비디오 플랫폼 내 인기 영상 분석을 통한 사용자 행동 연구

지도교수 : 권태경

이 논문을 공학학사 학위 논문으로 제출함.

2019년 12월 26일

서울대학교 공과대학
컴 퓨 터 공 학 부
이 창 영

2020년 2월

목차

1. Introduction

플랫폼이란 공급자와 수요자가 참여해 서로 얻고자 하는 가치를 교환할 수 있도록 구축된 환경이다. 나아가 비디오 플랫폼은 공급자가 영상을 제공하고 수요자는 그 영상을 소비할 수 있는 플랫폼이고, 이 플랫폼을 통해 사용자들은 좋아요, 싫어요, 또는 댓글 등의 반응을 표현할 수 있는 특징이 있다. 스마트폰과 같은 매체들의 공급으로 이러한 플랫폼에 대한 사용자들의 접근이 쉬워졌고, 그에 따라 비디오 플랫폼의 사용량이 급격하게 증가하고 있다. 그중 가장 대표적인 비디오 플랫폼인 유튜브는 국내에서 작년 대비 스마트폰 앱 사용시간이 50%나 상승할 정도로 엄청난 성장을 하고 있다. [1]

비디오 플랫폼 사용의 폭발적인 증가에 따라 이 플랫폼에 대한 연구도 활발히 진행되고 있다. 예를 들면 플랫폼 서버에서 생성되는 트래픽의 특징에 대한 연구가 있다. 이러한 연구는 사용자의 비디오 품질을 예측하고 네트워크 설계를 개선하는 데 이점을 가져다줄 수 있다. 또 다른 연구로는 영상 추천 시스템에 대한 연구가 있다. 이 연구는 사용자의 활동을 기반으로 한 영상 추천에 대한 것으로, 비디오 플랫폼의 품질 향상에 도움이 된다. [2][3]

하지만 이러한 연구들은 시스템 최적화 및 플랫폼 개선을 위한 연구에 집중되어 있고, 인기 있는 영상들의 다양한 특징 및 그 영상에 대한 사용자들의 소비 패턴을 연구한 것은 찾아보기 힘들다. 따라서 본 연구에서는 사용자들에게 인기 있는 동영상들의 경향성과 사용자들이 어떻게 소비하는지 연구하기 위해 대표적인 비디오 플랫폼인 유튜브에서 제공하는 인기 영상들을 분석하기로 하였다. 이를 위해 크롤러를 개발하여, 한 달 동안 유튜브에서 제공하는 인기 동영상 페이지를 주기적으로 수집하였다. 이때 국가별 차이도 확인하기 위해 한국의 인기 동영상 페이지, 미국의 인기 동영상 페이지를 각각 수집하였고, 수집할 때마다 각 영상의 '좋아요', '싫어요', '조회수' 변화를 수집하였다, 또한 카테고리와 같은 메타데이터도 수집하여 카테고리 별 차이도 확인할 수 있도록 하였다. 이러한 연구는 인기 동영상의 특징 및 사용자들의 반응에 대한 이해를 높이는 것을 통해 유튜브 채널의 홍보 효과 증진을 위한 목표 설정에 도움이 될 것으로 판단된다, 또한 국가별 유튜브 이용에 관한 특성을 비교함으로써, 각 국가별 이용자에 대한 이해를 높일 수 있으며 그 차이를 이해할 수 있다.

본 논문의 구성은 다음과 같다. 제 2장에서는 본 논문에서 이용하기 위한 데이터를 수집한 방법과 수집한 데이터에 대하여 설명하고, 제 3장은 국가별, 카테고리별 영상들의 특징과 그 영상들에 사용자들이 어떻게 반응하는지에 대해 수집한 데이터셋을 가공하여 그래프로 나타낸 모습을 볼 것이다. 제 4장에서는 3장으로부터 도출된 결론과 이에 기반하여 할 수 있는 향후 연구에 대해서 이야기할 것이다.

2. Data Description

본 연구를 수행하기 위해서 가장 대표적인 비디오 플랫폼 중 하나인 유튜브의 인기

영상 정보를 수집하였다. 유튜브는 자체적인 알고리즘을 통해 현재 인기 있다고 판단되는 영상 50개를 인기 영상 목록으로 제공하고 있다. 이 영상들은 접속 국가별로 다르게 나타나는데 이 연구를 위해서 크게 한국의 인기 영상, 미국의 인기 영상으로 나누어 데이터를 수집했다. 각 국가 별로 수집된 데이터는 다시 두 구성으로 이루어져 있다. 첫 번째 구성은 매 30분 마다 인기 목록에 올라온 영상들과 그들의 순위 정보를 수집한 내용이다. 두 번째 구성은 매 30분마다 최근 3일 이내에 인기 목록에 올라왔던 모든 영상들에 대해 조회수, 좋아요, 싫어요, 카테고리 등을 수집한 내용이다. 이 데이터 셋에 대해 정리한 내용은 표 1과 같다.

