www.volsu.ru

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В БЕЗОПАСНОСТИ И ТЕЛЕКОММУНИКАЦИЯХ

DOI: https://doi.org/10.15688/NBIT.jvolsu.2025.4.1

УДК 004
ББК 32.81

ИССЛЕДОВАНИЕ ПЕРСПЕКТИВЫ ОПТИМИЗАЦИИ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ МЕТОДОМ ОГРАНИЧЕНИЯ РЕТРОСПЕКТИВЫ ОБУЧАЮЩИХ ДАННЫХ

Николай Сергеевич Аушкап

Магистрант, центр дополнительного профессионального и онлайн-образования «Пуск»,
Московский физико-технический институт
n.aushkap@gmail.com
пер. Институтский, 9, 141701 г. Долгопрудный, Российская Федерация

Олеся Александровна Какорина

Кандидат физико-математических наук,
заведующий кафедрой информационной безопасности,
Волгоградский государственный университет
kakorina@volsu.ru
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Аннотация. В настоящей статье представлены результаты эмпирического исследования перспективности оптимизации рекомендательных систем при помощи уменьшения количества обучающих данных методом ограничения временной ретроспективы данных о взаимодействиях пользователей. Для проведения исследований был использован открытый набор данных yambda, предоставленный с платформы потокового вещания Яндекс Музыки. Набор данных содержит анонимизированные данные о взаимодействиях пользователей с платформой с временными метками за десятимесячный период. Эксперименты проводились с использованием baseline моделей decayPop и itemKNN. Точность моделей оценивалась при помощи метрик NDCG@N, DCG@N, DCG@N, Coverage@N, где N \ in (10, 0, 100).

Ключевые слова: машинное обучение, датасет, тестовая выборка, модель DecayPop, модель itemKNN.

© Аушкап Н.С., Какорина О.А., 2025

Введение

Рекомендательные системы стали неотъемлемой частью многих современных цифровых платформ, предлагающих пользователям выбор, таких как: музыкальные и видео стриминговые платформы, платформы онлайн-торговли и прочие. Ядром таких систем являются модели машинного обучения, обучающиеся на данных о взаимодействиях с пользователями [1–5].

С ростом продолжительности работы сервисов и их популярности, объем данных, используемый для обучения рекомендательных моделей, стремительно увеличивается, что порождает ряд проблем.

1. Увеличение вычислительной сложности обучения и применения моделей, что приводит к более долгому обучению моделей, уменьшению скорости применения моделей для получения рекомендации, более высоким требованиям к применяемому оборудованию и в конечном итоге к удорожанию системы в целом, а иногда и к принципиальной невозможности применения некоторых моделей из-за слишком высоких требований к памяти вычислительного оборудования.

2. Концептуальный дрейф, возникающий из-за изменения предпочтений пользователей и других факторов.

Следовательно, возникают естественные исследовательские вопросы: как временная глубина обучающей выборки влияет на качество различных рекомендательных моделей в условиях реальных промышленных данных; существует ли оптимальный временной горизонт ретроспективы данных, за пределами которого данные могут не только не приводить к улучшению точности моделей, но и уменьшать точность работы рекомендательных моделей?

При этом важно отметить, что речь идет не об учете в моделях влияния периодических трендов (сезонных, краткосрочных и прочих), а именно об ограничении всего тренировочного набора данных.

Цель настоящей статьи – представить первоначальную оценку перспективности оптимизации рекомендательных систем методом ограничения ретроспективы обучающих данных. И для этого проведено эмпирическое исследование влияния глубины временной ретроспективы обучающей выборки на точность рекомендаций базовых рекомендательных моделей, примененных к реальным промышленным данным.

Описание данных

Описание полного набора данных приведено в статье [6]. Для проведения экспериментов был использован набор данных yambda 50m и его подмножество likes (таблица).

Минимальный timestamp: 0.

Максимальный timestamp: 26000000.

Продолжительность: ~300 дней (см. рис. 1).

Всего уникальных пользователей: 10 000 (см. рис. 2).

В среднем новых пользователей в день: 35.

