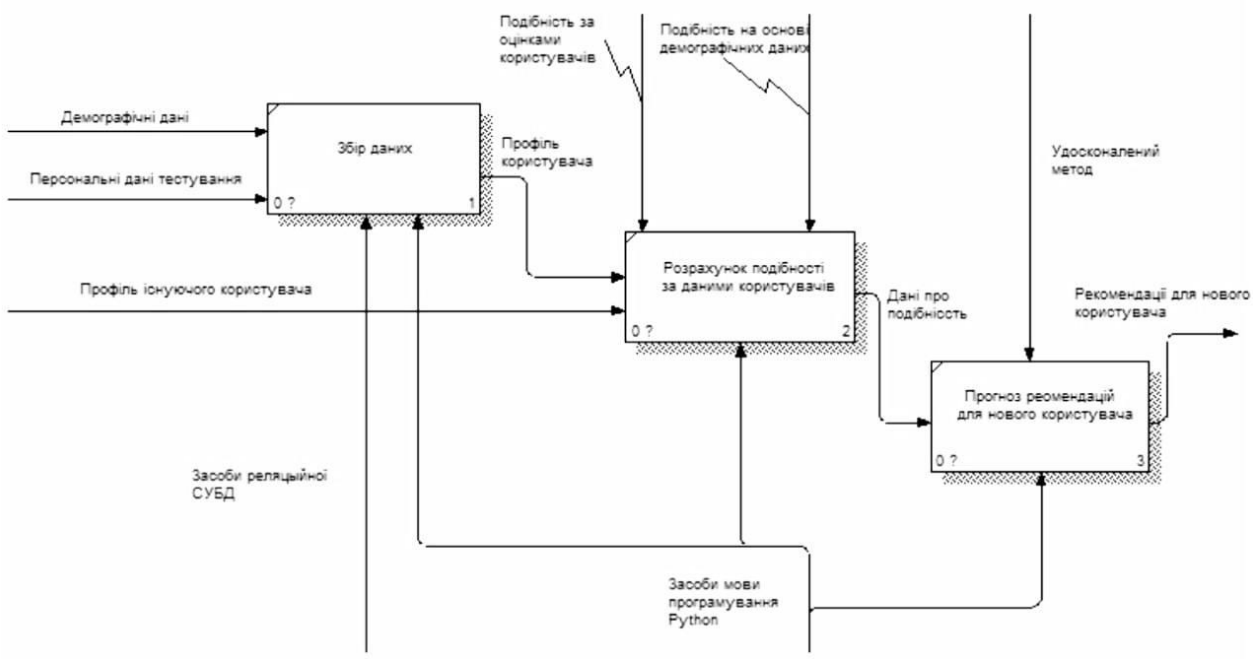

Рисунок 3.1 – Схема технології удосконаленого методу

Збір, зберігання та обробка даних є етапом для отримання та обробки інформації про користувачів з врахуванням демографічних та особистісних даних нових користувачів. Для надання якісної рекомендації під найперших взаємодій користувачів з рекомендаційною системою необхідно зареєструватися

та пройти анкетування, адже РС необхідно мати якомога більше інформації про користувача. Це дасть змогу РС отримати початкові дані для створення профілю нового користувача. Щойно створений користувач повинен чітко визначити свою демографічну інформацію і пройти особистісне тестування для оцінювання, а також окремо пройти особистісний тест для визначення особистісного бала Big5 (сумлінність, відкритість, екстраверсія, доброзичливість, емоційна стабільність) [18].

Профілем користувача є набір особистих даних кожного із користувачів, а саме: уподобання користувача, його інтереси, демографічні дані, особисті риси, історію оцінок тощо. Його використовують для побудови моделі користувача. Успішність рекомендаційної системи залежить в більшій частині від її здатності представляти поточні інтереси користувача. Точні моделі незамінні для отримання відповідних і надання точних рекомендацій із використанням будь-яких методів [2].

Метод спільної фільтрації на основі сусідства використовувався на етапі розрахунку подібності, тоді як розрахунок подібності на основі демографічних даних використовувався для визначення кількості користувачів з подібними демографічними характеристиками та подібними оцінками, для яких цільовий користувач створив сусідство.

При отриманні висновку самої рекомендації, прогнозування, її оцінки, ми отримуємо позитивно оцінені рекомендації від сусідніх користувачів і надаємо їх цільовим користувачам. Очікується, що користувачі з подібними демографічними показниками та рисами особистості оцінюватимуть варіанти рекомендацій однаково. На основі результатів отриманих прогнозів створюється ряд впорядкованих об'єктів з урахуванням інтересів цільового користувача.

4 ДОСЛІДЖЕННЯ МЕТОДІВ ВИРІШЕННЯ ПРОБЛЕМИ ХОЛОДНОГО СТАРТУ

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

Виникнення технології рекомендацій можна простежити до винаходу колаборативної фільтрації. З того часу було опубліковано багато наукових праць на різні теми. Були винайдені такі важливі прориви, як матрична факторизація [11], ранжування [12] та глибоке алгоритмічне навчання [13], а технічна точність та користувацький досвід поступово покращувалися. Апаратна підтримка алгоритмів, необхідних для побудови комп'ютерів, стає дедалі складнішою.

Майже всі соціальні сіті, сайти або додатки з медіа-контентом мають персоналізовані рекомендаційні системи, котрі допомагають користувачам вибрати конкретний елемент з десятків тисяч рекомендованих.

Основне завдання рекомендаційної системи - знайти зв'язки між елементами, які користувачі зазвичай дивляться разом. На основі отриманих даних користувачі групуються за схожими вподобаннями. З цього можна зробити висновок, що користувачі, які зазвичай переглядають одні й ті ж елементи в одному районі, можуть рекомендувати одні й ті ж елементи. Тобто це може використовуватися для прогнозування вподобань користувачів щодо нових елементів, або для налаштування продуктивності системи шляхом порівняння фактичних і прогнозованих оцінок користувачів.

Найважливішим завданням рекомендаційної системи, заснованої на користувачах, є пошук схожих сусідніх користувачів за допомогою відповідних мір схожості.

