SINH VĂN BẢN TĂNG CƯỜNG TRUY XUẤT CÓ NHẬN THỨC ĐỒ THỊ: KHAI THÁC ĐỒ THỊ TRI THỨC VÀ ĐÁNH GIÁ DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN

GRAPH-AWARE RETRIEVAL-AUGMENTED GENERATION: LEVERAGING KNOWLEDGE GRAPHS AND LLM-BASED EVALUATION

GV: PGS.TS. Lê Đình Duy

Họ tên: Âu Hồng Ân - 240101033

Tóm tắt

- Lớp: CS2205.FEB2025
- Link Github: https://github.com/anah19uit/CS2205.FEB2025
- Link YouTube video:
- Tên thành viên: Âu Hồng Ân - MSSV : 240101033

Giới thiệu

Vấn đề

- LLMs & RAG phát triển nhanh → vẫn gặp thách thức khi khai thác tri thức ngoài dữ liệu huấn luyện
- Naive RAG [1] có cải thiện truy xuất nhưng còn hạn chế là truy xuất sai, kém tổng quát hóa, lặp phản hồi, thiếu suy luận quan hệ giữa thực thể → gây "ảo giác".

Giải pháp

- Xây dựng đồ thị tri thức với GPT-4 kết hợp với RAG là GraphRAG [2] và tổng hợp ý nghĩa toàn cục (global sensemaking).
- Kỹ thuật LLM-as-a-judge [3] được dùng để tạo câu hỏi đa dạng và đánh giá giữa hai hệ thống RAG.

Kết quả

- Graph RAG vượt trội hơn Naive RAG, Vector Rag, đặc biệt câu hỏi mở, không có đáp án cố định.

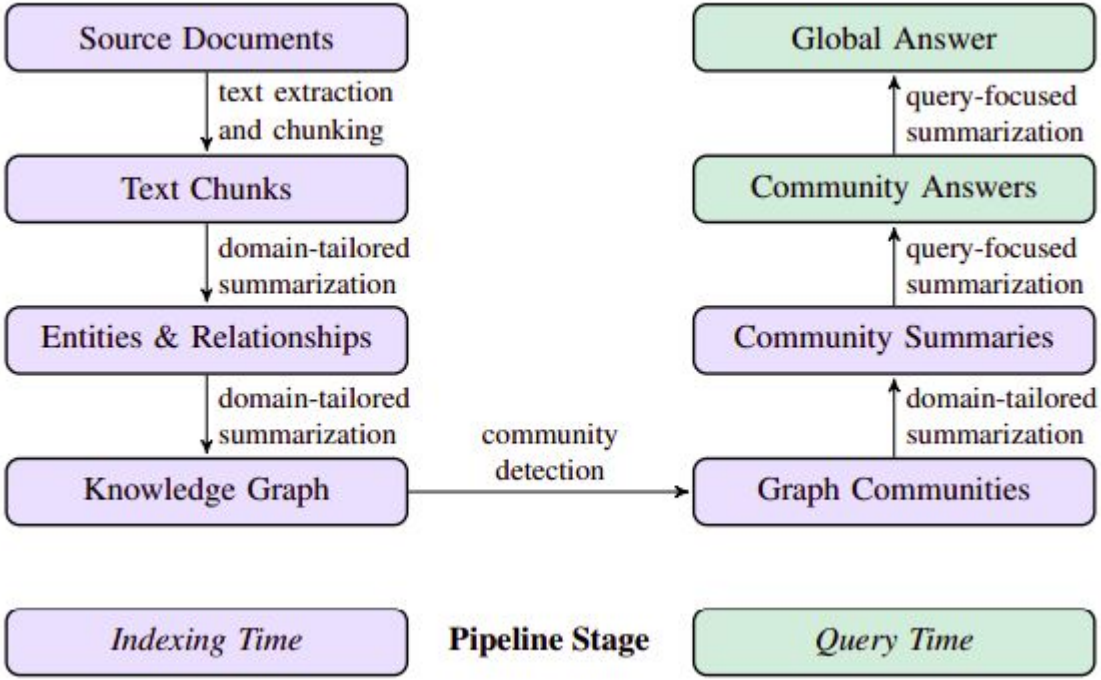
Input: Yêu cầu từ user, đồ thị tri thức

Output: Sinh ra đoạn văn bản đáp án ứng yêu cầu người dùng

Mục tiêu

- Đảm bảo phản hồi từ LLMs phù hợp yêu cầu người dùng
- Xây đồ thị tri thức không cần GPT-4 như [2], dùng phương pháp từ [4], cho phép mở rộng đồ thị từ phản hồi người dùng trong tương lai.
- Giải quyết bài toán: xác định mức độ liên quan giữa văn bản và ảnh trong văn bản nguồn.

