

АНАЛІЗ ТА ПРОЄКТУВАННЯ ПЕРСОНАЛІЗОВАНИХ РЕКОМЕНДАЦІЙ ДЛЯ КНИЖКОВОГО ІНТЕРНЕТ-МАГАЗИНУ НА БАЗІ МАШИННОГО НАВЧАННЯ

Пискір С.В.¹; Гаврилова К.М.²

^{1,2} Національний університет «Одеська політехніка»

^{1,2} Україна, Одеса

¹ 6868970@stud.op.edu.ua; ² havrylova.8864944@stud.op.edu.ua

Анотація. У цій роботі представлено аналіз та проєктування рекомендаційної системи для книжкового інтернет-магазину, яка базується на методах машинного навчання. Основна мета системи – підвищити задоволеність користувачів, пропонуючи персоналізовані рекомендації, що відповідають їхнім уподобанням.

Ключові слова: гібридна система; машинне навчання; колаборативна фільтрація; система рекомендацій; книжковий інтернет-магазин.

Вступна частина. Завдяки інтернету, книжкові магазини здобувають змогу пропонувати ширший асортимент та забезпечувати доступ до своїх товарів цілодобово, відкриваючи перед собою величезний ринок. Окрім того, використання персоналізації та інтерактивності дозволяє підвищити задоволеність клієнтів, створюючи для них унікальний досвід покупки. Однак, разом з перевагами, постає і виклик: як зробити так, аби покупці легко знаходили книги, які їх цікавлять, серед тисяч доступних назв?

Саме тут і виникає проблема – забезпечення зручного пошуку та вибору книг з-поміж широкого асортименту. Ця проблема особливо актуальна у випадку онлайн-магазинів, де кількість назв може досягати десятків тисяч, а тематика книг варіюється від популярної науки до спеціалізованої літератури. Актуальність теми зумовлена популярністю онлайн книжкового шопінгу.

Ефективним рішенням цієї проблеми стає впровадження рекомендаційної системи, яка аналізує інтереси та вподобання користувачів, щоб пропонувати їм книги, найбільш відповідні їхнім перевагам. Така система не тільки полегшує пошук потрібної літератури серед величезного асортименту, але й допомагає відкрити для покупців нові книги, які вони могли б проігнорувати. Використання алгоритмів машинного навчання та штучного інтелекту в рекомендаційних системах дозволяє адаптуватися до змінних вподобань користувачів, забезпечуючи високу точність рекомендацій.

Метою роботи є аналіз існуючих методів та проєктування системи рекомендацій для книжкового інтернет-магазину.

Основна частина. У сучасних системах персоналізованих рекомендацій для інтернет-магазинів книг використовуються різні методи. Два основні підходи — це контентно-орієнтовані (content-based) та колаборативні (collaborative filtering) методи.

Контентно-орієнтована фільтрація заснована на аналізі атрибутів або характеристик книг, таких як жанр, автор, опис, ключові слова та рецензії [1]. Цей метод дозволяє системі створювати профіль користувача, який відображає його інтереси на основі того, які книги він уже переглядав, купував чи оцінював. Відповідно до цього профілю, система рекомендує книги, схожі на ті, які вже були обрані користувачем.

Ключовим компонентом контентно-орієнтованої фільтрації є векторизація текстових даних, яка дозволяє перетворити неструктуровані тексти (опис книг, анотації) у числову

форму, зрозумілу для алгоритмів. Один із найпопулярніших методів для цього — це TF-IDF (Term Frequency-Inverse Document Frequency).

TF-IDF використовується для оцінки важливості окремих слів у документі, зважаючи на те, як часто це слово зустрічається в конкретному тексті та в усіх інших текстах бази даних. Формула TF-IDF для кожного слова виглядає так:

$$TFIDF(t, d) = TF(t, d) \times IDF(t), \quad (1)$$

де $TF(t, d)$ – частота слова t у документі d .

Складова IDF обчислюється за такою формулою :

$$IDF(t) = \log \frac{N}{df(t)}, \quad (2)$$

де N – загальна кількість документів у корпусі;

$df(t)$ – кількість документів, у яких зустрічається слово t .

