НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2025 Том 25 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2025 Vol. 25 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-6-1117-1124
УДК 004.05

Комбинированная модель качества рекомендательных систем

Алексей Михайлович Цыплов[1]✉, Александр Валерьевич Бухановский[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] tsyplov80@mail.ru✉, https://orcid.org/0009-0009-5211-3019
[2] avbukhanovskii@itmo.ru, https://orcid.org/0000-0003-1588-8164

Аннотация

Введение. Рассмотрены подходы к количественной оценке различных эффектов, таких как позиционный сдвиг (Position Bias), сдвиг в сторону популярных объектов (Popularity Bias) и другие, в рекомендательных системах. Предложена новая модель качества рекомендательных систем, которая приводит выбранные метрики к одной единице измерения и определяет для каждого эффекта его влияние на систему. Полученные оценки позволяют проводить более глубокий сравнительный анализ различных систем, а также исследовать поведение системы на разных сегментах пользователей. **Метод.** Для каждой метрики в рамках предложенной модели строится две условные маргинальные плотности распределения: отдельно на релевантных и нерелевантных рекомендациях. На основе сравнения этих плотностей множество возможных значений метрики разделяется на нормальную и критическую. Модель оценивает влияние каждого эффекта на систему на основе частоты попадания значений соответствующей метрики в свою критическую область. **Основные результаты.** Для демонстрации работы модели проведен анализ четырех алгоритмов построения рекомендаций на академическом наборе данных MovieLens-100K. В ходе тестирования оценивались Popularity Bias, отсутствие новизны в рекомендациях и склонность систем рекомендовать объекты исключительно на основе демографических данных пользователей. Для каждого эффекта построена оценка его влияния на систему, приведен пример прогнозирования верхней оценки качества системы в случае устранения соответствующего эффекта. **Обсуждение.** Показано, что метрики таких эффектов, как Popularity Bias или Position Bias, могут менять распределение абсолютных значений в зависимости от рекомендательной системы. Одним из способов более надежно сравнивать разные рекомендательные системы является предложенная модель качества. Модель подходит для оценивания персональных рекомендаций независимо от сферы применения и алгоритма, который был использован для их построения.

Ключевые слова

рекомендательные системы, ранжирование, оценка качества рекомендаций, Popularity Bias, Position Bias, машинное обучение

Ссылка для цитирования: Цыплов А.М., Бухановский А.В. Комбинированная модель качества рекомендательных систем // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 6. С. 1117–1124. doi: 10.17586/2226-1494-2025-25-6-1117-1124

Compound quality model for recommender system evaluation

Aleksei M. Tsyplov[1]✉, Alexander V. Boukhanovsky[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] tsyplov80@mail.ru✉, https://orcid.org/0009-0009-5211-3019
[2] avbukhanovskii@itmo.ru, https://orcid.org/0000-0003-1588-8164

Abstract

The study examines approaches to quantifying various effects, such as Position bias, Popularity Bias, and others, in recommender systems. A new quality model of the recommendation algorithms is proposed which reduces the selected metrics to one unit of measurement and determines its impact on the system for each effect. The obtained scores allow for a deeper comparative analysis of various algorithms as well as investigation the behavior of the algorithm in different

user segments. For each metric, two conditional marginal distribution densities are built within the framework of the model: separately based on relevant and irrelevant recommendations. Based on the comparison of these densities, the set of possible metric values is divided into normal and critical. The model evaluates the impact of each effect on the system based on the frequency of hitting the values of the corresponding metric in its critical area. To demonstrate how the model works, four recommendation algorithms were analyzed on the MovieLens-100K academic dataset. During the testing, Popularity Bias, the lack of novelty in recommendations, and the tendency of algorithms to recommend objects solely based on user demographic data were evaluated. For each effect, an assessment of its impact on the system is constructed, and an example of predicting an upper estimate of the system quality is given if the corresponding effect is eliminated. The study demonstrated that metrics of effects such as Popularity or Position Bias can change the distribution of absolute values depending on the system. One of the ways to compare different recommendation algorithms more reliably is the proposed quality model. The model is suitable for evaluating personal recommendations, regardless of the scope of application and the algorithm that was used to build them.

Keywords
recommendation systems, ranking, evaluation of the quality of recommendations, popularity bias, position bias, machine learning

For citation: Tsyplov A.M., Boukhanovsky A.V. Compound quality model for recommender system evaluation. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 6, pp. 1117–1124 (in Russian). doi: 10.17586/2226-1494-2025-25-6-1117-1124

Введение

На сегодняшний день область применения рекомендательных систем включает в себя множество разнообразных направлений: медиа-системы, социальные сети, онлайн-торговые площадки и т. д. Несмотря на то, что каждое направление обладает своими уникальными особенностями (разные ключевые показатели, формат контента, паттерны поведения пользователей и т. п.), существует ряд общих эффектов, которым могут быть подвержены рекомендательные системы [1, 2]. К таким эффектам можно отнести позиционный сдвиг (Position Bias), склонность к рекомендации популярных объектов (Popularity Bias), усиление цикла обратной связи (Feedback Loop) и другие [3].

Position Bias — эффект, при котором на вероятность взаимодействия пользователя с объектом в большей степени влияет позиция этого объекта в списке рекомендаций, чем его релевантность этому пользователю [1].

Popularity Bias — эффект, при котором на вероятность взаимодействия пользователя с объектом в большей степени влияет популярность этого объекта в списке рекомендаций, чем его релевантность. Под популярностью объекта подразумевается количество взаимодействий с ним за период, предшествующий показу объекта в рекомендациях [4, 5].

