Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ
КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

КУРСОВАЯ РАБОТА

НА ТЕМУ:

Определение критериев оценки качества работы рекомендательных систем

Студент ИУ5-34М (группа) _____ (подпись, дата) В.А. Андреев (И.О. Фамилия)

Руководитель курсовой работы _____ (подпись, дата) Ю.Е. Гапанюк (И.О. Фамилия)

2025 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ5
(индекс)
В.И. Терехов
(И.О. Фамилия)
(подпись)
(дата)

З А Д А Н И Е
на выполнение курсовой работы

по дисциплине НИР по обработке и анализу данных
Студент группы ИУ5-34М Андреев Виктор Алексеевич
(Фамилия, имя, отчество)

Тема курсовой работы
Определение критериев оценки качества работы рекомендательных систем

Направленность КР (учебная, исследовательская, практическая, производственная, др.)
УЧЕБНАЯ

Источник тематики (кафедра, предприятие, НИР) КАФЕДРА

Задание

Систематизировать критерии оценки качества РС, установить их взаимосвязи с методами реализации, а также провести анализ современных практических и этических аспектов оценки.

Оформление курсовой работы:

Расчетно-пояснительная записка (Отчет по КР) на _____ листах формата А4.

Дата выдачи задания « _____ » _____ 202 ___ г.

Студент _____ В.А. Андреев
(подпись, дата) (И.О. Фамилия)

Руководитель курсовой работы _____ Ю.Е. Гапанюк
(подпись, дата) (И.О. Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Содержание

Введение

В условиях экспоненциального роста объемов информации рекомендательные системы (РС) стали неотъемлемым инструментом навигации в цифровом пространстве. Их экономическое влияние огромно: по данным исследований, около 80% времени просмотра на платформе Netflix генерируется именно рекомендациями [1]. Несмотря на широкое распространение, в научном сообществе до сих пор отсутствует унифицированный подход к оценке их эффективности, что порождает серьезную научную проблему.

Ключевой аспект этой проблемы – методологический разрыв между результатами оффлайн-оценки, проводимой на исторических данных, и реальной производительностью систем в онлайн-среде [2]. Высокие показатели традиционных метрик точности не гарантируют достижения бизнес-целей или повышения удовлетворенности пользователей. Это приводит к риску неоптимального распределения ресурсов при разработке и невозможности адекватного сравнения результатов различных исследований. Таким образом, выбор критериев оценки должен быть не произвольным, а строго систематизированным и обоснованным с учетом как целей, так и архитектуры конкретной системы.

Целью данного исследования является систематизация критериев оценки качества РС, установление их взаимосвязи с методами реализации, а также анализ современных практических и этических аспектов оценки.

1. Методы реализации рекомендательных систем

1.1. Коллаборативная фильтрация (Collaborative Filtering)

Коллаборативная фильтрация (CF) основана на анализе коллективного поведения пользователей. Современные реализации CF часто используют методы матричной факторизации (например, сингулярное разложение – SVD), которые позволяют выявлять скрытые, или латентные, факторы в предпочтениях пользователей [3]. Суть метода заключается в представлении матрицы «пользователь-товар» в виде произведения двух матриц меньшей размерности, которые отражают скрытые предпочтения пользователей и скрытые характеристики товаров. Преимуществом метода является способность находить неочевидные рекомендации, однако он страдает от проблемы «холодного старта» для новых пользователей и объектов, для которых отсутствует история взаимодействий.

1.2. Контентная фильтрация (Content-Based Filtering)

Контентная фильтрация анализирует атрибуты самих объектов. Процесс включает два основных этапа:

1. Извлечение признаков (Feature extraction):

   Описание каждого объекта представляется в виде вектора его характеристик (например, жанр и актеры для фильма, ключевые слова для статьи).

2. Построение профиля пользователя:

   На основе объектов, которые пользователь оценил положительно в прошлом, строится агрегированный профиль его интересов.

Данный подход решает проблему «холодного старта» для новых объектов, но несет в себе риск создания «пузыря фильтров» и однообразия рекомендаций [4].

1.3. Гибридные системы

Гибридные системы комбинируют несколько подходов для взаимной компенсации их недостатков. Стратегии гибридизации многообразны и, согласно классификации [5], включают:

- Взвешивающие (Weighted):

  Результаты нескольких моделей комбинируются с использованием весовых коэффициентов.

