JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 21, No. 1, pp. 165-172, Jan. 2020

Check for updates

블록체인 활용 웹콘텐츠 유사도 검색 플랫폼 개발

서 태 후[1] · 변 문 경[2*] · 이 홍 석[1]
[1]서울시립대학교 컴퓨터과학부 학사과정
[2]서울시립대학교 창업지원단 교육연구교수

Development of Web Content Similarity Search Platform using Blockchain

Tae-Hoo Seo[1] · Moonkyoung Byun[2*] · Hong-Seok Lee[1]
[1]Undergraduate course, Department of Computerscience, University of Seoul, Seoul, 02504, Korea
[2]Education & Research Professor, Startup Support Foundation, University of Seoul, Seoul, 02504, Korea

[요 약]

초연결 초지능 시대를 맞아 플랫폼의 성장은 최고의 콘텐츠를 유치하는 것과 맥락을 같이 하게 되었다. 흥행성 있는 컨텐츠는 방송사 매출의 주요 원천이며 OSMU(One Source Multi Use)가 활발하게 이루어지며 큰 부가가치를 창출하게 된다. 하지만 콘텐츠의 배포 속도가 빠르고 유통량도 많을수록 표절 문제, 유사한 모티브 차용 문제가 더 많이 발생한다. 특히 표절은 창작자의 창작동기를 저해하는 가장 큰 요인 중 하나이다. 하지만 저작권 등록 외에는 컨텐츠의 배포 과정에서 원작자를 보호하는 시스템이 없다. 따라서 배포 과정에서 우수한 컨텐츠가 표절되고, 모티브가 차용되는 것을 시스템적으로 막을 수 없는 실정이다. 따라서 본 연구에서는 블록체인을 사용하여 웹 컨텐츠의 무결성을 보장하는 플랫폼을 개발하여 창작자의 저작권을 보호하는 방법에 대해 창작자의 관점에서 탐색해 보았다. 이 플랫폼은 표절 문제에서 창작자들을 보호할 뿐만 아니라, 의도적 비의도적 표절을 예방하며, 각종 공모전 및 개인 저작 활동을 촉진하기 위해 널리 활용될 수 있을 것이다.

[Abstract]

In the age of hyper-connected super-intelligence, platform organizations are trying to attract the best contents of creators. One piece of outstanding content is the dominant source of broadcaster sales, and OSMU (One Source Multi Use) is also active. However, the more active distribution of content, the more plausible problems caused by plagiarism and the creation of similar works. Particularly, plagiarism is one of the biggest detrimental factors for creators. There is no system that protects original authors in the distribution process of content, so it is not possible to prevent excellent content from being plagiarism and motive borrowing in distribution process. Therefore, in this study, we developed a platform for verifying the similarity of web contents using blockchain and explored ways to protect the copyrights of creators. The platform will not only protect creators from plagiarism issues, but will also be widely used to prevent intentional and unintentional plagiarism, and to promote competitions and individual authoring activities.

색인어 : 표절 검색, 블록체인, 플랫폼 개발, 콘텐츠 창작자
Key word : Plagiarism Search, Blockchain, Platform Development, Content Creator

http://dx.doi.org/10.9728/dcs.2020.21.1.165

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-CommercialLicense(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 14 November 2019; **Revised** 16 December 2019
Accepted 23 January 2020

***Corresponding Author; Moonkyoung Byun**

Tel: ※ 개인정보 표시제한
E-mail: curiomoonlight@gmail.com

Ⅰ. 서 론

1-1 연구 목적

최근 웹콘텐츠 공유 플랫폼의 발달로 웹콘텐츠 유통 플랫폼은 지속적인 성장세에 있다[1]. 대표적인 국내 웹콘텐츠 유통 플랫폼은 네이버 시리즈, 카카오페이지, 문피아, 버프툰 등이다[2]. 이 중에서 네이버, 카카오페이지의 비중이 상대적으로 클 뿐, 각 플랫폼 기관들이 각자의 특성을 가지고 사용자 중심 서비스를 개선하며 성장하고 있는 추세이다[3], [4]. 양분화 되어 있는 국내와 달리 미국, 중국 등은 시장 논리에 의해 이미 독점 플랫폼이 구축되어 있다[4].

플랫폼 기관들은 판권을 보유하고 있는 오리지널 콘텐츠 IP를 글로벌 마켓을 겨냥해 OSMU(One Source Multi Use) 한다[5]. 특히 4차 산업 혁명의 초연결성으로 전 세계로 동시에 콘텐츠가 유통되면서, 글로벌 시장에서 기존보다 더 큰 수입을 창출하며 창작에만 집중하는 웹창작자들이 등장하게 되었다[6]. 웹콘텐츠 창작물에 대한 글로벌 마켓의 거대한 보상 시스템은 실력 있는 스토리텔러들을 웹콘텐츠 창작으로 대거 유입하고 있다[2], [4].

