논문 2025-2-4 http://dx.doi.org/10.29056/jsav.2025.06.04

대규모 언어모델을 활용한 OTT 콘텐츠 불법 유통 탐지 및 증거 수집 자동화 방법

박병찬*, 이재청**, 김석윤*, 김영모*†

An Automated Method for Detecting and Collecting Evidence of Illegal OTT Content Distribution Using Large Language Model

Byeong-Chan Park*, Jae-Chung Lee**, Seok-Yoon Kim*, Young-Mo Kim*†

요 약

본 논문은 대규모 언어모델(LLM)과 대규모 액션모델(LAM)을 기반으로, OTT 콘텐츠 불법 유통 사이트로부터 증거를 자동으로 수집할 수 있는 탐지 및 수집 자동화 방법을 제안한다. 기존의 수작업 중심 웹 크롤링 방식은 구조 변경에 취약하고 유지보수 비용이 높은 한계가 있다. 이를 해결하기 위해 본 연구에서는 웹 구조 분석, 크롤링 코드 자동 생성, 유사도 판단, 증거 채증까지의 전 과정을 자동화하는 시스템 구조를 설계하였다.

제안된 방법은 LLM이 HTML 구조를 분석하여 크롤링 코드를 생성하고, LAM이 이를 실행하여 데이터를 수집한다. 또한 RAG 기반 정보 보강과 유사도 판단을 통해 의미 기반 비교를 수행하며, 판단 신뢰도가 낮을 경우 자동 보완 루프를 통해 반복적으로 개선된다. 이를 통해 다양한 웹사이트 구조에 유연하게 대응하고, 증거 수집의 신뢰성과 효율성을 높일 수 있는 기대 효과가 있다.

Abstract

This paper proposes an automated method for detecting and collecting evidence of illegal OTT content distribution using a Large Language Model (LLM) and a Large Action Model (LAM). Traditional manual web crawling methods are limited in their ability to adapt to structural changes in websites and incur high maintenance costs. To address these issues, we design a fully automated system architecture encompassing web structure analysis, code generation, similarity-based judgment, and evidence preservation.

In the proposed method, the LLM analyzes the HTML structure of a website and generates code for information extraction, while the LAM executes the code to collect data. Additionally, the system incorporates Retrieval-Augmented Generation (RAG) to enhance semantic similarity comparison and includes a feedback loop that iteratively improves the results when confidence is low. This approach allows the system to flexibly adapt to various web structures and enhances the reliability and efficiency of evidence collection.

한글키워드 : 대규모 언어모델, 대규모 액션모델, 불법 콘텐츠 유통, 웹 구조 분석, 증거 수집 자동화

keywords : Large Language Model(LLM), Large Action Model(LAM), Illegal Content Distribution, Web Structure Analysis, Automated Evidence Collection

* 숭실대학교 컴퓨터학과
** ㈜비욘드테크
† 교신저자: 김영모(email: ymkim828@ssu.ac.kr)

접수일자: 2025.05.07. 심사완료: 2025.06.11.
게재확정: 2025.06.20.

1. 서 론

OTT(Over-The-Top) 콘텐츠 시장의 급격한 성장에 따라 다양한 영상 콘텐츠가 디지털 플랫폼을 통해 광범위하게 유통되고 있다[1][2]. 그러나 이와 동시에 해당 콘텐츠가 저작권자의 동의 없이 복제·배포되는 불법 유통 문제도 심각해지고 있다[3]. 특히 웹사이트를 통한 불법 콘텐츠 유통은 형태가 다양하고 구조가 자주 변동되기 때문에, 기존의 수작업 중심 증거 수집 방식이나 정적 규칙 기반 탐지 시스템으로는 적절한 대응이 어렵다.

기존 대응 방식은 다음과 같은 한계를 지닌다. 첫째, 불법 유통 사이트의 HTML 구조나 게시글 템플릿 등 특정 정보를 사람이 직접 분석해 수집 스크립트를 작성해야 하므로, 사이트 구조가 변경될 때마다 반복적인 유지보수가 필요하다[4][5]. 둘째, 탐지 속도가 느리고 적시성이 떨어지며, 셋째, 자동화 수준이 낮아 대규모 불법 유통에 효과적으로 대응하기 어렵다[6][7][8].

이러한 한계를 극복하기 위해, 본 논문에서는 AI 기반의 지능형 웹 정보 처리 기술을 접목하여, 사람의 개입 없이도 변화무쌍한 불법 유통 사이트로부터 효율적이고 정확한 증거 수집이 가능하도록 하는 대규모 언어모델을 활용한 OTT 콘텐츠 불법 유통 탐지 및 증거 수집 자동화 방법을 제안한다.

