Improving Retrieval with Graph Pruning

Aaron Toh, Frederick Amal Emerson, Hong Yi, Yong Ee

CP3108B-13

Abstract

This project investigates the use of graph pruning algorithms to enhance both the performance and efficiency of graph-based retrieval augmented generation (RAG) systems.

Motivation

Graph-structured data are used across domains such as social networks, biological systems, and recommendation engines. However, the scale and complexity of modern graphs can make it inefficient and hard to scale. There is a strong need for scalable approaches that retain the benefits of graph-structured data while reducing their size and complexity.

Methodology

1) Pruning Strategies

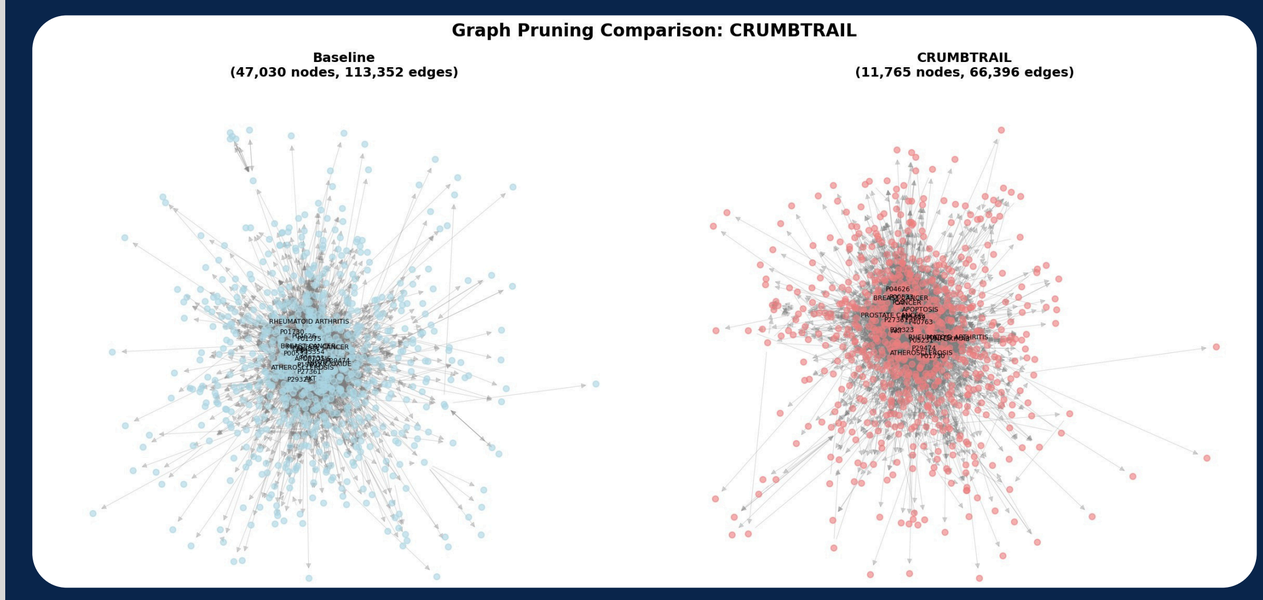
- **CrumbTrail**: A bottom-up, iterative layering algorithm that progressively removes cyclic or weakly supported nodes while guaranteeing that all protected entities remain reachable from the designated root.
- **KGTrimmer**: A single-pass knowledge-graph pruning method that combines community, structural, and semantic importance scores to excise the lowest-ranked nodes, explicitly preserving hubs and bridge vertices to maintain connectivity.
- **PathRAG Hybrid**: A flow-aware strategy that first retains the highest-scoring PathRAG paths and then supplements them with the strongest standalone nodes, yielding a balanced subgraph that is able to keep critical relational corridors.
- **POG Hybrid**: A semantic-path approach that preserves the top SBERT-ranked POG paths and augments them with individually important nodes, ensuring both concept coverage and path coherence in the pruned graph.

2) Evaluation Methods

- **Goal**: measure whether pruning improves RAG retrieval quality while maintaining efficiency.
- **Evaluation Metrics**:
  - **Answer Accuracy** (0-1, higher better): The number of questions answered correctly.
  - **Mean Reciprocal Rank (MRR)** (0-1, higher better): Retrieval quality metric measuring rank of relevant documents.
  - **Recall** (0-1, higher better): Retrieval quality metric measuring the number of ground truth documents retrieved.
  - **Response Time**: Average query latency in seconds (lower better).

Dataset

The **PubMedQA** dataset was initially used, a question-answering dataset where each question has one supporting document. Pruning showed minimal impact, both pruned and unpruned graphs achieved near-perfect accuracy.
Thus, the **MedHop** dataset was used instead. It evaluates multi-hop reasoning. Each question requires context from approximately 50 documents. This dataset was sufficiently complex for the graph RAG system to be fully utilised. It was used in our final pruning evaluation.



