ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.89

Текущее состояние метода Generative Retrieval в рекомендательных системах

М.В. Морозов, В.В. Мисюра
Донской государственный технический университет, г. Ростов-на-Дону, Российская Федерация

Аннотация
В современной индустрии рекомендательных систем существует проблема ограничения производительности классических моделей при масштабировании объёмных таблиц эмбеддингов. Цель настоящей статьи — обзор парадигмы Generative Retrieval как концептуального решения этой архитектурной проблемы, её текущего состояния и основных тенденций. Объект исследования — архитектуры рекомендательных систем, методология — анализ научной литературы. На основе изученных публикаций систематизированы подходы к токенизации каталогов и рассмотрены особенности адаптации больших языковых моделей (LLM) к задачам рекомендаций. Практическая значимость работы состоит в формировании целостного представления о новом классе алгоритмов, что открывает перспективы их внедрения в высоконагруженные продуктовые среды для оптимизации памяти и вычислений.

Ключевые слова: рекомендательные системы, Generative Retrieval, машинное обучение, генеративные нейронные сети

Для цитирования. Морозов М.В., Мисюра В.В. Текущее состояние метода Generative Retrieval в рекомендательных системах. *Молодой исследователь Дона.* 2026;11(4):75–78.

State-of-the-Art of the Recommender Systems with Generative Retrieval Model

Mikhail V. Morozov, Valentina V. Misyura
Don State Technical University, Rostov-on-Don, Russian Federation

Abstract
Modern recommender systems face a problem of limited performance capacity of conventional models upon scaling up massive embedding lookup tables. The aim of the article is to conduct a review of the Generative Retrieval paradigm, which is considered to be a conceptual solution for this architectural problem, as well as to define the state-of-the-art and key trends of this paradigm. The object of the research is the Recommender System architecture, and research methodology is literature review. Based on the publications studied, the approaches to catalog tokenization have been systematized and features of large language models (LLM) adaptation to recommendation tasks have been analysed. The practical value of the study is included in the development of comprehensive understanding of the new class of algorithms, which opens up prospects for their deployment in high-traffic production environments for optimization of memory and computation.

Keywords: recommender systems, generative retrieval, machine learning, generative neural networks

For Citation. Morozov MV, Misyura VV. State-of-the-Art of the Recommender Systems with Generative Retrieval Model. *Young Researcher of Don.* 2026;11(4):75–78.

Введение. Рекомендательные системы стали фундаментом пользовательского опыта в современных цифровых продуктах. С экспоненциальным ростом объёмов контента (товаров, публикаций, видеозаписей) традиционные алгоритмы сталкиваются с серьёзным архитектурным ограничением. Классические подходы, основанные на моделировании последовательностей и использовании больших таблиц эмбеддингов (Embedding Tables), достигают предела производительности при масштабировании каталогов. Каждое увеличение каталога на сотни миллионов записей кратно повышает требования к памяти и усложняет вычислительные пайплайны [1]. Выходом из этого инженерного тупика стала новая парадигма — Generative Retrieval. Переосмыслив рекомендатель-

© Морозов М.В., Мисюра В.В., 2026

ную задачу как процесс авторегрессионной генерации семантических токенов, метод позволяет полностью отказаться от громоздких матриц векторов. В настоящей статье рассматриваются принципы построения алгоритмов Generative Retrieval, анализируются современные методы токенизации и исследуется влияние больших языковых моделей (LLM) на архитектуру рекомендательных систем.

С момента появления трансформерных архитектур в 2017 году [2] началось активное внедрение подобных моделей в рекомендательные системы. Результатом стали алгоритмы SasRec [3], BERT4Rec [4], HSTU [5], ARGUS [6] и другие, которые применяются в том числе в российских информационных системах компаний МТС, Yandex, hh.ru и Сбера. Общей чертой этих и других глубоких рекомендательных моделей (DLRM) является наличие крупных таблиц обучаемых векторов [1] (Embedding Table) — матриц, в которых каждая строка соответствует одному товару каталога, а сами алгоритмы, по существу, решают задачу ранжирования таких таблиц. Задача сводилась к следующему: опираясь на историю взаимодействий, сформировать скрытое состояние, максимально близкое к вектору того товара, с которым пользователь с наибольшей вероятностью провзаимодействует следующим. Мерой близости обычно служит скалярное произведение векторов.

$$I^{i+1} = argmax_j\, R\left(E^{I^0}, \ldots, E^{I^i}\right) \cdot E^j, \tag{1}$$

где, I^i — идентификатор i-ого товара в истории пользователя, R (·) — скрытое состояние следующего товара на основе истории, E — матрица обучаемых векторов.

