JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 23, No. 11: 1321~1329, Nov. 2019

인플루언서 속성 분석 기반 추천 시스템

박정련[1] · 박지원[1] · 김민우[2] · 오하영[3*]

Influencer Attribute Analysis based Recommendation System

JeongReun Park[1] · Jiwon Park[1] · Minwoo Kim[2] · Hayoung Oh[3*]

[1]Undergraduate Student, Department of English Language and Literature, Ajou University, Suwon, 16499, Korea
[2]Undergraduate Student, Department of Digital Media, Ajou University, Suwon, 16499, Korea
[3*]Assistant Professor, DASAN University College, Ajou University, Suwon, 16499, Korea

요 약

소셜 정보망의 발달로 마케팅의 방법도 다양하게 변화되고 있다. 기존의 유명인, 경제적 지원 기반의 성공적인 마케팅방법론과 달리, 최근 인플루언서 기반 유튜브 마케팅이 큰 대세를 이루고 있다. 본 논문 에서는 처음으로 유튜브 양적 정보 및 댓글분석 기반 다각도 질적 분석을 활용하여 54개 이상의 유튜브 채널에서 인플루언서 특징을 추출하고 대표적인 주제들을 모델링하여 개인 맞춤형 영상 만족도 극대화는 물론 기업체가 새로운 아이템을 마케팅 할 때 기존의 인플루언서 특징을 참고하여 새로운 아이템의 영상을 제작하고 배포함으로써 성공적인 홍보 효과를 누릴 수 있도록 보조 수단 제공을 목적으로 한다. 유튜브 채널 별 다양한 영상의 모든 댓글을 각 문서로 가정하고 TF-IDF 및 LDA알고리즘을 적용하여 성능 극대화 향상을 보였다.

ABSTRACT

With the development of social information networks, the marketing methods are also changing in various ways. Unlike successful marketing methods based on existing celebrities and financial support, Influencer-based marketing is a big trend and very famous. In this paper, we first extract influencer features from more than 54 YouTube channels using the multi-dimensional qualitative analysis based on the meta information and comment data analysis of YouTube, model representative themes to maximize a personalized video satisfaction. Plus, the purpose of this study is to provide supplementary means for the successful promotion and marketing by creating and distributing videos of new items by referring to the existing Influencer features. For that we assume all comments of various videos for each channel as each document, TF-IDF (Term Frequency and Inverse Document Frequency) and LDA (Latent Dirichlet Allocation) algorithms are applied to maximize performance of the proposed scheme. Based on the performance evaluation, we proved the proposed scheme is better than other schemes.

키워드 : 인플루언서 속성분석, 추천 시스템, 단어출현빈도에 따른 중요도 측정 기법, 잠재 디리클레 분석

Key word : Influencer Attribute Analysis, Recommender System, TF-IDF, LDA

Received 16 July 2019, **Revised** 11 August 2019, **Accepted** 26 August 2019

*** Corresponding Author** Hayoung Oh (E-mail:hyoh@ajou.ac.kr, Tel:+82-31-219-3560)
Assistant Professor, DASAN University College, Ajou University, Suwon, 16499, Korea

Open Access http://doi.org/10.6109/jkiice.2019.23.11.1321 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

