생성형 인공지능과 검색 증강 생성을 활용한 학과 정보 제공 시스템 개발

천민우, 정세준, 신혜민, 김시현, 최정열*

성결대학교

{minwoce, jungsj0524, shm7696, huen2003, passjay}@sungkyul.ac.kr

Development of an Academic Information Provision System Using Generative AI and Retrieval-Augmented Generation

Chun Min Woo, Jung Se Jun, Shin Hye Min, Kim Si Hyun, Choi Jung Yul*

Sungkyul Univ.

요 약

본 논문은 전공자율선택제에서 신입생의 전공 탐색 지원 및 교육과정 소개를 위해 생성형 인공지능과 검색 증강 생성에 기반한 학과 정보 제공 시스템 개발 내역을 소개한다. 교육과정 문서 내용을 바탕으로 청킹과 임베딩을 수행하고 데이터베이스에 저장하여 의미 기반 검색이 가능한 파이프라인을 구축하였다. 사용자 질의를 바탕으로 관련 정보를 검색한 결과를 생성형 인공지능에 전달하여 개인화된 응답을 생성하도록 구성하였다. 제안하는 시스템은 다양한 전공 체계에도 확장할 수 있는 학습 지원 도구로서의 가능성을 제시한다

Ⅰ. 서 론

대학 내 전공자율선택제의 확대로 신입생은 관심과 적성에 맞춰 전공을 선택할 수 있으나 교과목 간 선후수 관계나 진로 연계성에 대한 명확한 안내 부족으로 인해 수강 계획 수립이 어려운 경우가 많다. 이러한 상황은 학습 동기 저하와 전공 이수의 비효율성을 초래할 수 있어 신입생이 보다 직관적이고 개인화된 정보를 바탕으로 전공을 탐색할 수 있는 지원 도구가 필요하다.

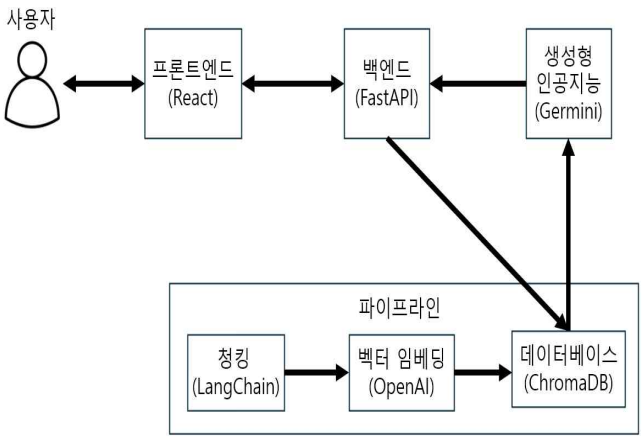
본 논문에서는 학생들의 전공 및 수강 과목 선택 과정에서 발생하는 정보 탐색의 어려움을 해소하고 개인화된 학습 경로 추천을 지원하기 위해 생성형 인공지능 기반의 학과 정보 제공 시스템을 제안한다. 기존의 정적 안내 웹페이지나 수동적 상담 방식과 달리 사용자의 다양한 질문 의도와 맥락을 실시간으로 파악하고 맞춤 과목 구성, 이수 순서, 진로 연계 추천 등에 대해 개인화된 답변을 제공할 수 있다는 점에서 차별성을 가진다.

Ⅱ. 본론

2.1 시스템 구조

본 논문에서 제안하는 맞춤형 학과 정보 안내 시스템은 생성형 인공지능과 검색 증강 생성 기술을 기반으로 한다 [1]. 제안하는 시스템은 랭체인(LangChain) 프레임워크와 ChromaDB를 활용한 지식 검색 및 응답 생성 파이프라인을 이용하여 구축하였다. 파이프라인은 교육과정 문서를 의미 단위로 청킹(Chunking)한 뒤, OpenAI API를 활용하여 임베딩(Embedding)하고, 그 결과 생성된 벡터 데이터를 데이터베이스에 저장하는 과정으로 진행된다.

