Бакин Руслан Романович

Студент, Байкальский государственный университет

г. Иркутск, Российская Федерация

АНАЛИЗ МЕТОДОВ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ ДЛЯ ПОСТРОЕНИЯ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ В КНИЖНОЙ ОТРАСЛИ: SVD, KNN И NMF

Аннотация. В данной статье представлен сравнительный анализ трех методов коллаборативной фильтрации, применяемых в рекомендательных системах книжной отрасли: сингулярного разложения матрицы (SVD), метода k-ближайших соседей (KNN) и неотрицательной матричной факторизации (NMF). Основное внимание уделяется оценке эффективности этих методов в контексте предсказания предпочтений пользователей и формирования рекомендаций на основе их предыдущих взаимодействий с контентом. Для количественной оценки качества предложенных моделей используются стандартные метрики, такие как RMSE (корень из средней квадратичной ошибки), MAE (средняя абсолютная ошибка) и MSE (среднеквадратичная ошибка). В процессе исследования анализируются преимущества и ограничения каждого из методов в рамках задач книжной рекомендации. Результаты работы позволяют выделить наиболее подходящие алгоритмы для разных типов данных, а также предоставить рекомендации по выбору оптимального метода в зависимости от специфики данных и требований к системе. Данное исследование представляет практическую ценность для разработчиков рекомендательных систем в книжной отрасли и может быть полезным в области персонализированных рекомендаций в других индустриях.

Abstract. This paper presents a comparative analysis of three collaborative filtering methods used in recommendation systems within the book industry: Singular Value Decomposition (SVD), k-Nearest Neighbors (KNN), and Non-negative Matrix Factorization (NMF). The primary focus is on evaluating the effectiveness of these methods in predicting user preferences and generating recommendations based on their previous interactions with content. To

quantitatively assess the quality of the proposed models, standard metrics such as RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and MSE (Mean Squared Error) are used. The study analyzes the advantages and limitations of each method in the context of book recommendation tasks. The results highlight the most suitable algorithms for different types of data and provide recommendations on selecting the optimal method based on the specific characteristics of the data and system requirements. This research is of practical value for developers of recommendation systems in the book industry and can be useful in the area of personalized recommendations in other industries.

Ключевые слова: коллаборативная фильтрация, SVD, KNN, NMF, рекомендательные системы, книжная отрасль, метрики качества, RMSE, MAE, MSE.

Keywords: collaborative filtering, SVD, KNN, NMF, recommendation systems, book industry, quality metrics, RMSE, MAE, MSE.

С развитием цифровых технологий и увеличением объемов данных рекомендательные системы становятся ключевыми компонентами в различных областях, таких как электронная коммерция, музыкальная индустрия, кино и книжная отрасль [1]. В частности, в контексте книжной индустрии такие системы играют решающую роль в повышении пользовательского опыта, обеспечивая персонализированные рекомендации, что способствует как увеличению продаж, так и повышению удовлетворенности клиентов.

Актуальность разработки высокоэффективных алгоритмов для рекомендательных систем в данной области обусловлена растущей конкуренцией на рынке и потребностью в удовлетворении интересов и предпочтений пользователей.

Рекомендательные системы функционируют на основе анализа данных о взаимодействиях пользователей с контентом, что позволяет предсказать их предпочтения. Одним из наиболее распространенных подходов к созданию таких систем является метод коллаборативной фильтрации. Данный метод

предполагает, что пользователи, имеющие схожие предпочтения в прошлом, с высокой вероятностью будут иметь схожие интересы и в будущем [2]. Несмотря на широкое использование коллаборативной фильтрации, она сталкивается с рядом существенных проблем, включая проблему «холодного старта», высокую вычислительную сложность и необходимость обработки больших объемов данных, что ограничивает её применение в некоторых сценариях.

В рамках данной работы рассматривается три основных метода коллаборативной фильтрации — сингулярное разложение матрицы (SVD)[1], метод k-ближайших соседей (KNN)[2] и неотрицательная матричная факторизация (NMF)[1]. Эти методы анализируются с точки зрения их применимости в книжной отрасли, где важным аспектом является как точность рекомендаций, так и их способность к персонализации. Для экспериментальной проверки данных алгоритмов был разработан веб-сайт на платформе Streamlit[3], что позволило интерактивно тестировать их на реальных данных о книгах. Платформа предоставляет возможность сравнивать различные модели и оценивать их эффективность с использованием метрик, таких как RMSE, MAE и MSE.