Навіть в наш час рекомендаційним системам притаманні проблеми, котрі потребують вирішення. Проблема "холодного старту" - це виклик для нових користувачів, які не мають історії переглядів або особистих уподобань, щоб увійти в систему рекомендацій.

Можна з упевненістю сказати, що CF - один з найпопулярніших алгоритмів фільтрації. Це рекомендаційна система, заснована на рейтинговій структурі, представлена у вигляді матриці рейтингів користувачів і продуктів.

CF на основі ДЗ аналізує та використовує дані користувачів. Кожен користувач відноситься до певної групи зі схожими, тому передбачається, що користувачі, котрі оцінили об'єкт однаково, будуть оцінювати інші об'єкти так само в майбутньому [14].

Значення кожної клітинки являє собою оцінку користувача для елемента рекомендації. Рейтинги оцінюються на основі мір подібності, таких як косинусна подібність, евклідова відстань [15] та коефіцієнт кореляції Пірсона.

Однак у класичному вигляді CF стикається з проблемою холодної ініціалізації. Вирішенням даної проблеми дуже часто використовуються гібридні алгоритми на основі CF з використанням атрибутивних даних користувача.

Можемо розділити методи CF на дві категорії:

- на основі сусідства;

- демографічна.

У фільтрації на основі сусідства, рейтинг елемента e для користувача x оцінюється з урахуванням поточних рейтингів інших користувачів, котрі ідентифіковані системою як ті, що мають схожі вподобання з цільовим користувачем. Підхід ґрунтується на врахуванні сусідства вибраної кількості користувачів x при оцінюванні рейтингів. Ступінь, до якої користувачі мають однакові вподобання, обчислюється за допомогою кореляції Пірсона [16] - міри подібності. Перейдемо до розгляду трьох типів обчислення подібності.

На основі оцінок двох користувачів зі схожими продуктами, алгоритми на основі близькості обчислюють кореляції між користувачами і беруть середньозважене значення всіх оцінок для створення прогнозу користувача. Можна побачити, що в цьому підході розраховується схожість між користувачами, тому потрібно використати кореляцію Пірсона (рис. 4.1).

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

Рисунок 4.1 - Формула розрахунку кореляції Пірсона.

де $R_{x,e}$ і $R_{y,e}$ - оцінки користувачів x і y для позиції e;

R_x та R_y - відповідні середні оцінки, n - кількість елементів.

Розрахунок подібності на основі змін в інтересах користувачів, схожості на основі кореляції Пірсона між користувачами X та Y показано на рисунку 4.2:

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

Рисунок 4.2 - Формула розрахунку подібності на основі кореляції Пірсона.

де $s_{(x,\,y)}$ - схожість між користувачами x та y;

$E_{sim(e,\,o)}$ - подібність між елементами e та o;

$t_{(y,\,o)}$ та $t_{(x,\,o)}$ - часові ваги.

Чим ближче час до теперішнього, тим більша ймовірність того, що відображені дані є релевантними (рис. 4.3)

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

Рисунок 4.3 - Формула обчислення ймовірності релевантності даних.

де ε - коефіцієнт, який контролює часову вагу$\varepsilon \in (0,1)$.

Чим більше ε, тим більша часова вага при обчисленні схожості. Розглянемо докладніше методи демографічної фільтрації.

Атрибути користувачів відіграють важливу роль у створенні груп користувачів для рекомендацій; РС часто стикається з проблемою "холодного старту", коли додаються нові користувачі, через недостатню кількість даних про їхні рейтинги та вподобання. У таких випадках необхідно використовувати дані, отримані з моменту, коли користувачі починають користуватися системою. Однією з таких даних є демографічні дані про користувачів. Ці дані можна використовувати для надання більш точних рекомендацій користувачам, які не мають історії. Адже, знаючи вік, стать і країну проживання користувача, легше віднести його до певної групи [16].

Методи демографічної фільтрації ґрунтуються на принципі розподілу користувачів на групи за певними характеристиками. Ці групи є вхідною інформацією для генерації рекомендацій. Цей процес необхідний для виявлення груп людей, які віддають перевагу одному й тому ж продукту. Наприклад, якщо користувач з групи G віддає перевагу продукту P, а користувач h з тієї ж групи ще не бачив продукт P, цей продукт можна порекомендувати користувачеві h.

Подібність на основі атрибутів користувача. Атрибути користувача - це дані, надані користувачем при реєстрації, зазвичай це ім'я, країна, мова,якою він розмовляє тощо. Однак ці елементи можуть змінюватися залежно від системних вимог. Цей метод враховує лише демографічні атрибути користувача. Цей метод широко використовується для пом'якшення проблем холодного старту. Схожість на основі атрибутів користувача показано на рисунку 4.4.

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

Рисунок 4.4 - Формула для обчислення схожості користувачів.

де D_{sim} - демографічна подібність;

$\cos_{sim}$ - косинусна подібність;

x_e і y_e - компоненти векторів x та y відповідно.

У свою чергу, кожен користувач x_e і y_e у системі має такі демографічні характеристики (рис. 4.5).

$$x_e \; and \; y_e = \{Id, a, g, c, l\},$$

Рисунок 4.5 - Демографічні характеристики користувачів.

де Id - унікальний ідентифікатор користувача;

a - вік користувача;

g - стать користувача;

c - країна проживання;

l - мова спілкування.

Але що відбувається, коли вподобання нових користувачів, класифікованих системою як такі, що мають схожі демографічні характеристики, не збігаються? Більшість досліджень використовують демографічну фільтрацію для вирішення проблеми "холодного старту". Якщо користувачі не мають історії оцінок, схожість між ними розраховується на основі демографічних характеристик, але коли два користувачі одного віку, статі та місцезнаходження мають різні особистісні риси, то можемо отримати дану проблему. Це пов'язано з тим, що інтереси користувачів часто відрізняються в залежності від їх особистості.

Саме через це рекомендації, засновані виключно на демографічних характеристиках, як правило призводять до неправильних рішень, таких, як рекомендувати продукти, які не відповідають вимогам користувача. Поєднуючи демографічні характеристики та дані опитування користувачів для розрахунку схожості, можна отримати більш точні результати рекомендацій, навіть якщо для обох користувачів немає історії оцінок.

