유튜브 댓글 뷰어 시스템

구태형
경희대학교 소프트웨어융합학과
kth0321@khu.ac.kr

Youtube Comment Viewer System

Taehyung Koo
Department of Software Convergence, Kyung hee University

요 약

유튜브는 가장 높은 점유율을 갖고 있는 동영상 플랫폼이다. 유튜브는 동영상에 대한 사람들의 반응을 확인하기 위해 댓글 기능을 제공한다. 댓글의 개수가 많지 않다면 사용자들의 반응을 확인하기 쉽지만, 댓글이 많아지면 사용자들은 댓글창 상단에 있는 댓글에만 주로 노출이 되고 다른 대다수의 댓글 내용을 확인하기 어렵다는 문제가 있다. 이는 사용자들에게 노출되는 일부 댓글이 전체 댓글을 대표하는 것처럼 보여 정보의 왜곡이 발생될 수 있다는 문제점이 있다. 본 연구에서는 유튜브에서 제공하는 댓글을 모아서, 사용자에게 시각화 정보를 제공하는 크롬 확장 프로그램, 유튜브 댓글 뷰어를 개발하였다.

1. 서 론

유튜브는 한국에서 가장 높은 사용시간을 보이며 사람들 사이에서 높은 사용률을 자랑하는 플랫폼 중 하나이다.[1] 유튜브는 사용자들의 의견 공유와 양질의 콘텐츠 생성을 위해 댓글 기능을 제공한다. 댓글은 영상 크리에이터가 콘텐츠의 방향을 결정하는 데 도움을 줄 수 있고, 시청자들이 댓글을 통해 추가적인 정보를 얻거나 영상에 대한 다른 사람들의 반응을 확인할 수 있도록 도와준다.

영상에 대한 사용자들의 평가가 50개 이하라면 사용자는 큰 힘을 들이지 않고 댓글을 확인할 수 있다. 하지만 사람들의 댓글이 많아질수록 사용자가 직접 댓글을 확인하는 데는 한계가 생긴다. 유튜브의 댓글은 좋아요 수를 기반으로, 인기순 정렬로 사용자들에게 노출이 된다. 노출이 되는 댓글은 계속해서 사용자에게 노출이 되지만, 다른 사용자들에게 노출되지 않는 댓글은 점점 노출 빈도가 줄어들게 된다. 일부 댓글만이 상단에 노출되게 되면 사용자들은 영상에 대한 다양한 의견을 확인할 수 없게 되고, 이는 정보의 왜곡이나 편향으로 이어줄 수 있다. 본 연구에서는 사용자가 상단에 노출되는 일부 댓글뿐만 아니라 더 다양한 댓글까지도 확인할 수 있는 방법을 제안한다.

댓글을 분석하기 위해 다양한 분석 방법이 제시되었다. [1], [2]은 긍부정 라벨링을 통해 긍정 댓글과 부정 댓글을 구분하였다.[2][3] [3]는 감정 클래스를 더 세밀하게 나누어 클래스를 구분하였다.[4] [4]는 관심주제에 따른 감성분석과 댓글 텍스트를 기반으로 한 유아동의 관심 주제 변화를 나타냈다.[5] [5]는 토픽 모델링을 통해 2개의 영상에 대해 주제를 추출하고 키워드 중요도를 비교하였다.[6] 리뷰 데이터 분석을 수행하는 (주)크리마는 키워드를 기반으로 리뷰의 분석을 하지만패션과 뷰티라는 한정된 도메인에서의 리뷰 분석을 수행한다.[7]

그러나 선행 연구들은 영상을 보며 실시간으로 실행할 수 없거나, 특정 영상만을 대상으로 비교를 진행한다. 또한, 연구들은 영상에 대한 사람들의 감정을 긍정과 부정으로 분류하여 나타내거나 주제(도메인)를 한정하여 분석을 진행하였다.

선행 연구들은 영상을 보며 실시간으로 다수의 사람들의 반응을 확인하는 데 한계가 있다. 또한 감성분석이나 토픽 모델링을 통해서는 다양한 사람들의 의견을 보여주는 데 한계가 존재한다. 본 연구에서는 사용자가 영상을 보며, 영상에 대한 다수의 사람들의 반응을 모아서 확인할 수 있는 크롬 확장 프로그램을 개발하였다.