Как видно из формулы (1), в такой парадигме объём вычислений линейно зависит от размера каталога. В популярных сервисах с пользовательским контентом это создавало значительные трудности. Так, размер Embedding Table в рекомендательной системе одной социальной сети достигал 13 триллионов параметров [7]. Для борьбы с проблемой часто применялись индексы приближённого поиска ближайших соседей (ANN), например FAISS [8], однако они усложняли обучение, поскольку при расчёте градиентов требовались сами таблицы. Эта проблема затрагивала и сервисы Google [9], и на конференции NeurIPS 2023 ими был предложен новый класс алгоритмов — Generative Retrieval, а также его реализация TIGER [10].

Предложенная парадигма заменяет ранжирование каталога авторегрессивной генерацией следующего товара посредством семантических токенов (Semantic IDs). Ключевая идея состоит в представлении каждого товара как набора идентификаторов, которые иерархически кодируют его семантические признаки. По аналогии с магазином: хотя количество уникальных товарных позиций может исчисляться десятками тысяч, каждый товар однозначно описывается небольшим числом обобщённых характеристик — одна из нескольких десятков категорий, один из сотен брендов и т.д. Вместо единственной огромной Embedding Table с записями о каждом товаре можно завести несколько компактных таблиц — по одной на каждую характеристику. Для разложения на семантические токены вводится дополнительный модуль — токенизатор. Алгоритм по-прежнему выполняет ранжирование, но многократно, и результатом оказывается не конкретный товар, а совокупность его свойств. Возвращаясь к аналогии с магазином: традиционные алгоритмы на основе истории покупок рекомендовали один товар из всего каталога, тогда как Generative Retrieval сначала выбирает категорию из каталога категорий, затем бренд из каталога брендов и так далее, после чего декодирует полученный набор свойств обратно в конкретный товар. Таким образом, справедлива система уравнений:

$$\begin{gathered} S^0 = argmax_j\, R\left(E_0^{I^0}, \ldots, E_k^{I^0}, \ldots, E_0^{I^i}, \ldots, E_k^{I^i}\right) \cdot E_0^j \\ S^k = argmax_j\, R\left(E_0^{I^0}, \ldots, E_k^{I^0}, \ldots, E_0^{I^i}, \ldots, E_k^{I^i}, E_0^{S^0}, \ldots, E_{k-1}^{S^{k-1}}\right) \cdot E_k^j \\ I^{i+1} = L\left(S^0, \ldots, S^k\right), \end{gathered} \tag{2}$$

где I^i — идентификатор i-ого товара в истории пользователя, R (·) — скрытое состояние следующего товара на основе истории, S^k — семантический токен k-ого ранга, E_k — матрица обучаемых векторов семантических токенов k-ого ранга, L (·) — декодирование товара по семантическим токенам.

Как видно из формулы (2), сложность поиска ближайших векторов зависит от числа кодируемых признаков (k) и размеров соответствующих таблиц ($|E_k|$). Несмотря на выполнение нескольких последовательных операций, этот подход позволяет заменить трудоёмкий поиск по обширному каталогу последовательностью более простых вычислений, что повышает эффективность системы при работе с масштабными наборами данных [10].

Предложенная группой Google система TIGER [10] продемонстрировала как возможность замены больших таблиц набором семантических токенов, так и лучшие (SOTA) метрики на открытом наборе данных на момент публикации.

В настоящее время известно несколько крупных компаний, уже внедривших данный подход в реальные сервисы с аудиторией в сотни миллионов пользователей, — Google, Pinterest, Kuaishou; ещё больше компаний, включая Baidu и Spotify, находятся на стадии исследований и разработки.

Этот интерес стимулировал активное развитие метода в академических и инженерных сообществах. В настоящей статье особое внимание уделяется совершенствованию методов токенизации и моделей анализа последовательностей.

Единый базовый вариант токенизатора для систем Generative Retrieval на сегодня отсутствует.

В канонической работе 2023 г. [10] применялась векторная квантизация посредством RQ-VAE и VQ-VAE поверх эмбеддингов Sentence-T5, построенных на текстовом представлении товаров. Последующие исследования [11] показали, что более простые подходы, в частности RK-Means (Residual K-Means), не уступают, а иногда и превосходят VAE-методы.

