DOI 10.37882/2223-2966.2026.01.11

ГИБРИДНЫЕ АЛГОРИТМЫ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ ДЛЯ РЕКОМЕНДАЦИИ ТОВАРНЫХ ГРУПП НА ОСНОВЕ НЕОДНОРОДНЫХ ПОЛЬЗОВАТЕЛЬСКИХ ДАННЫХ

HYBRID COLLABORATIVE FILTERING ALGORITHMS FOR PRODUCT GROUP RECOMMENDATIONS BASED ON HETEROGENEOUS USER DATA

M. Verbova
V. Podolnyy
S. Suvorov

Summary. With the rapid growth of e-commerce, the ability to hold the user's attention and provide relevant personalized content is becoming an important success factor. The purpose of this work is to develop and experimentally validate a hybrid recommendation system algorithm for a marketplace that solves the problems of incomplete interaction history and disparate metadata. The proposed solution combines two classical approaches: collaborative filtering based on matrix decompositions (Alternating Least Squares) to identify hidden patterns in implicit interactions and a modified popularity-based recommendation method that considers the individual user story.

Keywords: recommendation systems, hybrid algorithm, collaborative filtering, Alternating Least Squares.

Вербова Мария Андреевна
Московский политехнический университет
358maryv@gmail.com
Подольный Владимир Александрович
Московский политехнический университет
nl.podolnyy@vk.com
Суворов Станислав Вадимович
кандидат экономических наук, профессор,
Московский политехнический университет
kafedrapi12@mail.ru

Аннотация. В условиях стремительного роста электронной торговли способность удерживать внимание пользователя и предоставлять релевантный персонализированный контент становится важным фактором успеха. Целью данной работы является разработка и экспериментальная валидация гибридного алгоритма рекомендательной системы для маркетплейса, решающего проблемы неполной истории взаимодействий и разрозненных метаданных. Предлагаемое решение комбинирует два классических подхода: коллаборативную фильтрацию на основе матричных разложений (Alternating Least Squares) для выявления скрытых паттернов в неявных взаимодействиях и модифицированный метод рекомендаций на основе популярности, учитывающий индивидуальную историю пользователя.

Ключевые слова: рекомендательные системы, гибридный алгоритм, коллаборативная фильтрация, Alternating Least Squares.

Введение

В настоящее время наблюдается стремительный рост числа и объема онлайн-платформ электронной торговли. В этих условиях одним из ключевых фактора успеха является способность удерживать внимание пользователя. Персонализация предложений перестала быть дополнительной опцией и стала чем-то привычным и повсеместным. Пользователи ожидают, что система «понимает» их потребности и предлагает релевантные товары, что напрямую влияет на лояльность покупателей и на итоговую выручку.

На практике построение таких персонализированных рекомендаций сталкивается с некоторым проблемами, например: неполная история взаимодействий, разрозненные метаданные.

Рекомендательные системы применяют методы анализа данных, чтобы помогать пользователям находить товары на платформах электронной коммерции, формируя прогнозы или списки рекомендованных позиций. Рекомендации могут строиться на основе демографии, популярности товаров или истории покупок пользователя. Наиболее эффективным современным методом является коллаборативная фильтрация (CF), которая предлагает рекомендации или прогнозы, опираясь на оценки и поведение схожих пользователей [4].

Целью данной работы является разработка и экспериментальная валидация гибридного алгоритма рекомендаций, который комбинирует два классических подхода: коллаборативную фильтрацию для выявления паттернов схожести пользователей и товаров на основе их неявных взаимодействий и контентные методы (Content-based) для обогащения моделей признаками из метаданных и учета свойств самих сущностей.

Разработанное решение может обладать прямой практической значимостью и предназначено для внедрения в промышленные рекомендательные системы крупных маркетплейсов, что может позволить: суще-

ственно увеличить конверсию посетителей в покупателей за счет предоставления более точных и разнообразных персонализированных предложений, повысить средний чек за счет кросс-продаж и рекомендации сопутствующих групп товаров.

Методы

В данной работе была разработана двухэтапная гибридная рекомендательная система для предсказания актуальных товарных групп для пользователей маркетплейса. Алгоритм объединяет коллаборативную фильтрацию на основе матричных разложений и контентно-независимые методы, что позволяет учитывать как историю взаимодействий пользователей, так и общую популярность товарных групп.

Система состоит из следующих ключевых компонентов:

1. Модель Alternating Least Squares (ALS).
2. Модифицированный метод рекомендаций на основе популярности, который учитывает индивидуальную историю просмотров каждого пользователя, исключая уже просмотренные товарные группы.
3. Алгоритм объединения рекомендаций от различных методов с весовой схемой, где рекомендациям от ALS присваивается больший вес.