Feedback Loop — эффект, который заключается в том, что рекомендательная система со временем «забывает» интересы пользователя, по которым он не дает достаточного количества обратной связи. Иногда этот эффект также называют Selection Bias [6, 7].

Склонность к рекомендации кликбейтов (Clickbait Bias) — эффект, связанный с тем, что пользователи активнее взаимодействуют с объектами с яркими обложками или «кричащими» заголовками, в результате чего история взаимодействий перестает отражать истинные интересы пользователя [1, 6].

Каждый из перечисленных эффектов отдельно рассмотрен в работах [8–10]. Для измерения работы эффекта вводится количественная оценка (обычно такие оценки называют прокси-метриками), а затем предлагается способ ее улучшения. Однако влияние эффектов на рекомендательные системы носит разнонаправленный характер, что затрудняет их совместное устранение [7]. Кроме того, предлагаемые прокси-метрики имеют разные единицы измерения и могут принимать значения из разных диапазонов, что не позволяет использовать их для проведения сравнительного анализа [1]. На данный момент в научных работах отсутствуют универсальные методики измерения, позволяющие не только выявить воздействие различных эффектов, но и сравнить их влияние между собой.

Таким образом, возникает задача построения модели качества рекомендательных систем, которая позволяла бы унифицировать часто используемые метрики и устанавливала бы однозначно интерпретируемую взаимосвязь между этими метриками и качеством рекомендаций. Способ приведения метрик должен работать независимо от диапазона, в котором находятся исходные значения метрик, а также от того, являются они дискретными или непрерывными. Также следует учесть, что значительная часть рассматриваемых метрик вычисляется как среднее значений, рассчитанных для каждого пользователя или объекта рекомендательной системы [1]. В этом случае модель качества должна сохранять возможность рассчитать приведенные метрики для каждого пользователя (или объекта) в отдельности для проведения сравнительного анализа работы рекомендательной системы между различными группами пользователей.

Обзор существующих подходов

Работы, связанные с исследованием характеристик, косвенно влияющих на рекомендательные системы, можно разделить на несколько направлений: исследования, направленные на измерение и устранение конкретных негативных эффектов, обзорные работы с перечислением метрик, а также работы, в которых выполняется сравнительный анализ различных подходов к оценке качества рекомендаций.

В работе [1] приводится 13 различных направлений, по которым можно оценивать рекомендательные системы. Предложено использовать метрики Average Recommendation Popularity (ARP) и Average Coverage of Long Tail items (ACLT) для измерения Popularity

Bias, Discounted Cumulative Gain (DCG) и Click-Through Rate (CTR) — для Position Bias, а для Clickbait Bias строят модель логистической регрессии вовлеченности пользователей по времени взаимодействия с объектом. В исследовании [11] подробно рассмотрена устойчивость системы к Feedback Loop, которая оценивается как способность предсказывать новые предпочтения пользователей.

В работах [8, 12, 13] изучена взаимосвязь между качеством и разнообразием рекомендаций, и предложены подходы, направленные на их увеличение. Низкое разнообразие рекомендаций может свидетельствовать как о сильном воздействии Feedback Loop, так и о высоком Popularity Bias модели [2].

Работы [4, 5, 10, 14, 15] посвящены изучению эффекта Popularity Bias в рекомендательных системах. В [4] для его оценки предлагается использовать частоты попадания в рекомендации среди наиболее популярных объектов, а также применять специальные метрики Popularity Lift (PL) и Miscalibration (MC), основанные на расстоянии Хеллингера между распределениями популярности в рекомендациях и исторических взаимодействиях. В [5] предложено переопределять порядок объектов в списке рекомендаций, смещая популярные объекты в конец списка. В [10] для оценки эффекта применена метрика ARP, а для повышения качества рекомендаций — регуляризация. Работа [15] предлагает подход, позволяющий получить хорошо кластеризуемые эмбеддинги товаров. Это позволяет рекомендовать специализированные непопулярные объекты без снижения качества.

В работе [3] применена коллаборативная информация для предсказания будущих трендов в рекомендациях. Предложенная модель отдает предпочтение объектам, которые будут популярны в будущем, разрывая таким образом Feedback Loop. Для измерения качества прогноза в [3] введены две новых метрики Acceleration (ACC) и Trendiness-Normalized-DCG.

Работы [6, 16] содержат сведения об основных метриках качества рекомендательных систем, которые позволяют получить количественные оценки удовлетворенности пользователей независимо от проявления различных факторов. К ним относятся Hit Rate (HR), Recall, Precision, CTR, Coverage и другие.

В [14, 17] исследовано влияние неравномерного распределения популярности между объектами на качество обучения модели рекомендаций. В [14] сделан акцент на том, как сэмплировать негативные примеры для обучения модели, чтобы нивелировать нежелательные эффекты от неравномерного количества обратной связи, а в [17] для этой цели предсказания модели скорректированы с учетом популярности объектов.

Таким образом, в научных работах представлено множество метрик, позволяющих оценивать рекомендательные системы с различных сторон, однако их сопоставление затруднено из-за разницы в диапазонах значений. Для оценки и интерпретации совокупного влияния различных эффектов можно было бы использовать значения Шепли [18], однако такой подход требует построения объясняющей регрессионной модели, выбор которой будет влиять на результат не меньше, чем сама рекомендательная система [8]. Следовательно, возникает необходимость построения модели качества рекомендательных систем, позволяющей сравнивать прокси-метрики с разными единицами измерения напрямую.