- Переключающие (Switching):

  В зависимости от контекста (например, наличие данных о пользователе) система переключается между разными моделями.

- Каскадные (Cascade):

  Одна модель производит грубый отбор кандидатов, а вторая выполняет их точное ранжирование.

- Смешанные (Mixed):

  Рекомендации от разных систем представляются пользователю одновременно. Такие системы демонстрируют значительный прирост точности [6].

2. Классификация и анализ критериев оценки

Оценка качества рекомендательных систем является многогранным процессом, требующим применения различных классов метрик, выбор которых детерминирован как архитектурой системы, так и ее целевой функцией. В данной главе приводится классификация и подробный анализ основных критериев оценки, разделенных на три категории: оффлайн-метрики, онлайн-метрики и бизнес-метрики.

2.1. Оффлайн-метрики

Оффлайн-метрики вычисляются на статичных, исторических наборах данных (датасетах), где взаимодействия пользователей с объектами уже известны. Данный класс метрик является основным инструментом на этапе разработки и итеративного улучшения моделей, поскольку позволяет быстро и без затрат на реальных пользователей сравнивать различные алгоритмы и подбирать их гиперпараметры.

Данный подкласс метрик применяется в тех случаях, когда рекомендательная система решает задачу предсказания точного значения рейтинга, который пользователь мог бы поставить объекту.

- **RMSE (Root Mean Square Error – Среднеквадратичная ошибка):**

  Является одной из наиболее распространенных метрик для задач регрессии. RMSE вычисляется как квадратный корень из среднего значения квадратов разностей между предсказанными и реальными оценками. Благодаря возведению в квадрат, данная метрика непропорционально сильно штрафует за большие ошибки, что делает ее полезной в тех случаях, когда даже единичные, но грубые промахи являются крайне нежелательными.

- **MAE (Mean Absolute Error – Средняя абсолютная ошибка):**

  Вычисляется как среднее абсолютных значений разностей между предсказанными и реальными оценками. В отличие от RMSE, MAE не так

чувствительна к выбросам и дает более интуитивно понятную интерпретацию – среднюю величину ошибки в единицах рейтинга.

Применимость и ограничения: Метрики RMSE и MAE наиболее релевантны для классических моделей коллаборативной фильтрации, основанных на матричной факторизации (SVD), где целевая функция алгоритма часто заключается в минимизации именно этих показателей [3]. Однако их главным недостатком является то, что они полностью игнорируют порядок элементов в итоговом списке рекомендаций. Система с низким RMSE может давать точные предсказания, но при этом ранжировать релевантные объекты ниже нерелевантных, что делает ее бесполезной с точки зрения пользователя.

Большинство современных РС решают задачу ранжирования, формируя для пользователя упорядоченный список из N лучших предложений. Для оценки качества такого списка применяются следующие метрики.

- **Precision@k и Recall@k:**
    - Precision@k (Точность в топ-k): Показывает долю релевантнымих объектов среди k первых рекомендованных.
    - Recall@k (Полнота в топ-k): Показывает, какую долю релевантных объектов, которые пользователь действительно оценил, система смогла найти и поместить в топ-k.

Между этими метриками существует компромисс: увеличение k (длины списка рекомендаций) как правило, повышает полноту, но снижает точность. В практических исследованиях, например, при сравнении различных мер схожести в CF, часто используются кривые «полнота-точность» для визуализации этого компромисса [7].

- **nDCG (Normalized Discounted Cumulative Gain):** Является более совершенной метрикой ранжирования, поскольку учитывает позиции релевантных элементов в списке. Ее расчет происходит в три этапа:
    1. Cumulative Gain (CG): Суммируется релевантность всех объектов в списке.

2. Discounted Cumulative Gain (DCG): Вводится дисконтирующий множитель (обычно логарифмический), который понижает вес элементов, находящихся ниже в списке. Таким образом, релевантный объект на первой позиции вносит больший вклад в метрику, чем тот же объект на десятой.
3. Normalized DCG (nDCG): Полученное значение DCG делится на "идеальное" значение (IDCG), которое было бы у списка, если бы все релевантные объекты находились вверху. Это позволяет получить нормированное значение от 0 до 1, удобное для сравнения. Данная метрика широко используется для оценки современных систем [6].

Помимо точности, важно оценивать и качественные аспекты рекомендаций, которые напрямую влияют на пользовательский опыт.