결국 실력있는 콘텐츠 창작자를 확보하고, 플랫폼의 핵심 콘텐츠인 오리지널 콘텐츠를 생산, 공급하는 것이 플랫폼의 성장 발전 전략이 되었다[7]. 실제 중국의 텐센트는 최고의 스타작가 영입 정책을 고수하면서 결국 독점 플랫폼으로 자리 잡을 수 있었다. 미국의 디즈니는 마블, 애니메이션 등의 독자적인 오리지널 콘텐츠 IP를 소유하며 디즈니 플러스라는 OTT(Over-The-Top) 서비스를 2020년 제공하게 된다. 텐센트, 디즈니의 성공사례를 거울삼아 국내의 모든 플랫폼 기관은 오리지널 콘텐츠 개발 및 유통에 집중하고 있는 실정이다[8]. 또한 신진 및 기성 작가의 고유 콘텐츠를 선 독점하기 위한 나름의 전략을 마련하고 있다. 예를 들어 고가의 상금을 내 건 공모전 및 OSMU(One source Multi Use) 정책으로 창작자에게 최고의 수익을 보장하는 정책들이 신진 작가 유치를 위해 활용되고 있다[9].

영상콘텐츠 유통 플랫폼 넷플릭스 측은 창작자에게 자유를 보장하는 것이 혁신적 콘텐츠 생산의 원동력이라고 설명했다[10]. 하지만 작가 유치를 위한 플랫폼 기관의 전략은 스타 작가에 대한 금전적 보상에 집중하고 있을 뿐, 보유 콘텐츠에 대한 저작권을 보호하는 시스템은 아직 미흡한 실정이다[9]. 김민영 한국콘텐츠 총괄디렉터도 “창작자들이 비전을 실현할 수 있도록 창작자의 자유를 중요하게 생각한다”며 “창작자가 충분히 자유를 구현할 수 있도록 지원하는 것이 넷플릭스의 궁극적 목표인 이용자를 즐겁게 하는 것과 일맥 상통한다”고 강조했다. 각자가 운영하는 플랫폼에서 작가 성공사례를 홍보하고, 프로 창작자들의 노하우와 인사이트를 신인작가 및 예비창작자에게 공유하여 플랫폼의 가치제고에 힘쓰고 있다[10], [11], [12].

플랫폼 기관들이 창작자들을 유입하기 위한 정책들은 정작 실제 창작자의 창의성을 강화하기 보다는 성장 및 운영 방식의 효율성에 집중하고 있다[9]. 웹출판 발전을 위한 향후 과제로 웹소설의 질적인 문제, 표절 문제, 해외 진출 문제 등은 해결해야 할 시급한 문제이다[4]. 특히 웹소설과 같은 웹콘텐츠의 경우 유사한 클리세의 반복, 캐릭터, 플롯 설정이 많아 표절 시비가 온, 오프라인에서 흔하게 제기되는 편이다[2].

저작권 등록 정책이 있긴 하지만 표절 시비는 당사자 간의 민사적인 손해배상시비로만 가려진다. 또한 웹소설의 경우 콘텐츠 유통 플랫폼들은 표절 시비가 발생한 작품에 대한 유통을 중단하고 저작자에 대한 인세결재를 미루는 행동을 취할 뿐, 근본적으로 원천 창작자를 보호하기 위한 조치는 취하지 않고 있는 실정이다. 플랫폼 기관에 가해지는 실질적인 금전적 불이익은 없기 때문이다[10].

또한 표절 작품의 창작자도 억울함을 호소한다. 의도적인 표절이 아니라, 초기에 모티브에 대한 보상을 했으나 절차가 체계적이지 않았고, 기억 속에 남아 있는 대사들을 모방하다보니, 자신도 모르게 표절시비에 휘말렸다는 것이다[13]. 이는 플랫폼 안에서 웹소설, 소설 및 영상화를 위한 시나리오의 출판 전 표절 검증 서비스가 미비한 까닭이다. 현재 논문 및 텍스트 표절검사에 최적화 된 플랫폼은 “카피킬러”만이 유일하게 존재한다. 그 외에 자기소개서와 같이 각자의 플랫폼의 요구에 따라 자체 데이터베이스를 활용한 표절확인 서비스도 있다. 하지만 가장 범용적으로 사용되고 있는 플랫폼은 “카피킬러”이다. 하지만 가장 범용적으로 사용되고 있는 플랫폼 “카피킬러”도 웹소설과 같은 웹콘텐츠의 유사도를 검사하기에 부적합한 요소들이 있다.

첫째, 비교 데이터가 드라마 대본, 웹소설 혹은 영화시나리오 등의 창작물이 아니기 때문에 유사도 검정 결과를 신뢰할 수 없다. 결국 유사도 비교 대상 자체가 다르다. 단 비교하려고 하는 시나리오를 먼저 등록하고 비교대상이 되는 시나리오를 등록하는 경우에는 예외가 될 것이다.