Постановка задачи

Пусть U — множество пользователей и I — множество объектов рекомендательной системы. Рекомендательная система $R = \{R_u\}_{u \in U}$, $R_u \subset I$ предоставляет для каждого пользователя набор персональных рекомендаций. Для каждого элемента $i \in R_u$ вычислен набор прокси-метрик $\boldsymbol{\mu}_{ui} \in \mathrm{M} \subseteq \mathbb{R}^k$, характеризующий данную рекомендацию с точки зрения различных эффектов, и некоторая основная метрика качества $q_{ui} \in Q \subseteq \mathbb{R}$. Единой оценкой качества рекомендаций является значение $\overline{q} \in \overline{Q}$:

$$\overline{q}(R) = \frac{1}{|U|} \sum_{u \in U} \sum_{i \in R_u} q_{ui}.$$

Качество рекомендации q_{ui} обычно вычисляется по некоторой тестовой выборке $S = \{S_u\}_{u \in U}$, где $S_u \subset I$ — объекты, с которыми взаимодействовал пользователь u. На практике часто используют индикаторную функцию множества S_u:

$$q_{ui} = \frac{[i \in S_u]}{|R_u|}, \tag{1}$$

где [] — скобка Айверсона.

Соответствующее таким q_{ui} значение $\overline{q}$ называют уровнем попадания HR [19]. В зависимости от области применения рекомендательной системы формула (1) может быть модифицирована, например, на торговых площадках эту величину умножают на стоимость товара, получая Gross Merchandise Value, а в медиасервисах — на длительность взаимодействия, и называют соответствующее значение Time Spent [16].

Чем больше значение $\overline{q}$, тем лучше рекомендательная система. Множества U и I — конечные, а значит, конечно множество $\mathcal{R} = \{R\}$ всех возможных рекомендательных систем.

$$|\mathcal{R}| = 2^{|U||I|}.$$

Следовательно, для фиксированных U, I, S существует система с наилучшим возможным качеством

$$R^* = \arg\max_{R \in \mathcal{R}} \overline{q}(R) = S.$$

Аналогично основной метрике качества, можно вычислить вектор средних значений $\overline{\boldsymbol{\mu}} \in \overline{\mathrm{M}}$, который характеризует рекомендательную систему с точки зрения исследуемых эффектов.

$$\overline{\boldsymbol{\mu}}(R) = \frac{1}{|U|} \sum_{u \in U} \sum_{i \in R_u} \boldsymbol{\mu}_{ui}.$$

Обычно в работах, связанных с устранением таких эффектов, как Popularity или Position Bias, в экспериментах сравниваются пары $(\overline{\boldsymbol{\mu}}, \overline{q})$ построенных рекомендательных систем, а также эмпирически проверяется, что наилучшему значению прокси-метрики соответствует лучшая рекомендательная система [5, 10, 12].

Прокси-метрики могут иметь разнонаправленный характер, и в общем случае точки экстремума прокси-метрик и основной метрики качества не совпадают [4]. Оптимизация разных прокси-метрик по-разному меняет значение $\bar{q}$, поэтому при работе одновременно с несколькими эффектами выбор прокси-метрики для улучшения играет ключевую роль [2].

Пусть R — некоторая рекомендательная система. Необходимо выявить такой эффект j в системе, устранение которого обеспечит максимальный прирост качества рекомендаций. Пусть $\bar{\mu}^{(j)}$, $j = 1, \ldots, k$ — компоненты вектора $\bar{\boldsymbol{\mu}}$, а $\mathrm{M}_j^* \subseteq \bar{\mathrm{M}}$ — множество всех векторов $\bar{\boldsymbol{\mu}}$, для которых $\bar{\mu}^{(j)}$ принимает наилучшее возможное значение. Тогда:

$$\sup_{R' \in \mathcal{R} | \bar{\boldsymbol{\mu}}(R) \in \mathrm{M}_j^*} (\bar{q}(R') - \bar{q}(R)) \to \max_j. \qquad (2)$$

Рассмотрим этапы решения задачи (2).

Этап 1. Построение проекции P: $\mathrm{M} \to \mathrm{M}'$ пространства метрик M в пространство M′, где все размерности измеряются по одной шкале, у которой наилучшее значение равно нулю, а наихудшее или 1, или $+\infty$, и зависимость между $\bar{\boldsymbol{\mu}}(R)$ и $\sup(\bar{q}(R))$ монотонная. Проекция будет построена на основе некоторой выборки $(\boldsymbol{\mu}_{ui}, q_{ui})$.

Этап 2. Построение аддитивного разложения вида

$$\bar{q}(R) = 1 - \varepsilon - \sum_{j=1}^{k} w_j,$$

где w_j характеризует влияние эффекта j на качество рекомендаций, а ε — случайная ошибка системы. Если j-й эффект не влияет на систему, то $w_j = 0$. Соответственно, для оценки прироста качества после устранения j-го эффекта можно воспользоваться равенством

$$\sup_{R' \in \mathcal{R} | \bar{\boldsymbol{\mu}}(R) \in \mathrm{M}_j^*} (\bar{q}(R') - \bar{q}(R)) = w_j.$$

Таким образом (2) сводится к нахождению наибольшего w_j в аддитивном разложении. Такая задача построения модели качества рекомендаций ранее не рассматривалась в известных работах и представляет собой научную новизну.