Целью данного исследования является проведение сравнительного анализа методов SVD, KNN и NMF с точки зрения их применения для построения рекомендательных систем в книжной отрасли. В работе будет осуществлена детальная оценка каждого метода, проведен анализ их производительности и точности, а также сформулированы рекомендации по выбору наиболее подходящих методов в зависимости от характеристик данных и задач системы. Полученные результаты имеют практическую значимость для разработчиков рекомендательных систем в книжной

---

[1] Матричная факторизация // Yandex Education. URL: https:// education. yandex.ru/ handbook/ ml/article /matrichnaya-faktorizaciya.

[2] Метричкеские методы // Yandex education. URL: https:// education. yandex.ru/ handbook /ml /article/metricheskiye-metody.

[3] Streamlit. URL: https://streamlit.io/.

индустрии и могут быть применены в других отраслях, где требуются высокоэффективные решения для персонализированных рекомендаций.

Коллаборативная фильтрация представляет собой одну из наиболее широко используемых техник в области рекомендательных систем[4]. Этот подход основывается на предположении, что пользователи, имеющие схожие предпочтения в прошлом, будут иметь схожие предпочтения в будущем. В коллаборативной фильтрации, как правило, используются различные методы для анализа взаимодействий между пользователями и товарами, такими как сингулярное разложение матрицы (SVD), метод k-ближайших соседей (KNN) и неотрицательная матричная факторизация (NMF). Каждый из этих методов реализует различные аспекты коллаборативного подхода и может быть использован в зависимости от требований системы и характеристик данных.

Сингулярное разложение матрицы (SVD — Singular Value Decomposition) — это метод линейной алгебры, который используется для разложения исходной матрицы предпочтений пользователей на несколько меньших матриц, что позволяет выделить скрытые факторы, влияющие на предпочтения пользователей [3]. Основная идея заключается в представлении матрицы предпочтений как произведения трех матриц: пользовательской, товарной и матрицы сингулярных значений. Такой подход позволяет выявить латентные зависимости между пользователями и товарами, что способствует более точному предсказанию интересов пользователей.

Применение SVD дает возможность значительно уменьшить размерность данных, что позволяет улучшить вычислительную эффективность при обработке больших наборов данных. Тем не менее, одним из ограничений метода является проблема холодного старта, когда для новых пользователей или товаров недостаточно информации для построения качественных рекомендаций. Также SVD требует наличия большого объема данных, чтобы эффективно выявить скрытые факторы, что может быть

[4] Коллаборативная фильтрация. URL: https://habr.com/ru/articles/150399/.

затруднительно для системы с ограниченным числом пользователей или товаров.

Метод k-ближайших соседей (KNN — K-Nearest Neighbors) является одним из наиболее интуитивно понятных и простых методов коллаборативной фильтрации. Этот метод основан на идее, что для каждого пользователя система находит k пользователей, чьи предпочтения наиболее похожи на предпочтения целевого пользователя, и на основе их предпочтений формирует рекомендации. В KNN могут быть использованы два подхода: фильтрация на основе пользователей и фильтрация на основе товаров [4].

В случае фильтрации на основе пользователей система ищет пользователей, схожих с целевым, и рекомендует товары, которые оценены высоко этими соседями. В фильтрации на основе товаров система ищет товары, похожие на те, которые были оценены пользователем, и рекомендует их. Этот метод прост в реализации и хорошо подходит для задач с небольшими объемами данных. Однако при большом объеме пользователей или товаров вычислительная сложность метода возрастает, так как для каждого нового запроса необходимо проводить сравнение с данными всех остальных пользователей или товаров.

Неотрицательная матричная факторизация (NMF — Non-negative Matrix Factorization) представляет собой метод факторизации, аналогичный SVD, но с важным ограничением: все элементы в результирующих матрицах должны быть неотрицательными. Это свойство делает NMF особенно полезным для анализа данных, где все значения имеют только положительный характер, такие как рейтинги, оценки и другие типы взаимодействий с контентом [5].

Процесс факторизации в NMF заключается в нахождении двух неотрицательных матриц, которые аппроксимируют исходную матрицу предпочтений. Это разложение позволяет выявить скрытые факторы, которые могут быть использованы для формирования рекомендаций. Основное преимущество NMF перед SVD заключается в том, что оно более эффективно работает с данными, содержащими только положительные значения. Однако

NMF также имеет свои ограничения, такие как высокая вычислительная сложность и необходимость оптимизации гиперпараметров, что может повлиять на его производительность при работе с большими наборами данных.

