

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ПОИСКА В RAG-СИСТЕМАХ

В работе рассматриваются основные подходы к поиску контекста в системах Retrieval-Augmented Generation. Сопоставлены разреженный, плотный и гибридный поиск. Показано, что выбор retrieval-механизма влияет на полноту извлечения релевантных фрагментов, устойчивость работы системы и итоговое качество ответа языковой модели. Сформулированы практические рекомендации по выбору метода поиска в зависимости от характера запросов, особенностей корпуса документов и требований к вычислительным ресурсам.

Введение

В последние годы большие языковые модели активно применяются в интеллектуальных системах поиска, анализа документов и консультирования пользователей. Однако использование только параметрических знаний модели не всегда обеспечивает достаточную точность и актуальность ответа, особенно в тех случаях, когда требуется опора на внешние документы, инструкции, базы знаний или регулярно обновляемые данные. Одним из наиболее перспективных способов решения этой проблемы является подход Retrieval-Augmented Generation, при котором перед генерацией ответа система извлекает релевантные фрагменты из внешнего корпуса документов и передает их модели в качестве контекста [1].

Эффективность RAG-системы во многом определяется качеством этапа retrieval. Если система извлекает нерелевантные или неполные данные, итоговый ответ становится менее точным, а вероятность фактических ошибок возрастает. По этой причине сравнение методов поиска представляет как теоретический, так и практический интерес. Целью работы является сравнительный анализ разреженного, плотного и гибридного поиска в RAG-системах, а также определение условий, в которых каждый из этих подходов оказывается наиболее целесообразным.

I. Методы поиска в RAG-системах

Разреженный поиск (sparse retrieval) основан на лексическом совпадении слов запроса и документа. На практике он хорошо работает в случаях, когда важны точные формулировки, специальные термины, имена сущностей и технические обозначения. Его преимуществами являются простота реализации, низкая вычислительная стоимость и понятная интерпретация результатов. По этой причине данный подход широко применяется в системах, где необходимо быстро получать релевантные документы по коротким и конкретным запросам.

Плотный поиск (dense retrieval) использует векторные представления запросов и документов. Это позволяет учитывать не только совпадение слов, но и смысловую близость текстов. Такой подход особенно полезен, когда запрос и релевантный документ сформулированы разными словами. Исследование Dense Passage Retrieval показало высокую эффективность данного метода в задачах open-domain question answering, что сделало dense retrieval важной частью современных RAG-систем [2]. Дополнительным пре-

имуществом плотного поиска является способность находить документы, близкие по смыслу даже при отсутствии прямого совпадения ключевых слов.

Гибридный поиск объединяет оба подхода. В этом случае система одновременно учитывает точные лексические совпадения и семантическое сходство. Такой вариант оказывается особенно полезным в прикладных системах, где пользовательские запросы могут быть как короткими и точными, так и более свободными по формулировке. На практике именно гибридный подход часто позволяет достичь более устойчивого качества извлечения контекста на неоднородных коллекциях документов.

II. Сравнительный анализ подходов

С практической точки зрения разреженный поиск целесообразно использовать в задачах, где корпус содержит много терминов, кодов, названий библиотек, команд или нормативных формулировок. В таких условиях точное совпадение слов нередко оказывается более важным, чем общая семантическая близость. Кроме того, этот подход удобен в эксплуатации, поскольку не требует построения векторного индекса и сложной предварительной обработки данных.

Плотный поиск предпочтителен в ситуациях, когда запросы формулируются естественным языком и могут существенно отличаться от текста документа. Он лучше справляется с перефразированием и поиском по смыслу, однако требует эмбединговой модели, векторного индекса и больших вычислительных ресурсов. Это делает его более требовательным к инфраструктуре, но одновременно и более гибким при работе со сложными пользовательскими формулировками.

Гибридный поиск можно рассматривать как компромисс между двумя стратегиями. Он обычно обеспечивает более устойчивый результат на разнородных коллекциях документов, поскольку компенсирует недостатки каждого из отдельных методов. Недостатком гибридного подхода является более сложная инфраструктура и увеличение стоимости обработки запроса. Тем не менее в прикладных системах, где важно одновременно учитывать и точность терминов, и общий смысл запроса, такой подход часто оказывается наиболее оправданным.

Для повышения качества поиска в RAG-системах также применяется reranking. На первом этапе система отбирает небольшой набор кандидатов, а затем дополнительно перераспределяет их более

сложной моделью. Такой подход позволяет улучшить верхнюю часть выдачи и, следовательно, повысить качество конечного ответа. Одним из наиболее известных решений этого типа является passage re-ranking на основе BERT [3]. Использование reranking особенно полезно в тех случаях, когда изначально найденные документы в целом релевантны, но нуждаются в более точном упорядочивании.

Следовательно, выбор метода retrieval зависит от типа корпуса, формулировки запросов и допустимых вычислительных затрат. Если на первом месте стоят скорость и простота, более оправдан разреженный поиск. Если требуется лучшее понимание смысла запроса, предпочтительнее плотный подход. При необходимости получить более стабильный результат в реальных условиях наиболее рациональным вариантом становится гибридный поиск.

Заключение

В работе рассмотрены три основных подхода к поиску в RAG-системах: разреженный, плотный и гибридный. Установлено, что разреженный поиск эффективен при точном совпадении терминов, плотный

лучше учитывает смысловую близость, а гибридный обеспечивает более сбалансированный результат. Таким образом, универсального метода для всех сценариев не существует, а выбор retrieval-механизма должен определяться особенностями корпуса документов, характером пользовательских запросов и требованиями к производительности системы.

Именно качество извлеченного контекста во многом определяет содержательность и надежность итогового ответа.

Список литературы

1. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. – NeurIPS, 2020.
2. Karpukhin V., Oguz B., Min S. et al. Dense Passage Retrieval for Open-Domain Question Answering. – EMNLP, 2020.
3. Nogueira R., Cho K. Passage Re-ranking with BERT. – arXiv preprint arXiv:1901.04085, 2019.
4. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. – EMNLP-IJCNLP, 2019.
5. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge University Press, 2009.

Мухомедзянов Егор Сергеевич, студент 2 курса факультета информационных технологий и управления БГУИР, gens4458@gmail.com.

Аникеенко Андрей Дмитриевич, студент 2 курса факультета информационных технологий и управления БГУИР, anikeenko.andrew@list.ru.

Туровец Павел Дмитриевич, студент 2 курса факультета информационных технологий и управления БГУИР, eq1.eq@outlook.com.

Научный руководитель: Зотов Никита Владимирович, ассистент кафедры интеллектуальных информационных технологий БГУИР, n.zotov@bsuir.by