NLRAG:하이브리드 검색 구조 기반 RAG 프레임워크

송도연[1], 문에스더[1], 조준희[2], 문남미[3]
[1]호서대학교 빅데이터AI학과 학부생
[2]호서대학교 컴퓨터공학부 학부생
[3]호서대학교 컴퓨터공학부 교수
thdehdus0101@gmail.com, esterstst@gmail.com,
jjunhuui@gmail.com, nammee.moon@gmail.com

NLRAG:RAG Framework Based on Hybrid Search Structure

Doyeon Song[1], Esther Moon[1], Jun-Hui Cho[2], Nammee Moon[2]
[1]Dept. of Department of Big Data and AI, Hoseo University
[2]Dept. of Computer Engineering, Hoseo University

요 약

본 연구는 검색 증강 생성(Retrieval-Augmented Generation, RAG) 시스템의 고질적인 문제인 정확도와 효율성 간의 상충 관계를 해결하고자 새로운 모델 NLRAG(Node-Light RAG)를 제안한다. 기존의 그래프 기반 NodeRAG는 복잡한 관계 추론으로 정확도가 높지만 응답 속도가 느리고, 경량 벡터 기반 LightRAG는 빠르지만 문맥적 이해가 부족한 한계가 있다. 이를 해결하기 위해 NLRAG는 경량 벡터 인덱스를 이용한 1차 검색과 이종 그래프의 구조적 정보를 활용하는 2차 검색을 결합한 하이브리드 방식을 사용한다. 실험 결과, 제안 모델은 응답 속도의 저하를 최소화하면서도 베이스라인 모델 대비 F1-Score 기준 정확도를 0.0324 향상시켜 기존의 상충 관계를 효과적으로 해결했다.

1. 서론

최근 대규모 언어모델(LLM) 기반 검색 증강 생성은 외부 지식을 효과적으로 활용할 수 있는 방법으로 주목받고 있다. 그러나 기존의 단순 유사도 검색 방식은 문서 간의 복잡한 구조적 관계를 충분히 반영하지 못하는 한계가 있다[1]. 이를 극복하기 위해 제안된 NodeRAG는 문서 간의 관계를 이종 그래프 형태로 모델링하여 높은 검색 정확도를 제공하지만, 복잡한 연산으로 인해 속도가 느리고 시스템 부담이 크다는 단점이 있다[2]. 반면 LightRAG는 경량화된 검색 기법을 통해 빠른 응답 속도를 보장한다[3]. 이는 이원적 검색 시스템을 통해 특정 정보와 광범위한 주제를 모두 효과적으로 탐색하도록 설계되었지만 문서의 관계 정보를 충분히 활용하지 못해 정확도 면에서는 다소 한계가 존재한다[3]. 이처럼 기존 방법론들은 정확성과 효율성이라는 두 가지 중요한 요소가 상충하는 문제점을 갖고 있다. 따라서 RAG의 발전을 위해서는 두 요소를 균형 있게 달성할 수 있는 통합적 접근이 필요하다. 본 연구는 이러한 필요성에 따라 NodeRAG의 정확도와 LightRAG의 효율성을 동시에 확보할 수 있는 새로운 검색 방법 NLRAG를 제안한다.

2. NLRAG

NLRAG는 복잡한 질문에 대한 관계성 추론에 강점을 보이는 NodeRAG와 단순한 질문에 즉각적으로 응답하는 LightRAG의 장점을 통합한다.

2.1 데이터 전처리 및 이종 그래프 구축

노드 정의: 문서에서 엔티티, 속성, 관계, 그리고 텍스트 청크와 같은 다양한 유형의 노드를 정의한다.

엣지 연결: 정의된 노드들 사이의 의미적 연결 관계를 엣지로 표현하여 이종 그래프를 구축한다. 이 그래프는 복잡한 문서 구조와 정보를 효과적으로 담아낸다.

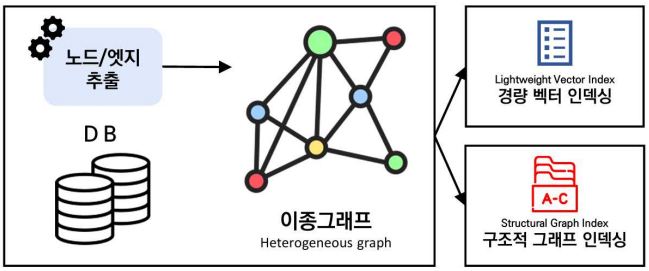
2.2 이중 레벨 인덱싱

NLRAG는 검색 효율성과 정확성을 동시에 확보하기 위해 다음과 같은 두 가지 인덱스를 구축한다.

경량 벡터 인덱스: 구축된 그래프의 각 노드를 임베딩하여 벡터 공간에 저장한다. 이 인덱스는 사용자의 질의와 노드 간의 빠른 유사도 검색에 활용된다.

구조적 그래프 인덱스: 노드 간의 복잡한 연결 관계를 담은 그래프 구조 자체를 저장한다. 이 인덱스는 복잡하고 다중 홉(multi-hop) 질의를 처리하는 데 사용된다.

