

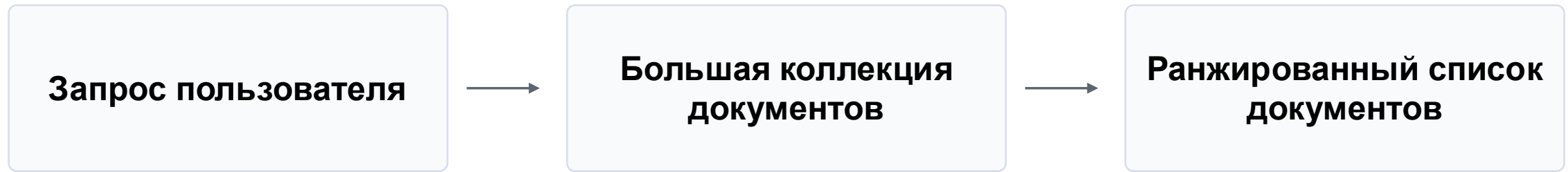
Тестирование гибридных методов информационного поиска для русского языка

RusBEIR Wikifacts · BM25 · Dense Retrieval · RRF · LLM reranking

А. Ю. Кабанов, Н. В. Лукашевич
МГТУ им. Н. Э. Баумана · МГУ им. М. В. Ломоносова

Суперкомпьютерные дни · 28 сентября 2026

Ранжированный информационный поиск



Современные этапы

Поиск — из большой коллекции выбираются наиболее подходящие документы.

Ранжирование — найденные документы упорядочиваются по релевантности.

Переранжирование — верхняя часть выдачи дополнительно уточняется более точной, но дорогой моделью.

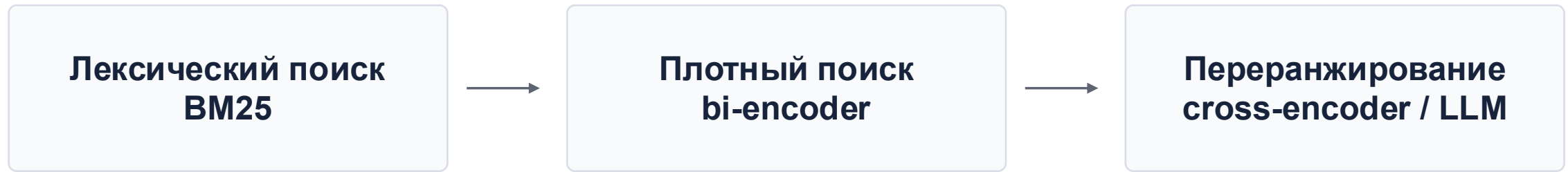
Определение

Информационный поиск — задача нахождения и ранжирования документов, наиболее релевантных пользовательскому запросу.

Цель

Релевантный документ должен оказаться как можно выше в поисковой выдаче.

Подходы к информационному поиску



Проблема

BM25 хорошо работает на длинных документах и терминологии.

Плотный поиск лучше улавливает смысл текста.

Эффективность методов зависит от длины документа.

Для русского языка

Сложная морфология повышает риск несовпадения формы слов.

Публичных русскоязычных бенчмарков меньше, чем англоязычных.

Вывод: разные методы лучше работают на разных типах документов, поэтому их объединяют в многоступенчатую систему.

Цель и вклад работы

Цель

Сравнить эффективность лексических, плотных, комбинированных подходов и методов переранжирования для русскоязычного информационного поиска на корпусах разной длины.

Исследовательские вопросы

Как длина документа влияет на качество BM25 и плотного поиска?

Дает ли комбинация нескольких методов поиска прирост качества?

Что эффективнее для переранжирования:
pointwise или listwise LLM-подход?

Вклад

Системное сравнение лексических, плотных, комбинированных методов и переранжирования.

Анализ нескольких вариантов Wikifacts: статьи, параграфы, предложения, окна.

Отдельная оценка pointwise и listwise LLM-переранжирования.

Сравнение с результатами исходного исследования Wikifacts v0.

Датасет: RusBEIR Wikifacts

**599
запросов**

**≈1.9 тыс.
статей**

**≈15.9 тыс.
параграфов**

**≈188 тыс.
предложений / окон**

Как устроена задача

Факт из Википедии используется как поисковый запрос.