2. 유튜브 댓글 처리 모델

[그림 1]에서 제시된 바와 같이 프로그램은 유튜브 페이지에서 DOM을 처리하는 클라이언트 프로그램과 텍스트 데이터를 처리하는 서버로 구성된다.

1) https://www.madtimes.org/news/articleView.html?idxno=18289
2) https://www-dbpia-co-kr-ssl.webgate.khu.ac.kr/journal/articleDetail?nodeId=NODE11183879
3) http://nciss.or.kr/xml/xmldom.asp?xmlidx=NCISS1091_767
4) https://www-dbpia-co-kr-ssl.webgate.khu.ac.kr/journal/articleDetail?nodeId=NODE11113834
5) https://www-dbpia-co-kr-ssl.webgate.khu.ac.kr/journal/articleDetail?nodeId=NODE09272518
6) https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002844363
7) https://store.cafe24.com/kr/story/1788

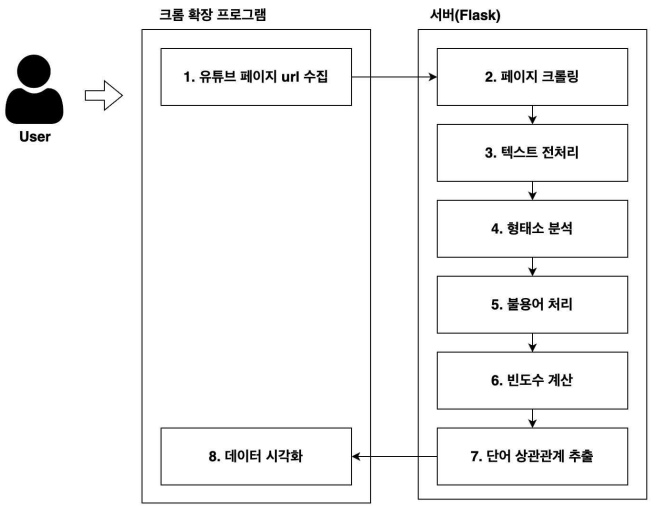

그림1. 유튜브 댓글 뷰어 기능 구조도

2.1. 클라이언트와 서버의 통신

크롬 브라우저의 manifest.json을 활용하여 크롬 확장 프로그램을 클라이언트 애플리케이션으로 활용했다. 크롬 확장 프로그램은 content script와 background script로 구성된다. content script는 사용자 페이지의 DOM을 조작하고, background script는 비동기 통신을 이용해 서버와 통신하며 필요한 정보를 json 형태로 반환받는다. content script는 서버로부터 응답받은 json 데이터를 d3.js를 이용해 시각화하였다.

서버 프로그램은 flask를 활용하였다. flask는 가볍고 url 라우팅이 쉽다는 장점이 있어 필요한 기능만을 빠르게 구현할 수 있다. 본 연구에서는 url을 전달받으면 해당하는 페이지의 크롤링을 진행하고 텍스트 처리만 진행하기 때문에 빠르게 구현할 수 있는 flask를 사용하였다.

2.2. url 수집

사용자의 PC가 아닌 서버에서 데이터의 처리를 하기 위해 사용자로부터 url을 전달받아야 한다. 크롬 확장 프로그램은 크롬 API를 이용해 페이지의 url을 전달받고 이를 서버로 전송한다.

2.3. 웹 페이지 크롤링

서버는 클라이언트로부터 전달받은 url을 이용해 크롤링을 수행한다. 유튜브의 웹 페이지는 사용자가 접속할 때 전체 페이지를 바로 로드하지 않고 사용자의 스크롤 동작에 따라 동적으로 페이지 요소를 로드한다. selenium을 이용해 웹 페이지 스크롤 동작을 수행하고 beautifulsoup를 이용해 페이지 요소를 파싱한다.

태그의 아이디를 이용해 댓글에 해당하는 텍스트를 크롤링한다. 크롤링하는 댓글의 개수는 최대 500개로 설정한다. 크롤링을 할 때는 최대한 다양하게 사람들의 반응을 수집하기 위해 댓글에 대한 좋아요 수는 배제하였다.

