Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.30, NO.6, Dec. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.6.1067

불법복제물 고속검색 및 Heavy Uploader 프로파일링 분석기술 연구*

황 찬 웅,[1†] 김 진 강,[2] 이 용 수,[2] 김 형 래,[2] 이 태 진[3‡]
[1,2,3]호서대학교 (대학원생, 학생, 교수)

High-Speed Search for Pirated Content and Research on Heavy Uploader Profiling Analysis Technology*

Chan-Woong Hwang,[1†] Jin-Gang Kim,[2] Yong-Soo Lee,[2] Hyeong-Rae Kim,[2] Tae-Jin Lee[3‡]
[1,2,3]Hoseo University (Graduate student, Student, Professor)

요 약

인터넷 기술의 발달함에 따라 많은 콘텐츠가 생산되고 그 수요가 증가하고 있다. 이에 따라 유통되고 있는 콘텐츠 수가 증가하였고, 반면에 저작권을 침해하는 불법복제물을 유포하는 건수도 증가하고 있다. 한국저작권보호원은 문자열 매칭 기반 불법복제물 추적관리시스템을 운영하고 있으며, 이를 우회하기 위해 다수의 노이즈를 삽입하므로 정확한 검색이 어려운 현실이다. 최근, 노이즈를 제거하기 위한 자연어 처리, AI 딥러닝 기술을 이용한 연구와 저작권 보호를 위한 다양한 블록체인 기술이 연구되어 있으나 한계가 있다. 본 논문에서는 온라인에서 수집한 데이터에 노이즈를 제거하고, 키워드 기반 불법복제물을 검색한다. 또한, heavy uploader 대상 프로파일링 분석을 통해 동일 heavy uploader를 추정해 간다. 향후, 불법복제물 검색기술과 heavy uploader 대상 프로파일링 분석 결과를 바탕으로 차단 및 대응기술이 결합하면 저작권 피해를 최소화할 것으로 기대한다.

ABSTRACT

With the development of internet technology, a lot of content is produced, and the demand for it is increasing. Accordingly, the number of contents in circulation is increasing, while the number of distributing illegal copies that infringe on copyright is also increasing. The Korea Copyright Protection Agency operates a illegal content obstruction program based on substring matching, and it is difficult to accurately search because a large number of noises are inserted to bypass this. Recently, researches using natural language processing and AI deep learning technologies to remove noise and various blockchain technologies for copyright protection are being studied, but there are limitations. In this paper, noise is removed from data collected online, and keyword-based illegal copies are searched. In addition, the same heavy uploader is estimated through profiling analysis for heavy uploaders. In the future, it is expected that copyright damage will be minimized if the illegal copy search technology and blocking and response technology are combined based on the results of profiling analysis for heavy uploaders.

Keywords: Copyright protection technology, Pirated contents search, heavy uploader profiling

Received(10. 13. 2020), Accepted(10. 22. 2020)
* 본 연구는 문화체육관광부 및 한국저작권위원회의 2020년도 저작권기술개발사업의 연구결과로 수행되었음.(No.2019-PF-9500)
† 주저자, hcw85123@gmail.com
‡ 교신저자, kinjecs0@gmail.com(Corresponding author)

I. 서 론

기술과 문화의 융합으로 콘텐츠 부가가치를 창출하고 있고, 세계와 소통하며 가치를 인정받고 있다. 최근에는 1인 미디어 사용과 모바일 영상 이용이 급증하면서 저작권 침해 우려 또한 확대되고 있다. 그러나 저작권 보호는 법에 중점을 두고 있으며, 저작권 보호를 위한 기술적인 OSP(Online Service Provider) 필터링 시스템[1]은 저작권 침해물을 완전히 차단하지 못하고 있다는 점을 문제점으로 인식된 데 따른 것으로 보고 있다[2-3].

불법복제물 유통환경은 오프라인인 CD/DVD에서 온라인인 웹하드, P2P, 토렌트의 다운로드 방식으로 변화했고, 현재는 페이스북과 유튜브 등을 통한 스트리밍 방식으로 옮겨가고 있지만, 2011년 웹하드 등록제 시행 이후 PC 시장규모는 줄었으나 모바일 환경 증가로 여전히 2019년 12월 말 기준 '특수한 유형의 부가통신 등록사업자 현황'에 따르면[4], 웹하드(P2P, 모바일환경 포함) 업체는 PC 서비스 기준으로 41개가 운영되고 있으며, 총 90개(모바일환경 포함) 사이트에서 불법복제물이 유통되고 있다. 2019년 한 해 동안의 콘텐츠 분야별 불법복제물 이용현황을 살펴보면, 불법복제물 이용 비율이 높게 나타난 분야는 방송 43.0%, 음악 41.6%, 영화 41.5% 큰 차이는 없으며, 음악영화 제작사들도 이러한 불법복제물 중심으로 법적 대응에 나서고 있지만, 사후대응을 이루고 있어서 저작권을 보호할 수 있는 기술 연구가 필요하다.

본 논문에서는 기존 필터링 기반의 저작권 보호 기술을 우회하는 웹하드, P2P 사이트에서의 데이터 유형을 분석하여, 변형된 게시물 제목을 정규화과정을 통해 노이즈를 제거하고, 키워드(keyword) 기반 블룸필터(bloom filter) 검색기술을 통해 불법복제물 추적관리시스템(illegal content obstruction program)의 성능을 개선을 위한 방법론을 제시하고 검증한다. 또한, 불법복제물을 다량으로 유포하는 동일 heavy uploader 프로파일링을 위한 OSP/게시자(ID)별 유포저작물 전반에 대한 정보가 담기도록 feature engineering을 통해 유사한 feature set을 생성하고, 클러스터링 기반 동일인으로 추정되는 heavy uploader 분석기술을 제안한다. 현재 불법복제물에 대해 노이즈 제거 및 다양한 검색기술이 적용되고 있음에도 불구하고, 이를 우회하고 정확한 검색을 회피하는 불법복제물 탐지를 개선하고, 주요 heavy uploader 분석을 통해 저작권 관련 피해를 최소화할 것이다. 또한, 불법복제물 뿐만 아니라 다양한 텍스트 검색 환경에서 응용할 수 있다.