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

Після проведення детального аналізу та дослідження комбінованих підходів, було вирішено об'єднати розглянуті раніше типи обчислення подібності для користувачів схожих між собою на основі даних про сусідів та на основі демографічних даних, додавши міру схожості на основі даних опитування.

Щоб вирішити проблему "холодного старту", я об'єднав дані опитування користувачів з демографічними атрибутами, щоб отримати m найближчих сусідів нових користувачів.

В результаті отримано вдосконалений гібридний підхід, який враховує дані про атрибути користувачів, дані про рейтинги користувачів та індивідуальні дані користувачів при розгляді підсистеми в рекомендаційній системі (рис. 4.6).

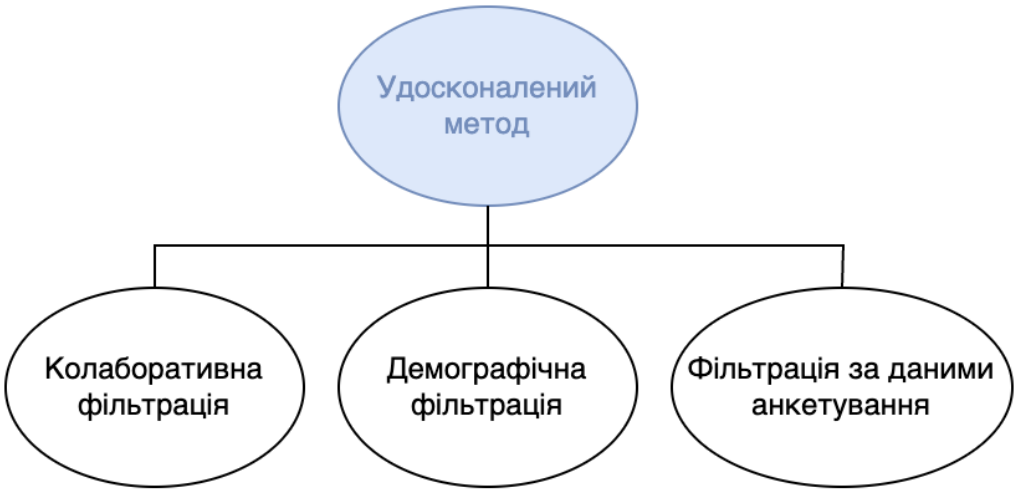

Рисунок 4.6 - Структура удосконаленого методу.

Розроблений метод об'єднує декілька методів фільтрації в один. Таким чином, поєднуючи переваги представлених методів, можна більш точно вирішувати завдання створення рекомендацій для нових користувачів.

Демографічні дані, дані опитування користувачів і розрахунок їхньої схожості може забезпечити кращі та точніші результати, навіть якщо пара користувачів не має історії оцінок, а рекомендації на основі лише демографічних характеристик можуть призвести до рекомендацій невідповідним користувачам або прийняття невірних рішень.

Але, нажаль, нічого ідеального не існує і мій просунутий метод також має недолік у тому, що при великому обсягу даних йому потрібно чимало обчислювальних витрат.

Перейдемо до кроків створення розробленого удосконаленого методу.

Крок 1 - збір даних. На цьому етапі створюється профіль користувача. Коли користувач вперше контактує з рекомендаційною системою, він повинен зареєструватися і заповнити анкету. Це надає RFS початкові дані для створення профілю нового користувача. Ці дані необхідні для подальших розрахунків.

Крок 2- розрахунок схожості. На цьому етапі демографічні дані та розрахунки схожості на основі балів використовуються для розрахунку ваг схожості для кожного користувача та кількості користувачів, пов'язаних з цільовим користувачем, що утворюють групу. Цей крок можна розбити на такі етапи:

Етап 1 - для кожного користувача x розраховується схожість з користувачем y. Це робиться за допомогою CF на основі сусідства шляхом обчислення схожості Пірсона. Цей метод обчислює кореляції на основі оцінок схожих об'єктів двома користувачами і розглядає час, коли користувачі оцінювали об'єкти, як один з факторів, що визначають схожість.

Оскільки рейтинги не є постійними величинами, оцінка, яку користувач дає елементу рекомендації, може змінюватися з часом через зміни в інтересах користувачів.

Якщо скористатися рівнянням на рисунку 4.1, то зі списку користувачів вибереться група певного розміру, яка є найближчою до користувача, про якого йде мова.

Етап 2 - використаємо проаналізовані дані про атрибути користувачів, отримані на минулому етапі. На основі демографічних характеристик (вік, стать, країна проживання тощо) розраховується вага схожості для кожного користувача. Для цього використовується рівняння зображене на рисунку 4.4.

Результатом цього етапу є кількість користувачів, які мають подібні демографічні характеристики (тобто належать до тієї ж групи), що й користувачі, які складають район.

Крок 3 - формування прогнозів та рекомендацій. Демографічні характеристики користувачів поєднуються з результатами анкетування, на які користувачі відповіли в реєстраційній формі. Цей крок також можна розбити на декілька етапів:

Етап 1 - для кожного користувача розраховується ваговий коефіцієнт на основі схожості з даними опитування (даними користувачів, які проходили особистісні тести "Великої п'ятірки" індивідуально) та використовується шкала схожості (рис. 4.7).

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

Рисунок 4.7 - Формула обчислення подібності.

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

P_x та P_y - середні особистісні характеристики у користувачів x та y.

Етап 2 - об'єднайти демографічні дані користувачів з даними опитування, наданими користувачами під час реєстраційного опитування, щоб знайти сусідніх користувачів у групі. На цьому кроці розраховується коефіцієнт подібності $Ed_{sim(x,y)}$ обчислюється як сума коефіцієнтів подібності демографічних даних та даних опитування (рис. 4.8):

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

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

Рисунок 4.8 - Формула розрахування коефіцієнту подібності.

де $Ed_{sim(x,\,y)}$ - розширена демографічна подібність користувачів x та y;

$D_{sim(x,\,y)}$ - демографічна подібність користувачів x та y;

$sp(x, y)$ - схожість на основі даних опитування користувачів.

В результаті формується група сусідніх користувачів, які більш схожі на користувача за демографічними показниками та анкетними даними.

