키워드 빈도 기반 동적 임베딩을 활용한 자율 학습형 연합 RAG 시스템

황효은, 허의남*
경희대학교, *경희대학교
hhe5361@khu.ac.kr, *johnhuh@khu.ac.kr

Self-Evolving Domain Routing for Ferderated RAG System using Keyword Frequency-based Dynamic Embeddings

HyoEun Hwang, EuiNam John Heo*
KyungHee Univ. , *KyungHee Univ.

요 약

검색 증강 생성(Retrieval-Augmented Generation, RAG) 기법이 확산되면서, 서로 다른 주제와 데이터를 다루는 다수의 RAG를 연합 형태로 운용하는 Federated RAG 구조가 주목받고 있다. 그러나 기존 라우팅 방식은 사람이 도메인을 사전에 정의하거나 대형 언어 모델을 추가로 사용해야 하므로, 도메인 확장 시 관리 비용이 증가하고 질의 처리 지연이 커지는 문제가 있다. 본 논문에서는 문서로부터 추출된 키워드의 문서 등장 빈도를 가중치로 사용하는 자율 학습형 도메인 임베딩 기법을 적용하여 RAG를 구축하였다. 제안 방식은 각 RAG 노드에서 키워드별 문서 수를 누적 관리하고, 해당 키워드 임베딩의 가중 평균으로 도메인 임베딩을 구성함으로써, 사람이 명시적으로 도메인을 정의하지 않아도 데이터 분포에 따라 도메인 표현이 점진적으로 진화하도록 설계되었다. 중앙 도메인 라우팅 서버는 사용자 질의를 임베딩한 벡터와 각 RAG 노드의 도메인 임베딩 간 코사인 유사도를 계산하여, 사전 정의된 임계값 이상인 노드에 대해서만 검색을 수행함으로써 효율적인 라우팅을 달성하고자 하였다. 프로토타입 구현 및 예비 실험을 통해 제안 기법이 지연 시간과 운용 비용을 크게 줄일 수 있음을 확인하였다.

Ⅰ. 서 론

대규모 언어 모델(LLM)의 등장 이후, 이의 환각(Hallucinaton) 문제를 해결하기 위해 검색 증강 생성(Retrieval Augemented Generator, RAG)[1] 기술이 주목받고 있다. 그러나 단일 RAG 시스템에 모든 도메인의 문서를 통합할 경우 검색 정확도가 저하되고 인덱스 관리가 어려워지는 문제가 발생한다. 이를 해결하기 위해 여러 도메인 특화 RAG를 연합하여 운영하는 Federated RAG [2]구조가 연구되고 있다.

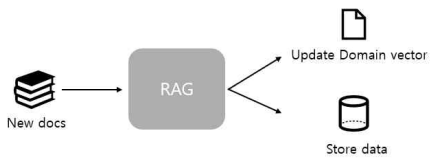
기존의 Federated RAG에서는 사용자 질의를 적절한 도메인 RAG로 보내기 위해 LLM 기반의 라우터를 주로 사용한다. 하지만 이 방식은 라우팅 시마다 LLM을 호출하여 비용과 지연이 발생하고, 새로운 도메인 추가 시 라우터를 재학습하거나 프롬프트를 수정해야 하는 사람의 수동적 개입이 필요하다는 한계가 있다.

본 논문에서는 이러한 문제를 해결하기 위해 자율 학습형 도메인 임베딩 기법(Self-Evolving Domain Embedding Method)를 제안한다. 제안하는 시스템은 문서를 학습할 때 추출된 키워드를 기반으로 도메인의 대표 임베딩 벡터를 동적으로 갱신한다. 이를 통해 별도의 LLM 없이도 실제 RAG 시스템의 데이터 성격에 따라 도메인 특성이 스스로 형성될 수 있도록 하며 경량화된 연산만으로 효율적인 질의 라우팅을 수행할 수 있도록 한다.

