УДК 004.8

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ПОИСКА И ГЕНЕРАЦИИ ОТВЕТОВ НА ОСНОВЕ НЕСТРУКТУРИРОВАННОГО НАБОРА ИНФОРМАЦИИ

Габов М.Д (ИТМО)

Научный руководитель – кандидат технических наук, доцент Федоров Д.А.

(ИТМО)

Введение. Генеративные возможности больших языковых моделей постоянно совершенствуются, но определение области их применения тесно связано с требованиями и ограничениями разрабатываемой системы. В условиях, требующих достоверности в обработке и передаче релевантной информации, требования к минимизации препятствующих признаков, таких как галлюцинации, использование неактуальной информации, являются необходимыми к соблюдению для соответствия результата ожидаемым критериям. Современные подходы, основанные на Retrieval-Augmented Generation (RAG) [4] подразумевают использование методов работы с селективными источниками данных, что позволяет создавать системы, обрабатывающие запросы на естественном языке и способные к структурированию разрозненной информации. Традиционные RAG системы используют фиксированные алгоритмы обработки запросов, что ограничивает их эффективность при работе с расплывчатыми формулировками. Парадигма Agentic RAG [1] предлагает решение данной проблемы, основанное на возможностях языковых моделей к пониманию намерений пользователя, при этом сохраняющее требуемую актуальность и достоверность извлекаемой информации.

В докладе будет рассмотрена реализация Agentic RAG [1] с дополнениями в виде адаптации исследований Microsoft GraphRAG [2], PankRAG [3], а также разработка динамического подхода, позволяющего решать retrieval задачи разного уровня сложности без необходимости дополнительной конфигурации.

Основная часть. Современные RAG системы, такие как LangChain, LlamaIndex, Haystack, требуют работы специалиста для настройки. В то же время, основной целью является достижение приемлемых показателей RAGAS относительно существующих решений с сохранением требований к универсальности системы — минимизации работы специалиста. Ключевые подходы, использованные в работе:

1) **Адаптивный подход.** Система использует единую систему первичной обработки документов, не требующую вызовов LLM. На стороне поиска адаптивность реализуется с помощью анализа запроса большой языковой моделью, на основе которого выбирается стратегия поиска — базовый поиск для простых запросов, декомпозиция для многоаспектных, multihop для многоэтапных запросов;
2) **Декомпозиция запроса.** Для запросов, классифицированных как многоаспектные, LLM составляет несколько независимых подзапросов для параллельного поиска, после чего результаты объединяются посредством Reciprocal Rank Fusion;
3) **Инкрементное построение графа знаний.** Вместо построения полного графа сущностей и связей согласно подходу GraphRAG [2], требующего больших вычислительных затрат, реализовано альтернативное “ленивое” построение графа, подразумевающее его формирование по мере обработки пользовательских запросов, побочным продуктом которых становятся наборы извлеченных сущностей и связей;
4) **Параллельное выполнение подзапросов (DAG).** Данный подход представляет собой адаптацию PankRAG [3] и применим к запросам, требующим многошагового рассуждения. LLM строит направленный ациклический граф подзапросов с указанием

зависимостей. Независимые подзапросы выполняются параллельно, зависимые с подстановкой промежуточных результатов.

Данные улучшения оптимизируют работу системы, позволяя обрабатывать запросы и документы различной степени сложности, не требуя дополнительной настройки, а также соответствуют требованиям работы production-системы.

Выводы. Предложенная комбинация подходов позволяет обрабатывать пользовательские запросы различной степени сложности в рамках единой системы без дополнительной конфигурации, при этом соответствует актуальным критериям точности по RAGAS. Были адаптированы современные подходы с фокусом на сохранение производительности и универсальности системы.

Список использованных источников:

1. Singh A. et al. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG // arXiv preprint arXiv:2501.09136. — 2025. — URL: https://arxiv.org/abs/2501.09136 (дата обращения: 15.02.2026).
2. Edge D. et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization // arXiv preprint arXiv:2404.16130. — 2024. — URL: https://arxiv.org/abs/2404.16130 (дата обращения: 15.02.2026).
3. Li Z. et al. PankRAG: An Efficient Retrieval-Augmented Generation Framework with Globally-Aware Planning and Dependency-Aware Reranking // Proc. IEEE ICASSP. — 2026.
4. Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv preprint arXiv:2005.11401. — 2020. — URL: https://arxiv.org/abs/2005.11401 (дата обращения: 15.02.2026).

Габов М.Д. (автор)

Федоров Д.А. (научный руководитель)