- **Diversity (Разнообразие):** Эта метрика оценивает, насколько разнообразны объекты в списке рекомендаций. Обычно она вычисляется как среднее попарное "непохожее" между всеми объектами в списке (например, 1 минус косинусное сходство их векторов признаков). Метрика Diversity напрямую связана с оценкой контентно-ориентированных систем, так как их основной недостаток – риск создания «пузыря фильтров» [4]. Оптимизация разнообразия часто вступает в противоречие с метриками точности, и нахождение баланса между ними является одной из центральных проблем при проектировании РС.
- **Novelty (Новизна) и Serendipity (Неожиданность):**
    - Новизна измеряет способность системы рекомендовать объекты, которые пользователь ранее не видел или о которых не знал.
    - Serendipity – более сложное понятие. Оно означает способность рекомендовать объекты, которые не только новы для пользователя, но и являются для него неожиданными, но при этом релевантными и полезными.

      Эти метрики используются для оценки уникального преимущества

коллаборативной фильтрации – ее способности находить неочевидные, межжанровые связи [3].

2.2. Онлайн-метрики и их роль в A/B-тестировании

Онлайн-метрики измеряются в ходе контролируемых экспериментов (A/B-тестирования) на реальных пользователях и являются единственным надежным способом измерить фактическое влияние изменений в системе. Методология A/B-тестирования предполагает разделение пользовательского трафика на две или более группы: контрольную (A), которая видит старую версию системы, и тестовую (B), которой показывается новая. Сравнивая поведение этих групп, можно сделать вывод о причинно-следственной связи между изменением и результатом.

Ключевыми онлайн-метриками являются **CTR (Click-Through Rate)** и **Conversion Rate**. Их преимущество в том, что они отражают реальное взаимодействие пользователя с рекомендациями, а не просто гипотетическую релевантность на исторических данных. Как было показано в работах, посвященных разрыву между оффлайн- и онлайн-оценкой, рост оффлайн-метрик далеко не всегда приводит к улучшению онлайн-показателей [2]. Практическим примером важности онлайн-метрик является опыт YouTube, где ключевым показателем успеха новых моделей является *увеличение времени просмотра* видеороликов пользователями [9].

2.3. Бизнес-метрики и экономический эффект

Конечная цель любой коммерческой РС – приносить пользу бизнесу. Поэтому технические метрики должны быть дополнены бизнес-показателями, отражающими долгосрочный экономический эффект.

- **Retention (Удержание пользователей)** и **LTV (Пожизненная ценность клиента)** являются ключевыми долгосрочными метриками. Система, которая предоставляет качественные и полезные рекомендации, повышает удовлетворенность пользователя, что, в свою очередь, ведет к его удержанию и увеличению LTV.

- **Разнообразие продаж (Sales Diversity):** Качественные рекомендации способны не только увеличивать прямые продажи популярных товаров, но и повышать разнообразие продаж, вовлекая пользователей в «длинный хвост» нишевых продуктов. Исследования показывают, что это может быть значительным источником дополнительной выручки для платформ с большим каталогом [10].

3. Практические и этические проблемы оценки качества

Успешная оценка рекомендательной системы выходит далеко за рамки расчета оффлайн-метрик и должна учитывать реальное взаимодействие с пользователем, долгосрочные эффекты и этические аспекты, которые могут оказывать существенное влияние на итоговое качество пользовательского опыта.

3.1. Разрыв между оффлайн- и онлайн-показателями: методологическая проблема

Одной из центральных и наиболее обсуждаемых проблем в области оценки РС является методологический разрыв между оффлайн- и онлайн-показателями. Он заключается в том, что улучшение метрик на исторических данных (оффлайн-оценка) не гарантирует пропорционального роста бизнес-показателей при взаимодействии с реальными пользователями (онлайн-оценка).

Данный разрыв обусловлен несколькими фундаментальными причинами:

1. **Статичность исторических данных:** Оффлайн-датасеты представляют собой «снимок» прошлого поведения пользователей и не могут моделировать динамическую природу пользовательских интересов и ответную реакцию на новые рекомендации.
2. **Игнорирование внешних факторов:** Оффлайн-тесты не учитывают такие факторы, как пользовательский интерфейс (UI), сезонные колебания спроса, маркетинговые кампании и **эффект новизны**, при котором пользователи на короткое время активнее взаимодействуют с любым новым интерфейсом, что может исказить результаты.
3. **Неявная предвзятость данных (Implicit Bias):** Исторические данные уже содержат в себе предвзятость, внесенную предыдущей версией рекомендательной системы. Модель, обучаясь на этих данных, может просто научиться воспроизводить старые паттерны, а не находить новые, более эффективные.

