JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 21, No. 4, pp. 673-681, Apr. 2020

Check for updates

데이터 마이닝을 이용한 유튜브 인기 동영상 콘텐츠 분석

김 희 숙
광주과학기술원 대우교수

Analysis of Popular YouTube Video Content using Data Mining

Hye-Suk Kim

Lecture Professor, Gwangju Institute of Science and Technology, Gwangju Cheomdangwagi-ro 123, Korea

[요 약]

본 논문에서는 유튜브 인기 동영상 콘텐츠를 기반으로 데이터 마이닝을 실행하여 인기 동영상 요소들의 상관성을 분석하였다. 데이터 마이닝의 과정은 python에서 제공하는 라이브러리를 활용하여 데이터의 수집, 정제, 적재 그리고 분석 및 시각화의 단계로 처리하였다. 유튜브 인기 동영상 콘텐츠 기반의 데이터들을 수집하여 키워드, 조회 수, 동영상의 likes와 dislikes의 수, 댓글 수 등의 요소를 추출하여 데이터를 정제하였다. 정제된 요소들을 기반으로 유튜브 인기 메뉴에 등장하는 단어를 수집하여 빈도수를 분석한 결과, 매일 전해지는 주요 뉴스에 해당하는 동영상이 인기 메뉴 범주의 상위권을 차지하였으며 상관 분석 결과 양의 상관 관계를 보였다. 본 논문에서 구축된 빅데이터는 향후 유튜브 수익 예측을 위한 딥러닝 구현으로 확장 연구가 가능할 것이다.

[Abstract]

As the activities of creators based on Youtube are vigorously activating, video content market is also expanding in recently. In this paper, I analyze several popular YouTube contents using data mining technique, and figure out concepts and correlation going through those popular video elements. The process of data mining was done using the library provided by python. The data was collected, purified, loaded, analyzed, and visualized. First, I collected data based on popular video content on YouTube and then refined and clarified the data by extracting elements such as keywords, views, the number of likes and dislikes of videos, and the number of comments. After analyzing those refined factors by collecting the words repeatly appearing to popular videos and menus on Youtube, the result shows that the videos including the burning issues and news of the day occupy the top ranks at youtube platform. Also the result of correlation analysis, using python-based library on the loaded big data, shows positive correlation. In this paper, big data loaded by data mining based on popular YouTube video content is expected to be used for artificial intelligence service for predicting YouTube profit by executing deep learning based on regression analysis.

색인어 : 데이터 마이닝, 빅데이터, 유튜브, 상관 분석, 회귀 분석, 딥러닝

Key word : Data Mining, Big Data, YouTube, Correlation Analysis, Regression Analysis, Deep Learning

http://dx.doi.org/10.9728/dcs.2020.21.4.673

Received 23 February 2020; **Revised** 15 April 2020
Accepted 25 April 2020

***Corresponding Author; Hye-suk Kim**

Tel: +82-62-530-0147
E-mail: iamtinakhs@naver.com

Ⅰ. 서 론

4차 산업 혁명과 더불어 일상에서도 변화가 시작되고 있다. 산업 현장 및 가정에서도 인공지능(Artificial Intelligence)에 관심이 높아지고 있으며 스마트폰이나 자동차는 물론이고 TV, 냉장고, 정수기, 화분 등과 같이 일상생활에서 쉽게 접하고 있는 다양한 제품에도 인공지능 서비스가 장착되어 삶의 질이 향상되고 있다.

인공지능과 더불어 현재 1인 미디어 분야에서도 양질의 미디어 콘텐츠 제작 환경이 제공되고 있으며 콘텐츠 시장의 확대로 인하여 1인 크리에이터(Creator)는 개인의 브랜드 가치를 얻는 것은 물론이고 수익을 창출하기 시작하였다. 아날로그 신문이나 음반 시장이 축소된 사이 동영상 뉴스와 동영상 음악, 동영상 방송 등의 미디어 시장은 나날이 확대되고 있다. 그 이유 중 하나는 스마트폰의 대중화에서 찾을 수 있으며 우리의 일상에서 정적인 이미지보다 동적인 동영상의 편리함을 느끼고 있기 때문이다[1]. 유튜브(YouTube)의 경우 인터넷 이용자가 가장 선호하는 미디어 플랫폼이자 많은 사람들이 이용하는 검색 플랫폼이다. 이러한 유튜브 플랫폼이 1인 미디어 매체로 활용되면서 1인 크리에이터 인구 또한 급속하게 증가하며 새로운 사회적 수익 구조로 고착화되고 있는 상황이다. 유튜브에 업로드된 동영상 콘텐츠들은 언어적, 지리적, 문화적 한계를 뛰어넘어 전 세계 시청자들의 관심을 받고 있으며 이로 인하여 기존 미디어 시장과 콘텐츠 산업이 재편성되고 있다. 유튜브에서는 동영상에 삽입된 광고를 통하여 광고 상품의 브랜드 인지도, 구매 예상 확률 등을 확인해 볼 수 있으며 이는 경제의 흐름에도 영향을 미치고 있다[2].

