Алгоритми оцінки подібності об’єктів у рекомендаційних системах для медіа-контенту

Андрій Лепський, бакалавр[1] (ORCID: 0009-0001-4899-971X)
Володимир Хроленко, канд. техн. наук, доц.[1], (ORCID: 0009-0007-2157-2023)
[1] *Київський національний університет будівництва та архітектури, Київ, Україна*

АНОТАЦІЯ

У роботі розглянуто проблему обчислення подібності між об’єктами медіа-контенту як ключового етапу в інтелектуальних рекомендаційних системах. Проаналізовано контентні, колаборативні та гібридні методи, серед яких косинусна схожість, коефіцієнт Жаккара. Запропоновано комбінований підхід, що враховує як атрибутивні характеристики об’єктів, так і статистичні дані. Реалізація базується на модульній архітектурі, яка забезпечує інтеграцію нових алгоритмів і підвищує адаптивність системи до змін інтересів користувачів.

Ключові слова: рекомендаційна система, персоналізація, косинус подібності, коефіцієнт Жаккара.

1. ВСТУП

В умовах стрімкого зростання обсягів цифрового медіа-контенту зростає потреба у використанні інтелектуальних рекомендаційних систем, здатних адаптивно підлаштовуватися під інтереси користувачів. Ключовим елементом таких систем є алгоритми оцінки подібності між об’єктами, які дозволяють формувати персоналізовані пропозиції на основі аналізу характеристик контенту та історії взаємодії з ним. Саме ефективність і точність методів визначення схожості значною мірою визначають якість рекомендацій, що робить цю проблему актуальною як з наукової, так і з прикладної точки зору.

2. МЕТА РОБОТИ

Мета цієї роботи полягає в висвітленні та розгляді підходів до оцінки подібності об’єктів у рекомендаційних системах з використанням контентних, колаборативних та гібридних алгоритмів, а також створення комбінованого методу, що враховує як атрибутивні характеристики об’єктів, так і статистичну інформацію про їх використання.

3. СИСТЕМИ РЕКОМЕНДАЦІЙ

Сучасні інтелектуальні системи рекомендацій активно застосовуються для персоналізації взаємодії користувачів із медіа-контентом. Одним із ключових етапів у процесі формування рекомендацій є обчислення ступеня подібності між об’єктами, що представляють собою фільми, книги чи серіали. Якість оцінки подібності безпосередньо впливає на точність та корисність запропонованих рекомендацій. Проблематика ускладнюється багатовимірністю опису об’єктів медіа-контенту, оскільки вони можуть характеризуватися жанрами, акторами, ключовими словами, роком випуску, тривалістю та іншими параметрами.

Рекомендаційні системи поділяються на кілька основних видів залежно від принципу роботи. Контентно-орієнтовані системи формують рекомендації на основі аналізу характеристик об’єктів, з якими вже взаємодіяв користувач. Наприклад, якщо користувач переглядав фільми певного жанру або з конкретним актором, система підбирає схожі об’єкти за цими ознаками.

Колаборативні системи ґрунтуються на досвіді та вподобаннях інших користувачів. Вони знаходять схожих користувачів за історією переглядів або оцінками та рекомендують об’єкти, з якими знайомі ці "сусіди". Такий підхід дозволяє виявляти нові інтереси, але страждає від проблеми "холодного старту".

Демографічні системи формують рекомендації на основі соціально-демографічних характеристик користувача, таких як вік, стать, освіта, місце проживання чи професія. Перевагою цього підходу є простота реалізації та відсутність потреби в історії взаємодій користувача з контентом. Однак недоліком є поверхневий характер рекомендацій, оскільки вони часто узагальнені та не враховують індивідуальні інтереси конкретної людини.

Контекстно-орієнтовані системи враховують додаткові фактори, що характеризують ситуацію, у якій користувач взаємодіє з контентом. До таких факторів можуть належати час доби, геолокація, пристрій, сезонність або навіть настрій користувача. Основною перевагою цього підходу є підвищена релевантність рекомендацій у конкретних ситуаціях, проте реалізація вимагає збирання та обробки великого обсягу контекстних даних, що може створювати виклики у сфері конфіденційності та обчислювальних ресурсів.

Гібридні системи поєднують кілька підходів, частіше за все колаборативний і контентно-орієнтовний, комбінуючи інформацію про характеристики об’єктів та спільні вподобання користувачів. Це дозволяє підвищити точність і зменшити обмеження кожного з методів окремо. Гібридні системи найчастіше застосовуються у сучасних програмних рішеннях для медіа-контенту, оскільки вони забезпечують більш комплексну персоналізацію.

4. АЛГОРИТМИ ОЦІНКИ ПОДІБНОСТІ ОБ’ЄКТІВ

Одним із ключових завдань у побудові інтелектуальних систем рекомендацій є визначення ступеня схожості між об’єктами. Традиційно для цього застосовуються різні метрики подібності, що дозволяють формалізувати поняття «схожість» у кількісній формі. Серед найбільш поширених методів можна виділити косинусну міру, яка добре працює при представленні об’єктів медіа-контенту у вигляді векторів ознак, наприклад, жанрових міток чи тегів. Її перевагою є стійкість до різної довжини векторів, однак недоліком виступає обмежене урахування вагових характеристик атрибутів.

