동적 지식 갱신을 지원하는 완전 오프라인 온디바이스 RAG 아키텍처 연구

김진혁

숭실대학교 일반대학원 인공지능IT융합학과

cancon@naver.com

지도교수: 박민호 교수

An On-Device RAG Architecture for Fully Offline Operation and Dynamic Knowledge Updating

Jinhyuk Kim

Soongsil Univ.

Advisor: Prof. Minho Park

요 약

본 논문은 기존 RAG(Retrieval-Augmented Generation) 시스템이 가진 클라우드 의존성 및 정적 지식 기반의 한계를 극복하고, 온디바이스 환경에서 완전 오프라인으로 동작하며 동적 지식 갱신이 가능한 통합 RAG 아키텍처를 설계하고 구현하는 것을 목표로 한다. 기존 연구가 대부분 서버형 LLM 및 벡터 데이터베이스에 의존하는 것과 달리, 본 시스템은 소형 언어모델(SLM), 임베딩 모델, 그리고 온디바이스 벡터 데이터베이스(ObjectBox)를 단일 안드로이드 기기 내에 통합한다.

특히, 사용자의 반복적인 질의와 그에 대한 LLM의 응답을 증분 개인화 메모리에 저장하고, 유사 질의 발생 시 기존 지식과 새로운 정보를 합성하여 답변을 갱신하는 메커니즘을 제안한다. 이는 네트워크 연결 없이도 시스템이 사용자 패턴을 학습하고 시간이 지남에 따라 응답 품질을 점진적으로 향상시키는 동적 지식 갱신을 가능하게 한다.

Ⅰ. 서 론

본 논문에서는 동적 지식 갱신을 핵심 목표로 하며, 네트워크 연결 없이 완전 오프라인으로 동작하는 RAG 파이프라인을 온디바이스 환경에 최적화하여 설계하고 구현하는 것을 목표로 한다.

1.1. 연구 배경 및 필요성

대규모 언어모델(LLM)이 다양한 분야에서 활용되면서, 지식 기반 생성(RAG)에 대한 수요가 빠르게 증가하고 있다.[3] 그러나 기존 RAG 시스템은 대부분 클라우드 기반 LLM과 서버형 벡터 데이터베이스에 의존하여, 네트워크 없는 환경에서는 정상적으로 동작할 수 없다. 또한, 사용자 질의와 검색 대상 데이터가 외부 서버로 전송되는 구조로 인해 개인정보 및 민감 정보 노출 위험이 존재하며, 보안 측면에서도 취약하다. 더불어, 정적인 문서 집합을 기반으로 하기 때문에 사용자의 반복적 상호작용을 반영한 지식의 동적 갱신이 불가능하다.

본 연구는 이러한 한계를 해결하기 위해 온디바이스 환경에서 완전 오프라인으로 실행되는 RAG 아키텍처를 제안한다. 특히, LLM의 품질을 크게 향상시키는 RLHF(Reinforcement Learning from Human Feedback) 기법은 대규모 데이터/반복학습/서버 연산을 필요로 하므로 온디바이스에서는 적용 자체가 불가능하다.[4] 따라서, 클라우드 연결 없이 사용자의 반복 질의와 응답을 점진적으로 축적하여 개인화된 지식을 형성하는 오프라인 개인화 RAG 시스템의 필요성이 크다.

1.2. 기존 연구의 한계

기존 RAG 연구는 다음과 같은 세 가지 주요 한계를 가진다.

첫째, 클라우드 의존성과 프라이버시 문제이다. 서버 연결을 전제로 하므로, 오프라인 환경에서 동작이 제한되고 사용자 데이터가 외부에 노출될 위험이 있다.

둘째, 오프라인 RAG의 정적 지식 구조이다. 기존 연구된 오프라인 RAG은 대부분 고정된 문서 집합을 기반으로 하며, 사용자 패턴을 반영하는 개인화된 구조가 미흡하다. 특히, LLM의 응답 품질 개선에 효과적인 RLHF 메커니즘은 지속적인 서버 연결 및 대규모 데이터셋 갱신을 필요로 하므로, 완전 오프라인 환경에서는 현실적으로 적용이 어렵다는 근본적인 제약을 가진다.

셋째, 통합 아키텍처 부재이다. 생성 모델, 임베딩 모델, 검색 엔진, 쿼리 라우팅 및 증분 저장을 단일 기기 내에서 완전히 동작시키는 통합 시스템 사례가 매우 제한적이다.

1.3. 연구 목적 및 기여

본 연구의 핵심 아이디어는 사용자의 반복적인 질의-응답 상호작용 자체를 지식 갱신을 위한 암묵적 피드백으로 활용할 수 있다는 점이다. 기존의 RLHF 기반 강화 학습법이 외부 서버, 대규모 데이터, 반복 학습을 필요로 하는 반면, 본 연구는 온디바이스 환경에서 발생하는 사용자 질의 패턴과

응답의 누적을 통해 개인화된 지식을 점진적으로 형성하는 대안적 지식 갱신 방식을 제안한다.