1인 크리에이터로 활동하는 유튜버(YouTuber)의 경우 유튜브에서 수익을 올리기 위해서는 파트너 프로그램(Partner Program)에 신청하고 참여를 승인받으면 수익을 창출할 수 있는 자격을 획득하게 된다. 광고 수익, 채널 멤버십, 상품 라이브러리, Super Chat 및 Super Sticker, YouTube Premium 수익과 같은 요소 등이 수익 창출 요소로 활용되고 있으며 이러한 요소들을 기반으로 유튜브 자체 알고리즘에 의해 수익 모델이 생성되고 있다. 유튜브 자체 알고리즘의 경우 구독자 수에 기반하여 인기 동영상 순위 혹은 댓글 반응, 2차 채널 확산 유무, 유튜브 내에서 브랜드 쿼리(Query) 상승 여부 등을 체크하여 수익 구조에 반영하고 있다. 이와 같은 이유로 현재 유튜브에서 수익을 창출하기 위해서는 업로드된 동영상이 인기 메뉴에 포함되도록 하는 것이 급선무가 될 수 있다[3]. 이렇게 인기 메뉴에 포함된 동영상 콘텐츠의 경우 조회수 및 구독자 수를 높일 수 있는 지름길이 될 수 있을 것이다. 유튜브 인기 메뉴의 경우 조회수는 방문자 및 방문을 망설이는 예비 방문자에게 관심을 유도할 수 있는 변수가 될 수 있으며 여러 요소들과 맞물려 유튜브에서 수익을 창출할 수 있는 요인이 된다. 유튜브에서 조회수와 연관되는 변수들을 예측하는 것은 1인 미디어 시장에서 경쟁력을 높일 수 있는 요인이 될 것이다.

본 논문에서는 인기 메뉴 순위에 포함된 유튜브 동영상 콘텐츠는 유튜브 수익 창출 모델에서 중요한 요소에 해당하는 구독자수와 조회수를 높이는데 영향을 줄 수 있다는 가설을 설정한다. 설정된 가설을 기반으로 연관 변수들을 선택하여 데이터의 상관성을 분석하고 인공지능 딥러닝 알고리즘을 적용하여 학습시키는 것은 향후 유튜브에서 창출되는 수익성 예측에 도움이 될 수 있을 것이다. 이를 위하여 현재 유튜브 수익에 영향을 주고 있는 인기 메뉴에 해당하는 유튜브 동영상 콘텐츠의 조회수, 이 동영상이 마음에 듭니다, 이 동영상이 마음에 들지 않습니다, 댓글 수 등을 대상으로 데이터를 수집하고 분석하여 향후 유튜브 크리에이터들의 수익성 예측을 위한 인공지능 서비스 구현의 기반을 마련하는 것은 의미가 있을 것이다.

본 논문에서는 유튜브 내에서 인기 동영상 콘텐츠에 포함될 수 있는 요인들을 고려하여 필요한 데이터들을 수집하고 상관성을 분석하여 딥러닝을 이용한 인공지능 서비스 구현의 초석을 마련하고자 한다. 논문의 구성은 다음과 같다. 2장에서는 빅데이터 수집 및 분석 방식을 이용한 데이터 마이닝(mining)에 관한 기존 연구를 소개하고 3장에서는 본 논문에서 제안된 유튜브 인기 동영상 콘텐츠 구성 요소의 상관성 분석을 이용한 데이터 마이닝 과정을 설명한다. 그리고 4장에서는 데이터 마이닝 과정으로 적제된 빅데이터를 기반으로 정성적 및 정량적 데이터 분석을 실시한 결과를 평가한다. 마지막으로 5장에서는 결론 및 향후 연구 방향에 관해 서술한다.

Ⅱ. 관련 연구

네트워크 속도의 향상과 메모리 가격의 하락으로 텍스트 뿐만 아니라 다양한 멀티미디어 데이터까지 데이터베이스에 저장되면서 데이터의 크기는 방대해지고 있다. 빅데이터는 정보와 결합된 다양한 속성에 대하여 실시간으로 통합 분석을 수행하면서 의미있는 정보를 추출하여 금융 거래, 로그 파일 또는 네트워크 트래픽 등을 분석하여 이상하거나 의심스러운 활동을 식별하는 데 사용될 수 있으며 미래의 사고에 대응할 수 있는 기술까지 함축하고 있다[4].

데이터 마이닝이란 대용량의 자료로부터 쉽게 드러나지 않는 유용한 정보들을 찾아내는 과정을 말하며 이러한 데이터 분석 과정을 통해 유의미한 패턴과 규칙을 찾아내는 기술[5]-[6]을 의미한다. 빅데이터 기반의 데이터 마이닝은 기본적으로 3Vs의 특성에서 출발하였지만 현재 5Vs, 7Vs 및 9Vs로 확장되고 있다. [7]과 [8]에서는 빅데이터 마이닝의 7Vs 특성을 다음과 같이 분석하였다.

(1) 크기(volume) : 빅데이터의 물리적 크기는 일반적으로 테라바이트(terabyte), 페타바이트(petabyte), 엑사바이트(exabyte) 등을 의미한다.

(2) 다양성(variety) : 빅데이터는 정형, 반정형, 비정형 데이터 등의 다양한 종류의 데이터를 수용해야 한다. 정형 데이터란

데이터베이스와 같은 고정형 필드에 저장되는 데이터를 의미한다. 반정형 데이터는 XML이나 HTML 같이 메타데이터나 스키마 등을 포함하는 데이터를 의미하며 비정형 데이터는 동영상, SNS 메시지, 사진, 오디오 등 고정된 형태가 없는 데이터를 의미한다. 빅데이터를 처리하기 위해서는 손글씨 데이터, 음성 데이터, 사진 데이터 등의 비정형 포맷의 데이터까지를 처리할 수 있는 기술이 포함되어 있음을 의미한다.

(3) 속도(velocity) : 데이터를 빠르게 처리하고 분석할 수 있는 능력을 빅데이터의 속도라고 정의할 수 있다. 즉, 데이터를 얼마나 빨리 생산하고 처리하는지를 의미한다. 오늘날 빅데이터에서 직면한 주요 과제는 가능한 한 가장 빠른 속도로 필요한 데이터를 찾아서 분석할 수 있어야 함을 의미한다[9].

