1. 연구 개요

이 연구에서는 **영상 → 오디오 → 텍스트 → 키워드/엔티티 → 그래프+벡터 RAG**로 이어지는 파이프라인을 설계했다.

단순히 관련 문서 몇 개를 보여주는 RAG 시스템이 아니라, **역사적 개념·사건·인물 간의 관계를 지식 그래프 수준에서 실시간으로 탐색할 수 있는 프레임워크**를 목표로 삼았다.

핵심 목표는 다음 세 가지로 정리되었다.

(1) 멀티모달 사료 처리: 영상·오디오에서 텍스트를 추출하고, 이를 1차 사료 텍스트와 함께 통합 검색·분석할 수 있는 기반을 마련했다.

(2) 지식 그래프 기반 RAG: 텍스트에서 엔티티와 관계 트리플을 추출하여 **지식 그래프와 FAISS 벡터 인덱스**를 함께 구축하는 구조를 설계했다.

(3) 실시간 지식 탐색 UI: 영상 전사 텍스트에서 키워드를 추출하고, 해당 키워드와 연결된 그래프 노드를 실시간으로 시각화하여, 사용자가 노드를 클릭했을 때 관련 1차 사료 문서·단락을 즉시 열람하도록 하는 인터랙티브 뷰어를 구상했다.

또한 LLM이 추출한 엔티티와 관계를 그대로 사용했을 때 발생할 수 있는 **노이즈·일관성 문제**를 줄이기 위해,
한국민족문화대백과사전의 공공 데이터에 기반한 역사 용어 사전을 먼저 구축하고,
이를 그래프 DB 구축 이전 단계의 필수 선행 자원으로 사용하도록 설계를 재구성했다.

2. 공공 데이터 기반 역사 용어 사전 설계

2.1 설계 방향과 역할

현재 프레임워크는 LLM을 통해 텍스트에서 직접 엔티티·관계 트리플을 추출하는 방식을 전제로 두고 있었다.
그러나 도메인 지식이 명시적으로 구조화된 용어 사전 계층이 존재하지 않을 경우,
엔티티 품질 관리, 명칭의 일관성, 관계 유형의 통일성 측면에서 한계가 나타날 가능성이 높았다.
이 문제를 해결하기 위해, 공공 데이터 기반 역사 용어 사전을 먼저 구축한 뒤, 이를 다음과 같은 역할로 활용하는 구조를 설계했다.

(1) 엔티티 후보 필터링 역할

- LLM이 추출한 엔티티 후보를 용어 사전에 존재하는 항목 중심으로 선별하도록 했다.
- 사전에 없는 엔티티는 "추가 검증 필요" 또는 "low-confidence 엔티티"로 태깅해,

이후 그래프 분석·시각화 단계에서 구분되도록 설계했다.

(2) 엔티티 정규화 역할

- 용어 사전에 수록된 표제어, 동의어, 이명(異名) 정보를 활용하여, "한산도 대첩", "한산도 해전", "한산 대첩"과 같은 다양한 표현을 하나의 canonical ID 아래로 정규화하도록 설계했다.

이를 통해 그래프 상에서 동일 개념이 여러 노드로 분산되는 현상을 줄이고자 했다.

(3) 카테고리 및 속성 부여 역할
- 사전에서 제공하는 분류 정보를 이용해 인물, 지명, 사건, 문헌, 제도, 날짜 등 카테고리를 부여했다.
- 단순 문자열 규칙으로 구현된 기존의 엔티티 분류 로직보다 정밀한 타입 지정이 가능하도록 했다.

(4) 그래프 시맨틱 강화 역할
- 각 용어의 기본 설명, 연대, 소속, 연관 항목을 이용하여, LLM이 추출한 트리플과 교차 검증을 수행할 수 있도록 했다.
- 더 나아가, 용어 사전만으로도 "이순신 −소속→ 조선 수군", "임진왜란 −연도→ 1592년"과 같은 기초 관계를 선행 구축하는 그래프 시드로 활용하는 방안도 함께 고려했다.

2.2 공공 데이터 수집 및 전처리 파이프라인
용어 사전은 오프라인 배치 파이프라인으로 구축하는 방향으로 설계했다.

(1) 데이터 소스 선정
- 한국민족문화대백과사전 같이 신뢰도 높은 공공 데이터를 주요 소스로 선정했다.

