너비 우선 탐색(BFS) 기법 기반 국내 유튜브(Youtube) 동영상 정보 크롤링 기법 및 국내 Youtube 데이터셋 구축

김태영○ 박하명[1)]
국민대학교 소프트웨어융합대학 소프트웨어학부
tykim75@kookmin.ac.kr hmpark@kookmin.ac.kr

Crawling Techniques for Video Information on Youtube in Korea Based on BFS and Construction of Youtube video dataset in Korea

TaeYoung Kim○ Ha-Myung Park[1)]
College of Computer Science, Kookmin University

요 약

Youtube는 사용자 간 동영상을 게시 및 시청하며 반응을 표현할 수 있는 동영상 기반의 SNS 서비스 플랫폼이다. 최근 조사에 따르면 Youtube는 한국인이 가장 많은 시간을 소비하는 플랫폼이며, 이에 따라 다양한 관점의 Youtube 관련 연구들이 지속적으로 발표되고 있다. Youtube에 관한 해외 연구의 경우 고품질의 풍부한 데이터셋을 기반으로 연구가 진행되는 반면, 국내 연구는 충분한 데이터셋을 확보하지 못하여 현재로서는 깊이 있는 연구가 사실상 불가능한 실정이다. 본 연구에서는 고품질의 국내 Youtube 데이터셋을 구축하는 너비 우선 탐색 (BFS) 기반 크롤링 방법을 제안한다. 또한, 다양한 후속 연구에서 활용할 수 있도록 제안하는 방법으로 미리 구축한 데이터셋을 공개한다.

1. 서 론

Youtube는 사용자 간 동영상을 공유하며 서로 의견을 나눌 수 있는 동영상 기반 SNS 서비스 플랫폼이다. 이용자들은 다양한 정보의 획득과 재미를 위해 Youtube를 일상적으로 사용하고 있다. 앱 분석업체인 와이즈앱에 따르면 2020년 11월 기준 Youtube는 한국인이 가장 많이 사용하는 플랫폼이다.[1]

Youtube의 인기에 힘입어 Youtube에 관한 다양한 연구가 진행되고 있다.[2-6] Youtube와 관련한 해외 연구의 경우 고품질의 풍부한 데이터셋을 기반으로 활발한 연구가 진행되고 있는 반면,[5,6] 국내에서 공유되는 Youtube 동영상과 관련한 양질의 데이터셋이 존재하지 않아 국내 연구는 지지부진한 실정이다. 지금까지 발표된 국내 연구는 각기 나름대로의 방식으로 동영상 정보를 수집하는데, 데이터의 양이 많지 않거나 다양성의 관점에서도 부족하다.[2,3]

본 논문에서는 Youtube 영상의 연관 동영상을 찾을 수 있는 Youtube Search API를 이용해서, 너비 우선 탐색(BFS)에 기반한 효과적인 Youtube 동영상 정보 크롤링 기법을 제안한다. 또한, 다양한 후속 연구에서 활용할 수 있도록 제안하는 방법으로 미리 구축한 데이터셋을 공개한다.[2)]

1) 교신저자

2) https://drive.google.com/file/d/1WmGxATD7meJhneqeI9qXS2UuchJoElRa

2. 배 경

2.1 관련 연구

국내에서 Youtube를 이용하는 사용자가 많아지면서 국내에서도 Youtube 서비스의 네트워크 분석과 Youtube의 영상 데이터를 활용한 알고리즘 연구가 진행되었다.[2-4] 해당 연구들은 국내 Youtube의 동영상 데이터를 획득하기 위해 각자 다른 크롤링 기법을 사용하고 있다. 한 연구[2]에서는 Youtube의 인기 메뉴 동영상 콘텐츠라는 웹페이지[3)]를 크롤링해 동영상의 ID를 얻어 Youtube의 Video API(Application Programming Interface)를 이용해 동영상 정보를 수집한다. 해당 연구에서 제시한 데이터 수집 방법은 'Youtube Trend' 라는 인기 동영상 데이터만 수집되어 인기 동영상이 아닌 동영상과 과거 동영상의 정보를 수집하는데 제한이 된다.