Связанные статьи или фрагменты статьи считаются релевантными документами.

Релевантность размечена по трехуровневой шкале: 0, 1, 2.

Почему Wikifacts подходит для этой работы

Русскоязычный бенчмарк, а не перевод англоязычного эксперимента.

Есть варианты корпуса с разной длиной документов.

Можно сравнить поведение BM25, плотного поиска и переранжирования в одной предметной области.

Пример запроса и разметки Wikifacts

Пример запроса

Дядя Булгакова так и не узнал, что является прототипом профессора Преображенского.

Релевантность 2 – полное подтверждение

... так и не узнал до самой смерти Николай Михайлович Покровский, что послужил прообразом ... Филиппа Филипповича Преображенского ...

Релевантность 1 – частичное подтверждение

Николай Михайлович Покровский ... дядя писателя Михаила Булгакова, основной прототип профессора Филиппа Филипповича Преображенского ...

Пример запроса

Памятник на месте рождения Пушкина стоит не там.

Релевантность 2 – полное подтверждение

... исследования, проведённые уже после установки памятника, показывают, что место рождения Пушкина скорее находится на соседней ... улице.

Релевантность 1 – частичное подтверждение

Памятник Пушкину ... скульптурное изображение ... установленное на месте предполагаемого рождения русского поэта ...

Фрагменты сокращены для представления на слайде.

Варианты корпуса и метрики

Корпус	Единица поиска	Ожидаемая сложность
articles	полная статья	много лексических совпадений
para	параграф	баланс лексики и семантики
sents	предложение	мало слов, важна семантика
window_2..6	2-6 предложений	контекст между предложением и абзацем

Метрики качества

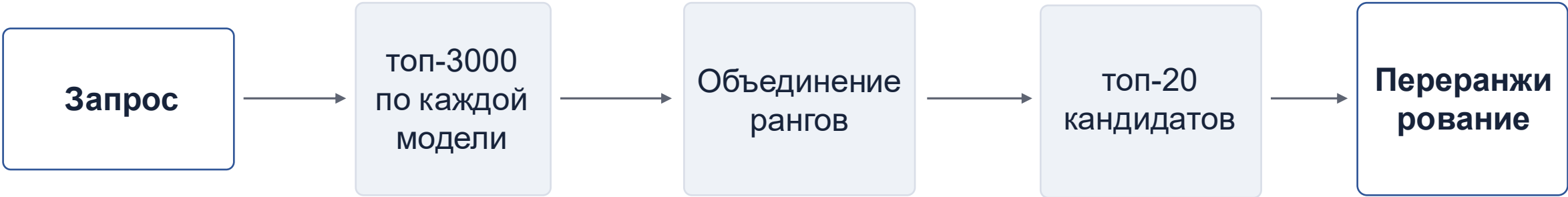
nDCG@10
учитывает позиции и уровни релевантности в топ-10

MRR@10
позиция первого релевантного документа

Интерпретация

Ключевая идея эксперимента: не менять предметную область, а менять длину индексируемого документа и смотреть, какие методы от этого выигрывают.

Общий пайплайн эксперимента



Поисковые компоненты

Группа	Модели
Разреженный	BM25, BM25_Anserini
PRF	BM25 + LLM-расширение запроса
Плотный	E5, GTE, ArcticEmbed-M/L
Переранжирование	BGE, T-lite, RuadaptQwen3

Что сравнивалось

- Начальное качество отдельных поисковых моделей.
- Эффект объединения выдач.
- Качество cross-encoder и LLM-переранжирования.
- Различия между pointwise и listwise стратегиями.

Методы первичного поиска

BM25
лексическое
совпадение
термов

BM25

Сильный базовый метод для полнотекстового поиска.

Учитывает частоты слов, длину документа и редкость терминов.

BM25 + PRF
LLM генерирует
расширенный запрос

LLM PRF

Напишите текст, который отвечает на заданный запрос, основываясь на контексте:

Контекст: {doc1}
{doc2}
{doc3}
Запрос: {query}

Плотный поиск
семантическая
близость
эмбеддингов

Плотный поиск

Использовались E5, GTE, ArcticEmbed-M и ArcticEmbed-L.

