https://doi.org/10.33778/kcsa.2019.19.4.133

저작권 보호를 위한 변형된 파일 제목 정규화 기법★

황 찬 웅*, 하 지 희, 이 태 진**

요 약

토렌트 및 P2P 사이트나 웹 하드는 쉽고 편리하게 무료로 다운로드 받거나 가격을 저렴하게 이용할 수 있다는 이유만으로 사용자들은 자주 이용하지만 국내 토렌트 및 P2P 사이트나 웹 하드는 저작권과 관련되어 매우 민감하기 때문에 저작권 보호를 위한 기술들이 연구되고 적용 되어지고 있다. 이 중에서 파일의 제목이나 주요 단어의 조합 등 경우의 수를 금칙어로 설정하여 차단하는 제목 및 문자열 비교방식 필터링 기술은 제목 변경, 띄어쓰기 등을 통해 우회가 용이하다. 저작권 보호를 위한 불법저작물을 검색하고 차단하기 위해서는 변형된 파일 제목을 정규화 하는 기술이 필수적이다. 본 논문에서는 불법저작물의 변형된 파일 제목을 정규화 하는 기법과 파일 제목을 정규화를 진행 전과 후에 따른 검색에 의한 탐지율을 비교하였다. 정규화를 진행하기 전 탐지율은 77.72%로 아쉬운 탐지율이 보인 반면에 정규화를 진행한 후 90.23%로 정규화가 필수적이라고 말할 수 있다. 향후, 공통으로 나타나는 날짜와 화질 표시 같은 무의미한 용어들을 처리하면, 더욱 좋은 결과가 산출될 것으로 기대한다. 국문 요약입니다.

Modified File Title Normalization Techniques for Copyright Protection

Hwang Chan Woong*, Ha Ji Hee, Lee Tea Jin**

ABSTRACT

Although torrents and P2P sites or web hard are frequently used by users simply because they can be easily do wnloaded freely or at low prices, domestic torrent and P2P sites or web hard are very sensitive to copyright. Techni ques have been researched and applied. Among these, title and string comparison method filtering techniques that bl ock the number of cases such as file titles or combinations of key words are blocked by changing the title and spaci ng. Bypass is easy through. In order to detect and block illegal works for copyright protection, a technique for norm alizing modified file titles is essential. In this paper, we compared the detection rate by searching before and after no rmalizing the modified file title of illegal works and normalizing the file title. Before the normalization, the detection rate was 77.72%, which was unfortunate while the detection rate was 90.23% after the normalization. In the future, i t is expected that better handling of nonsense terms, such as common date and quality display, will yield better resu lts.

Key words : P2P Sites, Banned Word, Filtering, File Title Normalization, Detection

접수일(2019년 9월 3일), 게재확정일(2019년 9월 19일)

* 주저자, hwang85123@naver.com
** 교신저자, kinjecs0@gmail.com (Corresponding author)

★ 본 연구는 문화체육관광부 및 한국저작권위원회의 2019년도 저작권기술개발사업의 연구결과로 수행되었음.
(No. 2019-PF-9500)

1. 서 론

한국저작권보호원 통계에 따르면 2017년 불법복제물 이용량은 총 약 20억 8천 3백만 개에서 온라인 불법복제물 이용량은 약 18억 7천 7백만 개로 전체 불법복제 이용량의 90.2%를 차지한 것으로 조사되었다. 온라인 불법복제물 이용에 대한 유통 경로별 비중을 살펴보면, 토렌트가 28.8%, 모바일 21.9%, 웹 하드 17.9% , 포털 16.9%, P2P 9.8%, 스트리밍 전문 서비스 5.8% 순으로 집계되었다. 콘텐츠별 침해율은 영화 22.9%, 음악 20.3%로 가장 높았다. 또한, 2017년 불법복제물로 인한 직·간접적인 생산 감소는 콘텐츠 산업에서 약 3조 원, 우리나라 전체 산업에서 약 4조 8천억 원으로 분석되었으며, 이에 따른 고용손실은 콘텐츠 산업에서 약 3만 명, 전체 산업에서 약 4만 3천명에 달하는 것으로 분석되었다[5]. P2P 사이트에서 콘텐츠를 내려받는 행위는 저작권법상 복제에 해당되는 사항이며, 영리를 목적으로 하지 않고 개인적인 이용이나 가정에 준하는 한정된 범위도라도 불법적으로 저작물을 다운받는 경우 사적인 복제 면책 규정이 적용되지 않아 저작권 침해에 해당된다. P2P 사이트는 다운을 받는 동시에 업로드가 이루어지기 때문에 저작권 위반 및 처벌에 해당된다.