Іншим методом є коефіцієнт Жаккара, який оцінює схожість на основі перетину множин. Він ефективний для категоріальних даних, таких як акторський склад чи ключові слова, але не враховує частоту появи ознак. Евклідова

відстань, у свою чергу, є придатною для числових атрибутів, наприклад рейтингу чи року випуску, проте її застосування ускладнюється різною шкалою вимірювання ознак, що вимагає попередньої нормалізації. Огляд існуючих підходів показав, що використання лише однієї метрики не дає змоги отримати високу якість рекомендацій у багатовимірному середовищі медіа-контенту.

5. Можливі реалізації в персональних системах

5.1. Використання персональної бази знань для формування ознак медіа-контенту

У процесі дослідження особлива увага була приділена побудові персональної бази знань, яка має зберігати структуровану інформацію про медіа-контент і користувацькі вподобання. Кожен об'єкт може бути описаний через набір атрибутів: жанр, режисер, актори, студія, ключові слова, рік випуску та інші характеристики. Додатково до бази знань планується включати дані про користувацькі рейтинги й історію переглядів, що дозволить сформувати індивідуальний профіль користувача.

Передбачається, що така база знань забезпечить можливість гнучкого розширення: у разі додавання нових атрибутів система зможе враховувати їх у процесі обчислення подібності. Окремий інтерес становить застосування онтологічних моделей, де між об'єктами та їх характеристиками встановлюються семантичні зв'язки. Це створює передумови для того, щоб рекомендаційна система у майбутньому могла враховувати не лише формальні параметри, а й приховані семантичні відношення між медіа-об'єктами.

5.2. Реалізація комбінованого алгоритму оцінки подібності

У ході дослідження розглядається можливість використання комбінованого підходу до оцінки подібності, що поєднує контентні та статистичні характеристики об'єктів. Основна ідея полягає у застосуванні вагових коефіцієнтів для різних категорій ознак. Наприклад, для жанрів і тегів може використовуватися косинусна міра з урахуванням ваги кожної категорії. Вага при цьому визначається залежністю від оцінок, які залишив користувач. Для категоріальних даних, таких як актори чи студія, доцільно застосовувати коефіцієнт Жаккара, тоді як числові атрибути (рік випуску, тривалість, рейтинг) обробляються за допомогою нормалізованої евклідової відстані.

З технічної точки зору доцільним є створення модульної архітектури, у якій кожен окремий модуль відповідає за обчислення подібності в межах певної групи ознак. Підсумкова оцінка формується шляхом агрегування результатів усіх модулів із застосуванням вагових коефіцієнтів. Такий підхід дозволить гнучко поєднувати різні алгоритми та забезпечує можливість розширення: у майбутньому до структури можна буде додавати нові методи оцінки без потреби змінювати основну логіку.

Окреме місце в архітектурі займає індивідуальний профіль користувача. Він формується на основі історії взаємодії з медіа-контентом: переглядів, оцінок, часу використання, вибору улюблених категорій чи тегів. Профіль зберігає не лише явні вподобання (наприклад, жанри, яким користувач ставить найвищі оцінки), але й приховані закономірності, що виявляються під час аналізу частоти та послідовності переглядів. Таким чином, профіль виступає динамічною моделлю користувача, яка постійно уточнюється у процесі накопичення нових даних.

Використання індивідуального профілю у процесі рекомендації передбачає адаптацію загальних алгоритмів під конкретні потреби користувача. При розрахунку подібності між об'єктами враховуються ті атрибути, які мають найбільшу значущість для конкретного користувача, що дозволяє формувати більш релевантні рекомендації. Наприклад, якщо користувач частіше обирає фільми з певним режисером чи студією, ці параметри отримують підвищену вагу у функції схожості. У результаті рекомендаційна система здатна краще враховувати індивідуальний контекст і надавати більш персоналізовані результати.

6. Висновки

Отже, було проаналізовано існуючі підходи до оцінки подібності медіа-об'єктів та окреслено можливості використання персональної бази знань для їх структурованого опису. Розглянуто концепцію комбінованого алгоритму, який поєднує контентні та статистичні характеристики з урахуванням вагових коефіцієнтів, що дозволяє підвищити точність рекомендацій. Запропоновані підходи створюють підґрунтя для подальшої розробки інтелектуальної системи рекомендацій медіа-контенту, здатної адаптуватися до індивідуальних інтересів користувачів та забезпечувати більш релевантні результати.

Список літератури

[1] Machine Learning: Cosine Similarity for Vector Space Models (Part III). URL: https://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii (дата звернення 17.09.2025).

[2] Introduction to Data Mining by Tan, Steinbach, Kumar. URL: https://www-users.cse.umn.edu/~kumar001/dmbook/dmslides/chap2_data.pdf (дата звернення 17.09.2025).

[3] Matrix Factorization Techniques for Recommender Systems. URL: https://web.archive.org/web/20150226085949/http://research.yahoo.com/files/ieeecomputer.pdf (дата звернення 17.09.2025).