둘째, 기존 유사도 검색 플랫폼에서 웹소설, 영상 시나리오의 데이터베이스를 강화한다 하더라도 유사도 검증 방식이 부적합 할 수 있다. 웹소설의 경우 따옴표 안에 대사가 있다. 더구나 시나리오는 일반적인 글과 달리 대사, 지문, 인물 등의 구분이 더 명확하다. 특히 시나리오에서 지문의 경우 행동 지시문이다보니 반복적인 표현들이 많아 유사도가 높게 나올 가능성이 크다. 따라서 시나리오의 경우 “카피킬러” 와는 다른 비교 검색 알고리즘을 필요로 한다.

셋째, 창작자들이 자신의 콘텐츠의 유사도를 검증하고 싶어도 콘텐츠가 외부로 유출될 것에 대한 우려가 많다. “카피체커”는 지속적으로 방송용 대본이 수집되어야 플랫폼으로써 성장 발전할 수 있으므로, 사용자들이 안심하고 데이터를 업데이트 할 수 있는 장치가 필요하다. 하지만 “카피킬러”에 입력한 데이터를 보호하는 기술적인 장치는 마련되어 있지 않다.

이에 본 연구에서는 블록체인 기술을 적용한 웹콘텐츠 유사도 검증 플랫폼을 개발하기로 하였다. 블록체인은 네트워크 내의 모든 참여자가 공동으로 거래 정보를 검증·기록·보관할 수 있는 분산원장기술(DLT: Distributed Ledger Technology)

이다[14]. 블록체인 기술을 적용하면 블록체인 상에 콘텐츠가 저장되어 있어 위변조가 불가능하다. 보다 범용적으로 사용하게 되는 경우 불법 복제한 내용도 쉽게 추적도 가능해지기 때문에 저작권 보호가 수월하다는 사실을 이미 알려져 있다[14], [15]. 따라서 블록체인 기반 콘텐츠 서비스는 창작자들이 직접 자신의 콘텐츠를 작성하고 게시하거나, 사진, 동영상을 업로드할 때, SNS서비스, 뉴스 미디어, 음악 산업, 물류, 선거, 부동산 매매 등 다양한 분야에서 응용될 수 있다[14], [16].

따라서 본 연구팀이 개발한 “카피체커”에서도 블록체인을 이용하면 사용자들이 입력한 콘텐츠 데이터의 무결성(data integrity)을 보장할 수 있다. 데이터 무결성은 임의로 편집 불가능하게 데이터를 보호하고, 항상 정상인 데이터를 유지하는 것을 말한다. 데이터 무결성 보장을 위한 블록체인은 창작자들이 유출에 대한 걱정 없이 자신의 데이터를 업데이트할 수 있도록 도울 것이다.

본 연구자들은 이러한 문제 인식을 토대로 웹콘텐츠 창작자들의 권익을 보호하고, 의도적 또는 비의도적 표절을 예방할 수 있는 블록체인 기반 유사도 검증 플랫폼을 개발하고자 한다. 본 연구의 결과물인 “카피체커” 플랫폼이 기존 웹콘텐츠 플랫폼 안에서 긍정적인 웹콘텐츠 문화 형성에 기반이 될 뿐만 아니라, 창작자들의 비의도성 표절까지도 사전에 검증하고, 이후 웹콘텐츠 플랫폼 콘텐츠 관리, 공모전, 개인 저작 활동에 널리 활용될 수 있기를 기대해 본다.

1-2 연구 방법 및 범위

본 연구의 범위는 블록체인을 적용하여 웹소설, 시나리오의 유사도 검증 플랫폼 “카피체커1.0 버전”을 개발하는 것이다. 개발을 위한 주요 로직을 구성하기 위해서 “카피체커 1.0”버전에서의 표절 판단 기준을 먼저 설정하기로 했다. 한국학술단체 총연합회는 “타인의 출판 또는 미출판된 저작물의 일부를 출처 표기 없이 무단으로 사용한 모든 사례에 대해서는 표절”이라고 규정한 바 있다. “카피체커”에서는 대사 지문의 경우 출처를 표시했다 하더라도, 인용한 양과 내용이 정당한 범위를 넘어서 저작자의 고유한 가치로 인정받기 어려운 짜깁기, 말 바꾸기, 생각의 단위가 되는 주요 내용을 다른 사람의 저작물에서 사용한 경우를 표절이라고 규정하고 있었다. 하지만 이는 정량적으로 표절여부를 검증하기 위한 로직으로 설정할 수 없다.

이에 교육인적자원부의 한국학술진흥재단의 정책연구 보고서에 제시된 표절 규정 “주요 단어를 중심으로 여섯 단어 이상의 연쇄적인 표현이 남의 것과 일치하는 경우, 출처 표시 없이 그대로 가져다 쓴 경우”를 본 플랫폼 개발에 적용 하였다. 마지막으로 생성된 인덱스 파일들을 비교했을 때 연속된 3어절이 일치할 경우 표절로 판단하는 로직을 구성하기로 했다.