Распространённой практикой стало использование контрастного обучения (Contrastive Learning), которое позволяет токенизатору фиксировать коллаборативные зависимости между товарами.

Перспективный алгоритм ActionPiece [12] был представлен на конференции ICML 2025. Токенизация в нём строится аналогично subword-уровневым LLM, но вместо корпуса текстов используется история взаимодействий, в которой товары заданы набором низкоуровневых признаков. Задача алгоритма — итеративно сжимать наиболее часто совместно встречающиеся признаки до тех пор, пока словарь не сократится до заданного размера. Такой подход позволяет учитывать контекст при токенизации без контрастного обучения; метод показал улучшение метрик по сравнению с RQ-VAE.

Резюмируя состояние токенизации в алгоритмах Generative Retrieval, можно отметить, что исследователи продолжают поиск наиболее эффективного способа представления товаров семантическими токенами. На сегодняшний день RK-Means ближе прочих к роли стандартного токенизатора (baseline), поскольку демонстрирует достойные метрики и, в отличие от VAE, не требует значительных вычислительных ресурсов.

Рекомендации в парадигме Generative Retrieval по своей сути близки к генерации текста в LLM. Естественной идеей было взять предобученную LLM, расширить её словарь семантическими токенами товаров и дообучить на историях взаимодействий; однако модель крайне плохо обобщает такие токены из-за их сильного отличия от текстовых.

В работе 2025 г. исследователи Google [13] решили указанную проблему путём дообучения LLM на доменных данных. Чтобы обучить модель рекомендовать видео, они сначала дообучили её на текстовых описаниях роликов и пользовательских историях, постепенно вводя семантические токены из токенизатора. Полученная модель совместила общие знания LLM со способностью давать рекомендации. Итоговая система PLUM продемонстрировала рост метрик точности и бизнес-показателей в AB-тестах.

Таким образом, доказано, что вместо обучения рекомендательной системы с нуля целесообразно применять перенос знаний (Transfer Learning) на базе готовой LLM. На практике это способно решить серьёзную инженерную проблему — потребность в колоссальных объёмах исторических данных. Современные рекомендательные системы обучаются на записях о миллиардах взаимодействий [9], однако такая возможность имеется не всегда. Использование LLM потенциально позволяет создавать рекомендательные системы на значительно меньших выборках, компенсируя нехватку данных знаниями из смежных доменов.

Парадигма Generative Retrieval успешно разрешает инженерные трудности, связанные с большими каталогами, демонстрируя высокие показатели точности и бизнес-метрики. Эти системы активно внедряются в западных и китайских компаниях. Вместе с тем в России интереса к данному направлению не отмечено: отсутствуют как упоминания о внедрении со стороны бизнеса, так и академические публикации (поиск проводился по открытым базам eLibrary).

Заключение. Таким образом, Generative Retrieval представляет собой не очередное эволюционное улучшение, а полноценный парадигмальный сдвиг в области рекомендательных систем. Переход от дискретного ранжирования необъятных таблиц к авторегрессионной генерации семантических токенов решает ключевую проблему DLRM-моделей — масштабируемость. Инновации последних лет, включая контекстно-зависимую токенизацию (ActionPiece) и перенос знаний посредством адаптации готовых LLM (PLUM), подтверждают значительный потенциал подхода. Опыт ведущих технологических компаний свидетельствует о жизнеспособности Generative Retrieval в высоконагруженных продуктовых средах. Для бизнеса это означает возможность предоставлять пользователям более точные, разнообразные и учитывающие контекст рекомендации при снижении затрат на хранение раздутых Embedding Tables.

Список литературы

1. Desai A, Shrivastava A. The Trade-Offs of Model Size in Large Recommendation Models: 100GB to 10MB Criteo-tb DLRM model. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS'22. Article No.: 246*; 2022. P. 33961–33972. URL: https://arxiv.org/abs/2207.10731 (дата обращения: 05.05.2026).