구분		구성
한국	인기영상 목록 및 순위 정보	수집 시간, 영상 ID, 각 영상 순위
	인기영상에 올라온 비디오 정보	수집 시간, 영상 ID, 조회수, 좋아요, 싫어요, 카테고리
미국	인기영상 목록 및 순위 정보	수집 시간, 영상 ID, 각 영상 순위
	인기영상에 올라온 비디오 정보	수집 시간, 영상 ID, 조회수, 좋아요, 싫어요, 카테고리

[표 1] 데이터셋 구성

위 정보들은 직접 개발한 크롤러를 통해 수집되었다. 크롤러는 매 30분 마다 작동되도록 설정되었고 한국과 미국의 인기 영상 목록 페이지와 각 영상의 재생 페이지에서 필요한 정보들을 수집하였다. 그리고 수집된 데이터는 데이터베이스에 저장하여 보관하였다. [표 2]에 나와 있듯이 데이터는 2019년 10월 26일 11시부터 2019년 11월 26일 11시까지 30분 간격으로 총 1489번 수집되었고, 그 결과 수집된 영상들은 한국은 1359개, 미국은 1121개이다.

	수집 횟수	수집된 영상 수
한국	1489	1359
미국	1489	1121

[표 2] 2019년 10월 26일 11:00 ~ 2019년 11월 26일 11:00 수집된 데이터

3. Analysis

본 단락에서는 인기 목록에 올라온 영상이 보이는 특징과 그에 대한 사용자들의 반응에 대해서 크게 두 가지 기준, 국가별 차이, 그리고 카테고리별 차이가 어떻게 나타나는지 분석하였다.

3-1. 국가별 분석

1) 인기 목록에 올라오는 영상들의 카테고리 점유율

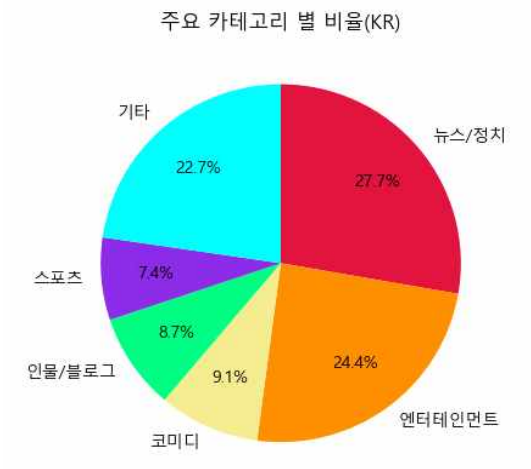
첫 번째로, 각 국가에서 어떤 카테고리가 인기 있는지 분석해 보았다. 이를 위해 전체 수집된 영상들을 카테고리별로 나누어 인기 있는 상위 5개 카테고리에 대해 수를 세어 보았다. [표 3], [표 4]을 통해 그 수를 확인할 수 있으며, 그 비율은 [그림 1], [그림 2]를 통해 확인할 수 있다. 각 국가에서 인기 있는 카테고리 5개씩 확인해 보니 한국은 '뉴스/정치', '엔터테인먼트', '코미디', '인물/블로그', '스포츠' 순으로 나타났고, 미국은 '엔터테인먼트', '스포츠', '음악', '코미디', '인물/블로그' 순으로 나타났다. 미국은 한국에 비해 상위 5개 카테고리 이외의 비율이 높은 것으로 보아 더 다양한 카테고리가 관심을 받는다는 것을 알 수 있다. 그리고 미국은 '뉴스/정치' 카테고리가 거의 관심을 받지 못하지만, 한국은 '뉴스/정치' 카테고리가 압도적인 관심을 받는다는 특징을 확인할 수 있다. 이러한 특징은 관련 기사에서도 찾아볼 수 있다. 한 기사에 따르면 정치/시사 분야의 영상은 선거기간뿐만 아니라 그렇지 않을 때도 인기 영상의 약 20%를 차지하는 것으로 나타났다. 이를 통해 한국에서는 '뉴스/정치' 카테고리가 큰 인기를 끌고 있는 것을 다시 확인할 수 있다. [4]

카테고리	개수
뉴스/정치	395
엔터테인먼트	347
코미디	129
인물/블로그	124
스포츠	106
기타	258
계	1359

[표 3] 수집된 데이터의 카테고리별 영상 개수(한국)

카테고리	개수
뉴스/정치	316
엔터테인먼트	157
코미디	111
인물/블로그	101
스포츠	74
기타	362
계	1121

[표 4] 수집된 데이터의 카테고리별 영상 개수(미국)



[그림 1] 카테고리별 영상 개수 비율(한국)