전 세계 76억 인구 중 소셜 미디어(SNS)를 사용하는 인구가 약 34억 명에 달한다. 이를 2018년과 비교하면 약 2억 8천만명이 일년 사이에 SNS를 시작한 것임을 알 수 있다[1]. 특히 소셜 미디어 마케팅 분야에서는 새롭게 등장한 인플루언서(Influencer) 마케팅이 새로운 마케팅 전략으로 떠오르고 있다. 인플루언서란 영향력 있는 개인을 의미하는데 연예인이나 SNS 스타 뿐 만 아니라 일반인도 이에 포함될 수 있다. SNS를 이용하면 일반인도 컨텐츠를 쉽게 소비할 수 있고 동시에 컨텐츠 생산자가 될 수도 있다. 인플루언서 마케팅은 기존의 광고와 다르게 소비자가 직접 콘텐츠를 찾아 소비한다. 그러므로 능동적인 광고 노출이 일어나며, 광고에 대한 거리감 또한 기존 광고들보다 가깝게 다가온다. 따라서 인플루언서 마케팅은 새롭지만 견고한 마케팅 전략으로 자리 잡고 있다. 이에 의해 인플루언서 마케팅은 다양한 플랫폼에서 진행되고 있는데, 우리가 주목해야 할 플랫폼은 '유튜브'이다. 2017년 2월 기준 한국인이 동영상 플랫폼인 유튜브에 소비한 시간이 257억분에 달하는 것으로 나타났다. 이는 2015년과 비교하였을 때, 2년 사이에 유튜브 소비 시간이 3배 이상 늘어난 것임을 알 수 있다[2]. 즉 시간이 지날수록 사람들은 텍스트 컨텐츠보다 영상 컨텐츠에 관심이 많다는 것을 나타내기에 영상 및 사진 기반 유튜브 인플루언서 분석 기법 연구는 매우 중요하다.

제안하는 기법은 사용자들에게는 양적 및 질적으로 인지도 높은 유튜브 채널 및 영상을 추천해 주기 위해, 기업체 들에게는 사용자들이 주로 선호했던 유튜브 채널 및 영상의 속성들 (예: TF-IDF(Term Frequency - Inverse Document Frequency) 및 LDA(Latent Dirichlet Allocation))이 인플루언서 속성과 어떤 연관성이 있는지 제안하여 향후 상품별 유튜브 채널 및 영상 생성시 이를 반영하여 마케팅에 성공할 수 있도록 유도할 수 있기 때문에 의미가 크다. 이는 일반 인플루언서 사용자들이 경험담을 바탕으로 자연스럽게 생성한 영상에 대한 사용자의 반응에서부터 시작됐기 때문에 다음과 같이 분석될 수 있다. "소비자가 이 인플루언서를 좋아할까?", "만약 좋아한다면 이와 비슷한 인플루언서를 추천해주면 추천해준 인플루언서가 홍보한 상품에 대한 관심도 높지 않을까?", "이는 결국 광고 효과를 향상시키고 소비자도 자신이 원하는 광고를 볼 수 있는 것이 아닐까?". 즉, 제안하는 기법에서는 TF-IDF와 LDA 기반 소비자의 실제 반응 분석을 통해 인플루언서 속성 기반 인지도 높은 채널 및 영상을 추천하여 사용자 및 기업체의 만족도를 다각도에서 높인다.

제안하는 기법에서 인플루언서기반 채널 및 영상 데이터를 활용하기 위해, 우리는 유튜브에서 현재 가장 많은 인플루언서를 동원하여 광고를 하고 있는 소개팅 어플인 '스와이프'를 모델로 삼았다. 좋아요 수, 조회 수 등은 영상을 충분히 보지 않거나 끝까지 다 안보고도 단 한번의 클릭으로 유튜브에 표시되는 가시적 데이터이기 때문에, 우리는 영상에 대한 좀 더 정확한 사용자들의 반응을 분석 해보기 위해 함축적 데이터인 사용자들의 댓글 분석을 사용하기로 했다. 즉, 스와이프에 홍보영상에 대한 댓글들을 홍보한 인플루언서에 대한 사용자의 반응이라고 간주하고 TF-IDF와 LDA활용해서 이를 질적으로 분석했다. 이를 위해, 먼저 댓글 토큰화 기반 속성 분석과정(feature analysis) 및 TF-IDF를 통해 유튜브 채널 별 특징 키워드를 뽑는다. 다음, 해당 결과 및 LDA를 통해 영상 간 속성 별 유사도를 구한다.