(2) 데이터 수집 단계
- 각 소스에서 다음 필드를 중심으로 데이터를 수집하도록 설계했다.
- 항목 제목(표제어)
- 분류(인물·지명·사건·제도·문헌 등)
- 짧은 요약/설명
- 연대 정보(시작/종료 연도, 연표 관련 정보)
- 관련 항목(연관 인물, 관련 사건, 관련 지명 등)
- 키워드 혹은 주제어

(3) 통합 전처리 단계
- 서로 다른 스키마를 가진 데이터를 단일 용어 사전 스키마로 매핑하기 위해,
키 이름을 통일하고 값 타입을 정규화하는 전처리 과정을 설계했다.
- 이 과정에서 중복 항목을 통합하고, 외부 식별자(ID)를 하나의 내부 term_id로 연결하는 규칙을 정의했다.

(4) 저장 방식
- 최초 버전에서는 terms/ 디렉터리 내에 JSON 파일 또는 경량 DB 형식으로 저장하는 방식을 채택했다.
- 이후 규모가 커질 경우를 대비해, 별도의 전용 그래프 DB 또는 키-값 스토어로 확장할 여지도 남겨 두었다.

2.3 용어 레코드 스키마
각 용어는 다음과 같은 스키마를 갖도록 설계했다.
- term_id: 고유 식별자 (예: "KMK_00012345")
- label: 대표 표제어 (예: "임진왜란")
- aliases: 다른 이름·표기(예: "임진란", "정유재란" 등)
- category: "사건", "인물", "지명", "문헌", "제도", "연도/기간" 등
- description: 짧은 설명 문단 (요약 수준의 서술)

- period: 연대 정보 (예: "1592-1598년", 또는 구조화된 {start_year, end_year} 형태)
- related_terms: 연관 용어의 term_id 리스트
- source: 데이터 출처(예: "한국민족문화대백과사전")

이 스키마를 기반으로, 이후 엔티티 정규화·카테고리 부여·그래프 시드 생성 작업이 일관되게 수행되도록 했다.

2.4 엔티티 추출 단계와의 결합 전략

역사 용어 사전은 지식 그래프 추출 파이프라인의 여러 지점에서 사용되도록 설계되었다.

(1) LLM 프롬프트 레벨에서의 활용

- 텍스트에서 트리플을 추출할 때 사용하는 LLM 프롬프트에
"다음은 이 문맥에서 특히 중요하게 고려해야 할 역사 용어 목록이다"라는 식으로 대표 용어 리스트를 힌트로 주는 방식을 도입했다.
- 예를 들어, 주요 역사 용어: 임진왜란, 한산도 대첩, 진주대첩, 선조, 이순신, ... 와 같이 제공하여, LLM이 의미 있는 엔티티 후보에 집중하도록 유도하는 구조를 설계했다.

(2) 후처리 단계에서의 엔티티 필터링

- LLM이 출력한 엔티티 후보에 대해, 기본 정제 규칙을 통과한 뒤에 용어 사전과의 매칭 단계를 추가했다.
- 이때 다음과 같은 매칭 전략을 사용하도록 설계했다.
- label과의 완전 일치
- 정규화된 label과의 일치
- aliases에 포함된 문자열과의 일치 혹은 유사도 기반 매칭
- 사전에 존재하지 않는 엔티티는 그래프에서 제외하거나 type='unknown', confidence='low' 등의 속성을 부여해 신뢰도 차등 표시를 하도록 했다.

(3) 엔티티 정규화·병합 과정과의 결합

- 엔티티 이름을 정규화하는 과정에서 용어 사전의 canonical label·alias 집합을 직접 활용하도록 설계했다.
- 같은 term_id에 매핑되는 엔티티들은, 그래프 상에서 하나의 노드로 병합되도록 병합 규칙을 강화했다.

(4) 용어 사전을 이용한 초기 그래프 시드 구축

- 용어 사전에 기록된 기본 관계(예: 인물-소속, 사건-연도, 사건-장소)를 별도의 배치 작업을 통해 초기 지식 그래프로 먼저 구축 하는 전략을 상정했다.
- 이후 1차 사료 텍스트에서 추출된 관계는 이 기초 그래프 위에 누적되는 형태로 설계하여, 공공 데이터 기반 지식과 1차 사료에서 관찰된 관계가 자연스럽게 통합되도록 했다.

