ОПТИМИЗАЦИЯ RAG-АРХИТЕКТУР ДЛЯ ПРЕДМЕТНО-ОРИЕНТИРОВАННЫХ СПРАВОЧНЫХ СИСТЕМ ВУЗОВ

Завалей В. А., Орлов Д. В., Скиба И. Г.
Центр информатизации и инновационных разработок,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {v.zavalej, d.orlov, i.skiba}@bsuir.by

В статье рассматриваются подходы к повышению эффективности RAG-архитектур в справочных системах вузов. Проанализированы проблемы наивных реализаций, связанные с обработкой гетерогенных академических данных. Предложен комплекс методов оптимизации, включающий семантическое чанкирование документов, гибридный поиск и динамическое построение промпта. Разработана система критериев оценки, учитывающая специфику образовательной среды. Результаты работы представляют методическую основу для создания интеллектуальных ассистентов в учебных заведениях.

Введение

Современные высшие учебные заведения характеризуются сложной организационной структурой и значительным объемом внутренней документации, включающей учебные планы, академические регламенты, расписания занятий и методические указания. Обеспечение оперативного и точного доступа к этой информации для студентов, преподавателей и сотрудников представляет собой важную практическую задачу. Универсальные большие языковые модели, такие как GPT, не имеют доступа к актуальной и часто конфиденциальной информации конкретного учреждения, что может приводить к генерации недостоверных или устаревших сведений [1].

Архитектура Retrieval-Augmented Generation (RAG) представляет собой перспективный подход для решения этой проблемы, позволяя сочетать генеративные способности языковых моделей с точным доступом к релевантным внешним источникам данных. Однако стандартные реализации RAG-систем демонстрируют недостаточную эффективность при работе со специализированным контентом высших учебных заведений. Основные проблемы включают неудачное извлечение контекста вследствие неоптимального разделения документов на фрагменты, неспособность адекватно обрабатывать составные запросы и игнорирование специфической терминологии предметной области.

Целью данной работы является теоретическое обоснование и разработка методов оптимизации RAG-архитектуры, адаптированных для нужд справочных систем высших учебных заведений. В качестве основных задач определены: анализ ключевых проблемных мест в конвейере RAG, предложение конкретных методов оптимизации на этапах поиска и генерации, а также разработка системы критериев для сравнительного анализа эффективности предлагаемых решений.

I. Анализ проблем RAG-архитектур

Стандартные реализации RAG-архитектур демонстрируют ряд системных проблем при работе с корпоративными данными вуза. Академическая среда характеризуется высокой структурной и семантической гетерогенностью документов, что существенно усугубляет традиционные ограничения RAG-конвейера.

Ключевой проблемой на этапе извлечения информации является неоптимальное чанкирование документов. Статические методы разбиения текста на фрагменты фиксированной длины часто разрушают логические связи внутри документов. Например, при обработке типичного учебного плана информация о конкретной дисциплине, её кредитной стоимости, семестре изучения и форме контроля знаний может оказаться распределенной по разным чанкам. Это делает невозможным формирование целостного ответа на запрос студента о содержании конкретного курса [2].

Другой существенной проблемой является ограниченная эффективность чистого семантического поиска. Векторные модели, лежащие в основе такого поиска, могут недостаточно точно ранжировать документы при работе с узкопредметными терминами, общепринятыми аббревиатурами и составными запросами, характерными для академической среды. Типичным примером может служить запрос «Как пересдать ЭВМ осенью?», где «ЭВМ» представляет собой аббревиатуру дисциплины «Электронные вычислительные машины» [3].

На этапе генерации ответов возникает проблема игнорирования языковой моделью части предоставленного контекста. Модель может чрезмерно полагаться на собственные параметрические знания, полученные в ходе предварительного обучения, что особенно критично в условиях быстро меняющейся информации образовательного учреждения.

II. Методы оптимизации RAG-конвейера

Для решения выявленных проблемных аспектов предлагается комплекс методов оптимизации, направленных на повышение эффективности ключевых этапов работы RAG-архитектуры в условиях высшего учебного заведения.