2장에서는 저작권 보호 기술 배경과 한국저작권보호원의 불법복제물 추적관리시스템을 설명하고, 3장에서는 불법복제물 탐지를 위한 텍스트 정규화 및 검색기술과 heavy uploader 프로파일링 분석을 제안한다. 4장에서는 실험 결과와 5장에서는 결론으로 끝을 맺는다.

II. 관련 연구

2.1 저작권 보호 기술 배경

1990년대 등장한 대표적인 디지털 저작권 보호 기술은 디지털 워터마킹(digital watermarking)과 DRM(Digital Rights Management) 기술이다[5-6]. 그러나 보호해야 하는 저작물이 증가하고, 저작물이 아날로그 환경에서는 한계가 있다.

최근 저작권 보호를 위해 블록체인(blockchain) 기술을 결합한 연구가 진행되고 있다[7-8]. 블록체인과 네트워크를 연결하여 인가된 사용자 간의 콘텐츠 거래가 가능하며 등록된 콘텐츠는 체인 링크를 통해 주기적으로 블록이 생성되어 블록체인을 활용한 저작권 보호에 관한 연구가 진행 중이다[9]. 블록체인 기술은 놀라운 장점이 있지만 당장 현실적으로 적용하기에는 몇 가지 문제가 발생한다. 첫째, 블록체인은 이미 블록에 기록된 거래내역을 바꿀 수 없지만, 기록되기 전의 기록 대상인 저작물의 진위에 대하여는 위변조 여부를 확인할 수 없다. 둘째, 영상과 같은 대용량 저작물의 경우 그 데이터 용량이 블록에 담을 수 없는 정도이므로 결국 거래내역과 내용만 블록에 기록하고 실제 영상 데이터는 중앙화된 서버를 이용할 수밖에 없다는 기술적 한계도 존재한다. 따라서, 저작권 보호 기술이 끊임없이 연구되고 있음에도 불구하고 현재까지 사용하고 있는 기술은 필터링(filtering) 기술이다[10-13].

필터링 기술은 검색어 기반 필터링, 해시 기반 필터링, 특징기반 필터링으로 분류된다. 저작물을 식별할 수 있는 특징 데이터베이스를 확보함으로써 웹하드와 같은 특수 유형의 온라인서비스제공자에게서 저작물의 불법유통을 검색하여 차단하는 데 유용하게 활용할 수 있다[14]. 특징 데이터베이스는 다양한 방법으로 구축할 수 있으며 합법적인 사용자에게 이

Table 1. Collection Title Noise Type Analysis

Noise Type	Collection Title	Original Title
Special Characters	〔인기 애니〕 디즈니 /모/아/나 - 고화질(BPRip), 우리말 더빙	모아나
Letter	유쭌상X문종원--〔두얼굴의인간ㅅr냥-성난 호r가〕FHD초고화질	성난화가
	ㅅr기꾼 도둑 폭ㅍr 프로들의 ㄱ|상천외한 작전 그들의 통쾌한 반격〔ㅇ|딸ㄹ|안짭〕 한글	이탈리안 잡
	--전미 박스오피스 1위 〔크.ㄹㅐ.ㅇ ㅣ.ㅈㅣ. 리치 아시안〕	크레이지 리치 아시안
Actor	〔 톰 행크스 실화 -(해.적.소.탕.작.전)-1080P 한글자막	캡틴필립스
Keyword	2O2O.O8 신작꿀잼x평점굿((----타임루프:시간무한반복----))완벽자막	팜 스프링스
	08월 무더위쿨한액션 전직 미 국.정.원〔-비밀요원-〕완벽자막 고화질	가짜 암살자의 진짜 회고록
Delete	08월 〔개봉예정〕신비롭고 아름다운 정원(씨 크 릿) 자체자막	시크릿가든
	2020. 디즈니 대작 떳다 (종. 족. 전. 쟁)ㅇㅏ르테ㅁㅣ스	아르테미스 파울
Director	2020 마 이 클 베 이 〔바이러스공격의 최 후 의 보류〕 완벽한글자막	더 라스트 쉽
English mix	〔공포.스릴러〕 키링타임용-- 〔The 랜타알〕-자체자막.고화질	더 렌탈
Pronunciation	상상 경계가 무너진다 〔시 작 인 샙 송〕 고화질 완벽자막	인셉션
	02월.〔一 쮸. 만. 지. 3 - 넥스트 래벨 一〕 뜨웨인 존쓴. 초고화질	쥬만지3: 넥스트레벨

용 편의성을 극대화 시킬 방법으로 매우 효과적이다.

2.2 불법복제물 추적관리시스템(ICOP)

한국저작권보호원의 불법복제물 추적관리시스템은 온라인 불법복제물 모니터링 정보와 긴급대응저작물과의 검색기술을 결합하여 불법복제물을 차단한다. 불법복제물 추적관리 시스템의 운영 프로세스는 Fig. 1.과 같다.

웹하드 등 온라인 사이트에서 수집한 게시물 등의 정보와 긴급대응 저작물, 중점 보호 저작물 데이터베이스와의 문자열 매칭으로 동작하기 때문에 시스템 성능이 수집 게시물 정보의 노이즈를 제거하는 것과 검색기술에 의존하고 있다[15]. 불법복제물 게시자

