ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.827

ПОВЫШЕНИЕ КАЧЕСТВА РЕЗУЛЬТАТОВ ПОИСКА НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ В КНИЖНОМ ФОНДЕ ЗА СЧЁТ ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ ИИ И МЕТОДА RAG

Винокуров Алексей Алексеевич,
Студент группы ИУК5-71Б
vinokurovaa1@student.bmstu.ru

Осипов Андрей Александрович,
Студент группы ИУК5-71Б
osipovaa@student.bmstu.ru

Вершинин Евгений Владимирович,
Кандидат физико-математических наук, доцент и заведующий кафедрой ИУК5 «Системы обработки информации»
vershinin@bmstu.ru

Ильичев Владимир Юрьевич,
Кандидат технических наук, доцент кафедры ИУК5 «Системы обработки информации»
ilyichev.vyu@bmstu.ru

Калужский филиал Московского государственного технического университета имени Н.Э. Баумана

Калуга, Россия

Аннотация

Рассмотрена проблема неэффективности традиционных методов поиска научно-технической информации в условиях экспоненциального роста объемов данных. Исследована технология RAG (Retrieval-Augmented Generation) для повышения качества семантического поиска. Выявлены преимущества RAG-подхода в сравнении с классическими методами поиска по метаданным. Проведена экспериментальная оценка различных способов индексирования книжного фонда с применением векторных представлений текста. По результатам проведенной работы и на основании полученных данных сделаны выводы о значительном повышении эффективности поиска и снижении количества нерелевантных результатов при использовании RAG-архитектуры. Результаты исследования применимы для модернизации библиотечных каталогов и систем управления научно-технической документацией.

Ключевые слова: RAG, семантический поиск, информационный поиск, векторные базы данных, библиотечные системы.

IMPROVING THE QUALITY OF SCIENTIFIC AND TECHNICAL INFORMATION SEARCH RESULTS IN BOOK COLLECTIONS THROUGH THE APPLICATION OF AI TECHNOLOGIES AND THE RAG METHOD

Vinokurov Alexey Alexeyevich
Student of group IUK5-71B
vinokurovaa1@student.bmstu.ru

Osipov Andrey Alexandrovich
Student of group IUK5-71B
osipovaa@student.bmstu.ru

Vershinin Evgeny Vladimirovich
Candidate of Physical and Mathematical Sciences, Associate Professor and Head of Department IUK5 "Information Processing Systems"
vershinin@bmstu.ru

Ilyichev Vladimir Yurievich
Candidate of Technical Sciences, Associate Professor of Department IUK5 "Information Processing Systems"
ilyichev.vyu@bmstu.ru

Kaluga Branch of Bauman Moscow State Technical University

Kaluga, Russia

ABSTRACT

The problem of inefficiency of traditional methods for searching scientific and technical information under conditions of exponential growth in data volumes is considered. The RAG (Retrieval-Augmented Generation) technology for improving the quality of semantic search is investigated. The advantages of the RAG approach compared to classical metadata-based search methods are identified. An experimental evaluation of various methods for indexing book collections using vector text representations was conducted. Based on the results of the work performed and the data obtained, conclusions were drawn about a significant increase in search efficiency and reduction in the number of irrelevant results when using RAG architecture. The research results are applicable for modernizing library catalogs and scientific and technical documentation management systems.

Keywords: RAG, semantic search, information retrieval, vector databases, library systems.

Современные библиотечные и информационные системы сталкиваются с критической проблемой низкого качества поиска в растущих массивах научно-технической литературы. Традиционные методы поиска, основанные на точном соответствии ключевых слов и метаданных, не способны понимать семантическое содержание запросов и возвращают нерелевантные результаты. Пользователи сталкиваются с необходимостью

угадывать точную терминологию каталогов для получения релевантных результатов, что приводит к пропуску ценных источников информации [1][4].

Технология RAG решает эту проблему за счёт объединения семантического поиска и генерации ответов на основе обширных и регулярно обновляемых баз знаний. RAG (Retrieval Augmented Generation — генерация с дополненной выборкой) сочетает языковую модель с внешней базой знаний: ИИ-помощник на базе RAG сначала находит релевантные документы во внешних источниках (например, с помощью векторного поиска), ранжирует найденную информацию, а затем формирует ответ на запрос пользователя, интегрируя полученные данные и свой смысловой анализ [7]. Такой подход учитывает как контекст запроса, так и извлечённую информацию, что позволяет понимать сложные и неточные формулировки, находить релевантные источники независимо от терминологии пользователя и тем самым существенно повышать полноту и точность поиска в масштабных коллекциях научно-технической литературы.