2.4. 텍스트 처리

텍스트는 한국어만을 포함하는 댓글로 한정하였다. 텍스트는 3단계에 걸쳐 진행한다. 먼저 텍스트에 포함되는 모음 또는 자음으로만 구성된 문자, 2개 이상의 공백 문자, 개행문자, 특수기호, url주소, 시간를 제거한다. 다음으로 Okt 형태소 분석기를 이용해 명사에 해당하는 키워드만을 추출한다. 불용어 사전에 포함되는 키워드를 제외시킨다. 전처리를 끝낸 데이터는 댓글별로 키워드를 포함한 리스트로 저장된다.

2.4.1. 불용어 사전

불용어는 2가지 방법으로 수집하였다. 먼저 오픈소스에 공유된 불용어 사전을 수집하였다.[8] 다음으로 무작위로 유튜브 페이지를 선별하고, 형태소 분석기를 이용해 각 단어의 빈도수를 계산하여 BoW를 생성, 단어와 문서들에 대해 TF-IDF를 수행하였다. 단어의 임계값을 설정하고 해당 임계값보다 낮은 단어 중요도를 갖는 단어들을 불용어로 설정한다.

2.5. 상위 빈도수 키워드 추출과 상관관계 분석

전처리된 데이터를 이용해 빈도수 계산과 상관관계 분석을 진행한다. 리스트로 된 모든 데이터를 합쳐 빈도수가 가장 높은 상위 10개의 키워드를 선정한다. 다음으로 각 댓글들을 순회하며 해당하는 키워드와 상관관계가 0.3 이상인 키워드들을 뽑아 json 형태로 저장한다. json 데이터는 {기준키워드(상위 빈도수 키워드), 연관된 키워드, 상관계수}를 포함한다. 이를 재귀적으로 반복하며 연관된 키워드를 저장해간다. 데이터의 양이 많아지면 사용자가 데이터를 확인하는 데 어려움이 있기 때문에 기준이 되는 키워드는 10개, 연관된 키워드는 최대 5개가 되도록 제한하였다.

2.6. 데이터의 시각화

클라이언트는 서버로부터 응답받은 json 데이터를 d3..js를 이용해 워드 네트워크로 시각화한다.

3. 실 험

본 연구는 상단에 노출되는 댓글과 상단에 노출되지 않는 댓글간에는 차이점이 있다는 가정을 하고 진행하였다. 먼저 가설을 확인하기 위해 댓글 상단에 노출되는 댓글에 자주 나타나는 키워드와 상단에 노출되지 않는 댓글에 자주 나타나는 키워드의 빈도수 차이를 확인하였다.

다음으로 오픈소스에 올라와 있는 불용어 사전에 추가적으로 목적에 맞게 키워드를 추가해주었다.

마지막으로 댓글에 자주 나타나는 키워드와의 상관관계가 높은 키워드들을 선택하는 실험을 진행하였다.

3.1. 댓글 개수에 따른 상위 빈도수 키워드의 차이

실험은 상단에 노출되는 사람들의 반응과 상단에 노출되지 않는 사람들의 반응에 차이가 있다고 가정한다. 상

8) https://www.ranks.nl/stopwords/korean?fbclid=IwAR2ExNUknGf4bOHA3cECFrv50f8YO2WOTEV4XKP5iDFAANYFWJ1PbMu9j_k

단에 노출되는 댓글의 수는 스크롤 2번을 내려 얻을 수 있는 50개로 설정하였다. 상단에 노출되지 않는 댓글의 수는 5000개로 설정하였다. 표1은 모병제를 주제로 한 동영상에 대한 상위 20개의 키워드이다.