저작권 보호를 위한 기술적 조치로 검색어 기반 필터링 방법을 사용한다. 불법저작물은 이러한 저작권 위반 및 처벌을 우회하기 위해서 불필요한 기호들을 추가하거나 문자를 변경하는 방법을 사용한다. 따라서, 검색어 기반 필터링 작업이 어려운 문제이다. 변경된 제목의 예는 아래와 같다.

- O1월. 중국 박스오피스 1위 범죄 액션 정.ㅇ1.건[금.괴.털.이]초고화질.한글자막
- 2019.01월 (신작) [--- ㅇr.쿠.ㅇㅏ.맨 ---]완벽한글.정식 DVD릴.초고화질.무삭제 1080P
- 미스터리액션[ㅡㅁㅐㅇㅣㅈ.ㄹㅓ너ㅡ]완벽한글.108oP
- 2018.05.떠따!!초SF-불ㄱrㅅrㄹI 6- ㅈI옥으I 추운날.초고화질.한글자막 aa
- 2018.o4 (리빙빙) SF.액션 [ㅡ 고ㄷH . ㅈ lㅎ r . 무 덤 ㅡ] 한글자막zz

변형된 제목은 사용자는 쉽게 어떤 콘텐츠를 포함하는지를 알 수 있지만 컴퓨터는 어떤 콘텐츠를 포함하는지 알 수 없어 자동 필터링 적용이 어려움을 격는다. 이 연구에서 변형된 제목 문장을 복원하기 위하여 한글 자음, 모음과 유사한 문자를 한글로 변환하거나, 한글 자음과 모음으로 구성된 문자열을 음절로 결합하는 방법을 적용한다.

2. 관련 연구

효과적인 저작권 보호 방법에 관한 연구로 불법 웹툰을 자동으로 모니터링과 IP기반 접속차단위한 신속한 증거수집이 가능하다. 880만 개의 웹툰 특징점을 등록하고 이를 식별하는데 평균 1.48초의 시간이 소요된다[2].

P2P 환경에서 파일 공유를 할 때 발생할 수 있는 보안 위협도 존재한다.[3] 또한 P2P 환경에서 DHT기반 다중 키워드 검색시 발생되는 노드간의 전송되는 역리스트의 양을 효과적으로 줄이기 위해 Bloom-fiter기법을 적용하였다[4].

연구 [5]에서 제안한 MAUCA는 사용자의 컨텍스트 정보를 속성별로 분류하고, 사용자 컨텍스트 정보와 서비스의 상관 관계를 분석하여 사용자에게 필요한 서비스를 제공하도록 한다.

P2P 사이트에서 불법저작물을 공유하려면 파일 ID을 생성해야 한다. 이 영상 파일의 경우, 저작권 보호 요청에 의해 파일 ID이 공유 금지 리스트에 추가되어 있다. 따라서 다운로드를 수행하면 저작권 보호 요청 또는 유해정보로 분류되어 경고창과 함께 다운로드가 진행되지 않는다. 하지만 파일을 압축하거나 변형하여 공유할 경우, 파일 제목이 달라지고, 기존 파일 ID에 대한 공유 금지 설정을 우회할 수 있게 되어 실제로 다운로드가 가능하다[6].한 글자로 작성하되 들여쓰기 없이 작성한다. 정리, 보조정리, 따름정리의 증명 끝에 증명 끝 표시문자 □를 표시한다.

정리 1. [홍길동 정리] 논문은 이 양식을 만족해야 한다.

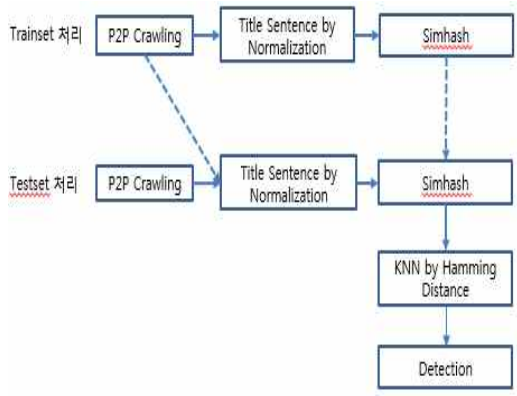
스팸 문자 메시지에서는 “첫!가!입!”, “㉮ㅈ①N 0”, “ㅇ♯ ㅁ✝ 토가”, “ㄱ ㅏ입 ”, “つr입즉ΛI”와 은 한글단어들의 초성, 중성, 종성을 유사한 형태의 영어나 다른 나라의 언어, 특수기호, 숫자 등으로 대치하여 단어를 다양한 형태에 단어로 왜곡시킨다. 그리하여 보편적인 스팸 문자 메시지 필터링 시스템에서는 이러 한 왜곡된 단어들을 “첫가입”, “카지노”, “가입즉시”와 같은 정규화 과정을 거쳐 스팸 문자 메시지 필터링을 위한 어휘 사전과의 비교를 통해 스팸 문자 메시지를 차단하고 있다. 이 방법을 적용하면 변형된 문자열을 정규화함으로써 스팸 문자 차단 효과를 17% 향상시키는 효과를 보였다.[7,8] 현재 텍스트에 대한 백터의 표현 방식으로 TF-IDF가 가장 널리 사용되고 있지만, 단어 수가 증가함에 따라 차원의 수도 같이 증가하여 대용량 단어들을 처리하는데 어려움이 있다. 따라서 워드 임베딩과 딥러닝 기법을 이용하여 처리하여 벡터를 통한 단어 의미 유사성 추론에서 더 나은 성능을 보여준다[9,10,11,12].