Цель исследования — выявить, насколько применение технологии RAG повышает качество и полноту поиска научно-технической информации по сравнению с традиционным реляционным подходами, а также определить оптимальные условия и параметры внедрения RAG для библиотечных и информационных систем.

Актуальность данного исследования определяется необходимостью повышения качества поиска в условиях информационного изобилия и растущими требованиями пользователей к релевантности и полноте результатов поиска научно-технической информации.

Для проведения эксперимента был сформирован корпус из 100 искусственно сгенерированных книг научно-технической тематики, содержащих метаданные (название, авторы, жанры, год издания, язык), описания и полнотекстовое содержание. Такой подход обеспечивает объективность эксперимента — благодаря отсутствию упоминаний этих книг в открытых источниках исключается вероятность их наличия в датасетах обучения языковых моделей или поиска по внешним данным.

Текстовое содержание каждой книги предварительно сегментировалось на фрагменты размером 1000 токенов с перекрытием 100 токенов между соседними фрагментами для сохранения контекстной связности при чанкинге. Векторные представления текстовых фрагментов формировались с помощью библиотеки SentenceTransformer на базе модели cointegrated/rubert-tiny2, генерирующей эмбеддинги размерностью 312. Данная модель оптимизирована для русскоязычных текстов и обеспечивает эффективное преобразование семантики в векторное пространство [3].

В качестве векторной СУБД использовалась Qdrant с настройкой коллекции под размерность эмбеддингов 312 и метрикой сходства Cosine. Косинусная схожесть измеряет угол между векторами в многомерном пространстве, игнорируя их длину, что делает её оптимальной для семантического поиска текстов. Значения метрики варьируются от -1 (противоположная направленность) до 1 (полное совпадение направления), где большие значения соответствуют большему семантическому сходству [2].

Поисковые запросы выполнялись с параметрами top-k=5 (возврат 5 наиболее релевантных результатов) и score_threshold=0.5 для отсечения не релевантных фрагментов. Результаты ранжировались по убыванию косинусной схожести между векторным представлением запроса и индексированными фрагментами.

Для количественной оценки эффективности RAG-подхода был сформирован тестовый набор из 30 поисковых запросов, составленных вручную для обеспечения разнообразия тематик и формулировок. Качество поиска оценивалось по четырём стандартным метрикам информационного поиска:

Precision@k (точность на уровне k) — доля релевантных документов среди первых k результатов выдачи. Формально рассчитывается как отношение числа релевантных объектов в топ-k к общему числу k. Метрика отражает качество отобранных результатов и минимизацию информационного шума [6].

Recall@k (полнота на уровне k) — доля найденных релевантных документов от общего числа релевантных объектов в коллекции. Показывает способность системы обнаруживать все значимые результаты в пределах топ-k позиций [6].

NDCG@k (Normalized Discounted Cumulative Gain) — метрика ранжирования, учитывающая как релевантность результатов, так и их позицию в выдаче. Результаты на верхних позициях вносят больший вклад в итоговую оценку, что отражает поведение пользователей при просмотре результатов поиска. Нормализация обеспечивает сопоставимость значений между различными запросами [6].

MRR (Mean Reciprocal Rank) — среднее значение обратного ранга первого релевантного результата по всем запросам. Метрика показывает, насколько быстро пользователь находит первый подходящий ответ, что критично для практического применения поисковых систем [6].

Все метрики усреднялись по 30 тестовым запросам для получения агрегированной оценки качества системы. По результатам обработки данных эксперимента авторами внесены данные в таблицу 1.

Метрика	RAG	SQL	Разница
Precision@5	0.3933	0.2322	+0.1611 (+69.4%)
Recall@5	0.6784	0.7458	-0.0674 (-9.0%)
NDCG@5	0.6518	—	—
MRR	0.6983	—	—
Hit Rate@5	0.8667	—	—
F1-мера	—	0.3016	—
Общее покрытие	—	0.6542	—
Найдено документов	—	70	0
Ожидалось документов	—	107	—

Таблица 1 – метрики, собранные в результате эксперимента

Анализ стандартных метрик поиска, указанных в таблице 1, продемонстрировал контрастирующие профили эффективности двух подходов.

RAG-система показала точность Precision@5 = 0.3933 (≈40% релевантных документов в топ-5), что существенно выше точности SQL-поиска (0.2322 или 23%). На практике это означает, что пользователь почти вдвое реже сталкивается с нерелевантными результатами на первых пяти позициях.

Recall@5 для RAG составил 0.6784 (≈68% релевантных документов в топ-5), тогда как SQL-поиск обнаружил 70 из 107 ожидаемых релевантных фрагментов (≈65.4%). Это указывает, что RAG быстрее обеспечивает большую часть нужной информации уже в первой пятёрке, а SQL пропускает свыше трети релевантных фрагментов.