Модель

В большинстве метрик качества рекомендательных систем ключевую роль играет индикаторная функция $[i \in S_u]$. Пусть p — вероятность того, что случайная рекомендация (u, i), сделанная исследуемой рекомендательной системой, окажется релевантной. Пусть $\gamma \sim Ber(p)$ — соответствующая ей случайная величина Бернулли. На вероятность p оказывают влияние прокси-метрики $\boldsymbol{\mu}$, которые в свою очередь также являются случайными величинами. В таком случае γ имеет комбинированное распределение

$$p = \mathcal{P}(\gamma = 1) = \int_{\mathrm{M}} \mathcal{P}(\gamma = 1 | \boldsymbol{\mu}) \mathcal{P}(\boldsymbol{\mu}) d\boldsymbol{\mu}$$

и, соответственно,

$$1 - p = \int_{\mathrm{M}} \mathcal{P}(\gamma = 0 | \boldsymbol{\mu}) \mathcal{P}(\boldsymbol{\mu}) d\boldsymbol{\mu}, \qquad (3)$$

где $\mathcal{P}$ — вероятность; M — множество всех возможных значений прокси-метрик. В последующих шагах интеграл в правой части (3) будет разбит на составляющие w_j и ε.

Выполним этап 1 модели — построим проекцию P: $\mathrm{M} \to \mathrm{M}'$. Для каждой компоненты $\mu^{(j)}$ вектора $\boldsymbol{\mu}$ определим две маргинальные условные плотности распределения: $\mathrm{pdf}(\mu^{(j)} | \gamma = 1)$ и $\mathrm{pdf}(\mu^{(j)} | \gamma = 0)$. Сравнивая значения данных плотностей, множество $\mathrm{M}^{(j)}$ всех возможных значений j-ой метрики разделим на множество нормальных значений $\mathrm{M}_1^{(j)}$ и критических $\mathrm{M}_0^{(j)}$.

$$\mathrm{M}_1^{(j)} = \{\mu^{(j)} \in \mathrm{M}^{(j)} | \mathrm{pdf}(\mu^{(j)} | \gamma = 1) > \mathrm{pdf}(\mu^{(j)} | \gamma = 0)\},$$
$$\mathrm{M}_0^{(j)} = \mathrm{M}^{(j)} \backslash \mathrm{M}_1^{(j)}.$$

На рисунке представлена наглядная визуализация такого разбиения.

После того, как разбиение построено для каждого признака, в каждой точке M можно определить, сколько признаков принимают критические значения. Обозначим соответствующий счетчик Count Critical (CC).

$$CC(\boldsymbol{\mu}) = \sum_{j=1}^{k} [\mu^{(j)} \in \mathrm{M}_0^{(j)}].$$

Разобьем M на M_0 и $\mathrm{M}_{>0}$ в зависимости от того, какие значения принимает функция CC.

$$\mathrm{M}_0 = \{\boldsymbol{\mu} \in \mathrm{M} | CC(\boldsymbol{\mu}) = 0\},$$
$$\mathrm{M}_{>0} = \{\boldsymbol{\mu} \in \mathrm{M} | CC(\boldsymbol{\mu}) > 0\}. \qquad (4)$$

Для элементов множества $\mathrm{M}_{>0}$ построим проекцию P, используя функцию CC.

$$P^{(j)}(\boldsymbol{\mu}_{ui}) = \frac{[\mu_{ui}^{(j)} \in \mathrm{M}_0^{(j)}]}{CC(\boldsymbol{\mu}_{ui})} (1 - [i \in S_u]). \qquad (5)$$

Перейдем к этапу 2 модели. Проекция P имеет следующий смысл. Если хотя бы одна из метрик принимает критическое значение, то $\mathcal{P}(\gamma = 0 | \boldsymbol{\mu}) = 1$, причем каждая из метрик равновероятно является причиной

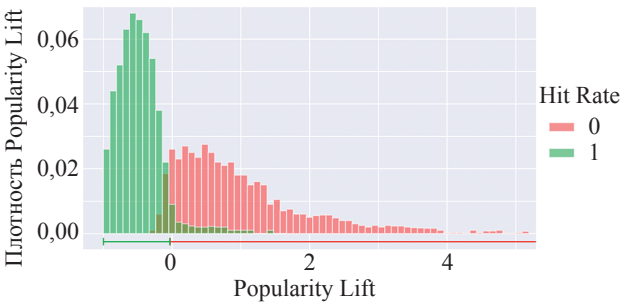

Рисунок. Гистограммы условных распределений Popularity Lift некоторой рекомендательной системы, построенной для набора данных MovieLens100K

Figure. Histograms of the conditional Popularity Lift distribution of a recommender system built for the MovieLens100K dataset

нерелевантности. Если все метрики принимают нормальное значение, то $\mathcal{P}(\gamma = 0|\boldsymbol{\mu}) = 0$. Тогда $\forall \boldsymbol{\mu} \neq 0$

$$\mathcal{P}(\gamma = 0|\boldsymbol{\mu}) = \sum_{i=1}^{k} P^{(j)}(\boldsymbol{\mu}), \tag{6}$$

и интеграл (3) можно разделить по множествам (4).

$$1 - p = \int_{M_0} \mathcal{P}(\gamma = 0|\boldsymbol{\mu}) \times \\ \times \mathcal{P}(\boldsymbol{\mu}) d\boldsymbol{\mu} + \int_{M_{>0}} \mathcal{P}(\gamma = 0|\boldsymbol{\mu}) \mathcal{P}(\boldsymbol{\mu}) d\boldsymbol{\mu}. \tag{7}$$