본 방법은 대규모 언어모델(Large Language Model, LLM)을 중심으로 한 인공지능 기술을 활용하여, 웹사이트 구조를 이해하고 불법 유통 정보를 탐지하며, 자동으로 크롤링 코드 생성 및 실행을 통해 증거를 수집하는 전 과정을 자동화한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 기존 웹 크롤링 및 콘텐츠 탐지 방법 및 LLM/RAG 기반 정보 추출 방법을 기술한다. 3장에서는 본 논문에서 제안하는 대규모 언어모델을 활용한 OTT 콘텐츠 불법 유통 탐지 및 증거 수집 자동화 방법을 기술한다. 4장에서는 실험 및 결과를 살펴보고 5장에서는 결론으로 마무리한다.

2. 관련 연구

2.1 웹 크롤링 및 콘텐츠 탐지 기술

웹 크롤링(Web Crawling)은 인터넷 상의 특정 정보를 자동으로 수집하는 대표적인 기술로, 불법 유통 콘텐츠 탐지에도 널리 활용되어 왔다. 일반적으로 크롤러는 사이트의 HTML 구조, DOM 트리, XPath 등의 경로 정보를 기반으로 데이터를 추출하며, 정규 표현식이나 CSS 선택자 등을 활용하여 콘텐츠를 필터링한다[9][10].

그러나 이러한 방식은 웹사이트 구조가 변경되거나 의도적으로 우회되는 경우 쉽게 무력화된다. 또한, 다양한 불법 유통 사이트가 서로 다른 게시판 구조, 콘텐츠 템플릿, 동적 로딩 기법 등을 사용하는 경우, 각각에 대해 수작업으로 별도의 크롤러를 구성해야 하며, 유지보수 비용이 매우 높다. 이로 인해 규모 확장성, 적시성, 신뢰성의 측면에서 큰 한계를 가진다.

2.2 대규모 언어모델(LLM) 및 RAG 기반 정보 추출 기술

최근에는 대규모 언어모델(Large Language Model, LLM)을 활용한 자연어 이해 및 정보 추출 기술이 주목받고 있다. GPT, BERT 등의 LLM은 비정형 텍스트 데이터를 이해하고, 주어진 질의(Prompt)에 대해 유연하게 응답할 수 있는 능력을 갖추고 있어, 웹 페이지 상의 다양한 형태의 텍스트 및 구조적 정보를 처리하는 데 유리하다.

특히 RAG(Retrieval-Augmented Generation) 방식은 외부 지식 또는 벡터 검색을 통해 관련 정보를 보완하고, 이를 기반으로 보다 정교한 추론을 가능하게 한다. 이 방식은 단순한 문자열 매칭이나 패턴 인식 기반의 기존 탐지 방식에 비해, 의미 기반의 비교·판단이 가능하다는 점에서 우수하다[11][12].

LLM을 활용한 정보 추출 기술은 현재 다양한 산업 분야에서 활용되고 있으나, 이를 불법 콘텐츠 증거 수집 시스템에 적용하여 코드 생성, 실행, 판단까지 자동화한 사례는 거의 없다. 본 연구는 바로 이 지점에서 기존 연구와 명확히 차별화된다.

3. LLM을 활용한 불법 유통 탐지 및 증거 수집 자동화 방법

3.1 개요

본 논문에서는 제안하는 LLM을 활용한 불법 유통 탐지 및 증거 수집 자동화 방법은 대규모 언어모델과 대규모 액션 모델을 활용하여 OTT 콘텐츠의 불법 유통 웹사이트로부터 증거를 자동으로 수집하는 방법으로 그림 1과 같다.