Ⅱ. 관련연구

최근, 빅 데이터가 기하급수적으로 증가된 환경에서 개인 맞춤 추천은 정보를 효율적으로 검색하고 만족도가 높은 콘텐츠를 발견하기에 좋은 방법이다. 다양한 콘텐츠 중에서 동영상과 같은 동적 데이터를 효율적으로 제공해주는 유튜브는 전 세계적으로 큰 인기를 끌고 있다. 비록 유튜브 플랫폼에서 기본적인 추천 시스템은 탑재되어 인지도 높은 영상들을 추천해주고 있지만, 개인 맞춤형 만족도는 떨어진다. 추천시스템의 ranking을 결정하는데 크게 3가지 요소가 사용될 수 있다. 이는 각각 비디오의 품질, 사용자의 특수성 그리고 다양성이 있다. 비디오의 품질의 경우 조회수, 댓글 수, 즐겨 찾기, 업로드 시간 등과 같은 양적정보들이 포함된다. 사용자의 특수성은 개인의 취향과 선호도와 연관이 있다. 이를 확인하는 방법에는 시청 기록의 조회수나 시청 시간을 등을 고려하는 방법이 있다. 이후 비디오의 다양성을 유지하기 위해서는 동일한 채널의 영상 수를 제한하는 방법이 있다. [2] 저자들은 해당 요소를 고려해서 추천시스템을

디자인했다.

다른 기존 연구로는 휴리스틱 한 판단이 아닌 R기반 데이터 분석을 통해 인플루언서 결정의 유의미함을 검증하려 시도한 기법들이 존재한다. 이때 공통적으로 사용된 데이터는 채널의 총 조회수, 채널의 영 상 개수, 채널의 평균 영상 조회수, 채널 구독자 수, 각 영상의 별, 좋아요 수, 싫어요 수, 영상 길이, 해시 태그 수로 총 9개의 특성(feature)이다. [3] 논문에서는 각 변수의 공분산을 정규화 시킨 후 조회수를 종속 변수로 각 변수간의 상관관계를 분석했으며 채널의 평균 조회수와 채널의 구독자 수가 조회수와 양의 관계를 가진다는 것을 입증했다. 하지만 상관관계의 정도가 미비했으며 해당 통계 자료만으로 영상의 인지도 판단등과 같은 유의미한 결과를 도출 하는 데는 한계점이 있다.

[4] 논문에서는 소개팅 앱 '스와이프'를 이용하여 동영상을 만든 50여개의 채널에 수반된 데이터를 이용했다. 해당 데이터를 통해 영상의 발매 일, 좋아요 수, 싫어요 수, 영상 길이, 해시 태그 수, 채널의 총 조회수, 채널의 영상 개수, 채널의 평균 영상 조회수, 채널 구독자 수를 변수로 주성분 분석(Principal Component Analysis), 다중 변량 회귀 분석을 진행했다. 해당 논문에서도 R 기반 데이터분석을 진행했으며 PCA_1과 PCA_2의 누적 기여율은 전체의 80%이상으로 이 두개의 요소를 비교하여 각 변수들의 중요도를 판단했다. 첫번째 성분의 경우 채널의 평균 영상 조회수와 스와이프 영상의 댓글, 스와이프 영상의 조회수, 스와이프 영상의 좋아요 수 는 각각 -1.25, -2.6, -2.49, -2.83으로 강한 음의 부하량을 보여주었으며 스와이프 비디오의 영상 길이와 스와이프 영상의 해시 태그 개수는 각각 3.31과 1.95로 강한 양의 부하량을 보여 주었다. 두번째 성분의 경우는 조회수와 영상의 구독자수가 강한 음의 부하량을, 영상의 길이와 태그개수가 약한 양의 부화량을 보여주었다. 또한 두 성분을 종합적으로 판단하였을 때 채널의 영상 개수가 변수 중에서 인플루언서 속성 결정 모델에 가장 영향을 많이 미친다는 결론을 도출했다. 다중 변량 회귀 분석의 경우 영상의 개수만이 유의미성을 드러낸다고 한다. 하지만 주성분 분석의 경우 각각의 변수들 간의 차이가 크다는 점이 제한 점이었으며 다중 변량 회귀분석의 경우 일반 화시키기엔 주성분 분석 결과와 불일치성이 존재하기에 해당 연구 역시 한계점이 존재한다.

