Разработка метода гибридного информационного поиска для RAG систем

Аушев И.А.

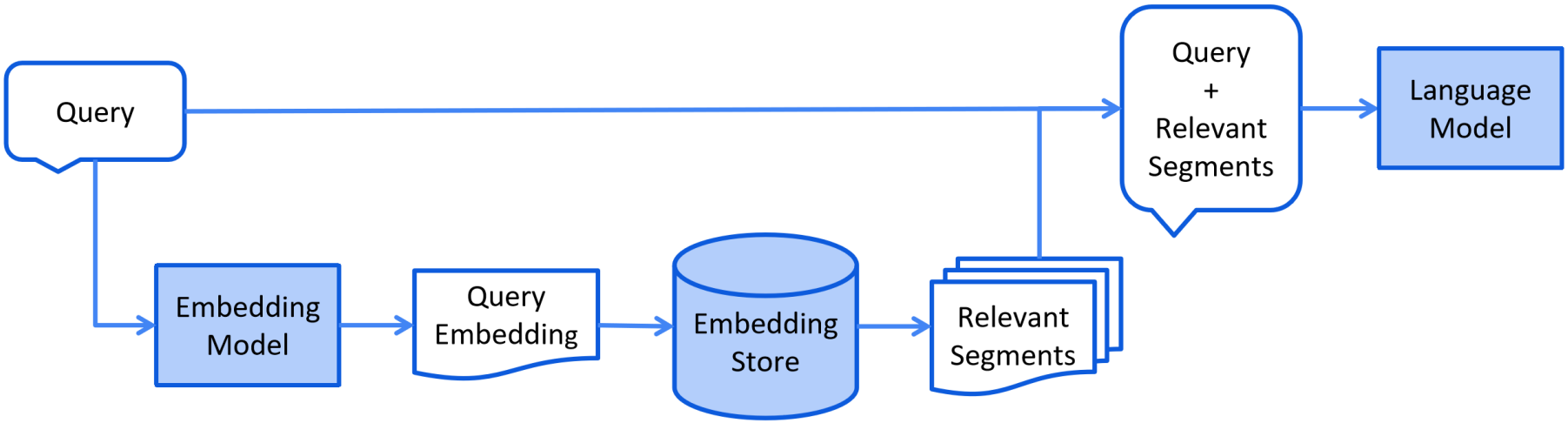
Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар

RAG

Архитектура RAG состоит из двух частей:

1. Информационный поиск
2. Генерация на основе поиска

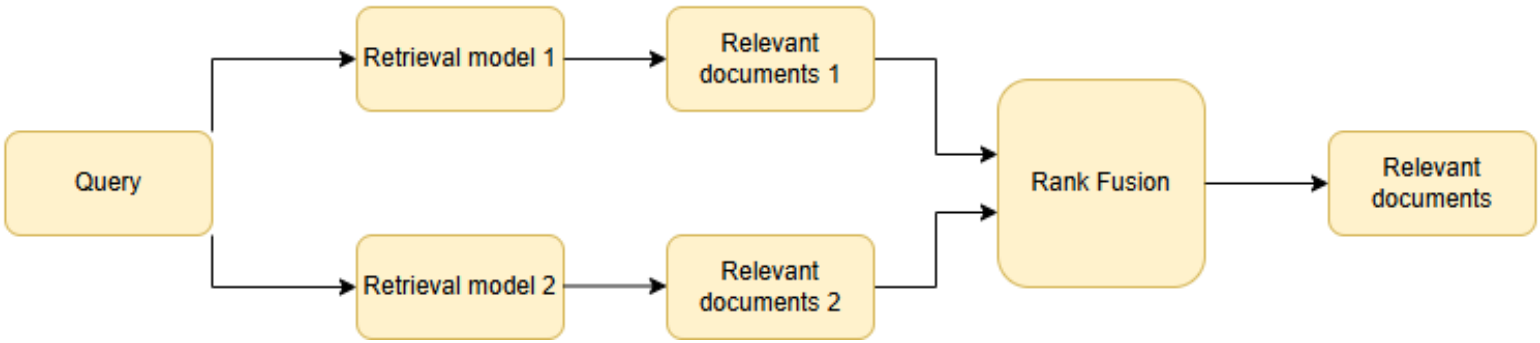


Информационный поиск непосредственно влияет на генерацию, поэтому во многих задачах для более качественной генерации требуется использовать продвинутые методы поиска.

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 сформулированы следующие **задачи**:

1. Изучить имеющиеся подходы для гибридного информационного поиска.
2. Выбрать датасеты для проведения экспериментов.
3. Предложить архитектуры гибридного информационного поиска и проанализировать их качество на выбранных наборах данных.

Постановка задачи гибридного информационного поиска

Пусть

- Q – текстовые запросы
- D – база данных документов
- M – метрика оценки качества поиска

Задача: построить гибридную модель поиска $R(q, D)$ (нейросетевую или классическую), которая для запроса q возвращает список документов $\{d_1, d_2, .., d_k\}$, максимизируя M.

