[IMEN891K] 특론 : 데이터엔지니어링

AI Research Assistant
: arXiv 기반 연구 보조 서비스

Term Project

IDS 20252552 김송일

SDS 20242629 박기나

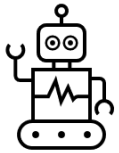
IME 20252322 박현민

IDS 20252732 박지원

Contents

1. Introduction
2. Overview
3. Database Schema
4. Project Progress
5. Future Work

Introduction



Background

주제 탐색	방법론 설계	실험	결과 해석	논문 작성

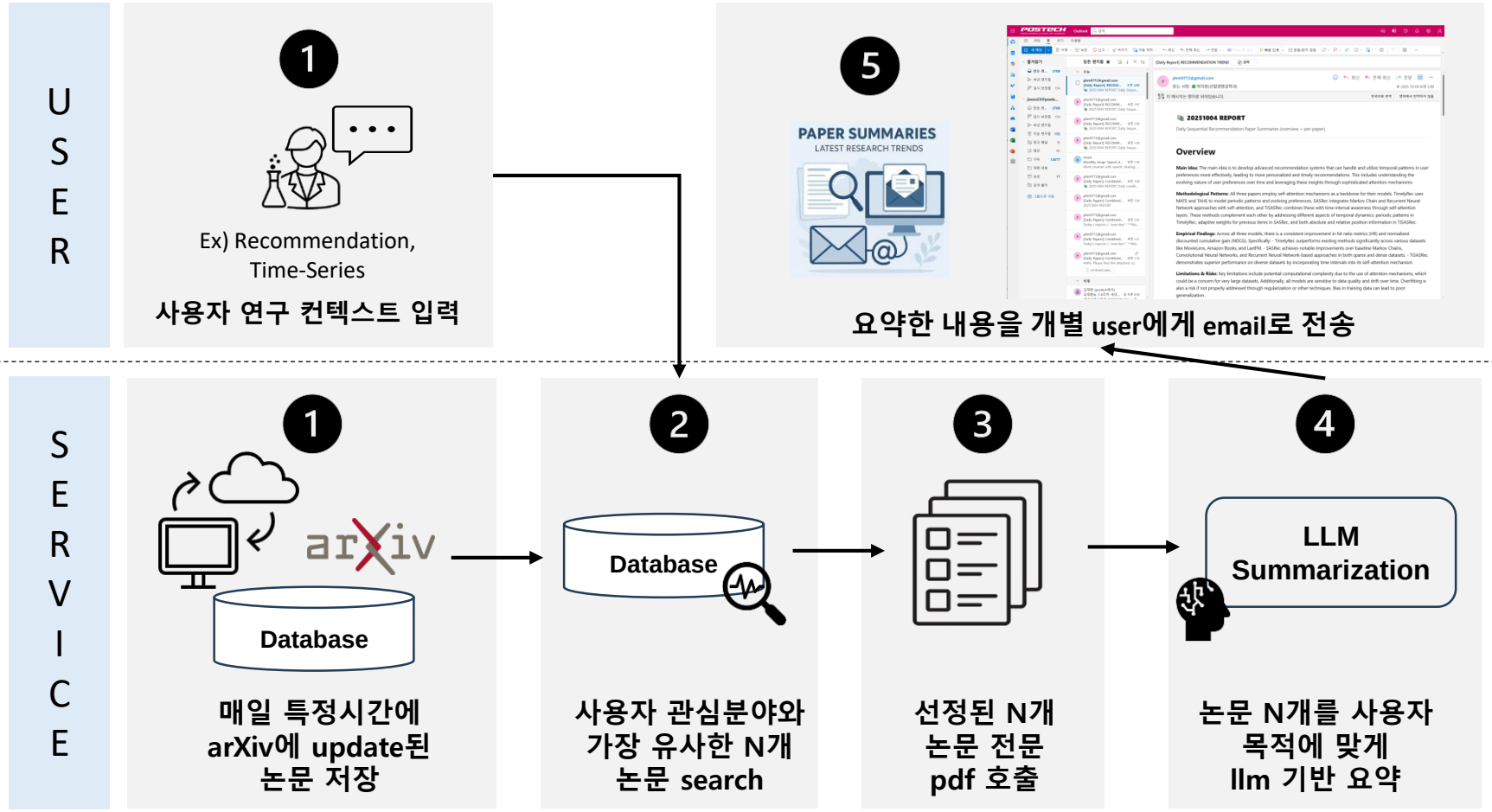
- AI, ML 발전으로 연구 발전 속도가 가속화되며, 최신 연구동향을 **지속적으로 추적**하기 어려워짐
- 개인의 연구 목적에 맞춘 맞춤형 논문 선별, 요약 도구 부재
- 연구자들의 연구 생산성을 향상시켜주는 연구 보조 Assistant 필요성 증대