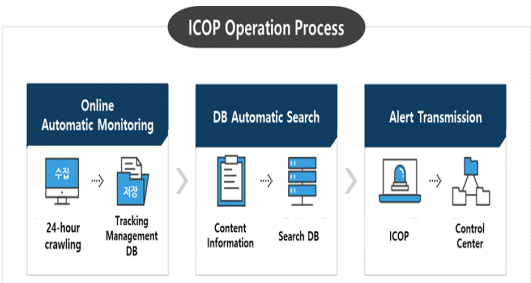


Fig. 1. Illegal Content Obstruction Program

들은 ICOP를 우회하기 위해 형태소 파괴, 특수문자 삽입 등 텍스트 제목을 변형하는 다양한 노이즈를 삽입한다[16]. 주요 상용 웹하드, P2P 사이트에서 수집된 저작물 제목을 분석한 노이즈 유형은 Table. 1.처럼 다양하다. 특수문자 삽입은 제거하면 되지만 문자 변형, 영문 혼합, 발음 변형은 대부분 한글 자음과 모음, 영문, 숫자 등 다양한 문자를 활용하여 변형하기 때문에 영문과 숫자를 한글 자음과 모음으로 변형하면 노이즈가 없는 영문과 숫자도 변형되는 이슈가 있고, 제거하면 실제 저작물 제목이 손상된다. 출연배우 언급, 주요 키워드 제공, 삭제, 감독명 언급 등 모든 수집 저작물 제목에 실제 저작물 제목을 포함하지 않는 경우도 존재하여 정확한 검색이 이루어지지 않고 있다. 또한, 검색 데이터베이스의 양이 많으면 속도가 저하되는 단점이 존재하고, 문자열 매칭방식으로 동작하기 때문에 노이즈가 많아 정확한 저작물 제목을 선별하기가 쉽지 않다. 따라서, 필터링 검색기술에 대한 성능개선이 필요한 이유이다.

III. 제안 모델

3.1 전체 시스템 구성

저작권 보호를 위해 상용 환경에서 공유되고 있는

콘텐츠가 합법적으로 제작되었는지 불법적으로 복제되었는지 분석하기 위해서 웹하드와 P2P 사이트에서 콘텐츠 정보를 수집하고, 저작권이 있는 공공데이터와 얼마나 유사한지 검색 결과에 따라 불법복제물로 분류한다. 본 논문에서 제안하는 방법은 웹하드와 P2P 사이트에서 크롤링(Cwaling)을 이용하여 콘텐츠 정보를 수집한다. 수집된 정보는 텍스트 형식의 다수의 노이즈가 존재하는 데이터이기 때문에 정확한 검색을 위해서 패턴(pattern) 기반의 텍스트를 정규화하는 전처리 과정을 거친다. 정규화된 텍스트 제목에서 키워드를 추출하고, 키워드가 포함된 고유한(unique) hash bit-array를 생성하여, Bloom filter 검색기술을 통해 저작권이 있는 데이터베이스를 검색한다. 유사한 콘텐츠가 검색될 경우 이를 불법복제물로 분류한다. 또한, heavy uploader가 다양한 OSP와 다수의 ID를 사용하여 불법복제물을 유포하기 때문에 heavy uploader 프로파일링 분석을 위한 불법복제물과 유포저작물 전반적인 특징이 담긴 OSP/게시자(ID)별 유사한 벡터(vector)를 가지도록 feature engineering 과정을 통해 동일 인물로 추정되는 heavy uploader를 선별하고, OSP/게시자별 프로파일링을 통해 유포 관계를 분석한다. 전체 시스템 구성은 Fig. 2.와 같다.

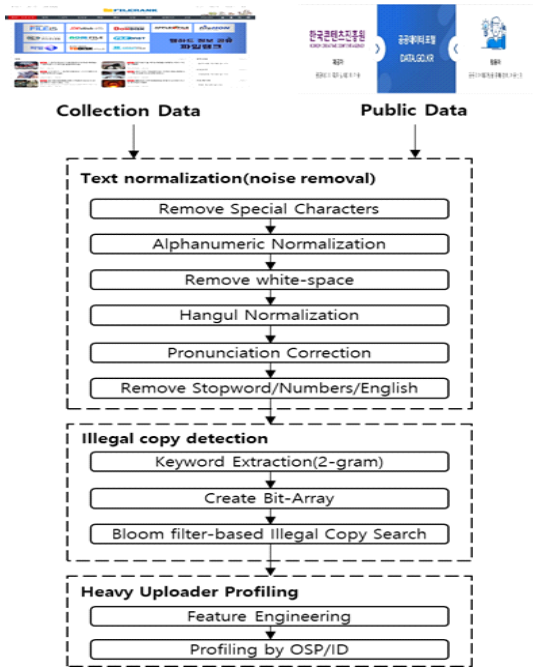


Fig. 2. System Configuration

3.2 불법복제물 탐지

3.2.1 텍스트 정규화

불법복제물 게시자들은 문자열 매칭 방식의 불법복제물 추적관리시스템을 우회하기 위해 해당 콘텐츠 제목에 노이즈를 삽입한다. 저작권 보호를 위한 기술적 조치로 검색어 기반 필터링 기술의 성능은 불필요한 문자나 변형된 문자와 같은 노이즈를 얼마나 제거하는 것에 좌우한다. 최근 자연어 처리, AI 딥러닝 기술을 이용한 네이버(Naver) 검색어 오타변환 기술은 단어 단위로 오타를 인식해 동작하고 Table 1.과 같은 모음이 변형된 데이터는 인식할 수 없다. 예를 들어, Naver 검색어 오타변환 기술은 '떠미네이터'를 '터미네이터'로 정규화하나 '신ㅂl ㅇrㅍr트'는 인식할 수 없다. 따라서, 본 논문에서 제안하는 불법복제물 추적관리시스템의 성능을 개선하기 위한 텍스트 정규화 과정은 Fig. 3.와 같이 여러 단계로 동작한다.

정규화 과정은 총 4개의 패턴(pattern)으로 변형

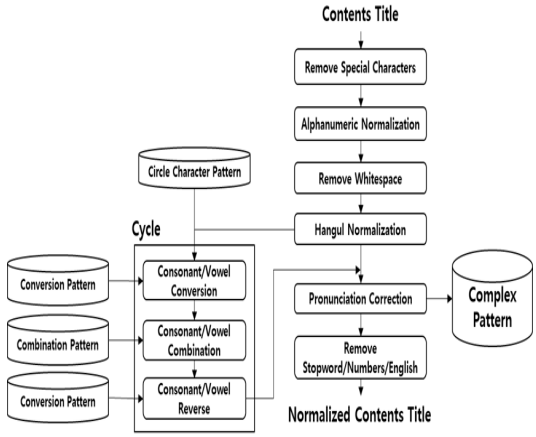


Fig. 3. Text Normalization Process

된 문자나 단어에 대응하는 알맞은 문자나 단어로 변환시켜 정규화한다. 특수문자를 제거하고, 조건에 맞는 영숫자(alphanumeric)를 정규화한다. 영숫자 정규화는 숫자 표현을 영문자로 변형한 경우에 변형된 영문자를 숫자로 변형한다. 수집된 제목에는 실제 제목뿐만 아니라 '2O2O', '1O8OP'처럼 날짜와 가격 등 다양한 숫자를 포함하기 때문에 영문자를 기준으로 앞, 뒤 문자가 숫자이면 해당 영문자를 숫자로 변경하는 것이 영숫자 정규화이다.