Виды моделей поиска

Модели информационного поиска можно глобально разделить на два класса:

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

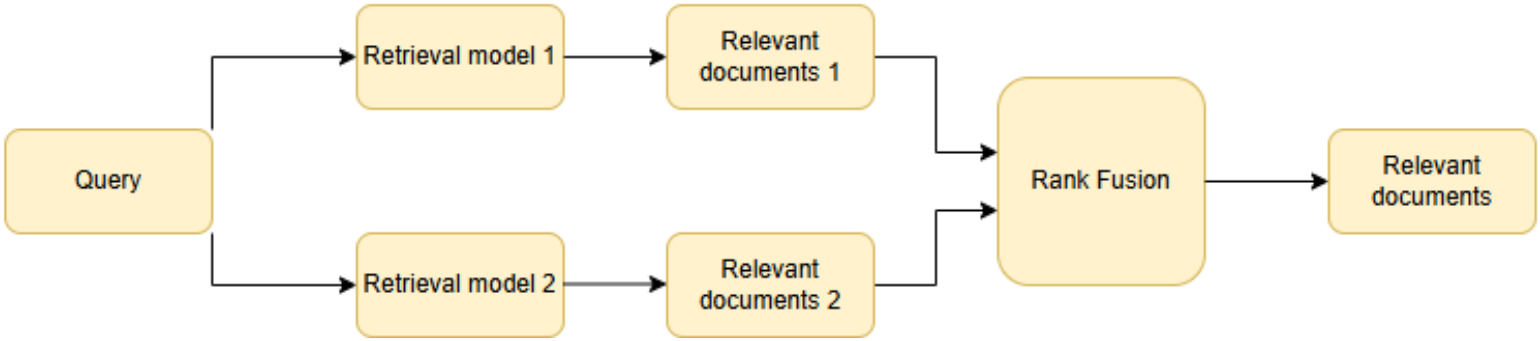
[2] Thibault Formal, Benjamin Piwowarski и Stéphane Clinchant. «SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking». В: Association for Computing Machinery, 2021. ISBN: 9781450380379.

[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.

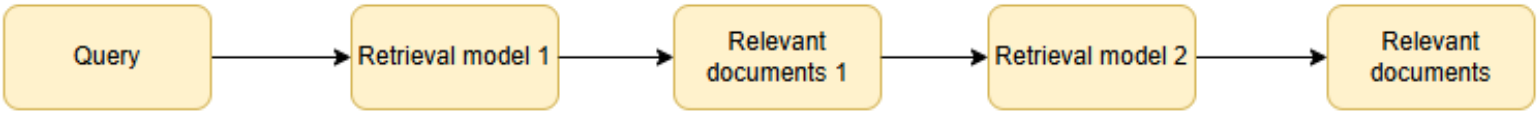
[4] Chaofan Li и др. «Making Text Embedders Few-Shot Learners». В: *Submitted to The Thirteenth International Conference on Learning Representations*. under review. 2024. URL: https://openreview.net/forum?id=wfLuiDjQ0u.

Гибридный информационный поиск

Пример параллельной стратегии:



Пример последовательной стратегии:



Эксперименты

Для экспериментов был взят датасет ObliQA[5].
ObliQA состоит из финансово-нормативных документов и вопросов.

На этих данных был реализован гибридный информационный поиск, включающий в себя 3 ключевые идеи:

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

[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.