Етап 3 - на основі прогнозів генеруються рекомендації для цільових користувачів. В алгоритмі CF на основі користувачів підмножина найближчих користувачів з тієї ж групи, що і цільовий користувач, обирається на основі їхньої схожості з цільовим користувачем, і зважений набір їхніх рейтингів використовується для генерації рекомендацій для цільового користувача. Для прогнозування рейтингів цільових користувачів використовується модифікована версія простого методу прогнозування на основі користувачів (рис. 3.9):

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

Рисунок 4.9 - Формула знаходження прогнозу користувача.

де $P_{x,e}$ - прогноз користувача x для елемента e;

$R_{y,e}$ - оцінка e поточним користувачем y;

R_x та R_y - середня оцінка існуючого користувача x та y;

b - кількість користувачів;

Ed_{sim} - розширена подібність, що враховує демографічні та особисті характеристики.

Під час цього етапу було сформульовано структурований перелік рекомендованих елементів, який враховує інтереси користувача і перелік найпопулярніших елементів подібних користувачів.

5 ПРАКТИЧНЕ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

5.1 Програмна реалізація удосконаленого методу

Програмна реалізація рекомендаційної системи буде представлена у вигляді веб-додатку, котрий буде складатися з клієнтської на серверної частини.

Серверна частина включатиме в себе два модулі:

1) Модуль для обробки інформації про користувачів з анкетування, вподобань тощо.

2) Модуль, відповідний за формування рекомендацій.

На клієнтській частині будуть створені форми для запису інформації про користувача та його вподобань, на основі чого, результатом запиту да серверної частини буде створено та відображено списки рекомендацій за вище-вказаними даними.

В даному дослідженні було використано універсальну модель, завдяки якій в майбутньому можна здійснити число модифікацій та вдосконалень для розширення та покращення функціоналу рекомендаційної системи.

Архітектура визначена наступними кроками:

- Отримання інформації про користувача та збереження його в основній пам'яті (для створення користувацької моделі).
- Вимірювання подібності на основі історії оцінок;
- Подібність користувачів поблизу;
- Подібність на основі демографічних даних та даних опитування;
- Порівняння вибору елементів для надання рекомендацій користувачам;
- Відображення підбірки рекомендацій.

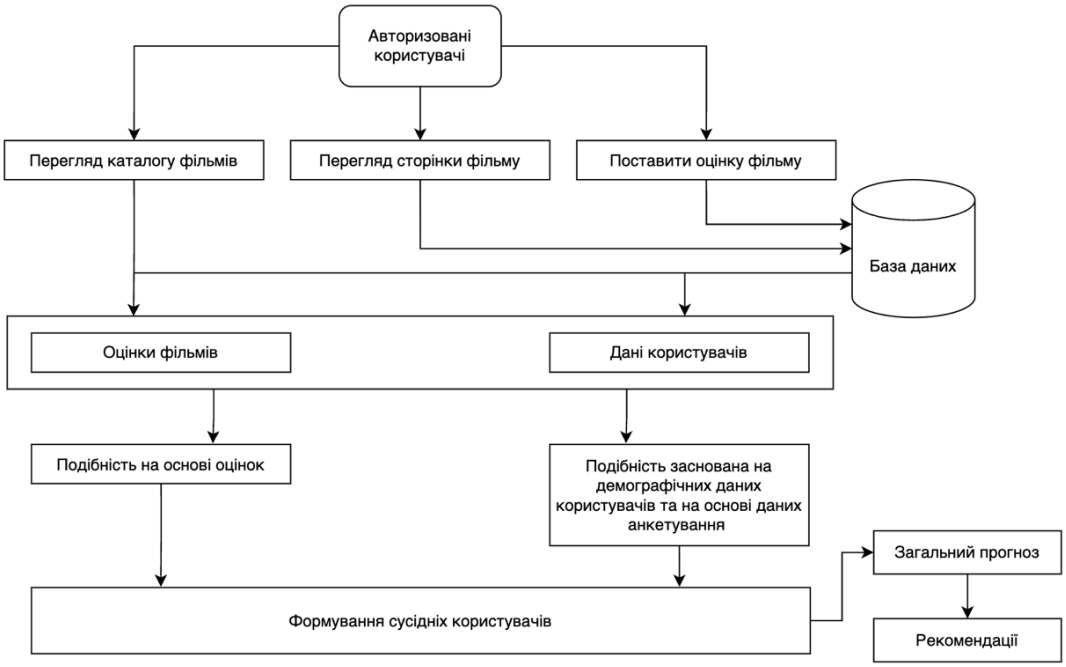

Рисунок 5.1 - Архітектура з визначеними кроками

У таблиці 5.1 наведено вхідні дані, збережені у таблиці бази даних, для створення рекомендаційної вибірки, вказані користувачами при реєстрації та проходженні анкетування/опитування.

Таблиця 5.1 – Приклад демографічних даних користувачів

ID	Name	Age	Gender	Country	Language
1	Artem	18	M	Ukraine	Ukrainian
2	Alex	39	M	Estonia	Estonian
3	Anton	22	M	France	French
4	Diana	14	F	Germany	German
5	Dasha	17	F	Spain	Spanish

Нижче приведено опис стовбців таблиці з отриманою інформацією про користувача:

- ID - унікальний ідентифікатор;
- Name - ім'я;

- Age- вік;
- Gender - стать користувача;
- Country - країна проживання користувача;
- Language - мова спілкування користувача.

Дані отримані з анкетування та процесу реєстрації зберігаються у формі ідентифікаторів особистості, створених на основі результатів тесту оцінки особистості Big5 (співчуття, екстраверсія, емоційна стабільність, совість, відкритість). В першу чергу потрібно врахувати час який користувач витрачає на дослідження, адже в іншому випадку система може потребувати затрати великого часу в процесі його проходження, і виглядати не привабливою для користувача.

Відповіді користувача, в залежності від результату оцінки обраної формули, перетворюються в бали (таблиця 5.2).

Таблиця 5.2 – Оціночна шкала

Бали	1	2	3	4	5
Значення	2	1	0	-1	-2

Кожен з п’яти основних факторів складається з основних множників-факторів. Наприклад, основним є "екстраверсія-інтроверсія", котрий складається з такої послідовності - 1.1, 1.2, 1.3, 1.4, 1.5. Кількісно основний фактор визначається шляхом підсумовування трьох балів.Дані засновані на результатах анкетування користувача можна побачити у таблиці 5.3.