사용자는 리액트(React) 기반의 프론트엔드 인터페이스를 활용하여 자연어 질의를 진행한다. 질의는 FastAPI 기반의 백엔드 서버로 전달되며, 사전에 구축한 파이프라인을 활용하여 의미적 유사도 검색을 실행한 뒤, 상위 검색 결과를 기반으로 적절한 컨텍스트(Context)를 구성하여 전달한다. 이후 생성된 컨텍스트는 프론트엔드로 반환되어 사용자의 질의에 대한 답변을 제공한다. 다음 절에서는 시스템 주요 구성 요소별 기능 및 과정을 상세히 설명한다. <그림 1>은 시스템의 주요 흐름을 요약한다.



<그림 1> 시스템 아키텍처

2.2 데이터베이스 설계 및 구축

데이터베이스는 컴퓨터공학과 교육과정 문서를 기반으로 구축하였다. 교육과정 문서는 PDF 형식으로 제공되었으며, 랭체인 라이브러리를 활용하여 문단 단위로 텍스트를 추출하고, 의미 단위별로 청킹 과정을 수행하였다. 청킹 데이터는 OpenAI의 text-embedding-3-small 모델을 이용해 벡터 임베딩 처리를 거친 후, ChromaDB에 컬렉션 형태로 저장하였다 [2]. 이때 각 청킹 블록에는 과목명, 페이지 번호, 교과목명, 전공역량과 같은 메타데이터가 함께 저장되어 검색 시 필터링 및 세부 검색이 가능하도록 설계하였다. 이러한 데이터 구축 과정은 사용자의 질의에 대해 보다 정확하고 개인화된 정보를 검색하고 제공하기 위한 기반이 된다.

2.3 검색 증강 생성(RAG) 파이프라인

본 시스템은 검색 증강 생성을 기반으로 구현하였다. 외부 지식베이스에서 관련 정보를 검색한 후, 이를 생성형 인공지능과 결합하여 정확하고 풍부한 응답을 생성할 수 있도록. 이를 통해 기존의 단순 생성형 언어 모델이 가지는 정적 지식 한계와 최신성 부족 문제를 보완할 수 있다. 사용자

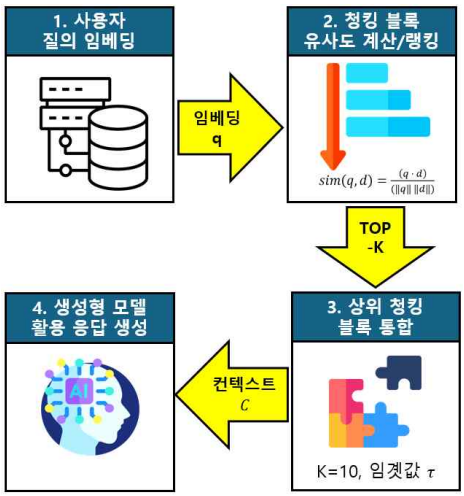
가 자연어 질의를 입력하면 백엔드 서버는 ChromaDB 내부의 컬렉션을 활용한 의미적 유사도 검색을 수행한다. 이때 의미 유사도는 벡터 임베딩 간의 코사인 유사도를 기준으로 계산되며, 그 수식은 다음과 같다.

$$sim(q,d) = \frac{(q \cdot d)}{(\|q\| \, \|d\|)} \quad (1)$$

q는 사용자 질의의 임베딩 벡터, d는 청킹 블록의 임베딩 벡터, $\| \cdot \|$는 벡터의 크기(norm), $\cdot$은 벡터의 점곱(dot product)을 의미한다. 상위 유사도를 가지는 k개의 청킹 블록은 컨텍스트를 구성하는 청킹 블록 집합 C로 정의되며 그 조건은 다음과 같다.

$$C = \{ d^1, d^2, \ldots, d_k \} \; such \; that \; sim(q, d_i) \geq \tau \quad (2)$$

여기서 τ는 유사도 임곗값이며 C는 최종적으로 생성형 인공지능에 전달되어 응답 생성을 위한 컨텍스트로 활용된다. 현재 시스템은 컬렉션 내에서 상위 유사도를 보이는 10개의 청킹 블록을 검사 후 컨텍스트로 통합한다. 이러한 파이프라인 구조를 통해 사용자는 자신의 질문에 가장 적합한 정보를 기반으로 한 응답을 실시간으로 제공받을 수 있으며, 기존의 키워드 기반 검색이나 단순 생성 방식에 비해 높은 정확도와 개인화된 정보 전달이 가능하다. <그림 2>는 이러한 과정을 네 단계로 시각화한 것이다. 먼저 사용자 질의를 바탕으로 임베딩하여 특정 조건(학년, 관심분야 등)에 따라 해당하는 컬렉션을 호출·선택한다. 이후 선택된 컬렉션 내부의 모든 청킹 블록을 대상으로 질의 벡터와의 코사인 유사도를 계산하여 관련성이 높은 순으로 정렬한다. 다음으로 상위 10개의 청킹 블록을 통합하여 컨텍스트를 구성한다. 마지막으로 컨텍스트를 생성형 인공지능 모델에 전달하여 최종 응답을 생성한다.