50개 댓글 키워드	500개 댓글 키워드
가지 *	국방 *
강제 *	군대
군대	군인
군복무 *	나라
군인	남녀 *
나라	남성 *
남자	남자
문제	모병 *
병사	문제
사람	병사
생각	사람
소리 *	사병 *
여성	생각
여자	여성
예산 *	여자
월급	월급
전쟁 *	의무 *
포로 *	이준석 *
한국 *	징병제 *
혜택 *	페미 *

표 1. 댓글 개수에 따른 키워드 차이
(* 은 하나의 집단에만 포함된 키워드를 나타낸다)

상위 노출 댓글에만 포함되는 키워드에는 **(가지, 강제, 군복무, 소리, 예산, 전쟁, 포로, 한국, 혜택)**이 있다. 반면 5000개 댓글을 대상으로 키워드의 빈도를 계산했을 때만 포함되는 키워드에는 **(국방, 남녀, 남성, 모병제, 사병, 의무, 이준석, 징병제, 페미)**가 있다. 11개의 키워드가 공통되지만 두 댓글 집단은 9개의 다른 키워드를 포함한다. 5000개 댓글 집단에서 가장 적은 빈도로 나온 키워드는 305번 등장한 사병이란 키워드이다. 반면 500개 키워드 집단에서는 많은 빈도로 나왔지만 5000개 키워드 집단에서는 나오지 않은 키워드들과 빈도수는 ('전쟁', 298), ('소리', 295), ('예산', 251), ('혜택', 233), ('한국', 139), ('강제', 126), ('가지', 100), ('군복무', 90), ('포로', 51)이다. 이는 상단에 노출되는 빈도가 높은 키워드가 상단에 노출되지 않는 곳에는 노출되는 빈도가 낮을 수 있음을 나타낸다.

3.2. 불용어 사전 생성

오픈소스에 공유된 불용어 사전은 명사 이외에도 다른 형태소를 갖는 키워드가 많이 포함되어 있어 무작위로 유튜브 페이지를 선별하고 형태소 분석기를 이용한 후 페이지와 키워드들에 대해 TF-IDF를 계산, 임계값보다 낮은 키워드 중요도를 갖는 키워드들을 선별하였다.

유튜브 페이지는 50개 이상의 댓글을 포함하고 한국어로 작성된 댓글을 포함하는 페이지로 수집하였다. 댓글은 2.4에서의 텍스트 처리와 동일하게 수행하였다. 초 500개의 유튜브 페이지를 대상으로 실험을 진행하였다.

```
각자 각종 거의 게다가 겨우 고로 과연 그때 기타 끙끙
남짓 너희 누구 다른 다만 다섯 다소 다수 다음 단잠
단지 당신 당장 더구나 더군다나 더라도 더욱더 독창 동안 둥둥
등등 따라서 따위 또한 뚝뚝 로부터 로써 리에티 마음대로 마저
마치 만약 만일 매번 모노노케 모노노케히메 모두 모두의마블 무드좋앙 무렵
무슨 무엇 바로 반드시 버금 부터 비로소 비록 삐걱 설령
설마 소인 스앵님 시각 시간 실로 심지어 아야 아이 아이야
아하 아홉 약간 어디 어이 어째서 어찌 언제 얼마 얼마나
엉엉 여기 여덟 여러분 여부 여섯 영차 오직 오히려 와르르
우르르 우리 우선 윙윙 은폐 이번 이상 인어공주 일곱 일단
임꼭 자기 자마자 자신 잠깐 잠시 저기 저쪽 저희 전부
전자 전후 조금 조리기구 조차 졸졸 즉시 차라리 차용 참나
창의성 콸콸 타인 하나 하물며 하하 한마디 행각 행방불명 허걱
혹시 혼자 흐흐
```

그림2. TF-IDF를 이용해 추가한 불용어 목록

단어 중요도의 임계값을 0.0001로 설정하였을 때 1114개의 단어를 불용어로 얻을 수 있었다. 반면, 임계값을 0.00007로 설정하였을 때는 134개의 추가적인 불용어를 얻을 수 있었다. 하지만 0.0001로 설정하였을 때는 중요도가 낮은 단어뿐만 아니라 일반적인 명사 단어도 많이 포함되어, 임계값을 0.00007로 설정하여 결과적으로 134개의 불용어를 구할 수 있었다. 그림2는 0.00007로 설정하였을 때 구해진 불용어이다.

3.3. 댓글의 상관관계 분석

3.3.1. TTM(Document-Term Matrix)

문서 단어 행렬(Document-Term Matrix, DTM)[9]은 다수의 문서에서 등장하는 각 단어들의 빈도를 행렬로 표현한 것을 말한다. 각 문서에 대한 BoW를 하나의 행렬로 만든 것으로 생각할 수 있으며, 각 문서의 BoW를 하나의 벡터로 보고, 전체 문서를 하나의 행렬로 표현한다.