Базовая статистика данных набора yambda 50m

Подмножество	Уникальные пользователи	Уникальные музыкальные композиции	Всего взаимодействий с пользователями
listens «прослушивания»	9 238	877 168	46 467 212
Likes «установка “лайка” пользователем»	8 283	181 304	881 456
Dislikes «установка “дизлайка” пользователем»	5 951	53 413	107 776
Unlikes «отмена “лайка” пользователем»	6 406	53 413	312 972
Undislikes «отмена “дизлайка” пользователем»	2 911	15 399	21 033
multi_event «все действия пользователей»	10 000	934 057	47 790 449

Подход к отбору тренировочных данных и ограничению глубины временной ретроспективы

Для разделения данных на тренировочные и тестовые наборы применяется подход global temporal split. При таком подходе выбирается момент времени, после которого взаимодействия попадают в тестовый набор данных, а до которого попадают в тренировочный набор данных, то есть модели обучаются на прошлом и тестируются на будущем.

При разделении данных на тренировочные и тестовые наборы гарантируется отсутствие возникновения проблемы холодного старта. Это значит, что гарантируется то, что все пользователи из тестового набора есть также и в тренировочном наборе.

Подход к ограничению временной ретроспективы данных

Для ограничения временной ретроспективы данных использовалось два метода.

1. Глобальная стратегия. Глубина данных по времени отсчитывалась от максимальной даты взаимодействия среди всех пользователей.

При применении данного метода важно отметить, что так как при отборе пользователей исключается проблема холодного старта, то при изменении глубины данных в тестовой выборке может меняться набор пользователей. Поэтому, строго говоря, сравнение метрик в экспериментах с разной глубиной ретроспективы данных не позволяет оценить изменение точности модели. Однако данный