TF-IDF дозволяє визначати слова, які є важливими в контексті конкретного опису книги, і знижувати вагу загальних слів, таких як "книга", "автор", тощо. Отримані вектори, що представляють кожен опис книги, використовуються для порівняння книг між собою.

Для вимірювання подібності між книгами за допомогою цих векторів, як один з популярних методів, застосовується метод косинусної подібності (Cosine Similarity). Косинусна подібність визначає кут між двома векторами, які представляють книги, і вимірює, наскільки ці вектори "схожі" [2]. Формула для розрахунку косинусної подібності виглядає так:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}, \quad (3)$$

де A, B – вектори, подібність між якими визначається;

$\|A\| \|B\|$ – евклідові норми (довжини) векторів A та B відповідно.

Косинусна подібність дає значення в діапазоні від -1 до 1, де 1 означає ідеальну схожість, а 0 – повну відсутність схожості.

На відміну від контентно-орієнтованої фільтрації, колаборативна фільтрація базується на аналізі поведінкових даних користувачів. Вона передбачає, що користувачі, які в минулому купували або оцінювали схожі товари, ймовірно, матимуть схожі інтереси в майбутньому [3].

Колаборативна фільтрація працює за двома основними напрямками:

– User-Based Collaborative Filtering – система шукає користувачів з подібними вподобаннями та рекомендує книги на основі того, що інші користувачі зі схожими інтересами купували або оцінювали високо;

– Item-Based Collaborative Filtering – система аналізує схожість між товарами на основі їхніх оцінок користувачами. Якщо кілька користувачів оцінили певні книги подібним чином, ці книги рекомендуються новим користувачам з подібними вподобаннями.

Одним із популярних методів, що використовується для обчислення схожості між користувачами або предметами, є кореляція Пірсона (Pearson Correlation Coefficient) [4]. Вона визначає, наскільки оцінки двох користувачів схожі:

$$\text{Pearson Correlation}(u, v) = \frac{\sum_{i \in I} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I} (R_{u,i} - \bar{R}_u)^2 \sum_{i \in I} (R_{v,i} - \bar{R}_v)^2}}, \quad (4)$$

де u та v – два користувачі;

$R_{u,i}$ та $R_{v,i}$ – оцінки, які дали користувачі u та v для предмета i ;

$\bar{R}_u$ та $\bar{R}_v$ – середні оцінки користувачів u та v відповідно;

I – набір книг, які оцінили обидва користувачі.

Кореляція Пірсона дає значення в діапазоні від -1 до 1. Значення, близьке до 1, означає високу подібність в оцінках користувачів, що дозволяє робити точніші рекомендації.

Хоча контентно-орієнтована та колаборативна фільтрація мають значні переваги, кожен із цих методів стикається з певними проблемами:

- Контентно-орієнтовані системи обмежуються лише атрибутами книг, тому якщо для певної книги не вистачає даних або опису, система не зможе надати точну рекомендацію. Крім того, такі системи часто рекомендують дуже схожі книги, що може звужити коло інтересів користувачів;

- Колаборативна фільтрація стикається з проблемою "холодного старту", коли нові користувачі або нові книги не мають достатньо даних для точних рекомендацій. Крім того, цей метод залежить від наявності великої кількості оцінок, що не завжди можливо в нових або нішевих платформах.

Для подолання цих обмежень часто використовують гібридну рекомендаційну систему (Hybrid Recommender System). Гібридні системи поєднують переваги контентної та колаборативної фільтрації, дозволяючи компенсувати недоліки кожного методу [5]. Така система може використовувати контентні дані для нових користувачів, які ще не мають історії взаємодії з платформою, а також аналізувати поведінкові дані для користувачів із наявною історією покупок або переглядів.

Для ефективного впровадження гібридної рекомендаційної системи в книжковий інтернет-магазин необхідно врахувати кілька етапів проектування, включаючи вибір та підготовку даних, оцінку точності моделі та інтеграцію.

Для якісного навчання моделі потрібні різні типи даних.

