대규모 언어 모델과 벡터 데이터베이스 기반 교차문화 간 한식 추천 시스템 설계

윤하은, 박은수, 김윤희
숙명여자대학교 소프트웨어학부 컴퓨터과학전공

{haeun728, psrkeunsoo, yulan}@sookmyung.ac.kr

A Design of a Cross-Cultural Korean Food Recommendation System Based on Large Language Models and Vector Databases

Ha Eun Yun, Eunsoo Park, Yoonhee Kim
Department of Computer Science, Sookmyung Women's University

요 약

최근 한국 문화에 대한 관심 증가로 외국인의 국내 미식 체험 수요도 빠르게 확대되고 있으나, 기존 키워드 기반 검색은 사용자의 의도와 음식의 의미적 특성을 충분히 반영하지 못한다. 이에 본 연구는 LLM(Large Language Model) 기반 음식 프로파일링과 FAISS(Facebook AI Similarity Search) 기반 벡터 검색을 결합한 이종 문화 음식 추천 시스템을 제안한다. 사용자가 입력한 음식의 의미적 특성을 LLM으로 분석하고 벡터 유사도 검색을 통해 유사한 한식을 도출한다. Kubernetes 기반으로 구현되어 부하에 대한 확장성과 장애 허용성을 보장한다.

Ⅰ. 서론

최근 한국 문화에 대한 관심 증가로 2025년 외래관광객이 역대 최대치인 1,870만 명을 돌파할 것으로 전망되며[1], 방한 외국인의 국내 미식 체험 수요도 확대되고 있다. 그러나 기존 검색 엔진의 키워드 기반 방식은 사용자의 의도를 충분히 반영하지 못한다[2]. 이러한 문제를 해결하기 위해 본 연구는 LLM을 활용한 지능형 음식 프로파일링과 고성능 벡터 데이터베이스인 FAISS[5]를 결합한 이종 문화 간 음식 추천 시스템을 제안한다. 이 시스템은 Kubernetes로 구현하여 부하에 대한 확장성과 장애 허용성을 갖는다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 선행 연구를 고찰하고, 3장에서는 제안 시스템의 아키텍처를 설명한다. 4장에서는 구현 결과를 5장에서는 시스템의 확장성과 장애 허용성을 실제 운영 환경에서 실험을 통해 검증하였다. 마지막으로 6장에서는 결론 및 향후 과제를 제시한다.

Ⅱ. 선행 연구 및 기술 개요

1) 기존 시스템 한계- 검색 엔진에 입력하는 단순 키워드는 동의어 처리 및 다국어 환경에서의 의미 소실 등의 한계를 지닌다[2]. 최근 검색 엔진에서 AI 요약 서비스가 제공되나, 특정 도메인에 대한 근거 결여로 인해 환각 현상이 발생할 수 있다[3].

2) LLM 기반 프로파일링- 본 시스템은 LLM을 활용하여 사용자가 입력한 단일 키워드에 대해서 음식의 특징을 능동적으로 프로파일링 한다. 이를 통해 사용자의 질의 의도를 고차원 벡터로 변환할 수 있다.

3) 추천 시스템의 설명 가능성- Zhang & Chen (2020)은 설명 방식에 따라 모델 내재적 방식과 후처리적 방식으로 분류하였다[4]. 본 연구에서는 벡터 검색 엔진의 블랙박스 특성을 보완하기 위해 LLM 기반 후처리적 설명 방식을 채택하였다.

4) RAG 기반 추천- FAISS는 고차원 임베딩 벡터 간의 유사도 검색을 실시간으로 수행하는 엔진으로, 전수 조사 방식에서 높은 정확도와 낮은 지연 시간을 보인다[5]. 본 연구에서는 IndexFlatIP 인덱스와 L2 정규화를 결합한 코사인 유사도를 구현하였다.

Ⅲ. 제안 시스템 구조

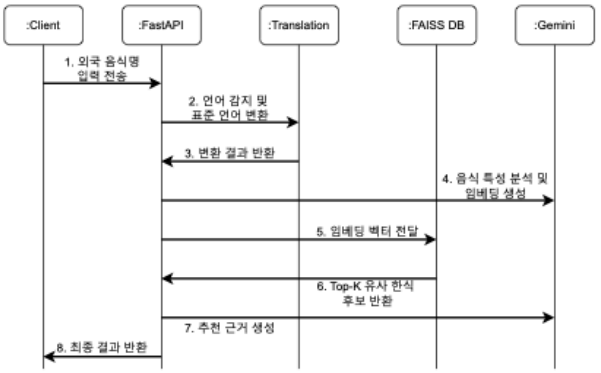


그림 1. 전체 시스템 시퀀스 다이어그램

그림 1은 본 시스템의 전체 구조를 나타낸다. 사용자는 웹 기반 Frontend 인터페이스를 통해 자국 음식 이름을 입력한다(1). 입력된 텍스트는 Backend API로 전달 되며(2), Backend API는 먼저 Translate Service를 호출하여 입력 언어를 분석 및 표준 처리 언어로 변환한다(3).