3. 전체 파이프라인: 단계별 흐름

3.1 영상·오디오 전처리 파이프라인

(1) 영상 업로드 단계

사용자가 웹 또는 데스크톱 UI에서 영상 파일(mp4, mkv 등)을 업로드하도록 구성했다.
업로드된 파일은 세션 ID 또는 영상 ID 기준 디렉터리에 저장되도록 했다.

(2) Whisper 전사 단계

OpenAI Whisper ASR 모델을 사용해 영상에서 문장/발화 단위 텍스트와 타임스탬프를 함께 추출 하도록 설계했다.
- 영상 길이가 긴 경우, 일정 길이 단위로 오디오를 분할 저장하도록 설계하여, 이후 전사 및 그래프 연동 시 구간 단위 분석이 가능하도록 했다.
- 화자 분리가 가능한 경우, 화자 ID를 함께 기록했다.
이 단계에서 영상 타임라인과 텍스트가 1:1로 매핑되는 구조가 완성되었다.

3.2 전사 텍스트 정제 및 질의 구조화

(1) LLM 1차 정제 단계

Whisper 전사 결과를 LLM에 입력하여, 다음과 같은 정제 작업을 수행하도록 설계했다.
- 불필요한 반복, 추임새, 감탄사 제거
- 구어체 표현을 표준어 맞춤법에 가까운 서면체로 변환
- 지나치게 긴 문장을 적절한 길이로 분할
- 필요한 경우, 발화 앞에 화자 태그를 추가
- 각 발화에 대해 간단 요약과 키워드 리스트를 생성
이 과정을 통해 영상에서 얻은 텍스트가 사료 분석에 적합한 형태로 정제되도록 했다.

(2) 질의 재구성 및 키워드 기반 그래프 구성 단계

- 사용자가 질문을 입력한 경우,
전사 텍스트와의 임베딩/키워드 유사도를 이용해
질문을 더 명확하고 구체적인 형태로 재구성하는 과정을 설계했다.

- 사용자가 별도의 질문을 입력하지 않은 경우, 전사 텍스트 전체에서 중요 키워드들을 시간 순으로 도출하여, 시간의 흐름에 따라 그래프를 구성하는 모드를 설계했다.
이 단계에서 도출된 질문 또는 키워드 시퀀스가 이후 GraphRAG 검색과 지식 그래프 탐색의 입력으로 사용되도록 했다.

4. 문헌 데이터 로딩 및 정규화 (1차 사료 처리)

(1) 문헌 로딩 단계

1차 사료는 output/ 디렉터리 이하에 JSON 또는 PDF 형식으로 저장된다고 가정했다.
시스템은 이 디렉터리 구조를 재귀적으로 탐색하여,
각 사료별로 본문 텍스트와 메타데이터를 읽어오는 문서 로더를 사용하도록 설계했다.

(2) 메타데이터 정규화 단계

서로 다른 JSON 포맷과 키 이름을 공통 스키마로 통합하기 위해, 아래와 같은 메타데이터 구조를 기준으로 정규화 과정을 설계했다.
- 필수 필드:
- source (파일명)
- file_path
- index (문서 내 인덱스 또는 페이지 번호)
- type (예: json, pdf)
- collection (사료명)
- 선택 필드: title, author, category, topic, keywords

- 관리용 필드: raw_metadata_keys (원본 메타데이터의 키 목록), document_id (제목이 있으면 제목, 없으면 collection_filename_index 형태의 식별자)
주석(annotations)이 존재하는 경우, 이를 별도로 추출해 본문 뒤에 주석 블록을 붙이거나, 주석 자체를 독립 문서로 다루는 구조를 설계했다.
이러한 정규화 과정을 통해, 이후 그래프 구축·검색 시 사료·문헌 단위 필터링과 통계 분석이 가능해지도록 했다.

5. GraphRAG 인덱스 및 지식 그래프 구축

역사 용어 사전이 선행적으로 구축되었다는 가정하에, 다음 단계에서 1차 사료 텍스트를 이용한 GraphRAG 인덱스와 지식 그래프를 구축하도록 설계했다.

