УДК 004.896

Исследование и внедрение технологий Retrieval-Augmented Generation (RAG)

Шперлинг М.А. (ИТМО), **Тамразов В.К.** (ИТМО)
Научный руководитель – **Фёдоров Д.А** (ИТМО).

Введение Retrieval-Augmented Generation (RAG) — современный подход, объединяющий методы поиска информации и генерации текста для повышения качества и релевантности ответов. В исследовании изучены и внедрены две ключевые технологии: RAPTOR (иерархическая организация данных) и Contextual Retrieval (контекстный поиск). Основная цель — создание гибридной системы, способной эффективно обрабатывать большие объёмы данных (до 500 ГБ) без потери качества.

Основная часть 1. Изучение RAPTOR. RAPTOR (Recursive Abstractive Processing for Tree-Organized Retrieval) — метод иерархической организации данных, структурирующий информацию в виде дерева. - Реализован алгоритм рекурсивной кластеризации текста. - Внедрена оптимизация индексации, сократившая время поиска на 30%. 2. Изучение Contextual Retrieval. Contextual Retrieval учитывает контекст запроса для повышения релевантности поиска. - Интегрированы модели Vikhr для семантического анализа. - Реализован механизм ранжирования результатов на основе контекста. 3. Создание гибридной системы. Гибридная система объединяет RAPTOR и Contextual Retrieval. - Оптимизированы процессы индексирования и распределённой обработки данных. - Внедрён механизм автоматического выбора метода поиска в зависимости от типа запроса. 4.Оптимизация производительности - Внедрены методы сжатия данных (квантование векторов, Delta Encoding). - Реализовано кэширование запросов (LRU-кэш). - Проведено тестирование на датасетах объёмом до 500 ГБ.

Выводы - RAPTOR доказал эффективность для иерархической организации данных. - Contextual Retrieval повысил релевантность поиска. - Гибридная система показала высокую производительность и стабильность. - Оптимизация сжатия и кэширования ускорила обработку запросов.

Список использованных источников:

1.Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., & Manning, C. D. (2024). RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. arXiv preprint arXiv:2401.18059.
arxiv.org
2. Nikolich, K., Shavrina, T., Fenogenova, A., & Mikhailov, V. (2024). Vikhr: Constructing a State-of-the-art Bilingual Open-Source Instruction-Following Language Model. arXiv preprint arXiv:2405.13929.
arxiv.org
3. Vikhrmodels. (2024). Vikhr-Nemo-12B-Instruct-R-21-09-24. Hugging Face.
huggingface.co
4.Vikhrmodels. (2024). Vikhr-Llama3.1-8B-Instruct-R-21-09-24. Hugging Face.
huggingface.co
5. Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., & Manning, C. D. (2024). RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. OpenReview.
openreview.net

Авторы Шперлинг М.А, Тамразов В.К
Научный руководитель Фёдоров Д.А.