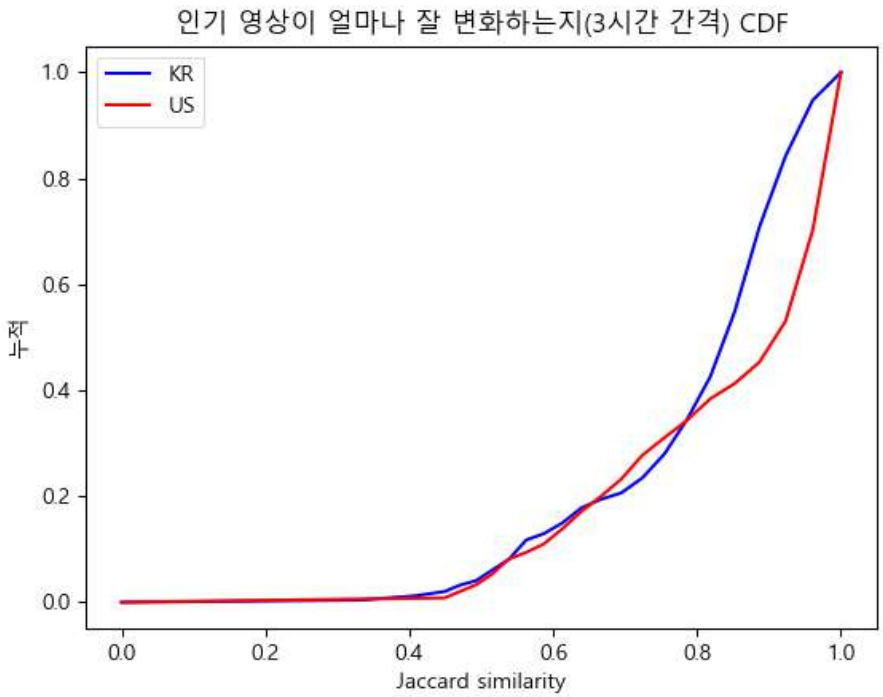
[그림 2] 카테고리별 영상 개수 비율(미국)

2) 인기 영상의 목록 변화 정도

두 번째로, 각 국가에서 인기 목록에 올라오는 영상들의 변화에 얼마나 차이가 있는지 분석해 보았다. 이 분석을 통해 이용자들의 관심이 변화하는 정도가 얼마나 차이나는지 알아보고자 하였다. 수집된 데이터에서 3시간 간격으로 인기 영상 목록을 set으로 뽑은 뒤 두 set 간의 차이들을 보기로 하였다. 데이터는 30분 간격으로 수집하였지만 30분 사이에서는 큰 차이를 보이는 경우가 거의 없었으므로 적당한 차이가 드러나는 3시간 간격으로 정하였다. [그림 3]은 3시간 간격으로 뽑은 set들의 Jaccard similarity를 계산하여 유사도를 구하고 그 결과를 누적분포함수로 그린 것이다. Jaccard similarity는 다음과 같은 식으로 구할 수 있다.

$$J(X,Y) = |X \cap Y| \ / \ |X \cup Y|$$

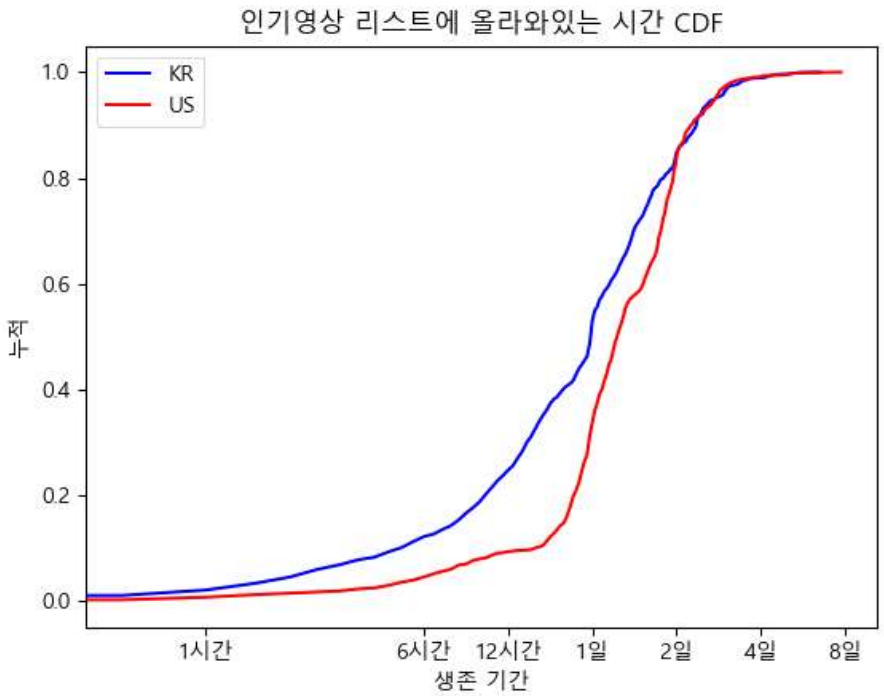
[그림 3]을 통해 보았을 때 두 나라 모두 Jaccard similarity가 0.4 이하로 내려가는 경우가 거의 없었다. 즉 매 3시간 간격에서 보통 60% 이상의 영상들은 인기 목록에서 유지된다는 것이다. 두 나라 모두 인기 목록이 급격하게 변하지 않았고, 그 양상도 80%까지는 비슷하였지만, 90% 이상으로 거의 변화가 없는 경우는 미국이 더 많았음을 알 수 있다.