II. 연구방법

2- 1 . 표절 비교 데이터 수집 방법

본 연구는 유사도 검증 플랫폼을 개발하기 위해서, 우선 표절 비교 대상이 될 수 있는 데이터베이스를 구축하는 작업으로 시작하였다. 카피체커 1.0 버전 개발을 위해서, 우선 “대본창고”에서 무상공여를 약속한 드라마 작가들의 대본집을 한글파일 및 이미지 파일 형태로 구할 수 있었다. 또한 기 출판된 대본집을 구입하여 직접 엑셀에 입력하는 방식으로 10건의 드라마 대본을 추가 구축하였다. 이후 유사도 검정을 실시할 시나리오와 기 입력된 데이터베이스의 시나리오와 정량적 유사도 분석을 가능하게하기 위해서 시나리오의 텍스트를 분류하는 작업을 시작하였다. 먼저 텍스트에서 씬(Scene), 상황 대한 설명(Explanation), 인물(Actor), 대사(Script) 등을 먼저 구분했다. 구분된 데이터를 바탕으로 어절 테이블과 어절들의 위치 정보가 기록된 인덱스 파일을 생성할 수 있었다.

2-1 시나리오 파일 파싱

1) 상태 값(Status) 설정

표 1. 상태 값
Table 1. Status value

```
from enum import Enum


class StatusEnum(Enum):
    DEFAULT = 0
    SCENE = 1
    TIME = 2
    EXP = 3
    SCRIPT = 4
    NEWLINE = 5
    ACTOR = 6
    NOT_DEFINE = 7
```

텍스트를 좀 더 정확하게 구분하기 위해서는 직전 상태에 대한 정보가 필요했다. 예를 들어 상태 값을 설정하지 않는다고 가정했을 때, 대사가 두 라인 이상에 걸쳐서 나온다면 두 번째 라인부터는 지문으로 구분될 것이다. 하지만 각 라인에 대한 상태 값을 지정한다면 동일한 상황에서 이전 상태 값이 대사라는 것을 확인한다면 현재 라인은 지문이 아닌 이어지는 대사로 구분할 수 있다.

2) 씬 분류

‘#’, ‘S#’ 등과 같이 씬을 지칭하는 구분자들로 씬을 인식한다. 단 한 라인 전체에서 위와 같은 구분자를 탐색할 경우 지문이나, 대사에 ‘#’ 기호들이 있을 경우 이를 씬으로 인식할 수 있으므로, 공백을 제외한 앞에서 두 번째 문자까지만 탐색한다.

3) 인물 및 대사 분류

현재 라인이 인물 및 대사 라인이라는 것을 확인하기 위해 라인에 인물과 대사를 구분하는 구분자(‘/’, ‘:’, ‘\t’)가 있는지를 확인한다. 해당 구분자가 있다면, 구분자의 이전은 인물, 이후

는 대사로 분류할 수 있다. 또한 대사 중에 해당 구분자들이 나올 수 있으므로, 좀 더 정확도를 높이기 위해서는 공백을 제외한 앞에서부터 15문자까지만 구분자가 있는지 확인한다.

4) 지문 분류

앞서 분류했던 문장들 이외에는 지문으로 분류한다.

	juer	fileId_FK	sceneNo_FK	time	type	lineIndex	actor	script
		필터	필터			필터		필터
6884	6	6	93	밤	지문	727		나란히 누워있는 홍길이와 순아.
6885	6	6	93	밤	대사	463	순아	그런데 너는 어떻게 황해도에서 여기까지 내려왔니?
6886	6	6	93	밤	대사	464	홍길	국방군 아저씨들을 따라 왔다
6887	6	6	93	밤	대사	465	순아	국방군....
6888	6	6	93	밤	대사	466	홍길	참 좋은 아저씨였는데.... 내일은 다시 그 자리에 올
6889	6	6	93	밤	대사	467	순아	...
6890	6	6	95	낮	지문	728		먼저 홍길이 도망치던 길
6891	6	6	95	낮	지문	729		순아와 홍길이가 온다
6892	6	6	95	낮	대사	468	순아	참 좋은 아저씨구나
6893	6	6	95	낮	대사	469	홍길	그런데 난 비겁하게 혼자서 도망쳤어...
6894	6	6	95	낮	지문	730		하는데 비명-
6895	6	6	95	낮	지문	731		비명) 엄마 사람 살려요오-
6896	6	6	95	낮	대사	470	홍길	...
6897	6	6	96	null	지문	732		괴뢰군 서너명이 여자를 나무에 묶어놓고 희롱하고
6898	6	6	96	null	지문	733		탕-총질을 하는 괴뢰병
6899	6	6	96	null	지문	734		겁에 질리는 여자
6900	6	6	96	null	지문	735		홍길이와 순아가 와서 숨어서 본다
6901	6	6	96	null	지문	736		계속해서 총질을 하며 희롱하는 괴뢰군들
6902	6	6	96	null	지문	737		홍길이 눈이 분노에 이글거린다
6903	6	6	96	null	지문	738		여자는 이미 기절한 상태다
6904	6	6	96	null	지문	739		괴뢰군의 추행이 계속 된다
6905	6	6	96	null	지문	740		총을 들어진 홍길의 손이 떨린다
6906	6	6	96	null	지문	741		차마 정시하지 못하고 고개를 돌린다
6907	6	6	96	null	지문	742		소리없이 눈물을 흘리고 있는 순아의 얼굴이 뛰어든다
6908	6	6	96	null	지문	743		홍길 더 이상 참지 못하고 총을 들고 달려나간다
6909	6	6	96	null	지문	744		놀라는 순아