(4) 신뢰성(veracity) : 수집된 데이터에 부여하는 신뢰 수준을 의미한다. 수집된 데이터에 편견이나 오류 또는 거짓이 있을 경우 데이터의 진실성을 판단할 수 있는 지식을 갖추거나 관련 지식을 갖는 전문가와 협업을 통해 신뢰할 수 있는 데이터만 선별하여 빅데이터를 구성해야 함을 뜻한다.

(5) 가치(value) : 빅데이터는 대량의 데이터에서 값을 추출한 결과가 서로 낮은 밀도를 갖는 데이터 값이어야 유의미한 가치를 가지는 영향력에 관한 지표가 될 수 있음을 의미한다.

(6) 변동성(variability) : 빠른 데이터 처리 속도와 더불어 데이터의 흐름이 지나치게 급변하는 경우가 발생하면 데이터의 의미 있는 분석을 수행하기 위해서 이상 유무에 관한 탐지 방법이 필요함을 뜻한다.

(7) 휘발성 (volatility) : 빅데이터는 장기적인 관점에서 유용한 가치를 창출할 수 있어야 하므로 수집된 데이터의 양이 방대하고 의미있는 정보를 포함하고 있었더라도 시간이 지나면서 가치를 상실하거나 데이터의 크기가 물리적 저장 장치에 보관하기 어려울 경우에는 빅데이터로서의 수명과 활용성을 점검해야 함을 의미한다.

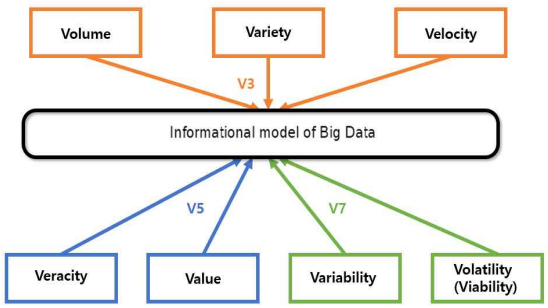


그림 1. 빅데이터의 7Vs
Fig. 1. 7Vs in Big Data

[8], [10]에서는 빅데이터 마이닝의 특성 7Vs가 확장된 9Vs를 표 1과 같이 분석하였고, 9Vs를 데이터 수집, 데이터 처리, 무결성 데이터, 데이터 시각화 및 데이터 가치의 5가지 클러스터 그룹의 범주로 그림 2와 같이 빅데이터 처리 과정을 구분하였다.

표 1. 빅데이터 마이닝 기법
Table 1. Big Data Mining Techniques

Parameter Name	Description
Veracity	It is referred to the biases, noise, and abnormality in data.
Variety	Various types of data like structured, semi-structured and unstructured.
Velocity	It is referred to, how fast data is to be produced and processed to meet the demand.
Volume	Volume is a size of data such as terabyte, petabyte, exabyte, zettabyte etc.
Validity	The data is correct and accurate for the intended.
Variability	Along with the velocity, the data flows may be highly inconsistent with the data. The need to be found by anomaly and outlier detection methods in order for any meaningful analytics to occur.
Volatility	Recall the retention policy of structured data that the implement every day in the businesses.
Visualization	It makes all that huge amount of data comprehensible and easy to understand and read.
Value	It has a low-value density as a result of extracting value from massive data.

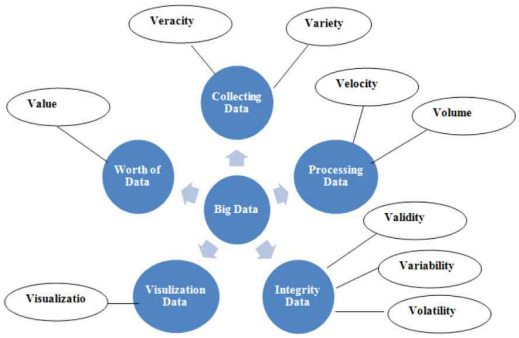


그림 2. 5가지 범주로 분류한 빅데이터 특성
Fig. 2. Big data characteristics classified into five categories

III. 인기 유튜브 동영상 기반의 데이터 마이닝

본 논문에서는 테이터 마이닝의 9Vs특성을 고려하여 그림 3과 같이 수집, 정제, 적재, 분석, 시각화와 같은 순서로 데이터 마이닝 과정을 처리한다. 수집된 비정형 형식의 데이터 중에서 불필요한 부분은 제거하는 정제 과정을 거쳐서 정형화된 구조의 데이터베이스에 저장하고 적재하는 단계를 수행한다. 그리고 적재한 정형화된 데이터셋을 대상으로 분석하고 그 결과를 도표 및 그래프로 시각화하는 처리를 수행한다.

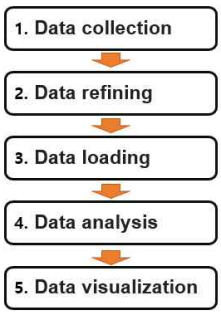


그림 3. 빅데이터 마이닝 처리 단계
Fig. 3. Big Data Mining Processing Steps

3-1 빅데이터 수집

유튜브의 인기 동영상 콘텐츠를 이용하여 제공되는 데이터를 수집하고 상관성을 분석하여 유튜브 인기 메뉴 콘텐츠 진입에 필요한 빅데이터셋을 구축하기 위하여 본 논문에서는 그림 4와 같이 6단계의 처리 과정을 수행하여 관련 빅데이터를 수집하였다.