Особенности реализации: используются только контактные события (is_contact = 1), исключаются товарные группы, с которыми пользователь уже взаимодействовал, каждый пользователь получает ровно 40 рекомендаций, при отсутствии персонализированных рекомендаций используется глобальная популярность.

Стратегия, на которой фокусируется данная работа, опирается только на прошлое поведение пользователей, не требуя создания явных профилей. Этот подход известен как Коллаборативная фильтрация (Collaborative Filtering, CF). CF анализирует взаимосвязи между пользователями и взаимозависимости между товарами, чтобы выявить новые ассоциации «пользователь-товар». Единственной необходимой информацией является прошлое поведение пользователей, которое может включать их предыдущие транзакции или способ оценки товаров. Главное преимущество CF заключается в том, что она не зависит от предметной области, но при этом может учитывать аспекты данных, которые часто неуловимы и очень сложно описать с помощью контентно-ориентированных методов [1].

Наиболее распространенный подход к CF основан на моделях соседства. Его первоначальная форма, которую использовали практически все ранние системы CF, ориентирована на пользователей. Такие методы, оценивают неизвестные оценки на основе записанных оценок единомышленников. Позже стал популярен аналогичный подход, ориентированный на товары [1].

Центральным для большинства подходов, ориентированных на товары, является мера сходства между товарами, где s_{ij} обозначает сходство i и j. Часто она основана на коэффициенте корреляции Пирсона. Цель — предсказать ненаблюдаемое значение для пользователя u и товара i. Используя меру сходства, мы идентифицируем k товаров, оцененных пользователем u, которые наиболее похожи на i. Этот набор из k соседей обозначается как $S^k(i;u)$ Предсказанное значение принимается как взвешенное среднее оценок для соседних товаров [1]:

$$\hat{p}_{ui} = \frac{\sum j \in S^k(i;j)^{S_{ij} r_{uj}}}{\sum j \in S^k(i;j)^{S_{ij}}} \quad (1)$$

Нужно найти вектор $x_u \in R^f$ для каждого пользователя u и вектор $y_i \in R^f$ для каждого товара i, которые будут факторизовать предпочтения пользователей. Другими словами, предполагается, что предпочтения являются скалярными произведениями: $r_{ui} = x_u^T y_i$. Эти векторы будут известны как пользовательские факторы и товарные факторы соответственно. Это похоже на методы матричной факторизации, популярные для данных с явной обратной связью. Соответственно, факторы вычисляются путем минимизации следующей функции стоимости:

$$\min_{x,y} \sum_{u,i} c_{ui} \left(r_{ui} - x_u^T y_i \right)^2 + \lambda \left(\sum_u \left\| x_u \right\|^2 + \sum_i \left\| y_i \right\|^2 \right) \quad (2)$$

Оптимальные параметры можно найти с помощью градиентного спуска, но есть более быстрые и надёжные способы. Если мысленно заморозить параметры, соответствующие латентным факторам пользователей, задача оптимизации латентных представлений объектов сведётся к задаче наименьших квадратов, для которой мы знаем точное решение [2].

Итоговый процесс оптимизации функции потерь будет иметь следующий вид.

В цикле до сходимости:

1. Фиксируем матрицу X (скрытые представления пользователей);
2. Решаем задачу L2-регуляризованной регрессии для каждого товара и находим оптимальную матрицу Y;
3. Фиксируем матрицу Y (скрытые представления объектов);
4. Решаем задачу L2-регуляризованной регрессии для каждого пользователя и находим оптимальную матрицу X.

Решение, получаемое путём попеременного вычисления точных аналитических решений, обычно точнее тех, что получаются с помощью наивного градиентного спуска. Более того, данное решение имеет эффективную реализацию, позволяющую использовать преимущества параллельных вычислений [2].

Преобразуя формулу (2), получим:

$$x_u^* = (\sum_{i:(u,i)\in R} y_j y_j^T + \lambda C_i I)^{-1} (\sum_{j:(i,j)\in R} r_{ij} y_j) \quad (3)$$

Таким образом, мы получили аналитическое выражение для вычисления каждого x_u на шаге алгоритма. Отметим, что каждый вектор x_u можно вычислить независимо от других x_v. Данное наблюдение позволяет нам использовать всю мощь параллельных вычислений для эффективного решения оптимизационной задачи. Распределив данные так, что на каждой вычислительной машине хранятся все y_i для некоторого подмножества x_u , на одной итерации алгоритма мы можем параллельно вычислить все x_u. На следующей итерации аналогичным образом вычисляем все y_i [2].