그림 1. 구문 분석 시나리오 테이블
Fig. 1. Parsed Scenario Table

2-2 어절(Word) 테이블 생성

분류된 텍스트를 한 개의 스페이스 문자를 단위로 구분한 다음, 이 어절들을 키로 관리할 수 있도록, key와 word 컬럼을 구성한다. 이후 분리된 어절이 테이블의 word에 존재하는지 확인하여 테이블에 존재하지 않는 어절이라면 이를 테이블에 등록하고 새로운 키(Key)를 생성하고 이 키를 가진다. 그리고 이미 존재하는 어절이라면, 키만 획득한다.

표 2. 문장 예시를 통한 단어 테이블 만들기
Table 2. Create word table for example sentences
예) 철수야 농구하러 가자

word key	word
65	철수야
47	농구하러
235	가자

2-3 키 데이터 파일 생성

한 문장의 단어들은 키들의 집합으로 나타낼 수 있다. 위와 같이 '철수야 농구하러 가자' 라는 문장은 [65, 47, 235] 의 집합으로 만들어진다. 하나의 시나리오의 대사와 지문들을 이런 키 집합들로 변환하고, 아래와 같이 JSON 형태로 생성한다.

표 3. 키 데이터 파일 / JSON
Table 3. Key Data File / JSON

```
{"data": [
        [65, 47, 235],
        [15, 37, 2352, 54, 21],
        [9436, 28124, 13675, 8514, 28125, 28126, 28127],
        [28128, 28129, 28130, 28131, 28132],
       ...(중략)
       ]
}
```

이러한 형태의 데이터 파일은 한 시나리오에 대해 두 개가 생성된다. 첫 번째 파일은 대사에 대한 데이터 파일이고, 다른 하나는 지문에 대한 데이터 파일이다. 지문과 대사를 구분하여 데이터 파일을 생성하기 때문에 이후 속성(지문, 대사)이 같은 파일끼리 비교할 수 있다.

2-4 어절의 인덱스(Index) 생성

테이블에 데이터를 삽입하는 과정에서 획득한 키와 해당 어절이 파일 내에서 몇 번째 라인에서 몇 번째 위치에 존재하는지 정보를 담은 인덱스 파일을 생성한다.

표 4. 문장에 대한 색인 맵핑 구조 예시
Table 4. Index mapping structure for example sentences

word key	[location, line]
65	[1,1]
47	[2,1]
235	[3,1]

마찬가지로 JSON 형태로 파일을 생성한다.

표 5. 인덱스 파일 예제 / JSON
Table 5. Example index file / JSON

```
{"index": {
      "1": [[1, 0]],
      "2": [[2, 0]],
      "3": [[3, 0]],
      "4": [[1, 3], [15, 2], [15, 4], [29, 2]],
      "5": [[1,4]],
      "6": [[1, 5], [250, 3], [267, 5]],
      "7": [[4, 0]],
      "8": [[4, 1]],
      "9": [[5, 2]],
      "10": [[6, 2]],
      ...(Omission)
      }
}
```

이후 유사도 검증을 수행 해야하는 시나리오의 데이터 파일에서 키를 가져와 비교 대상이 되는 기존 시나리오들의 인덱스

파일에서 위치들을 트래킹(tracking)해 나간다.

Ⅲ. 유사도 검증 방법

"카피체커"의 특성은 대사와 지문을 구분하여 유사도를 분석한다는 점이다. 시나리오의 경우 지문은 배우들에게 씬의 배경과 상황 등을 설명하는 목적을 가지고 있다. 대사는 생각을 행동으로 표현할 때의 의사전달의 목적을 가지고 있다. 이처럼 각 텍스트가 가지는 목적과 속성이 다르기 때문에 비교를 수행할 때 지문 영역과 대사 영역을 구분해야 좀 더 의미 있는 결과를 얻을 수 있다.

3-1 키 데이터 파일에서 인덱스 파일 트래킹

표절률 검사 대상인 파일이 업로드 되었을 때 생성되는 키 데이터 파일에서 키들을 가져와, 기존에 생성되어 있는 인덱스 리스트를 추적해 나간다. 표절을 판단하는 기준은 연속으로 세 단어가 일치했을 때로 설정해 두었다. 검사 대상인 파일의 키 데이터 파일이 아래[표 6]과 같다고 가정한다.