Таблиця 5.3 – Дані засновані на результатах анкетування користувача

ID	A	C	E	ES	O
1	3.6	4.3	3.2	4.4	2.6
2	4.5	3.2	3.8	3.6	3.4

Кінець таблиці 5.3

3	2.2	2.6	3.2	4.6	2
4	3	2.6	2.7	2.8	4.4
5	2.6	4	2.6	4.1	3.3

Таблиця має такі стовбці:

1) ID – унікальний ідентифікатор;
2) A – доброзичливість;
3) C – сумлінність;
4) E – екстраверсія;
5) ES – емоційна стабільність;
6) O – відкритість.

Алгоритм буде реалізовано з використанням мови програмування Python, яка є інтерпретованою та об'єктно-орієнтованою мовою високого рівня. Вибір на користь Python зроблено через його швидкість та зручність у процесі розробки. Python також має безліч бібліотек для обробки даних, що сприяє швидкому створенню програм.

Спочатку підключимо необхідні бібліотеки Python, зокрема pandas, numpy та scipy.stats, які використовуються для роботи з даними та проведення обчислень. Додатково підключимо cosine_similarity для розрахунку показників схожості.

```
1   import pandas
2   import numpy
3   import scipy.stats
4
5   from sklearn.metrics.pairwise import cosine_similarity
6
```

Рисунок 5.2 – Скріншот програмного коду

Для формування списку рекомендованих фільмів використовується база даних "MovieLens". Цей ресурс містить свіжі оцінки фільмів від користувачів і охоплює рейтинги понад 100 тисяч фільмів. Завдяки постійним оновленням, база забезпечує точність та актуальність рекомендацій для користувачів.

Для аналізу даних буде використано хмарну платформу Google Colaboratory, котра є інтегрованим середовищем розробки для написання та виконання коду прямо в браузері. Цей інструмент підтримує Python і ідеально підходить для завдань машинного навчання, аналізу даних та освітніх проєктів.

На рисунку 5.3 показано приклад коду, що використовує Colab.

```
1   from google.colab import drive
2   drive.mount('/content/drive')
3
4   import os
5   os.chdir('drive/My drive/contents/recommendation_system')
6
```

Рисунок 5.3 – Програмний код

Нижче показано, яким чином зачитуються дані про фільми.

```
19    movies = pandas.read_csv('latest/movies.csv')
20    movies.head()
21
```

Рисунок 5.4. Програмний код

Отримані об’єкти фільмів з MovieLens мають унікальний ідентифікатор, власні назви 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 5.5 – Отримані об’єкти фільмів з MovieLens

Щоб оптимізувати управління пам'яттю в Google Colab, слід відфільтрувати фільми з щонайменше 100 оцінками. Для цього згрупуємо фільми за назвою, підрахуємо кількість їхніх оцінок і залишимо лише ті, що мають понад 100 оцінок. Також обчислимо середній рейтинг для кожного з цих фільмів.

```
22  agg_ratings = df.groupby('title').agg(mean_rating = ('rating', 'mean'), number_of_ratings = ('rating, mean')).reset_in
23
24  agg_ratings_GT100 = agg_ratings[agg_ratings['number_of_ratings']>100]
25  agg_ratings_GT100.info()
26
```

Рисунок 5.6 – Підрахунок та фільтрація фільмів

Далі можна побачити вибірку отриману у результаті (див. рис. 5.7).

	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 5.7 – Вибірка найпопулярніших фільмів

Далі буде перетворено набір даних у вигляд матриці, де рядки відображають користувачів, а стовпці - фільми. Якщо дані відсутні, вони будуть позначені як "NaN".

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

Рисунок 5.8 – Код перетворення даних у вигляд матриці

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 5.9 – Матриця елемент-користувач

Для визначення схожих користувачів ми виконаємо розрахунок матриці подібності користувачів, застосовуючи кореляцію Пірсона. Цей метод дозволяє виявити ступінь схожості між різними користувачами на основі їх взаємодії з вмістом або даними. Чим ближче значення кореляції Пірсона до 1, тим більша ймовірність того, що користувачі мають схожі вподобання або поведінку. Такий підхід дозволяє побудувати більш точні та змістовні рекомендації на основі спільних інтересів користувачів.

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

Рисунок 5.10 – Розрахунок матриці подібності

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 5.11 – Отримана матриця подібності

У поточному етапі ми можемо надавати рекомендації лише для активних користувачів, але новим користувачам ця можливість недоступна.

Щоб урахувати таких користувачів у процесі формування матриці подібності, ми збираємося використовувати їх демографічні дані та особисті характеристики, що покращить точність нашої моделі, оскільки ми зможемо адаптувати рекомендації під індивідуальні вподобання і потреби кожного користувача, навіть якщо вони ще не мають історії взаємодії з платформою.

```
recommendation_df = pandas.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 = pandas.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 5.12 – Розрахунок рекомендацій для користувачів без історії

У завершальній фазі ми визначимо, які фільми можна рекомендувати новому користувачеві, створивши перелік тих, що зазвичай здобувають позитивні оцінки від користувачів з подібними демографічними характеристиками та особистісними рисами. Мається на увазі, що люди з схожими профілями оцінювання схильні робити подібні висновки про фільми, тому рекомендації для них будуть аналогічні.