Ⅱ. 본론

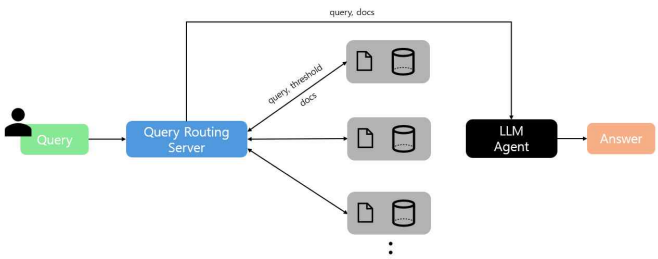
2.1. 시스템 구조

제안하는 시스템은 중앙 도메인 라우팅 서버와 다수의 도메인 특화 RAG 노드로 구성된다. 각 RAG 노드는 문서를 수집할 때마다 자신의 도메인 성격을 나타내는 Domain embedding vector를 갱신한다.



[그림 1. 도메인 자율 학습 흐름]

[그림 2] 와 같이 도메인 라우팅 서버는 사용자 쿼리 벡터와 유사도 임계점을 각 RAG로 전송한다. RAG는 자신의 domain embedding vector와 사용자 query 간의 코사인 유사도를 계산한다. 도메인 유사도 스코어가 임계점을 넘을 경우, RAG는 벡터 데이터 베이스에서 관련 문서들을 반환한다. 라우팅 서버는 문서들과 query를 종합하여 LLM으로 전송한다.



[그림 2. 시스템 구조 아키텍처]

2.2. 키워드 기반 동적 도메인 임베딩

$$V_{domain} = \frac{\sum_{i=1}^{N} (w_i E_{k_i})}{\sum_{i=1}^{N} (w_i)} \quad \text{...식(1)}$$

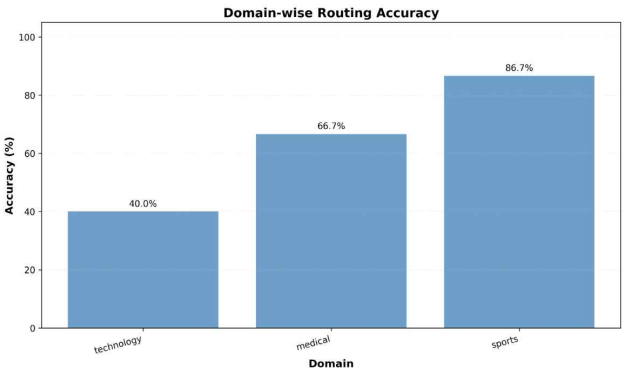
여기서 E_{k_i}는 키워드 i의 임베딩 벡터이며 w_i는 해당 키워드가 포함된 누적 문서 수이다. 이 방식은 특정 키워드가 여러 문서에서 반복적으로 등장할수록 해당 키워드의 의미가 도메인 벡터의 방향성을 결정하도록 유도한다.

2.3. 질의 라우팅 및 유사도 산출

사용자 쿼리 Q가 입력되면, 중재 서버는 쿼리 임베딩 V_{query}와 각 RAG 노드의 도메인 벡터 V_{domain} 간의 코사인 유사도(Cosine Similarity)를 계산한다.

2.4 실험 결과 요약

실험 연구에서는 자율 학습형 도메인 임베딩을 기반으로 한 Federated RAG 라우팅 기법의 유효성을 검증하기 위해, 3개의 상이한 도메인(technology, medical, sports)을 대상으로 실험을 수행하였다. 각 도메인마다 10개의 임의의 학습 문서를 사용하여 도메인 임베딩을 구성하였으며, 도메인별 15개의 테스트 쿼리(총 45개)를 통해 라우팅 성능을 평가하였다. 유사도 임계값(Threshold)은 0.4로 설정하였다.



도메인별 성능을 분석한 결과, medical 도메인은 66.7%, sports 도메인은 86.7%, technology 도메인은 40%의 정확도를 기록하였다. technology 도메인의 상대적으로 낮은 정확도는 해당 도메인이 포함하는 주제 범위가 넓은 반면, 학습에 사용된 문서 수가 제한적이었기 때문에 도메인 임베딩이 충분히 대표성을 확보하지 못한 데에서 기인한 것으로 판단된다. 이는 단일 도메인 임베딩이 다양한 하위 주제를 포괄할 경우 표현력에 한계가 발생할 수 있음을 시사한다.