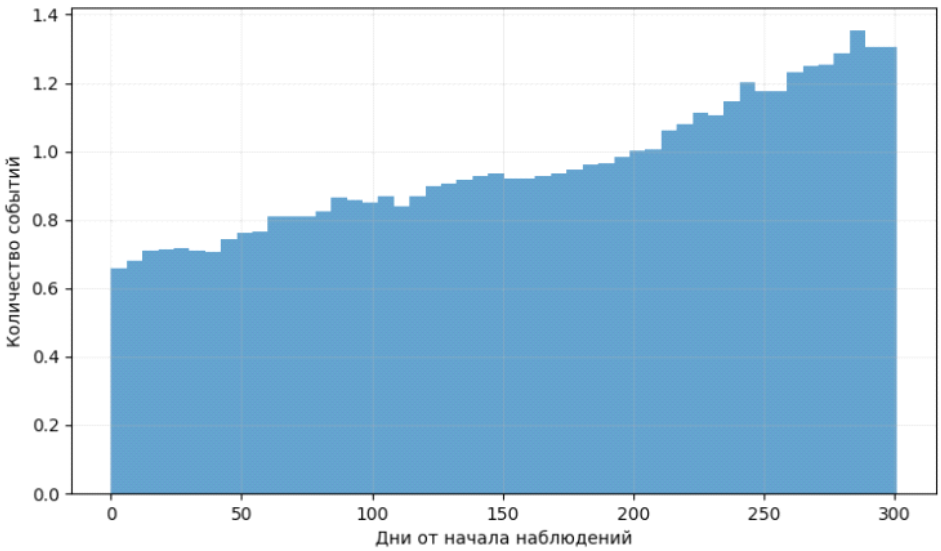


Рис. 1. Распределение событий взаимодействий пользователей с платформой

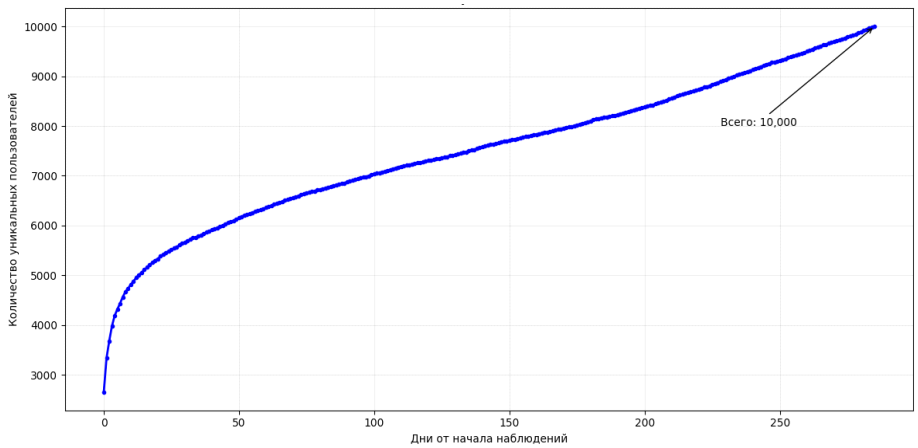


Рис. 2. График общего количества уникальных пользователей платформы в датасете

метод позволяет оценить перспективность исследования влияния временной ретроспективы на точность моделей.

2. Персонализированная стратегия. Глубина данных по времени отсчитывалась для каждого пользователя отдельного от его максимальной даты взаимодействия.

Данный метод лишен недостатка первого метода – тестовые наборы данных не меняются.

Результаты проведенных экспериментов представлены на рисунках 3–6.

Из графиков, представленных на рисунках 3–6, видно, что уменьшение временной ретроспективы действительно может приводить к увеличению точностей моделей. В данных экспериментах оптимальная глубина данных составила 7–8 недель.

Однако так же стоит отметить, что в эксперименте «Персонализированная стратегия ограничения временной ретроспективы. Модель DecayPop» рост точности не произошел, так же, как и не произошло уменьшение точности модели.

Заключение

Проведенные эксперименты показали перспективность темы исследования возможности оптимизации рекомендательных систем методом ограничения ретроспективы обучающих данных. В то же время при исследовании данной темы необходимо учитывать особенности и принципы работы рекомендательных моделей. Результаты экспериментов, полученные применительно к одной модели, не гарантированно можно применять к другим моделям, поэтому важно проверить устойчивость выявленной закономерности на других более сложных моделях.