상위 빈도수를 가진 키워드를 대상으로 빈도수를 이용해 다른 키워드들과의 상관관계를 분석한다.

4. 유튜브 댓글 뷰어

그림3. 크롬 확장 프로그램 실행 화면

크롬 확장 프로그램은 그림3과 같이 유튜브 동영상 페이지 우측 상단에 워드 네트워크를 시각화한다. 사용자는 전체적인 내용을 확인하기 위해 동영상을 멈출 필요 없이, 동영상을 시청하는 중에 우측 상단에 시각화된 결과

9) https://wikidocs.net/24559

를 이용해 빈도수가 높은 키워드, 그리고 해당 키워드와의 상관관계가 높은 단어를 확인할 수 있다.

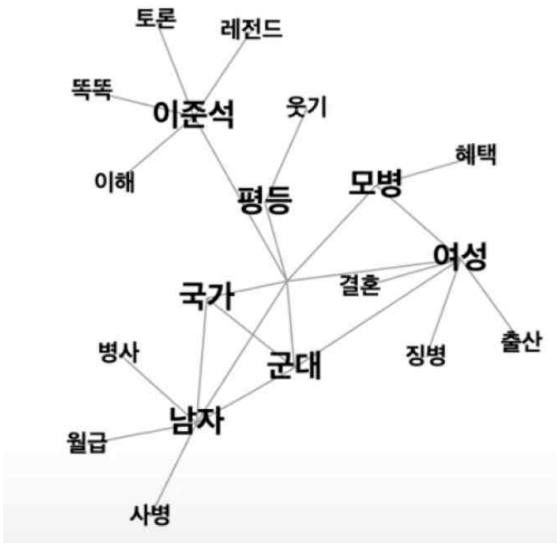

그림4. 사용자에게 노출되는 워드 네트워크

기준이 되는 키워드는 상위 빈도수를 가지는 5개의 키워드를 사용하였다. 그리고 기준 키워드와의 상관관계가 0.3 이상인 단어 중 상관관계가 높은 키워드를 간선을 이용해 연결하고 있다.

그림4는 그림3의 사용자 페이지에 추가된 워드 네트워크를 확대한 것이다. 댓글의 상단에 노출되는 단어도 포함하고 있지만 “이준석”, “모병”과 같은 키워드는 댓글 상단에 노출되지 않는 키워드가 포함됨을 확인할 수 있다.

5. 결과 및 제언

본 연구는 사용자들의 반응을 모아서 요약해주는 시스템을 구축하였다. 시스템은 페이지의 상단에 노출되는 댓글뿐만 아니라 더 아래쪽에 있는 댓글까지도 대신 스크롤을 내려주며 사용자들의 반응을 모으고 정리해준다.

시스템은 상위의 노출되는 단어만을 모아서 보여주는 대신에, 댓글을 모아서 보여주는 기능을 통해 사람들이 유사한 반응을 보이는 영상에서는 내용을 간결하게 정리해주는 역할을 하고, 의견이 분분하거나 토론이 필요한 영상에서는 편향 없이 다양한 의견을 확인할 수 있도록 내용을 제시해준다.

댓글의 수가 많아지면 댓글을 처리하는 시간이 증가한다. 여기에 더해 키워드를 기반으로 한 상관관계 분석은 모든 키워드를 대상으로 연산을 수행하기 때문에 처리 시간이 급증하게 된다는 문제가 있다. 사용자가 영상을 시청하며 동시에 결과를 빠르게 확인하기 위해서는 키워들간의 관계를 나타내는 다른 접근 방식으로의 개선이 필요할 것으로 보인다.

참고 문헌

[1] 유튜브 크리에이터를 위한 유튜브 댓글 분석 시스템
[2] 칼림바 유튜브 채널의 동영상 콘텐츠와 감정표현 댓글 분석
[3] 유튜브 댓글 감정 분석 시각화
[5] 텍스트 마이닝(Text-mining)을 이용한 댓글 분석 연구