(1) 임베딩 및 벡터 인덱스 구축

- 한국어 및 다국어를 지원하는 임베딩 모델(intfloat/multilingual-e5-large-instruct)을 사용해 사료 텍스트 청크를 임베딩 공간에 매핑하는 구조를 설계했다.
- 텍스트는 일정 길이의 청크로 분할한 뒤, FAISS 기반 벡터 인덱스로 저장되도록 했다.
- 이를 통해 벡터 기반 의미 검색이 가능해지도록 했다.

(2) 지식 그래프 추출 단계

- LLM을 이용해 텍스트에서 "주어 | 술어 | 목적어" 형식의 트리플을 추출하는 파이프라인을 설계했다.
- 추출된 트리플은 다음과 같은 구조로 변환되도록 했다.
- 엔티티 레코드: 이름, 정규화 이름, 카테고리(인물·지명·사건 등), 출처 목록
- 관계 레코드: 출발 엔티티, 도착 엔티티, 관계 타입(술어), 근거 스니펫·문서 정보
이때, 앞서 구축한 역사 용어 사전을 사용해 엔티티 후보를 필터링하고, canonical label로 정규화하며, 사전 상의 카테고리·연대·연관 용어 정보를 그래프 속성으로 덧붙이는 구조를 도입했다.

(3) 배치 처리 및 자원 관리

- 다수의 문서를 LLM에 한 번에 보내는 배치 기반 트리플 추출 방식을 도입했다.
- 배치별 처리 시간과 GPU VRAM, 시스템 메모리, 스왑 사용량을 모니터링하여 안정적인 추출 작업이 이루어지도록 설계했다.
- 중복 엔티티와 관계는 엔티티 병합 규칙과 관계 시그니처를 통해 통합되도록 했다.

(4) 엔티티 벡터 인덱스 구축

- 그래프 DB에 저장된 모든 엔티티에 대해, 이름 + 타입 + 일부 출처 정보를 결합한 짧은 설명 텍스트를 생성했다.
- 이 텍스트를 임베딩하여 별도의 엔티티 벡터 인덱스를 구축하고, 엔티티 이름 기반 KNN 검색을 지원하도록 설계했다.
- 이를 통해 사용자 질의나 영상 키워드와 가까운 엔티티를 신속하게 찾는 기능을 구현할 수 있도록 했다.

6. 실시간 뷰어 및 인터랙티브 그래프 설계

사용자 요구 사항의 핵심은 다음과 같았다.
> "영상의 텍스트에서 키워드를 뽑고 그 키워드와 관련된 노드들을 실시간으로 보여주고,

> 사용자가 노드를 클릭하면 그 노드에 해당하는 문서를 보여준다."
이를 위해 기존 GraphRAG 구조 위에 UI·API 계층을 추가하는 방식으로 설계를 정리했다.

6.1 실시간 키워드 추출 및 노드 매핑

(1) 전사 텍스트 스트림 기반 키워드 추출

- LLM 1차 정제 단계에서 각 발화에 keywords를 미리 부여해 두었다.
- 영상 재생 시, 현재 재생 구간에 해당하는 발화들의 키워드를 모아 실시간 키워드 집합을 구성하는 방식을 설계했다.
- 필요 시, 최근 30초~1분 구간을 기준으로 키워드 빈도 및 가중치를 계산해 상위 N개의 키워드를 선택하는 실시간 랭킹 방식도 고려했다.

(2) 키워드 → 그래프 엔티티 매핑

- 추출된 키워드 리스트에 대해, 먼저 역사 용어 사전의 label/aliases와 매칭을 수행하고, 매칭되지 않는 경우 엔티티 벡터 인덱스를 활용해 근접 엔티티를 KNN 검색하도록 설계했다.
- 이 과정을 통해 현재 재생 구간과 관련된 엔티티 노드 집합을 도출하도록 했다.
- 각 노드에 대해서는 엔티티 타입, 요약 설명, 대표 출처 문헌 정보 등을 함께 조회하도록 했다.

(3) 노드 주변 이웃 탐색 및 시각화

- 각 중심 엔티티에 대해, 그래프 DB에서 깊이 1~2의 이웃 노드와 관계를 조회하는 과정을 추가했다.
- 조회된 이웃 정보를 기반으로, UI 상에서 작은 서브그래프 형태로 시각화하도록 설계했다.