따라서, 연구 [6]에서처럼 불법저작물을 등록을 우회하기 위해 파일 ID 즉, 제목을 변경하거나 압축하여 등록하면 쉽게 사용자가 다운로드가 가능하다는 문제점이 있다. 파일 ID를 정규화하여 불법저작물을 등록을 차단하는 기술이 필요하다. 본 연구에서는 연구 [7,8]처럼 스팸 문자를 정규화하는 과정과 달리 불법저작물은 한글 뿐만 아니라 영어로 된 제목도 존재하고 날짜나 화질을 나타내는 숫자도 변형하기 때문에 영숫자 정규화 처리와 한글 정규화 처리를 나누어 진행하였다. 이후 연구[9,10,11,12]처럼 텍스트에 대한 TF-IDF나 워드 임베딩 딥러닝 기법을 이용하여 불법저작물 검색 및 탐지가 가능하다.

3. 제안 모델

앞서 P2P 사이트나 웹 하드에서 불법저작물들은 특정 단어 기반으로의 검색과 탐지되는 것을 우회하기 위해 제목을 변경한다고 하였다. 이번 장에서는 변경된 제목을 컴퓨터가 학습할 수 있게 제목 문장을 정규화하는 기법을 소개하고 정규화한 문장을 이용하여 Simhash를 거쳐 Hamming 거리 기반 유사 파일 검색을 통하여 불법저작물을 탐지한다.

3.1 전체 구조



(그림 1) 전체 구조

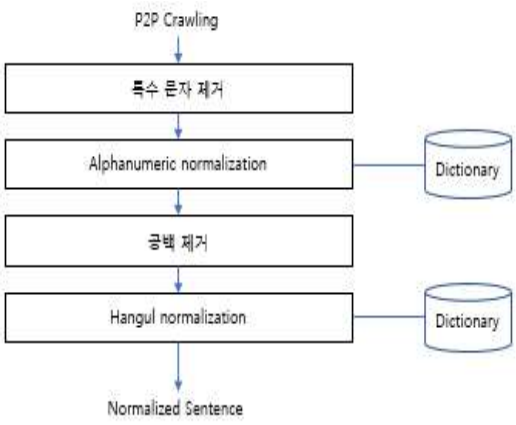
그림 1은 불법저작물 탐지 과정을 기술한 것이다. P2P Crawling 작업을 통하여 데이터를 생성하고 제목을 정규화한다. Simhash를 통해 0과 1로 vector화를 진행하고 Hamming Distance를 통해 유사한 파일을 검색하여 불법저작물을 탐지한다.

3.2 P2P Crawling

사용할 데이터는 P2P 사이트에서 크롤링을 진행하여 100,000개로 구성된 제목, 용량을 표시하는 사이즈, 거래되는 가격(캐시), 드라마나 영화등 콘텐츠 종류, 배포자의 아이디를 포함하는 데이터셋을 생성한다. 이 중에서 본 연구는 제목만을 가지고 실험을 진행한다. 제목에는 대부분 날짜와 화질, 인코딩 닉네임을 표시한다.

3..3 Title Sentence by Normalization

앞서 크롤링한 데이터는 변형된 제목을 포함하고 있기 때문에 정규화 과정이 필요하다. 그림 2는 변형된 제목을 정규화 하는 기법을 기술한 것이다.