[그림 3] 인기 영상 목록 set 간의 Jaccard similarity CDF

3) 영상이 인기 목록에서 생존한 시간

세 번째로, 각 국가에서 인기 영상이 인기 목록에서 얼마나 생존하는지 분석하였다. 이 분석을 통해 인기 영상의 생존 시간의 국가별 차이를 확인할 수 있다. 목록에 올라와 있던 기간을 측정하기 위하여 인기 영상 목록을 30분에 한 번씩 수집하였으므로 어떤 영상이 인기 목록에 한 번 올라올 때마다 그 영상이 생존한 시간을 30분 추가하여 계산하였다. [그림 4]는 위와 같은 방법으로 구한 영상들의 생존 시간을 누적분포함수로 그린 것이다. 한국, 미국에서 공통적으로 하루에서 이틀 동안 생존하는 경우가 가장 많았고, 이틀 이상 생존한 영상들은 각 국가에서 비슷한 비율로 나타났다. 차이점은 20시간을 기준으로 보았을 때 한국은 그 이하 생존한 영상도 40% 정도로 많은 비율을 차지하고 있지만, 미국은 10%로 매우 낮은 것을 볼 수 있었다. 따라서 평균적으로 한국의 영상들이 생존 시간이 더 짧은 것을 알 수 있다.

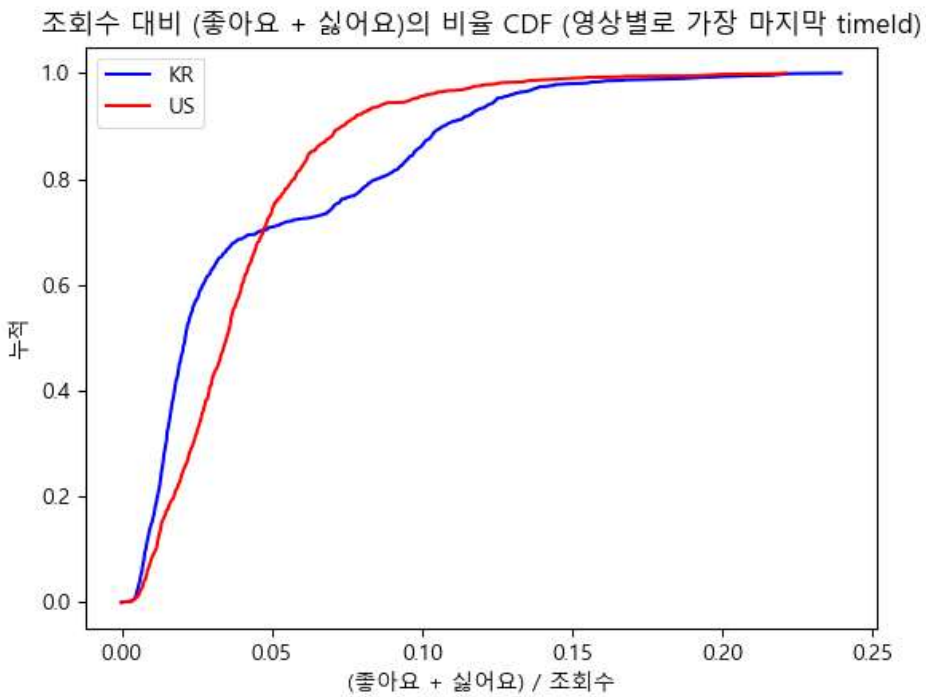


[그림 4] 영상이 인기 목록에서 생존한 시간 CDF

4) 영상에 사용자가 반응하는 정도

네 번째로, 사용자가 영상에 얼마나 반응하는지에 대해 분석하였다. 이는 사용자가 영상을 본 뒤 그에 대한 주관적인 의사 표현을 얼마나 하는지 알아보기 위함이다. 사용자는 의사 표현을 하기 위해 단순히 동영상 시청에서 끝나는 것이 아닌 댓글을 달거나 '좋아요', '싫어요' 버튼을 누를 것이라 보았고, 본 연구에서는 그중에서 간단하게 수치를 구할 수 있는 사용자가 영상을 보고 '좋아요' 또는 '싫어요'를 누른 경우 반응한 것으로 정하였다. 따라서 모든 영상에 대해 마지막에 수집된 메타 정보에서

((‘좋아요’ + ‘싫어요’) / ‘조회수’) 값을 구하였다. [그림 5]는 이렇게 구한 값을 누적 분포함수로 그린 것이다. 양쪽 국가 모두 90% 이상의 영상들이 10% 이내에서 반응하는 편이며, 일부는 20%~25%까지 반응하기도 하였다. 미국은 반응하는 정도가 5%~6%인 영상들의 수까지는 일정하다가 점점 줄어드는 평범한 모습을 보인 반면, 한국은 반응 정도가 낮은 영상과 반응 정도가 높은 영상으로 분명하게 나뉘는 모습을 보인다. 이렇게 분명하게 나뉘는 부분에 대해서는 3-2의 3)에서 분석하였다.