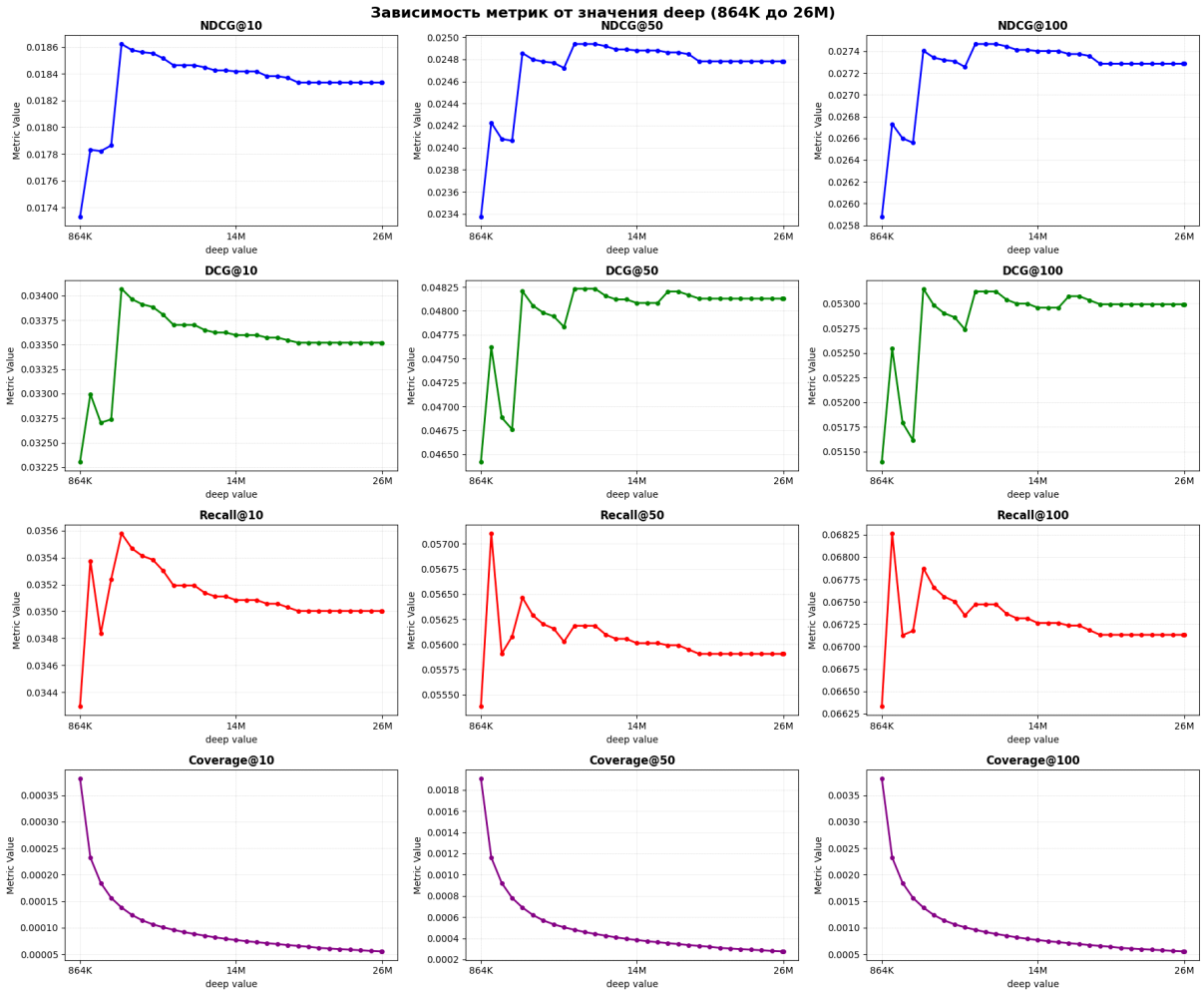


Рис. 3. Глобальная стратегия ограничения временной ретроспективы. Модель DecayPop

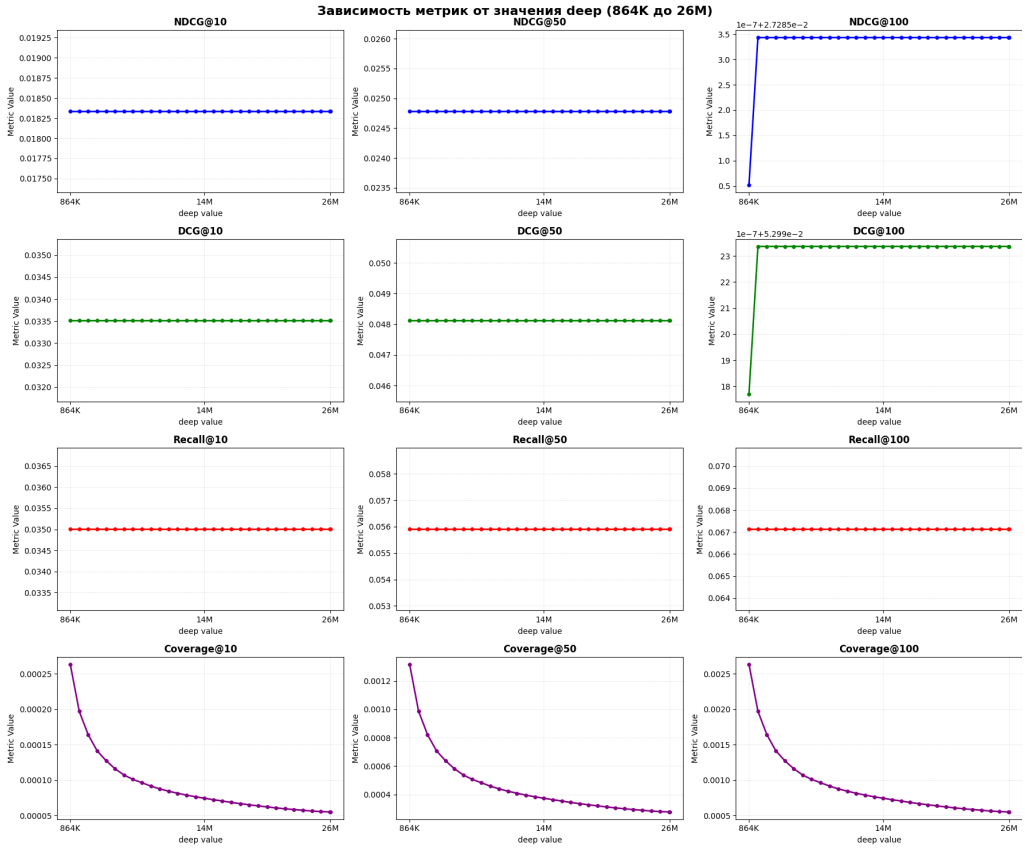


Рис. 4. Персонализированная стратегия ограничения временной ретроспективы. Модель DecayPop