(그림 2) 제목 문장 정규화

3.3.1 특수 문자 제거

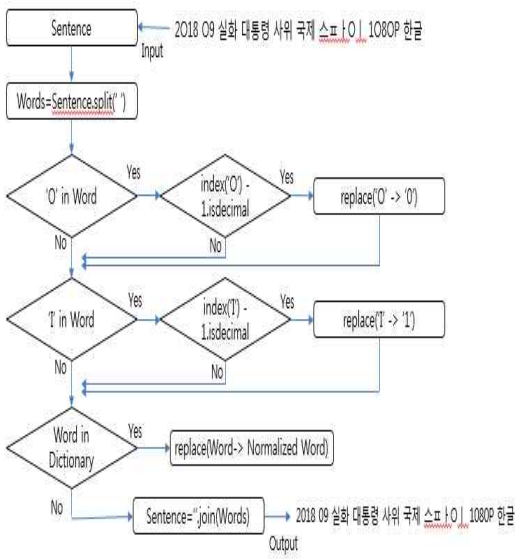
모든 불법저작물은 특수 문자를 삽입하여 검색 필터링 과정을 우회한다. 또한 특수 문자를 이용하여 제목의 핵심 단어를 부각하기 때문에 불필요한 특수 문자를 먼저 제거해야 한다.

3.3.2 Alphanumeric Normalization

스팸문자와 달리 토렌트, P2P, 웹하드에서는 제목이 영숫자와 한글 모두를 포함하고, 날짜와 화질을 표현하는 영숫자 또한 변형시키는 대상이기도 하다. 예를 들어 1O8OP나 2O18 I2 같이 화질이나 날짜를 표기할 때 1080P와 2018년 12월을 나타내는 단어를 변형하고 대부분 0과 1을 대문자 O와 대문자 I로 많이 표현한다. 대문자 O와 I뿐만 아니라 소문자 o와 소문자 l로 표현하기도 하고 한글 자음 ㅇ과 모음 ㅣ도 대부분 대문자 O와 I를 사용하여 변형한다.

영숫자 정규화 과정에서 대문자 O와 I를 0과 1로 변환하면 제목의 모든 대문자 O와 I를 0과 1로 변환하기 때문에 3가지의 조건을 확인하고 변환해야 한다. 첫째, 숫자가 변형한 형태여야 한다. 둘째, 영어로 된 제목이나 제목에 영어 단어가 존재하면 대문자 O와 I에 대하여 0과 1로 변환되어지면 안된다. 예를 들어 'OCN', WITH 같이 영어 단어를 포함하는 제목들은 '0CN', 'W1TH'로 변환되어 본래의 제목을 잃어버리게 된다. 셋째, 대문자 O와 I가 한글 자음 ㅇ과 모음 ㅣ를 변형시킨 것은 0과 1로 변환되어지면 안된다. 예를 들어 'Oㅣ'나 'ㄱI'처럼 한글을 변형한 형태가 '0ㅣ'나 'ㄱ1'처럼 변형되어지면 안되고 대문자 O와 I가 한글 자음 ㅇ과 모음 ㅣ로 변경되어져야 하기 때문이다.

그림3은 3가지 조건을 확인하기 위하여 문장에서 공백을 기준으로 단어로 나눈다. '2 O 1 8'처럼 공백을 기준으로 나누게 되면 '2', 'O', '1', '8'로 나눠지고 대문자 O와 I에 대해서 독립적으로 존재하면 0과 1로 변경한다. 또한 단어의 대문자 O와 I를 찾고 앞 문자가 숫자일 경우 0과 1로 변환 한다. 대문자 O와 I가 맨 앞에 있는 경우에는 단어의 맨 뒤 문자로 판단한다. 이로써 영어 단어는 영어 문자로 이루어져 있어 변경되지 않고 1O8OP과 2O18 I2는 1080P와 2018 12로 변경되어진다. 하지만 더 교밀한 불법저작물은 IO8OP나 'IO월'처럼 대문자 O와 I에 대해 앞 문자가 영문일 경우 1080P나 '10월'으로 변경되지 않는다 따라서, 단어를 기준으로 대응대는 단어로 변경해주는 Dictionary를 생성하여 변경해 준다.



(그림 3) Alphanumeric Normalization Process

대문자 O와 I는 0과 1로 변경되어지는지 한글 자음 ㅇ과 모음 ㅣ로 변경되어지는지 구분을 하기 위하여 Alphanumeric normalization(영숫자 정규화)과정과 Hangul normalization(한글 정규화)를 나누어 진행하는 이유이다.

3.3.3 공백 제거

Hangul normalization을 진행하기 전에 공백을 제거한다. 불법저작물은 음절마다 공백을 삽입하거나 한글을 변형시키고 자음과 모음 사이에 공백을 삽입하는 경우가 있다. 예를 들면 'O ㅏ'처럼 자음 ㅇ을 대문자 O로 변형시키고 모음 ㅏ와 사이에 공백이 존재한다. 이후 3.3.4의 Hangul normalization을 수월하게 진행하기 위해선 공백 제거를 진행 해야만 한다.