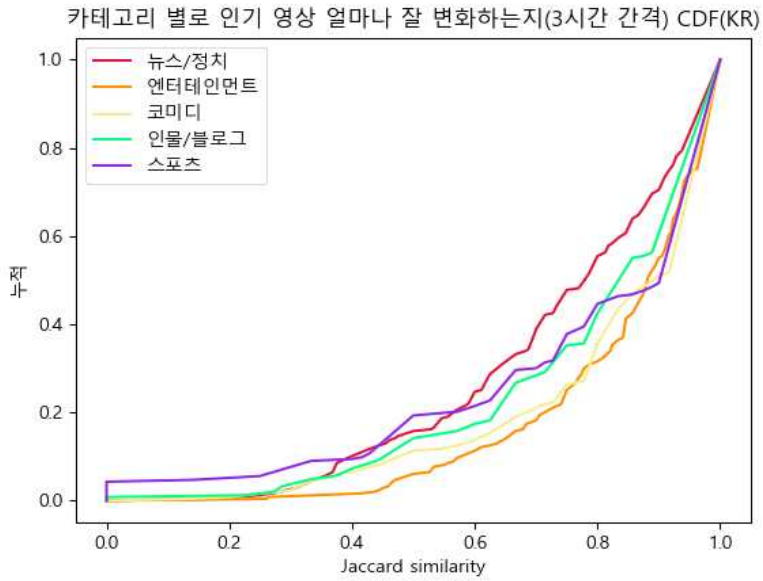
[그림 5] 영상의 ((좋아요 + 싫어요) / 조회수) CDF

3-2. 국가별 + 카테고리별 분석

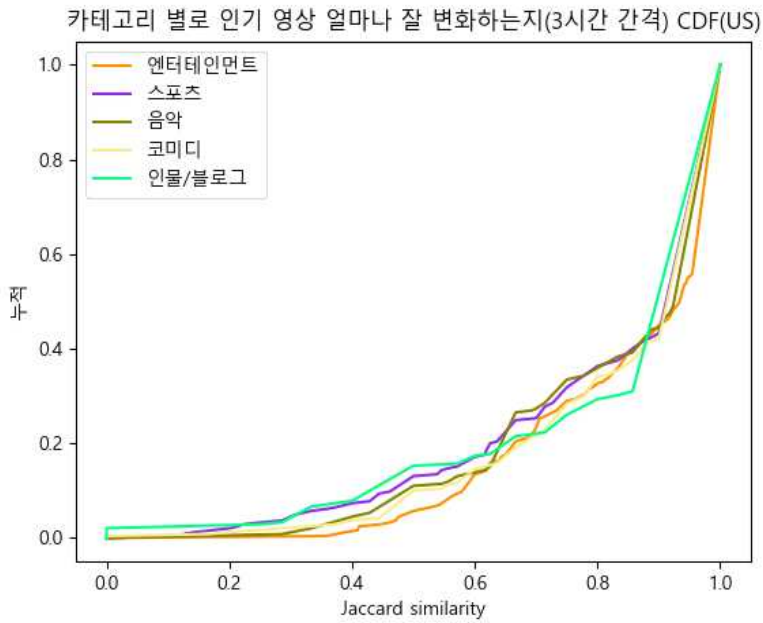
3-1의 여러 분석에서 인기 영상들의 국가별 의미 있는 차이를 확인할 수 있었으며 3-1의 4)에서와 관찰한 것과 같이 국가 간에 유의미한 경향성의 차이를 볼 수 있는 부분 역시 존재하였다. 이러한 차이에 대해 좀 더 심도 있는 분석을 위해서 각 국가별 영상들에 대해서 이를 영상의 범주로 세분화하여 그 사이에서 어떤 차이를 보이는지 더 자세하게 관찰하였다. 유튜브에서는 다양한 카테고리의 영상을 찾아볼 수 있는데 본 논문에서는 3-1의 1)에서 분석한 나라별 인기 있는 카테고리 상위 각 5개씩을 대상으로 분석하였다.

1) 인기 영상의 목록 변화 정도

첫 번째로, 3-1의 2)에서 본, 인기 영상의 목록 변화 정도를 같은 방식으로 카테고리별로 나누어 분석해 보았다. [그림 6]을 통해 봤을 때 한국은 평균적으로 뉴스/정치 카테고리가 비교적 빨리 변화하고 엔터테인먼트 카테고리는 덜 변화하는 것으로 나타난다. 반면 [그림 7]을 통해 봤을 때 미국은 카테고리별로 의미 있는 차이를 보이지 않는 것을 확인할 수 있었다.