웹사이트 HTML 데이터를 수집하고, 이를 바탕으로 LLM이 자동으로 크롤링 코드를 생성한 뒤, LAM이 해당 코드를 실행하고 결과를 평가

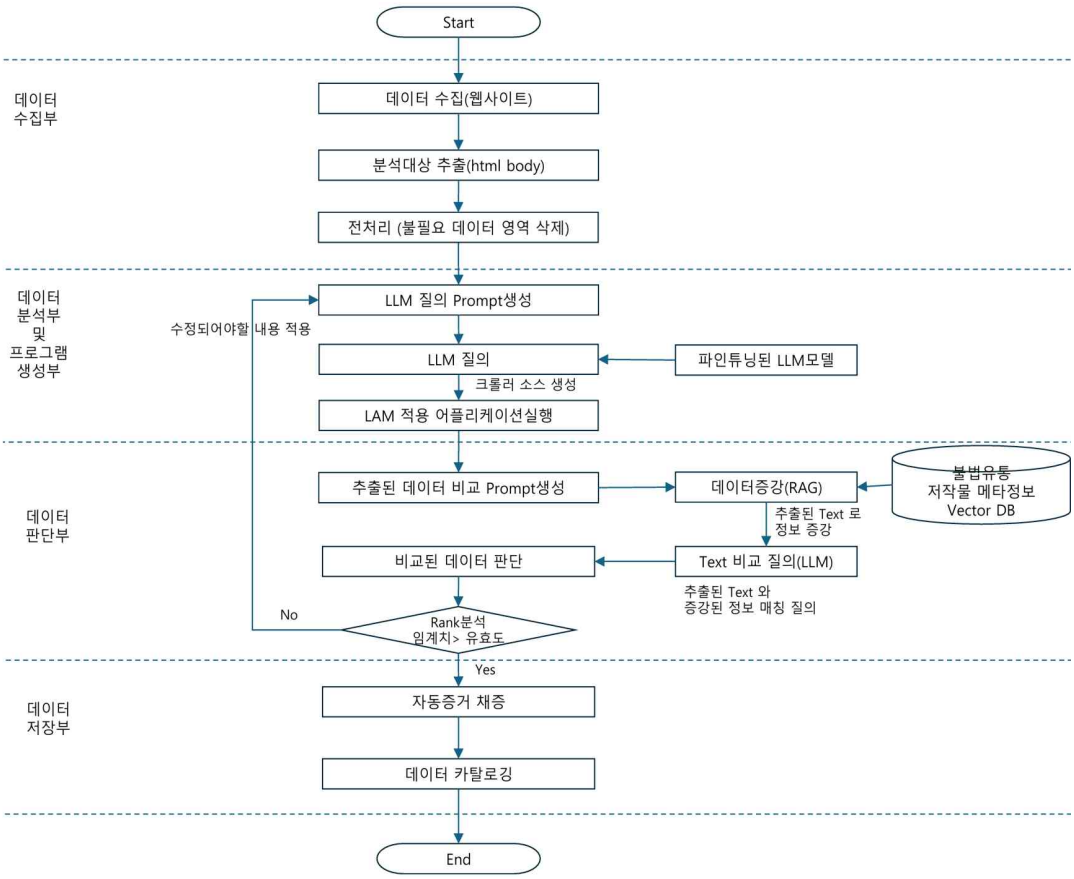

그림 1. OTT 콘텐츠 불법 유통 탐지 및 증거 수집 자동화 처리 흐름도

Fig. 1. Automated Workflow for Detecting and Collecting Evidence of Illegal OTT Content Distribution

하여 증거를 저장하는 자동화된 프로세스로 구성된다.

데이터 수집: 웹사이트 HTML 정보 수집 및 전처리

데이터 분석 및 프로그램 생성: LLM 기반 구조 분석 및 코드 생성

데이터 판단: 수집 정보와 기준 메타데이터 비교

데이터 저장: 증거 채증 및 데이터베이스 저장

반복 개선 루프: 정확도 임계치 미달 시 자동 보완

3.2 데이터 수집 단계

데이터 수집 단계에서는 불법 콘텐츠 유통이 의심되는 웹사이트에서 HTML 기반 정보를 수집한다. 대상 사이트의 게시글 구조를 중심으로 필요한 영역을 식별하고, 불필요한 광고나 이미지 등의 요소를 제거한다. 전처리를 통해 LLM 입력에 적합한 정제된 HTML 데이터를 생성한다. 스크립트, 스타일, 비정형 데이터는 제거되며, 주요 정보가 포함된 영역만 추출된다. 이 데이터는 이후 LLM이 코드 생성을 수행할 수 있도록 표준화된 형식으로 저장된다. 수집된 메타정보는 추후 증거 판단 및 분석 단계에서 활용된다.

3.3 데이터 분석 및 프로그램 생성 단계

전처리된 HTML 데이터를 기반으로 LLM은 웹페이지 구조를 이해하고, 필요한 정보를 추출하기 위한 크롤링 코드를 자동으로 생성한다. 게시글 목록, 제목, 작성시간, 링크 등의 위치를 추론하며, XPath, CSS Selector 등을 포함한 코드를 생성한다. 이후 LAM은 이 코드를 실행해 실제 데이터를 수집하며, 각 목적별 프롬프트 설계 예시는 표 1과 같다.