3.3.4 Hangul normalization

3.3.2에서 영숫자(Alphanumeric)를 정규화를 진행했다면 이번 장에서는 한글을 정규화 하는 과정을 설명한다. 'ㄱㅏ'와 같이 한글 자음과 모음이 변형되지 않고 분리되어 있는 경우도 있지만 변형된 한글은 대부분 자음과 모음을 변형한 형태이다. 자음 ㅇ는 대문자 O와 소문자 o 그리고 숫자 0로 표현하고 자음 ㄴ은 대문자 L, 자음 ㅌ는 대문자 E로 다양하게 표현이 가능하다. 모음 ㅣ는 대문자 I, 소문자 i, l, 숫자 1 표현이 가능하고 모음 ㅏ는 소문자 r과 소문자 t, 모음 ㅐ는 대문자 H와 소문자 h,로 다양하게 표현이 가능하다. Hangul normalization(한글 정규화)는 변형된 한글 자음과 모음을 대응되는 한글 자음과 모음으로 변경하기 위한 표 1에 해당되는 Dictionary1과 한글 자음과 모음을 결합하여 음절을 생성하는 표 2에 해당하는 Dictionary2 총 2개의 해당되는 Dictionary를 사용한다.

<표 1> Dictionary1

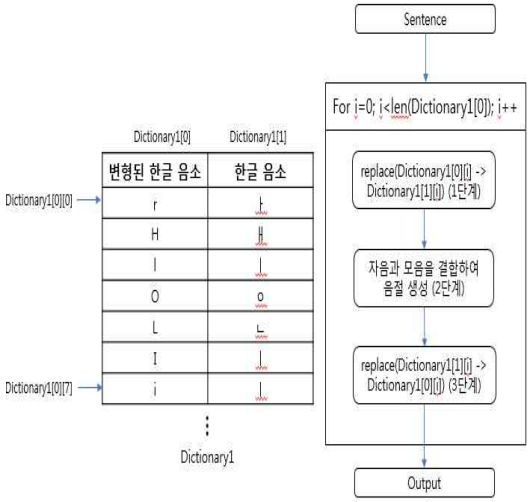
Before	After
i	ㅣ
H	ㅐ
h	ㅐ
t	ㅏ
r	ㅏ
I	ㅣ
l	ㅣ
1	ㅣ
O	ㅇ
L	ㄴ
o	ㅇ

<표 2> Dictionary2

ㄱ	ㄱ'	ㄴ	ㄴ'	…
ㄱㅏ	가	ㄴㅏ	나	
ㄱㅐ	개	ㄴㅐ	내	
ㄱㅑ	갸	ㄴㅑ	냐	
ㄱㅒ	걔	ㄴㅒ	냬	
ㄱㅓ	거	ㄴㅓ	너	
ㄱㅔ	게	ㄴㅔ	네	
ㄱㅕ	겨	ㄴㅕ	녀	
ㄱㅖ	계	ㄴㅖ	녜	
ㄱㅣ	기	ㄴㅣ	니	

기본 개념은 Dictionary1로 변형된 한글 자음과 모음을 대응되는 한글 자음과 모음으로 변경시켜주고 Dictionary2로 변경시킨 한글 자음과 모음을 결합하여 음절로 변경하는 것이다. Dictionary2는 빠른 검색을 위해 제목에 ㄱ, ㄴ과 같은 자음을 찾은 다음 ㄱ, ㄴ열에 해당하는 모음을 찾고 대응하는 ㄱ', ㄴ' 열로 변경해 준다. 변형된 한글 자음과 모음은 대부분 영문자로 변형시키기 때문에 Dictionary1을 한번에 문장에 적용하게 되면 영어로 된 제목과 정상적인 영어 단어가 존재 하면 영어 단어가 한글 자음과 모음으로 변경되고 음절까지 생성될 수 있다. 예를 들어 Light 라는 영어 단어를 가지는 제목이 있다고 가정해보면 Light는 Dictionary1을 통해 ㄴㅣgㅐㅏ변경되고 Dictionary2를 통해 니gㅐㅏ로 변경된

다. 이러한 문제 때문에 본 연구에서는 그림 4와 같이 총 3단계로 진행한다. Dictionary1이 한 문자에 대해 Before에서 After로 진행(1단계)되면 Dictionary2를 거쳐(2단계) After에서 Before 다시 역변환(3단계)을 반복한다.