6.2 노드 클릭 시 문헌/문서 연동

(1) 노드 → 출처 문헌 매핑

- 각 엔티티는 sources 필드를 통해 doc(파일명/경로), page 또는 index, snippet, collection 등의 정보를 보유하도록 설계했다.
- 사용자가 특정 노드를 클릭했을 때, 이 sources 정보를 기반으로 관련 1차 사료 문서/페이지 목록과 스니펫 프리뷰를 제공하도록 했다.

(2) 문헌 텍스트 상세 조회

- 사용자가 출처 리스트에서 하나를 선택하면, 해당 doc/page/index에 해당하는 원문 텍스트를 로딩하여 보여주는 UI를 구성했다.
- 동시에, 선택된 엔티티·키워드를 기반으로 추가 GraphRAG 검색을 수행해 인접한 다른 증거 문단도 함께 제시할 수 있도록 설계했다.

(3) 영상 타임라인과의 연동

- 전사된 텍스트를 그래프에 포함할 때, 메타데이터에 video_id, start_time, end_time을 함께 기록하도록 설계했다.
- 이를 바탕으로, 노드를 클릭했을 때 "해당 엔티티가 언급된 영상 구간으로 바로 점프"하는 기능을 구현할 수 있게 했다.

6.3 실시간 뷰어 UI 구성

(1) 좌측 패널: 영상 플레이어(재생/일시정지/타임라인)

(2) 우측 상단 패널: 지식 그래프 뷰어

- 현재 구간 키워드와 매핑된 중심 엔티티 노드
- 주변 이웃 엔티티 및 관계 엣지

- 노드 hover 시 간단 설명, 클릭 시 우측 하단 패널 갱신

(3) 우측 하단 패널: 문헌·답변 패널
- 선택된 노드에 대한 대표 설명(용어 사전/그래프에서 가져온 요약), 관련 1차 사료 스니펫 리스트
- 질문 시 GraphRAG generate_answer를 통해 생성한 해설 텍스트

7. GraphRAG 질의 및 결과 저장·활용

7.1 GraphRAG 질의 수행

질문 또는 사용자가 클릭한 특정 엔티티 이름을 기반으로, GraphRAG 질의를 수행하도록 설계했다.
- 질의가 들어오면,
- (1) 벡터 인덱스를 통해 상위 관련 문서 청크를 검색하고,
- (2) 엔티티 KNN 및 그래프 이웃 탐색을 통해 관련 엔티티·관계를 수집하고,
- (3) 이를 통합한 하이브리드 컨텍스트를 기반으로 LLM에게 답변 생성을 요청하도록 했다.

7.2 결과 저장 및 재사용

(1) 영상별 하이라이트/요약 생성
- 전사 요약 영상별 학습 노트/강의 노트 문서를 자동으로 생성하는 활용 방안을 설정했다.

(2) 후속 분석
- 어떤 엔티티·관계가 자주 조회되었는지,
어떤 사료가 증거로 가장 자주 인용되었는지를 통계로 집계해, 그래프 품질 개선, 용어 사전 확장, 프롬프트 튜닝 등에 재투입하는 피드백 루프를 구상했다.

8. 정리

이 프레임워크는 다음과 같은 특징을 가진 구조로 정리되었다.
- 멀티모달-지식그래프 통합 구조를 통해, 영상/오디오 전사와 1차 사료 텍스트를 하나의 GraphRAG 인프라 위에 올려 시청 경험과 사료 탐색 경험이 자연스럽게 연결되도록 했다.

- 공공 데이터 기반 역사 용어 사전 계층을 선행 구축함으로써, LLM 기반 엔티티·관계 추출의 노이즈를 줄이고, 역사 도메인 특화 지식을 체계적으로 반영하려 했다.

- 실시간 인터랙티브 뷰어 설계를 통해, 단순한 문서 목록이 아닌 지식 그래프 노드 단위의 상호 작용(노드 클릭 → 문헌/영상 구간/해설 연결)을 지원하는 방향으로 구조를 잡았다.

- 확장 가능한 GraphRAG 인프라를 전제로, 새로운 사료·영상이 추가될 때마다 인덱스를 점진적으로 확장하고, 인덱스 저장/로드를 통해 재계산 비용을 줄일 수 있도록 설계했다.