УДК 004.8:004.934.2

АНАЛІЗ ТА РОЗРОБКА ГІБРИДНОЇ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ДЛЯ КНИЖКОВОГО ІНТЕРНЕТ-МАГАЗИНУ

Пискір С.В.
Національний університет «Одеська політехніка»
Україна, Одеса
stasikpyskir2@gmail.com

Анотація. *У цій роботі представлено результати дослідження та реалізації гібридної рекомендаційної системи для онлайн-книгарні. Модель об'єднує контентно-орієнтований підхід із колаборативною фільтрацією. Такий підхід дозволяє долати обмеження окремих методів і забезпечує користувачам релевантні рекомендації, адаптовані до їхніх інтересів.*
Ключові слова: *гібридна рекомендаційна система; контентно-орієнтований підхід; колаборативна фільтрація; TF–IDF; косинусна подібність; SVD; персоналізація; книжковий інтернет-магазин.*

Вступна частина. Стрімкий розвиток електронної комерції сприяв перетворенню онлайн-книгарень на один із головних каналів поширення літератури. Широкий асортимент, що налічує десятки тисяч видань, створює нові можливості для читачів, але водночас ускладнює процес вибору книги. Звичайні механізми пошуку та категоризації не враховують індивідуальних інтересів користувачів, через що процес знаходження потрібної літератури часто стає тривалим і незручним.

Ефективним інструментом вирішення цієї проблеми є рекомендаційні системи, які здатні аналізувати поведінку користувачів та їхні вподобання. Використання гібридних підходів, що поєднують контентний аналіз і колаборативну фільтрацію, дозволяє компенсувати недоліки кожного методу окремо й формувати більш точні та різноманітні добірки. Саме тому створення такої системи для книжкових онлайн-магазинів є актуальним науково-практичним завданням.

Метою роботи є аналіз існуючих методів та розробка гібридної рекомендаційної системи для книжкового інтернет-магазину, яка поєднує контентно-орієнтований підхід із колаборативною фільтрацією. Запропонована комбінація дозволяє підвищити точність і різноманітність рекомендацій, зменшити вплив проблеми «холодного старту» та покращити персоналізацію для користувачів.

Основна частина. У рекомендаційних системах застосовуються різні методи для формування персоналізованих пропозицій. Найпоширенішими є три підходи: контентно-орієнтований, колаборативна фільтрація та гібридний метод. Кожен з них має свої переваги та обмеження, що визначають сферу їх ефективного застосування.

Контентно-орієнтований підхід ґрунтується на аналізі характеристик об'єктів – у випадку книжкових онлайн-магазинів це жанри, автори, ключові слова та описи книг. Така система формує рекомендації на основі схожості між атрибутами товарів і може працювати навіть за відсутності великої кількості даних про користувача [1]. Водночас, її недоліком є «вузький профіль»: користувач зазвичай отримує пропозиції лише подібних до вже переглянутих або придбаних книг, що обмежує різноманітність рекомендацій.

Системи колаборативної фільтрації (Collaborative filtering) – ґрунтуються на схожості користувачів (user-based) або схожості об'єктів (item-based) [2]:

– user-based – рекомендації формуються на основі подібності користувачів: користувачу пропонуються об’єкти, які були позитивно оцінені іншими користувачами зі схожими вподобаннями;

– item-based – рекомендації ґрунтуються на подібності між самими об’єктами: користувачу пропонуються ті об’єкти, які мають схожий патерн оцінок з тими, що йому вже сподобалися.

Цей підхід дозволяє враховувати колективний досвід, але має проблему «холодного старту» коли для певних користувачів або елементів невідомі попередні події, такі як рейтинги чи кліки.

Гібридні системи – поєднують переваги кількох методів, зазвичай контентного та колаборативного. Це дозволяє зменшити недоліки кожного з підходів окремо: контентна складова частково вирішує проблему «холодного старту», а колаборативна забезпечує більшу різноманітність рекомендацій. Гібридні системи вважаються найбільш ефективними у випадках, коли необхідно обробляти великі каталоги з різноманітними характеристиками об’єктів.