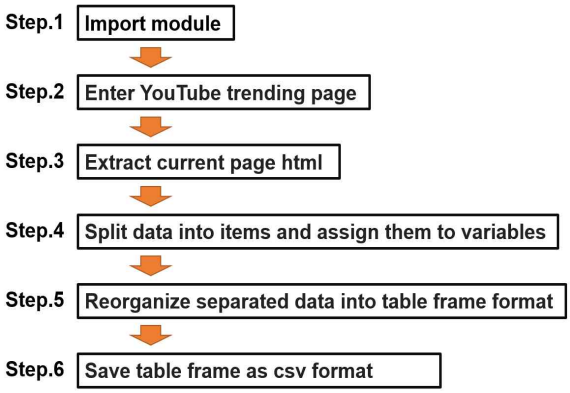


그림 4. 유튜브 수익성 예측에 필요한 빅데이터 수집 과정
Fig. 4. Big data collection process for predicting YouTube profitability

1단계에서는 bs4의 BeautifulSoup과 selenium의 webdriver 그리고 유튜브에서 제공하는 API(Application Programming Interface)를 이용하여 데이터 크롤링에 활용한다. 2단계에서는 webdriver.Chrome(chromedriver)를 활용하여 유튜브의 인기 메뉴 동영상 콘텐츠(https://www.youtube.com/feed/trending)로 진입한다. 3단계는 soup.find('div',id="grid-container") 태그를 중심으로 인기 메뉴 동영상 콘텐츠 페이지에서 필요한 데이터를 추출한다. 4단계에서는 추출된 데이터를 순번, 제목, 요약, 조회수, 닉네임 항목으로 분리하고, 각 분리된 데이터를 해당 변수에 할당하여 정제한 후 정형화한다. 그림 5는 추출된 데이터를 항목별로 분리한 결과 중에서 10번과 11번에 해당하는 데이터셋을 보여준다.

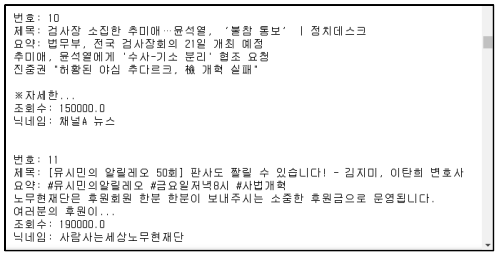
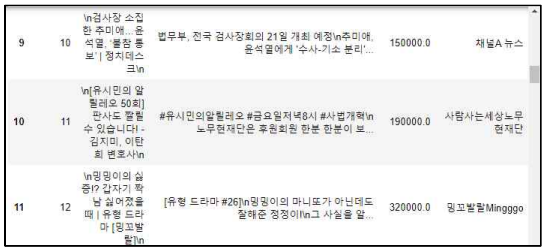


그림 5. 추출된 데이터를 항목별로 분리한 데이터셋
Fig. 5. Dataset of results of dividing extracted data by item

5단계에서는 pandas의 DataFrame 라이브러리를 이용하여 항목별로 분리된 데이터를 테이블 형식의 데이터 프레임으로 재구성한다. 아래의 그림은 데이터 프레임으로 구조화된 데이터셋을 보여준다.



9	10	\n검사장 소집한 추미애…윤석열, '불참 통보' | 정치데스크\n	법무부, 전국 검사장회의 21일 개최 예정\n추미애, 윤석열에게 '수사-기소 분리'...	150000.0	채널A 뉴스
10	11	\n[유시민의 알릴레오 50회] 판사도 짤릴 수 있습니다! - 김지미, 이탄희 변호사\n	#유시민의알릴레오 #금요일저녁8시 #사법개혁\n노무현재단은 후원회원 한분 한분이 보...	190000.0	사람사는세상노무현재단
11	12	\n밍꼬의 심중(?) 갑자기 확 남 싫어졌을 때 | 유형 드라마 | 밍꼬발랄\n	[유형 드라마 #26]\n밍꼬의 마니또가 아닌데도 잘해준 정정이!\n그 사실을 알...	320000.0	밍꼬발랄Mingggo

그림 6. 테이블 프레임으로 재구성된 데이터셋
Fig. 6. Dataset reorganized into table frames

마지막 6단계에서는 재구성된 테이블 프레임 형식의 데이터셋을 utf-8-sig 인코딩 기반의 csv(comma separated values)파일로 저장한다.

위와 같이 유튜브 인기 동영상 콘텐츠를 기반으로 본 논문에서 제안한 6가지 단계를 이용하여 2020년 1월 17일부터 2020년 2월 15일까지 총 30일 동안 18.3GB의 정형화된 형식의 빅데이터를 수집하였다.

3-2 빅데이터 기반의 정성적 분석

정성적 데이터 분석에 이용되는 데이터는 자연어에 의한 서술로 표현되는 범주형 데이터로 본 논문에서는 이를 텍스트 데이터로 간주하여 분석한다. 테스트 데이터 분석의 경우 키워드 중심의 빈도수 분석을 다음과 같이 4 단계로 수행한다.

(1) 1단계 : 수집된 텍스트 데이터셋을 기반으로 Konlpy 및 NLTK(Natural Language Toolkit) 라이브러리를 활용하여 형태소 분석을 실행한다.

(2) 2단계 : okt.nouns 라이브러리를 활용하여 명사를 중심으로 키워드를 추출한다.

(3) 3단계 : 불용어 사전을 준비하여 추출된 키워드를 대상으로 불용어를 제거한다.

(4) 4단계 : collections 라이브러리를 활용하여 불용어가 제거된 키워드의 빈도수를 출력한다.

위와 같이 4단계의 과정으로 각 날짜 별로 수집된 데이터를 대상으로 키워드 중심의 빈도수를 산출하였다. 2020년 2월 13일 날짜로 수집한 610.8MB 분량의 데이터셋에서 추출된 키워드를 분석한 결과 빈도수 높은 10개의 키워드로 갤럭시, 영상, 방송, 구독, 플립, 삼성, 미국, 한국, 스타, 수상 등의 순서로 출력되었다. 그 결과를 워드 클라우드(Word Cloud)[11]로 시각화하면 다음 그림 7과 같다.

