질의 의도 인식 기반 RAG를 통한 수사 지원 데이터 검색 성능 향상

오지연[1], 임정민[2], 박현호[3,4]

[1]명지대학교, [2]한림대학교, [3]한국전자통신연구원, [4]과학기술연합대학원대학교

yeonaf6e@mju.ac.kr, limjm1617@gmail.com, hyunhopark@etri.re.kr

Improving Search Performance of Investigation Support Data via Query Intent-Aware RAG

Jiyeon Oh[1], Jeongmin Lim[2], Hyunho Park[3,4]

[1]Myongji University, [2]Hallym University, [3]Electronics and Telecommunications Research Institute, [4]University of Science and Technology

요 약

본 논문은 경찰 수사를 지원하기 위한 수사 지원 데이터(예: 수사 매뉴얼 및 법령 지식)의 검색 성능 향상을 목표로, 질의 의도 인식 기반 검색 증강 생성(RAG, Retrieval-Augmented Generation) 기법을 제안한다. 제안 기법은 수사 지원 데이터에 포함된 문서 정보를 “주제 영역”과 “내용 영역”으로 구조적으로 분리하여 검색에 활용한다. 수사 지원 데이터 검색 질의가 입력되면, 질의 의도 인식 기반 RAG는 거대언어모델을 활용하여 수사관의 질의 문맥을 분석하고, 질의 의도에 따라 주제 영역과 내용 영역의 상대적 중요도를 동적으로 조정함으로써 질의 목적에 부합하는 정보를 효과적으로 검색한다. 실제 수사 매뉴얼 및 법령 데이터를 활용한 실증 실험 결과, 제안 기법은 고정 가중치를 사용하는 기존 Baseline 방식 대비 평균 검색 스코어 기준으로 8.48%의 성능 향상을 달성하였다. 본 연구는 질의 의도에 따른 영역별 정보 중요도 조절을 통해 보다 정확한 수사 지원 데이터 검색을 제공함으로써, 경찰 수사의 정확성과 실효성 향상에 기여한다.

Ⅰ. 서 론

최근 거대언어모델(LLM, Large Language Model)의 성능 향상과 함께, 외부 지식을 효과적으로 활용하기 위한 검색 증강 생성(RAG, Retrieval-Augmented Generation) 기법이 다양한 전문 도메인에서 주목받고 있다[1]. 특히 복잡한 수사 매뉴얼과 수시로 개정되는 법령을 정확하게 참조해야 하는 경찰 수사 도메인에서는, 수사 매뉴얼과 법령 지식으로 구성된 수사 지원 데이터의 검색을 지원하는 RAG 기술의 중요성이 점차 커지고 있다[2]. 그러나 범용 검색 알고리즘 기반 RAG는 방대한 수사 지원 데이터를 벡터 임베딩하는 과정에서 이질적인 정보 유형이 단일 표현 공간에 혼합됨으로써, 핵심 식별자 중심 정보의 표현력이 약화되고 검색 성능이 저하되는 정보 희석(Information Dilution) 문제가 발생할 수 있다[3].

본 논문은 이러한 정보 희석 문제를 극복하기 위한 질의 의도 인식 기반 RAG를 제안한다. 수사 지원 데이터는 크게 법령의 조항이나 죄명과 같은 식별자 정보를 담은 “주제 영역”과, 구체적인 수사 기법과 절차를 설명하는 “내용 영역”으로 구분할 수 있다. 질의 의도 인식 기반 RAG는 LLM을 활용하여 수사관의 질의 의도가 “주제 영역”에 해당하는지 또는 “내용 영역”에 해당하는지를 추론하는 질의 속성에 맞춰 두 영역의 검색 가중치를 동적으로 제어한다. 소규모 핵심 수사 데이터를 활용한 사전 실험 결과, 제안 기법은 고정 가중치 방식(Baseline) 대비 평균 검색 스코어 기준으로 약 8.48%의 성능 향상을 달성하였다. 이러한 결과는 질의 의도에 적합한 수사 지원 데이터를 제공함으로써, 수사 지원 RAG의 검색 정확도와 실효성을 향상시킬 수 있다.

Ⅱ. 질의 의도 인식 기반 RAG의 동적 순위 재지정 기술

본 연구에서는 수사 데이터가 지니는 “주제 영역”과 “내용 영역”이 병합된 계층적 특성을 보존하기 위해, ‘이원화 인덱싱(Dual-Field Indexing)’ 구조를 제안한다. 제안 기법은 문서를 “주제 영역”과 “내용 영역”으로 물리적으로 분리하여 관리한다. 사용자의 질의가 입력되면, 지능형 의도 분류 모델이 질의의 문맥을 분석하여 “주제 영역”과 “내용 영역” 중 어느 필드에 상대적으로 집중할지를 결정한다. 이 모델은 질의의 문맥을 분석하여 “주제 영역”과 “내용 영역” 중 어느 필드에 상대적으로 집중할지를 결정한다.

