# 효과적인 RAG Document Data 구조화 전략

손영진 [1], 임유경 [2], 박민정 [3] 채상미 [4]
[1] 이화여자대학교 경영학과 박사수료
[2] 이화여자대학교 데이터사이언스학과 석사과정
[3] 금오공과대학교 경영학과 교수
[4] 이화여자대학교 경영학과 교수

teumdal@ewhain.net, limyg512@gmail.com, mjpark@kumoh.ac.kr, smchai@ewha.ac.kr

# Effective RAG Document Data Structuring Strategy

Young Jin Son[1], Yugyung Lim[2], Minjung Park[3], Sangmi Chai[4]
[1]Dept. of Business Administration, Ewha Womans University
[2]Dept. of Data Science, Ewha Womans University
[3]Dept. of Business Administration, Kumoh National Institute of Technology
[4]Dept. of Data Science, Ewha Womans University

**요 약**

대규모 언어 모델의 발전은 텍스트 생성 및 정보 제공 분야에서 큰 진전을 이루었으며 사용자와의 원활한 소통을 가능하게 했다. 그러나 언어 모델은 특화된 정보 제공에 한계를 가지며 때때로 부정확한 정보를 생성할 수 있다. RAG(Retrieval-Augmented Generation) 기법은 이러한 한계를 극복하기 위해 제안되었다. 본 연구에서는 RAG 의 답변품질과 효율성을 높이기 위해 외부 문서 정보와 단어 단위로 카테고리화된 인덱싱 데이터 세트를 함께 제공하여 보다 정확하고 신뢰성 있는 문서 생성을 가능하게 하는 접근법을 제시한다.

## 1. 서론

GPT 와 같은 대규모 언어 모델(LLM, Large Language Model)들의 발전으로 텍스트 생성 및 정보 제공 분야의 엄청난 발전이 있었다. 특히 사용자와 기계 간의 원활한 소통을 가능하게 하여 일반 사용자들이 인공지능을 보다 손쉽게 사용할 수 있게 하였다.

LLM 으로 생성한 텍스트는 LLM 훈련 데이터의 한계로 인해 특정 도메인, 회사, 조직 등과 같은 특화된 정보에 있어서는 부정확한 정보를 제공하거나 답변 생성에만 중점을 둔 Hallucination 발생의 위험이 있기 때문에 기업에서 충분히 활용하기 어렵다. 특히 언어 모델이 학습하지 않은 기업의 특성과 내부 정보가 필요한 작업에서 더욱 그러하다.

검색증강생성(RAG, Retrieval-Augmented Generation) 기법[1]은 언어모델의 학습 데이터 한계로 발생하는 정보 생성 한계를 해결하고자 제시된 전략 중 하나로 입력 정보에 대한 사전지식을 제공, 즉 법률, 의학과 같은 도메인 특화 정보, 회사나 조직 내부 정보를 반영한 문서 정보를 LLM 프로세스 외부에서 제공하여 LLM 을 재훈련시키지 않고도 LLM 이 특화된 텍스트와 정보를 효율적으로 생성할 수 있도록 한다[2].

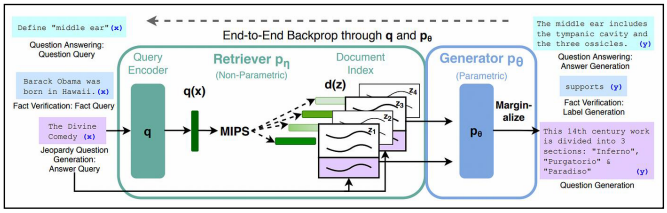
RAG 인덱싱 과정의 쿼리분석 및 문서 분석 단계에서 사용자의 쿼리와 Database document 는 특히 특정 분야에 맞는 정보를 제공하는데 매우 중요하다. 본 연구에서는 RAG 의 답변 품질 효율성을 높이기 위해 데이터 인덱싱을 구조화하는 방법을 제안하여 답변 최적화를 돕고자 한다.

## 2. RAG (Retrieval-Augmented Generation)

RAG 기법은 언어모델에서 응답을 생성하기 전에 기존에 학습된 데이터의 외부에서 신뢰할 수 있는 데이터베이스를 참조(retrieve)하도록 하는 기법이다[2]. 사용자 쿼리에 대한 답변의 맥락으로 사전 지식을 제공함으로써 언어 모델의 출력을 향상시킨다. RAG 는 특히 생성된 텍스트 결과물의 유효성을 검사하기 위해 증거가 필요한 법적 질의응답과 같은 지식 집약적이고 전문가 의존적인 작업에 유용하다. [3]

RAG 의 주요 구성요소로는 인덱싱(Indexing), 검색(Retrieve), 생성(Generation)이 있다. 인덱싱 과정은 원본 데이터를 정제, 추출하여 검색에 필요한 데이터벡

터로 인코딩하는 과정(Embedding)이다. 검색 과정은 사용자의 쿼리 입력을 벡터로 변환하여 해당 값과 인덱싱에서 얻은 벡터 값 간의 유사성을 계산하여 상위 K 값(Top K)을 선택하는 과정이다. 생성과정은 인덱싱, 검색 과정을 통해 쿼리와 정보를 결합하여 새로운 프롬프트를 생성하고 이를 기반으로 언어모델이 사용자에게 답변을 제공하는 과정이다[2].