그림 7. 2020년 2월 13일 날짜로 수집된 데이터셋의 워드클라우드 시각화
Fig. 7. Wordcloud visualization of the dataset collected on February 15, 2020

아래 그림은 2020년 2월 13일 날짜로 수집한 610.8MB 분량의 유튜브 인기 동영상 콘텐츠 데이터셋에서 키워드 빈도수를 그래프로 표현하면 그림 8과 같다.

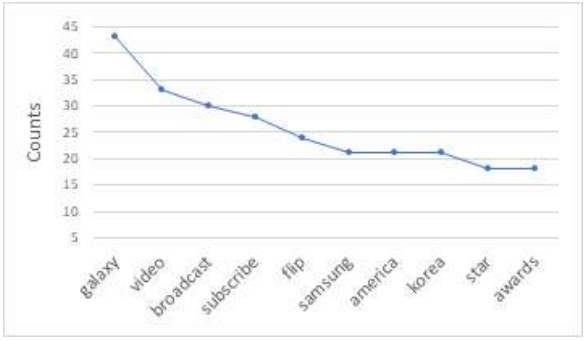


그림 8. 키워드 빈도수 그래프
Fig. 8. Graph of keyword frequency

2020년 2월 13일 삼성전자는 톰브라운과 협업한 프리미엄 패키지 '갤럭시 Z 플립 톰브라운 에디션(Galaxy Z Flip Thom Browne Edition)'을 21일부터 판매할 계획을 발표하자 유튜브의 인기 메뉴로 급상승함은 물론이고 가장 높은 빈도수의 키워드로 분석되었다. 92회 아카데미 시상식이 개최된 2020년 2월 9일 날짜의 유튜브에서는 작품상, 감독상, 국제영화상, 각본상을 수상한 영화 기생충과 봉준호 감독이 가장 높은 빈도수의 키워드로 분석되었다. 2020년 2월 9일 날짜에 최고 빈도수를 차지한 기생충과 봉준호 감독에 해당하는 키워드는 5일이 지난 후에도 여전히 인기 키워드의 범주 안에 머물고 있음이 분석되고 있다.

주요 뉴스와 연관된 키워드가 포털 사이트의 인기 검색어 상위 순위에 올라오는 것처럼 유튜브의 인기 동영상 콘텐츠의 인기 요소가 주요 뉴스와 관련성이 높음을 알 수 있었다.

3-3 빅데이터 기반의 정량적 분석

정량적 데이터란 숫자로 표현되는 수치 데이터를 의미하며 본 논문에서는 유튜브의 인기 동영상 콘텐츠를 기반으로 구조화된 csv 형식의 데이터셋 파일에서 조회 수, 이 동영상이 마음에 듭니다, 이 동영상이 마음에 들지 않습니다, 댓글 수에 해당하는 4가지 변수를 기반으로 숫자 데이터 항목들 사이의 상관성을 분석한다. 이러한 4가지 항목의 데이터 분석을 위하여 python 기반의 pandas, matplotlib, seaborn[12] 라이브러리를 사용한다. 그림 9는 유튜브 인기 메뉴를 대상으로 2020년 2월 13일 날짜로 수집된 610.8kb 분량의 데이터셋을 분석하기 위하여 위에서 정의한 4가지 항목을 기반으로 재구성한 데이터셋을 보여준다. 그림에서 view는 조회 수, likes는 이 동영상이 마음에 듭니다, dislikes는 이 동영상이 마음에 들지 않습니다 그리고 comment는 댓글 수를 의미한다.

	view	likes	dislikes	comment
1	241880	4815	103	1237
2	437013	12674	112	2075
3	344813	12621	367	1521
4	637799	9092	1372	1978
5	112700	15927	219	1081
⋮				

그림 9. 정량적 분석을 위한 데이터셋의 재구성
Fig. 9. Reconstruction of Datasets for Quantitative Analysis

2020년 2월 13일 날짜로 수집된 1730 rows × 4 columns로 구성된 610.8MB 분량의 데이터셋을 기반으로 4개의 변수를 대상으로 pandas의 describe 라이브러리를 이용하여 샘플 수(count), 평균(mean), 표준편차(std), 최소값(min), 백분위 수 25%, 50%, 75%에 해당하는 값, 최대값(max)에 관한 통계 처리 결과는 아래의 그림과 같다.

	view	likes	dislikes	comment
count	1.730000e+02	173.000000	173.000000	173.000000
mean	3.270994e+05	9076.982659	247.612717	1158.312139
std	4.932928e+05	14629.945929	427.293518	1666.906408
min	1.841600e+04	0.000000	0.000000	0.000000
25%	8.642600e+04	1792.000000	50.000000	224.000000
50%	1.859920e+05	4815.000000	104.000000	571.000000
75%	3.228670e+05	11379.000000	259.000000	1318.000000
max	3.997128e+06	144030.000000	3427.000000	10486.000000

그림 10. describe 라이브러리를 이용한 통계 처리 요약
Fig. 10. Summary of Statistics Processing with the describe Library

위에서 정의한 4개의 변수를 기반으로 matplotlib, seaborn 라이브러리를 이용하여 피어슨 상관 분석(pearson correlation analysis)[13] -[14]을 수행한 결과는 그림 11과 같다.

