

Домашнее задание 3. Гипотеза, новизна, актуальность

Гладышев Виталий Владимирович

1 Анализ источников

При исследовании источников выявлено несколько ключевых направлений для проведения исследования. Важным аспектом является изучение метрик и методов для сравнения графов, что служит основой для дальнейшего анализа в разных областях их применения. Графовые нейронные сети играют важную роль в обработке текстов и их анализе, существенно повышая качество извлечения информации и выполнения классификации. Также анализ структуры графов способствует более глубокому пониманию сложных взаимосвязей в данных, что критично для задач, связанных с RAG. Исследование комплексных графовых моделей, включая взвешенные и ориентированные графы, открывает новые возможности для их адаптации в задачах генерации текста и извлечения информации. Наконец, интеграция методов RAG с подходами, основанными на графах, способна повысить качество генерации и понимания текстов во множестве доменных областей.

Определена новая методика построения семантических моделей построенных на графах – применение больших языковых моделей для получения более качественных и сложных моделей [1].

2 Формулировка гипотезы

Гипотеза: Использование семантических графов для моделирования контекста в RAG-конвейере реализовать сравнение семантической структуры вопроса пользователя и контекста с привязкой к отдельным элементам вопроса и контекста в отличии от существующих метрик таких как Faithfulness или Answer Accuracy. Соответственно можно связать количественные параметры со конкретными элементами семантической структуры.

Объект исследования: Конвейер RAG (retrieval-augmented generation).

Предмет исследования: Семантическое моделирование с помощью графов для оценки качества RAG.

Исследовательский вопрос: Как графовые представления влияют на объяснимость и точность оценки этапов извлечения и генерации в RAG? Гипотеза конкретизирует связь между графовыми моделями и метриками качества, сохраняя фокус на объекте и предмете.

Проблема: Недостаточная объяснимость традиционных методов оценки RAG, основанных на векторных представлениях, затрудняет анализ ошибок в retrieval и генерации. В литературе обсуждаются: преимущества графов для структурирования контекста [1], важность интерпретируемости в NLP [2]. Однако не исследовано, как графы улучшают прозрачность оценки RAG. Гипотеза предлагает решение через интеграцию семантических графов, позволяющих явно отслеживать связи между извлечёнными данными и сгенерированным текстом.

Работы по RAG [3] и графам [4] рассматривают эти области изолированно. Гипотеза объединяет их, фокусируясь на объяснимости — аспекте, критичном для доверия к AI-системам [5].

3 Обоснование новизны исследования

Исследование источников выявило наличие теоретических лакун. Отсутствие унифицированных метрик для оценки графов в контексте RAG. Большинство работ [4, 6] предлагают общие метрики сравнения графов, но не адаптируют их для оценки качества извлечения и генерации в RAG.

Выявлен недостаток исследований по интеграции графовых нейронных сетей (GNN) с RAG. Хотя GNN активно применяются в NLP [2], их роль в улучшении объяснимости оценки RAG остаётся неисследованной. Обнаружена слабая связь между структурой графов и семантикой контекста. Работы по текстовым графам [7] фокусируются на классификации, но не на интерпретации связей для оценки качества генерации.

Обнаружен недостаток средств анализа и визуализации - практические лакуны. Отсутствие инструментов для визуализации и анализа графов в RAG. Методы вроде graphkernels [8] не интегрированы в конвейеры RAG, что затрудняет их применение для объяснимой оценки. Ограниченное использование графов для улучшения фазы извлечения. Исследования RAG [1] акцентируют генерацию, игнорируя оптимизацию retrieval через графовые представления. Нет стандартизированных датасетов для тестирования графовых подходов в

RAG. Существующие бенчмарки [9] не включают аннотированные графовые структуры, что усложняет валидацию.

Данные лакуны должны быть закрыты в ходе моего исследования. Предусмотрена разработка специализированного аналитического средства оценки графов для RAG, сочетающих структурный и семантический анализ включая создание инструмента визуализации, связывающего графовые представления контекста и вопроса пользователя.

4 Актуальность исследования

4.1 Контекст исследования

Рост использования RAG-систем в критически важных областях (медицина, финансы, юридический анализ) требует повышения доверия к их результатам. Однако, как отмечают [5], большинство пользователей сомневаются в обоснованности решений, сгенерированных LLM, из-за низкого уровня интерпретируемости. В научной сфере актуальность подтверждается увеличением числа публикаций по RAG кратно за 2023–2024 гг. (по данным Semantic Scholar), а также ростом интереса к графовым методам в NLP.

4.2 Своевременность исследования

Исследование заполняет пробел между теоретическими работами по графам [4] и прикладными задачами RAG [1], предлагая метод, который одновременно улучшает точность и прозрачность оценки. Появление мультимодальных LLM (например, GPT-4o) требует новых методов оценки, способных работать с гетерогенными данными. Графы, в отличие от векторов, сохраняют структурные и семантические связи между текстом, изображениями и знаниями [10].

5 Методы для проверки гипотезы

Предполагается применение и развитие подхода предложенного в работе [1] и реализованного в библиотеке GraphRAG применительно к конвейеру RAG для численного анализа семантических признаков вопроса пользователя и контекста получаемого семантическим поиском.

В настоящее время не существует аналитических средств для объяснимой оценки качества конвейера RAG на базе семантического моделирования с помощью графов.

Большая часть метрик оценки RAG либо узкоспециальна (оценивает один узкий аспект), либо сложны в интерпретации. Например многие метрики используют большие языковые модели как судью и это вносит субъективность в оценку.

В качестве источника данных будет использован датасет Vikhrmodels/Grounded-RAG-RU-v2, а также специально сгенерированные синтетические наборы данных.

В отличие от традиционных методов вместо оценки через LLM-судью (как в RAGAS) будут использованы графы для объективной количественной интерпретации. Соответственно данный функционал будет реализован вновь и обладает высоким уровнем актуальности и новизны.

СПИСОК 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. Wu L. et al. Graph neural networks for natural language processing: A survey //Foundations and Trends® in Machine Learning. – 2023. – Т. 16. – №. 2. – С. 119-328.
3. He X. et al. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering //Advances in Neural Information Processing Systems. – 2024. – Т. 37. – С. 132876-132907.
4. Wills P., Meyer F. G. Metrics for graph comparison: a practitioner's guide //Plos one. – 2020. – Т. 15. – №. 2. – С. e0228728.
5. Sui Y. et al. Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering //arXiv preprint arXiv:2410.08085. – 2024.
6. Shervashidze N. et al. Efficient graphlet kernels for large graph comparison //Artificial intelligence and statistics. – PMLR, 2009. – С. 488-495.
7. Huang L. et al. Text level graph neural network for text classification //arXiv preprint arXiv:1910.02356. – 2019.
8. Sugiyama M. et al. graphkernels: R and Python packages for graph comparison //Bioinformatics. – 2018. – Т. 34. – №. 3. – С. 530-532.
9. Auer S. et al. The sciqqa scientific question answering benchmark for scholarly knowledge //Scientific Reports. – 2023. – Т. 13. – №. 1. – С. 7240.
10. Peng C. et al. Knowledge graphs: Opportunities and challenges //Artificial Intelligence Review. – 2023. – Т. 56. – №. 11. – С. 13071-13102.