```
59  item_score = {}
60  for i in similar_user_movies.columns:
61      movie_rating = similar_user_movies[i]
62      total = 0
63      count = 0
64      for u in similar_users.index:
65          if pandas.isna(movie_rating[u]) == False:
66              score = similar_users[u] * movie_rating[u]
67              total += score
68              count += 1
69      item_score[i] = total / count
70  item_score = pandas.DataFrame(item_score.items(), columns=['movie', 'movie_score'])
71
72  ranked_items_score = item_score.sort_values(by='movie_score', ascending=False)
73  m = 10
74  ranked_items_score.head(m)
75
76  avg_rating = matrix[matrix.index == picked_user_id].T.mean()[picked_user_id]
77
```

Рисунок 5.13 – Розрахунок рекомендацій для нових користувачів

Після виконання ми матимемо перелік фільмів, які можна порекомендувати новому користувачу, враховуючи його демографічні атрибути та особистість.

	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 5.14 – Вибірка рекомендованих фільмів

Після успішного створення логіки створення рекомендацій для нових користувачів ми готові переходити до наступного етапу: впровадження цієї системи у онлайн кінотеатр. Цей крок є ключовим для забезпечення персоналізованого досвіду для нових користувачів та підвищення їхньої задоволеності від використання платформи. На рисунках 5.15-5.16 наведено макети сайту, які відображають, як система рекомендацій буде інтегрована в інтерфейс користувача та як користувачі зможуть зручно отримувати персоналізовані рекомендації щодо перегляду фільмів.

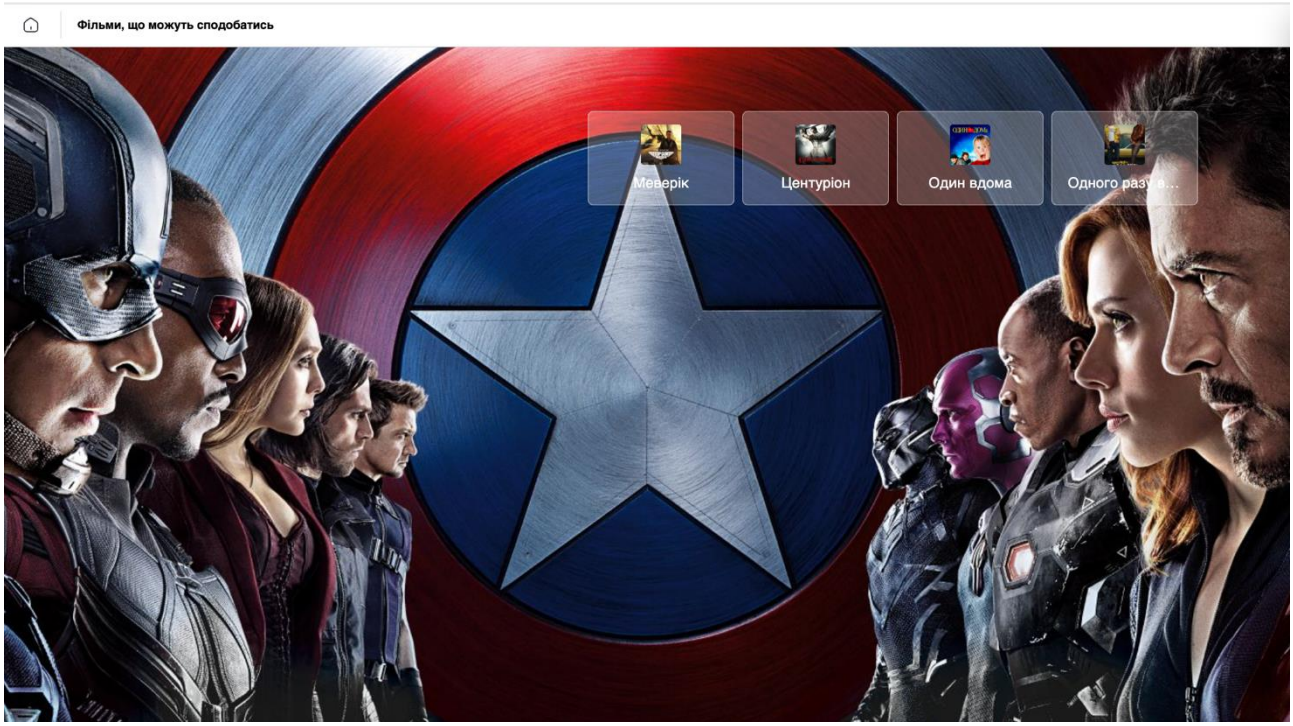

Рисунок 5.15 – Сторінка з рекомендаціями для нового користувача

На головній сторінці можна побачити набір із чотирьох рекомендацій фільмів, які були відібрані з урахуванням різноманітних факторів, включаючи демографічні дані та особисті характеристики користувачів.

Починаючи з першого відвідування сайту, рекомендації формуються лише на основі особистих даних. Проте цей список може постійно еволюціонувати: коли ви переглядаєте та оцінюєте фільми, система адаптується і надає вам нові рекомендації, які враховують ваші нові вподобання та інтереси. Після кожного

оновлення сторінки ви автоматично отримуєте свіжий набір рекомендацій, що відповідають вашим оновленим уподобанням та переглядам.

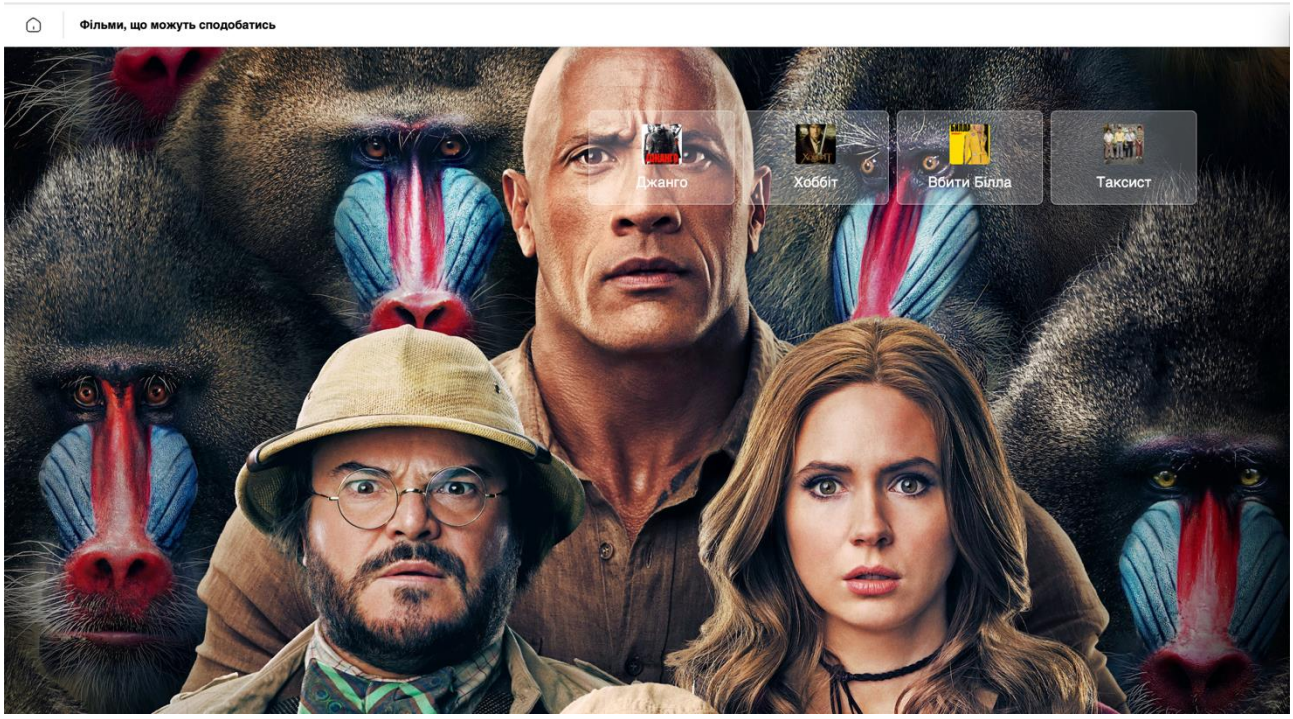

Рисунок 5.16 – Отримані рекомендації після оновлення сторінки

Подальший розвиток цієї системи можна здійснювати в декількох ключових напрямках, щоб підвищити її ефективність та точність.

По-перше, варто зосередитися на оптимізації структур даних, щоб забезпечити швидший доступ до інформації та зменшити витрати пам'яті.

По-друге, важливо працювати над покращенням продуктивності системи, зокрема шляхом впровадження більш ефективних алгоритмів обробки даних і рекомендацій.

Третім напрямком є вибір найбільш підходящих метрик для вимірювання схожості користувачів та фільмів, що дозволить підвищити релевантність рекомендацій.

Нарешті, корисним буде вдосконалення алгоритмів фільтрації, щоб враховувати різноманітні фактори та забезпечити більш персоналізований підхід до кожного користувача.

Розвиток у цих напрямах допоможе зробити систему більш адаптивною, швидкою і точною, що підвищить задоволеність користувачів та ефективність роботи платформи в цілому.

5.2 Експериментальна перевірка удосконаленого методу

Метою дослідження є підвищення продуктивності систем рекомендацій при моделюванні вибору користувача в умовах обмежень холодного старту шляхом розробки нової гібридної архітектури систем рекомендацій, що буде поєднувати в собі схожість користувачів та CF для вирішення проблеми холодного старту.