[5] 논문에서는 유튜브 인플루언서의 지속적인 영향력을 입증하기 위해 각 영상의 시간 별 댓글을 분석했다. 전체 댓글의 수를 기준으로 하여 그룹을 "Large group"과 "Small group"이라 하여 분류했으며, "Large group"의 경우 시간 별 댓글의 수가 비슷한 비율이고, 시간의 흐름에 따라 감소하는 폭이 일정하지만 "Small group"의 경우 시간 별 댓글의 수가 각각 다르고, 시간의 흐름에 따라 감소하는 폭이 일정하지 않다고 가정했다. 이에 따라 논문에서는 group의 "Big3"들을 선출해 감소폭을, 기존 "Large group", "Small group"을 통해 구한 일정함을 평가 지표로 삼았다. 감소폭의 경우 가장 작은 비율일 경우 3%, 가장 큰 경우일 때 2%로 작은 차이를 보여 평가 지표에서 제외되었다. 일정 함의 경우 표준 편차를 이용했으며 "Large group"의 표준편차가 "Small group"의 표준편차에 비해 상대적으로 작다는 결과를 도출했으며 이는 "Large group"의 인플루언서가 지속적으로 사람들의 관심을 안정적으로 받는다는 것을 증명했다. 하지만, 인지도 높은 영상 분석을 통해 추천시스템과의 연결성이 부족하기에 한계점이 존재한다.

[6] 논문에서는 댓글의 분석의 중요성을 설문을 통해 수치화 하며 댓글 분석과 수익 창출의 연관성에 대해 제안했다. 설문조사에 따르면 이용자의 12%인원이 정기적으로 의견을 게시한다고 분석결과를 제시했으며 이는 유튜브 커뮤니티의 규모가 8억 명으로 추정된다는 점을 생각했을 때 약 9천6백만 명이 적극적인 의견을 작성하는 것을 의미한다. 또한 조사 인원의 34%가 의견을 자주 읽었으며 53% 동영상을 본 후 코멘트를 2-3번씩 읽어본다고 분석했다. 댓글의 경우 타당하고 실질적이고 내용 맞춤형인 경우 좋아요 를 많이 받았으며 결국 이것이 동영상의 인기를 높여 수익을 올릴 수 활용될 수 있음을 보였다. 하지만, 해당 기존 연구 역시 인지도 높은 영상 분석을 통해 추천시스템과의 연결성이 부족하기에 한계점이 존재하며, 인간의 모순적인 행동을 분석하는 것에는 한계가 있다.

[7] 논문에서는 인플루언서를 선별을 위해 유튜브 댓글 분석, TF-IDF (Term Frequency inverse Document Frequency)와 연관 규칙 분석을 활용했다. 댓글 분석은 유튜브 랭킹 시스템을 제공하는 social blade에서 각각 "English study", "English education", "English learn"을 검색했을 때 B등급을 받은 유튜브 채널이 사용됐다. 이후 TF(단어 빈도수)와 TF-IDF분석을 비교해 TF-IDF 분석이 보다 유의미한 단어를 도출한다는 점을 입증했다.

예를 들어, TF(단어 빈도수)으로 단어 분석을 진행한 경우 “learn”, “lesson”등 학습에 관련된 일반적인 단어들이나 “love”, “video”, “channel” 등 영상 내용과 아예 상관없는 단어들이 많은 비중을 차지했다. 그에 반해 TF-IDF 분석 기법의 경우 “resort”, “plan”, “proposal” 그리고 “vegan”, “vegetarian” 등의 영상 내용과 관련된 단어들이 나왔으며 이를 통해 첫번째는 여행을, 두번째의 경우 채식주의를 주제로 영상에서 이야기하고 있음을 예상해 볼 수 있다. TF분석에 비해서 TF-IDF는 각 영상의 특징을 추출해서 잘 보여줌을 알 수 있다. 마지막으로, 연관 규칙 분석을 통해 단어 간 연관도를 분석했다. 단어 연관 규칙의 분석의 결과 일반적으로 TF가 높은 단어가 규칙으로 생성되는 것을 볼 수 있었으며 TF-IDF만 이용할 경우 단어 간의 연관성을 휴리스틱하게 판단해야하는 한계가 존재함을 알 수 있었다. 또한 단어 연관 규칙의 분석의 경우에는 일반적으로 단어들의 빈도수에 기인해 규칙을 생성하기 때문에 유튜브 환경내에서는 큰 의미 없는 단어들만 반복되는 것을 알 수 있었다. 하지만, 해당 논문에서는 두 방법을 접목해 인플루언서 선정에 활용해 볼 수 있다는 것을 발견했기에 제안하는 기법의 초석을 제공한다고 생각된다.