Алгоритм

Поставленная задача представляет собой разработку рекомендательной системы на python для онлайн-платформы, специализирующейся на товарных объявлениях. Платформа функционирует как маркетплейс, где пользователи размещают объявления о продаже различных товаров, а другие пользователи могут просматривать эти объявления и совершать различные действия с ними.

В реализации алгоритма для каждого пользователя $u \in U_{test}$ нужно было построить упорядоченный список из K=40 групп товаров $R_u(i_1, i_2, \ldots, y_K)$, которые пользователь вероятнее всего просмотрит, при условии $i \notin H_u$, где H_u — история взаимодействий пользователя u.

Для оценки использовалась метрика Recall@K (Recall@40):

$$Recall@K = \frac{1}{|U|} \sum_{u\in U} \frac{\left(\left|I_u^{rel} \bigcap I_u^{pred@K}\right|\right)}{\left(\left|I_u^{rel}\right|\right)} \quad (4)$$

Где: I_u^{rel} — релевантные группы товаров для пользователя u.

$I_u^{pred@K}$ — топ-K предсказанных групп товаров.

Данная метрика измеряет способность системы обнаруживать и показывать пользователю подходящие группы товаров в пределах первых 40 рекомендаций. Выбор именно @40 обусловлен практическими ограничениями пользовательского интерфейса — исследования показывают, что пользователи редко просматривают больше 40 рекомендаций, особенно на мобильных устройствах.

Для обучения и валидации использовались следующие данные:

1. История взаимодействий — 1,966,247 записей
2. Тестовые пользователи — 92,319 пользователей
3. Уникальных групп товаров — 150,000+

Алгоритм представляет из себя следующие блоки: базовые популярные рекомендации — глобально популярные товарные группы. Персонализированные популярные рекомендации — популярные товары с учетом истории пользователя. Модель коллаборативной фильтрации ALS — для выявления скрытых предпочтений.

Среди необходимых библиотек для работы использовались следующие: Polars для обработки табличных данных, NumPy для численных операций, SciPy для работы с разреженными матрицами и implicit для реализации ALS алгоритма. Важным этапом предобработки данных является фильтрация только контактных событий (просмотры, добавления в корзину, покупки), что повышает релевантность данных для построения рекомендаций.

Для объективной оценки качества модели реализована строгая процедура валидации с временным разделением. Функция выделяет последние 14 дней взаимодействий в качестве тестового набора, исключая при этом пользовательско-товарные пары, присутствовавшие в тренировочных данных, что предотвращает информационную утечку. Основной метрикой качества выбран Recall@40, который измеряет долю релевантных товаров среди 40 рекомендованных. Реализация метрики включает агрегацию по пользователям с последующим усреднением, что обеспечивает устойчивость к разной степени активности пользователей. В процессе оценки система сравнивает три подхода: базовые популярные рекомендации (топ-40 самых популярных товаров для всех пользователей), ALS-рекомендации и предложенный гибридный метод.

Ключевым компонентом системы является модель коллаборативной фильтрации ALS, реализованная через библиотеку implicit. Процесс обучения начинается с преобразования пользовательско-товарных взаимодействий в разреженную матрицу формата CSR, где строки соответствуют пользователям, столбцы — товарам, а значения указывают на факт взаимодействия.

В формате CSR для представления разреженной матрицы используется три массива, Val, Col и Row. Массив Val содержит значения всех ненулевых элементов матрицы, упорядоченных построчно, а в массиве Col хранятся номера столбцов соответствующих элементов. Размер этих массивов равен количеству ненулевых элементов,

nnz. Массив Row размером $n + 1$ содержит смещения от начала матрицы для первых ненулевых элементов в каждой строке, а также общее число ненулевых элементов [3].

Для эффективного маппинга между идентификаторами и индексами создаются словари в обоих направлениях. Модель настраивается с параметрами: 64 латентных фактора, 15 итераций обучения и регуляризация 0.01 для предотвращения переобучения. Алгоритм поочередно оптимизирует пользовательские и товарные эмбеддинги, минимизируя ошибку реконструкции матрицы взаимодействий.