Первое слагаемое в правой части выражения (7) соответствует ситуации, когда рекомендация оказалось нерелевантна, а все прокси-метрики приняли нормальное значение. Вероятность данного события легко можно оценить по выборке, в рамках модели качества она обозначена ε и отражает долю неудачных рекомендаций, которые нельзя объяснить с точки зрения выбранных прокси-метрик. Второе слагаемое (7) можно разложить в сумму, используя (6).

$$1 - p = \varepsilon + \sum_{i=1}^{k} \int_{M_{>0}} P^{(j)}(\boldsymbol{\mu}) \mathcal{P}(\boldsymbol{\mu}) d\boldsymbol{\mu}. \tag{8}$$

Равенство (8) определяет модель качества рекомендательной системы. Внедрение метода, направленного на устранение j-го эффекта, позволит устранить критические значения в соответствующей метрике. В результате качество рекомендаций возрастет не более, чем на

$$w_j = \int_{M_{>0}} P^{(j)}(\boldsymbol{\mu}) \mathcal{P}(\boldsymbol{\mu}) d\boldsymbol{\mu}.$$

Таким образом, решением задачи (2) является пара (j, w_j) с наибольшим w_j.

Объекты экспериментов

Для демонстрации работы модели качества для набора данных MovieLens-100K были построены и проанализированы четыре рекомендательные системы: Top by CTR, Top by CTR within Cohort, Item KNN и MF-BPR.

Набор данных MovieLens-100K содержит информацию 100 тыс. оценок 943 пользователей 1682 фильмам. По каждому пользователю даны демографические данные: пол, возраст, тип занятости, а также не менее 20 его оценок. Для построения выбранных рекомендательных систем данные были разделены на обучающую и тестовую выборки таким образом, чтобы в тестовой выборке оказались последние 10 оценок каждого пользователя. Для всех пользователей подготовлено по 10 рекомендаций.

Top by CTR — рекомендательная система, в которой все фильмы сортируются по убыванию CTR, т. е., отношения количества оценок к количеству пользователей, рассчитанному на обучающей выборке. Каждому пользователю предлагаются фильмы, которые имеют наибольшую оценку кликов (CTR) среди тех, которые он еще не оценивал.

Top by CTR within Cohort использует такую же механику, что и Top by CTR, с небольшим усложнением. Все пользователи разбиваются на группы по полу и возрасту. Чтобы групп было не слишком много, вместо реального возраста используются возрастные группы, составленные на основе децилей возрастов. Внутри каждой группы реализуется система Top by CTR.

Item KNN — рекомендательная система, которая для каждой пары фильмов считает количество пользователей, посмотревших их.

$$\mathrm{sim}(i, i') = \frac{1}{|U|} \sum_{u \in U} [i \in S_u \wedge i' \in S_u].$$

Для оценки релевантности фильма i пользователю u вычисляется суммарное сходство между i и всеми фильмами S_u.

$$\mathrm{rel}(u, i) = \sum_{i' \in S_u} \mathrm{sim}(i, i').$$

Фильмы сортируются по убыванию $\mathrm{rel}(u, i)$, из списка удаляются фильмы, оцененные пользователем ранее.

Matrix Factorization, Bayesian Pairwise Ranking (MF-BPR) — рекомендательная система, в которой каждому пользователю и каждому фильму ставятся в соответствие эмбеддинги e_u, e_i. Релевантность оценивается как скалярное произведение эмбеддингов. Модель обучалась по методологии BPR, при которой для каждого взаимодействия i пользователя u из дискретного равномерного распределения выбирается случайный фильм i', который пользователь не оценивал, а функция потерь вычисляется по формуле

$$\mathrm{loss} = \log_2 \sigma(e_u^{\top}(e_i - e_{i'})).$$

Модель была обучена на 40 эпохах методом Adagrad.

Эксперименты

Каждая рекомендательная система оценивалась по четырем метрикам.

HR для оценки общего качества рекомендаций (1).

PL для оценки влияния Popularity Bias [4].

Novelty Lack (NL) для оценки эффекта недостатка новизны, при котором рекомендательная система предпочитает рекомендовать старые объекты новым. NL рассчитывается так же, как PL, только вместо популярности используется возраст фильма [1].

Коэффициент ранговой корреляции Спирмена между предсказаниями рекомендательной системы и CTR внутри социально-демографической группы (SCbSaCP) [6, 20, 21].

Результаты расчетов по всем пользователям представлены в табл. 1.

С точки зрения абсолютных значений метрик алгоритмы показали предсказуемые результаты. Наилучшее качество продемонстрировала самая сложная система, самый высокий PL показала Top By CTR, основанная на популярности объектов, а самый высокий SCbSaCP — система Top by CTR within Cohort. Применим модель качества к полученным значениям метрик, чтобы оценить влияние соответствующих эффектов на системы (табл. 2).

Таблица 1. Абсолютные значения метрик
Table 1. The absolute scores of the metrics

Рекомендательная система	HR	PL	SCbSaCP	NL
Top By CTR	0,0768	**2,59**	0,55	2,05
Top By CTR within Cohort	0,0721	2,22	**1,00**	**2,47**
Item KNN	0,0965	2,31	0,52	2,10
MF-BPR	**0,1047**	1,85	0,48	1,24

Примечание: жирным шрифтом выделены строки с наибольшими значениями метрик.