Ⅲ. 본 론

3.1. 연구의 배경 및 제안하는 기법의 초기 실험 결과

본 실험을 위해 사용한 데이터는 어플 ‘스와이프’ 채널에 홍보 영상을 제작한 54개 채널 인플루언서 영상 데이터를 이용했다. ‘스와이프’ 영상은 일반 인플루언서들이 직접 제작하고 유튜브에서 다른 사용자들의 반응을 다양한 수치 정보 및 댓글 등으로 관찰해볼 수 있기 때문에 인플루언서 속성 분석 활용에 용이하다. 유튜브에서 가시적으로 제공하는 수치 데이터로는 각 영상의 생성 일, 좋아요 수, 싫어요 수, 영상 길이, 해시 태그 수, 채널의 총 조회수, 채널의 영상 개수, 채널의 영상 평균 조회수 및 채널 구독자 수 등에 해당된다.

제안하는 기법의 1차적인 분석은 유튜브에서 제공하는 다양한 수치 데이터들을 이용하여 이들의 상관 관계를 살펴보는 것이다. 이때, 설정한 가설은 “각 수치 데이터들의 상관관계가 높으면 무조건 해당 영상의 인지도가 높다고 판단할 수 있는가?”이다. 일반적으로 일반 사용자들은 유튜브의 수치 데이터 기반 랭킹 시스템만 믿고 채널 및 영상의 질을 판단하기 쉽다. 하지만, 유튜브 영상은 한번의 클릭으로 채널의 영상 평균 조회수가 증가되기 때문에 사용자가 실제로 영상을 관심있게 봤는지, 영상에서 어떤 부분이 특히 인기가 높았는지, 단순히 클릭만 한 게 아니고 일정시간 머물렀는지 등의 실제 재생 시간은 알 수 없다. 게다가 유튜브의 경우, 각 영상별 주로 시청한 시청자들의 연령층, 성별 등의 통계 및 시각화 정보 (예: demographic information)에 접근이 불가능하기 때문에 영상이 주로 어떤 사용자 계층에서 인지도 높게 시청 됐는지도 알 수 없다.

결과, 본 연구에서는 1차적으로 각 변수들을 대상으로 다 변량 분석(Multivariate Data Analysis: MDA) 분석 중 주성분 분석(Principal Component Analysis: PCA) 및 회귀 분석을 통해 주요 변수를 뽑아내고, 변수 간의 관계를 살펴보기 위해 변수들 간 상관 관계 분석을 진행한다. PCA는 확률론에서 2개의 확률변수의 상관 정도를 나타내는 값인 공분산 행렬을 이용해 차원 축소하는 분석 기법이다. 해당 기법을 통해 각 확률 변수 별로 가장 높은 상관관계를 지니는 결과 값을 알 수 있고 이에 따라 중요 확률 변수를 선별할 수 있다. 공분산이 양의 수치를 지닐 때 두개의 확률 변수는 하나의 변수가 증가할 때 다른 변수도 증가함을 의미한다. 따라서 음의 수치를 지닐 경우 두 확률 변수는 반비례한다고 볼 수 있으며 확률 변수 $E[X]$, $E[Y]$가 있을 때 공분산의 기대값은 식(1)과 같이 나타낼 수 있다.