Ожидаемое преимущество - короткие тексты и слабое совпадение слов.

Reciprocal Rank Fusion

Модельно-независимое объединение ранжированных списков

$$\text{RRF}(d) = \sum_i 1 / (k + r_i(d))$$

где $r_i(d)$ - позиция документа d в выдаче i -й модели, k - сглаживающий коэффициент.

Почему RRF удобен

Не требует обучения.

Не зависит от абсолютных оценок разных моделей.

Подходит для объединения разреженного и плотного поиска.

Проверялось ограничение

Больше моделей $\neq$ всегда лучше

Если слабая модель добавляет шум, объединение может ухудшить выдачу.

Эффект объединения зависит от длины документов и состава моделей.

Переранжирование: pointwise и listwise

После RRF переранжировались top-20 документов

Pointwise

запрос + один документ
→ оценка релевантности

Listwise

запрос + группа документов
→ перестановка документов

Pointwise

Модель оценивает каждый документ независимо.

Простой и относительно дешевый подход.

Ограничение: документы не сравниваются напрямую.

Listwise

Модель видит сразу несколько кандидатов.

Может напрямую сравнивать документы между собой.

Для контроля контекста использовалось скользящее окно размера 5 с шагом 3.

LLM: T-lite-it-2.1 и RuadaptQwen3-8B-Hybrid; базовая модель: bge-reranker-v2-m3.

Как LLM выполняет переранжирование

Pointwise

Прочитайте запрос и отрывок ниже и укажите, насколько отрывок соответствует запросу. Используйте следующую шкалу:

3 – идеально соответствует: отрывок полностью посвящен запросу и содержит точный ответ.

2 – высокая релевантность: отрывок содержит ответ на запрос, но он может быть не совсем ясным или скрыт среди второстепенной информации.

1 – по теме: отрывок кажется связанным с запросом, но не дает ответа на него.

0 – не релевантно: отрывок не имеет никакого отношения к запросу.

Запрос: {query}

Отрывок: {passage}

Listwise

SYSTEM

Ты – RankGPT, интеллектуальный помощник, который ранжирует пассажи по их релевантности запросу. Я предоставлю тебе {num} пассажиров, каждый из которых помечен идентификатором в скобках []. Твоя задача – отранжировать их в порядке убывания релевантности запросу: {query}.

USER

[{i}] {short_p}...

Запрос: {query}

Отранжируй эти {num} пассажиров выше. Самые релевантные должны быть первыми.

Экспериментальная настройка

Единая схема оценки для всех вариантов корпуса

Этап	Настройка
Первичный поиск	топ-3000 документов на каждый запрос для каждой модели
Комбинация выдач	RRF по спискам BM25_Anserini, BM25+PRF, GTE, Arctic-M, Arctic-L
Кандидаты для переранжирования	топ-20 документов после RRF
Базовая модель переранжирования	bge-reranker-v2-m3 (cross-encoder)
LLM-переранжирование	T-lite-it-2.1 и RuadaptQwen3-8B-Hybrid
Метрики	nDCG@10 и MRR@10

**Промпты были зафиксированы заранее
и не подбирались отдельно под модель**

Результат 1: качество зависит от длины документа

Корпус	Сильнейший компонент	nDCG@10
articles	BM25	0.843
para	T-lite listwise	0.698
sents	T-lite listwise	0.534
window_2	T-lite listwise	0.550
window_6	BGE-переранжировщик	0.625

Интерпретация

На полных статьях BM25 остается сильнейшим: документы достаточно длинные, лексических совпадений много.

На коротких текстах плотный поиск становится конкурентнее, потому что прямых совпадений слов меньше.

Наибольший прирост качества дает переранжирование, особенно на коротких документах.

**Вывод: нет универсальной конфигурации
для всех типов документов**

Результат 2: RRF полезен не всегда

Корпус	Arctic-L	Полный RRF	Комментарий
sents	<u>0.450</u>	0.448	не улучшает
window_2	<u>0.493</u>	0.486	добавляется шум
window_5	0.544	<u>0.554</u>	небольшой прирост
articles	0.729	<u>0.786</u>	большой прирост

Что это означает

RRF может объединять полезные сигналы разреженного и плотного поиска.

