УДК 004.89

АДАПТИВНЫЙ ВЫБОР МЕТОДОВ ПОИСКА И ОБРАБОТКИ В СИСТЕМАХ ИИ НА ОСНОВЕ КЛАССИФИКАЦИИ ЗАПРОСА И КОЛЛЕКЦИИ

Максимов А.В.
Научный руководитель – канд. техн. наук, Бутаков Н.А.
Университет ИТМО

Введение

Современные системы искусственного интеллекта используют различные методы извлечения и обработки информации. Помимо классических моделей информационного поиска применяются Retrieval-Augmented Generation (RAG) [1], его графовые расширения (GraphRAG) [2], многошаговые стратегии поиска (DeepSearch) [3], структурированное и полуструктурированное представление, предполагающие использование SQL и NoSQL баз данных, а также генерация ответов на основе знаний, заложенных в модель на этапе обучения. Каждый из этих режимов демонстрирует различную эффективность в зависимости от характера запроса и данных в наличии.

Исследования показывают, что выбор представления существенно влияет на качество поиска. Классическая эвристическая модель BM25 эффективна для лексического поиска по текстовым корпусам. Векторные представления и поиск демонстрируют преимущество при семантическом сопоставлении текстов. Для задач, требующих анализа связей, применяются модели с использованием графов знаний.

Несмотря на наличие этих результатов, выбор режима поиска и обработки в системах ИИ часто фиксирован архитектурно. Отсутствует формализованный механизм, который учитывает одновременно характеристики запроса и тип имеющихся данных для рекомендации подходящих методов.

Основная часть

В работе предлагается метод выбора стратегий поиска в системах искусственного интеллекта на основе анализа двух компонентов: свойств запроса и характеристик коллекции данных.

Метод включает два этапа. На первом выполняется предварительный анализ данных в коллекции с целью выбора способа их представления и соответствующей подготовки. На втором этапе анализируется пользовательский запрос, на основании чего формируется рекомендация по выбору стратегий поиска и формирования ответа.

Запрос анализируется по намерению пользователя, например на требования к точности извлечения, необходимости анализа связей между сущностями, агрегированию, необходимости обобщения или анализа структурированной информации. Эти признаки определяют, полезен ли лексический либо семантический поиск, требуется ли многошаговый поиск.

Коллекция описывается через характеристики данных и доступные их представления. Может варьироваться структура самих данных (например, много текста, таблицы, графовые представления), наличие в них элементов разных модальностей (изображения, таблицы, текст), существует ли возможность выполнять специфические виды поиска (по документам, по фрагментам, по графу знаний, по метаданным)

На основе классификации запроса и коллекции формируется рекомендация по выбору методов поиска и обработки. Рекомендация может включать один или

несколько режимов, таких как лексический поиск, векторный поиск, GraphRAG, DeepSearch, генерация без использования поиска.

Система формирует приоритетный набор методов, которые потенциально обеспечивают корректную обработку конкретной комбинации запроса и данных. Такой механизм позволяет использовать сильные стороны различных подходов без архитектурной фиксации одного режима.

Планируется оценить влияние использования предложенного механизма выбора методов на корректность формируемых ответов.

Выводы

Предложен механизм выбора методов поиска и обработки информации в системах ИИ на основе классификации запроса и данных. В основе классификации лежит анализ намерения пользователя и специфики доступной для поиска информации. Предлагаемый механизм формирует рекомендации по применению различных режимов поиска и обработки данных. Это обеспечивает адаптивность системы при работе с гетерогенными коллекциями данных и различными типами задач.

Литература

1. Karpukhin V., Oğuz B., Min S. и др. Dense Passage Retrieval for Open-Domain Question Answering // EMNLP 2020 arXiv:2004.04906
2. Hu Y., Lei Z., Zhang Z и др. GRAG: Graph Retrieval-Augmented Generation // NAACL 2025 [Электронный ресурс]. - Режим доступа https://aclanthology.org/2025.findings-naacl.232/
3. Alzubi S., Brooks C., Chiniya P. и др. Open Deep Search: Democratizing Search with Open-source Reasoning Agents // arXiv - 2025 - arXiv:2503.20201