(그림 4) Hangul Normalization Process

3단계 역변환을 해주면서 2단계의 한글 자음과 모음이 조합하여 음절이 되지 않은 경우는 제목에서 영어 단어가 변경되어 진 경우이기 때문에 영향을 미치지 않는다. 예를 들어 Light는 L ㅣ ght로 변경되었다가 2단계인 음절로 변환이 되지 않고 3단계인 Light로 다시 변경되어진다. 다른 숫자와 영어 단어에 영향을 끼치지 않고 한글 자음과 모음으로 음절로 변환하는 것이 핵심이다. 하지만 한글 자음과 모음을 한 개씩 변환하다 보니 자음과 모음 한번에 변형된 것은 음절로 변환이 되질 않는다. 예를 들어, 'OI'나 'OrOl' 같은 한글 자음과 모음이 모두 변형된 상태는 'ㅇ'나 '아이'로 변형되지 않는다. 또한, 이중 자음을 같은 음절과 종성이 분리된 음절을 생성하지 못한다. 예를 들어, '고r'나 '으i' 같은 이중 모음을 갖는 표현과 '따O'이나 '고O'같은 종성을 분리하여 표현하는 것은 하나의 음절로 변환이 어렵다. 이러한 문제를 해결하기 위하여 Dictionary3를 생성한다. Dictionary3는 표 3과 같다. 앞서 정규화되지 않은 것을 Dictionary3로 정규화가 가능하였다.

<표 3> Dictionary3

Before	After
OrOl	아이
OI	이
OH	애
고r	과
으1	의
주ㅓ	줘
따O	땅
도O	동

본 논문에서는 Dictionary 표의 일부를 보여주었다. Dictionary는 삽입, 삭제, 수정이 자유롭게 작성할 수 있다. 화이트 리스트 ,Database를 이용하여도 무방하다.

3.4 Simhash , KNN by Hamming

SImhash[13] 기반 알고리즘은 대량의 데이터에서 중복 제거를 처리하기 위한 LSH[14](Locality Sensitive Hashing) 알고리즘으로, 수억 개의 웹 페이지에서 중복된 문서를 검색하고 제거하기 위해 설계되었다[15]. Simhash기반 알고리즘 적용시 텍스트 데이터에 대해 Vector화 처리가 가능하며, 유사한 텍스트 데이터에 대해서는 유사한 Hash값을 가지게 되는 특징이 존재한다. 본 논문에서는 불법저작물의 정규화 처리된 제목을 활용하여 Feature를 추출하고 Hash Funciton을 적용하여 Simhash 값을 산출한다. bit로 표현된 Simhash값 간의 Hamming distance 비교를 통해 유사한 정도를 확인 할 수 있으며 사전에 생성한 Simhash값을 기반으로 분석대상과 유사한 데이터 검색을 진행한다.

4. 실험 결과

본 논문에서는 불법저작물 탐지 시 정규화 작업의 중요성에 대해 검증하기 위해 정규화 처리를 하지 않은 데이터와 정규화 처리를 진행한 데이터로 Simhash기반 유사 게시글 검색을 통한 불법저작물 탐지 실험을 진행하였다.

4.1 Dataset 구성

불법 저작물 탐지를 위해 구성한 데이터 셋은 P2P 사이트에서 수집한 10만개 데이터를 활용해 실험을 진행하였다. 유사 게시글 검색시 기준이 되는 Train 데이터 셋은 100,000개로 구성하였고, 그중 300개를 선별하여 분석하고자 하는 대상이 되는 Test 데이터로 구성하였다. 데이터 셋 모두 제목을 가지고 정규화를 진행한다. 제목에는 영숫자와 한글 모두 포함하기 때문에 영숫자를 먼저 처리하고 변형된 한글에 대해 정규화를 진행하였다. 변형되지 않은 문장도 불필요한 특수문자를 제거하고 공백이 존재 하지 않는 문자열 형태로 정규화가 진행된다. 정규화된 문장을 가지고 다양한 방식으로 실험을 진행할 수 있다. 본 실험은 정규화 처리된 제목을 활용하여 Simhash기반 유사 검색을 진행하였다. Feature를 추출하기 위해 문자열 데이터에 2-gram을 적용하였으며, 각 Feature에 Hash Funciton을 적용하여 50bit의 Simhash값을 추출하였다. 분석 대상에 대한 50bit의 Simhash값과 사전에 생성한 Train 데이터들의 50bit Simhash값을 비교하여 Hamming distance 5 이내의 데이터를 유사한 게시글로 판단한다. 아래 표 4는 유사 게시물 검색에 대한 예시를 나타낸다.