2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: *Proceedings of 31st International Conference on Neural Information Processing Systems (NIPS 2017).* Long Beach, CA, USA; 2017. 30p. https://doi.org/10.48550/arXiv.1706.03762

3. Kang WC, McAuley J. Self-Attentive Sequential Recommendation. In: *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Nov 17–20, 2018*. Singapore: IEEE; 2018. P. 197–206. https://doi.org/10.48550/arXiv.1808.09781

4. Sun F, Liu J, Wu J, Pei C, Lin X, Ou W, et al. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM'19. November 3–7, 2019*. Beijing, Chin; 2019. P. 1441–1450. https://doi.org/10.48550/arXiv.1904.06690

5. Zhai J, Liao L, Liu X, Wang Y, Li R, Cao X, et al. Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations. In: *Proceedings of the 41st International Conference on Machine Learning, ICML'24.* JMLR.org; 2024. P. 58484 – 58509 (Article No.: 2414). URL: https://arxiv.org/abs/2402.17152 (дата обращения: 05.05.2026).

6. Khrylchenko K, Matveev A, Makeev Sergei, Baikalov V. Scaling Recommender Transformers to One Billion Parameters. In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, KDD'26*. New York, USA: Association for Computing Machinery; 2026. P. 2255 – 2265 https://doi.org/10.1145/3770854.3783916

7. Mudigere D, Hao Y, Huang J, Jia Z, Tulloch A, Sridharan S, et al. Software-Hardware Co-design for Fast and Scalable Training of Deep Learning Recommendation Models. In: *Proceedings of the 49th Annual International Symposium on Computer Architecture, ISCA'22*. New York, US: Association for Computing Machinery. P. 993–1011. https://doi.org/10.1145/3470496.3533727

8. Johnson J, Douze M, Jégou H. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data.* 2019;7(3):535–547. https://doi.org/10.1109/TBDATA.2019.2921572

9. Covington P, Adams J, Sargin E. Deep Neural Networks for YouTube Recommendations. In: *Proceedings of the 10th ACM Conference on Recommender Systems, RecSys'16.* New York: United States Association for Computing Machinery; 2016. P. 191–198. https://doi.org/10.1145/2959100.2959190

10. Rajput S, Mehta N, Singh A, Keshavan R, Vu T, Heidt L, et al. Recommender Systems with Generative Retrieval. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS'23*. Red Hook, NY, United States: Curran Associates Inc; 2023. P. 10299 –10315 (Article No.: 452). URL: https://arxiv.org/abs/2305.05065 (дата обращения: 05.05.2026).

11. Ju CM, Collins L, Neves L, Kumar B, Wang LY, Zhao T, et al. Generative Recommendation with Semantic IDs: A Practitioner's Handbook. In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM'25.* New York, US: Association for Computing Machinery; 2025. P. 6420–6425. https://doi.org/10.1145/3746252.3761612

12. Hou Y, Ni J, He Z, Sachdeva N, Kang WC, Chi EH, et al. ActionPiece: Contextually Tokenizing Action Sequences for Generative Recommendation. In: *Proceedings of the 42nd International Conference on Machine Learning, ICML'25.* JMLR.org; 2025. P. 24004–24024 (Article No.: 941). URL: https://arxiv.org/abs/2502.13581 (дата обращения: 05.05.2026).

13. He R, Heldt L, Hong L, Keshavan R, Mao S, Mehta Nikhil, et al. PLUM: Adapting Pre-trained Language Models for Industrial-scale Generative Recommendations. In: Proceedings of the ACM Web Conference 2026 (WWW'26). New York, US: Association for Computing Machinery; 2025. P. 8093–8104. https://doi.org/10.1145/3774904.3792802

Об авторах:

Михаил Валерьевич Морозов, магистрант кафедры «Математика и информатика» Донского государственного технического университета (344003, Российская Федерация, г. Ростов-на-Дону, пл. Гагарина, 1) yahryn@gmail.com

Валентина Владимировна Мисюра, кандидат физико-математических наук, доцент кафедры «Математика и информатика» Донского государственного технического университета (344003, Российская Федерация, г. Ростов-на-Дону, пл. Гагарина, 1) vvmisyura2011@gmail.com

Конфликт интересов: **авторы заявляют об отсутствии конфликта интересов.**

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the Authors:

Mikhail V. Morozov, Master's Degree Student of the Mathematics and Computer Science Department, Don State Technical University (1, Gagarin Sq., Rostov-on-Don, 344003, Russian Federation), yahryn@gmail.com

Valentina V. Misyura, Cand.Sci. (Physics and Mathematics), Associate Professor of the Mathematics and Computer Science Department, Don State Technical University (1, Gagarin Sq., Rostov-on-Don, 344003, Russian Federation), vvmisyura2011@gmail.com

Conflict of Interest Statement: **the authors declare no conflict of interest.**

All authors have read and approved the final manuscript.