$$\mathrm{Cov}[X,Y] = E[(X-E[X])(Y-E[Y])] \quad (1)$$

회귀 분석이란 변수 사이의 관계를 알아내는 방법으로써, 특히 선형 관계를 밝히는 데 유용하다. 다중회귀 분석은 하나의 종속변수에 복수의 독립변수가 미치는 영향을 파악하는 회귀분석으로 공식은 식 (2)와 같다.

$$\mathrm{E}(y_i|x_{li}, \cdots, x_{ki}) = \beta_0 + \beta_i x_{li} + x_{li} + \cdots + \beta_k x_{ki} \quad (2)$$

제안하는 기법에서 다음과 같이 활용했다. 첫째, 각 채널의 평균 조회수가 해당 영상의 조회수에 영향을 끼치는지 유무를 검증하기 위해 공분산과 상관관계를 구했다. 공분산과 상관관계는 Pearson상관계수를 이용하였으며 이를 추출하기 위한 결측 값의 처리는 use의 세부 옵션 중 Complete.obs를 이용하여 결측 값이 있는 경

우는 모두 제거된 상태에서 상관계수를 계산했다. 그림 1과 같이 공분산 값은 1182963412246이라는 값이 나왔으며 이는 두 변수가 양의 상관관계를 가진다는 것을 알 수 있다. 하지만, 두 변수 간 상관관계의 정도를 알기 위해 공분산을 표준화한 상관 계수를 이용하였을 때 상관계수의 값은 0.3543527이라는 낮은 양의 값으로 서로 상관관계가 존재하나 약하다는 것을 알 수 있다. 둘째, 채널의 전체 구독자 수와 해당 영상의 조회수의 공분산 및 상관관계를 구하다. 그림 2와 같이 공분산 값은 84210636999라는 값이 나왔고 이때 상관관계 값은 0.3917762라는 채널 평균 조회수보다는 살짝 높지만 여전히 낮은 양의 상관관계를 보여주고 있다.

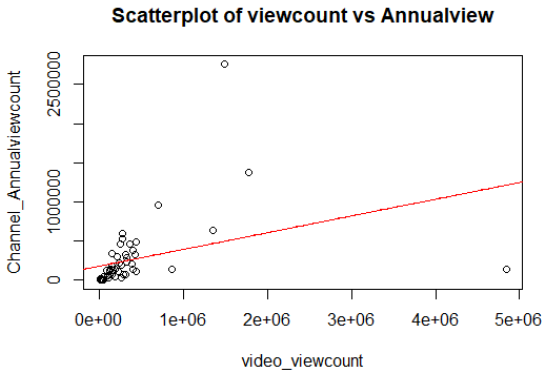


Fig. 1 Scatterplot of Video View Count vs Channel Annual View Count

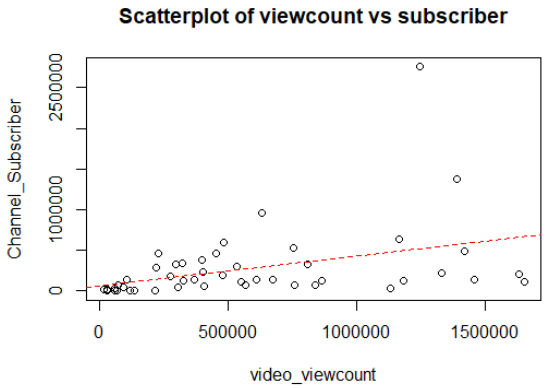


Fig. 2 Scatterplot of Video View Count vs Channel Subscriber

결과, 유튜브 가시적 수치 데이터 속성분석들만으로는 인지도 높은 인플루언서 채널 및 영상 추천이 한계가 있음을 알 수 있다. 즉, 1차적인 연구 결과에서 볼 수 있듯이 유튜브 시스템에서 명시적으로 보이는 단순한 통계정보기반 가정 사항들(예: 평균 조회 수가 높은 채널에 등록된 동영상은 인지도가 높을 것이다. 혹은 채널 구독자 수가 높은 채널에 등록된 동영상은 인지도가 높을 것이다. 등등)은 큰 의미가 없으며 이 정보만으로 유튜브 채널 인플루언서라고 결론지을 수 없다는 것이다. 하지만, 본 연구의 1차적인 결과를 바탕으로 향후 유튜브 댓글 분석을 활용한 함축적 데이터 분석이 다각도로 수행 되야 하며 이를 바탕으로 유튜브 인플루언서 속성 결정 연구에 초석(motivation)이 될 수 있다는 것이다.