표 6. 일반적인 인덱스 파일 예시 / JSON
Table 6. Example of General Index / JSON

```
{"data": [
      [1, 2, 3, 4],
      [8, 2, 16, 4]
     ]
}
```

```
{"index": {
     "1": [[1, 1]],
     "2": [[2, 1], [4, 3]],
     "3": [[3, 1]],
     "4": [[4, 5], [15, 2], [15, 4], [29, 2]],
     "5": [[5,4]],
     "6": [[1, 5], [10, 3], [267, 5]],
     "7": [[4, 2]]
     }
}
```

먼저 검사 대상 시나리오의 데이터 파일에서 첫 번째 라인인 [1, 2, 3, 4]를 불러온 다음 첫 번째 요소인 1에 대해 확인할 것이다. 인덱스 파일에서 키가 1인 리스트를 불러와 위치를 확인하면 기존 파일의 [1, 1](1 번째 라인의 1 번째 위치)에 키 1이 존재한다는 것을 알 수 있다. 다음으로 다시 키 데이터 파일에서 두 번째 요소인 2가 그 다음 위치인 [2, 1]에 있는지 확인한다. 인덱스 파일에서 키가 2인 리스트를 확인했을 때 [2, 1] 이 존재하므로, 현재까지 두 단어의 배치가 동일하다는 것을 알 수 있다.

다시 데이터 파일에서 번째 요소인 3이 기존 파일의 [3, 1]의 위치에 있는지 확인한다. 마찬가지로 인덱스 파일에서 키가 3인 리스트에 [3,1]이 존재하므로, 세 단어의 순서가 일치한다는 것을 확인할 수 있다. 마지막으로 데이터 파일에서 첫 번째 라인의 마지막 요소인 4가 그 다음 위치인 [4, 1]에 위치하는지 확인한다. 인덱스 파일에서 키가 4인 리스트에 [4, 1]이 존재하지 않기 때문에 다음으로 넘어간다.

검사 대상인 파일에서 세 키의 배열([1, 2, 3])이 기존 파일에 이미 있기 때문에, 이 세 단어의 집합은 표절된 것으로 간주한다. 이 과정을 모두 반복한 다음 아래의 수식으로 표절률을 계산한다.

$$\frac{(\text{표절한 어절 수})}{(\text{전체 어절 수})} \times 100\,(\%)$$

여기서 한 문장이 3 단어 이하로 이루어 진 경우 표절 여부를 판단할 수 없기 때문에, 전체 어절 수는 한 라인이 3 단어 이상으로 이루어진 문장들에 대해 모든 단어 수의 합으로 계산한다.

한 파일에 대한 비교 과정이 끝날 때마다 메모리를 초기화하고, 비교 결과를 커밋(commit)한다. 구축된 "카피체커"에 드라마 도깨비 1, 2회를 업로드하고 표절률을 확인해 보았다. [표 3]처럼 드라마 도깨비 1회, 2회간의 표절률이 카피킬러에서는 52%로 나타나고, 카피체커 서비스에서는 3% 이내로 나타나므로, 분명한 차별화가 가능한 서비스임을 확인할 수 있었다.

그림 2. 샘플 시나리오 파일에 대한 표절 검사
Fig. 2. Plagiarism check for sample scenario file

업로드의 경우 현재 한글 파일에서 텍스트로 변환하는 방법을 고려해 보았으나, 오류가 너무 많아 최종적으로 업로더가 문서를 직접 txt 파일로 변환해서 올리는 방법을 채택하였다. 작가에게서 공여받은 파일에 있는 시나리오 대사는 (탭 : \t) 로 구분되어 있어서, "철수 \t 영희야 놀자!~"로 구성되어 있는 경우 '철수'라는 인물과 '영희야 놀자'라는 대사가 구분되어 업로드 된다. 또한 웹소설에서 인용부호(") 또는 이에 준하는 기호(')로 구분되어 있는 경우 모두 대사로 인식한다. 하지만 향후

업로드 과정에서 대사, 서사 구분에서의 오류를 최소화하기 위해서는 창작자들이 입력할 때부터 분류하여 입력할 수 있도록 화면을 설계하는 것이 필요하다. 또한 기존 한글, 워드 또는 이미지 파일에서 씬, 서사, 대사를 분리하는 알고리즘을 개발하여 파싱하는 기술을 개발한다면, 빅데이터를 구축하기 용이할 것이다. 웹페이지로 구축하여, [그림 4]와 같이 웹 상에서 확인이 가능하도록 하였다.

그림 3. 웹페이지에서 샘플 시나리오에 대한 표절 검사 결과
Fig. 3. Result of plagiarism check for the sample scenario on webpage

3-2 이더리움 네트워크에 커밋

카피체커는 [표 7]과 같이 파일에 대한 정보(파일의 이름, 시나리오 텍스트, 인덱스 파일, 타임스탬프 등)와 표절률에 대한 결과를 업로드한 유저의 개인키로 해시한 트랜잭션을 이더리움 네트워크에 커밋하는 방식으로 블록체인을 활용한 것이 특징이다.