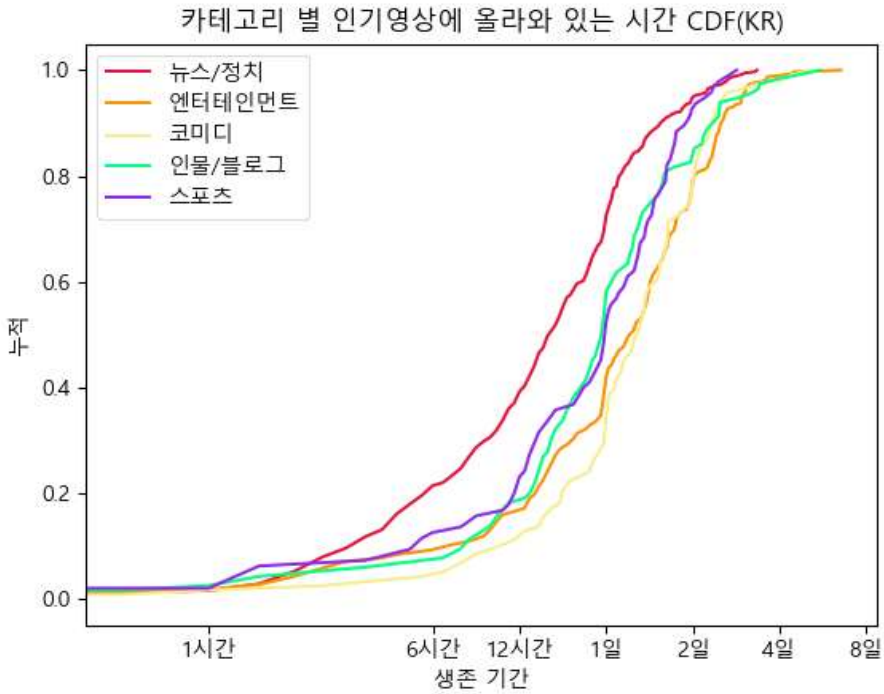
[그림 6] 한국의 카테고리별 인기 목록 Jaccard similarity CDF



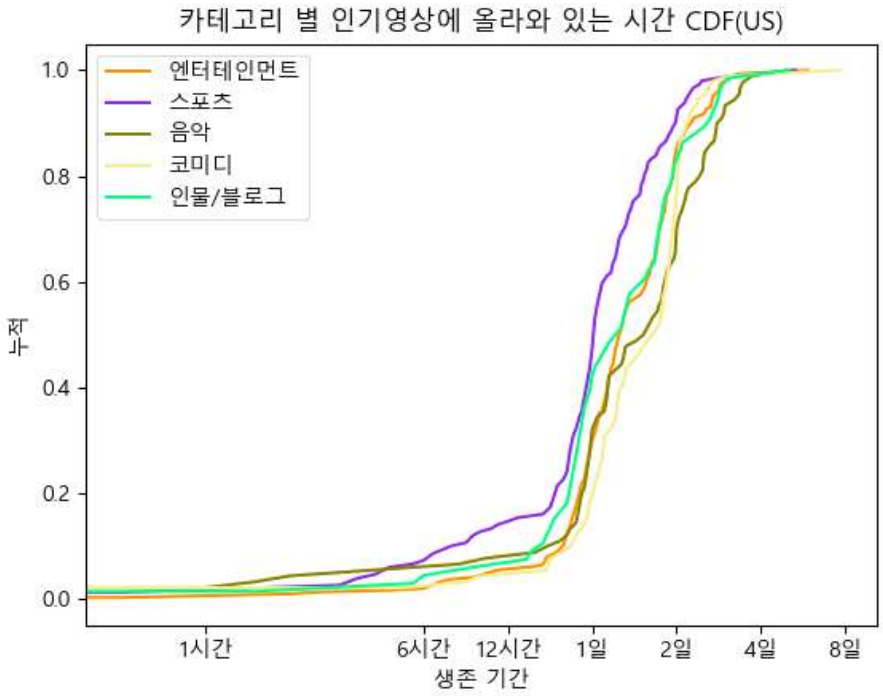
[그림 7] 미국의 카테고리별 인기 목록 Jaccard similarity CDF

2) 영상이 인기 목록에서 생존한 시간

3-1의 3)에서 본, 영상이 얼마나 오래 인기 목록에 생존하였는지에 대해 같은 방식으로 카테고리별로 나누어 분석해 보았다. [그림 8]을 보면 한국에서는 평균적으로 뉴스/정치 카테고리의 생존 시간이 짧고 코미디, 엔터테인먼트 카테고리의 생존 시간이 비교적 길게 나타났다. [그림 9]를 통해 본 미국의 데이터는 카테고리별 생존 시간은 큰 차이는 없지만 '뉴스/정치'와 '음악' 카테고리만 빼면 평균적으로는 한국의 카테고리별 생존 기간 순위랑 비슷하게 나타났다. 이 분석을 통해 볼 때 3-1의 3)에서 분석한 한국의 그래프는 '뉴스/정치' 카테고리의 영향을 많이 받은 모습으로 보인다.