다른 연구에서[3]는 Youtube 채널에서 구독자 수와 총 영상 수를 자체적인 기준을 두어 크롤링할 채널을 선정하고 Youtube Channel API를 통해 각 채널의 최고 인기 영상의 통계 데이터를 수집한다. 이 방법은 [2]의 연구보다 Youtube API 기준으로 나눈 데이터셋의 카테고리가 더 다양하게 출력될 수 있다는 장점은 있으나, 전처리 기준을 채널로만 나누다 보니 적은 데이터를 가져올 수밖에 없다는 한계가 존재한다.

이러한 국내의 Youtube 동영상 정보를 수집하는 방법의 한계를 넘어 본 연구는 Youtube Search API와 너비 우선 탐색 기법을

3) https://www.youtube.com/feed/trending

Youtube 동영상 정보 수집 과정에 적용하여 국내의 수많은 다양한 Youtube 동영상 정보를 수집하는 방법을 제안하고자 한다.

2.2 Youtube API

본 연구에서는 Youtube Video API와 Youtube Search API를 사용한다. Youtube Video API는 특정 동영상의 ID를 입력으로 받아 영상 게시에 관한 기본 정보, 조회 수, 댓글 수 등 영상에 관한 통계 정보를 출력한다. 해당 출력 데이터로 본 연구는 입력 동영상의 게시 국가가 국내인지, 국외인지 파악할 수 있다. Youtube Search API는 노드의 Video ID, maxResult 등의 입력을 받아 Video ID에 해당하는 영상과 연관된 추천 영상을 maxResult 수 만큼 출력한다. maxResult는 최대 50까지 설정이 가능하다.

3. 크롤러 시스템 구성

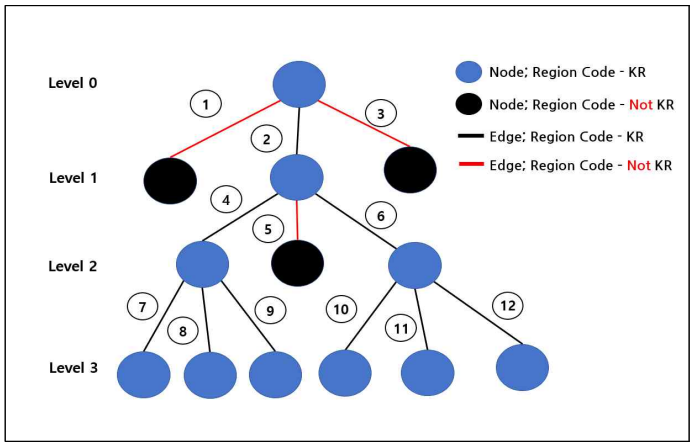

그림 1. BFS Tree를 활용한 Youtube 동영상 크롤러 시스템 아키텍처

본 연구에서 제안하는 Youtube 동영상 정보 크롤러 시스템은 그림 1과 같은 너비 우선 탐색 기법으로 작동한다. Tree에서 각 노드는 Youtube 영상을 의미하며, 노드와 연결된 자식 노드들은 Youtube Search API를 통해 획득한 연관 동영상들을 의미한다. 제안 기법은 복수의 Seed 노드를 Root노드로 하여 너비 우선 탐색을 수행하며 국내에서 시청하는 Youtube 영상을 수집한다. Youtube에는 다양한 국가의 영상들이 존재하는데 그 중에 한국 영상으로 판별된 노드는 파란색, 외국 영상으로 판별된 노드는 검정색으로 표현했다. 외국 영상에서는 더 이상의 깊이 우선 탐색을 수행하지 않는다. 간선 옆의 숫자는 탐색 순서의 예시이다.

국내 영상을 효과적으로 수집하기 위해서는 세 가지 문제를 해결해야 한다. 첫째, 탐색의 시작 정점인 Seed 영상은 어떻게 획득하는가? 둘째, 탐색 중인 영상이 국내 영상인지 외국 영상인지 어떻게 판별이 가능한가? 셋째, Youtube API 사용에 제약이 있는데 이를 어떻게 활용하여야 최대한 많은 영상을 수집할 수 있는가?