표 1. 프롬프트 설계 예시
Table 1. Prompt Design Examples

목적	프롬프트 예시
게시글 제목 추출	다음 HTML에서 게시글 제목의 XPath를 찾아 Python 코드로 추출해줘
제목+시간 정보 포함 추출	이 HTML에서 게시글 제목, 시간 정보를 포함해 리스트로 추출하는 크롤러 코드를 생성해줘
구조 이해 및 설명	아래 HTML 코드에서 게시글 목록 영역을 찾아 구조를 분석하고 ID나 XPath로 설명해줘

이러한 프롬프트 설계는 자동 코드 생성의 정확도를 높이며, 다양한 구조를 갖는 사이트에 유연하게 대응할 수 있게 한다.

3.4 데이터 판단 과정

데이터 판단 과정은 수집된 콘텐츠 데이터가 실제 불법 유통 콘텐츠에 해당하는지를 정량적·정성적으로 판별하는 단계이다. 이 단계는 단순한 문자열 일치 비교를 넘어, 대규모 언어모델(LLM)을 활용한 의미 기반 비교와 RAG(Retrieval-Augmented Generation) 방식의 지식 보강을 통해 높은 정확도를 확보한다.

비교는 다음 두 가지 입력 간에 수행되며 표 2와 같다.

표 2. 콘텐츠 유사도 판단 기준
Table 2. Content Similarity Evaluation Criteria

A (수집 데이터)	LAM을 통해 수집된 웹사이트 게시글 정보 (제목, 텍스트, 작성자, 게시 시간 등)
B (기준 정보)	사전 정의된 불법 유통 콘텐츠 메타정보 (예: 정품 콘텐츠 제목, 에피소드 번호, 시즌 정보 등)

두 정보는 벡터화(Vector Embedding)되어 의미 기반의 비교를 가능하게 하며, 벡터화는 보통 OpenAI, Hugging Face 등에서 제공하는 사전학습 모델 또는 커스터마이징된 LLM을 사용한다.

LLM 기반의 비교 질의는 다음과 같은 형식으로 구성된다면,

> 다음 게시글 제목이 다음 기준 제목과 유사한지 판단해줘. 만약 동일 콘텐츠일 가능성이 높다면 유사도 점수(0~1)를 반환해줘.

이와 함께 RAG(Retrieval-Augmented Generation) 방식으로 Vector DB에서 불법 콘텐츠 정보를 검색하고, 추출된 정보와 의미적 비교를 수행한다. 기준은 유사도(식(1)), 키워드 포함률(식(2)), XPath 일치율으로 판단한며, 표 3과 같다.

$$Similartiy(A,B) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\cdot\|\vec{B}\|} = \cos(\theta) \quad \text{식(1)}$$

$\vec{A}$: 추출된 웹 데이터의 벡터 표현

$\vec{B}$: 기준 콘텐츠 메타데이터 벡터

θ: 벡터 간 각도

$$Keyword\ Match\ Rate = \frac{Matched\ Keywords}{Total\ Keywords} \quad \text{식(2)}$$

표 3. 불법 콘텐츠 판별을 위한 정량 평가 기준
Table 3. Quantitative Evaluation Criteria for Illegal Content Detection

항목	기준값	설명
텍스트 유사도	≥ 0.80	기준 콘텐츠와 의미적 일치 여부 (Cosine Similarity)
키워드 포함률	≥ 60%	콘텐츠 제목 내 주요 키워드 포함 비율
XPath 일치율	≥ 70%	구조 기반 정보 추출 성공률

이 수치는 특히 콘텐츠 제목 내에 핵심 단어(작품명, 시즌, 화수 등)가 얼마나 포함되어 있는지를 파악하는 데 사용된다.

만약 유사도나 포함률이 기준 미달일 경우, 시스템은 다음의 보완 절차를 자동 수행한다.

1) Prompt 수정: 추출이 부정확했을 가능성에 대비하여 새로운 지시어를 재구성
2) 코드 재생성: LLM이 새로운 XPath/코드로 크롤러 코드를 재작성
3) 재수집 및 재판단 루프: 수정된 코드로 다시 수집 → 비교를 반복 수행

이 과정은 임계치를 만족할 때까지 반복되며, 불필요한 수작업 개입 없이도 높은 신뢰도의 판단이 가능하게 한다.

이에 대한 유효성 판단 예시 시나리오는 표 4와 같다.