Следствием этого разрыва является высокий риск принятия неверных решений. Модель, показывающая лучший результат по метрике nDCG в оффлайн-эксперименте, может в реальности привести к снижению вовлеченности пользователей. Как показывают многочисленные исследования, рост оффлайн-метрик зачастую демонстрирует слабую или даже нулевую корреляцию с улучшением онлайн-метрик [2].

Несмотря на свои ограничения, оффлайн-оценка остается необходимым этапом. Она используется как быстрый и дешевый способ **предварительного отбора и валидации моделей**. На этом этапе можно отсеять заведомо слабые гипотезы и подобрать гиперпараметры для нескольких наиболее перспективных кандидатов. На рисунке 1 показан типичный пример кривой обучения, где качество модели на валидационной выборке отслеживается в процессе обучения для определения оптимального момента его остановки [8].

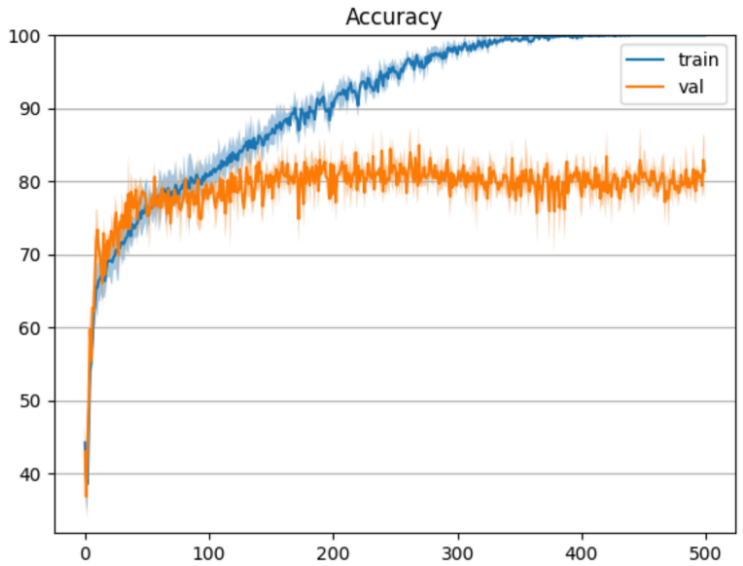

Рис. 1. Пример кривой обучения модели в оффлайн-режиме

Однако единственным надежным способом измерить реальное причинно-следственное влияние изменений является **A/B-тестирование**. Этот метод позволяет оценить изменение ключевых бизнес-показателей, таких как время

просмотра на платформах для просмотра видеороликов, что является более надежным прокси-показателем удовлетворенности, чем простые клики [9].

3.2. Этические аспекты: предвзятость и справедливость

Современная оценка РС не может обходиться без анализа этических проблем. Алгоритмы не являются нейтральными: они могут непреднамеренно воспроизводить и усиливать существующие в данных искажения.

Термин, введенный Эли Паризером, описывает феномен, при котором персонализация приводит к интеллектуальной изоляции пользователя [11]. Алгоритм, стремясь максимизировать релевантность, начинает предлагать пользователю только тот контент, который соответствует его устоявшимся взглядам и предпочтениям. Это создает информационный кокон, который ограничивает доступ к альтернативным точкам зрения, что может приводить к усилению поляризации мнений и обеднению пользовательского опыта.

Другая серьезная проблема – предвзятость популярности. Как показано на рисунке 2, небольшое количество очень популярных товаров получает непропорционально большую долю всех рекомендаций. Этот феномен создает положительную обратную связь, где популярные объекты становятся еще популярнее, а нишевые остаются невидимыми [12]. Результатом является снижение разнообразия, несправедливое распределение внимания и потенциальная гомогенизация культурного потребления.