Strategy	Entity Reduction	Answer Accuracy	Doc MRR	Doc Recall	Response Time/s
UNPRUNED	0.00%	0.1576	0.5053	0.1784	53.61
CRUMBTRAIL	74.98%	0.1636	0.4277	0.1633	31.66
POG HYBRID	54.98%	0.1536	0.4615	0.1697	38.06
PATHRAG HYBRID	50.20%	0.1497	0.4721	0.1722	40.39
KGTRIMMER	55.23%	0.1437	0.4721	0.1726	38.12

Results

“ Pruning the knowledge graph significantly reduces RAG system latency and size. ”

Efficiency

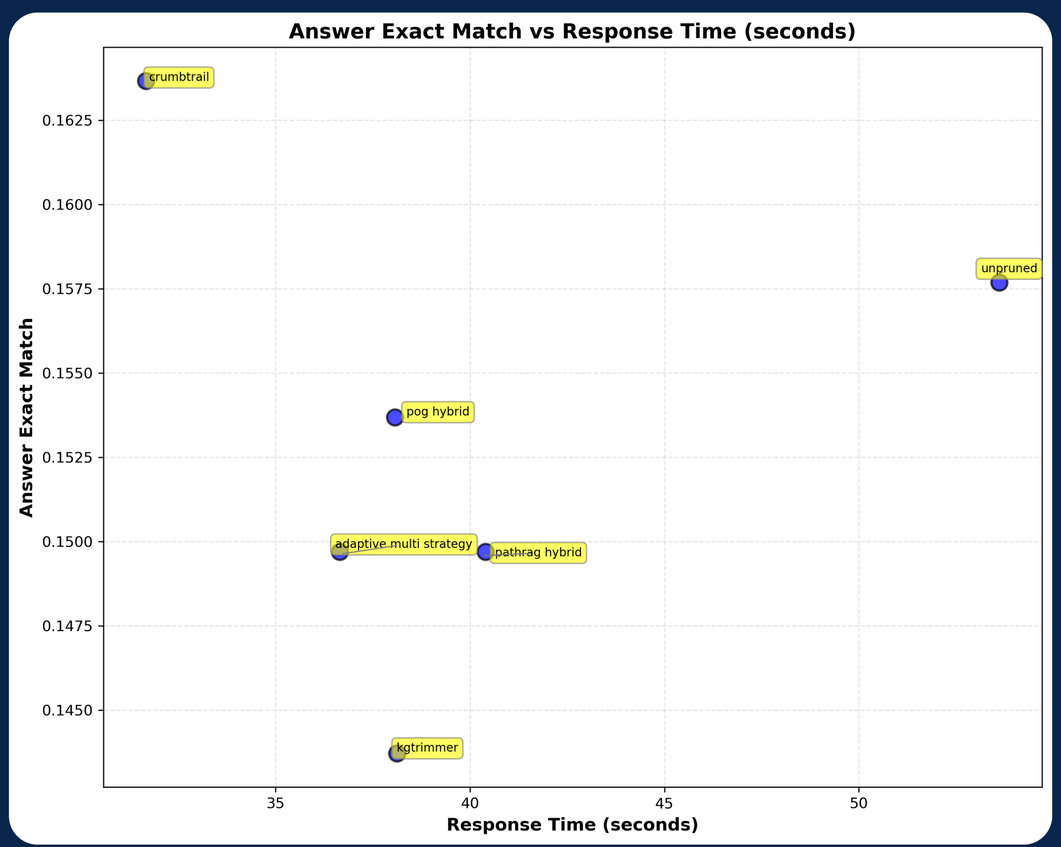
Response times were dramatically reduced by all pruning methods. Crumbtrail was the fastest, with a 41% speed up over the unpruned baseline.
Graph density was also reduced by 50-75%, thus lowering the computation and memory overhead for running the retrieval

The Trade-off

This speed gain came at a cost. Most methods saw a slight decline in retrieval quality (MRR and Recall). However, this was not directly correlated with answer accuracy

Key Insight

Crumbtrail pruning was the only one to improve final answer accuracy (+3.8%), despite its lower intermediate retrieval metrics.
This could have been caused by a combination of crumbtrail limiting retrieval to the most relevant ground truth documents, and its lower number of documents retrieved, reducing the impact of the “lost in the middle” effect.



Questions to Ponder

Does pruning create a "sweet spot" in graph density where retrieval speed doubles and long-tail recall improves, and can we predict this point from the original graph’s structure alone?

Why does removing weak links sometimes boost answer accuracy, and can we design a pruning rule that reliably concentrates evidence without ever seeing the queries?

References

1. Chen, B., Guo, Z., Yang, Z., Chen, Y., Chen, J., Liu, Z., Shi, C., & Yang, C. (2025, February 18). PathRAG: Pruning Graph-based Retrieval Augmented Generation with Relational Paths. arXiv.org. https://arxiv.org/abs/2502.14902
2. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024, April 24). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv.org. https://arxiv.org/abs/2404.16130
3. Faralli, S., Finocchi, I., Ponzetto, S. P., & Velardi, P. (2018). CrumbTrail: An efficient methodology to reduce multiple inheritance in knowledge graphs. Knowledge-Based Systems, 151, 180–197. https://doi.org/10.1016/j.knosys.2018.03.030
4. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019, September 13). PubMedQA: a dataset for biomedical research question answering. arXiv.org. https://arxiv.org/abs/1909.06146
5. Lin, F., Zhu, X., Zhao, Z., Huang, D., Yu, Y., Li, X., Zheng, Z., Xu, T., & Chen, E. (2024, May 19). Knowledge graph pruning for recommendation. arXiv.org. https://arxiv.org/abs/2405.11531
6. Welbl, J., Stenetorp, P., & Riedel, S. (2018). Constructing datasets for multi-hop reading comprehension across documents. Transactions of the Association for Computational Linguistics, 6, 287–302.