데이터 전처리와 이중 레벨 인덱싱의 개념도는 (그림1)과 같다.



(그림 1) 이중 레벨 인덱싱 개념도

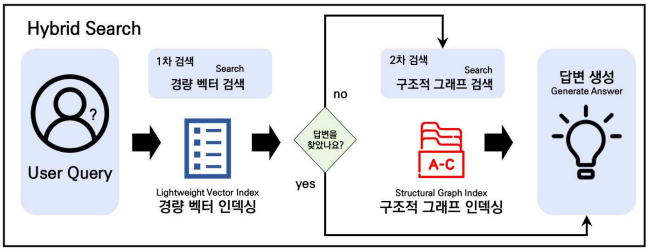
2.3 하이브리드 검색 및 답변 생성

NLRAG는 사용자의 질의에 대해 두 가지 인덱스를 활용한 하이브리드 검색을 수행하여 최적의 답변을 생성한다.

경량 검색 (1차 검색): 먼저 사용자의 질의를 경량 벡터 인덱스와 비교하여 빠르게 유사도 검색을 수행한다. 만약 간단한 질문이거나 명확한 답변이 즉시 발견되면, 이 단계에서 답변을 생성하고 검색을 종료한다.

구조적 검색 (2차 검색): 1차 검색만으로 답변을 찾기 어렵거나, 질의가 여러 노드의 복잡한 관계를 필요로 할 경우, 구조적 그래프 인덱스를 활용한 그래프 탐색을 시작한다. 이 단계에서는 여러 노드를 순차적으로 따라가는 다중 홉 검색을 수행하여 더 넓고 정확한 정보를 찾을 수 있다.

하이브리드 검색 시스템 개념도는 (그림2)와 같다.



(그림 2) 하이브리드 검색 시스템 개념도

3. 실험

3.1 실험 설계

본 연구에서는 모델의 복합적 추론 능력을 평가하기 위해 다중 홉 질의응답(multi-hop question answering)에 특화된 Hotpot-QA 데이터셋을 활용한다[4]. 이 데이터셋은 위키백과를 기반의 113,000개의 질문-답변 쌍으로 구성된다. RAG 파이프라인의 핵심인 임베딩 모델로는 허깅페이스(Hugging Face)에서 제공하는 Qwen/Qwen3-Embedding-0.6B를, 생성 모델로는 OPENAI의 GPT-4o mini를 활용한다[5,6]. 비교 대상 모델로는 기본적인 RAG 구조를 따르는 NaiveRAG를 포함하여, 효율성을 개선한 LightRAG와 정확도를 높인 NodeRAG를 선정한다.

3.2 실험 결과

모델의 성능 평가는 정확성과 효율성 두 가지 지표를 중심으로 구성한다. 정확성 지표로는 모델의 답변과 정답 간의 의미적 유사성을 측정하는 F1-Score를 사용한다. 효율성 측면에서는 답변 생성에 필요한 검색된 토큰 수와 질의 처리 완료에 소요되는 평균 응답 시간(Latency)을 측정하여, 실제 환경에서의 성능을 종합적으로 평가한다. 각 모델에 대한 50개 샘플 질문 테스트 결과는 <표1>과 같다.

<표 1> 성능비교표

Model	F1-score	Tokens	Latency
Naive RAG	0.0250	103k	1.36s
Light RAG	0.0250	103k	0.86s
NodeRAG	0.0248	108k	0.93s
NLRAG	0.0574	132k	0.90s

4. 결론

본 연구는 NLRAG의 성능을 정량적으로 평가한 결과, 기존 베이스라인 대비 정확도를 0.0324 향상시키면서도 평균 응답 속도(0.90s)를 기존 모델과 유사한 수준으로 유지함으로 확인하였다. 이는 속도와 정확도 간의 절충을 요구하던 기존 RAG 접근과 달리, NLRAG가 두 요소를 동시에 충족시킬 수 있음을 입증한다. 더 나아가 적응형 검색 구조를 통해 다양한 질의 환경에서 유연하게 대응할 수 있는 가능성을 보여주었다. 향후 연구에서는 대규모 데이터 환경과 다양한 도메인 적용을 통해 NLRAG의 범용성과 확장성을 더욱 강화할 필요가 있다.

ACKNOWLEOGEMENT

본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음(No. 2025-0-00040).

참고문헌

[1] LI, Huayang, et al, “A survey on retrieval-augmented text generation”, arXiv preprint arXiv:2202.01110, 2022.

[2] XU, Tianyang, et al, “NodeRAG: Structuring graph-based rag with heterogeneous nodes”, arXiv preprint arXiv:2504.11544, 2025.

[3] GUO, Zirui, et al, “Lightrag: Simple and fast retrieval-augmented generation”, arXiv preprint arXiv:2410.05779, 2024.

[4] YANG, Zhilin, et al, “HotpotQA: A dataset for diverse, explainable multi-hop question answering”, arXiv preprint arXiv:1809.09600, 2018.

[5] CHU, Yunfei, et al, “Qwen2-audio technical report”, arXiv preprint arXiv:2407.10759, 2024.

[6] HURST, Aaron, et al, “Gpt-4o system card”, arXiv preprint arXiv:2410.21276, 2024.