표 7. 블록체인에서 해쉬 함수를 사용하는 예
Table 7. Example of using hash function in blockchain

Keccak256(fileName, fileText, indexFile, timeStamp, userPrivateKey)
Keccak256(fileName, timeStamp, plagiarismCheckResultFile, userPirvateKey)

이미 이더리움 네트워크에 커밋된 시나리오를 수정할 경우 수정 전 트랜잭션의 ID와 함께 수정된 시나리오가 커밋되며, 필요할 경우 트랜잭션을 역추적해 수정된 내역을 확인할 수 있다. 이와 같이 업로드 한 유저의 개인키로 본인에 대한 인증과, 표절률과 타임스탬프 정보가 함께 커밋된다. 때문에 이후 누구나 그 사람이 해당 시간에 시나리오를 커밋했다는 것을 확인할 수 있고, 이는 시나리오의 저작권에 대한 증명도 쉽게 가능하다는 것을 의미한다.

Ⅳ. 카피체커의 구현 및 결과

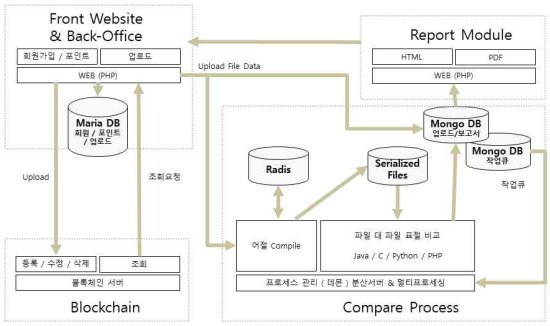

그림 4. 카피체커 시스템 구성도 v0.1. Cent OS 7.x / UTF-8
Fig. 4. Copychecker System Configuration Diagram v0.1. Cent OS 7.x / UTF-8

"카피체커"의 특성은 기존 표절 검사 시스템과 달리 대사와 지문을 구별하여 유사도를 분석한다는 점이다. 시나리오의 경우 지문은 배우들에게 씬의 배경과 상황 등을 설명하는 목적을 가지고 있다. 반면 대사는 생각을 행동으로 표현할 때의 의사전달의 목적을 가지고 있으므로, 기존 유사도 검증 플랫폼으로는 정확한 유사도 검사 데이터를 갖출 수 없다. 드라마 도깨비의 대본 1회, 2회간의 유사도가 카피킬러에서는 52%로 나타난다. 하지만 카피체커 서비스에서는 대사 지문, 인물 등이 분리되어 있어 해당 부분에 대해서 서로 유사도를 검증하지 않게 된다. 대사는 대사끼리, 지문은 지문끼리 대조하게 되어 유사도가 3% 이내로 나타나므로, 분명히 차별화 가능한 서비스임을 확인할 수 있었다.

Ⅴ. 결 론

본 연구는 "카피체커"의 개발 전 과정에 대한 사례 연구로, 데이터 파싱 후 블록체인을 사용하여 웹 컨텐츠의 유사성을 검증하는 플랫폼을 개발하고 제작자의 저작권을 보호하는 방법에 대해 창작자의 관점에서 탐색하며 서비스를 구축해 보았다. 기존 "카피킬러"가 논문, 보고서에 특화되어 있다면, "카피체커"는 대본, 지문을 분리해서 데이터베이스로 활용하여 유사도를 검증하는 차별성을 가지고 있다. 이처럼 방송용 시나리오에 최적화 된 유사도 검증 결과를 도출할 수 있다는 것이 본 연구를 통해 확인되었다.

본 연구팀은 "카피체커" 플랫폼을 더 발전시켜 창작자들의 기본적인 권리를 보장해 줄 수 있는 웹콘텐츠 플랫폼에 적용하려고 한다. 특히, "카피체커" 플랫폼을 통해서 표절률을 검색하게 되면, 개인이 보유한 시나리오를 업로드 하는 효과가 있어,

빅데이터를 구축할 수 있게 되어 표절률 검색의 효율성이 높아진다. 또한 개인키를 이용하여 블록체인에 저장하기 때문에 표절률의 결과에 대한 증명과 개인의 시나리오를 인증하는 효과가 있어 향후 서비스를 추가한다면, 방송사에서 주관하는 각종 공모전 및 개인 저작권 보호에 널리 활용될 수 있을 것이다.

블록체인 플랫폼을 개발하여 콘텐츠를 유통할 때 유사도 검증 서비스 뿐 아니라 콘텐츠 생태계의 유통 구조를 개선할 수도 있다. 이는 기존 서비스를 저해하지 않으면서 새로운 수익 구조를 창출해 내어 특정 플랫폼의 중개 수수료 독점 현상도 완화할 수 있다[14].

최근 흥행성 있는 컨텐츠는 방송사 매출의 주요 원천이며 OSMU (One Source Multi Use)도 활발하게 이루어지고 있다. 콘텐츠의 배포 속도가 빠르고 유통량도 많을수록 표절 문제, 유사한 모티브 차용 문제가 더 많이 발생하기 때문에, “카피체커”는 유통 과정에서 표절 문제를 해소하고, 컨텐츠의 배포 과정에서 원작자를 보호하는 시스템으로 활용할 수 있을 것이다. 이 플랫폼은 창작자들의 기본적인 권리를 보장해 줄 수 있는 웹콘텐츠 플랫폼에 적용 가능하고, 각종 공모전에 참여하려는 개인 저작 활동에 널리 활용될 수 있을 것이다.

