Разработка метода гибридного информационного поиска для RAG систем

Аушев И.А.

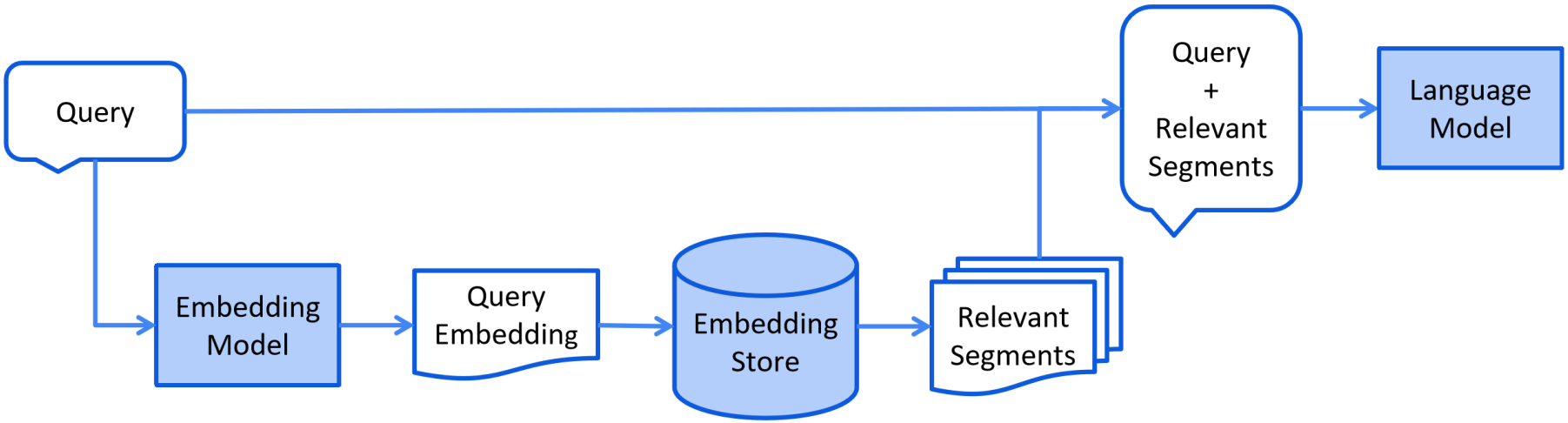
Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар

RAG

Архитектура RAG состоит из двух частей:

1. Информационный поиск
2. Генерация на основе поиска

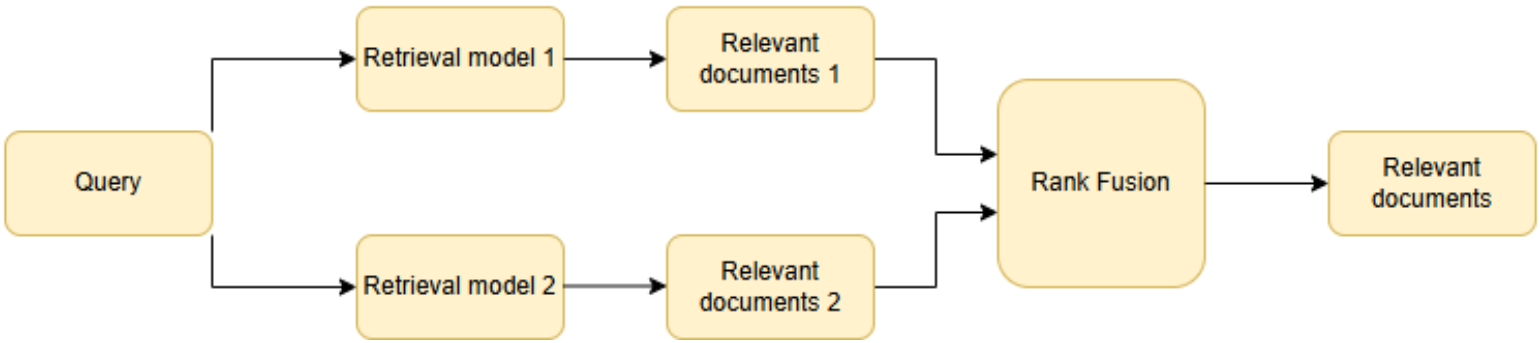


Информационный поиск непосредственно влияет генерацию, поэтому во многих задачах для более качественной генерации требуется использовать продвинутые методы поиска.