В контексте книжной отрасли использование методов коллаборативной фильтрации позволяет строить системы рекомендаций, которые могут точно предсказывать предпочтения пользователей, основываясь на их взаимодействиях с книгами. Каждый из рассмотренных методов — SVD, KNN и NMF — может быть адаптирован для различных сценариев и типов данных. Метод SVD, благодаря своей способности выявлять скрытые закономерности и эффективно работать с большими данными, подходит для задач, где важно учитывать разнообразие жанров, авторов и другие факторы, влияющие на предпочтения пользователей. Однако его применение ограничено проблемой холодного старта, особенно для новых пользователей и товаров.

Метод KNN подходит для ситуаций, где необходимо учитывать схожесть между пользователями или товарами. Он является простым в реализации и может быть полезен при работе с небольшими наборами данных. Однако его высокая вычислительная сложность при увеличении объема данных может ограничить его использование в масштабируемых системах.

NMF, с другой стороны, особенно полезен при анализе только положительных данных, таких как рейтинги, что делает его хорошим выбором для работы с отзывами пользователей о книгах. Однако его необходимость в подстройке и высокая вычислительная нагрузка могут быть ограничениями для его применения на больших наборах данных.

Для оценки эффективности рекомендательных систем важно использовать количественные метрики, которые позволяют объективно измерить точность и производительность различных алгоритмов. В рамках данного исследования будут использоваться несколько метрик: RMSE (корень из средней квадратичной ошибки), MAE (средняя абсолютная ошибка) и MSE

(среднеквадратичная ошибка), каждая из которых имеет свои особенности и области применения[5].

Корень из средней квадратичной ошибки (RMSE, Root Mean Squared Error) является одной из наиболее популярных метрик для оценки точности предсказаний в рекомендательных системах. RMSE измеряет среднеквадратичное отклонение между предсказанными и фактическими значениями, что позволяет оценить, насколько хорошо модель предсказывает рейтинги пользователей. Формула для вычисления RMSE выглядит следующим образом:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_{ui} - \hat{r}_{ui})^2} \quad (1)$$

где:

- N – Общее количество оценок,
- r_{ui} – фактический рейтинг пользователя u для книги i,
- $\hat{r}_{ui}$ – предсказанный рейтинг пользователя u для книги i.

RMSE дает представление о среднем отклонении предсказаний от реальных значений и является хорошей метрикой для оценки общей точности модели. Однако RMSE чувствителен к большим ошибкам, так как ошибки, возведенные в квадрат, усиливают влияние крупных отклонений.

Средняя абсолютная ошибка (MAE, Mean Absolute Error) — это метрика, которая измеряет среднее значение абсолютных отклонений между предсказанными и фактическими оценками. MAE дает прямое представление о средней ошибке модели, выраженной в тех же единицах, что и данные. Формула для вычисления MAE следующая:

[5] Простыми словами про метрики в ИИ. Регрессия. MSE, RMSE, MAE, R-квадрат, MAPE // Habr. URL: https:// habr.com/ru/articles/820499/.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|r_{ui} - \hat{r}_{ui}| \quad (2)$$

где:

- N – Общее количество оценок,
- r_{ui} – фактический рейтинг,
- $\hat{r}_{ui}$ – предсказанный рейтинг.

MAE проще интерпретировать, чем RMSE, так как оно не усиливает ошибки, и может быть полезной метрикой для задач, где важен не столько масштаб ошибок, сколько их среднее значение. Однако MAE может игнорировать крупные ошибки, что делает его менее чувствительным к экстремальным отклонениям, чем RMSE.

Среднеквадратичная ошибка (MSE, Mean Squared Error) является схожей с RMSE, однако она не включает извлечение квадратного корня, что делает её менее интерпретируемой в тех же единицах, что и исходные данные. MSE оценивает среднее квадратичное отклонение между предсказанными и фактическими значениями. Формула для вычисления MSE следующая:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(r_{ui} - \hat{r}_{ui})^2 \quad (3)$$

где:

- N – Общее количество оценок,
- r_{ui} – фактический рейтинг,
- $\hat{r}_{ui}$ – предсказанный рейтинг.

MSE схожа с RMSE, но поскольку она не извлекает квадратный корень, её значение всегда будет больше, чем значение RMSE для тех же данных. Эта метрика чувствительна к большим ошибкам и дает более высокую степень наказания за крупные отклонения. MSE полезна для случаев, когда требуется усилить влияние больших ошибок на итоговую оценку.