Objective

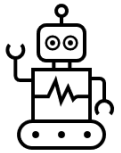
- 연구자 개별 목적에 따른 연구 수행 과정 전반을 보조하는 AI Assistant 개발
  - **연구동향 탐색 보조 :** 사용자 관심 분야의 최신 동향을 자동으로 추적, 관련 논문을 선별하여 요약
  - **방법론 설계 보조:** 유사 연구의 비교, 분석을 통해 연구자가 적절한 방법론을 설계할 수 있도록 지원
  - **결과 해석 보조 :** 유사 연구의 실험 결과 제공 및 분석 내용 요약
- 연구 Trend report 생성 AI Assistant
  - 연구 전반에 대한 최신 동향을 자동으로 추적, 관련 논문을 선별하여 요약 제공

Overview

- 전체 파이프라인



Database Schema – User



- **User DB**
  - 사용자 별 개인정보 및 연구 Context 저장

Table	Column	설명
users	id	사용자 ID
	email	로그인 이메일
	password_hash	해시된 비밀번호
	created_at	생성 시각

Table	Column	설명
research_ contexts	id	연구 컨텍스트 ID
	user_id	사용자 ID
	research_abstract	사용자의 연구를 기술한 내용
	keywords	연구와 관련된 추적 키워드
	is_active	현재 사용 중인지 표시
	created_at	생성 시각
	updated_at	마지막 수정 시각

Database Schema – Service

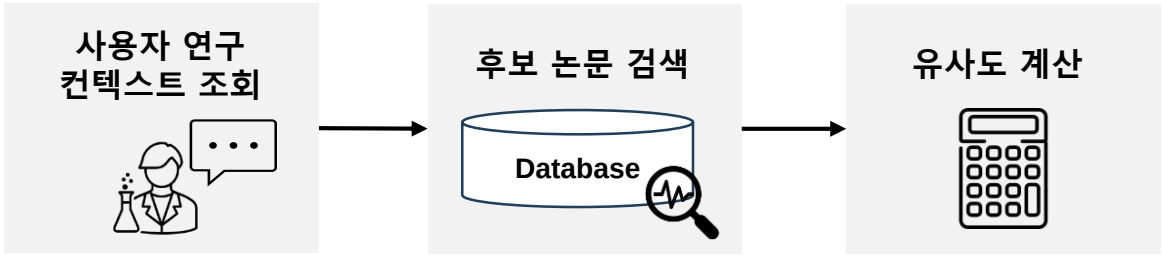
- **Paper DB**
  - 매일 update 하는 형태; 일일 배치 스케줄링



Table	Column	설명
paper	arxiv_id	arXiv ID
	title	논문 제목
	abstract	논문 초록
	authors	저자 목록
	primary_category	대표 카테고리
	categories	전체 카테고리
	published	최초 게시 시각(UTC)
	updated	최신 갱신 시각(UTC)
	doi	DOI 식별자
	comment	저자 코멘트/페이지 수 등
	journal_ref	저널 게재 정보
	url_abs	arXiv 초록 페이지 URL

Table	Column	설명
paper_summaries	arxiv_id	arXiv ID
	one_line	핵심 내용 요약
	contribs	핵심 기여사항
	methodology	사용된 방법론
	data_metrics	사용 데이터셋과 평가 지표
	main_results	대표 결과
	limits_future	한계점과 제안된 향후 연구
	overview	논문 전반의 개요
	model_name	요약 생성에 사용된 모델명
	updated_at	마지막 수정 시각

Project Progress



CURRENT

❶ 사용자 연구 컨텍스트 조회

- 기존 사용자가 정의한 컨텍스트를 참조 (DB 내 테이블 참조)

❷ 논문 저장 및 후보 논문 검색

- arXiv에서 일 단위 api 호출 후 DB에 논문 저장
- DB에 저장된 논문들 중, 분석할 후보군 지정 ; 날짜 기반 filtering

❸ 유사도 계산

- 연구 컨텍스트와 후보 논문 abstract 간의 유사도를 계산하여 가장 관련성 높은 문서를 선별
- 임베딩 모델(gemini-embedding-001) 을 사용하여 텍스트를 벡터로 변환 후 코사인 유사도로 비교