Гибридный информационный поиск

Подход, использующий композицию моделей поиска, называется гибридным информационным поиском. Такой подход на практике показывает более качественные результаты[1]

Пример архитектуры гибридного информационного поиска:



Отдельные модели поиска будем называть базовыми моделями.

[1] Priyanka Mandikal и Raymond Mooney. «Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval». В: *CoRR* abs/2401.04055 (2024). DOI: 10.48550/ARXIV.2401.04055. arXiv: 2401.04055. URL: https://doi.org/10.48550/arXiv.2401.04055.

Цель

Цель: Улучшение методов информационного поиска.

Для достижения этой цели была сформулирована **задача**:
Исследование и улучшение гибридных подходов для информационного поиска.

Постановка задачи

Пусть Q – множество текстовых запросов (например, вопросы от пользователей). Имеется база знаний D, элементы D будем называть документами или контекстами.

$R(q, D)$ – модель поиска релевантных документов для $q \in Q$
$R(q, D) = \{d_1, d_2, .., d_k\}$ упорядоченный по релевантности набор документов для запроса q

Задача:
Построить модель $R(q, D)$, используя методы гибридного информационного поиска, улучшая качество выбора релевантных документов.

Виды моделей поиска

Модели информационного поиска можно глобально разделить на два класса:

- Разреженные (BM25, SPLADE[2])
- Плотные (NV-Embed[3], BGE-en-icl[4])

[2] Thibault Formal, Benjamin Piwowarski и Stéphane Clinchant. «SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking». В: Association for Computing Machinery, 2021. ISBN: 9781450380379.

[3] Chankyu Lee и др. «NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models». В: *Submitted to The Thirteenth International Conference on Learning Representations*. under review. 2024. URL: https://openreview.net/forum?id=lgsyLSsDRe.

[4] Chaofan Li и др. «Making Text Embedders Few-Shot Learners». В: *Submitted to The Thirteenth International Conference on Learning Representations*. under review. 2024. URL: https://openreview.net/forum?id=wfLuiDjQ0u.

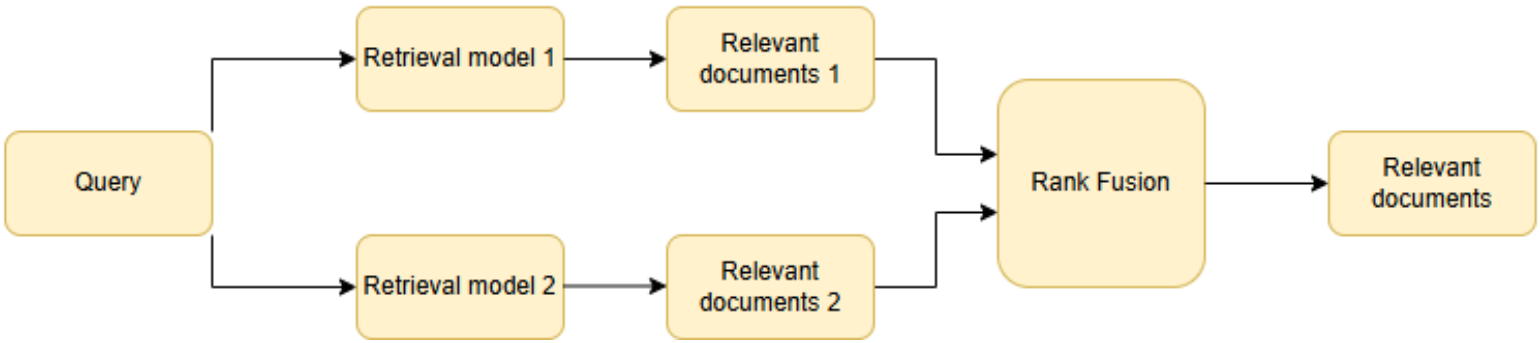
Гибридный информационный поиск

Подходы к гибридному поиску разделяются на 3 стратегии:

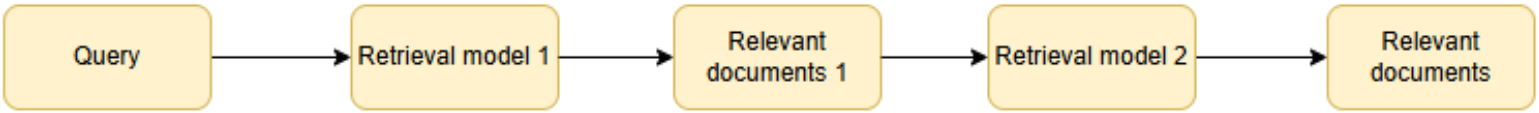
1. Параллельные
2. Последовательные
3. Комбинации 1 и 2 в рамках одной архитектуры

Гибридный информационный поиск

Пример параллельной стратегии:



Пример последовательной стратегии:



Эксперименты

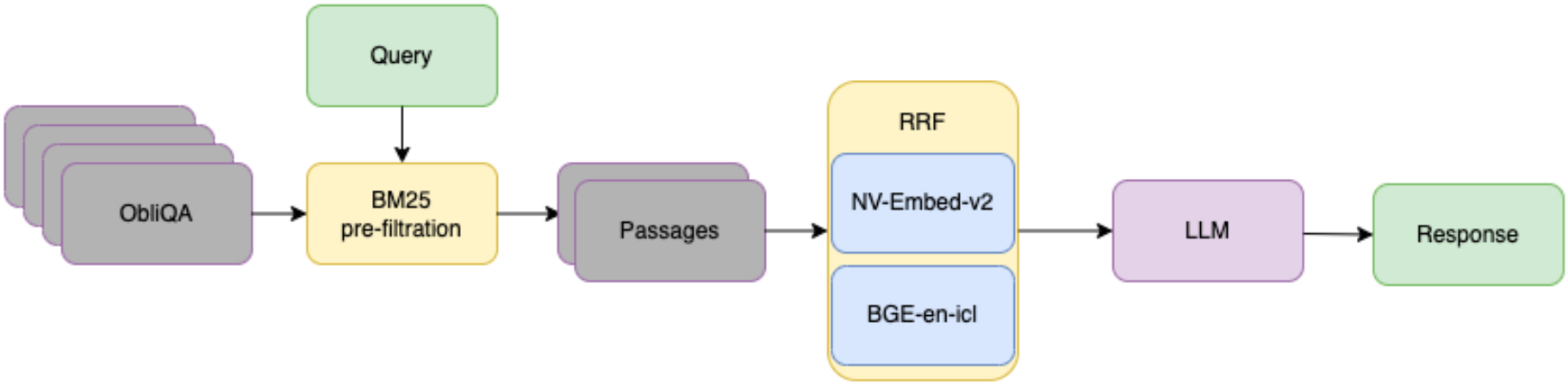
Для экспериментов был взят датасет ObliQA[5].
ObliQA состоит из финансово-нормативных документов, регулирующих правила поведения участников на финансовом рынке и вопросов по этим правилам.

На этих данных был реализован гибридный информационный поиск, включающий в себя две ключевые идеи:

1. Префильтрация исходных документов с помощью BM25, для получения более узкого набора.
2. Параллельный информационный поиск с помощью нескольких моделей(BM25, NV-Embed-v2 и BGE-en-icl). Далее Reciprocal Rank Fusion между всеми моделями.

[5] Tuba Gokhan и др. *RegNLP in Action: Facilitating Compliance Through Automated Information Retrieval and Answer Generation*. Сент. 2024. DOI: 10.48550/arXiv.2409.05677.

Схема предложенного мной метода



Результаты

Модель	F@0	F@100	F@200	F@300	F@500
BGE-en-ICL	80.04	79.69	80.11	**80.22**	80.16
NV-Embed-v2	80.48	79.92	80.45	**80.59**	80.55
BGE-en-ICL + NV-Embed-v2	80.76	80.72	**81.10**	81.03	80.88

Recall@10 предложенного метода. F@n показывает префильтрацию, где топ-n документов получены с помощью BM25. F@0 показывает отсутствие префильтрации. Значения в таблице показывают результат RRF модели с BM25.

Статья с более подробным описанием подхода[6].

[6] Islam Aushev и др. «RAGulator: Effective RAG for Regulatory Question Answering». В: *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*. Под ред. Tuba Gokhan и др. Abu Dhabi, UAE: Association for Computational Linguistics, янв. 2025, с. 114—120. URL: https://aclanthology.org/2025.regnlp-1.18/.