Но на коротких текстах лексическая выдача иногда ухудшает результаты плотного поиска.

Простое добавление моделей в RRF не гарантирует рост качества.

Наибольшая эффективность наблюдается на полных документах.

RRF стоит применять селективно: важен состав моделей и длина документа

Результат 3: переранжирование дает основной прирост качества

sents · T-lite
0.463 → 0.534
pointwise → listwise

window_2 · T-lite
0.504 → 0.550
pointwise → listwise

sents · Ruadapt
0.489 → 0.524
pointwise → listwise

Наблюдения

BGE является сильным базовым методом на коротких и средних корпусах.

Listwise-переранжирование часто превосходит pointwise для той же модели.

На нескольких коротких корпусах listwise LLM обходит BGE, но это не универсально.

Почему listwise помогает

Модель сравнивает кандидаты между собой, а не выставляет независимые оценки.

Особенно важно, когда документы короткие и похожи по отдельным словам.

Ограничение: высокая вычислительная стоимость.

Сравнение с предыдущей работой по Wikifacts v0

Корпус	Лучшее ранее	Лучшее в этой работе	Разница
articles	BM25+BGE: <u>0.853</u>	BM25: 0.843	-0.010
para	BM25+BGE: 0.666	T-lite listwise: <u>0.698</u>	+0.032
sents	FRIDA+BGE: 0.494	T-lite listwise: <u>0.534</u>	+0.040
window_2	FRIDA+BGE: 0.522	T-lite listwise: <u>0.550</u>	+0.028
window_3	BM25+BGE: 0.533	BGE: <u>0.563</u>	+0.030
window_6	BM25+BGE: 0.599	BGE: <u>0.625</u>	+0.026

Ключевой вывод сравнения

На статьях результат близок к предыдущей работе: BM25 остается сильным методом.

На коротких документах прирост заметнее: sents и window-корпуса выигрывают от переранжирования.

ArcticEmbed-L превосходит ранее приведенную модель FRIDA на большинстве вариантов корпуса.

**Улучшения связаны прежде всего с моделями
плотного поиска и переранжирования**

Вычислительная стоимость пайплайна

Компонент	Стоимость	Комментарий
BM25	низкая	инвертированный индекс
Плотный поиск	средняя	эмбединги и векторный поиск
RRF	низкая сама по себе	требует запуска нескольких моделей поиска
Cross-encoder	высокая	оценка каждой пары запрос-документ
LLM-переранжирование	очень высокая	инференс LLM; listwise зависит от окна

Практические ограничения

Полные документы не переранжировались LLM из-за контекста и стоимости.

Переранжирование применялось только к топ-20 после RRF.

Для практических систем нужен баланс между качеством и задержкой ответа.

Основная вычислительная проблема – LLM-переранжирование

Основные выводы

- 1** Длина документа - ключевой фактор. BM25 силен на статьях, плотный поиск и переранжирование важнее на коротких текстах.
- 2** RRF не является гарантированным улучшением: эффект объединения зависит от состава моделей и может добавлять шум.
- 3** Переранжирование дает самый заметный прирост качества; listwise-переранжирование на основе LLM чаще превосходит pointwise-подход на коротких документах.
- 4** Эффективный русский поисковый конвейер должен подбираться под тип корпуса, а не строиться как одна универсальная схема.

Ограничения и дальнейшая работа

Ограничения

Использовался один фиксированный промпт для каждой LLM-стратегии.

Систематическое сравнение промптов не проводилось.

LLM-переранжирование не применялось к полным статьям из-за стоимости и длины контекста.

Работа сфокусирована на Wikifacts v0, а не на более полной новой версии датасета.

Дальнейшая работа

Проверить устойчивость listwise-преимущества на других промптах и LLM.

Расширить анализ на другие датасеты RusBEIR.

Исследовать более легкие схемы переранжирования для практических сценариев.

Оценить компромисс качество / задержка ответа / вычислительная стоимость.

Спасибо за внимание

Вопросы?