Таблица 2. Результат применения комбинированной модели качества
Table 2. The compound quality model projection

Рекомендательная система	p	$P^{(PL)}(\boldsymbol{\mu})$	$P^{(SCbSaCP)}(\boldsymbol{\mu})$	$P^{(NL)}(\boldsymbol{\mu})$	ε
Top By CTR	0,0768	0,0161	0,0454	0,0502	0,8115
Top By CTR within Cohort	0,0721	0,0116	0,0000	0,0469	**0,8694**
Item KNN	0,0965	**0,0212**	**0,0523**	0,0640	0,7660
MF-BPR	**0,1047**	0,0192	0,0380	**0,0684**	0,7697

Примечание: жирным шрифтом выделены строки с наибольшими значениями.

В каждой строке табл. 2 сумма значений равна единице. Результаты проекции значительно отличаются от исходных значений метрик, приведенных в табл. 1. Например, влияние Popularity Bias на систему оказалось наибольшим у системы Item KNN, а не у Top By CTR. Действительно, система Top By CTR часто ошибается из-за того, что предлагает пользователям слишком популярные объекты, но и релевантные рекомендации он подбирает по той же причине. Следовательно, данную систему нельзя улучшить, за счет снижения популярности в рекомендациях. Зато значение $P^{(NL)}(\boldsymbol{\mu}) = 0{,}0502$ указывает на возможность улучшения системы Top By CT за счет удаления старых фильмов из рекомендаций. Верхняя граница p для Top By CTR в случае оптимизации NL составляет 0,0768 + 0,0502 = = 0,1270.

Аналогичные выводы можно сделать и относительно модели MF-BPR. С точки зрения абсолютных значений может показаться, что Popularity Bias имеет более сильное влияние на качество, чем NL. Но на самом деле такое различие объясняется тем, что у PL и NL отличаются выборочные распределения, и в действительности недостаток новизны имеет большее влияние. У моделей Item KNN и MF-BPR меньше ε, что говорит о том, что остальные метрики сильнее разделяются на нормальные и критические значения.

Для более детального анализа все пользователи были разделены на четыре группы (сегмента) на основе квартилей по количеству оценок на обучающей выборке. Первая группа — самые активные пользователи, четвертая группа — наименее активные пользователи. Сравним абсолютные значения метрик и проекцию (5) на примере модели Item KNN (табл. 3).

У активных пользователей высокое абсолютное значение PL и NL. Это объясняется тем, что большая часть оценок новым и популярным фильмам попала в обучающую выборку, а старые и непопулярные фильмы пользователи стали смотреть в последнюю очередь, именно они оказались в тестовой выборке. Для малоактивных пользователей новые популярные фильмы оказались в тестовой выборке, поэтому PL и NL для них имеют более кучный характер. Как следствие, такие пользователи оказываются более чувствительными к исследуемым эффектам и для них можно значительно улучшить рекомендации.

Обсуждение

Рассмотрим подробнее, чем отличается предложенная модель качества от использования прокси-метрик.

Исходные значения метрик необходимы для описания рекомендательной системы с различных сторон.

Таблица 3. Метрики Item KNN для разных сегментов пользователей
Table 3. Item KNN scores for different user segments

Сегмент	HR	PL	SCbSaCP	NL	$P^{(PL)}(\boldsymbol{\mu})$	$P^{(SCbSaCP)}(\boldsymbol{\mu})$	$P^{(NL)}(\boldsymbol{\mu})$	ε
1	0,0615	**3,1137**	0,4930	**3,3596**	0,0141	0,0333	0,0350	**0,8560**
2	0,0853	2,4510	**0,5412**	1,4872	0,0151	0,0440	0,0509	0,8047
3	0,1000	1,9052	0,5280	1,9001	0,0212	0,0511	0,0654	0,7623
4	**0,1370**	1,7684	0,5101	1,6500	**0,0337**	**0,0793**	**0,1028**	0,6472

Примечание: жирным шрифтом выделены строки с наибольшими значениями.

Если одна из метрик близка к экстремуму, это помогает быстро выявить влияние соответствующего эффекта. Прокси-метрики, как правило, разрабатываются таким образом, чтобы их единицы измерения имели четкую интерпретацию. Тем не менее, при значениях, удаленных от экстремальных, оценивание влияния эффекта становится затруднительным. Кроме того, различие величин измерения у разных прокси-метрик делают невозможным их сравнение между собой.

Предложенная модель качества унифицирует прокси-метрики, что позволяет решить проблему их сопоставления. Кроме того, полученные значения вклада каждого эффекта напрямую связаны с количеством ошибочных рекомендаций, что позволяет быстро оценить прирост качества рекомендательной системы в случае его устранения.

Таким образом, совместное использование прокси-метрик и комбинированной модели качества позволяют всесторонне описать работу рекомендательной системы.

Заключение

В работе предложена новая модель, позволяющая численно измерить взаимосвязь между качеством рекомендательной системы и ее произвольными прокси-метриками. Модель названа комбинированной, поскольку в ее основе лежит предположение о том, что факт успешной рекомендации — случайная величина с комбинированным распределением. Разработан алгоритм, позволяющий численно оценивать вклад различных факторов в вероятность ошибки рекомендательной системы. Знание таких вкладов позволяет определить список наиболее важных параметров, которые следует учитывать при дальнейшей разработке системы. Данный подход позволяет сравнивать качество рекомендательной системы на различных сегментах пользователей и сравнивать поведение различных рекомендательных систем между собой.

