LLM 기반 차세대 추천 시스템: 아키텍처 설계와 구현에 관한 연구

문민석
고려대학교 SW.AI대학원 인공지능학과 석사과정
mpd2000@korea.ac.kr

LLM-based next generation recommender systems: A study on architectural design and implementation.

Moon, Min Suk
Korea Univ.

요 약

본 논문은 두 개의 LLM 모델과 Retrieval-Augmented Generation(RAG) 기법을 결합한 추천 시스템 아키텍처를 제안하였다. 이를 통해 도메인 특화된 추천과 자연스러운 문장 생성을 동시에 달성하며, 바이어스 문제를 해결하고 추천 결과의 신뢰성을 높였다. 또한, 로드 밸런싱 및 요청 큐 관리를 통해 자원을 효율적으로 사용하는 방안을 제시했고, 다양한 설정 값 조정을 통해 맞춤형 추천이 가능함을 입증하였다. 제안된 시스템은 오픈소스를 활용해 구현 사례를 제공함으로써 비즈니스 및 개인의 진입 장벽을 낮추고, 추천 시스템의 실질적인 활용 가능성을 보여준다.

Ⅰ. 서 론

최근 생성형 AI(Generative AI)를 활용한 추천 시스템 연구가 활발히 진행되고 있으며, 특히 Collaborative Filtering 과 대규모 언어 모델(LLM)을 결합한 아키텍처[1], 특정 생성형 모델을 기반으로 한 구현 사례[2] 등 다양한 연구가 이루어지고 있다. 하지만, 대부분의 연구가 기업 내부 프로젝트로 진행되거나 비공개 데이터를 활용하고 있어 구체적인 구현 방법과 성능 최적화 전략에 대한 정보는 제한적이다.

LLM 기반 추천 시스템은 기존 알고리즘 기반 추천 시스템과 비교해 자연어 처리와 문맥 이해 능력을 활용하여 더 개인화되고 정교한 추천을 가능하게 하며, 비정형 데이터를 처리하고 복합적인 사용자 의도를 파악해 유연한 응답을 제공한다. 또한, 다양한 데이터 소스를 결합하고 자연스러운 문장으로 결과를 전달하여 사용자 경험을 향상시킬 수 있다.

그러나 이러한 시스템의 구축과 운영에는 도메인 특화 추천의 어려움, 자원 제약, 응답 시간 지연 문제, 그리고 높은 인프라 비용과 복잡성 등 여러 기술적 과제가 존재한다. 이에 본 논문에서는 오픈소스 기반 LLM 추천 시스템 아키텍처를 설계하고 구현한 사례를 제시하며, 도메인 특화 추천, 자원 최적화, 응답 시간 개선을 위한 구체적인 방안을 연구한다. 또한 특정 도메인 적용 사례를 통해 비즈니스 활용 가능성을 검증하고, 기업 및 개인이 보다 쉽게 LLM 을 활용한 추천 시스템을 구현하고 사용할 수 있도록 지원하는 것을 목표로 한다.

Ⅱ. 본론

본 논문에서는 오픈소스를 기반으로 LLM 추천 시스템을 구축하고 인프라 최적화 방안을 연구하였다.

1. 추천 시스템 아키텍처 설계

두 개의 LLM 모델과 RAG 시스템을 결합한 아키텍처를 제안하였다.

첫 번째 모델: 도메인 특화 추천을 위해 특정 데이터로 파인튜닝된 LLM 을 활용한다. 이 모델은 추천할 상품 목록, 설명, 주요 고객 특성(예: 연령, 성별), 과거 구매 이력 등을 학습해 정교한 추천 결과를 제공한다.

두 번째 모델: 기본 Pretrained 모델(예: LLaMA, Qwen)을 사용해 추천 결과를 자연스러운 문장으로 변환한다. 이는 파인튜닝된 모델이 특정 바이어스에 치우쳐 자연어 질의에 제대로 답변하지 못하는 문제를 해결하기 위함이다.