표 4. 유효성 판단 예시
Table 4. results of evaluation

수집 제목 (A)	기준 제목 (B)	유사도	판단 결과
킹덤 시즌2 3화 다시보기 무료	킹덤 시즌2 에피소드 3	0.92	불법 유통
킹덤 전편 무료	킹덤 시즌1~2 전체	0.83	불법 유통
킹덤 리뷰 모음	킹덤 시즌2	0.41	무관

3.5 데이터 저장 단계

유효성이 확인된 콘텐츠 정보는 자동으로 증거로 채증되며, HTML 원문, 추출 텍스트, 이미지 캡처, 비교 결과 등을 포함한 증거 패키지가 생성된다. 이 데이터는 정형화된 메타데이터로 구성되어 Meta DB에 저장되며, 이후 법적 대응 자료로 활용할 수 있도록 구조화된다.

저장 항목으론 원본 HTML, 추출된 제목/링크/시간, 비교 결과 및 유사도 점수, 스크린샷 이미지, 수집 일시 및 URL 등이 될 수 있다.

이러한 데이터 저장 방식은 보고서 자동 생성, 증거 추적, API 기반 검색 시스템 구현 등으로 확장 가능하다.

4. LLM · LAM 기반 불법 콘텐츠 탐지 시스템 구성

4.1 개요

본 논문에서 제안하는 LLM을 활용한 불법 유통 탐지 및 증거 수집 자동화 방법을 제안하는 시스템인 LLM·LAM 기반 불법 콘텐츠 탐지 시스템은 크게 네 개의 처리 단계로 웹사이트의 정보를 수집하는 데이터 수집 단계, 구조를 이해하고 추출 코드를 생성하는 데이터 이해 및 추출 단계, 수집된 정보와 기준 정보를 비교하는 데이터 판단 단계, 그리고 증거를 채증하고 관리하는 데이터 카탈로깅 단계이며, 그림 2와 같다.

먼저 웹사이트의 HTML 구조는 크롤러를 통해 수집되며, 불필요한 광고나 이미지, 스크립트 등이 제거되어 전처리된다. 전처리된 HTML은 LLM의 입력으로 사용되며, 이를 통해 LLM은 웹페이지의 반복 구조나 패턴을 이해하고 필요한 정보를 추출할 수 있는 코드를 자동 생성한다. 해당 코드는 Python 기반의 크롤링 코드로 변환되며, LAM 모듈이 이를 실제로 실행한다.

코드 실행 결과로 수집된 데이터는 다시 LLM의 비교 모듈로 전달되며, 기준 메타정보와의 유사도를 분석하여 불법 유통 여부를 판단하게 된다. 판단 결과가 임계치 이상일 경우, 증거로 저장되며 메타 데이터베이스 및 이미지 저장소 등에 등록된다.

4.2 코드 자동 생성 및 실행 흐름

웹사이트 구조가 다양하고 수시로 변경되는 특성에 대응하기 위해, 본 시스템은 LLM 기반의 동적 코드 생성을 중심으로 구현되었으며, 그림 3과 같다.

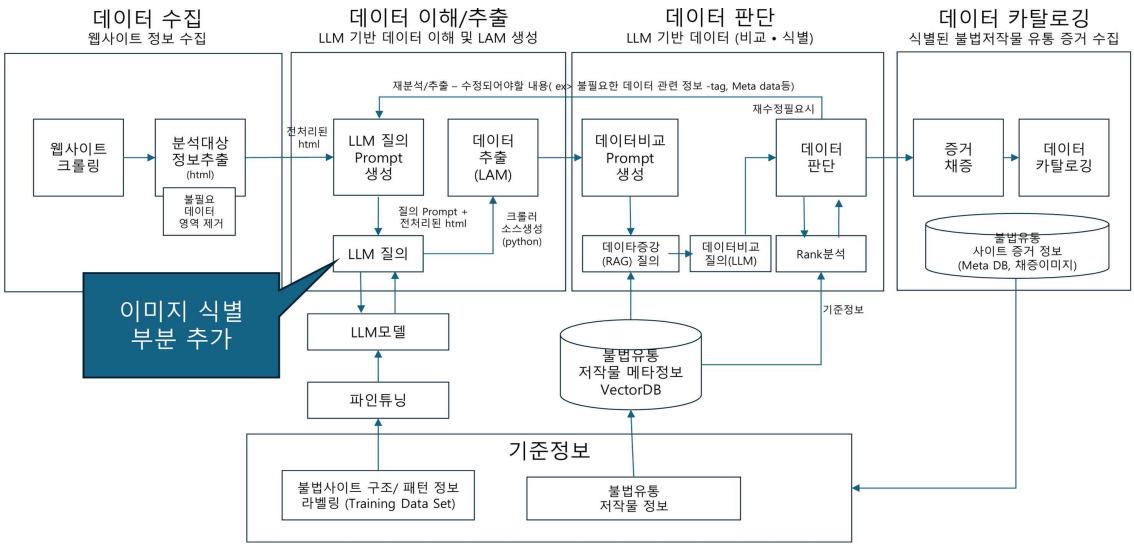

그림 2. 이미지 식별 기능이 포함된 불법 OTT 콘텐츠 탐지 및 증거 수집 시스템 아키텍처
Fig. 2. System Architecture for Illegal OTT Content Detection and Evidence Collection with Image Identification