Рекомендательная система реализует многоэтапный процесс генерации предсказаний. На первом этапе создаются улучшенные популярные рекомендации, которые учитывают индивидуальную историю просмотров каждого пользователя — из общего пула популярных товаров исключаются уже просмотренные пользователем позиции. На втором этапе, при наличии достаточного объема данных (более 10,000 взаимодействий и 100 уникальных пользователей), обучается ALS-модель, которая генерирует персонализированные рекомендации на основе латентных предпочтений. Эти два типа рекомендаций объединяются с весовыми коэффициентами (ALS-рекомендации получают двойной вес), после чего для каждого пользователя выбирается топ-40 товаров с наибольшим суммарным скором.

Финальный этап включает гарантирование корректности выходного формата: для каждого пользователя должно быть ровно 40 рекомендаций. При недостатке персонализированных рекомендаций система добирает недостающие позиции из глобально популярных товаров, исключая уже рекомендованные.

Результаты

Основной метрикой качества выбрана Recall@40, измеряющая долю релевантных товаров среди 40 рекомендованных позиций. Проведенная оценка на валидационной выборке, выделенной методом временного разделения (последние 14 дней), показала следующие результаты:

1. Базовые популярные рекомендации продемонстрировали значение Recall@40 = 0.1215. Этот результат достигнут благодаря значительному объему тренировочных данных (1,278,625 записей) и валидационной выборке, охватывающей 56,762 уникальных пользователя.
2. Гибридный подход (ALS + популярные) достиг Recall@40 = 0.1286, что на 5.9 % превышает показатель базовых популярных рекомендаций. Улучшенные популярные рекомендации, в отличие от базовых, исключают из рекомендаций уже просмотренные пользователем товары, что повышает их релевантность.

Анализ внутренней структуры рекомендаций через визуализации подтверждает сбалансированность гибридного подхода.

Распределение типов рекомендаций демонстрирует оптимальное соотношение: 56,4 % популярных рекомендаций обеспечивают стабильность и покрытие, а 43,6 % ALS-рекомендаций обеспечивают персонализацию (рисунок 1). Такое распределение минимизирует проблему холодного старта при сохранении индивидуального подхода для активных пользователей.

Рис. 1. Распределение типов рекомендаций при генерации

Анализ скоров рекомендаций выявил бимодальное распределение с медианой 1.00 и средним 0.66 (рисунок 2). Выраженный пик на максимальном score (1.0) соответствует фиксированному весу популярных рекомендаций, тогда как более плавное распределение ALS-скоров (0–1) отражает вариативность уверенности модели. Такая структура подтверждает консервативность системы в присвоении высоких скоров, что снижает риск рекомендации нерелевантного контента.

Для понимания достигнутых результатов важно отметить, что в подобных задачах рекомендательных систем для e-commerce типичные значения Recall@40 находятся в диапазоне от 0.05 до 0.20 в зависимости от специфики данных и сложности задачи. Полученное значение 0.1286 соответствует среднему уровню эффективности для подобных систем.

Система успешно обработала 1,966,247 записей взаимодействий и сгенерировала персонализированные рекомендации для 92,319 тестовых пользователей, что