한글 정규화 과정은 자음과 모음의 변형 및 분리되어있는 것을 알맞은 자음과 모음으로 변환하고 결합한다. 수집된 제목에는 'ⓐ', '①' 등 원문자를 삽입하여 영문자와 숫자를 표현한 것이 존재하여 원문자 패턴으로 대응하는 'a', '1'로 변환한다. 또한, 'O'는 'ㅇ'으로 'i'는 'ㅣ'로 변환하는 자모 변환 패턴과 'ㅇㅏ'를 '아'로 결합하는 패턴을 사용하여 한글을 정규화한다. 수집된 제목은 한글뿐만 아니라 영어단어도 존재하기 때문에 'O'는 'ㅇ'으로 'i'는 'ㅣ'로 변환했으나 결합 패턴에 의해 자모가 결합하지 않는 유형은 영문자일 가능성이 있어서 자모 변환 패턴을 다시 역으로 'o'는 'O'로 'ㅣ'는 'i'로 역 변환한다.

'인샙숑'처럼 발음을 변형시킨 경우나 'The 랜타알'처럼 영문을 혼합한 경우 지금까지의 패턴으로는 정규화가 불가능하다. 따라서, 최신 유형의 패턴을 추가해야 하는 단점이 있지만 불법복제물을 탐지하기 위해 패턴 관리가 필요한 복합 패턴으로 '인셉션', '더 렌탈'로 정규화한다.

마지막으로 불용어, 숫자, 영문을 제거한다. 수집된 제목은 '자막', '초고화질', '더빙판' 등 다양한 불용어가 존재하고, 정규화된 날짜를 표현하는 '2020', '03월' 등 숫자가 존재하고, 검색에 필요 없는 영어 단어와 같은 이들을 모두 제거한다.

3.2.2 키워드 추출 및 비트 배열(bit-array) 생성

웹하드 및 P2P에서 수집된 제목은 실제 제목만 담고 있지 않으며 다양한 노이즈가 포함되어 있어 패턴 기반의 텍스트 정규화를 통해 노이즈를 제거한다. 정규화된 제목에서 키워드를 추출하기 위해 2-gram 언어 모델링 기반 단어 길이가 2로 분리한다. 예를 들어 '내안의그놈'의 주요 키워드는 '내안', '안의', '의그', '그놈'으로 총 4개의 키워드로 분리한다. 분리된 키워드를 해시(hash) 함수를 이용하여 콘텐츠별 고정된 bit-array를 얻는다. 해시 함수는 SAH256을 사용한다. Fig. 4.은 콘텐츠별 bit-array를 생성하는 의사 코드(pseudo code)이다. 콘텐츠별 각각의 키워드에 대한 해시값을 생성하고 고정된 크기로 모듈러(modular) 연산을 통해 해당 인덱스(index)의 비트(bit)를 1로 설정한다. 고정 크기가 클수록 bit가 중복되지 않아 주요 키워드별 고유함을 유지하기 위해 고정 크기를 100,000으로 설정한다. 따라서, 콘텐츠별 고유한 100,000개의 bit로 구성되어 있다.

Function1 - Bit-array generation function with keywords as elements

Description. This function converts to a fixed size bit-array for each content.

SET S : Fixed Bit-array Size
SET keywords : Array of Extracted Keywords
SET HASH : SHA-256 HASH FUNCTION

```
1. BitArray Vector[S]=0
2. For keyword in Keywords :
3.     Index=HASH(keyword) Mod S
4.     Vector[Index]=1     # bit setting
5. Return Vector
```

Fig. 4. Bit-array Generation Pseudo-code

3.2.3 Bloom Filter 기반 불법복제물 검색

공공데이터의 bit-array의 bit가 수집 데이터 bit-array의 bit에 일치하는지 검색 과정에 Bloom filter 알고리즘을 사용하여 검색 효율을 증가시켰다. Bloom filter는 원소가 집합에 속하는지 아닌지를 판별하는 확률적 자료구조이다[17-18]. Bloom filter 특성상 어떤 원소가 집합에 속한다고 판단된 경우 실제로는 원소가 집합에 속하지 않는 긍정 오류(false positive)가 발생하는 것이 가능하지만, 반대로 원소가 집합에 속하지 않는 것으로 판단되었는데 실제로는 원소가 집합에 속하는 부정 오류(false negative)는 절대로 발생하지 않는다는 특성을 이용하여 고속검색이 가능하다. Bloom filter가 동작하는 방법은 Fig. 5.와 같다.

수집 데이터의 bit-array는 원소의 집합에 해당하고, 하나의 bit는 원소에 해당한다. 그러나, 웹하드, P2P 사이트에서 수집한 콘텐츠 제목에는 다양한 노이즈가 존재하기 때문에 키워드가 간결하지 않

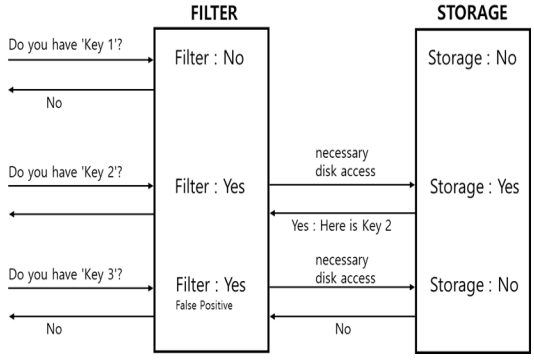


Fig. 5. Bloom Filter Operation Process

아 탐지가 어려우므로 공공데이터의 bit가 수집 데이터 bit에 포함하는지 검색한다. 또한, 공공데이터 콘텐츠 제목의 길이가 짧을수록 키워드 개수가 적어 오검출이 발생할 확률이 증가하고, Bloom filter 방식을 그대로 사용하면 공공데이터의 bit가 수집 데이터 bit에 모두 존재해야 하므로 공공데이터에 존재하나 검색되지 않는 불법복제물이 존재한다. 본 논문에서는 공공데이터 콘텐츠 제목의 길이에 따라 특정 임계치(threshold)를 다르게 설정한다. 공공데이터 콘텐츠 제목의 길이와 두 콘텐츠의 bit가 일치 여부에 따라 유사도(similarity)를 계산하고, 유사도가 특정 threshold 이상일 경우 불법 복제물로 판단한다. 일치하는 bit 개수에 따른 유사도 계산 식은 다음과 같다.