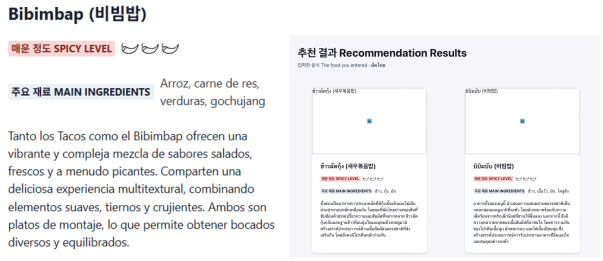
이후 변환된 음식명을 기반으로 LLM(Gemini)을 호출하여 해당 음식의 맛, 식감, 주요 재료, 조리 방식 등의 의미적 특성을 분석하고, 이를 고차원 임베딩 벡터로 변환한다(4). 생성된 임베딩 벡터는 FAISS 기반 벡터 데이터베이스로 전달되며(5), FAISS DB Service는 사전에 구축된 한식 임베딩 인덱스와의 코사인 유사도 계산을 통해 가장 유사한 한식 후보를 검색한다(6).

검색 결과는 다시 Backend API로 반환되며, Backend API는 LLM을 활용하여 추천 근거를 자연어로 생성한다(7). 필요 시 해당 설명은 Translate Service를 통해 사용자 입력 언어로

재번역되며, 최종 결과는 Frontend로 전달되어 사용자에게 시각적으로 제공된다(8).

IV. 구현 결과

(a) 자국 음식 입력 화면 (b) 한식 추천 결과 화면(스페인어)



(c) 결과에 대한 추천 근거 (d) 한식 추천 결과 화면(태국어)

그림 2. 사용자 화면

그림 3. 실제 검색 시스템에 동일한 질의를 수행한 결과

그림 2는 시스템의 사용자 인터페이스 화면을 나타낸다. 사용자는 'Tacos'를 입력하였다(a). 사용자는 자국 언어로 음식 이름을 입력한다. 시스템은 해당 입력을 자동으로 인식하여 내부 처리 언어로 변환한 뒤 추천 과정을 수행한다. Backend API는 LLM을 활용하여 입력 음식의 의미적 특성을 분석하고 FAISS 벡터 검색을 통해 가장 유사한 한식 후보 중 상위2개를 도출한다. 시스템은 'Tacos'와 유사한 한식으로 'Empanadilla de carne(고기만두)'와 'Bibimbap(비빔밥)'을 추천하고 결과에 대해 음식 간 유사성을 설명하는 추천 근거(c)를 함께 제공한다.(b) 스페인어를 입력하면 결과가 스페인어(b). 태국어는 태국어(d)로 출력한다. 그림 3은 Google 검색 엔진에 'tacos와 비슷한 한식'을 입력한 화면이다. 비빔밥, 쌈밥, 김밥 등의 한식을 단순 나열하여 제시하며, 음식 간 유사성에 대한 구체적인 기준이나 추천 근거는 제공되지 않는다.

V. Kubernetes 시스템 운영

1) 자동 확장(HPA) 테스트

그림 4. Pod의 자동 확장

Backend API에 대한 확장 성능을 검증하기 위해 사용자 요청이 증가하는 부하 테스트를 수행하였다. (그림 4) 테스트 결과, 요청 증가로 인해 CPU 사용률이 임계값을 초과하자 Backend Pod 수가 2개에서 최대 5개까지 자동으로 확장되었다. 확장 과정에서도 요청 처리 지연 없이 서비스가 안정적으로 유지되며 부하 상황에서도 응답 성능이 유지됨을 확인하였다. 이후 부하가 감소하자 Pod 수는 다시 최소 개수로 축소되어, 시스템이 부하 변화에 따라 안정성과 자원 효율성을 동시에 확보함을 검증하였다.

2) 자가 복구(Self-healing) 테스트

그림 5. 예기치 않은 pod 종료 시 자가 복구

Kubernetes의 Self-healing 기능을 검증하기 위해, 그림 5와 같이 FAISS DB Pod를 의도적으로 삭제하는 실험을 수행하였다. 그 결과 pod 종료 직후 새로운 Pod가 자동으로 재생성되며, PVC에 저장된 인덱스를 다시 로드하여 서비스 중단 없이 정상 동작을 유지한다. 이 과정에서 데이터 유실이나 서비스 중단 없이 검색 기능이 지속되었으며, 시스템이 예기치 않은 장애 상황에서도 안정적으로 복구됨을 검증하였다.

VI. 결론

본 연구는 LLM과 FAISS 기반 벡터 데이터베이스를 결합한 문화 간 음식 추천 시스템을 제안하였다. 제안된 시스템은 음식의 의미적 특성을 기반으로 유사도를 계산하고, LLM 기반 설명을 제공함으로써 기존 키워드 중심 추천 방식의 한계를 보완하였다. 또한 Kubernetes 기반으로 구현하여 시스템의 확장성과 안정성을 보장하였다. 향후 연구에서는 IVF, HNSW 등 FAISS의 다양한 인덱스 구조를 적용한 성능 비교를 진행하고자 한다.

참 고 문 헌

[1] 문화체육관광부, "2025년 외래관광객 역대 최다 1,870만 명 돌파 전망," 보도자료, 2025. 12. 23.

[2] Tina Gross, Arlene G. Taylor & Daniel N. Joudrey., "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching", Cataloging & Classification Quarterly, 53:1, 1-39, 2015.

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung., "Survey of Hallucination in Natural Language Generation". ACM Comput. Surv. 55, 12, Article 248, 38 pages. 2023.

[4] Yongfeng Zhang and Xu Chen, "Explainable Recommendation: A Survey and New Perspectives", Foundations and Trends® in Information Retrieval: Vol. 14, No. 1, pp 1–101., 2020.

[5] Rusum, G. P., & Anasuri, S., "Vector Databases in Modern Applications: Real-Time Search, Recommendations, and RAG,", IJAIBDCMS, 2024.