제안하는 기법의 2차적인 분석은 TF-IDF(Term Frequency - Inverse Document Frequency) 및 LDA (Latent Dirichlet Allocation)을 활용하여 댓글에 대한 텍스트 질적 분석을 진행하는 것이다. TF-IDF는 여러 문서로 이루어진 문서 군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 이는 문서들 사이의 비슷한 정도를 구하거나, 문서의 핵심어 추출 및 검색 결과 순위 결정을 위해 단어의 특정 문서 내 중요도를 산출하는 통계적 가중치 알고리즘으로 주로 사용된다. 즉, TF(Term Frequency)는 단어 빈도를 의미하며, 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있지만 단어 자체가 여러 문서 군 내에서 자주 사용되는 경우, 이것은 그 단어가 흔하게 등장한다는 것을 의미하며 tf(t,d)와 같이 함수로 표현할 수 있다. tf(t,d)에서 t는 특정 단어를, d는 문서를 의미하기 때문에 결과값은 d에서 t가 등장하는 빈도수를 의미한다. 이와 상응하는 개념으로 DF(Document Frequency), 문서 빈도가 있다. 이는 특정 단어 t가 얼마나 많은 문서 등장하는지를 나타내는 수치이다. 즉, DF가 높은 단어일수록 많은 문서에서 사용되었기 때문에 흔한 단어가 되고 이 단어로는 문서들 간의 차이를 확인하기 힘들어진다. 즉, DF가 낮을수록 핵심어가 될 확률이 높아지는데, 이 점을 이용하여 DF의 역수인 IDF(Inverse Document Frequency), 역 문서 빈도를 식 (3)과 같이 동시에 고려한다.

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{3}$$

여기서 |D|는 전체 문서 집합 D의 크기, 즉 전체 문서의 수를 의미하며, |{d∈D : t∈d}|는 단어 t가 포함된 문서의 수를 의미한다. 특정 단어가 전체 문서 안에 존재하지 않을 경우 분모가 0이 되기 때문에 이를 방지하기

위하여 보통 1+|{d∈D : t∈d}|를 사용하게 된다.

위의 내용을 바탕으로 중요한 핵심어를 추출할 때 주로 사용되는 TF-IDF를 구할 수 있게 된다. 이는 TF와 IDF를 곱한 통계적 수치로 이 값이 클수록 핵심어일 확률이 높아진다. 이 수치는 IDF를 반영하기 때문에 식(4)와 같이 모든 문서에서 흔하게 나타나는 공통 단어를 쉽게 걸러낼 수 있게 된다.

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D) \tag{4}$$

제안하는 기법에서는 유튜브 채널 별 동영상들의 모든 댓글들을 크롤링하여 한 개의 문서로 가정한다. 해당 문서에 TF-IDF적용 결과, 유튜브 채널 별 인플루언서 속성을 추출할 수 있다. 그림 3과 4는 해당 결과들의 일부분을 보여준다. 빈도수가 높은 단어들은 주로 일반 사용자들이 일상적으로 사용하는 용어들 혹은 영상의 등장인물의 이름 등으로 인플루언서 채널 영상들의 속성 및 특징들이라고 볼 수 있다.