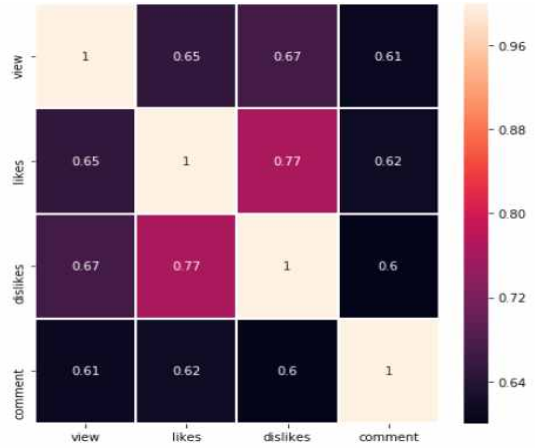


그림 11. 변수들의 피어슨 상관 분석
Fig. 11. Pearson Correlation Analysis of Variables

상관 분석(correlation analysis)이란 다변량 통계분석에서 변수들 사이의 상관 계수를 측정하여 유의한 선형적인 상관 관계가 있는지 파악하는 것을 의미한다[15]. 상관 분석에 이용되는 공분산(covariance)은 두 확률변수 사이의 선형적인 상관 정도를 나타내는 정량적인 값으로 $(-\infty, \infty)$ 범위의 값을 갖는다. 공분산 $Cov(X, Y) > 0$ 는 X 가 증가할 때 Y 도 증가하는 선형적인 양의 상관 관계(positive correlation)를, $Cov(X, Y) < 0$ 는 X 가 증가할 때 Y 는 감소하는 선형적인 음의 상관 관계(negative correlation)를, $Cov(X, Y) = 0$ 과 같이 공분산이 0 이라면 두 변수간에는 선형 상관 관계가 없음을 의미한다. 공분산의 경우 상관 관계의 흐름을 파악할 수는 있으나 두 변수의 측정 단위 크기가 상이할 경우 상관도를 파악하기에는 적절하지 않다. 그러므로 이를 정규화하여 특정 범위 $(-1, 1)$ 범위 내에서 출력되도록 피어슨 상관 계수를 활용하는 것이 바람직하다[16]. 피어슨 상관 계수에 사용되는 표본 공분산(sample covariance)은 식 (1)과 같이 정의되고, 피어슨 상관 계수는 식 (2)와 같이 표본 공분산을 표준편차의 곱으로 나눈 값으로 계산한다.

$$Cov(X, Y) = \frac{\Sigma_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \quad (1)$$

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{\frac{\Sigma_i^n (X_i - \overline{X})^2}{n-1}} \sqrt{\frac{\Sigma_i^n (X_i - \overline{X})^2}{n-1}}} \quad (2)$$

식 (2)는 $-1 \le Corr(X, Y) \le 1$ 성질을 만족하며 피어슨 상관 계수의 값이 1에 가까울수록 두 확률변수는 양의 상관 관계를, -1에 가까울수록 음의 상관 관계를 지닌다.

본 논문에서 구성한 데이터셋을 기반으로 피어슨 상관 분석을 수행한 결과 그림 11에서 제시한 바와 같이 4개의 변수들 사이의 상관 계수 값으로 0.6 ~ 0.77의 지표를 얻었다. 이는 4개의 변수들이 양의 선형적인 상관 관계가 인정되고 있으며 비교적 긴밀하게 연관되어 있다는 것을 보여주고 있다. 특히 0.77의 상관 계수를 출력하고 있는 likes와 dislikes는 4개의 변수들 중에서 가장 높은 상관성이 나타나고 있다.

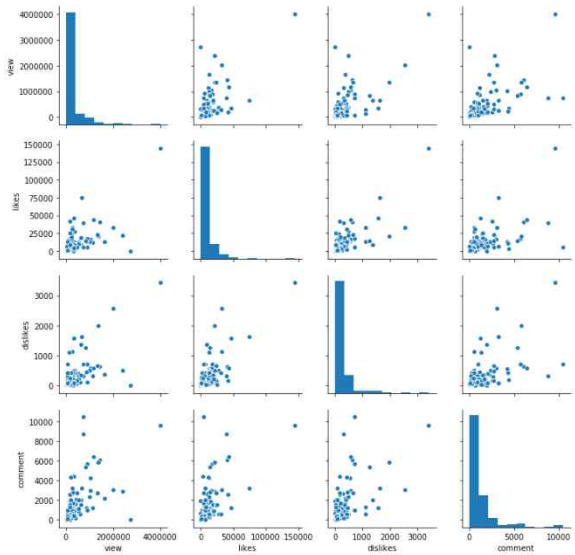


그림 12. 변수들의 상관 관계 시각화
Fig. 12. Visualize the Correlation of Variables

그림 12는 위에서 정의한 4가지 변수를 기반으로 seaborn의 pairplot 라이브러리를 이용하여 데이터 프레임을 인수로 받아 그리드(grid) 형태로 각 데이터 열의 조합에 대해 산점도(scatter plot)를 출력하고 데이터가 만나는 대각선 영역에는 해당 데이터의 히스토그램으로 시각화한 결과를 보여준다. 산점도에서 보여준 바와 같이 likes와 dislikes는 비교적 양의 선형적인 상관 관계가 인정되고 있음을 나타난다.

본 논문에서는 수집된 4개의 변수들을 대상으로 독립 변수와 종속 변수의 인과 관계를 정량화하기 하기 위하여 선형 회귀 분석(linear regression analysis)[17]을 병행하였다. 회귀 분석에서 두 변수의 인과 관계를 분석하기 위해서 사용된 독립 변수(independent variable)는 영향을 주는 변수이고, 종속 변수(dependent variable)는 영향을 받는 변수이다. 이 경우 종속 변수가 1개인 분석을 단변량 회귀 분석(univariate regression analysis), 종속 변수가 2개 이상이 경우의 분석을 다변량 회귀 분석(multivariate regression analysis)이라고 한다. 그리고 독립 변수가 1개인 분석을 단순 회귀 분석(simple regression analysis), 독립 변수가 2개 이상일 경우의 분석을 다중 회귀 분석(multiple regression analysis)이라고 한다.