$$A : Collected\ \ Data \tag{1}$$

$$B : Copyrighted\,Public\,Data \tag{2}$$

$$\sim ilarity = \frac{Count(B\&A)}{Count(B)} \tag{3}$$

식1은 웹하드, P2P에서 수집한 데이터를 의미하고, 식2는 저작권이 있는 공공데이터를 의미한다. 식3의 유사도 공식은 B의 bit가 1일 때 같은 인덱스에 A의 bit가 1인 개수를 B의 bit가 1인 개수로 나눈 것을 의미한다. 다시 말해, A와 B의 같은 인덱스에 bit가 1인 개수를 B의 키워드 수로 나눈 것을 의미한다. 공공데이터 제목의 길이가 짧을수록 유사도가 높게 측정되기 때문에 공공데이터의 키워드 개수에 따라 threshold를 다르게 설정하여 불법복제물을 탐지한다. 키워드 수가 작으면 threshold를 높게 설정하고, 클수록 낮게 설정한다. 키워드 수가 1, 2개일 경우 threshold는 1로 설정하고, 3개일 경우 0.33으로 설정하고, 4, 5개일 경우 0.5로 설정하고, 6, 7개일 경우 0.6으로 설정한다. 8개 이상부터 threshold를 0.7로 고정한다. 또한, 두 콘텐츠 사이의 유사도가 특정 threshold보다 높아 검색되었지만, 유사도에 따라 저작권 보호를 위한 k개의 우선순위를 정할 수 있다. 따라서, 공공데이터 제목에 대해 유사한 웹하드, P2P에서 수집한 콘텐츠의 제목을 고속검색하여 불법복제물 탐지가 가능하다.

3.3 Heavy Uploader 프로파일링

3.3.1 Feature Engineering

불법복제물을 다량으로 유포하는 heavy uploader는 동일 불법복제물을 약간의 제목만 변경하여 여러 OSP를 통해 유포하고, 다수의 전혀 다른 ID를 사용하거나 약간 변형하여 유포하기도 한다. 따라서, Tabel. 2.와 같이 수많은 콘텐츠가 난잡하게 수집되는 환경에서 다른 정보들을 분석하여 특정의 동일인물로 추정되는 heavy uploader를 탐지할 수 있다. 이에 적합한 시간 흐름에 따른 heavy uploader 대상 유포 관계를 분석하기 위해 웹하드, P2P 데이터에서 수집한 OSP/게시자(ID)별 유포저작물 전반에 대한 정보가 담긴 저작물명을 feature hashing 기법으로 100개의 feature set을 생성한다. feature hashing은 OSP/게시자(ID)별 유포

Table 2. Collection of contents information through crawling

Title	ID	Price	URL	Number	OSP
공포 SF 스릴러 B급의재미 〔〔에이리언 레이더스〕〕	파일콕이다	90	http://m.filekok.com/storage.php?act=view&idx=13409143&mSec=MOV&sSec=all	13409143	모바일 파일*
〔 나니아 연대기 캐스피언 왕자 〕 고화질 한자막	판매장	0	https://m.applefile.com/board/board_view.html?idx=18283287	18283287	모바일 애플*
분노의 질주 4편 더 오리지날 〔Fast And Furious 2009〕 한글자막!!	니자료	210	http://m.wedisk.co.kr/mobile/contents_view.jsp?id=28536930&strm_id=null	28536930	모바일 위디*
〔 쥬. 만. 지. X 넥. 스. 트. 레. 벨. 〕〔 자체자막고화질 〕	빠새빠새호	385	http://m.sedisk.com/storage.php?act=view&idx=25700011	25700011	모바일 새디*
3월 장쯔이x성룡 떠따! ((에 베 레 스 트)) 실화 . 오경 . FHD . 초고화질 . 자체자막	bks	190	http://m.tple.co.kr/?todo=storageView&idx=310761760	310761760	모바일 티*

저작물명을 해시 함수로 해시값을 계산하고, 100으로 모듈러 연산을 통해 해당 인덱스에 1씩 더하여 총 100개의 feature를 생성한다. 따라서, OSP/게시자별 유포저작물 개수 정보에 기반한 유사한 벡터를 가지는 feature set을 생성하며, 어떠한 게시자가 어떠한 OSP에서 몇 개의 저작물을 유포하였고, 그 중 불법복제물 탐지 결과를 통해 불법복제물 개수는 몇 개인지 분석할 수 있다. Table. 2.와 같이 수많은 콘텐츠가 난잡하게 수집되는 환경에서 Table 3.와 같이 OSP/게시자별 유포저작물 개수 정보에 기반한 유사한 벡터를 갖는 feature engineering 결과를 보여준다.

Table 3. OSP/ID Feature Engineering Example

OSP	ID	f-0	…	f-99	Spread	Illegal
미투디*	빨강호떡	0	0	0	201	50
위디*	kis2393	6	0	0	837	708
미투디*	alblc02	5	0	0	991	370
파일*	kkoem1	7	0	0	14	8
애플*	매떡스	11	0	0	141	13

3.3.2 OSP/게시자별 프로파일링

OSP/게시자별 feature engineering을 통해 100개의 feature로 구성된 feature set을 생성하였다. feature set을 사용하여 클러스터링 기반 동일 uploader인지 분석한다. 클러스터링은 k-평균 알고리즘(k-means Algorithm)을 사용한다. k-평균 알고리즘은 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작하여 유사한 feature set을 가지는 데이터끼리 k개의 클러스터로 묶어준다. 수많은 콘텐츠 환경에서 k를 증가시킬수록 정교한 동일 heavy uploader 분석이 가능할 것으로 예상한다. Fig. 6.은 k를 6으로 설정하여 OSP/게시자별 클러스터링 시각화를 보여준다. 동일 클러스터에 포함되었다면, 동일 heavy uploader로 예상한다. 동일 OSP에서 heavy uploader가 ID의 숫자만 변경한 ID로 유포했으나 동일한 유포저작물과 불법복제물을 유포했으므로 feature set의 편차가 크지 않고 동일 클러스터에 속할 것이다. 따라서, heavy uploader는 한 개의 OSP 환경에서 유사한 ID를 사용하거나 다수의 OSP 환경에서 동일 ID를 사용하므로 동일 클러스터에 포함되면 같은 동일 heavy uploader로 예상할 수 있다.

