Адаптация методов оценки для сравнения GraphRAG подходов в доменной области технической документации машиностроительной PLM

Науки о данных

Подготовил: Гладышев Виталий Владимирович

Проблема и контекст исследования

Ключевая проблема — сложная структура и фрагментация PLM-данных:

- Разнородная документация (CAD, CAE, CAPP, MDM) и обилие нормативной базы (ГОСТ) создают барьеры для поиска знаний.
- Стандартные RAG-системы не справляются с многостадийным извлечением (поиск операций для детали в сборке, разработанной по конкретному стандарту).

Контекст и пробелы — GraphRAG позволяет найти решение применяя графы знаний, однако:

- Отсутствуют специализированные методы оценки для инженерных доменов (PLM).
- Существующие бенчмарки (GraphRAG-Bench) ориентированы на нерелевантные сценарии (медицина, литература), игнорируя особенности PLM:
  - Иерархия «сборка-деталь»,
  - Нормативные ссылки и версионность.

Цель исследования

Реализовать методологию и программное решение для комплексной оценки и сравнения GraphRAG-систем, применяемых для анализа и обработки технической документации в машиностроительной PLM-среде.

Объект исследования:

Конвейер GraphRAG

Предмет исследования:

Семантическое моделирование с помощью графов

Исследовательский вопрос:
Как обосновать необходимость применения GraphRAG, выполнить оценку и выбор решений в доменой области PLM?

МФТИ

Задачи исследования

1. Провести анализ существующих подходов к построению и оценке GraphRAG-систем.
2. Разработать и собрать специализированный оценочный датасет на основе реальных и синтетических PLM-документов, включающий иерархическую структуру и задачи различной сложности (фактический поиск, многопереходное рассуждение, суммаризация).
3. Создать доменно-адаптированную методологию оценки, для учета специфики PLM-документации, включая метрики качества графа, поиска и генерации.
4. Интегрировать онтологическую валидацию в процесс оценки.
5. Реализовать прототип программного решения для сравнительного анализа различных GraphRAG-систем на созданном датасете.
6. Провести экспериментальное сравнение различных подходов в контексте PLM.

Анализ источников

Тренды и прорывы — применение графов знаний для RAG:

- Microsoft GraphRAG (Edge, 2024) и RAPTOR (Sarthi, 2024) используют иерархические графы для комплексного извлечения контекста.
- HippoRAG2 (Gutiérrez, 2025) демонстрирует высокую связность знаний.

Пробелы в исследованиях — Отсутствие PLM-специфичных методов:

- Обзор Zhang et al. (2025) подтверждает: 90% работ фокусируются на медицине, праве и литературе.
- Нет решений для строгих инженерных зависимостей (сборка-деталь, стандарты ГОСТ/ISO).
- HotpotQA и MultiHop-RAG (Tang & Yang, 2024) не охватывают иерархию, версионность и нормативные зависимости PLM.

Дефицит оценочных метрик:

- Существующие бенчмарки (GraphRAG-Bench, HotpotQA) игнорируют специфику технической документации.

МФТИ

Актуальность и новизна

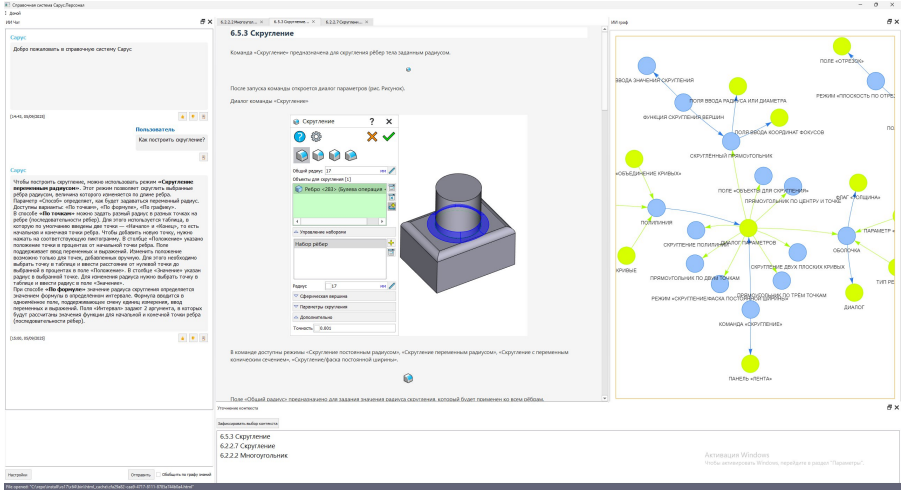
1. Специализированный датасет для PLM-домена - Будет создан оценочный датасет, имитирующий структуру и сложность реальной PLM-документации.
2. Доменно-адаптированная методология оценки - Будет предложена расширенная методология, которая адаптирует существующие метрики GraphRAG-Bench к уникальным характеристикам PLM-документации, включая иерархические и нормативные зависимости.
3. Интеграция онтологической валидации в оценку GraphRAG для проверки семантической целостности и корректности процесса извлечения знаний.
4. Реализация программного решения для практической оценки обоснованности применения GraphRAG и сравнения альтернативных решений для доменной области PLM.