4.2 정규화 여부에 따른 검색 성능 분석결과

유사 게시글 검색을 통한 불법저작물 탐지를 위해 수집된 원문에 정규화 처리를 하지 않은 데이터와 사전에 정규화 처리를 한 데이터간의 비교를 진행하였다. 분석대상 컨텐츠와 유사한 컨텐츠만 유사 게시글로 검색된 경우에 정탐하였다고 판단하였다. 분석대상 데이터의 경우 Train 데이터 셋에 존재하는 관계로 유사 게시글에 분석대상 데이터와 동일한 데이터가 포함되게 되는데 이를 포함하여 판단한 실험과 제거하여 판단한 실험, 두 가지 조건에서 결과를 산출하였다. 아래 표 5는 수집된 원문에 정규화 처리를 하지 않은 데이터에 대한 결과를 나타내고, 표 6은 정규화 처리를 진행한 데이터에 대한 결과를 나타낸다.

<표 4> Simhash기반 유사 게시물 검색 예시

Content_Name	SImhash	Similar Content Name	
남자친구.E15.190123.1080P	01001001011100110 00101101110101001 1010111111001000	〔남자친구〕 15화 .E15.190 123.1080p	남자친구.E15.190123.108 0P
6시 내고향.E6687.190102.720 p-NEXT	00000110001010011 11111001001110101 1010101100000101	6시 내고향.E6691.19010 8.720p-NEXT	6시 내고향.E6690.19010 7.720p-NEXT
신의 퀴즈-리부트.E13.190102. 720p-NEXT	00100110010011010 01111001011100011 1110101100010100	신의 퀴즈-리부트.E14.190 103.720p-NEXT	〔OCN〕 신의 퀴즈-리부트. E13.190102.720p-NEXT
〔머쓰마〕 남자친구.E09.19010 2.720p	01001110000010110 01101001000101001 0011101010000100	〔머쓰마〕 남자친구.E09.190 102.720p	〔tvN〕 남자친구.E09.1901 02.720p

<표 5> 정규화 처리를 하지 않은 데이터에 대한 불법저작물 탐지 결과

	Number of Data	Detection Rate(%)
Train 데이터에 분석대상 데이터 포함	300	86.33
Train 데이터에 분석대상 데이터 제거	184	77.72

<표 6> 정규화 처리한 데이터에 대한 불법저작물 탐지 결과

	Number of Data	Detection Rate(%)
Train 데이터에 분석대상 데이터 포함	300	94.33
Train 데이터에 분석대상 데이터 제거	157	90.23

표 5에서 보이는 것처럼 정규화 처리를 하지 않은 환경에서 불법저작물 탐지율은 77.72%로 낮은 탐지율을 보인 반면에 표 6은 정규화를 처리한 후 불법저작물을 탐지한 결과는 90.23%로 12.51%가 증가한 것으로 보인다. 정규화 처리후 진행한 실험에서 탐지하지 못한 5.67%에 해당하는 데이터를 살펴보면 '720P', '1080P', 'NEXT', 'HANrel' 등과 같이 영상의 화질이나 인코딩 닉네임을 나타내는 공통의 키워드가 포함되어 있는 것을 확인할 수 있다. 이는 변형된 제목을 정규화하는 작업과는 다른 관점에서 불법저작물 탐지를 혼탁하게 만드는 경향이 존재한다. 향후 연구에서 공통의 키워드 및 표현에 대해 사전에 조정한다면 검색 결과에 대한 개선이 가능할 것이다.

5. 결 론

토렌트나 P2P 사이트 및 웹하드를 사용자들은 편리하고 무료 혹은 저렴한 가격에 콘텐츠를 바로 다운로드 받을 수 있어서 많이 사용한다. 하지만 다운로드 받은 콘텐츠 파일은 대부분 저작권을 무시한 불법저작물이다. 하지만 끊임없이 쏟아지는 불법저작물들은 점점 더 필터링되는 과정을 우회하기 위하여 제목에 불필요한 특수문자나 공백을 삽입하거나 영숫자를 변형하거나 음절을 무시한 채 분리하거나 변형시키는 방법을 사용한다. 본 논문에서는 불법저작물들의 필터링 우회를 방지하기 위하여 제목 문장을 정규화하는 과정과 불법저작물을 탐지하는 방법을 제안하였다. 위 디스크라는 P2P 사이트에서 크롤링을 진행하였더니 이름, 날짜, 화질, 인코딩 닉네임을 포함한 제목이 대부분이었다. 제목만을 가지고 가장 먼저 불필요한 특수문자를 제거하였고, 영숫자 먼저 정규화를 진행하였다. 이후 한글 정규화를 하기 전에 공백을 제거 해주고 한글 정규화를 진행하였다. 한글 정규화는 변형된 한글 자음과 모음이 Dictionary1을 통해 하나씩 대치 될 때마다 한글 자음과 모음을 조합하여 음절로 변환해주는 Dictionary2와 영문에 영향을 미치지 못하게 하기 위하여 Dictrion1 역변환을 통해 총 3단계로 진행하였다. 또한, 여전히 한글 자음과 모음이 모두 변형되었거나 이중 모음을 갖거나 종성을 분리하여 표현한 정규화 되지 않은 문장이 존재하여 Dictionary3를 통해 변환한다. 최종 정규화된 문장을 이용하여 유사 콘텐츠 검색 시 제목을 정규화하기 전 탐지율은 77.72%이고, 정규화를 진행한 후 탐지율은 90.23%로 높은 탐지율을 보여주었다. 제목에 담고 있는 콘텐츠는 다르지만 '190808', '1080P', 'NEXT', 'HANrel' 등 일치할 수 있기 때문에 변형된 파일 제목을 더 이상 정규화를 통해 탐지율을 향상시키는 것은 불가능하다. 향후 연구에는 P2P 사이트 제목에 포함되는 '190808', '1080P', 'NEXT', 'HANrel' 등 독립적으로 화이트 리스트를 이용하거나 메타 데이터로 처리한다면 더 좋은 탐지율을 보일 것이다.