표 2. 회귀 분석 종류
Table 2. Regression types

Definition	Number of independent variable	Number of dependent variable
simple regression analysis	1	
multiple regression analysis	2 or more	
univariate regression analysis		1
multivariate regression analysis		2 or more

종속 변수가 1개이고 독립 변수도 1개인 회귀 모델은 단변량 단순 회귀 분석(univariate simple regression analysis) 이라고 정의한다. 단변량 단순 회귀 분석을 기반으로 선형 회귀 분석(linear regression analysis)을 사용할 경우 식 (3)으로 표현할 수 있다. X_1 이라는 독립 변수가 Y 라는 종속 변수에 주는 영향력을 수학 함수 형태로 나타낸 것이다. X_1 의 영향력은 a_1 이라는 계수의 크기와 부호로 나타낼 수 있으며 a_0 는 Y 에 영향을 주지 않는 상수이다.

$$Y = a_0 + a_1 X_1 \tag{3}$$

종속 변수가 1개이고 독립 변수가 2개 이상인 회귀 모델은 단변량 다중 회귀 분석(univariate multivariate multiple regression analysis) 이라고 정의한다. 단변량 다중 회귀 분석을 기반으로 선형 회귀 분석을 사용할 경우 식 (4)로 표현할 수 있다. X 가 n 개인 회귀 모델을 나타내고 있으며 a_0 는 Y 에 영향을 주지 않는 상수이고, $a_1, a_2 \ldots a_n$ 은 각각 $X_1, X_s, \ldots X_n$ 변수가 종속 변수 Y에 주는 영향력을 나타내고 있다.

$$Y = a_0 + a_1 X_1 + a_2 X_2 \quad \ldots \quad a_n X_n \tag{4}$$

본 논문에서는 종속 변수는 view로 설정하고, 독립 변수는 likes, dislies, comment로 설정하여 단변량 다중 회귀 분석을 실시하였다. 단변량 다중 회귀 분석에서 각 X 에 해당하는 a를 계산하기 위하여 선형성을 갖고 있으며 편향되지 않은 불편추정량이고 최소 분산성을 제공하는 최소 제곱법(least square method)[18]을 이용하여 선형 회귀 분석을 하였다. 최소 제곱법은 식 (5)와 같이 계산할 수 있다.

$$a = \frac{\sum_{i=1}^{n}(x - mean(x))(y - mean(y))}{\sum_{i=1}^{n}(x - mean(x))^2} \tag{5}$$

그림 13은 seaborn에서 제공하는 lmplot 라이브러리를 활용하여 likes, dislikes, comment 변수를 독립 변수로 설정하고 view를 종속 변수로 정의하여 단변량 다중 회귀 분석을 실행한 결과를 나타낸다. 단변량 다중 회귀 분석을 실시한 결과 그림 13과 같은 선형적인 회귀선을 보여주었다.

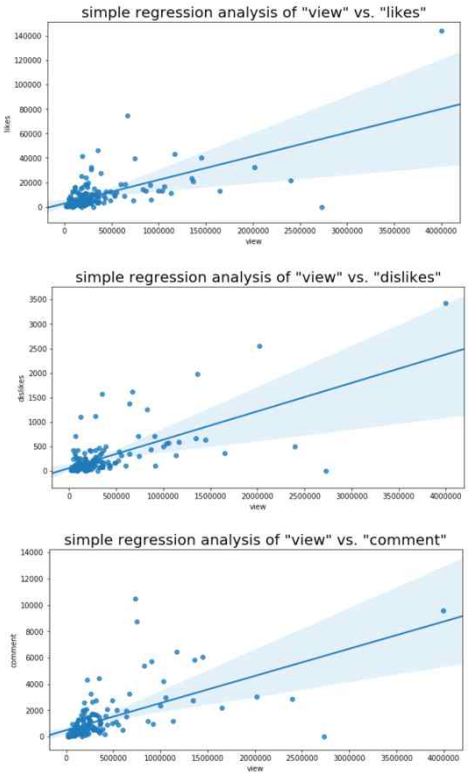


그림 13. 변수들의 선형 회귀 분석
Fig. 13. Linear Regression Analysis of Variables

Ⅳ. 결 론

데이터 마이닝은 빅데이터 분석을 통해 유의미한 정보와 규칙을 찾아내는 기술로 다양한 프로세싱을 통해 향후 결과를 예측하는 것을 목표로 한다. 본 논문에서는 python에서 제공하는 라이브러리를 활용하여 유튜브의 인기 동영상 콘텐츠를 대상으로 데이터 마이닝을 실행하였다.

유튜브의 인기 동영상 콘텐츠에 등장하는 단어의 빈도수를 기반으로 키워드를 추출하여 정성적 분석을 수행한 결과 현 시점의 주요 뉴스완 연관이 있는 영상이 인기 메뉴 범주에 포함되고 있음이 드러났다. 또한 수집된 데이터들은 조회 수, 이 동영상이 마음에 듭니다, 이 동영상이 마음에 들지 않습니다, 댓글 수 등을 변수로 설정하여 정형화된 데이터셋으로 정제하여 python 라이브러리를 활용하여 상관 분석을 실시한 결과 변수들은 양의 상관 관계가 있음이 나타났다.