참고문헌

[1] Hwang Hyun-chul (2019, October). business model makes the difference that leads to success, Startup Con X Next Contents Conference [internet] http://www.donga.com/news/article/all/20191017/97926886/1

[2] Yang Mi-sook, Ko Ji-wan, Byun Moonkyoung, “Explore 4C-level growth drivers for web content creators.” *Journal of the Digital Content Society of Korea,* Vol. 20, No. 3, pp. 469-480, 2019.

[3] Ahn Sang-Won, “A Study on the Changes of Platforms and Epistles by the Paying of Web Novels.” *Korean literary creation,* Vol. 16, No. 3, pp. 9-3, 2017.

[4] Lee Seung-hwan, Challenges of the development of web publishing. *South Korean publishing : Hagyeongu*, pp. 97-130, 2018.

[5] Kim Taek-hwan. *Web Content Big Bang*. Communication Books : Seoul, 2015.

[6] Blank, Grant. "Who creates content? Stratification and content creation on the Internet." *Information, Communication & Society*, Vol. 16, No. 4, pp. 590-612, 2013.

[7] Hwang Hyun-soo (2019, June). the future of K-Story dreamed by Kakao Page, *2019 SUMMER*, vol.3, p 42.

[Online] http://kocca.kr/cop/bbs/view/B0158947/1839536.do?searchCnd=&searchWrd=&cateTp1=&cateTp2=&useAt=&menuNo=203762&categorys=0&subcate=0&cateCode=&type=&instNo=0&questionTp=&uf_Setting=&recovery=&option1=&option2=&year=&categoryCOM062=&categoryCOM063=&categoryCOM208=&categoryInst=&morePage=&delCode=0&qtp=&pageIndex=1#

[8] Jeong Gil-joon (2019, September). Broadcasting company's Yonhap Audio Platform 'Tipod', [internet] http://www.viva100.com/main/view.php?key=20190925010008508

[9] Hwang Ji-yeon (2019, October). The crust of the platform, he web content creator's rice line with rope, [internet] http://www.snunews.com/news/articleView.html?idxno=17478

[10] Kim Hyung-won (2019, October). Kakao will hold an event to share know-how of professional creators. [internet] http://it.chosun.com/site/data/html_dir/2019/10/01/2019100101032.html

[11] Han Gwang-beom (2019, January). Netflix, which is targeting the Korean market with its contents and technology. "There is no increase in fares.". [internet] https://www.edaily.co.kr/news/read?newsId=04008166622361000&mediaCodeNo=257&OutLnkChk=Y

[12] Chae Min-sun (2019, October). share low price diversity...Content Business Core, Startup Con X Next Contents Conference 2019. [internet] http://www.bloter.net/archives/357793

[13] Jeong Yu-jin (2019, March). The Extreme Occupation Bae Se-young, said, "The next 11 films ...Drama Challenges in the Second Half." [internet] http://news1.kr/articles/?3583668

[14] Yang Danhee, Block Chain-based Contents Services through Decentralization of Trust Assurance, *Korean Institute of Information Technology*, Vol. 36, No. 12, pp. 35-40, 2018.

[15] Choi Hyun-jae (2017, November). Currently solves the copyright problem with 'blockchain technology', *Maeil Business Newspaper*, [internet] https://www.mk.co.kr/news/it/view/2017/11/762030/

[16] Kang Hee-jo, Various application services applying blockchain technology. *Proceedings of KIIT Conference*, pp. 545-547, 2018.

서태후 (Tae-Hoo Seo)

2014～현　재: 서울시립대학교 컴퓨터과학부 학부과정

2019년～현　재: 토스보험서비스(IT-Admin, 데이터 분석, 업무 자동화 및 MVP 개발)
관심분야 : 빅데이터 분석, 인공지능, 블록체인

변문경(Moonkyoung Byun)

2015 : 성균관대학교 대학원 (교육학석사)
2018 : 성균관대학교 대학원 교육공학 박사 (철학박사)

2016년～2018년: 성균관대학교 BK 플러스 교육인포매틱스 사업팀, 성균관대학교 산학협력단
2018년～2019년: 서울시립대학교 창업지원단 교육연구교수
관심분야 : 인공지능 Story Generation, Capstone Design, STEAM & Maker 교육, 빅데이터 분석, 블록체인, Creative Problem Solving,

이홍석 (Hong-Seok Lee)

2014～현　재: 서울시립대학교 컴퓨터과학부 학부과정

2018년～2019년: 디지털 서울 학생연구원(디지털 트윈기술 개발)
2019년 삼성전자 무선사업부 인턴(Android Framework 관련 Application 개발)
2020년 SK C&C Software Engineering
관심분야 : 블록체인, 인공지능