Tool-calling 기법을 활용해 결과를 체계적으로 관리하였다. Lang-Chain, ollama 등 오픈소스 프레임워크를 활용해 모델을 구축하고 배포하였다.

2. RAG 시스템 적용

Retrieval-Augmented Generation(RAG) 기법을 도입하여 LLM 이 생성하는 결과를 벡터 데이터베이스와 연동하였다. 이를 통해 추천 시스템이 사용자 질의와 데이터베이스의 정보를 효과적으로 결합하여 정확하고 도메인에 특화된 추천 결과를 제공하도록 하였다.

또한, RAG 기법을 활용함으로써 Fine-tuning 모델의 추천 결과를 검증하고 강화하여 추천 결과의 신뢰성을 높였다. RAG 를 활용해 사용자의 질의에서 주요 키워드를 추출하고 이를 프롬프트에 삽입하여 사용함으로써, 질의 응답의 정확성을 높이는 동시에 토큰 수를 줄여 GPU 자원 사용량을 절감하는 데 기여하였다.

3. 인프라 최적화 및 시스템 효율화

GPU 자원 부족 문제와 응답 시간 지연을 해결하기 위해 로드 밸런싱과 요청 큐 관리 전략을 적용하였다.

GPU 자원 사용량을 기존 대비 10% 절감 RAG 를 활용해 질의 프롬프트에서 불필요한 토큰 수를 줄이고 병렬 처리 최적화를 통해 자원 활용 효율성을 증대하였다.

그리고 요청 큐를 통해 동시 요청을 안정적으로 처리하면서 평균 응답 시간을 줄였다.

이러한 최적화 전략은 시스템 운영 비용을 줄이는 동시에 성능을 최대하 하는 방안을 제시한다.

4. 실제 구현 사례 및 비즈니스 응용

제안된 아키텍처를 특정 도메인에 적용한 구현 사례를 제시하였다. 이를 통해 비즈니스 응용 가능성을 검증하고, 오픈소스 기반 접근법을 통해 인프라 구축 비용과 진입 장벽을 낮추어 기업 및 개인이 보다 쉽게 추천 시스템을 구현할 수 있도록 제안하였다.

5. 추천 다양성 (Recommendation Diversity)

본 시스템은 Temperature, Top-K, Top-P 값의 조정을 통해 추천의 다양성을 쉽게 확보할 수 있도록 설계되었다.

해당 시스템을 활용하여 ILD (Intra-List Diversity) 값을 쉽게 조정 가능하게 했으며 ILD 값이 기존 시스템 대비 크게 증가시켜 다양한 추천 결과를 제공할 수 있었다.

이 시스템은 추천 목록의 다양성을 확보하기가 용이하며, 특정 추천 목표가 없는 경우에도 폭넓은 선택지를 사용자에게 제공할 수 있는 장점을 지녔다.

위 내용을 통해 제안된 시스템은 정교한 추천과 다양한 응답을 제공하며, 자원 효율성을 극대화한 인프라 설계로 실질적인 비즈니스 활용 가능성을 입증하였다.

Ⅲ. 결론

본 논문은 LLM 기반 추천 시스템의 아키텍처 설계와 최적화 전략을 제시하여 실제 서비스 적용 시의 기술적 문제를 해결하였다.

파인튜닝된 LLM 과 Pretrained LLM 을 결합해 도메인 특화 추천과 자연스러운 문장 생성을 달성 하였고 RAG 시스템으로 추천 결과의 신뢰성과 정확성을 강화 하였고 로드 밸런싱과 요청 큐 관리로 자원 최적화 및 응답 시간 개선 하였다. 그리고 다양한 도메인 사례를 통해 비즈니스 응용 가능성을 검증하고, 오픈소스 접근으로 진입 장벽을 낮춤. 향후 연구에서는 더 다양한 데이터와 도메인에 시스템을 적용해 확장성과 성능을 검증할 예정이다.

ACKNOWLEDGMENT

참 고 문 헌

[1] Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System. 2024

[2] LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. 2023