Высокая полнота SQL достигается ценой низкой точности, что отражено низкой F1-мерой (0.3016) — система находит значительное число релевантных документов, но при этом возвращает большое количество нерелевантных.

На рисунке 1 приведен построенный авторами график сравнения ключевых метрик RAG- и SQL-поиска, иллюстрирующее их различия по точности, полноте и общему покрытию релевантных документов.

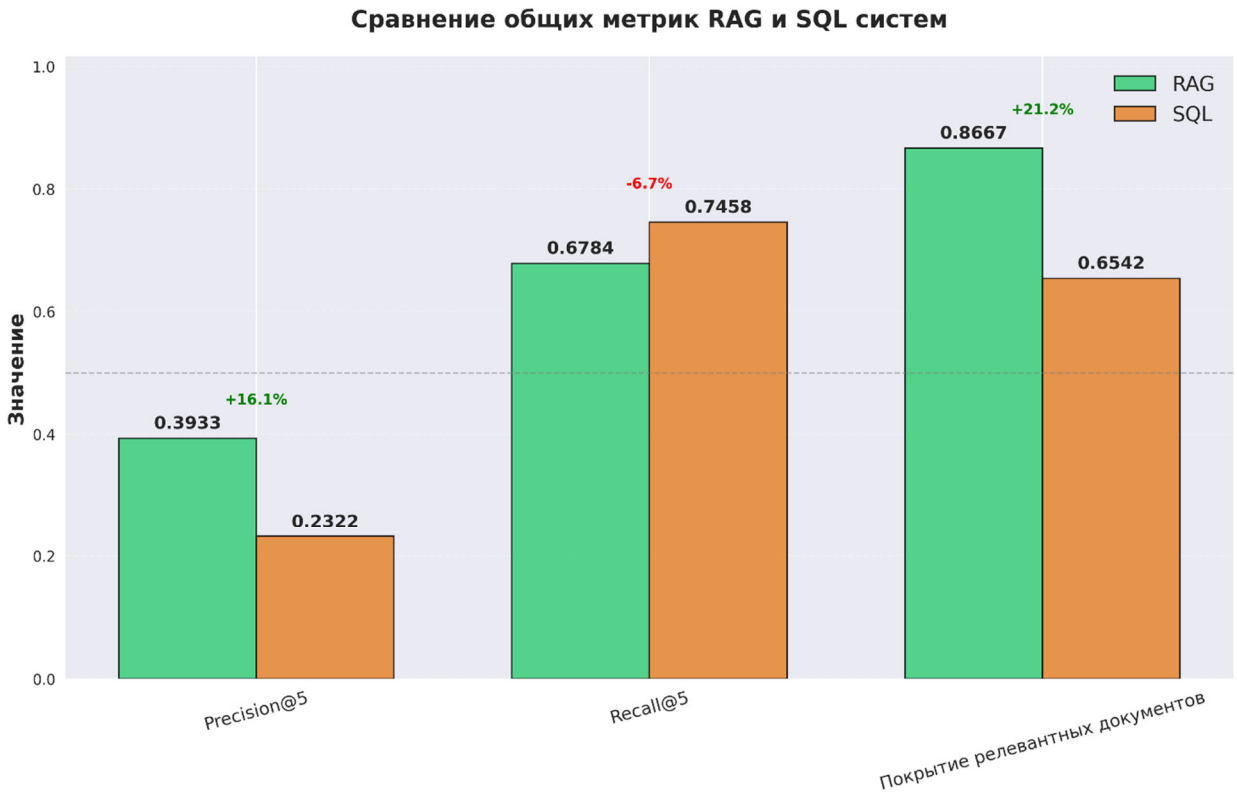

Рисунок 1 - Сравнение ключевых метрик RAG и SQL систем

Анализ рисунка 1 показывает, что система RAG превосходит SQL по двум из трёх метрик. Precision@5 для RAG составляет 0.3933, что на 16% выше по сравнению с результатом SQL (0.2322). Покрытие релевантных документов у RAG также значительно выше — 0.8667 против 0.6542 у SQL, что демонстрирует прирост на 21.2%. Однако по метрике Recall@5 наблюдается незначительное отставание: значение для RAG составляет 0.6784, что на 6.7% ниже по сравнению с SQL (0.7458). В целом, RAG-система показывает заметные преимущества по точности и полноте покрытия релевантных документов [5].