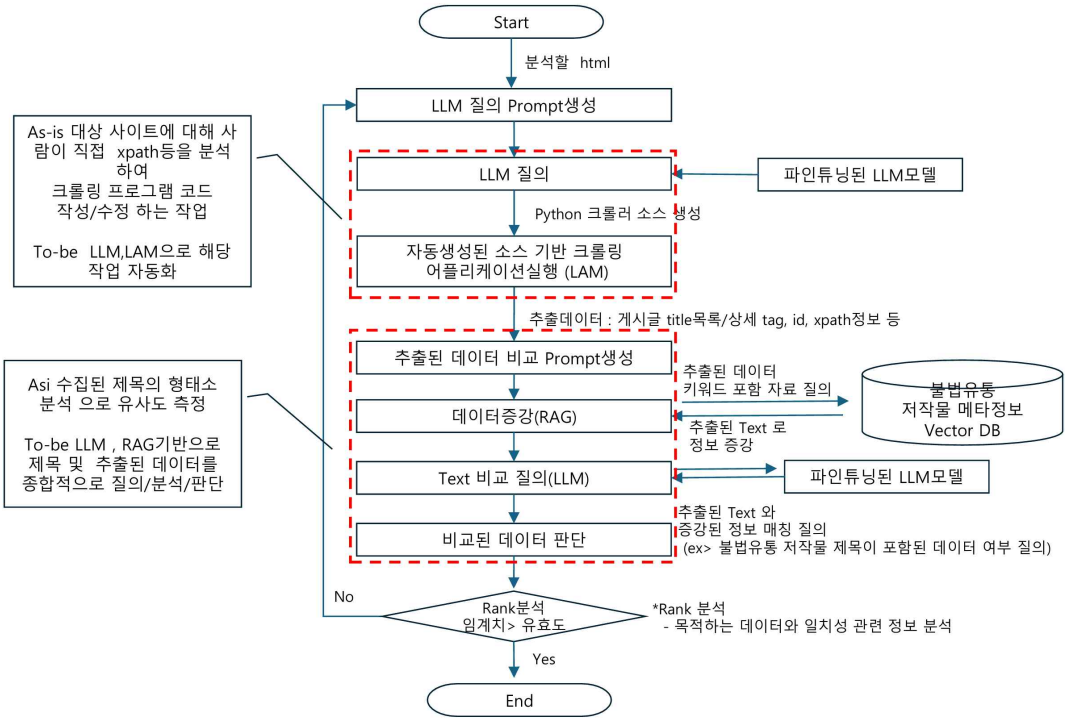

그림 3. 기존 수작업 방식 대비 LLM·LAM 기반 자동화 흐름 비교도
Fig. 3. Comparison Flow of Manual Process vs. LLM·LAM-Based Automation

전처리된 HTML을 기반으로 LLM은 질의(Prompt)를 통해 XPath 또는 CSS Selector를 추론하고, 이를 이용해 특정 게시글 영역을 추출하는 크롤링 코드를 자동으로 아래와 같이 작성한다.

> 이 HTML에서 게시글 제목과 작성 시간을 추출하는 코드를 만들어줘

위와 같은 형태의 Prompt를 통해 LLM은 Python 코드를 생성하며, HTML, CSS Selector, XPath 기반 예시 코드는 다음과 같다.

HTML 기반

```
<div class="post">
  <h2 class="title">게시글 제목입니다</h2>
  <span class="date">2025-06-17 14:23</span>
</div>
```

CSS Selector 기반

```
from bs4 import BeautifulSoup
html = ...
<div class="post">
  <h2 class="title">게시글 제목입니다</h2>
  <span class="date">2025-06-17 14:23</span>
</div>
...
soup = BeautifulSoup(html, 'html.parser')

# 게시글 제목과 작성 시간 추출
title = soup.select_one('.post .title').get_text(strip=True)
date = soup.select_one('.post .date').get_text(strip=True)

print("제목:", title)
print("작성 시간:", date)
```