Аналіз підходів показав, що використання лише одного методу не дозволяє забезпечити достатній рівень персоналізації. Контентно-орієнтований підхід має обмеження у різноманітності, тоді як колаборативний страждає від проблеми «холодного старту». Саме тому у даній роботі обрано гібридний підхід, що поєднує обидва методи:

– контентна частина формує рекомендації на основі опису, жанру та автора книги;

– колаборативна частина user-based враховує покупки користувачів та їхні оцінки.

Очікується, що така комбінація забезпечить баланс між точністю і різноманітністю рекомендацій та зробить процес вибору книг більш зручним для користувачів.

Почнемо з контентно-орієнтованого підходу. Найбільш поширеним методом у контентних рекомендаційних системах є представлення об’єктів у вигляді векторів ознак, після чого визначається їхня подібність. Для векторизації текстових описів книг (жанри, автори, опис) використовується метод TF–IDF (Term Frequency – Inverse Document Frequency).

TF–IDF дозволяє перетворити текст у набір числових значень, які відображають значимість слів у межах конкретного документа відносно всього корпусу текстів [3].Формула TF-IDF виглядає наступним чином:

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

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 $TF(t,d)$ загальна кількість документів у корпусі;

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

Після векторизації необхідно визначати подібність між книгами. Для цього використовується косинусна подібність, яка вимірює кут між двома векторами у багатовимірному просторі [4]. Формула для розрахунку косинусної подібності виглядає так:

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

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

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

Косинусна подібність має значення від -1 до 1, де 1 означає ідентичність векторів, 0 вказує на їх ортогональність – відсутність спільних ознак, а -1 – протилежність векторів.

Перейдемо до другого компоненту гібридної системи, колаборативної фільтрації, яка ґрунтується на аналізі вподобань користувачів. Суть user-based підходу полягає в тому, що користувачу рекомендуються книги, які сподобалися іншим користувачам зі схожими інтересами. Існує багато алгоритмів, які реалізують цей підхід, тому для обґрунтованого вибору алгоритму було проведено експериментальне тестування за допомоги бібліотеки Surprise, Вона надає готові реалізації основних алгоритмів колаборативної фільтрації та інструменти для їхньої оцінки за стандартними метриками RMSE – середньоквадратична похибка, яка відображає середній розмір відхилення прогнозованих значень від реальних. Вона особливо чутлива до великих похибок, оскільки помилки підносяться до квадрата та MAE – середня абсолютна похибка, яка показує середню різницю між фактичними та прогнозованими значеннями незалежно від напрямку відхилення. На відміну від RMSE, ця метрика менш чутлива до поодиноких великих помилок. Це дозволяє швидко визначити, який метод є найбільш ефективним для даних умов. Тестування проводилося на обробленому датасеті який містить користувачів та їх оцінки до книг після фільтрації та нормалізації даних, результати тестування наведені у таблиці 1.

Таблиця 1 – Результати тестування алгоритмів колаборативної фільтрації

Алгоритм	RMSE	MAE
SVD	0.7715	0.5893
SVD++	0.7774	0.5924
BaselineOnly	0.7828	0.6001
CoClustering	0.8326	0.6256
KNNBaseline	0.8545	0.6478
KNNWithMeans	0.8718	0.6452
NMF	0.9001	0.6962
KNNBasic	0.9209	0.7005
SlopeOne	0.9293	0.6909
NormalPredictor (random)	1.1262	0.8843

Аналіз результатів тестування показав, що найкраще співвідношення точності та продуктивності демонструє алгоритм SVD (Singular Value Decomposition). Саме його було обрано для реалізації користувацької частини рекомендаційної системи.

SVD є технікою матричної факторизації, яка розкладає матрицю оцінок R на три низькорозмірні матриці, що відображають приховані закономірності у вподобаннях, у моєму випадку матриця користувач-книга, де рядки – це користувачі, стовпці – книги, а елементи є оцінками, формула SVD виглядає так [5].

$$R \approx U \cdot \Sigma \cdot V^T, \tag{4}$$

де U – матриця користувачів, що містить латентні фактори для кожного користувача;

Σ – діагональна матриця сингулярних значень, яка відображає вагу відповідних латентних компонентів;

V^T – матриця книг, що містить латентні фактори для кожної книги.