Контекстуализация

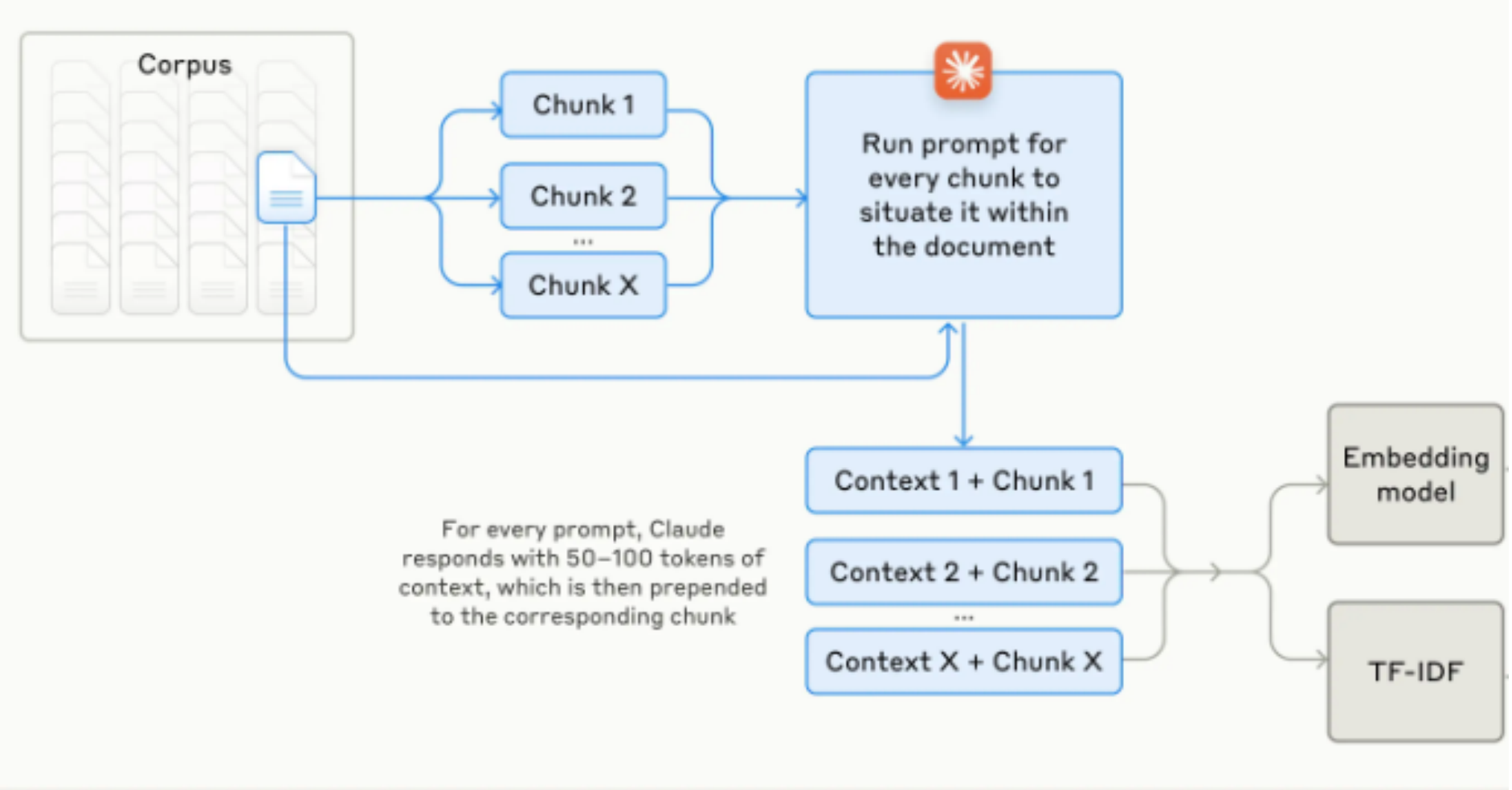
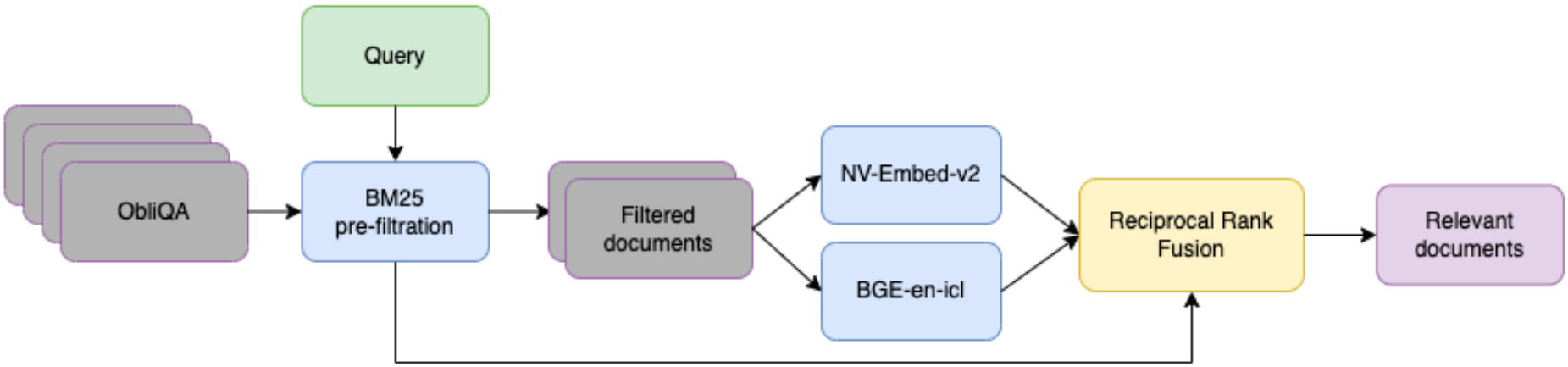


Схема предложенного мной метода



Результаты после контекстуализации

Model	Context	F@0	F@100	F@200	F@300	F@500
BGE-en-ICL	+	75.22	**77.82**	77.22	76.95	76.55
	-	77.39	**78.71**	78.51	78.37	78.02
NV-Embed-v2	+	74.34	**77.21**	76.91	76.36	75.87
	-	78.68	79.02	**80.45**	78.91	78.87

Recall@10 предложенного метода. Context означает есть ли контекстуализация, F@n представляет префильтрацию BM25, где топ-n документов получены с помощью BM25, F@0 обозначает отсутствие префильтрации.

Результаты RRF с BM25

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.