XPath 기반

```
from lxml import html

html_str = '''
<div class="post">
  <h2 class="title">게시글 제목입니다</h2>
  <span class="date">2025-06-17 14:23</span>
</div>
'''

tree = html.fromstring(html_str)

# XPath를 통한 추출
title =
tree.xpath('//div[@class="post"]/h2[@class="title"]/text()
')[0]
date =
tree.xpath('//div[@class="post"]/span[@class="date"]/tex
t()')[0]

print("제목:", title)
print("작성 시간:", date)
```

이렇게 생성된 코드는 LAM을 통해 실행된다. LAM은 단순한 스크립트 실행 엔진이 아니라, 실행 오류 발생 시에도 자동으로 오류 위치를 식별하고 프롬프트를 수정해 재생성 요청까지 처리할 수 있도록 설계되었다. 수집된 데이터는 게시글 제목, 링크, 작성 시각 등 메타정보이며, 이후 판단 모듈로 전달된다.

4.3 학습 기반 구조 및 데이터셋 구성

시스템의 정확도를 향상시키기 위해, 불법 유통 사이트의 구조 패턴 정보를 학습 데이터로 구성하고 LLM 모델을 파인튜닝하였으며, 그림 4와 같다.

이를 위해 먼저 수많은 게시판 형태의 사이트로부터 HTML을 수집하고, 유효한 게시글 목록이 포함된 영역을 라벨링하여 학습용 데이터셋을 구성하였다.

전처리 과정에서는 HTML의 불필요한 요소를 제거하고, 유의미한 콘텐츠 블록만을 추출하였다. 이후, 각 HTML에 대해 입력 데이터(A: 전처리 HTML), 출력 데이터(B: 게시글 제목 목록), 보조 정보(C: 추출 XPath 등)로 구성된 다중 입출력 형식의 학습 데이터를 설계하였다. 이를 통해 LLM은 단순한 자연어 모델을 넘어, 구조 인식 및 위치 지정 기능까지 수행할 수 있도록 학습되었다.

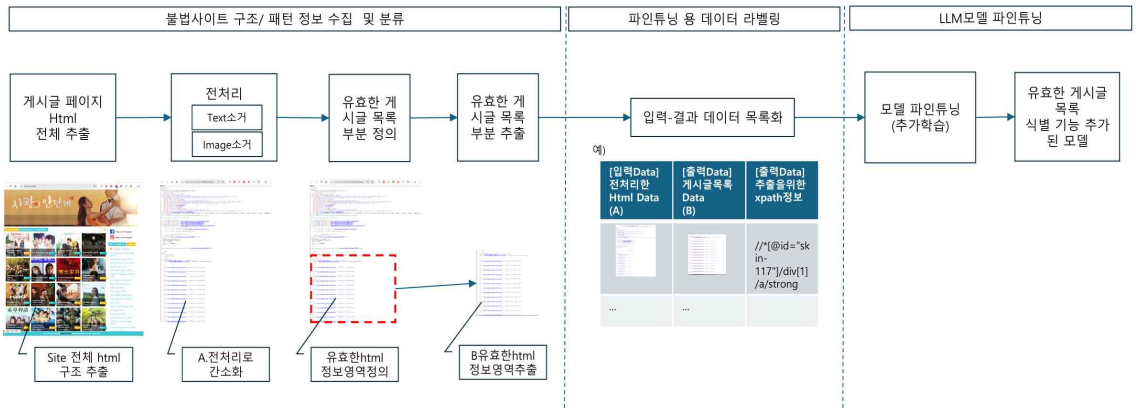

그림 4. 불법 유통 사이트 식별을 위한 학습 데이터 수집 및 LLM 모델 파인튜닝 과정
Fig. 4. Training Data Preparation and Fine-Tuning Workflow for Illegal Distribution Site Detection Using LLM

5. 결론

본 논문에서는 대규모 언어모델(LLM)과 액션 모델(LAM)을 기반으로, OTT 콘텐츠 불법 유통 탐지 및 증거 수집 과정을 자동화하는 시스템을 제안하였다. 기존의 수작업 기반 웹 크롤링 방식이 가진 구조 변화의 한계를 해결하기 위해, 본 논문에서 제안하는 방법은 HTML 구조 분석, 코드 자동 생성 및 실행, 유사도 판단, 반복 보완, 증거 채증의 전 과정을 자동화하였다.