본 연구는 다음 두 가지 핵심 목표를 가진다.

첫째, SLM, 임베딩 모델, SLM 기반 쿼리 라우팅, ObjectBox 벡터 DB를 모두 통합하여 완전 오프라인 RAG 파이프라인을 구축하여 클라우드 비의존 환경에서 검색/생성/저장 단계를 모두 실행한다.

둘째, 핵심 기여는 RLHF와 같은 외부 피드백 루프 없이 온디바이스 자체 환경에서 응답 품질을 향상시키기 위해, 사용자의 반복 질의를 기반으로 동적으로 지식을 합성하고 응답을 갱신하는 증분적 동적 지식 갱신 메커니즘을 제안 및 구현하는 데 있다. 이는 네트워크 의존 없는 개인화 지식 검색 및 생성을 실현함으로써, 온디바이스 AI 서비스의 실용성을 높이는 것을 목표로 한다.

이를 바탕으로, 본 연구는 서버 연동이나 RLHF와 같은 외부 학습 루프 없이도 오프라인 온디바이스 환경에서 개인화된 지식 축적이 가능함을 보임으로써, 향후 엣지 디바이스 환경에서 실용적으로 적용 가능한 RAG 시스템 설계의 새로운 방향성을 제시한다.

본 논문의 주요 기여는 다음과 같다.
(1) 완전 오프라인 환경에서 동작하는 온디바이스 RAG통합 아키텍처 제안
(2) 사용자 반복 질의 기반 증분적 동적 지식 갱신 메커니즘 제안
(3) 서버 연동 및 RLHF 없이 개인화된 지식 축적 가능성 실증

본 논문의 구성은 다음과 같다.
(1) Ⅱ장에서는 완전 오프라인 온디바이스 RAG 아키텍처 모듈들과 동적 지식 갱신 메커니즘을 설명한다.
(2) Ⅲ 장에서는 결론 및 향후 연구 방향을 제시한다.

Ⅱ. 본론

본 장에서는 완전 오프라인 환경에서 동작하는 온디바이스 RAG 아키텍처를 구성하는 핵심 모듈들과, 이들이 상호작용하여 동적 지식 갱신을 수행하는 구체적 메커니즘을 기술한다.

2.1 전체 아키텍처 구성

본 연구에서 제안하는 온디바이스 RAG 시스템은 총 4개의 핵심 컴포넌트로 구성된다.

1) SLM 기반 LLM 모듈 (Generation + Query Routing 통합)
Hugging Face 공개 모델인 Kanana 1.5B-2.1B-instruct-2505[1]를 기반으로 질의 도메인 분류(Query Routing)와 응답 생성을 동시에 수행한다. 단일 모델 구조를 통해 모델 수 증가 없이 온디바이스 메모리 점유를 최소화한다. Qualcomm QNN SDK 기반 INT8 양자화를 적용하여 안드로이드 기기에서 NPU를 활용하여 실시간 추론이 가능하다.

2) 임베딩 및 검색 모듈 (Embedding + Vector Retrieval)
Hugging Face 공개 모델인 BGE-M3 모델[2]을 ONNX Runtime 기반 INT8로 경량화하여 1024차원 임베딩 벡터를 생성한다. 생성된 임베딩은 ObjectBox 벡터 DB 내 ANN(Approximate Nearest Neighbor) 검색을 통해 관련성을 평가한다. 검색은 도메인 단위로 수행되며, 코사인 유사도 기반 Top-N 후보를 반환한다.

3) 증분 개인화 메모리 (Incremental Personalized Memory)
본 연구의 핵심 기여로, 다음 과정을 통해 동적 지식 갱신을 수행한다.

(1) 사용자의 질의와 LLM 응답을 벡터로 변환
(2) 동일 도메인 내 유사도 ≥ 0.85인 항목 존재 시 → 기존 지식과 병합
(3) 유사도 < 0.85인 경우 → 새로운 지식 엔트리 생성
(4) 엔트리 업데이트 시 기존 응답을 보강하며 사용자 맞춤 지식이 강화됨

이를 통해 실시간 사용자 행동 기반 지식 업데이트가 가능하며, 완전 오프라인 환경에서도 지속적으로 정밀도가 향상되는 RAG 구조를 실현한다.

4) 오프라인 실행 및 모델 최적화 엔진
완전 오프라인 구현을 위해 다음 최적화 전략을 적용하였다.