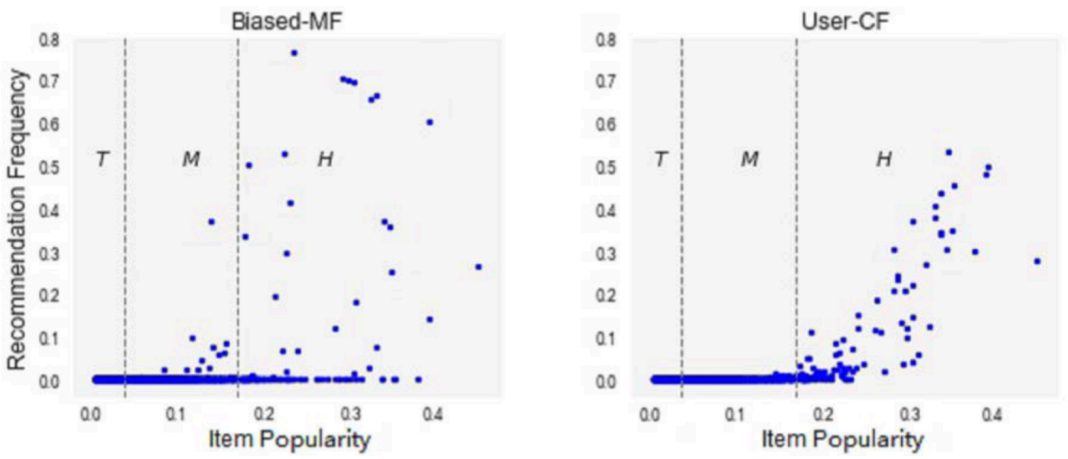

Рис. 2. Визуализация предвзятости популярности

Помимо вышеописанных, существуют и другие виды искажений, которые необходимо учитывать при оценке [12]:

- Предвзятость экспозиции (Exposure Bias):

  Система может обучаться только на тех объектах, которые она сама же и показывала пользователям. Объекты, которые по какой-то причине не попали в первоначальную выдачу, не имеют шанса получить взаимодействия и в дальнейшем также будут игнорироваться.

- Предвзятость позиции (Position Bias):

  Пользователи с большей вероятностью взаимодействуют с объектами, находящимися вверху списка, независимо от их релевантности. Если система не учитывает этот фактор, она может неверно интерпретировать клик на первой позиции как сигнал высокой релевантности.

Игнорирование этих аспектов приводит к созданию систем, которые, будучи формально точными, на практике оказываются несправедливыми и ограничивают выбор пользователя.

4. Комплексная модель оценки качества РС

Проведенный анализ демонстрирует, что для адекватной оценки качества рекомендательных систем необходима комплексная методологическая рамка, которая выходит за рамки оптимизации какой-либо одной метрики. В качестве решения предлагается многоуровневая модель оценки, условно называемая «панелью управления», которая объединяет четыре ключевых блока показателей. Такой сбалансированный подход позволяет избежать однобокой оценки и принимать взвешенные решения о качестве системы.

1. **Оффлайн-метрики:**

   Этот блок включает в себя метрики точности (nDCG, RMSE) и качества (Diversity), релевантные конкретному методу реализации. Данный уровень служит для быстрой итеративной разработки, подбора гиперпараметров и отсева заведомо слабых моделей на исторических данных. Это наиболее быстрый и наименее затратный способ проверить большое количество гипотез, однако его результаты, как было показано ранее, не могут служить финальным критерием качества.

2. **Онлайн-метрики:**

   Этот блок включает показатели, полученные в ходе контролируемых A/B-тестирований, такие как CTR и Conversion Rate. Данный этап является необходимым для валидации гипотез, отобранных на оффлайн-уровне, в реальных условиях. Онлайн-метрики измеряют непосредственное, краткосрочное взаимодействие пользователя с рекомендациями и служат "арбитром" для принятия решения о внедрении модели в промышленную эксплуатацию.

3. **Бизнес-метрики:**

   Этот блок оценивает долгосрочное влияние системы на ключевые бизнес-показатели, такие как Retention и LTV. Зачастую оптимизация краткосрочных онлайн-метрик (например, CTR) может приводить к негативным долгосрочным последствиям (например, снижению

разнообразия и, как следствие, уходу пользователей). Поэтому анализ бизнес-метрик позволяет оценить стратегическую ценность рекомендательной системы, а не только ее тактическую эффективность.