Можна розділити головні аспекти проекту на чотири етапи: збір даних, навчання і тестування, прогнозування і рекомендації, оцінка ефективності.

Запропонований підхід потребує експериментальної перевірки, щоб оцінити ефективність прогнозів на основі різних інтересів користувачів, особистих і демографічних характеристик.

Для оцінки запропонованого підходу використовуються експериментальні дані з набору даних MovieLens. Цей набір даних складається з 1000 209 оцінок (1-5), виставлених 6040 користувачами 3900 фільмам, а також демографічної інформації, такої як вік, стать і країна проживання. Кожен користувач оцінив щонайменше 20 фільмів. Набір даних є дуже розрідженим, так як її рівень становить 1-1000209(3900*6040) = 0,958.

70% даних було випадковим чином відібрано для навчання, а решту - для тестування.

Набір даних, зібраний з реєстраційного тесту, був використаний для генерації індивідуальних даних користувачів для системи рекомендацій. Тут користувачам було запропоновано відповісти на особистісний тест на основі

моделі "Великої п'ятірки" для створення профілів користувачів. Ці два набори даних були використані для порівняння запропонованого методу з іншими методами.

З цих наборів даних було випадковим чином відібрано 20% даних, які були використані як тестовий набір, а решта - як навчальний набір.

Якість рекомендаційної системи визначається за результатами оцінювання.

Тип використовуваних показників оцінки залежить від типу CF.

Для оцінки ефективності запропонованого методу використовується середня абсолютна похибка – MAE, – яка є мірою середнього абсолютного значення різниці між прогнозованими та виміряними значеннями в наборі даних; чим менша MAE, тим краще модель відповідає набору даних (рис. 5.17).

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

Рисунок 5.17 - Формула обчислення MAE.

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.

Нижче значення MAE означає, що якість рекомендацій, наданих алгоритмом, є вищою, точність прогнозування вищою, а рекомендації точнішими.

Після запуску експериментальної установки на обраному наборі даних були отримані результати оцінки покращеного методу шляхом порівняння його з іншими методами, розглянутими на основі подібності користувачів. Було отримано наступні результати (таблиця 5.4)

Таблиця 5.4 - Якість рекомендацій отримана за допомогою MAE

$C_{(user)}$	S	SP	Demg	Hybrid
5	1.132989	0.731642	0.853678	0.686302
10	1.128993	0.673141	0.743219	0.640271
15	1.127874	0.670047	0.717283	0.557533
20	1.084678	0.668341	0.710031	0.548215
25	1.000325	0.668102	0.702014	0.521703

де $C_{(user)}$ - кількість користувачів;

S - схожість на основі змін в інтересах користувача;

SP - схожість на основі особистих характеристик користувача;

Demg - схожість на основі демографічних даних користувача;

Hybrid - просунутий метод схожості користувачів, що поєднує демографічні та особисті дані опитування.

Результати приведені нижче на рисунку 5.18 показують, що метод гібридної схожості забезпечує найвищу якість рекомендацій з точки зору MAE.

Чим нижче значення MAE, тим краще, і, більше того, як показано на графіку, запропонований гібридний метод забезпечує нижчі значення MAE зі збільшенням кількості схожих користувачів.

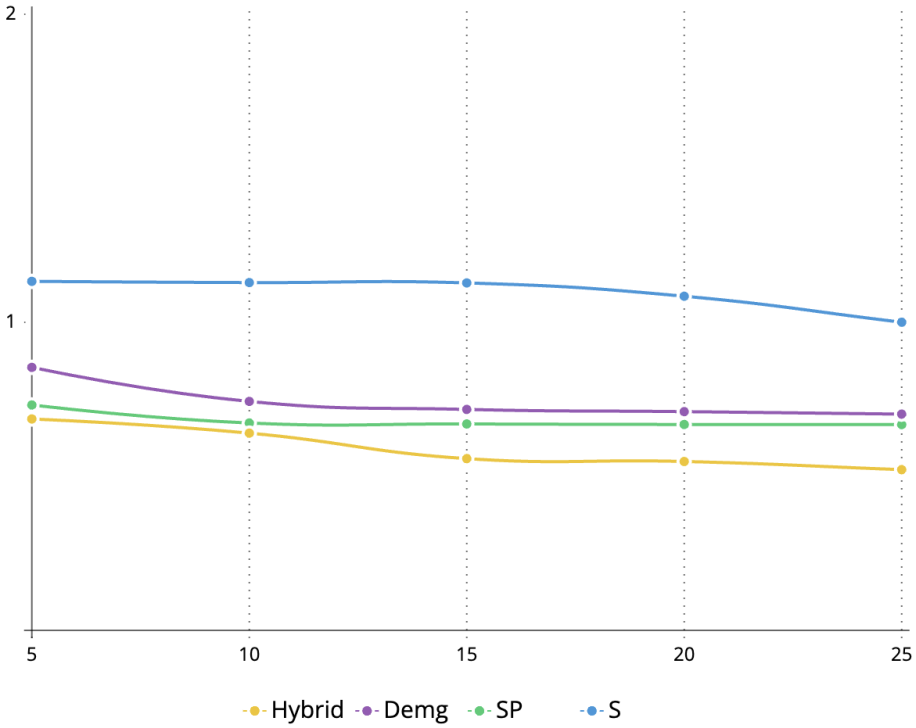

Рисунок 5.18 - Порівняння ефективності передових методів

Основна відмінність запропонованого методу від інших методів полягає в тому, що запропонований метод може надавати більш точні та якісні прогнози та рекомендації.

ВИСНОВКИ

Метою роботи є дослідження існуючих методів моделювання користувацького вибору в умовах обмежень холодного старту у рекомендаційних системах, а також створення удосконаленого методу.