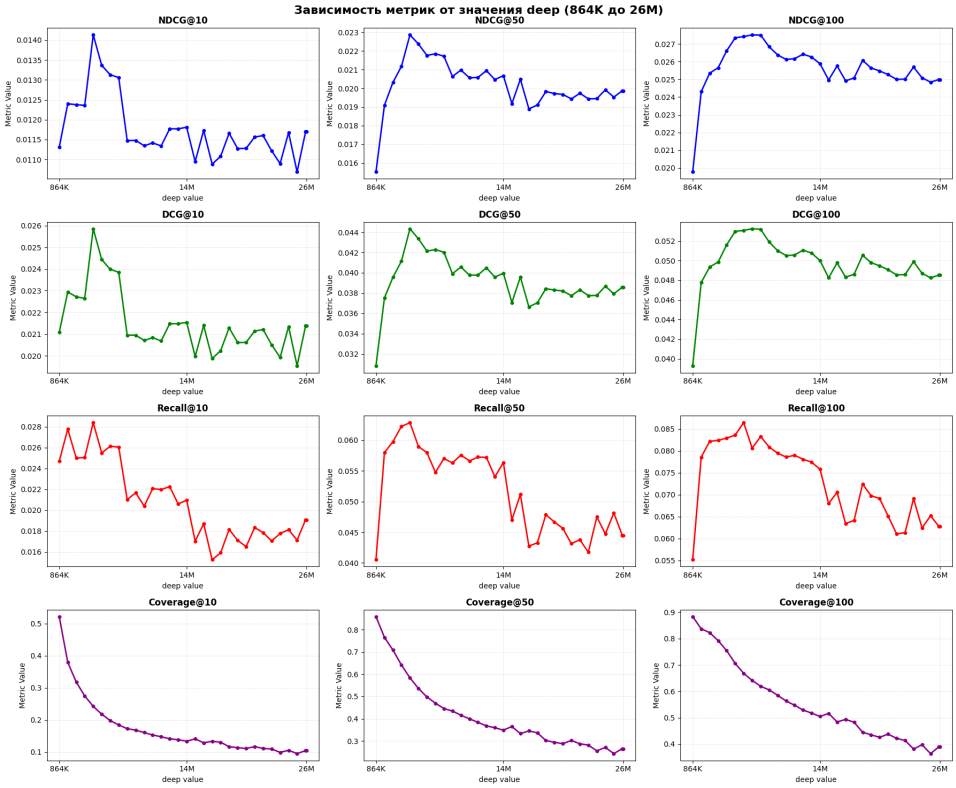


Рис. 5. Глобальная стратегия ограничения временной ретроспективы. Модель itemKNN

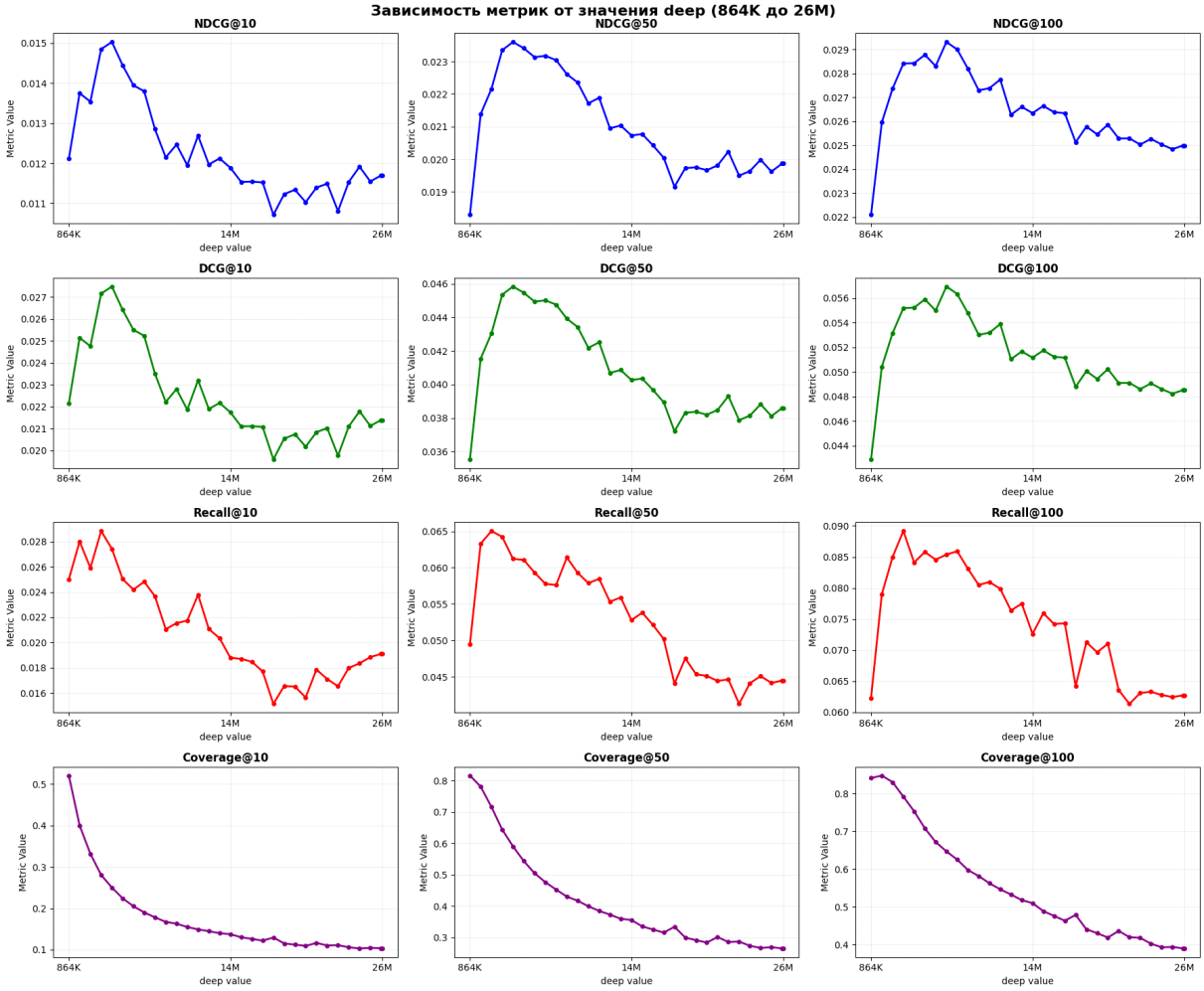


Рис. 6. Персонализированная стратегия ограничения временной ретроспективы. Модель itemKNN

Таким образом, работа подтверждает перспективность исследований в области оптимизации рекомендательных систем методом ограничения ретроспективы обучающих данных и потенциального включения данного этапа в pipeline проектирования рекомендательных систем.

СПИСОК ЛИТЕРАТУРЫ

1. Нефедов, Ю. В. Рекомендательные системы интернет-сервисов: этапы проектирования и разработки / Ю. В. Нефедов, И. А. Гальперина // Научно-техническая информация. Серия 2, Информационные процессы и системы. – 2018. – № 10. – С. 29–35.

2. Панфилов, А. В. Рекомендательные системы безопасности для риск-ориентированного подхода / А. В. Панфилов, В. В. Дерюшев, А. А. Короткий // Безопасность труда в промышленности. – 2020. – № 5. – С. 48–55. – DOI: 10.24000/0409-2961-2020-5-48-55

3. Тыжинов, Т. А. Рекомендательные системы в сфере туризма / Т. А. Тыжинов // Студенческий форум. – 2018. – № 11-2 (32). – С. 5–7.

4. Хасанов, В. И. Рекомендательные системы в автоматизированных библиотечных информационных системах / В. И. Хасанов, А. С. Карауш // Электронные средства и системы управления : материалы докл. Междунар. науч.-практ. конф. – 2009. – № 1. – С. 3–7.