Литература

1. Anderson A., Maystre L., Anderson I., Mehrotra R., Lalmas M. Algorithmic effects on the diversity of consumption on spotify // Proc. of the Web Conference. 2020. P. 2155–2165. https://doi.org/10.1145/3366423.3380281
2. Avazpour I., Pitakrat T., Grunske L., Grundy J. Dimensions and metrics for evaluating recommendation systems // Recommendation Systems in Software Engineering. 2014. P. 245–273. https://doi.org/10.1007/978-3-642-45135-5_10
3. Ding H., Kveton B., Ma Y., Park Y., Kini V., Gu Y., et al. Trending now: modeling trend recommendations // Proc. of the 17th ACM Conference on Recommender Systems. 2023. P. 294–305. https://doi.org/10.1145/3604915.3608810
4. Cai Y., Guo J., Fan Y., Ai Q., Zhang R., Cheng X. Hard negatives or false negatives: correcting pooling bias in training neural ranking models // Proc. of the 31st ACM International Conference on Information and Knowledge Management. 2022. P. 118–127. https://doi.org/10.1145/3511808.3557343
5. Abdollahpouri H., Mansoury M., Burke R., Mobasher B. The connection between popularity bias, calibration, and fairness in recommendation // Proc. of the 14th ACM Conference on Recommender Systems. 2020. P. 726–731. https://doi.org/10.1145/3383313.3418487
6. Beel J., Langer S., Genzmehr M., Gipp B., Breitinger C., Nürnberger A. Research paper recommender system evaluation: a quantitative literature survey // Proc. of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation. 2013. P. 15–22. https://doi.org/10.1145/2532508.2532512
7. Wasilewski J., Hurley N. Incorporating diversity in a learning to rank recommender system // Proc. of the 29th International Florida Artificial Intelligence Research Society Conference. 2016. P. 1–6.
8. Ricci F., Rokach L., Shapira B. Recommender Systems Handbook. Springer, 2010. 842 p.
9. Said A., Bellogin A. Comparative recommender system evaluation: benchmarking recommendation frameworks // Proc. of the 8th ACM Conference on Recommender Systems. 2014. P. 129–136. https://doi.org/10.1145/2645710.2645746
10. Wilhelm M., Ramanathan A., Bonomo A., Jain S., Chi E.H., Gillenwater J. Practical diversified recommendations on YouTube with determinantal point processes // Proc. of the 27th ACM International Conference on Information and Knowledge Management. 2018. P. 2165–2173. https://doi.org/10.1145/3269206.3272018
11. Chang Bo, Meng C., Ma H., Chang S., Gu Y., Peng Y., et al. Cluster anchor regularization to alleviate popularity bias in recommender systems // Proc. of the Companion Proceedings of the ACM Web Conference. 2024. P. 151–160. https://doi.org/10.1145/3589335.3648312

References

1. Anderson A., Maystre L., Anderson I., Mehrotra R., Lalmas M. Algorithmic effects on the diversity of consumption on spotify. *Proc. of the Web Conference*, 2020, pp. 2155–2165. https://doi.org/10.1145/3366423.3380281
2. Avazpour I., Pitakrat T., Grunske L., Grundy J. Dimensions and metrics for evaluating recommendation systems. *Recommendation Systems in Software Engineering*, 2014, pp. 245–273. https://doi.org/10.1007/978-3-642-45135-5_10
3. Ding H., Kveton B., Ma Y., Park Y., Kini V., Gu Y., et al. Trending now: modeling trend recommendations. *Proc. of the 17th ACM Conference on Recommender Systems*, 2023, pp. 294–305. https://doi.org/10.1145/3604915.3608810
4. Cai Y., Guo J., Fan Y., Ai Q., Zhang R., Cheng X. Hard negatives or false negatives: correcting pooling bias in training neural ranking models. *Proc. of the 31st ACM International Conference on Information and Knowledge Management*, 2022, pp. 118–127. https://doi.org/10.1145/3511808.3557343
5. Abdollahpouri H., Mansoury M., Burke R., Mobasher B. The connection between popularity bias, calibration, and fairness in recommendation. *Proc. of the 14th ACM Conference on Recommender Systems*, 2020, pp. 726–731. https://doi.org/10.1145/3383313.3418487
6. Beel J., Langer S., Genzmehr M., Gipp B., Breitinger C., Nürnberger A. Research paper recommender system evaluation: a quantitative literature survey. *Proc. of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, 2013, pp. 15–22. https://doi.org/10.1145/2532508.2532512
7. Wasilewski J., Hurley N. Incorporating diversity in a learning to rank recommender system. *Proc. of the 29th International Florida Artificial Intelligence Research Society Conference*, 2016, pp. 1–6.
8. Ricci F., Rokach L, Shapira B. *Recommender Systems Handbook*. Springer, 2010, 842 p.
9. Said A., Bellogin A. Comparative recommender system evaluation: benchmarking recommendation frameworks. *Proc. of the 8th ACM Conference on Recommender Systems*, 2014, pp. 129–136. https://doi.org/10.1145/2645710.2645746
10. Wilhelm M., Ramanathan A., Bonomo A., Jain S., Chi E.H., Gillenwater J. Practical diversified recommendations on YouTube with determinantal point processes. *Proc. of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 2165–2173. https://doi.org/10.1145/3269206.3272018
11. Chang Bo, Meng C., Ma H., Chang S., Gu Y., Peng Y., et al. Cluster anchor regularization to alleviate popularity bias in recommender systems. *Proc. of the Companion Proceedings of the ACM Web Conference*, 2024, pp. 151–160. https://doi.org/10.1145/3589335.3648312