Було проведено аналіз рекомендаційних систем, проаналізувано методи їх побудови, а також проблеми, з котрими стикаються користувачі при холодному старті. Детально вивчено проблему моделювання вибору користувача в умовах обмежень холодного старту рекомендаційних систем та розроблено підхід для вирішення. Удосконалено методи створення рекомендацій при холодному старті та розроблено методологію використання нового методу, завдяки проведенню експериментальної перевірки розробленого методу. Також, було розроблено програмну реалізацію.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Річчі Ф., Рокач Л., Шапіра Б. Довідник рекомендаційних систем. Нью-Йорк, 2015.

2. Рішення для холодного старту рекомендаційних систем - Бонн, Північний Рейн-Вестфалія, Німеччина - 2019. 35 с.

3. Рекомендаційна система для Netflix: Entertainment Made for You [Електронний ресурс] - Режим доступу: https://illumin.usc.edu/netflixsrecommendation-systems-entertainment-made-for-you/ (дата звернення: 05.11.2022).

4. Дослідження методів створення 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: Полтавський національний технічний університет імені Юрія Кондратюка, 2018.

5. розв'язання проблеми холодного користувача для рекомендаційних систем з використанням Multi-Armed Bandit [електронний ресурс] - режим доступу: https://towardsdatascience.com/solving-cold-user-problem-for-recommendationsystem-using-multi-armed-bandit-d36e42fe8d44 (дата звернення: 10.11.2022).

6. T. Chen, W. Zhang та ін. "SVDFeature: Інструментарій для спільної фільтрації на основі ознак", Journal of Machine Learning Research, 2012.

7. J. McAuley, R. He VBPR: Візуальне байєсівське персоналізоване ранжування на основі неявного зворотного зв'язку - Каліфорнійський університет, Сан-Дієго - 2016. 7 s. 62

8. І. Кантадор, П. Кремонезі Міждоменні рекомендаційні системи - Каліфорнія. 89 s.

9. М. Лобур, М. Шварц, Ю. Стех Моделі та методи прогнозування рекомендацій для колаборативних рекомендаційних систем - 2018. 8с.

10. Л. Сафурі та А. Салах, "Використання демографічних атрибутів користувачів для вирішення проблеми холодного старту в рекомендаційних системах". Lect. Інженерія програмного забезпечення, 2013. vol. 1, no. 1, pp. 3, no. 3.

11. ю. Корень, Р. Белл, С. Волинський. методи декомпозиції матриць для рекомендаційних систем - 2009. 42 s.

12. М. Тавакол, У. Брефельд. "Рекомендаційні" системи. Матеріали 8-ї конференції ACM з рекомендаційних систем - 2014. 33-40p.

13. 10 найкращих алгоритмів глибокого навчання, які повинен знати кожен ентузіаст ШІ [Електронний ресурс] - Режим доступу: https://ciksiti.com/uk/chapters/5902-top-10-deep-learning-algorithms-that-every-aienthusiast-sho (дата звернення: 21. 11.2022).

14. j. Basilico and T. Hofmann. інтеграція спільної фільтрації та фільтрації на основі контенту. Матеріали 21-ї Міжнародної конференції з машинного навчання, 2004 - с. 9-16.

15. Д. Асанов, "Алгоритми та методи в рекомендаційних системах". Інше, 2011.

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

17. Застосування активного навчання в ситуації циклічного холодного старту рекомендаційної системи / В.О. Лещинський, І.О. Лещинська 2018, № С. 66-70.

18. Представлення споживчого контексту в рекомендаційних системах. Праці 9-ї Міжнародної конференції з науки і техніки, С.35.

19. Доповнення вхідних даних рекомендаційної системи з часовими обмеженнями типу "next" у випадках періодичного холодного запуску / В.О. Лещинський, І.О. Лещинська // Системи управління, навігації та зв'язку. науковий журнал - Полтава: 4 (56). с. 105-109. doi: 10.1016/j.poltava.2011.09.004. - С. 105-109. doi:https://doi.org/10.26906/SUNZ.2019.4.105.s.