Для эксперимента используется датасет Amazon Popular Books Dataset[6], доступный на GitHub. Этот набор данных включает в себя информацию о популярных книгах на платформе Amazon, предоставляя подробности о названиях книг, авторах, рейтингах и изображениях, а также о других метаданных, связанных с книгами. Набор данных содержит следующие столбцы:

- asin: Идентификатор книги на платформе Amazon.
- title: Название книги.
- author: Автор книги.
- image_url: Ссылка на изображение обложки книги.
- rating: Рейтинг книги, присвоенный пользователями Amazon.
- rating_count: Количество пользователей, оставивших рейтинг.
- price: Цена книги.
- category: Категория книги (например, фэнтези, детективы и т. Д.).

Этот набор данных представляет собой идеальный источник для построения рекомендательных систем, так как он включает в себя не только информацию о рейтингах книг, но и метаданные, такие как авторы и категории, которые могут быть полезны при анализе предпочтений пользователей

Для оценки производительности различных методов коллаборативной фильтрации (SVD, KNN и NMF) были использованы метрики RMSE, MAE и MSE, которые позволяют объективно измерить точность рекомендаций для каждой модели [6]. Результаты этих метрик представлены в таблице ниже. Наилучшие показатели по всем оценочным метрикам продемонстрировала модель SVD, что свидетельствует о её высокой точности предсказаний и эффективности в задачах построения рекомендательных систем. Модели NMF и KNN уступили SVD по большинству метрик, продемонстрировав более высокие значения ошибок и меньшую точность прогнозирования.

[6] Amazon Popular Books Dataset // Github URL: https://github.com/luminati-io/Amazon-popular-books-dataset.

Таблица 1

Результатами метрик для каждой модели

Model	RMSE	MAE	MSE
NMF	0,072	0,061	0,005
SVD	0,050	0,045	0,004
KNN	0,135	0,056	0,018

Для каждой модели были построены графики распределения ошибок. Для SVD распределение ошибок оказалось наиболее сконцентрированным вблизи нуля, что свидетельствует о высоком качестве предсказаний и минимальных отклонениях (рис. 1). В то же время распределения ошибок для NMF и KNN были шире (рис. 2, рис. 3), что указывает на более высокие значения ошибок и меньшую точность предсказаний по сравнению с моделью SVD.