참고문헌

[1] [한국저작권보호원 보도자료] 한국저작권보호원, 2017년 불법복제물 유통 실태 발표.hwp

[2] 윤희돈, 조성환 “효과적인 웹툰 저작권 보호 방법에 관한 연구” 한국정보전자통신기술학회논문지(jkiiect)'19-2, Vol.12 No.1

[3] 김봉환 “파일 공유를 위한 P2P 어플리케이션 구조와 보안 위협” 한국콘텐츠학회지 7(1), 2009.3, 20-27(8 pages)

[4] 김병룡 “DHT 기반 P2P 네트워크에서 효과적인 다중 키워드 검색 기법 연구” 한국정보과학회 학술발표논문집 , 2014.6, 1236-1237(2 pages)

[5] 윤효근, 이상용 “협력적 필터링 기법을 이용한 P2P 모바일 에이전트 기반 사용자 컨텍스트 인식 및 서비스 처리 구조” 한국지능시스템학회 논문지 15(1), 2005.2, 104-109(6 pages)

[6] Changbin Lee, Kwangwoo Lee, Dongho Won and Seungjoo Kim “Weaknesses and Improvements of P2P File-sharing Filtering System”

[7] 강승식, 장두성, "SMS 변형된 문자열의 자동 오류 교정 시스템," 정보과학회논문지, 제35권, 제6호, 386-391쪽, 2008년 6월

[8] 강승식, "스팸 문자 필터링을 위한 변형된 한글 SMS 문장의 정규화 기법," 정보처리학회논문지, 제3권, 제7호, 271-276쪽, 2014년 7월

[9] 이현영, 강승식 “워드 임베딩과 딥러닝 기법을 이용한 SMS 문자 메시지 필터링” (No.NRF-2017M3C4A7068186)

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J., “Distributed Representations of Words and Phrases and their Compositionality,” In Advances in neural information processing systems, Lake Tahoe, the United States, pp.3111-3119, Dec. 2013

[11] Mikolov, Tomáš, et al., "Recurrent neural network based language model," Eleventh Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, pp. 1045-1048, Sep. 2010

[12] Mikolov, T., Yih, W. T., & Zweig, G., “Linguistic Regularities in Continuous Space Word Representations,” In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia , the United States, pp. 746-751, Jun. 2013

[13] M. S. Charikar, “Similarity estimation techniques from rounding algorithms,” in Proceedings of the 34th Annual ACM Symposium on Theory of Computing, pp. 380 - 388, ACM, New York, ,NY, USA, 2002

[14] DATAR, Mayur, et al. Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the twentieth annual symposium on Computational geometry. ACM, 2004. p. 253-262.

[15] Manku, Gurmeet Singh, Arvind Jain, and Anish Das Sarma. "Detecting near-duplicates for web crawling." Proceedings of the 16th international conference on World Wide Web. ACM, pp. 141-150, 2007.

〔저자소개〕

황 찬 웅 (Chan-woong Hwang)
2014년 3월~현재: 호서대학교 정보보호학과
<관심분야> 네트워크 보안, 악성코드 분석, 기계학습
E-mail: hwang85123@naver.com

하 지 희 (Ji-hee Ha)
2018년 2월: 호서대학교 정보보호학과 졸업
2018년 3월~현재: 호서대학교 정보보호학과 석사과정
<관심분야> 정보보호, 악성코드 분석, 암호학
E-mail: hjh1590@gamil.com

이 태 진 (Tea-jin Lee)
2003년 1월~2017년 2월: 한국인터넷진흥원 팀장
2017년 3월~현재: 호서대학교 컴퓨터정보공학부 교수
<관심분야> 시스템 보안, 악성코드 분석, 기계학습
E-mail: kinjecs0@gmail.com