5. Цурко, В. В. Рекомендательные системы в здравоохранении / В. В. Цурко // Управление большими системами : сб. тр. – 2019. – № 82. – С. 61–73.

6. Yambda-5B – A Large-Scale Multi-modal Dataset for Ranking And Retrieval / A. Ploshkin [et al.] // arXiv. – 2025. – URL: https://arxiv.org/abs/2505.22238

REFERECNES

1. Nefedov Yu.V., Galperina I.A. Rekomendatelnye sistemy internet-servisov: etapy

proektirovaniya i razrabotki [Recommender Systems for Internet Services: Design and Development Stages]. *Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnye protsessy i sistemy* [Scientific and Technical Information. Series 2: Information Processes and Systems], 2018, no. 10, pp. 29-35.

2. Panfilov A.V., Deryushev V.V., Korotkiy A.A. Rekomendatelnye sistemy bezopasnosti dlya risk-orientirovannogo podkhoda [Security Recommender Systems for a Risk-Based Approach]. *Bezopasnost truda v promyshlennosti* [Occupational Safety in Industry], 2020, no. 5, pp. 48-55. DOI: 10.24000/0409-2961-2020-5-48-55

3. Tyzhinov T.A. Rekomendatelnye sistemy v sfere turizma [Recommender Systems in the field of tourism]. *Studencheskiy forum* [Student Forum], 2018, no. 11-2 (32), pp. 5-7.

4. Khasanov V.I., Karaush A.S. Rekomendatelnye sistemy v avtomatizirovannyh bibliotechnyh informatsionnyh sistemah [Recommender Systems in Automated Library Information Systems]. *Elektronnye sredstva i sistemy upravleniya: materialy dokl. Mezhdunar. nauch.-prakt. konf.* [Electronic Means and Control Systems. Proceedings of the International Scientific and Practical Conference], 2009, no. 1, pp. 3-7.

5. Tsurko V.V. Rekomendatelnye sistemy v zdravookhranenii [Recommender Systems in Healthcare]. *Upravlenie bolshimi sistemami: sb. tr.* [Large Systems Control. Proceedings], 2019, no. 82, pp. 61-73.

6. Ploshkin A., Tytskiy V., Pismenny A., Baikalov V., Taychinov E., Permiakov A., Burlakov D., Krofto E., Savushkin N. Yambda-5B – A Large-Scale Multi-Modal Dataset for Ranking and Retrieval. *arXiv*, 2025. URL: https://arxiv.org/abs/2505.22238

RESEARCH ON THE PROSPECTS OF OPTIMIZING RECOMMENDATION SYSTEMS BY RESTRICTING THE RETROSPECTIVE OF TRAINING DATA

Nikolai S. Aushkap

Master's Student, Center for Further and Online Education "Pusk",
Moscow Institute of Physics and Technology
n.aushkap@gmail.com
Institutsky Lane, 9, 141701 Dolgoprudny, Russian Federation

Olesya A. Kakorina

Candidate of Sciences (Physics and Mathematics),
Head of the Department of Information Security,
Volgograd State University
kakorina@volsu.ru
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Abstract. This article presents an empirical study investigating the feasibility of optimizing recommendation systems by reducing the volume of training data through limiting the temporal retrospective of user interaction records. The research addresses key challenges arising from continuously growing datasets in operational services, such as increased computational complexity, model training costs, and conceptual drift in user preferences. Utilizing the open yambda dataset from Yandex.Music streaming platform, which contains anonymized timestamped user interactions over a ten-month period, the study experiments with two baseline models: DecayPop and itemKNN. Two primary strategies for restricting data retrospect were implemented and compared: a global strategy, which truncates data based on the latest timestamp across all users, and a personalized strategy, which sets individual cut-offs per user based on their last interaction. Model accuracy was evaluated using standard metrics, including NDCG@N, DCG@N, and Coverage@N for N values of 10, 50, and 100. The results demonstrate that strategically limiting the training data timeline can indeed improve recommendation accuracy for certain models, with an optimal retrospective depth identified

at approximately 7–8 weeks for the studied dataset. This work confirms the potential of this optimization method and suggests its prospective inclusion in the recommendation system design pipeline, while noting the need for further validation on more complex models.

Key words: machine learning, dataset, test set, DecayPop model, itemKNN model.