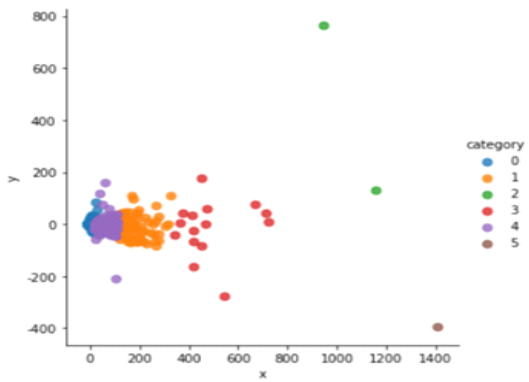


Fig. 6. Visualization of Clustering by OSP/ID

IV. 실험 결과

4.1 Dataset

불법복제물 탐지에 사용한 데이터셋은 한국저작권보호원에 의해 제공 받은 데이터를 사용하였다. 저작권이 있는 공공데이터 17,598개와 모니터링 요원에 의해 확인된 불법복제물 추적관리시스템(illegal content obstruction program) 채증자료 164,454중 성능 검증을 위해 1,000개 데이터로 실험했다. Table. 4.는 Dataset 구성을 보여준다. 공공데이터는 감독 정보, 콘텐츠 정보, 줄거리 등 다양한 메타데이터가 존재하나 실험에는 한글로 번역된 제목명만 사용했고, ICOP 채증자료도 검색 결과를 포함하나 원본콘텐츠 제목만 사용했다.

Table 4. Dataset configuration

Dataset	Count
Copyrighted Public Data	17,598
ICOPW_movies_2020-03	1,000

4.2 불법복제물 탐지 결과

불법복제물을 검색할 때 고려해야 할 사항은 수집 데이터는 노이즈가 다수 존재하여 키워드가 간결하지 않기 때문에 모든 키워드 검색 시 공공데이터에 존재하나 검색되지 않는 불법복제물이 존재하고, 짧은 글자 개수의 제목으로 인한 오검출 가능성이 존재한다. 예를 들어 '기생충'이라는 영화는 2-gram으로 분리 시 '기생'이라는 키워드가 존재하며, 웹하드나 P2P에는 많은 noise가 삽입되어 다른 콘텐츠지만 '기생'이

Table 5. Sample of Experimental results

Collection Data			Technology Verification					
Original Data	Text Normalization	Official Title	Result	k=1	k=2	k=3	k=4	k=5
2O19 실화를 바탕으로 하는〔 스킨 〕은 폭력적인 삶에 찌들어 있던 한 인간이 갱생하는 구원의 이야기이다	실화를바탕으로하는스킨은폭력적인삶에찌들어있던한인간이갱생하는구원의이야기이다	스킨	1	스킨	이다	밤의이야기	집 이야기	내이야기!!
〔_ 2월 완벽자막 떠따! - (미 . 뇨 . 샴 . 총 . 샤 III) - FHD . 초고화질 . 자체자막 . 스샷확인 _〕	미녀삼총사	미녀삼총사 3	NULL	삼총사	삼총사 2014	삼총사 3D	삼총사 2013	조선미녀삼총사
소년과 공룡의 감동 우정 어드벤쳐〔마이펫 다이노소어〕더빙판	소과공룡의감동우정어드벤쳐마이펫다이노소어	마이 펫 다이노소어	1	다이노소어 월드	마이 펫 다이노소어	다이	굿 다이노 3D	굿 다이노
2020.02〔 미스터 주 - 사라진 VIP 〕이성민 1080P	미스터주사라진이성민	미스터 주: 사라진VIP	1	미스터 주: 사라진 VIP	사라진 밤	미스터 캣	미스터 고	미스터 고3D
NY 한복판 다시 사랑할수있을까 (미드나잇 인 뉴욕)0416	한복판다시사랑할수있을까미드나잇인뉴욕	미드나잇 인 뉴욕	NULL	미드 90	다시	다시사랑할수있을까?	미드나잇 선	미드나잇 맨
〔NEW〕폭발이 시작됐다(e병헌.ha쩡우)_ 윗동네 흰머리산	폭발이시작됐다병헌쩡우윗동네흰머리산	백두산	0	단발머리	머리와 몸	여관발이	머리카락	
〔매직 오브 벨 아일〕SD 원하면 누구든 될 수 있어 aabb	매직오브벨아일원하면누구든될수있어	매직 오브 벨 아일	1	매직오브벨아일	인 디아일	매직울프	선오브갓	매직티팟
〔데스노트:라스트네임〕원작을 뛰어넘는 새로운 결말의 데스노트! 극찬	데스노트라스트네임원작을뛰어넘는새로운결말의데스노트극찬	데스노트: 라스트네임	NULL	스트라튼	라스트 홈	라스트씬	라스트 송	고스트 노트
2020. 스칼렛요한슨 신작〔〔 쪼 조 ㄹri 빗 〕〕 완벽한글자막 초 고화질	스칼렛요한슨신작조조래빗	조조래빗	1	조조래빗	래빗	스칼렛디바		
〔_ 3월 장쯔이x성룡 떠따! - (에 . 베 . 레 . 스 . 트) - 실화 . FHD . 초고화질 . 자체자막 _〕	장쯔이성룡에베레스트실화	에베레스트	1	에베레스트	에베레스트(3D)	언레스트	어레스트미	레스트리스
〔옹알스〕M 말없이 웃음으로 모두를 사로잡은 코미디팀 옹알스 jenm	옹알스말없이웃음으로모두를사로잡은코미디팀옹알스	옹알스	1	옹알스	이웃동서	숲속으로	자연으로	대리 남편
O3월. 돌아온.윌스미스〔 ㄴㅏ.쁜.색.ㅎㅣ.덜. 뽀.레.버 〕 HDTS.한글번역자막	돌아온윌스미스나쁜녀석들포에버	나쁜녀석들 :포에버	1	나쁜녀석들 : 포에버	나쁜녀석들	일진 나쁜녀석들	돌아온다	언틸 포에버
〔 바람의검심3 - 전설의최후 (2014) 〕	바람의검심전설의최후	바람의검심 3 - 전설의최후	1	전설의검	바람의검심	바람의색	바람 바람 바람	전설의 주먹
2O ㅣ 9. ㅣ O NEW 고골 악령과의 전쟁〔 미스터리 판타지 블록버스터 〕FHD자체자막	고골악령과의전쟁미스터리판타지블록버스터	고골: 악령과의전쟁	1	버스657	고골: 악령과의 전쟁	악령	용의 전쟁 1885	미스터 캣