<그림 2> 파이프라인 동작 과정

2.4 페르소나 설정 및 웹 인터페이스 구성

RAG 파이프라인을 통해 만들어진 컨텍스트를 생성형 인공지능에 전달하여 사용자에게 적합한 답변을 만들어낼 수 있다. 생성형 인공지능 모델로 gemini-1.5-pro를 사용하였으며 '컴퓨터공학과 멘토'라는 페르소나(Persona)를 설정하였다 [3]. 설정된 페르소나는 친근하고 신뢰감을 주는 조언자의 역할을 수행하며 과목 선택과 학습 경로 구성을 돕는 안내자의 입장에서 응답하도록 설계하였다. 대화 방식은 자연스러운 구어체를 채택하였으며, 기술 용어의 사용을 최소화하여 비전공자나 신입생도 쉽게 이해할 수 있도록 하였다. 특히 과목 이수 순서 안내 시에는 학습 경로를 명확하고 단계적으로 제시하도록 하였다.

프론트엔드는 React를 기반으로 웹 인터페이스를 구현하였으며, 직관적인 질문 입력창과 답변 표시창을 제공하여 사용자가 질의를 입력하고 응답을 쉽게 확인할 수 있도록 구성하였다. 이후 FastAPI 기반 백엔드와 REST API를 통해 통신하며, CORS(Cross-Origin Resource Sharing) 설정을 통해 프론트엔드-백엔드 간 연동이 원활하게 이루어지도록 하였다. 질의 요청은 /ask 엔드 포인트로 전달되며, 백엔드는 생성형 인공지능을 활용해 완성한 답변 결과를 프론트엔드로 다시 반환한다. 아래의 <그림 3>은 1학년 자율전공학부 학생이 인공지능 개발자와 관련된 교과목을 추천받는 상황을 가정하여 질의 후 응답 예시이다.

<그림 3> 시스템 질의 응답 예시

Ⅲ. 결론

본 연구에서는 생성형 인공지능과 검색 증강 생성을 기반으로 학과 정보 제공 시스템을 설계·구현하였다. 이어서 컴퓨터공학과 교육과정 문서를 청킹하여 임베딩 처리를 거쳐 데이터베이스를 완성하였다. 사용자가 질의를 하면 파이프라인을 통해 의미 기반 청킹 블록 검색을 수행하고 그 결과를 바탕으로 컨텍스트를 구성하여 생성형 인공지능에 전달함으로써 응답을 생성하도록 하였다. 생성형 인공지능에는 '컴퓨터공학과 멘토'라는 페르소나를 부여하여 신입생이 친근하고 이해하기 쉬운 방식으로 안내를 받을 수 있도록 하였고, 프론트엔드와 백엔드는 React와 FastAPI를 통해 연동되도록 구현하였다.

본 시스템은 컴퓨터공학과에 대한 정보만을 포함하고 있으나 향후 대학 내 모든 전공에 대한 정보를 포함하도록 시스템을 확장함으로써 융합전공, 복수전공, 부전공 등 유연한 학사제도에서 학생들이 활용할 수 있을 것이며 이를 통해 맞춤형 학습 지원 도구로서의 활용될 것으로 기대한다.

참 고 문 헌

[1] P. Lewis, E. Perez, A. Karpukhin, N. Piktus, F. Petroni, V. Karpukhin, P. Stenetorp, S. Riedel, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, Vol. 33, pp. 9459–9474, 2020.

[2] Q. Liu, M. J. Kusner, and P. Blunsom, "A Survey on Contextual Embeddings," arXiv preprint, arXiv:2003.07278, 2020.

[3] Gemini Team, A. Vyas, H. Xia, R. Urtasun, et al., "Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context," arXiv preprint, arXiv:2403.05530, 2024.