Project Progress

4 LLM 기반 요약 파이프라인

- 선정된 논문 N개를 사용자 목적에 맞게 LLM으로 요약하는 모듈
- N개의 논문 통합 요약 overview와 각 논문별 요약을 json & md 형태로 생성

Input : 선정된 N개 Paper pdf
Model : Qwen 2.5 – 3b or 7b
Output : Combined_report.json

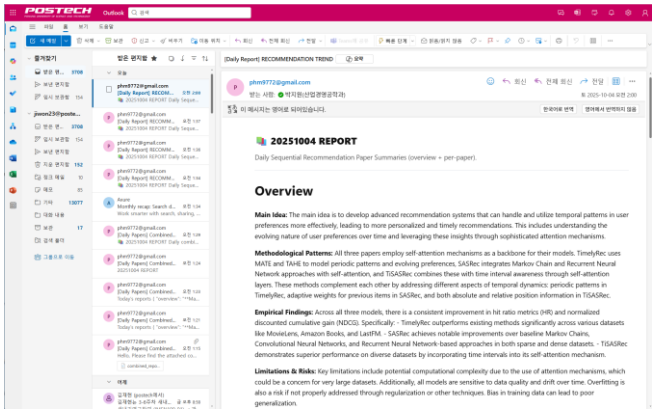
```
{ "overview": "Nowadays, sequential recommendation ...",
 "papers": [
  {
   "paper_title": "...",
   "one_line_summary": "...",
   "key_contributions": ["...", "...", "..."],
   "methodology": "...",
   "datasets_and_metrics": "...",
   "main_results": "...",
   "limitations_and_future_work": "..."
  },  ... ]}
```

5 LLM 기반 요약 결과 E-mail 전송

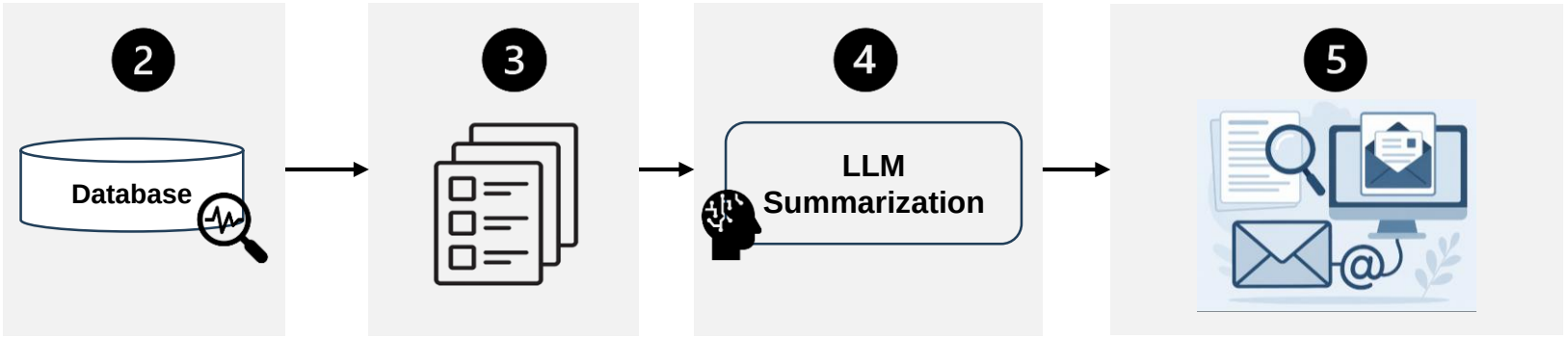
- User DB에서 메일정보를 추출하여, LLM 요약 결과를 Gmail SMTP를 통해 사용자에게 전송

Email UI Design

Gmail SMTP



Future Work



TO-DO

❷ 사용자 관심사 기반 유사 논문 검색

- 사용자 관심사를 embedding 한 뒤, DB의 논문 제목 + 초록과 유사도 비교로 Top-N 검색
  ; keyword matching & TF-IDF & embedding 점수 기반 유사도 측정

❸ 선정 논문 PDF 전문 호출 후 LLM Input으로 연결

❹ LLM 기반 요약 파이프라인

- PDF에서 텍스트를 추출해 섹션별로 나눈 뒤 LLM으로 목적에 맞게 요약
  ; prompt engineering & few shot example 생성 & model 변경하여 test

❺ 요약 메일 전송

- User DB에서 email 확인 후, 반응형 html 템플릿 기반으로 LLM 요약 결과 전송

EOD

Thank You ☺