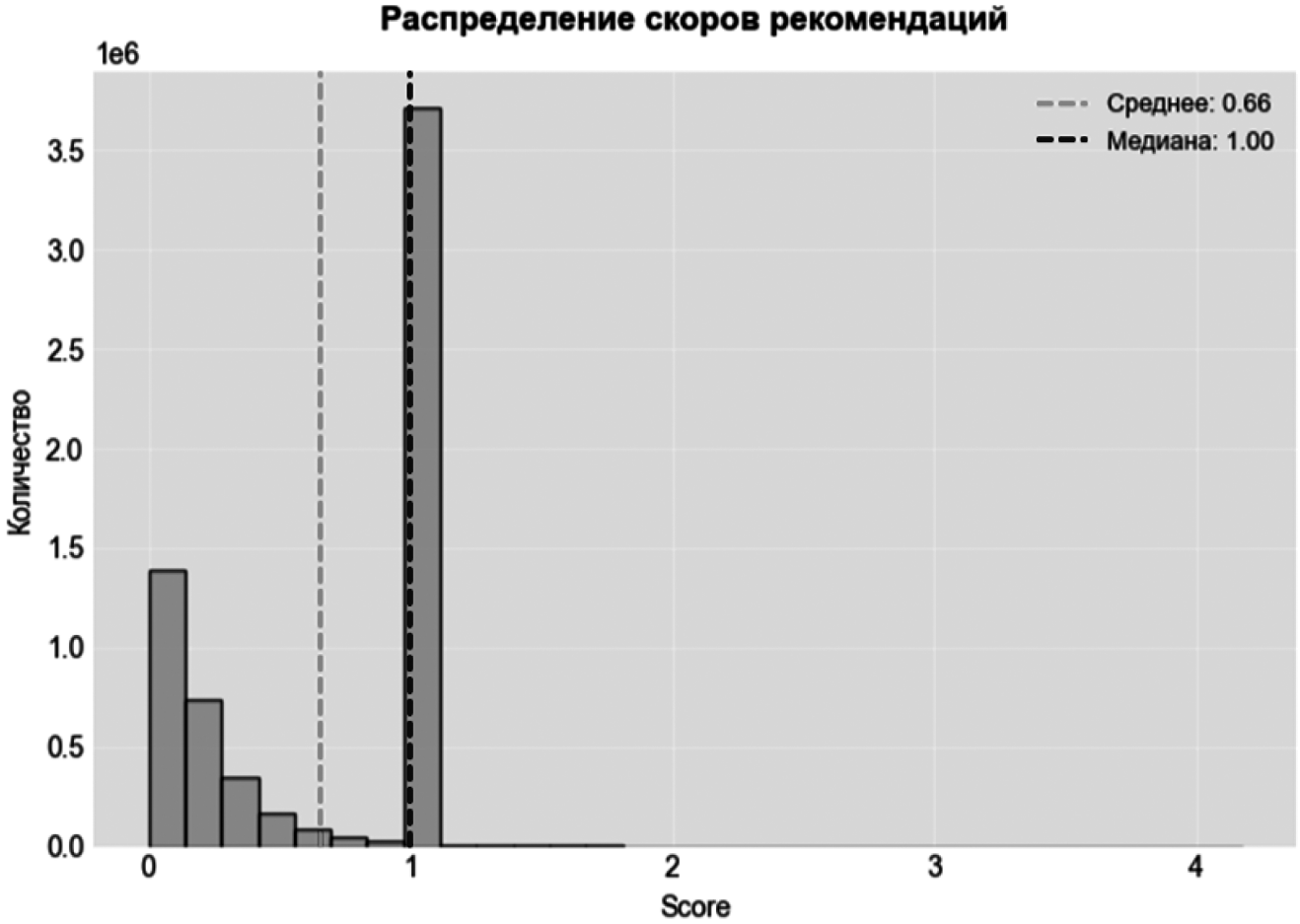

Рис. 2. Распределение скоров рекомендаций при генерации

подтверждает её производственную готовность. Все пользователи получили ровно 40 рекомендаций, соответствующих требуемому формату выходных данных.

Для повышения эффективности системы целесообразно сосредоточиться на следующих аспектах:

1. Устранение проблем обучения ALS-модели и тонкая настройка её гиперпараметров;
2. Эксперименты с альтернативными алгоритмами коллаборативной фильтрации;
3. Интеграция дополнительных признаков товаров и пользователей;
4. Разработка более сложных схем комбинирования различных рекомендательных подходов;
5. Оптимизация весовых коэффициентов при объединении рекомендаций от разных алгоритмов.

Заключение

Разработанная гибридная рекомендательная система представляет собой работоспособное и масштабируемое решение, демонстрирующее стабильную работу на производственных объемах данных. Система успешно сочетает преимущества персонализированных рекомендаций на основе коллаборативной фильтрации с надежностью и простотой популярных рекомендаций.

Полученные результаты подтверждают эффективность предложенного подхода и его практическую применимость в реальных условиях e-commerce. Дальнейшее развитие системы в указанных направлениях позволит достичь более высоких показателей точности рекомендаций при сохранении ключевых преимуществ текущей реализации: отказоустойчивости, масштабируемости и способности обрабатывать большие объемы данных.

ЛИТЕРАТУРА

1. Y. Hu, Y. Koren and C. Volinsky, «Collaborative Filtering for Implicit Feedback Datasets,» 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 263–272, doi: 10.1109/ICDM.2008.22.
2. Рекомендации на основе матричных разложений // Яндекс образование URL: https://education.yandex.ru/handbook/ml/article/rekomendacii-na-osnove-matrichnyh-razlozhenij#alternating-least-squares-als (дата обращения: 09.12.2025).
3. Куприй Р.М., Краснопольский Б.И., Жуков К.А. Оценка эффективности различных форматов представления разреженных матриц для вычислений на графических ускорителях // Параллельные вычислительные технологии — XIX всероссийская конференция с международным участием, ПаВТ'2025, г. Москва. — 2025. — С. 172–185.
4. Sarwar Badrul & Karypis George & Konstan Joseph & Riedl John. (2001). Item-based Collaborative Filtering Recommendation Algorithms. Proceedings of ACM World Wide Web Conference. 1. 10.1145/371920.372071.