기대 효과로 LLM을 활용한 웹 구조 분석 및 크롤링 코드 자동 생성, LAM을 통한 실행 자동화, RAG 기반의 의미 유사도 판단 기법 적용, 그리고 반복적 보완 루프 설계를 통해 다양한 웹사이트 구조에 유연하게 대응할 수 있다는 점이 있다. 특히, 실제 불법 콘텐츠 유통 사이트를 기반으로 한 시나리오에 적용 가능성이 높아, 구조 변화에 강한 탐지 시스템으로서의 활용이 기대된다.

한편 자바스크립트 기반의 동적 페이지나 이미지·영상 중심 콘텐츠의 탐지에는 아직 한계가 있으며, 향후 멀티모달 처리 기술과 추가 학습 데이터 확보를 통한 보완이 필요하다.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화기술 연구개발 사업으로 수행되었음(과제명 : OTT 콘텐츠 저작권 보호기술개발 및적용을 위한 저작권기술(+법)융합인재양성, 과제번호 : RS-2023-00225267)

참 고 문 헌

[1] Korea Communications Commission, “2024 Survey on Broadcasting Media Usage Behavior”, Broadcasting and Communications Policy Research Center, 2024. URL: https://www.mediin.or.kr/front/info/002/notice-detail.do?data_Id=2427

[2] Korea Information Society Development Institute (KISDI), “2023 Survey on the Distribution and Consumption of Internet Video Content”, KISDI Policy Research Report, 2023. URL: https://www.kisdi.re.kr/report/view.do?arrMasterId=3934581&artId=1777996

[3] Youngjoo Kim, “A Study on the Impact of OTT Service Expansion on Content Production, Distribution, and Consumption”, Journal of Broadcasting and Cultural Studies, 27(1), pp.75 - 102, 2015. DOI: https://doi.org/10.25024/kacs.2015.27.1.75

[4] Korea Internet & Security Agency (KISA), “Digital Threat Response and Illegal Content Response System”, 2024. URL: https://www.kisa.or.kr/10201

[5] Korea Creative Content Agency, “Analysis of Global OTT Trends”, KOCCA, 2023. URL: https://www.kocca.kr/globalOTT/vol01/document/all%20global%20OTT%20trend_230531.pdf

[6] G. D. Hong, H. K. Kim, “Sensor-based convergence system in Ubiquitous Environment”, Journal of Software Assessment and Valuation, 7(1), pp.1 - 6, 2017. DOI: 10.29056/jsav.2017.06.12

[7] S. Choi, Y. Lee, “Challenges in Detecting and Blocking Online Piracy in Dynamic Web Environments”, ACM Digital Threats: Research and Practice, 3(2), Article 12, 2021. DOI: https://doi.org/10.1145/3451217

[8] IFPI (International Federation of the Phonographic Industry), “Engaging with Piracy: A Global Overview of Online Infringement”, 2022. URL: https://www.ifpi.org/resources/

[9] ZenRows, “XPath for Web Scraping: Step-by-Step Tutorial for Beginners”, May

17, 2024. URL: https://www.zenrows.com/blog/xpath-web-scraping
[10] Scrapfly, “Parsing HTML with XPath”, August 2024. URL: https://scrapfly.io/blog/parsing-html-with-xpath/
[11] NVIDIA, “What is Retrieval-Augmented Generation (RAG)?”, 2024. URL: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/
[12] Scrapfly, “How to Power-Up LLMs with Web Scraping and RAG”, 2024. URL: https://scrapfly.io/blog/how-to-use-web-scaping-for-rag-applications/

저 자 소 개

박병찬(Byeong-Chan Park)

2015.2 학점은행제 졸업
2018.2 숭실대학교 컴퓨터학과 석사
2023.8 숭실대학교 컴퓨터학과 박사
2023.9-현재 숭실대학교 초빙교수
<주관심분야> 저작권 보호 및 이용활성화

이재청(Jae-Chung Lee)

1996.02 서울과학기술대학교 전자계산학과 학사
2012.07~현재 (주)비욘드테크 이사
<주관심분야> 저작권 보호 및 이용활성화

김석윤(Seok-Yoon Kim)

1980.2 서울대학교 전기전자 졸업
1990.2 University of Taxas at Austin Dept. of ECE 석사
1993.2 University of Taxas at Austin Dept. of ECE 박사
1982-1987 ETRI 연구원
1993-1995 모토로라 책임 연구원
1995-현재 : 숭실대학교 교수
<주관심분야> 저작권 보호 및 이용활성화

김영모(Young-Mo Kim)

2003.2 대전대학교 컴퓨터공학과 졸업
2005.2 대전대학교 컴퓨터공학과 석사
2011.2 대전대학교 컴퓨터공학과 박사
2012-현재 : 숭실대학교 교수
<주관심분야> 저작권 보호 및 이용활성화