최종 검색 스코어(S_{final})는 다음과 같은 “주제 영역”과 “내용 영역”에서 산출된 유사도 점수의 동적 가중치 선형 결합을 통해 산출된다.

$$S_{final} = (w_{topic} \cdot S_{topic}) + (w_{content} \cdot S_{content}) \quad (1)$$

여기서 S_{topic}은 주제 영역 벡터 공간에서 계산된 질의 - 문서 간 코사인 유사도 점수를 의미하며, $S_{content}$는 내용 영역에서 계산된 유사도 점수를 나타낸다. w_{topic}과 $w_{content}$는 각각 주제 영역과 내용 영역의 상대적 중요도를 반영하는 계수로, 질의 의도 분류 결과에 따라 질문 단위로 동적으로 결정된다. 두 계수는 다음의 수식 (2)를 만족한다.

$$w_{topic} + w_{content} = 1 \quad (2)$$

제안 기술의 핵심은 단순 검색을 넘어선 검색 이후의 ‘동적 순위 재지정(Dynamic Re-ranking)’이다. 파이프라인은 다음과 같이 구동된다. 사용자의 질의가 입력되면 시스템은 Qdrant DB로부터 유사도가 높은 상위 K개(Top-K)의 후보군을 우선 검색한다. 이후 LLM이 질의 문맥을 분석하여 의도를 인식하고, 최적 가중치 계수(w_{topic},$w_{content}$)를 스스로 결정한다. 설정된 가중치는 1차 검색된 Top-K 후보군들의 필드별 점수에 적용하여 최종 순위를 재정렬한다. 이 과정은 검색의 변별력을 높이고 전문 도

메인에서의 답변 정확도를 극대화한다. 본 연구에서는 수사 데이터의 구조적 특성을 보존하기 위해 영역별 분리 청킹(Field-specific Chunking) 전략을 수행한다. 수사 매뉴얼 및 법령 데이터를 단일 텍스트 뭉치로 처리하지 않고, “주제 영역”과 “내용 영역”으로 문서를 물리적으로 분리하여 청킹한다. 임베딩 모델은 Qwen/Qwen3-Embedding-8B를 채택하였으며, SentenceTransformer를 통해 실시간 임베딩을 수행한다. 추출된 벡터는 Qdrant DB에 이원화 인덱싱하여 저장된다. 즉, Payload 기능을 활용하여 “주제 영역”과 “내용 영역”을 서로 다른 속성값으로 연결하여 관리함으로써, 단일 벡터 공간 내에서 각 필드에 대한 독립적인 유사도 점수를 산출할 수 있는 환경을 구축한다.

본 연구의 자동화 타당성을 검토하기 위해 실제 수사 지원 데이터 3종을 대상으로 수동 가중치 탐색 실험을 수행하였다. 실험 질의는 특정 법령 명칭을 확인하는 “주제 영역” 중심(Topic-focused), 현장 대응 절차를 묻는 “내용 영역” 중심(Content-focused), 그리고 두 영역이 혼합된 “복합 영역” 중심(Topic+Content Combined)의 세 가지 질의 유형으로 구성하였다. 실험은 “주제 영역” 가중치 계수(w_{topic})를 0.0에서 1.0까지 0.1단위로 수동 조정하며 검색 성능의 변화를 측정하는 방식으로 진행되었다. 이는 질의 성격에 따라 LLM이 자동 설정해야 할 최적 가중치 계수가 가변적이며, 동적 가중치 제어의 당위성을 확보하기 위한 과정이다. 검색 성능은 검색된 상위 K개의 문서 중 질의와 실제로 연관된 문서의 비율인 ‘Top-K 결과의 정밀도’와 ‘질의-문서 간 코사인 유사도’ 점수를 종합한 평균 검색 스코어를 기준으로 평가하였다.

가중치 탐색 실험 결과, 그림 1과 같이 질의 유형별로 성능이 극대화되는 최적의 가중치 지점이 극명하게 대조됨을 확인하였다. “주제 영역” 중심(Topic-focused) 질의의 경우, w_{topic}이 증가함에 따라 검색 스코어가 향상되어 w_{topic} = 1.0에서 최고점(0.8004)를 기록하였다. 이는 고도의 전문 데이터를 검색할 때 핵심 식별자 정보의 비중을 높이는 것이 필수적임을 방증한다. 반면 “내용 영역” 중심(Content-focused) 및 “복합 영역” 중심((Topic+Content Combined) 질의의 경우, w_{topic} 비중을 최소화했을 때 최적의 성능이 나타났다. 두 유형 모두 w_{topic} = 0.0 지점에서 최고 스코어(0.6210, 0.8335) 를 기록하였다. 특히 “복합 영역” 중심(Topic+Content Combined) 질의의 경우 가중치 설정에 따른 성능 편차가 크게 관찰되었다. 이는 고정된 검색 전략을 사용 시 검색 품질이 급격히 저하될 위험이 있음을 시사하며, 제안하는 질의 의도 인식 기반의 동적 제어 기술의 필요성을 입증한다. 실험 데이터 전체에 대해 단일 고정 가중치(0.5 : 0.5)를 적용한 기존 Baseline 방식과 제안 기법의 성능을 비교 평가하였다.