3.1 Seed 영상 획득

본 연구 목표는 국내 Youtube 동영상 데이터셋 구축이다. 그래서 2013~2019년의 국내 인기 동영상 데이터는 Youtube에서 제공하는 'Youtube Rewind' 라는 채널[4]로 수집했다. 또한, 해당의 채널에 존재하지 않는 2020년의 국내 최고 인기 동영상에 대해서는 Youtube의 발표 자료를 인용한 YTN 기사[5]를 통해 파악했고 총 63개의 동영상을 시드 영상으로 선정했다.

3.2 국내 영상 판별

노드로 표현되는 하나의 영상이 국내에서 게시됐는지 구분하기 위해 고려할 수 있는 한 가지 방법은 Youtube Search API를 통해 얻을 수 있는 연관 영상의 지역 코드인 'RegionCode' 라는 필드를 확인하는 것이었다. 하지만 해당 API에서 나오는 RegionCode 필드가 영상이 게시된 국가가 아닌 Query를 요청한 IP 주소의 국가코드를 출력하는 문제가 있었다. 그래서 이를 영상 ID를 입력으로 받아 특정 영상의 정보를 출력하는 Youtube Video API로 해당 문제를 해결했다. Youtube Video API의 출력 데이터 중 영상 제목 'VideoTitle' 필드, 채널명 'ChannelTitle' 필드, 영상 게시자가 작성하는 해당 영상의 해시태그인 'Tags' 필드를 입력으로 하여 정규식과 한글 유니코드를 사용해 특정 영상이 국내에서 게시되었는지를 구분할 수 있었다.

3.3 너비 우선 탐색 방법

국내 Youtube 동영상 네트워크를 확장하기 위해 본 제안 방법은 너비 우선 탐색을 활용한다. 이를 위해 시스템의 기본 사항에서 언급한 영상의 ID를 입력으로 받아 특정 영상 정보를 출력하는 Youtube의 Video API, 특정 영상의 연관 영상을 출력하는 Search API를 사용하였다. 먼저 Tree의 level이 내려가기 전, Video API를 활용해 같은 레벨에 존재하는 동영상의 ID를 입력으로 해 이 영상들의 정보를 수집했다. 그리고 3.2에서 설명한 방법을 사용해 이 영상들이 국내의 영상인지, 아닌지를 구분한다. 이후, 해당 영상이 국내의 영상이라면 이 영상들의 ID를 입력으로 Search API를 이용해 이 특정 영상과 연관이 있는 영상 ID를 Youtube Video Tree의 다음 level에 배치한다.

또한, 국내 Youtube 동영상 네트워크를 확장하면서 하루에 수집되는 데이터의 크기에 대해 고려해야 했다. Youtube API에 따르면[6], Youtube API의 일일 Quota는 10,000인데 Search API의 Query 요청 1회당 비용이 100이어서 하루에 수집되는 데이

4) https://www.youtube.com/user/theyearinreviewKR
5) https://www.ytn.co.kr/_ln/0103_202012031430012800
6) https://developers.google.com/youtube/v3/determine_quota_cost

터가 적은 문제가 존재했다. 그래서 본 연구는 너비 우선 탐색과 Youtube Search API에서 출력하는 최대 항목 수 필드인 maxResult를 최대인 50으로 설정했다. 해당 방법을 통해 수집 데이터 크기 문제를 해결함과 동시에 연관성이 깊으면서 다양한 카테고리로 확장 가능한 국내 Youtube 네트워크를 구축할 준비를 할 수 있었다.

4. 실 험

본 연구의 데이터셋 구축 과정은 위의 제안한 크롤러 시스템을 이용하여 2021년 3월 3일부터 2021년 4월 7일까지 진행되었다. 해당 데이터셋을 통해 수집된 카테고리 15개를 통해 국내의 Youtube Network의 카테고리 분포를 파악할 수 있었고 해당 분포는 표 1을 통해 확인할 수 있다. 또한, 표 2에서 확인할 수 있듯, 방향 그래프로 국내의 Youtube 네트워크를 표현해 노드는 약 33만 개, 엣지는 약 161만 개가 수집되었음을 알 수 있었다.

