

ARAG: Agentic Retrieval Augmented Generation for Personalized Recommendation

<https://www.arxiv.org/pdf/2506.21931>

0. Introduction

- 사용자 장기 선호와 세션 기반 즉각적 의도는 추천 성능에 큰 영향을 주지만 기존 방법은 이를 통합적으로 모델링하기 어려움.
- RAG 파이프라인에 다중 LLM 에이전트 구조 도입
- 사용자 이해, 의미 정렬, 컨텍스트 요약, 랭킹을 역할별 에이전트로 분리
- 기존 RAG 및 recency 기반 추천 대비 추천 성능 크게 향상

1. Overview

- RAG 기반 추천 파이프라인에 에이전트 기반 추론 구조를 결합하여, 사용자 의도와 아이템 의미 정렬을 단계적으로 수행
- 기본 구조
 - 임베딩 기반 1차 후보 검색
 - 다수의 LLM 에이전트가 사용자 이해, 의미 정렬, 요약, 랭킹을 순차 수행
 - 최종 에이전트가 종합 신호 기반 추천 순위 생성
- 적용 범위

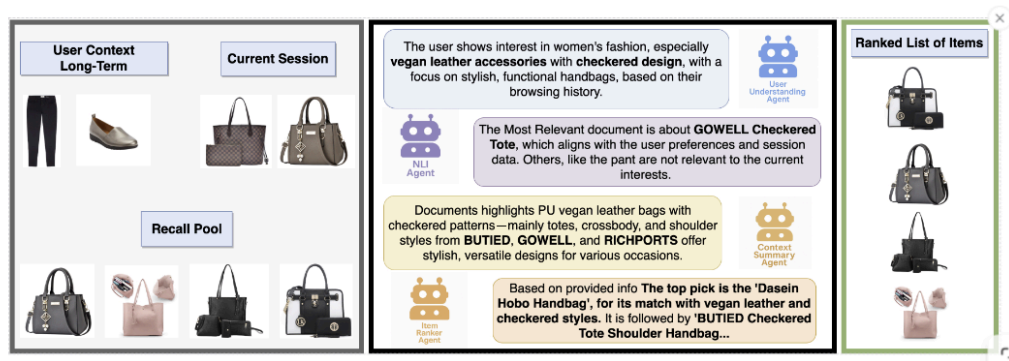
Amazon 카테고리 데이터셋(의류, 전자제품, 홈) 기반 개인화 추천 실험 수행
- 연구 목표

동적 사용자 의도 반영, 추천 해석 가능성 향상, agentic reasoning 기반 추천 구조 제안

2. Challenges

- 정적 검색 한계
임베딩 유사도 기반 검색은 세부 의미 정렬과 복합 의도 반영에 한계 존재
- 사용자 의도 변화 문제
장기 선호와 세션 기반 즉각적 의도를 동시에 반영하기 어려움
- 추천 정확도 문제
기존 RAG 및 recency 기반 접근은 복합 컨텍스트 환경에서 낮은 랭킹 성능을 보임

3. Method



- Initial Retrieval
 - 사용자 장기 행동과 세션 기록을 임베딩하여 초기 후보 아이템 집합 생성
- Multi-Agent Pipeline
 - User Understanding Agent
 - 장기 및 세션 행동 기록을 요약하여 사용자 선호 텍스트 표현 생성
 - NLI Agent
 - 사용자 의도와 각 후보 아이템 설명 간 의미적 일치 여부 평가
 - Context Summary Agent
 - 후보 아이템별 핵심 컨텍스트 신호 요약
 - Item Ranker Agent
 - 사용자 선호, 의미 정렬 결과, 요약 정보를 통합하여 최종 랭킹 생성

- Ranking Generation
 - 다중 에이전트 출력을 누적 반영하여 추천 순위 결정

4. Experiments

- 사용 데이터
Amazon 의류, 전자, 홈 카테고리 사용자 리뷰 및 아이템 메타데이터
- 실험 설정
 - 비교 대상: 기본 RAG 추천, recency 기반 추천
 - 평가 지표: NDCG@5, Hit@5
- 추가 실험
 - 에이전트 구성별 ablation study 수행
 - 각 에이전트 기여도 분석

5. Results

Performance of Benchmark Versus ARAG						
	Clothing		Electronics		Home	
	NDCG@5	Hit@5	NDCG@5	Hit@5	NDCG@5	Hit@5
Recency-based Ranking	0.30915	0.3945	0.22482	0.3035	0.22443	0.2988
Vanilla RAG	0.29884	0.3792	0.23817	0.321	0.22901	0.3117
Agentic RAG	0.43937	0.5347	0.32853	0.4201	0.28863	0.3834
% Improvement	42.12%	35.54%	37.94%	30.87%	25.60%	22.68%

Ablation Study						
	Clothing		Electronics		Home	
	NDCG@5	HIT@5	NDCG@5	HIT@5	NDCG@5	HIT@5
Vanilla RAG	0.29884	0.3792	0.23817	0.321	0.22901	0.3117
ARAG w/o NLI & CSA	0.3024	0.3859	0.2724	0.3559	0.2494	0.3308
ARAG w/o NLI	0.3849	0.4714	0.296	0.3878	0.2732	0.3582
ARAG	0.43937	0.5347	0.32853	0.4201	0.28863	0.3834

- 주요 성능 결과
 - NDCG@5 최대 약 42% 향상
 - Hit@5 최대 약 35% 향상
- 카테고리별 성능

의류 카테고리에서 가장 큰 개선 효과 확인

- Ablation 결과
 - User Understanding Agent 단독 추가 시 성능 상승
 - NLI + Context Summary 추가 시 추가 개선
 - 전체 파이프라인 통합 시 최고 성능 달성
- 에이전트 기반 구조가 의미 정렬 및 개인화 정확도를 효과적으로 향상시킴

6. Insight

- 단일 RAG 구조보다 역할 분리된 에이전트 협업 구조가 추천 품질에 효과적
 - 사용자 장기 선호와 세션 의도를 분리 모델링함으로써 동적 추천 문제 해결 가능
 - NLI 기반 의미 정렬은 임베딩 유사도 기반 검색을 정밀하게 보완
 - 다수 LLM 호출로 인한 추론 비용 및 지연 문제 존재 가능
 - 대규모 상용 시스템 적용 시 확장성 추가 검증 필요
 - 에이전트 간 상호작용 최적화
 - 비용 및 추론 효율 개선 연구
-