Контентні дані (характеристики книг):

- Назва книги, автор, жанр, ключові слова, опис;
- Метадані книг для формування контентного профілю кожної книги.

Користувацькі дані (взаємодія з системою):

- Рейтинги книг користувачами;
- Дані про покупки книг, додавання до кошика або у список бажань;
- Історія пошуку, переглядів і взаємодій з платформою.

Логи поведінки користувачів:

- Дані про кліки, переходи на сторінки книг, час перебування на сторінці.

Для оцінки ефективності рекомендаційної системи необхідно використовувати різні метрики точності та інструменти тестування:

Метрики точності:

- Precision (Точність) – визначає частку коректно рекомендованих книг з-поміж усіх рекомендацій;

- Recall (Повнота) – оцінює частку правильних рекомендацій з-поміж усіх релевантних книг;

- F1-міра – гармонійне середнє між точністю і повнотою для досягнення збалансованої оцінки;

- MAP (Mean Average Precision) – середня точність для кількох запитів, що оцінює загальну ефективність рекомендацій.

Тестування:

- Розбиття на тренувальну і тестову вибірки – частина даних (наприклад, 80%) використовується для тренування моделі, а інші (20%) – для тестування;

- Крос-валідація – розбиття на кілька підмножин для тренування і тестування моделі з метою перевірки стійкості результатів.

Для інтеграції гібридної рекомендаційної системи в книжковий інтернет-магазин необхідно створити чітку архітектуру, яка забезпечить зручний доступ до персоналізованих рекомендацій та їх безперебійну роботу. Основний елемент інтеграції — це побудова модульної системи, де рекомендаційний механізм буде працювати як окремий сервіс, який взаємодіє з іншими компонентами платформи через API.

Архітектура системи повинна передбачати поділ на мікросервіси, що дозволить легко масштабувати та оновлювати кожен компонент окремо. Рекомендаційна система, як сервіс, зможе взаємодіяти з базою даних книг і користувачів, отримуючи необхідні дані для аналізу та надання персоналізованих рекомендацій.

Ключовим елементом є створення RESTful API, який забезпечить зв'язок між рекомендаційною системою та іншими частинами магазину. API дозволить запитувати рекомендації для конкретного користувача в режимі реального часу на основі його дій та історії покупок. Це може бути виконано під час завантаження сторінки користувача або після перегляду конкретних книг, коли система генерує рекомендації відповідно до контентних або поведінкових даних.

Висновки. В результаті проведеного дослідження було проаналізовано різні підходи створення рекомендаційної системи та спроектовано гібридну рекомендаційну систему для книжкового інтернет-магазину, яка поєднує контентну та колаборативну фільтрації. Система враховує індивідуальні вподобання користувачів та їх поведінку, що дозволяє створювати релевантні рекомендації навіть за відсутності достатньої кількості даних про взаємодію нових користувачів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] Murel, J. Ph.D., Kavlakoglu, E. (2024, 21 березня). *What is content-based filtering? IBM - United States*. <https://www.ibm.com/topics/content-based-filtering>
- [2] Miesle, P. (2023, 29 вересня). *What is Cosine Similarity? A Comprehensive Guide*. DataStax. <https://www.datastax.com/guides/what-is-cosine-similarity>
- [3] Vihar, K. (2022, 11 жовтня). *What Is Collaborative Filtering: A Simple Introduction*. <https://builtin.com/data-science/collaborative-filtering-recommender-system>
- [4] McClenaghan, E. (2024, 29 квітня). *Pearson Correlation*. Technology Networks. <https://www.technologynetworks.com/tn/articles/pearson-correlation-385871>
- [5] Saturn Cloud. (2023). *Hybrid Recommender Systems*. <https://saturncloud.io/glossary/hybrid-recommendation-system/>

Pyskir S.V., Havrylova K.M.

Analysis and design of personalized recommendations for an online bookstore based on machine learning

Abstract. *This paper presents the analysis and design of a recommender system for an online bookstore based on machine learning methods. The main goal of the system is to increase user satisfaction by offering personalized recommendations that match their preferences.*

Keywords: *hybrid system; machine learning; collaborative filtering; recommendation system; online bookstore.*