Nội dung và Phương pháp



Hình 1 [2]: Các bước thực hiện trong GraphRAG

Nội dung và Phương pháp

Xây dựng đồ thị tri thức và phân cụm đồ thị bằng kỹ thuật tính toán trước (precomputing technique) sử dụng GPT4 tương ứng với tài liệu. Thực hiện 6 bước sau:

1. **Chia đoạn văn nhỏ**: Tách tài liệu thành các đoạn nhỏ; đoạn lớn tiết kiệm chi phí (ít gọi LLM hơn) nhưng có thể giảm khả năng nhớ lại thông tin.
2. **Trích xuất thông tin**: LLM trích xuất thực thể, mối quan hệ, khẳng định từ từng đoạn; mô tả chi tiết được tạo ra và sử dụng prompt để tăng độ chính xác.

Nội dung và Phương pháp (tt)

3. **Xây dựng đồ thị tri thức**: Các thực thể trở thành nút (nodes), mối quan hệ trở thành cạnh (edges), và claims, các thực thể trùng lặp được hợp nhất.Trọng số cạnh (edge weights) là tần suất xuất hiện của mối quan hệ.

4. **Phân cụm đồ thị từ đồ thị tri thức**: Các cộng đồng (communities - cụm các nút liên quan) được phát hiện theo cấu trúc phân cấp sử dụng thuật toán Leiden [5]. Cho phép tóm tắt theo phương pháp chia để trị (divide-and-conquer) ở nhiều mức độ chi tiết.

Nội dung và Phương pháp (tt)

5. Tóm tắt communities từ đồ thị:

- Mỗi cộng đồng được tóm tắt ưu tiên các nút/cạnh quan trọng và claims.
- Tóm tắt theo hai mức: cộng đồng lá (Leaf-level communities) và cấp cao ((Higher-level communities).

6. Trả lời câu hỏi từ tóm tắt:

- Xử lý truy vấn bằng cách so khớp với communities, chia thành các đoạn, dùng LLM khác để trả lời từng đoạn và chấm điểm (0-100).
- Các câu trả lời có điểm cao nhất sẽ được tổng hợp lại thành câu trả lời cuối cùng.

Kết quả dự kiến

- Hiệu quả nội dung và đa dạng phản hồi: Phương pháp vượt trội hơn RAG Naive nhờ khả năng bao quát thông tin đầy đủ và tạo ra các câu trả lời đa dạng về nội dung lẫn cấu trúc.
- Tối ưu hóa tài nguyên: Giảm số lượng token tiêu thụ nhờ áp dụng các cấp độ tóm tắt cộng đồng, đặc biệt ở cấp độ tóm tắt khái quát.
- Khả năng trích xuất thông tin: Truy xuất thông tin liên quan hiệu quả và điều chỉnh chiến lược tóm tắt phù hợp với loại dữ liệu.
- Hỗ trợ người dùng tốt hơn: Cung cấp ví dụ, trích dẫn chính xác và cải thiện trải nghiệm sử dụng, phụ thuộc vào chất lượng prompt trích xuất.

Tài liệu tham khảo

- [1] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, Haofen Wang:
- Retrieval-Augmented Generation for Large Language Models: A Survey. CoRR abs/2312.10997 (2023)
- [2] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Jonathan Larson:
- From Local to Global: A Graph RAG Approach to Query-Focused Summarization. CoRR abs/2404.16130 (2024)
- [3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Zhuang Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, et al.:
- Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 36 (2024)
- [4] Marvin Hofer, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, Erhard Rahm:
- Construction of Knowledge Graphs: State and Challenges. CoRR abs/2206.09508 (2022)
- [5] Vincent A. Traag, Ludo Waltman, Nees Jan van Eck:
- From Louvain to Leiden: Guaranteeing Well-Connected Communities. CoRR abs/1810.08473 (2018)