라는 키워드를 포함할 경우 오검출 가능성이 있다. 본 논문에서는 공공데이터에서 수집 데이터를 검색하고, 공공데이터의 제목 글자 수에 따라 threshold를 다르게 설정하고, 수식 3에 따른 유사도가 높은 순서대로 정렬하여 오검출을 최소화하였다. 또한, 유사도가 같으면 메타데이터를 활용하여 최신 콘텐츠를 기준으로 정렬한다. 불법복제물 탐지 결과는 Table 5.와 같다. k가 낮을수록 더 원본 데이터와 유사하다고 할 수 있고, k가 5까지의 검색된 결과를 가지고 정탐 여부를 판단했다. 공공데이터에 존재하지 않은 데이터는 검색할 수 없으므로 Null로 표기한다.

단점으로는 라벨이 존재하지 않아 여러 인력을 투입하여 수작업으로 검증해야 한다. 따라서, 실험에는 1,000개의 데이터만 검증했다. 1,000개 데이터 중에 40.7%가 공공데이터에 존재하지 않는 데이터이다. 나머지 59.3% 중에 52.7%가 정상으로 검색하고, 6.6%가 오검출 되었다. 공공데이터에 존재하지 않는 데이터를 제외하고, 88.8%의 탐지율을 가지며 이는 기존의 문자열 매칭방식을 개선할 수 있을 것으로 기대한다.

4.3 Heavy Uploader 프로파일링 결과

유포저작물 대상 동일인으로 추정되는 heavy uploader를 탐지하기 위해 OSP/게시자별 유포저작물 전반에 대한 정보가 담긴 feature engineering을 통해 100개의 feature set을 생성했고, 클러스터링을 통해 동일인으로 추정되는 heavy uploader를 탐지하는 방법을 제안했다. Table. 6.은 특정 클러스터의 OSP와 게시자별 유포저작물을 보여준다.

Table. 6.는 동일 OSP에서 약간의 ID를 변경하여 동일 콘텐츠를 유포한 동일인으로 추정되는 heavy uploader이다. 따라서, 수많은 콘텐츠가 난잡하게 수집되는 환경에서 동일인으로 추정되는 heavy uploader 프로파일링 분석으로 약간의 제목만 변경하여 여러 OSP를 통해 유포하고, 다수의 전혀 다른 ID를 사용하거나 약간만 변형하여 유포하는 동일인을 탐지할 수 있다.

OSP/게시자별 이외에 여러 가지 메타데이터로 다양한 프로파일링 분석을 할 수 있다. 첫째, 제안한 클러스터링 기반 동일 heavy uploader 분석 결과와 메타데이터를 통해 heavy uploader 유포시간 및 이동현황 분석 결과를 비교하면 정교한 동일 heavy uploader 분석이 가능하다. 따라서, 클러스터별 특정 날짜 기간에 유포저작물 개수가 같을수록 동일 heavy uploader로 판단한다. 둘째, heavy uploader는 한 개의 OSP에서 유사한 ID를 사용한다. 게시자 ID를 비교하여 유사도가 높은 것은 동일인으로 추정할 수 있다. 셋째, 유포저작물 개수로 저작권 침해사이트, 침해콘텐츠 중요도 분석이 가능하다. OSP별 불법복제물 유포 개수가 많은 것부터 우선순위가 높아 주요 분석대상 사이트가 된다. 또한, 저작물명별 불법복제물 유포 개수가 많은 것부터 우선순위가 높아 상용 환경에 가장 불법복제물이 유포되고 있다는 것을 알 수 있다. 따라서, 저작권 침해상황 파악 및 주요 대응 대상 식별이 가능하다.

Table 6. Same Heavy Uploader Profiling

OSP	Uploader	Original Data
미투*	imbc06	〔n번째이별중〕사랑하는 그녀의 마음을 되돌리기 위해 타임머신 어플 개발 imm
미투*	imbc05	〔n번째이별중〕사랑하는 그녀의 마음을 되돌리기 위해 타임머신 어플 개발 imm
미투*	imbc04	〔보이드갱〕 HD 은행털이범을 연기한 에디윈 보이드 MOF1
미투*	imbc07	〔보이드갱〕 HD 은행털이범을 연기한 에디윈 보이드 MOF1
미투*	imbc07	〔킬 케인〕 가족을 죽1인 갱들에게 복1수를 시작한다 MOF1
미투*	imbc08	〔킬 케인〕 가족을 죽1인 갱들에게 복1수를 시작한다 MOF1
애플*	goldtjdwhd12	인 타임 (저스틴팀버레이크아만다사이프리드주연)
예스*	goldtjdwhd12	인 타임 (저스틴팀버레이크아만다사이프리드주연) (4)

V. 결 론

웹하드 및 P2P 사이트의 사용자들은 편리하고 저렴한 가격에 콘텐츠를 바로 내려받을 수 있어서 편리하다. 하지만 내려받은 콘텐츠 파일은 대부분 저작권을 무시한 불법복제물이다. 불법복제물을 검색하고 차단하기 위해 한국저작권보호원에는 불법복제물 추적관리시스템을 운영하고 있으나, 이는 문자열 매칭 방식으로 동작하기 때문에 불법복제물 업로드 시 매