- SLM(Kanana 1.5B-2.1B-instruct-2505): INT8 정적 양자화(QNN SDK)
- BGE-M3 Embedding: INT8 동적 양자화(ONNX Runtime Mobile)
- Vector Storage: Float32 → Float16 변환 저장

이러한 최적화 기법은 온디바이스 환경에서 완전 오프라인으로 RAG를 실행할 수 있게 하는 핵심 기술적 기반이다.

2.2 동적 지식 갱신(Dynamic Knowledge Update) 메커니즘

본 연구에서 제안한 동적 지식 갱신 엔진은 RLHF의 핵심을 온디바이스 환경에 맞게 차용한 구조다. 사용자의 반복 질의를 implicit feedback(암묵적 보상)으로 간주하여 유사도 기반 지식 병합을 통해 기존 지식이 강화하고 신규 질의에 대해서는 새로운 지식 엔트리를 생성하여 개인화된 지식 데이터가 확장되도록 한다.

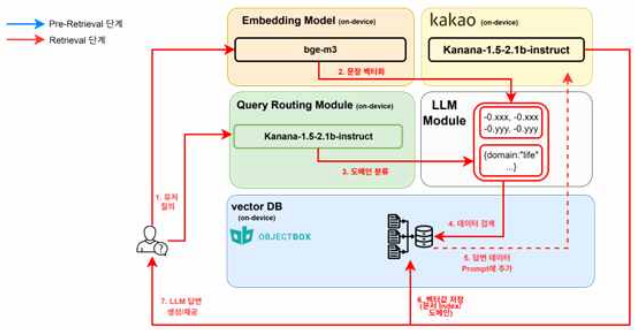

그림 1 〈전체 아키텍처 구성 및 동작 과정〉

동작 과정은 아래와 같다.
(4) 사용자 질의 입력
사용자가 자연어 질의를 입력한다.

(5) SLM 기반 Query Routing
SLM은 입력 질의를 분석하여 도메인 정보를 분류한다. 해당 도메인 정보는 이후 벡터 검색 범위를 제한하는 메타데이터로 활용된다.

(6) BGE-M3 임베딩 생성
도메인화 된 사용자 질의는 임베딩 모델(BGE-M3)을 통해 고차원 임베딩 벡터로 변환된다.

(7) ObjectBox에서 유사 벡터 검색
생성된 질의 임베딩은 ObjectBox 벡터 데이터베이스에서 동일 도메인 내 벡터들을 대상으로 HNSW 기반 ANN 검색을 수행한다. 이 과정에서 코사인 유사도 기준 Top-N 후보가 반환된다.

(8) 유사도 판단
검색 결과 중 최대 유사도를 기준으로 다음과 같이 처리한다.
- ≥ 0.85: 기존 응답과 병합하여 지식 업데이트
- < 0.85: 신규 지식 생성

(9) SLM 응답 생성
- 유사 지식이 존재하는 경우: 검색된 컨텍스트를 포함한 RAG 프롬프트로 SLM이 응답 생성 후 벡터 DB 지식 갱신
- 유사 지식이 없는 경우: 질의 단독 입력으로 SLM이 응답 생성 후 벡터 DB 지식 저장

(10) 지식을 벡터 DB에 반영하여 지속적 지식 갱신 효과 제공
최종 응답과 해당 질의 임베딩은 ObjectBox 벡터 DB에 저장되며, 이 과정이 반복됨에 따라 개인화된 지식 메모리가 점진적으로 강화된다.

이 과정을 반복함으로써 시스템은 사용자 패턴을 반영하는 개인화된 RAG 결과를 생성하게 된다.

Ⅲ. 결론

본 연구는 기존 RAG 구조가 가진 클라우드 의존성, 정적 지식 구조, 통합 아키텍처 부재 문제를 해결하기 위해 동적 지식 갱신과 완전 오프라인 구현이 가능한 온디바이스 RAG 아키텍처를 제안하였다.

제안된 시스템은 네트워크가 없는 환경에서도 RAG 전체 파이프라인이 동작하고 사용자의 반복 질의 기반으로 지식이 점진적으로 강화되며 RLHF 없이도 온디바이스 자체에서 동적 지식 갱신이 가능성을 확인하였다.

특히, 증분 개인화 메모리 구조는 차량용 AI 비서, IoT 단말기, 군사기기, 의료기기 등 실제 응용 환경에서 고비용 서버 인프라 없이도 개인화된 AI 서비스 기능을 구현할 수 있는 잠재적 활용 가능성을 제시한다.

참 고 문 헌

[1] https://huggingface.co/kakaocorp/kanana-1.5-2.1b-instruct-2505

[2] https://huggingface.co/BAAI/bge-m3

[3] Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP”, NeurIPS 2020.

[4] Ouyang et al., “Training Language Models to Follow Instructions with Human Feedback”, NeurIPS 2022.