Семантический RAG продемонстрировал заметное превосходство в аспектах пользовательского опыта. NDCG@5 = 0.6518 свидетельствует о высоком качестве ранжирования, при котором наиболее релевантные результаты надёжно находятся в верхних позициях выдачи, что сокращает время их нахождения. MRR = 0.6983 означает, что первый релевантный резуЦльтат появляется в среднем на второй позиции (1/0.6983 ≈ 1.43). Hit Rate@5 = 0.8667 указывает на стабильное обнаружение хотя бы одного релевантного документа в топ-5, исключая «пустые» выдачи [7].

Ключевое преимущество RAG заключается в устойчивости к вариациям формулировок, синонимам и семантическим перефразированиям, даже без точного лексического совпадения. SQL-поиск, опирающийся на строгий текстовый матчинг, не улавливает концептуальное сходство, что снижает его надёжность при изменении терминологии.

Результаты эксперимента убедительно демонстрируют качественное превосходство RAG-архитектуры для задач семантического поиска в библиотечных системах. При сопоставимых показателях полноты RAG обеспечивает значительно более высокую

точность и качество ранжирования, что напрямую улучшает пользовательский опыт. Для модернизации библиотечных каталогов и систем управления научно-технической документацией применение RAG-технологии представляется оптимальным решением, позволяющим минимизировать информационный шум и гарантировать стабильное получение релевантных результатов.

Список литературы:

1. Авдеева, Н. В. Научный поиск: методы тематически-ориентированного поиска научной информации : материалы конференции / Н. В. Авдеева, О. В. Никулина, А. С. Хританков, Ю. В. Чехович ; Российская государственная библиотека, ЗАО «Анти-Плагиат». – Москва, 2014. – URL: http://rcdl.ru/doc/2014/paper/RCDL2014_237-241.pdf (дата обращения: 06.10.2025). – Текст : электронный.
2. Всё про Qdrant. Обзор векторной базы данных. – URL: https://habr.com/ru/companies/amvera/articles/925206/ (дата обращения: 07.10.2025). – Текст : электронный.
3. Рейтинг русскоязычных энкодеров предложений. – URL: https://habr.com/ru/articles/669674/ (дата обращения: 07.10.2025). – Текст : электронный.
4. Стукалова, А. А. Функциональность электронного каталога: требования российских и зарубежных пользователей : отчет / А. А. Стукалова. – Новосибирск : ГПНТБ СО РАН, 2025. – URL: https://ntb.gpntb.ru/jour/article/download/638/561 (дата обращения: 06.10.2025). – Текст : электронный.
5. Тестирование качества работы RAG. Описание и сравнение метрик. – URL: https://habr.com/ru/articles/951222/ (дата обращения: 08.10.2025). – Текст : электронный.
6. Qdrant Similarity search. – URL: https://qdrant.tech/documentation/concepts/search/ (дата обращения: 08.10.2025). – Текст : электронный.
7. RAG: учим искусственный интеллект работать с новыми данными. – URL: https://yandex.cloud/ru/blog/posts/2025/05/retrieval-augmented-generation-basics (дата обращения: 07.10.2025). – Текст : электронный.

References:

1. Avdeeva, N. V. Scientific Search: Methods of Thematically-Oriented Scientific Information Search : conference proceedings / N. V. Avdeeva, O. V. Nikulina, A. S. Khritankov, Yu. V. Chekhovich ; Russian State Library, JSC "Anti-Plagiat". – Moscow, 2014. – URL: http://rcdl.ru/doc/2014/paper/RCDL2014_237-241.pdf (accessed: 06.10.2025). – Text : electronic.
2. Everything About Qdrant. Vector Database Overview. – URL: https://habr.com/ru/companies/amvera/articles/925206/ (accessed: 07.10.2025). – Text : electronic.
3. Rating of Russian-Language Sentence Encoders. – URL: https://habr.com/ru/articles/669674/ (accessed: 07.10.2025). – Text : electronic.
4. Stukalova, A. A. Electronic Catalog Functionality: Requirements of Russian and Foreign Users : report / A. A. Stukalova. – Novosibirsk : SPSTL SB RAS, 2025. – URL:

https://ntb.gpntb.ru/jour/article/download/638/561 (accessed: 06.10.2025). – Text : electronic.

5. Testing the Quality of RAG Performance. Description and Comparison of Metrics. – URL: https://habr.com/ru/articles/951222/ (accessed: 08.10.2025). – Text : electronic.
6. Qdrant Similarity Search. – URL: https://qdrant.tech/documentation/concepts/search/ (accessed: 08.10.2025). – Text : electronic.
7. RAG: Teaching Artificial Intelligence to Work with New Data. – URL: https://yandex.cloud/ru/blog/posts/2025/05/retrieval-augmented-generation-basics (accessed: 07.10.2025). – Text : electronic.