칭되는 과정을 우회하기 위하여 제목에 불필요한 특수문자와 공백을 삽입하거나 문자 변형, 음절 무시 등 노이즈를 추가하는 방식을 사용한다. 제안 모델은 문자열 매칭방식을 개선하기 위해 문자열에 노이즈를 제거하는 텍스트 정규화 처리 및 키워드 기반의 bloom filter를 활용한 불법복제물을 검색한다. 또한, heavy uploader 프로파일링 분석을 위해 유포 저작물 전반에 대한 정보를 담은 feature set을 생성하는 feature engineering 기술과 이를 활용하여 클러스터링 기반 동일인으로 추정되는 heavy uploader를 탐지한다. heavy uploader 대상 불법복제물을 검색하고 차단한다면 저작권 피해가 대폭 감소할 것으로 예상한다. 향후, 불법복제물 추적탐지 시스템의 성능을 개선을 위한 노력과 저작권을 보호하기 위한 연구를 이어갈 예정이다.

References

[1] Jung-Wook Cho, "OSP's Obligation to Adopt Technical Measures to Curtail Copyright Infringements", Journal of Korea Infomation Law, 12(2), pp.63-94, Dec. 2008.

[2] Yeong-Woo Oh, Gye-Hyun Jang, Hun-Yeong Kwon, Jong-In Lim, "A Study on the Copyright Protection Liability of Online Service Provider and Filtering Measure", Journal of the Korea Institute of Information Security and Cryptology, 20(6), pp.97-109, Dec. 2010.

[3] "2020 Annual Report on Copyright Protection," Korea Copyright Protection Agency, pp. 1-168, April. 2020.

[4] "Status of special types of value-added telecommunications providers," Ministry of Science and ICT(MSIT), Korea, Dec. 2019.

[5] Tae-Jung Kim, Jae-Ho Hwang, and Choong-Seong Hong, "A Wavelet Based Robust Logo Watermarking Algorithm for Digital Content Protection," *Joural of Internet Computing and Services*, 9(1), pp. 33-42, Feb. 2008.

[6] Jeon Seong-Tae, Jeon Soo-Jung, "A Study on the Legal Problems of DRM on the Use of copyright", Korea Copyright Commission, 75(19), pp.66-87, Oct. 2006.

[7] Jung-Sik Hwang and Hyun-Gon Kim, "Blockchain-based Copyright Management System Capable of Registering Creative Ideas," Journal of Internet Computing and Services, 20(5), pp. 57-65, Oct. 2019.

[8] "A Study on the Copyright New Service Model Using Blockchain Technology", Korea Copyright Commission, pp.1-174, Jan. 2018.

[9] Hwang Jung-Sik, Kim Hyun-Gon, "Blockchain-based Copyright Management System Capable of Registering Creative Ideas", The Korea Society of Science & Art, (35), pp.341-351, Oct. 2019.

[10] Ju-Seop Kim, Je-Ho Nam, "Analysis of illegal content filtering technology trends", Broadcasting and Media Magazine, 12(4), pp.53-63, 2007.

[11] C. Li, J. Lu and Y. Lu, "Efficient Merging and Filtering Algorithms for Approximate String Searches," 2008 IEEE 24th International Conference on Data Engineering, pp.257-266, April. 2008.

[12] Jong-An Kim, Jong-Heum Kim, Jin-han Kim and Yong-min Chin, "Development of the filtering technology of illegal IPTV contents," *Korea Institute of Information & Telecommunication Facilities Engineering*, pp. 108-111, Aug. 2009.

[13] Hyeon-Gu Son, Ki-su Kim and Young-seok Lee, "A File Name Identification Method for P2P and Web Hard Applications through

Traffic Monitoring" *Journal of KIISE*, 37(6), pp. 477-482, Dec. 2010.

[14] Bong-Gi Kim, Hae-Seok Oh, "A Feature-Based Retrieval Technique for Image Database", The transactions of the Korea Information Processing Society, 5(11), pp.2776-2785, Nov. 1998.

[15] Hee-Wan Park, "Design and Implementation of Server-based Resource Obfuscation Techniques for Preventing Copyrights Infringement to Android Contents," *Journal of the Korea Contents Society*, 16(5), pp. 13-20, May. 2016.

[16] Chan-Woong Hwang, Ji-Hee Ha and Tae-Jin Lee, "Modified File Title Normalization Techniques for Copyright Protection," Convergence Security Journal, 19(4), pp. 133-142, Oct. 2019.

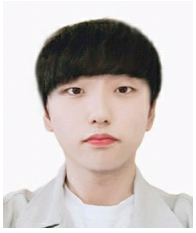
[17] Almeida, Paulo Sérgio, et al. "Scalable bloom filters." Information Processing Letters 101(6), pp.255-261, March. 2007.

[18] Broder, Andrei, and Michael Mitzenmacher. "Network applications of bloom filters: A survey." Internet mathematics 1(4), pp. 485-509, Jan. 2011.

〈저자소개〉

황 찬 웅 (Chan-woong Hwang) 학생회원
2014년 3월~2020년 2월: 호서대학교 정보보호학과 졸업
2020년 3월~현재: 호서대학교 정보보호학과 석사과정
〈관심분야〉 네트워크 보안, 악성코드 분석, 기계학습

김 진 강 (Jin-gang Kim) 학생회원
2016년 3월~현재: 호서대학교 정보보호학과
〈관심분야〉 포렌식, 악성코드 분석, 기계학습



이 용 수 (Yong-soo Lee) 학생회원
2016년 3월~현재: 호서대학교 정보보호학과
〈관심분야〉 네트워크 보안, 정보보호, 기계학습

김 형 래 (Hyeong-rae Kim) 학생회원
2016년 3월~현재: 호서대학교 정보보호학과
〈관심분야〉 악성코드 분석, 정보보호, 취약점 분석

이 태 진 (Tae-jin Lee) 종신회원
2003년 1월~2017년 2월: 한국인터넷진흥원 R&D 팀장
2017년 3월~현재: 호서대학교 컴퓨터정보공학부 교수
〈관심분야〉 시스템 보안, 악성코드 분석, 기계학습