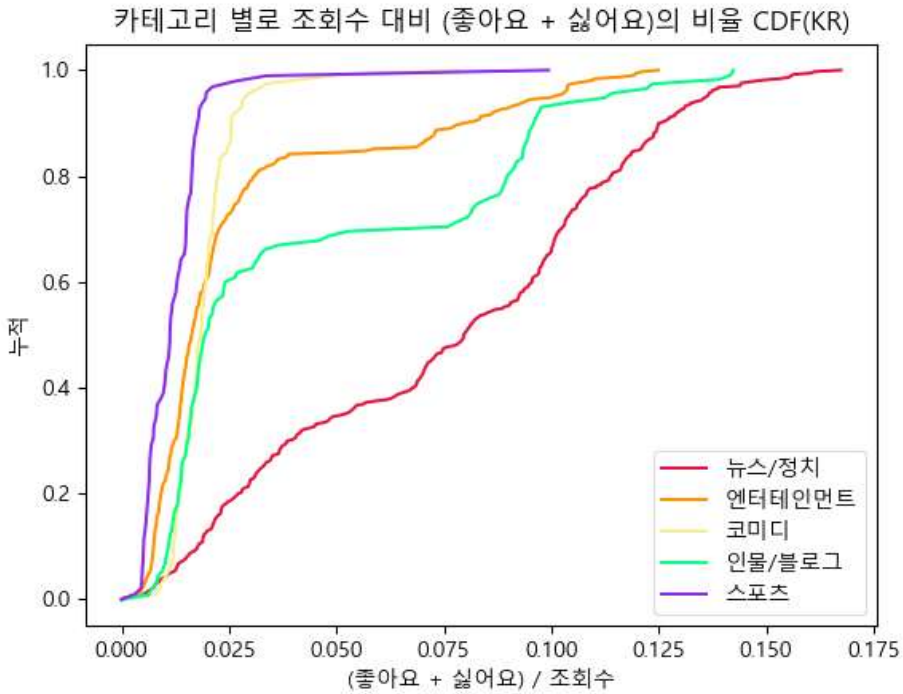
[그림 8] 한국의 카테고리별 생존 시간 CDF



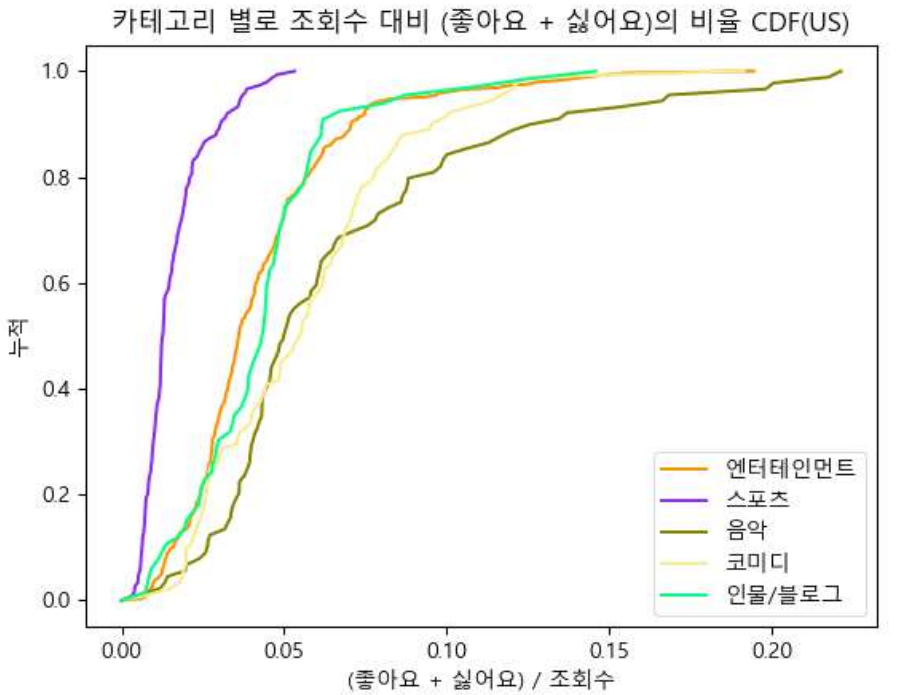
[그림 9] 미국의 카테고리별 생존 시간 CDF

3) 영상에 사용자가 반응하는 정도

3-1의 4)에서 본, 영상에 사용자가 반응하는 정도에 대해 같은 방식으로 카테고리별로 나누어 분석해 보았다. [그림 10]에서 보면 한국은 반응하는 정도가 분명하게 뉴스/정치 > 인물/블로그 > 엔터테인먼트 > 코미디 > 스포츠 순으로 나타났다. 3-1의 4) 분석을 보면 한국은 반응을 적게 하는 영상들의 집합과 반응을 많이 하는 영상들의 집합이 두드러지게 나뉘어 있었는데 이 분석을 통해 보니 인물/블로그 카테고리와 뉴스/정치 카테고리가 그 특징을 만들었다고 확인할 수 있었다. [그림 11]에서 보면 미국은 스포츠 카테고리에서 사용자들의 반응이 적었다는 점에서 한국과 비슷한 양상을 보였고 다른 카테고리에서는 평균적으로 엔터테인먼트 = 인물/블로그 < 코미디 < 음악 순으로 반응 정도를 보였으나 그 차이가 크지 않았다.