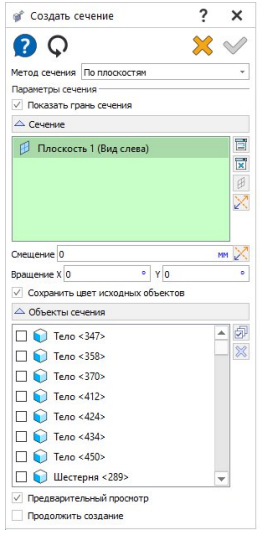
Гипотеза исследования

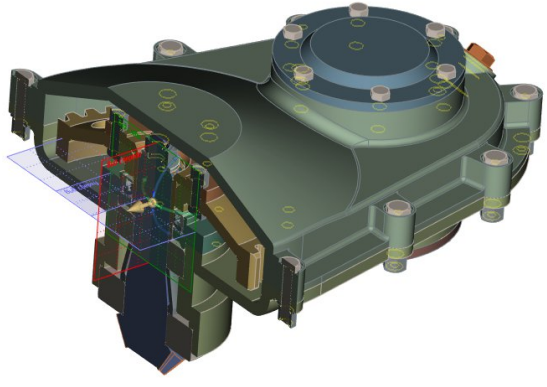
Применение GraphRAG в доменно-специфичном корпусе технической PLM-документации, демонстрирует более высокое качество ответов на сложные, структурированные запросы (требующие многопереходного рассуждения и понимания иерархий) по сравнению с классическим (векторным) RAG.

Специфика домена PLM и GraphRAG решений

- Сложная структура
- Неоднородность данных
- Наличие большого количества взаимосвязей
- Большое количество источников данных







Применение GraphRAG подходов для комплексного представления информации в домене PLM

МФТИ

Создание оценочного датасета PLM-RAG-Bench

Исходным корпусом послужит документация к модулям PLM системы

Выполняется разработка заданий по трём уровням сложности, адаптированным под PLM:

1. Фактический поиск: Вопросы, требующие прямого извлечения одного факта. Пример: «Какой материал указан для детали ‘Вал-12’?»
2. Простое рассуждение: Вопросы, требующие одного шага логического вывода через одну промежуточную сущность. Пример: «Какой стандарт регулирует изготовление детали ‘Корпус-34’, которая входит в сборку ‘Насос-56’?»
3. Сложное рассуждение и суммаризация: Вопросы, требующие синтеза информации из нескольких источников и выполнения логических операций И/ИЛИ. Пример: «Какие виды стали подходят для изделия ‘Вал-3’ в соответствии с ГОСТ на изделий и ТЗ.»

Дизайн эксперимента

В эксперименте будут сравниваться три системы:

1. **Обычный RAG (Baseline):** Документы разбиваются на чанки. Эмбеддинги создаются с помощью intfloat/multilingual-e5-large. Поиск — семантический с top-K=5. Генерация — gpt-oss-120.
2. **Microsoft GraphRAG:** Официальная реализация Microsoft GraphRAG. Используется та же LLM (gpt-oss-120) для всех этапов (извлечение сущностей и отношений, генерация сводок сообществ, финальная генерация ответа). Поиск будет выполняться в режиме Local Search.
3. **HippoRAG:** Реализация, основанная на архитектуре HippoRAG. Система сначала извлекает сущности и отношения из текста с помощью gpt-oss-120, строит на их основе знаменевой граф и применяет алгоритм PageRank для ранжирования узлов по их "важности".

Все три системы используют одну и ту же LLM (gpt-oss-120) для генерации конечного ответа, чтобы изолировать влияние архитектуры индексации и поиска от влияния генератора.

Метрики

Оценка извлечения

- **Evidence Recall:** Оценивается, какая доля эталонных ключевых фактов (сущностей/отношений) была успешно извлечена системой. Рассчитывается как Recall = (Извлеченные эталонные факты) / (Все эталонные факты).

Оценка финального ответа

- **Answer Accuracy:** Комплексная метрика, оценивающая, насколько ответ является корректным и полным. LLM-as-a-judge сравнивает ответ с эталоном и выставляет оценку по критериям фактической точности и полноты (шкала от 0 до 5).
- **Faithfulness:** Оценивает, насколько все утверждения в ответе подтверждены извлеченным контекстом. Проверка на наличие галлюцинаций с использованием LLM-as-a-judge.
- **ROUGE-L:** Метрика для измерения лексического перекрытия между ответом и эталоном.

Вычислительная эффективность

- Время от получения запроса до выдачи ответа (в секундах).
- Количество токенов в исходном запросе и извлеченном контексте, отправленных LLM для генерации.

Список ключевых источников

1. Edge D. et al. From local to global: A graph rag approach to query-focused summarization //arXiv preprint arXiv:2404.16130. – 2024.
2. Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., & Su, Y. (2024). HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. Advances in Neural Information Processing Systems (NeurIPS).
3. Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., & Su, Y. (2025). From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. International Conference on Machine Learning (ICML).
4. Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2024). LightRAG: Simple and Fast Retrieval-Augmented Generation. arXiv preprint arXiv:2410.05775.
5. Han, H., Ma, L., Shomer, H., Wang, Y., Lei, Y., Guo, K., & Li, Z. (2025). RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. arXiv preprint arXiv:2502.11371.
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Список ключевых источников

7. Sarthi, P., Abdullah, S., Tuli, A., & Manning, C. D. (2024). RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. The Twelfth International Conference on Learning Representations (ICLR).
8. Zhang, Q., Chen, S., Bei, Y., Liu, X., & Huang, X. (2025). A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models. arXiv preprint arXiv:2501.13958.
9. Edge, D., Trinh, H., & Larson, J. (2024). LazyGraphRAG: Setting a New Standard for Quality and Cost. Microsoft Research Blog.
10. Yang, Z., Qi, P., Zhang, S., Bengio, Y., & Cohen, W. W. (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
11. Tang, Y., & Yang, Y. (2024). MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. arXiv preprint arXiv:2401.15391.
12. Li, Z., Yu, X., Zhang, W., & Liu, S. (2024). StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-Time Hybrid Information Structurization. International Conference on Learning Representations (ICLR).

Спасибо
за внимание!

МФТИ