4. **Метрики справедливости:**

   Этот блок включает измерение таких аспектов, как предвзятость (Bias) и разнообразие (Diversity), и выступает в качестве регулятора для остальных уровней. Целью является не только максимизация точности или прибыли, но и обеспечение того, чтобы система была справедливой, не создавала «пузырей фильтров» и не усиливала диспропорции в популярности контента. Метрики справедливости должны контролироваться на всех этапах – от разработки до эксплуатации.

Таким образом, предложенная комплексная модель представляет собой не просто список, а целостный процесс оценки. Он предполагает переход от быстрой оффлайн-валидации к надежной онлайн-проверке, с постоянным контролем над долгосрочными бизнес-эффектами и этическими аспектами. Это позволяет сместить парадигму оценки от погони за максимальным значением одной метрики к поиску оптимального баланса между точностью, эффективностью, прибыльностью и ответственностью.

Заключение

В ходе работы было продемонстрировано, что выбор метрик не может быть универсальным и должен быть строго обоснован как решаемой задачей (предсказание рейтинга или ранжирование), так и имманентными свойствами используемого метода реализации. Так, для коллаборативной фильтрации важна оценка способности к неожиданным рекомендациям (serendipity), в то время как для контентного подхода критически важен контроль над риском снижения разнообразия.

Исследование подтвердило наличие существенного методологического разрыва между оффлайн- и онлайн-оценкой. Это делает A/B-тестирование не просто одним из этапов, а необходимым условием для принятия финального, экономически обоснованного решения о внедрении системы в промышленную эксплуатацию. Оффлайн-оценка, таким образом, должна рассматриваться как инструмент для предварительной валидации и отбора моделей, но не как финальный критерий истины.

Кроме того, было показано, что современные требования к рекомендательным системам выходят за рамки технической точности и должны включать этические аспекты. Контроль над алгоритмической предвзятостью, такой как предвзятость популярности, становится неотъемлемой частью процесса оценки, направленной на создание справедливых и ответственных систем.

Предложенная в работе комплексная модель оценки является методологической рамкой, позволяющей интегрировать эти разнородные аспекты – от точности до справедливости – в единый процесс. Такой подход обеспечивает создание более качественных, эффективных и ответственных рекомендательных систем, отвечающих как бизнес-целям, так и ожиданиям пользователей.

Список литературы

1. Gomez-Uribe C.A., Hunt N. The Netflix recommender system: Algorithms, business value, and innovation // ACM Transactions on Management Information Systems (TMIS). 2015. Т. 6. № 4. P. 1-19.
2. Shani G., Gunawardana A. Evaluating recommendation systems // Recommender systems handbook. Boston, MA: Springer, 2011. P. 257–297.
3. Королева Д.Е., Филиппов М.В. Анализ алгоритмов обучения коллаборативных рекомендательных систем // Инженерный журнал: наука и инновации. 2013. Вып. 6.
4. Андреева Я.А., Василевский К.А. Сравнительный анализ рекомендательных систем и методов оценки их качества // Международный журнал информационных технологий и энергоэффективности. 2022. Т. 7. № 4(26), часть 1. С. 59–66.
5. Ерёмин О.Ю., Моркулев Д.В. Методы реализации гибридных рекомендательных систем // E-Scio. 2023. № 3 (78).
6. Икласова К.Е., Шайханова А.К., Базарова М.Ж. и др. Обзор рекомендательных систем: модели и перспективы использования в образовательных платформах // Вестник университета Шакарим. Технические науки. 2025. № 1(17). С. 12–18.
7. Князева А.А., Колобов О.С., Турчановский И.Ю., Федотов А.М. Коллаборативная фильтрация для построения рекомендаций на основе данных о заказах // Вестник НГУ. Серия: Информационные технологии. 2018. Т. 16. № 2. С. 62–69.
8. Митина О.А., Жидков Я.А. Построение гибридной рекомендательной системы фильмов // Национальная ассоциация ученых (НАУ). 2023. № 92. С. 17–23.
9. Covington P., Adams J., Sargin E. Deep Neural Networks for YouTube Recommendations // Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16). ACM, 2016. P. 191–198.

10. Fleder D., Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity // Management Science. 2009. Vol. 55. No. 5. P. 697–712.
11. Pariser E. The filter bubble: What the Internet is hiding from you. London: Penguin UK, 2011. 294 p.
12. Chen J., Dong H., Wang X. et al. Bias and Debias in Recommender System: A Survey and Future Directions // ACM Computing Surveys. 2023. Vol. 55. No. 12. P. 1–39.