Після такого розкладу відсутні значення в матриці R можна апроксимувати, що дозволяє передбачати рейтинги для пар «користувач-книга», яких не було в початкових даних. Це створює можливість формувати персоналізовані рекомендації для кожного користувача, враховуючи його приховані інтереси та схожість з іншими користувачами.

Для реалізації рекомендаційних систем було використано бібліотеку scikit-learn, яка забезпечує роботу з розрідженими матрицями та надає інструменти для векторизації текстових даних і матричної факторизації.

Першим етапом створення рекомендаційної системи став аналіз і підготовка даних. Для контентно-орієнтованої моделі було використано набір, що містив ключові характеристики книг: назви, авторів, жанри та описи. Попередня обробка передбачала очищення текстів від артефактів, заповнення пропусків, стандартизацію жанрів та формування комбінованого текстового поля, яке включало автора, жанри та опис. Це дозволило сформувати повніші векторні подання книг та забезпечити коректність подальшої векторизації.

Для колаборативної моделі використовувався датасет із рейтингами користувачів. Первинний аналіз виявив високу розрідженість даних, тому було застосовано відбір активних користувачів та популярних книг, а також нормалізацію шкали оцінок. У результаті сформовано user–item матрицю з якісними рейтингами, придатними для навчання моделей колаборативної фільтрації.

У контентно-орієнтованій моделі використовувалися описи книг, інформація про авторів та жанри. Після попередньої обробки текст було перетворено у векторні представлення за допомогою алгоритму TF–IDF. Для визначення схожості між книгами застосовувалася косинусна міра. Система формувала рекомендації на основі останніх переглянутих користувачем видань, пропонуючи книги зі схожими змістовими характеристиками. Така модель забезпечує високу узгодженість за жанрами та здатність підтримувати різноманітність рекомендацій. Колаборативна модель побудована на основі матриці користувач–книга, що містить рейтинги. Через високу розрідженість даних було застосовано метод TruncatedSVD, який дозволяє зменшити розмірність і виявити латентні фактори. Це забезпечує прогнозування інтересів користувачів і формування персоналізованих рекомендацій на основі поведінкових патернів.

Висновки. У результаті проведеного дослідження було проаналізовано основні методи побудови рекомендаційних систем та розроблено гібридну модель для книжкового інтернет-магазину, що поєднує контентно-орієнтований підхід та колаборативну фільтрацію на основі SVD. Система враховує як змістові характеристики книг, так і поведінкові дані користувачів та є готовою до подальшого тестування й інтеграції у вебзастосунок.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Murel J., Kavlakoglu E. (2024, 21 березня). What is content-based filtering? IBM. https://www.ibm.com/topics/content-based-filtering

[2] Cechmanek J. (2025, 18 лютого). Collaborative filtering: How to build a recommender system. Redis. https://redis.io/blog/collaborative-filtering-how-to-build-a-recommender-system/

[3] Understanding TF-IDF (Term Frequency-Inverse Document Frequency) (2025, 13 серпня). GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/

[4] Krantz T., Jonker A. (2025, 30 липня) What Is Cosine Similarity? IBM. https://www.ibm.com/think/topics/cosine-similarity

[5] Singular Value Decomposition (SVD) (2025, 5 липня). GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/singular-value-decomposition-svd/

Pyskir S.V.

Analysis and development of a hybrid recommendation system for an online bookstore

Abstract. *This paper presents the results of research and development of a hybrid recommendation system for an online bookstore. The model combines a content-oriented approach with collaborative filtering. This approach overcomes the limitations of individual methods and provides users with relevant recommendations tailored to their interests.*

Keywords: *hybrid recommendation system; content-oriented approach; collaborative filtering; TF–IDF; cosine similarity; SVD; personalization; online bookstore.*