Проблема неоптимального чанкирования документов может быть успешно решена за счет перехода от статического разбиения текста к семантическому чанкированию. Данный подход предполагает интеллектуальное разделение документов на логически целостные фрагменты на основе анализа структуры текста и его смысловых границ. Для документов с жесткой структурой, таких как учебные планы или официальные приказы, эффективно использование специально разработанных шаблонов разметки. В случае работы с неструктурированными текстами применяются алгоритмы анализа когерентности, учитывающие тематические переходы.

Для повышения точности извлекаемых документов целесообразно внедрение гибридного поиска, комбинирующего семантический и ключевой методы. Семантический поиск на основе векторных представлений обеспечивает понимание смысловых оттенков запроса, а ключевой поиск гарантирует точное соответствие специфическим терминам и аббревиатурам. Результирующий список формируется на основе взвешенной суммы оценок релевантности [4].

На этапе генерации для уменьшения вероятности игнорирования контекста предлагается метод динамического построения промпта. В отличие от простого объединения чанков, промпт структурируется для направления внимания модели на ключевые аспекты запроса. Дополнительно используется контекстуальная компрессия для снижения семантического шума.

III. Критерии оценки эффективности

Для объективного сравнительного анализа эффективности базовой и оптимизированной RAG-архитектур предлагается использовать систему метрик, ориентированную на релевантность, точность и семантическую целостность ответов.

Оценка релевантности извлеченного контекста проводится с использованием метрик Precision@K и Recall@K. Для автоматизированной оценки может использоваться метрика NDCG (Normalized Discounted Cumulative Gain), учитывающая порядок следования релевантных документов в ранжированном списке, что особенно важно для обеспечения качества финального ответа системы [5].

Качество конечного ответа системы оценивается по точности представленных фактов и семантической согласованности. Точность может измеряться путем экспертного сравнения сгенерированных утверждений с эталонными данными из доверенных источников вуза. Семантическая согласованность оценивает, насколько логически связан и целостен представленный ответ.

Дополнительным практическим критерием является коэффициент успешного завершения диалога, который отражает процент взаимодействий с исчерпывающим ответом без необходимости переформулировать запрос. Эта метрика отражает удовлетворенность пользователя и общую эффективность системы в реальных условиях.

Заключение

В работе проведен анализ системных проблем, характерных для наивных реализаций RAG-архитектур при их применении в предметно-ориентированных справочных системах высших учебных заведений. Выявлены ключевые ограничения на этапах извлечения и генерации, связанные с неоптимальным чанкированием документов, недостаточной релевантностью поиска и некорректной обработкой контекста.

Для решения указанных проблем предложен комплекс методов оптимизации, включающий семантическое чанкирование с учетом типов учебных документов, гибридный поиск, комбинирующий семантический и ключевой подходы, и динамическое построение промпта для улучшения обработки контекста языковой моделью.

Разработана система критериев для оценки эффективности RAG-архитектур, включающая метрики релевантности извлеченного контекста, точности и семантической согласованности генерируемых ответов. Перспективными направлениями являются разработка адаптивных алгоритмов чанкирования и создание специализированных моделей для академической сферы.

Предложенные решения создают методическую основу для построения более надежных и эффективных интеллектуальных справочных систем, способных качественно улучшить информационную поддержку студентов и сотрудников вузов.

IV. Список литературы

1. Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis et al. // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459–9474.
2. Gao, L. et al. RAG vs Fine-tuning: A Comprehensive Comparison / L. Gao et al. // arXiv preprint arXiv:2305.16983. – 2023.
3. Hofstätter, S. et al. Hybrid Retrieval Models / S. Hofstätter et al. // ACM Transactions on Information Systems. – 2021. – Vol. 39, № 4. – P. 1–35.
4. Wang, L. et al. Precise Zero-shot Dense Retrieval without Relevance Labels / L. Wang et al. // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. – 2022. – P. 176–186.
5. Savchenko, A. V. Semantic Methods of Text Processing / A. V. Savchenko. – Minsk: BSU, 2020. – 215 p.