(그림 1) RAG 프레임워크[2]

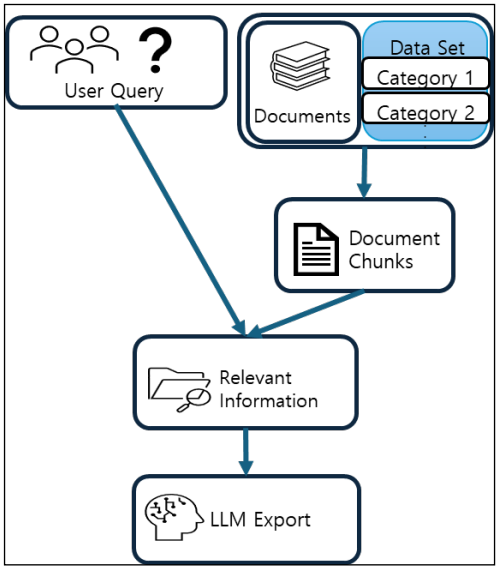
RAG 는 또한 검색 전 절차(Pre-Retrieval Process), 검색 후 절차(Post-Retrieval Process), RAG 파이프라인 최적화(RAG Pipeline Optimization)의 3 단계로 나눌 수 있다[4]. 검색 전 절차는 RAG 시스템의 검색 효율성과 결과의 질을 향상시키는데 필수적이며 데이터 인덱싱 최적화와 임베딩 과정을 통해 RAG 검색 결과의 관련성 및 정확성 향상에 중점을 둔다. 검색 후 절차는 데이터베이스에서 검색된 중요 정보를 쿼리와 결합하여 언어모델에 제공하는 과정이며, 파이프라인 최적화는 RAG 시스템 자체의 효율성과 정보 품질을 향상시키는데 중점을 둔다.

RAG 시스템의 검색 효율성과 결과의 질을 향상시키고 특정 분야의 정보를 제공하는데 있어 RAG 인덱싱 과정의 쿼리 및 문서 분석 단계에서 사용자의 쿼리와 Base 문서는 매우 중요하다. 이러한 정보를 제공하는 검색 전 절차를 강화하는 방법으로 데이터 인덱싱 최적화의 방법으로 데이터 세분화 강화, 인덱싱 구조 최적화, 메타데이터 정보 추가, 정렬 최적화, 혼합검색 등이 연구되고 있다[4].

### 3. 인덱싱 데이터 구조화 전략

인덱싱 과정에서 사용자 쿼리를 향상시키는 방법은 항상 정확한 결과 문서를 얻는다는 보장이 없으며 잘못된 Top K 문서의 도출은 언어모델이 답변을 생성할 때 정확성을 떨어뜨리는 원인이 된다. 또한, 사용자의 질문이 기존에 임베딩된 문서의 말투, 유의어, 대명사 등과 다를수록 응답의 정확성이 떨어질 수 있다[5]. 이를 해결하기 위해 RAG Database 제공 시 관련 문서뿐만 아니라 Document 내에서 동일한 의미이나 서로 다른 용어로 되어있는 단어 정보, 중복되거나 불필요한 단어 정보들을 일정 카테고리 항목으로 구분하고 명확한 단어로 수정한 인덱싱 데이터를 제공하여 RAG 시스템의 정확도를 높이고자 하였다(그림 2,3).

단어 단위로 카테고리화 된 인덱싱 데이터세트를 Database 로 함께 제공하여 검색 정확도를 높임으로써 사용자의 쿼리와 연관된 문서 검색 확률(Top K)를 향상시켜 사용자가 기대한 답변을 보다 잘 도출할 수 있다.




(그림 1) RAG 의 인덱싱 데이터 제공

```
primary_term_dict = {
    "계약기간": ["계약기간", "계약 기간", "용역수행기간"],
    "계약금액": ["계약금액", "계약금액 및 지급"],
    "계약의 변경": ["계약의 변경", "계약서의 변경", "협약의 변경"],
    "비밀 유지": ["비밀 유지", "비밀보장", "비밀유지"],
    "계약해지": ["계약의 무효", "계약의 해지", "계약의 해제 · 해지",
    "대금지급": ["대금지급", "연구비", "연구비 지급", "연구용역비",
    "업무수행범위":["업무수행범위", "용역업무 범위", "용역의 범위"],
    "분쟁해결": ["분쟁의 해결", "분쟁해결"],
    "손해배상": ["손해배상", "손해배상 책임"],
    "기타": ["기타", "기타사항"],
    "신의성실 및 상호협조": ["상호협조", '신의 성실 및 상호 협조',
    "양도의 제한": ["권리의무 양도금지", "양도금지 등", "양도의 제한
```

(그림 2) 단어 단위 인덱싱 샘플

### 4. 결론

RAG 의 Database 에 동일한 의미이나 서로 다른 용어로 되어있는 단어 정보, 중복되거나 불필요한 단어 정보를 가진 단어를 처리할 수 있는 단어 단위 인덱싱 데이터 세트의 추가 제공만으로 사용자 쿼리의 처리에 있어 보다 정확한 검색 문서를 도출하여 답변의 정확도를 개선하는데 도움이 될 수 있다. 추후 단어 단위 뿐만 아니라 문장 단위, 문단 수준까지 인덱싱 데이터 세트를 확장시켜 제공하고, Top K 값의 변화를 검증하여 RAG 를 통한 LLM 답변의 정확성을 높이고자 한다.

## 참고문헌

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. 2020..

[2] Musumeci, E., Brienza, M., Suriani, V., Nardi, D., & Bloisi, D. D.. LLM Based Multi-Agent Generation of Semi-structured Documents from Semantic Templates in the Public Administration Domain. arXiv preprint arXiv:2402.14871. 2024

[3] Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., ... & Fleisch, B. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. arXiv preprint arXiv:2404.04302. 2024.

[4] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. 2023.

[5] 강석훈, 김성진. M-RAG: 메타데이터를 이용한 RAG 방법의 성능향상. 한국정보통신학회논문지, 27(12), 1489-1500, 2023. 10.6109/jkiice.2023.27.12.1489