데이터 마이닝 기술은 검색 엔진 알고리즘 및 추천 시스템과 같은 최신 인공지능 응용 프로그램에 필요한 딥러닝 모델을 구

성하는 데 필요한 과정이다. 본 논문에서 설계하고 구현한 유튜브 인기 동영상 콘텐츠 기반의 데이터 마이닝 결과물을 활용하여 지도 학습(supervised learning) 기반의 딥러닝 프로그래밍으로 연결하여 훈련 데이터(training data)로 활용이 가능할 것으로 기대한다.

본 논문에서 구축한 빅데이터를 기반으로 유튜브 크리에이터를 위한 인공지능 서비스를 구현하여 유튜버들의 개인 브랜드 가치를 높이고 수익 창출에 도움이 될 수 있도록 연구를 확장할 계획이다.

참고문헌

[1] K. Y. Lee, "The Impact of Privacy Concerns on Smartphone-based Ad Blocker Use Intent: Mediated Moderating Effect of Smartphone Literacy via Attitude toward Online Video Advertising," *The Journal of Digital Contents Society*, Vol. 21, No. 1, pp. 111-119, Jan. 2020.

[2] E. J. Kim and S. C. Whang, "A Study on Advertising Effect Depending on Type of Information Source and Displaying of Economic Support in Influencer Marketing : Focusing on Youtube," *The Journal of Digital Contents Society*, Vol. 20, No. 2, pp. 297-306, Feb. 2019.

[3] J. W. Jeong, J. Y. Lee and C. S. Lee, "An Analysis of Characteristics and User Reactivity by Video Categories on YouTube," *The Journal of Digital Contents Society*, Vol. 20, No. 12, pp. 2573-2582, Dec. 2019.

[4] Cárdenas, Alvaro A., (2013). Big Data analytics for security intelligence. Cloud Security Alliance [Internet]. Available: http://www.cloudsecurity alliance.org/ research/big-data.

[5] J. Y. Lee, J. H. Kang, S. Y. Jang and S. J. Yoo, "Examining and Analyzing Influential Factors of Ego-resilience: By Applying Data Mining Analysis," *Counseling Psychology Education Welfare*, Vol. 6, No. 1, pp.125-136, Mar. 2019

[6] D. J. Park and W. S. Kim, "Improvement of the Parallel Importation Logistics Process Using Big Data," *The Journal of information and communication convergence engineering*, Vol. 17, No. 4, pp.267 − 273, Dec. 2019.

[7] A. Yu Dorogov, "Technologies of predictive analytics for big data," *in Proceeding of the International Conference on Soft Computing and Measurements*, pp. 182-183, 2015.

[8] A. Alim1 and D. Shukla, "Big Data: Myth, Reality and Parametric Relationship," *The International Journal of Advanced in Management, Technology and Engineering Sciences*, pp. 1235-1244, Vol. 8, Issue III, Mar. 2018

[9] Mukherjee, Samiddha, and Shaw, Ravi, "Big Data − Concepts, Applications, Challenges and Future Scope", *The International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, No. 2, pp. 66-74, Feb. 2016

[10] Owais, Sushil Sami and Hussein, Nada Seal, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data", *The International Journal of Advanced Computer Science and Applications*, Vol. 7, No. 3, pp. 254-258, 2016.

[11] H. S. Kim and D. G. Tak, "Analysis on the Effectiveness of Life-based Coding Education Using Big Data Analysis Method," *The Journal of Digital Contents Society*, Vol. 20, No. 10, pp. 1943-1952, Aug. 2019.

[12] R. Varley, "ExoData: A Python package to handle large exoplanet catalogue data," *The Journal of Computer Physics Communications*, Vol. 207, No. 10, pp. 298-309, Oct. 2016.

[13] P. Dutilleul, J. D. Stockwell, D. Frigon and P. Legendre, "The Mantel Test versus Pearson's Correlation Analysis: Assessment of the Differences for Biological and Environmental Studies," *The Journal of Agricultural, Biological, and Environmental Statistics.* Vol. 5, No. 2, pp. 131-150 , Jun. 2000.

[14] L. Xiaochuan, M. David and L. Tianran, "A Hybrid Framework Combining Data-level Fusion and Model-based Models for Remaining Useful Life Prediction," *in Proceeding of the Prognostics and System Health Management Conference (PHM-Qingdao) Prognostics and System Health Management Conference,* PHM:Qingdao, pp. 1-5, Oct. 2019.

[15] O. S. Sami and H. N. Seal, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data," *The International Journal of Advanced Computer Science and Applications*, Vol. 7, No. 3, pp. 254-258, 2016.

[16] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists,* 5th ed. MA: Elsevier Academic Press, 2014.

[17] S. Alvarez-Cortes, J. Serra-Sagrista, J. Bartrina-Rapesta and M. Marcellin, "Regression Wavelet Analysis for Near-Lossless Remote Sensing Data Compression," *The Journal of IEEE Transactions on geoscience and remote sensing,* Vol. 58, No. 2, pp.790 − 798, Feb. 2020.

[18] L. Tongying and Z. Hongbo, "Design and Implementation of Autoleveller Algorithm Based on Generalized Least Square Method," in *Proceeding of the Chinese Automation Congress (CAC)*, pp. 1851-1855, Nov. 2019.

김희숙(Hye-Suk Kim)

1999년 : 전남대학교 대학원 전산통계학과 (이학석사 - 멀티미디어)
2009년 : 전남대학교 대학원 전산학과 (이학박사 - 영상처리)

2003년~현 재: 전남대학교 전자컴퓨터공학부, GIST(Gwangju Institute of Science and Technology) Lecture Professor.
※관심분야 : 빅데이터, 데이터과학, 딥러닝, 인공지능, 영상처리, 멀티미디어콘텐츠 등