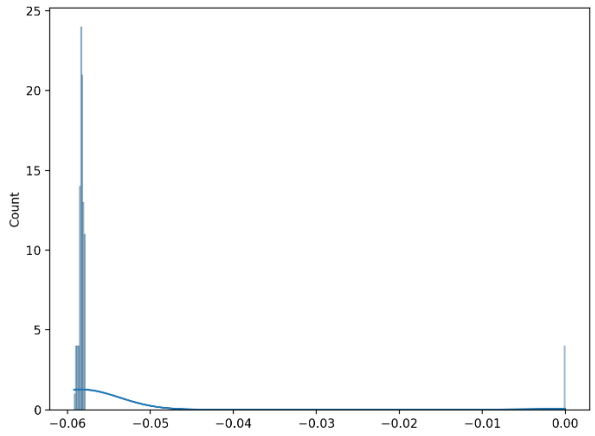

Рисунок 1. Распределение ошибок для модели SVD

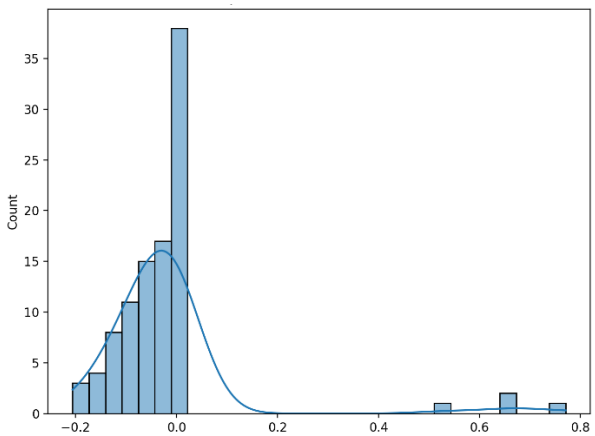

Рисунок 2. Распределение ошибок для модели NMF

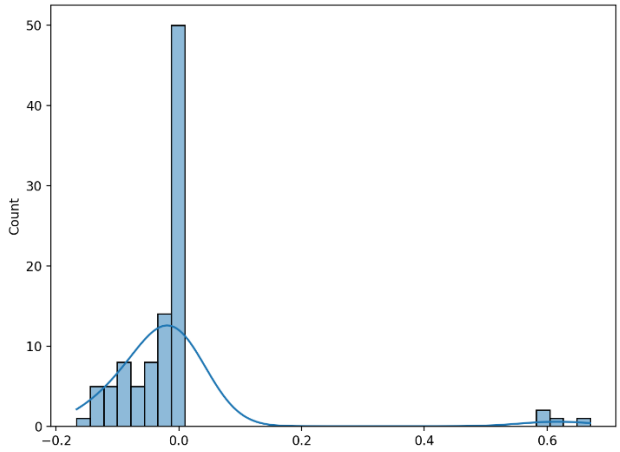

Рисунок 3. Распределение ошибок для модели KNN

Эксперименты проводились с использованием языка программирования Python, который является одним из наиболее популярных инструментов для анализа данных и разработки машинного обучения. Для реализации интерфейса веб-приложения использовалась библиотека Streamlit, которая позволяет легко создавать интерактивные веб-приложения для анализа данных. Для обработки данных и обучения моделей были использованы библиотеки Pandas[7], NumPy[8], а для визуализации результатов — Matplotlib[9] и Seaborn[10]. Методы коллаборативной фильтрации были реализованы с использованием библиотеки Surprise[11], которая предоставляет удобные средства для работы с такими моделями, как SVD, KNN и NMF, а также для вычисления метрик, таких как RMSE, MAE и MSE [7].

Полученные результаты эксперимента продемонстрировали, что метод SVD является наиболее эффективным для рассматриваемой задачи, обеспечивая минимальные значения метрик RMSE, MAE и MSE. Это свидетельствует о высокой точности рекомендаций, формируемых данной моделью коллаборативной фильтрации, и подтверждает её практическую

---

[7] Документация pandas. URL: http://pandas.geekwriter.ru/#pandas-documentation.

[8] NumPy documentation. URL: https://numpy.org/doc/stable/.

[9] Matplotlib: Visualization with Python. URL: https://matplotlib.org/.

[10] Seaborn: statistical data visualization. URL: https://seaborn.pydata.org/.

[11] A Python scikit for recommender systems. URL: https://surpriselib.com/; Surprise 1 documentation. URL: https://surprise.readthedocs.io/en/stable/getting_started.html; Recommender systems with Python - (3) Introduction to Surprise package in Python. URL: https://buomsoo-kim.github.io/recommender%20systems/2020/07/18/Recommender-systems-collab-filtering-3.md/.

применимость при построении рекомендательных систем для книжной отрасли. Преимущества SVD заключаются в его способности эффективно обрабатывать большие объемы данных, что особенно важно при наличии значительного количества информации о пользователях и товарах, а также при минимизации проблемы "холодного старта".

Метод NMF показал менее точные результаты по сравнению с SVD, однако остаётся конкурентоспособным для задач, связанных с обработкой положительных рейтингов. Несмотря на свою популярность, метод KNN продемонстрировал наиболее высокие значения ошибок среди рассматриваемых алгоритмов, что указывает на ограниченность его применения для крупных и разнородных наборов данных из книжной сферы.

Применение таких методов коллаборативной фильтрации в книжной индустрии может значительно улучшить качество персонализированных рекомендаций и повысить удовлетворенность пользователей. Это исследование также открывает перспективы для использования этих методов в других отраслях, таких как электронная коммерция, кино и музыка, где персонализированные рекомендации играют ключевую роль.

Список использованной литературы

1. Isinkaye F.O., Folajimi Y.O., Ojokoh B.A. Recommendation systems: Principles, methods and evaluation // Egyptian Informatics Journal. 2015. Vol. 16, Issue 4, Pp. 261–273. DOI: 10.1016/j.eij.2015.06.005.

2. Архипова З.В., Сорокин А.В. Анализ подходов к созданию рекомендательных систем в сфере предоставления образовательных услуг. // System Analysis & Mathematical Modeling. 2024. Т. 6, № 2. С. 133–145. DOI: 10.17150/2713-1734.2024.6(2).133-145.

3. Sarwar B., Karypis G., Konstan J., Riedl J. Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems // Proceedings of the ACM WebKDD'00 (Web-mining for E-Commerce Workshop), 2000.

4. Hssina, B., Grota, A., & Erritali, M. (2021). Recommendation system using the k-nearest neighbors and singular value decomposition algorithms. International Journal of Electrical and Computer Engineering (IJECE), 11(6), 5541-5548.

5. Hosseinzadeh Aghdam, M., Analoui, M., & Kabiri, P. (2015). A Novel Non-Negative Matrix Factorization Method for Recommender Systems. Applied Mathematics & Information Sciences, 9(5), 2721-2732.

6. Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2009). Scalable collaborative filtering approaches for large recommender systems. Journal of Machine Learning Research, 10, 623–656.

7. Hug, N. (2020). Surprise: A Python library for recommender systems. Journal of Open Source Software, 5(52), 2174.