12. Bellogin A., Castells P., Cantador I. Precision-oriented evaluation of recommender systems: an algorithmic comparison // Proc. of the 5th ACM Conference on Recommender Systems. 2011. P. 333–336. https://doi.org/10.1145/2043932.2043996
13. Cremonesi P., Koren Y., Turrin R. Performance of recommender algorithms on top-n recommendation tasks // Proc. of the 4th ACM Conference on Recommender Systems. 2010. P. 39–46. https://doi.org/10.1145/1864708.1864721
14. Abdollahpouri H., Burke R., Mobasher B. Managing popularity bias in recommender systems with personalized re-ranking // Proc. of the 32nd International Florida Artificial Intelligence Research Society Conference. 2019. P. 1–6.
15. Yi X., Yang J., Hong L., Cheng D.Z., Heldt L., Kumthekar A., Zhao Z., Wei L., Chi E. Sampling-bias-corrected neural modeling for large corpus item recommendations // Proc. of the 13th ACM Conference on Recommender Systems. 2019. P. 269–277. https://doi.org/10.1145/3298689.3346996
16. Silveira T., Zhang M., Lin X., Liu Y., Ma S. How good your recommender system is? A survey on evaluations in recommendation // International Journal of Machine Learning and Cybernetics. 2019. V. 10. N 5. P. 813–831. https://doi.org/10.1007/s13042-017-0762-9
17. Akiyama T., Obara K., Tanizaki M. Proposal and evaluation of serendipitous recommendation method using general unexpectedness // CEUR Workshop Proceedings. 2010. V. 676. P. 3–10.
18. Scott L.M., Su-In L. A unified approach to interpreting model predictions // Proc. of the 31st Conference on Neural Information Processing Systems. 2017. P. 1–10.
19. Isinkaye F.O., Folajimi Y.O., Ojokoh B.A. Recommendation systems: principles, methods and evaluation // Egyptian Informatics Journal. 2015. V. 16. N 3. P. 261–273. https://doi.org/10.1016/j.eij.2015.06.005
20. Rhee W., Cho S.-M., Suh B. Countering popularity bias by regularizing score differences // Proc. of the 16th ACM Conference on Recommender Systems. 2022. P. 145–155. https://doi.org/10.1145/3523227.3546757
21. Shani G., Gunawardana A. Evaluating recommendation systems // Recommender Systems Handbook. 2010. P. 257–297. https://doi.org/10.1007/978-0-387-85820-3_8

12. Bellogin A., Castells P., Cantador I. Precision-oriented evaluation of recommender systems: an algorithmic comparison. *Proc. of the 5th ACM Conference on Recommender Systems*, 2011, pp. 333–336. https://doi.org/10.1145/2043932.2043996
13. Cremonesi P., Koren Y., Turrin R. Performance of recommender algorithms on top-n recommendation tasks. *Proc. of the 4th ACM Conference on Recommender Systems*, 2010, pp. 39–46. https://doi.org/10.1145/1864708.1864721
14. Abdollahpouri H., Burke R., Mobasher B. Managing popularity bias in recommender systems with personalized re-ranking. *Proc. of the 32nd International Florida Artificial Intelligence Research Society Conference*, 2019, pp. 1–6.
15. Yi X., Yang J., Hong L., Cheng D.Z., Heldt L., Kumthekar A., Zhao Z., Wei L., Chi E. Sampling-bias-corrected neural modeling for large corpus item recommendations. *Proc. of the 13th ACM Conference on Recommender Systems*, 2019, pp. 269–277. https://doi.org/10.1145/3298689.3346996
16. Silveira T., Zhang M., Lin X., Liu Y., Ma S. How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 2019, vol. 10, no. 5, pp. 813–831. https://doi.org/10.1007/s13042-017-0762-9
17. Akiyama T., Obara K., Tanizaki M. Proposal and evaluation of serendipitous recommendation method using general unexpectedness. *CEUR Workshop Proceedings*, 2010, vol. 676, pp. 3–10.
18. Scott L.M., Su-In L. A unified approach to interpreting model predictions. *Proc. of the 31st Conference on Neural Information Processing Systems*, 2017, pp. 1–10.
19. Isinkaye F.O., Folajimi Y.O., Ojokoh B.A. Recommendation systems: principles, methods and evaluation. *Egyptian Informatics Journal*, 2015, vol. 16, no. 3, pp. 261–273. https://doi.org/10.1016/j.eij.2015.06.005
20. Rhee W., Cho S.-M., Suh B. Countering popularity bias by regularizing score differences. *Proc. of the 16th ACM Conference on Recommender Systems*, 2022, pp. 145–155. https://doi.org/10.1145/3523227.3546757
21. Shani G., Gunawardana A. Evaluating recommendation systems. *Recommender Systems Handbook*, 2010, pp. 257–297. https://doi.org/10.1007/978-0-387-85820-3_8

Авторы

Цыплов Алексей Михайлович — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0009-5211-3019, tsyplov80@mail.ru

Бухановский Александр Валерьевич — доктор технических наук, профессор, директор мегафакультета трансляционных информационных технологий, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 6603474810, https://orcid.org/0000-0003-1588-8164, avbukhanovskii@itmo.ru

Authors

Aleksei M. Tsyplov — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0009-5211-3019, tsyplov80@mail.ru

Alexander V. Boukhanovsky — D.Sc., Professor, Head of the School of Translational Information Technologies, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 6603474810, https://orcid.org/0000-0003-1588-8164, avbukhanovskii@itmo.ru

Статья поступила в редакцию 08.08.2025
Одобрена после рецензирования 17.09.2025
Принята к печати 24.11.2025

Received 08.08.2025
Approved after reviewing 17.09.2025
Accepted 24.11.2025