표 1. 카테고리별 Youtube 동영상 데이터 분포

Category	#Videos
Entertainment	96,699
People & Blogs	55,539
Music	54,235
Gaming	33,954
News & Politics	22,213
Film & Animation	18,259
Howto & Style	14,788
Education	14,565
Comedy	5,639
Travel & Events	3,352
Sports	3,102
Science & Technology	2,452
Pets & Animals	1,920
Nonprofits & Activism	1,697
Autos & Vehicles	1,059

표 2. 방향 그래프로 표현된 국내 Youtube Network의 특성

Graph Type	Directed Graph
#Node	329,289
#Edge	1,610,684
Average in degree	4.8914
Average out degree	4.8914
Maximum in degree	687
Maximum out degree	279

5. 결 론

본 연구는 Youtube Search API와 너비 우선 탐색(BFS)을 활용해 Youtube에 존재하는 국내의 동영상을 다양하게 수집하는 방법을 제안했다. 이를 통해 국내에서 진행되는 Youtube 연구에 있어서 본 연구가 제시하는 크롤러 시스템을 통해 국내에서 게시된 동영상의 정보를 어떻게 수집해야 하는지 방향성을 제시했다. 또한, 후에 국내 Youtube에 진행될 연구에 도움을 줄 수 있고 연구를 진행하며 설정했던 목표에 부합하도록 수집한 데이터를 공개한다.

본 연구에서 수집된 국내 Youtube 동영상 데이터는 다양한 카테고리의 영상을 수집하였지만, 상위 3개 카테고리인 Entertainment, People & Blogs, Music 의 영상이 다른 카테고리 영상보다 상대적으로 많이 수집됨을 확인했다. 향후 연구에서는 본 연구에서 수집된 카테고리를 포함한 Youtube에서 제공하는 모든 카테고리별로 인기 동영상을 수집해 Youtube 데이터셋을 확장하고자 한다. 그리고 언어권 별로 Youtube 동영상 정보를 수집하는 방법을 제안하고자 한다. 또한, 현재 데이터셋의 VideoID 필드를 활용해 각 영상의 댓글을 수집할 수 있는 Youtube Comment API를 이용해 한국어 댓글을 수집하고자 한다. 이를 통해 한국어 댓글 데이터셋을 구축해 한국어 감성 분석 연구에 이용하고자 한다.

References

[1] 조아라, "'먹통 논란' 유튜브…한국인 가장 오래 사용하는 앱 1위", 한국경제, 2020년 12월 15일 수정, 2021년 4월 6일 접속, https://www.hankyung.com/it/article/202012150863g

[2] 김희숙, "데이터 마이닝을 이용한 유튜브 인기 동영상 콘텐츠 분석", 디지털콘텐츠학회논문지, VOL. 21, NO. 4, PP. 673-681, 2020

[3] 정지원, 이재영, 임춘성, "Youtube 영상에 대한 카테고리별 특성 및 사용자 반응성 분석에 관한 연구", 디지털콘텐츠학회논문지, VOL. 20, NO. 12, PP. 2573-2581, 2019

[4] 유소엽, 정옥란, "사용자의 소셜 카테고리를 이용한 유튜브 동영상 추천 알고리즘", 정보과학회논문지, VOL 42, NO. 05 PP. 0664~0670, 2015

[5] William Hoiles, Anup Aprem, Vikram Krishnamurthy, "Engagement and Popularity Dynamics of Youtube Videos and Sensitivity to Meta-Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 7, PP. 1426-1437, 2017

[6] Tao Li, Lei Lin, Minsoo Choi, Kaiming Fu, Siyuan Gong, Jian Wang, "Youtube AV 50K: An annotated Corpus for Comments in Autonomous Vehicles", 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2018