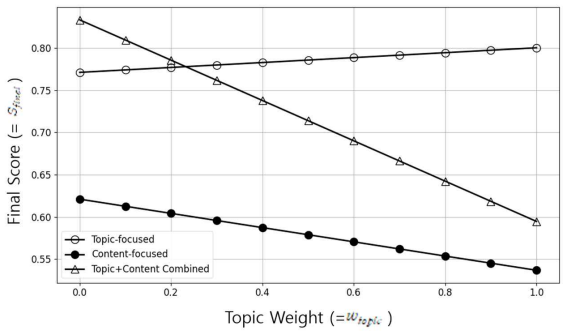

그림1. 질의 유형별 최적의 가중치 지점 분석

실험 결과, 표 1과 같이 최적 가중치를 선정한 방식이 Baseline 대비 평균 검색 스코어에서 8.48%의 성능 향상을 달성하였다. 이는 수사관의 질의 문맥을 사전에 파악하여, 정보의 밀도가 낮거나 검색 손실이 예상되는 영역을 배제하고 유효 정보가 집중된 영역의 비중을 강화했기 때문으로 분석된다.

Metric	Baseline(0.5)	Agentic (Optimal)	Improvement
Avg. Search Score	0.6929	0.7516	+8.48%

표1. 질의 의도 분류 유무에 따른 평균 검색 스코어(Avg. Search Score) 비교 분석

Ⅲ. 결론

본 연구에서는 전문적인 수사 지원 데이터 환경에서 발생하는 정보 희석(Information Dilution) 문제를 해결하기 위해, 질의 의도 인식 기반의 영역별 가중치 동적 제어 기술을 제안하고 그 실효성을 검증하였다. 실험 결과, 수사관의 질의 성격에 따라 최적의 검색 성능을 발휘하는 가중치 임계점이 극명하게 대조되는 ‘가중치 민감도 현상’을 확인하였다. 특히 “주제 영역” 중심 질의와 “내용 영역” 중심 질의에서 최적화되는 지점이 정반대로 나타난 결과는, 고정된 검색 전략을 사용하는 기존 RAG 시스템의 한계를 명확히 시사한다.

본 연구에서 제안한 이원화 인덱싱(Dual-Field Indexing) 및 동적 순위 재지정 기술(Dynamic Re-ranking) 파이프라인은 이러한 한계를 극복하고, 기존 Baseline 대비 평균 검색 스코어에서 8.48%의 유의미한 향상을 달성하여, 실제 수사 지원 환경에서 검색 정밀도를 개선하는 데 유효함을 입증하였다. 비록 본 연구는 소규모 수사 데이터를 활용한 수동 가중치 탐색 실험을 통해 수행되었으나, 도출된 성능 지표는 향후 LLM을 활용한 가중치 자동화 제어 기술의 기술적 당위성을 뒷받침한다. 데이터의 규모가 방대해지고 복합 질의의 비중이 높아질수록, 질의 문맥에 따른 적응형 가중치 최적화는 RAG 시스템의 성패를 결정짓는 핵심 요소가 될 것이다. 본 연구의 성과를 바탕으로 향후 연구에서는 다음과 같은 고도화를 추진할 예정이다. 먼저, 본 실험에서 확인된 가중치 최적화 로직을 LLM 엔진에 내재화하여 실시간 가중치 산출 및 제어 프로세스를 자동화할 계획이다. 또한, 수사 데이터 특유의 ‘편·장·절’ 체계를 반영한 다중 필드 인덱싱(Multi-Field Indexing)으로 구조를 확장하여 검색의 세밀도를 더욱 높일 예정이다.

ACKNOWLEDGMENT

이 논문은 26년도 정부(경찰청)*의 재원으로 과학치안진흥센터 사이버범죄 수사단서 통합분석 및 추론시스템 개발 사업의 지원을 받아 수행된 연구임(No. RS-2025-02218280)

참 고 문 헌

[1] Lewis, Patrick, et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in neural information processing systems* 33 (2020): 9459-9474.

[2] Cui, J., et al., “Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model,” arXiv preprint, arXiv:2306.16092v2, May 2024

[3] Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *Transactions of the Association for Computational Linguistics* 12 (2024): 157-173.