[그림 10] 한국의 카테고리별 ((좋아요 + 싫어요) / 조회수) CDF



[그림 11] 미국의 카테고리별 ((좋아요 + 싫어요) / 조회수) CDF

4. Conclusion & Future Work

(1) Conclusion

본 연구에서는 비디오 플랫폼에서 인기 있는 영상들의 특징과 그 영상들을 사용자들이 소비하는 패턴에 대하여 국가별, 카테고리별 어떤 특징이 있는지를 분석하고자 하였다. 가장 대표적인 비디오 플랫폼 중 하나인 유튜브에서 제공하는 한국과 미국의 인기 영상 목록을 수집하였고 수집된 데이터를 관찰하고자 하는 여러 기준에 맞게 가공하여 그래프를 그려보며 다양한 현상과 특징을 확인할 수 있었다.

먼저 인기 영상들의 국가별 분석에서 확인한 것은 한국, 미국에서 인기 있는 카테고리의 차이이다. 각 나라에서 인기 있는 다섯 가지 카테고리를 뽑았을 때 큰 차이를 보인 것은 '뉴스/정치' 카테고리이다. 한국에서는 '뉴스/정치' 카테고리가 1순위로 인기가 많았던 반면 미국에서는 거의 관심 밖이었다. 그리고 미국의 영상들이 더 다양한 카테고리가 고르게 관심을 받았다. 한국의 인기 영상 목록은 미국에 비해 평균적으로 더 잘 변화했고 그에 따라 자연스럽게 한국의 인기 영상이 인기 목록에서 생존하는 시간이 더 짧다는 사실도 확인할 수 있었다. 영상에 대한 사용자의 반응 정도는 한국이 특이한 양상을 보였는데 그 자세한 특징은 국가별 + 카테고리별 분석에서 알 수 있었다.

다음은 국가별 + 카테고리별 분석이다. 전체적으로 미국의 카테고리별 차이가 한국에 비해 적은 편이었다. 국가별 분석에서 한국에서만 특이하게 인기가 많았던 '뉴스/정치' 카테고리가 이 분석에서도 특이한 모습을 많이 나타냈다. 한국 데이터에서 '뉴스/정치' 카테고리는 인기 목록에서 가장 잘 변화하였으며, 자연히 생존 시간도 가장 짧았다. 그리고 사용자가 반응하는 정도도 다른 카테고리와 확연히 차이나게 많은 두드러진 모습을 보였다. '스포츠' 카테고리는 한국과 미국에서 공통적으로 가장 사용자의 반응이 없는 카테고리였으며 인기 목록에서 생존하는 시간도 짧아 사용자들이 가장 가볍게 소비하는 카테고리임을 알 수 있었다.

(2) Future work

본 연구에서는 인기 목록의 영상에 대한 특징 및 사용자의 소비 패턴을 국가, 카테고리의 측면에서만 분석하였다. 따라서 국가별 차이, 또는 카테고리별 차이의 존재 여부만 확인할 수 있었다. 따라서 차후 연구에서는 영상 제목이나 영상 설명, 영상 게시자 등등 다른 요소들은 어떤 영향을 끼치는지, 그리고 사용자의 반응성에 대해서는 '조회수', '좋아요', '싫어요'뿐만 아니라 댓글 수와 그 댓글의 의미까지 수집하여 더 깊이 있는 특징을 알아내고자 한다.

[1] “韓유튜브 사용시간 전세대 1위, ’50대↑’”, Bloter, 2019년 5월 14일 수정, 2019년 12월 12일 접속, http://www.bloter.net/archives/339870.
[2] Pablo Ameigeiras Juan J. Ramos-Munoz Jorge Navarro-Ortiz J.M. Lopez-Soler, 『Analysis and modelling of YouTube traffic』, Transactions on Emerging Telecommunications Technologies, Volume 23, Issue 4.
[3] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, Dasarathi Sampath, 『The YouTube video recommendation system』, RecSys '10 Proceedings of the fourth ACM conference on Recommender systems, Pages 293-296.
[4] “유튜브 인기영상 3건 중 1건이 정치·시사”, 미디어오늘, 2018년 11월 6일 수정, 2019년 12월 12일 접속,
http://www.mediatoday.co.kr/news/articleView.html?idxno=145362.