한편, 제안한 라우팅 과정은 단순한 벡터 유사도 계산만으로 수행되었으며, 그 결과 평균 응답 지연 시간은 8.17ms로 측정되었다. 이는 추가적인 대형 언어 모델 호출 없이도 빠른 속도의 라우팅이 가능함을 보여주며, 특히 연산 자원이 제한된 엣지 환경이나 대규모 분산 Federated RAG 구조에서도 적용 가능성이 높음을 의미한다.

종합하면, 본 실험 결과는 자율 학습형 도메인 임베딩 기반 라우팅 기법이 정확도와 경량성 측면에서 일정 수준 이상의 성능을 충족함을 입증한다. 제안 방식은 사람이 사전에 도메인을 명시적으로 정의하거나 추가적인 추론 비용을 요구하는 기존 접근법에 비해, 운영 복잡도를 낮추고 실질적인 시스템 효율을 향상시킬 수 있는 대안임을 보여준다.

Ⅲ. 결론

본 논문에서는 자율 학습형 도메인 임베딩을 활용한 Federated RAG 라우팅 기법을 제안하였다. 이 방식은 사람이 사전에 도메인을 정의하지 않더라도, 문서 데이터의 통계적 특성에 따라 각 RAG 노드의 도메인 표현이 점진적으로 형성되고 이에 기반한 라우팅 경로가 동적으로 최적화된다는 장점을 가진다. 또한 제안한 라우팅 과정은 별도의 대형 언어 모델 호출 없이 단순한 벡터 연산만으로 수행되므로, 추론 지연을 최소화하면서도 운영 비용을 획기적으로 절감할 수 있음을 확인하였다.

키워드 빈도 가중치를 활용한 자율 학습형 도메인 임베딩 생성 기법은 반복적으로 등장하는 핵심 키워드가 도메인 벡터의 방향성을 결정하도록 유도함으로써, 각 RAG 노드의 의미적 중심을 효과적으로 요약할 수 있었다. 다만 본 연구에서는 가장 단순한 형태의 self-evolving domain 방식을 채택하여, 각 RAG의 도메인을 단일 벡터로 표현하는 접근에 초점을 맞추었다. 이로 인해 도메인 내부에 존재하는 세부 주제 간의 다양성이나 의미적 분산을 충분히 반영하지 못하는 한계가 존재한다.

향후 연구에서는 이러한 한계를 극복하기 위해, 하나의 RAG 노드 내에서도 다수의 하위 도메인을 동시에 표현할 수 있는 multi-domain 또는 multi-prototype 기반 도메인 임베딩 확장 방안을 고려할 수 있다. 예를 들어, 클러스터링 기반의 복수 도메인 벡터 유지, 도메인 벡터의 분산 정보 활용, 또는 시간에 따른 도메인 변화 추적을 통해 보다 정교한 라우팅 정책을 설계할 수 있을 것으로 기대된다. 이러한 확장은 Federated RAG 환경에서의 확장성과 적응성을 더욱 강화하는 방향으로 기여할 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 2025년도 SW중심대학사업의 지원(2023-0-00042, 50%)과 2025년도 정부(교육부)의 재원으로 한국연구재단 대학기초연구소지원사업 (G-LAMP)의 지원을 받아 수행된 연구임(No. RS-2025-25442355, 50%)

참 고 문 헌

[1] Guerraoui, R. et al., “Efficient Federated Search for Retrieval-Augmented Generation (RAGRoute),” ACM (2025).

[2] Shojaee, P. et al., “Federated Retrieval Augmented Generation for Multi-Product Question Answering,” COLING Industry Track, 2025.

[3] Mao, X. et al., “Federated Retrieval-Augmented Generation: A Systematic Mapping Study (FedE4RAG),” 2025.

[4] “Dynamic Routing in RAG: Directing User Queries to the Right Vector Store with Open Source Models,” WebAI Blog, 2024.

[5] Lewis, P. et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS 2020.