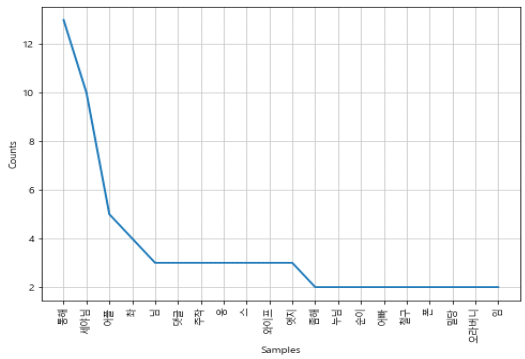


Fig. 3 Comments analysis of TF-IDF in a specific YouTube channel (1)

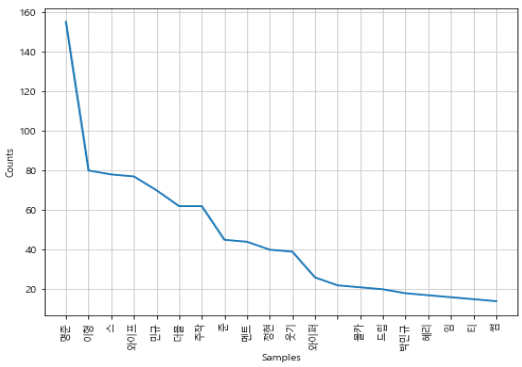


Fig. 4 Comments analysis of TF-IDF in a specific YouTube channel (2)

TF-IDF적용 후 각 채널의 영상들의 인플루언서 속성들의 유사도를 구해보면 주제로 묶일 수 있는데 이를 LDA로 분석한다. LDA는 주제들이 디리클레 분포를 따른다고 가정하기 때문에 잠재 디리클레 할당이라고 불리우며, 각 문서에 주제별 발견된 단어 수 분포를 분석함으로써 해당 문서가 어떤 주제들을 함께 다루고 있을지를 예측할 수 있는 확률적 토픽 모델 기법 중 하나이다. LDA는 일반적으로 문자 기반의 이산 자료들에 대한 확률적 생성 모형이지만, 사진(이미지), 소리 등의 다른 이산 자료들은 물론 연속 자료들에 대해서도 쓰일 수 있고 또한 다항 분포가 아닌 자료들에 대해서도 적용 할 수 있는 가능성이 있다. LDA에는 몇 가지 가정이 있는데 그 중 중요한 것은 단어의 교환성(exchangeability)이며 이는 '단어 주머니(bag of words)'라고 표현하기도 한다. 교환성은 단어들의 순서는 상관하지 않고 오로지 단어들의 유무만이 중요하다는 가정으로 예를 들어, 'Snow is white'와 'White is snow' 간에 차이가 없다고 생각하는 것이다. 결과, 단어의 순서를 무시할 경우 문헌은 단순히 그 안에 포함되는 단어들의 빈도 수만을 가지고 표현이 가능하게 되며, 이 가정을 기반으로 단어와 문서들의 교환성을 포함하는 혼합 모형을 제시한 것이 바로 LDA이다. 하지만 단순히 단어 하나를 단위로 생각하는 것이 아니라 특정 단어들의 묶음을 한 단위로 생각하는 방식(n-gram)으로 LDA의 교환성 가정을 추후 확장시킬 수도 있다.

예를 들어 [어밴져스, 스파이더맨, 헐크, 겨울왕국]등의 단어들이 많이 쓰인 문서와 [셰익스피어, 톨스토이, 파우스트, 안데르센]등의 단어들이 많이 쓰인 문서가 있다고 해보자. 우리는 첫번째 문서가 영화에 관련된 내용이고, 두번째 문서가 문학에 관련된 내용으로 추측할 수 있다. 그 이유는 우리는 문서에 들어가 있는 단어들이 해당 주제들을 표현하고 있음을 알기 때문이다. 이러한 사전 정보가 없다면 주제를 예측할 수 없다. 하지만 사전정보가 없어도 단어들과 주제들은 연관성이 있기 때문에 라벨링 된 단어들의 가능 도를 활용한 지도 학습은 물론 비지도 학습으로도 결정할 수 있다.

결과, 궁극적으로 본 연구에서는 54개의 영상 별 TF-IDF(Term Frequency - Inverse Document Frequency) 및 LDA(Latent Dirichlet Allocation)을 활용하여 주요 단어 및 주제를 선별하고 인플루언서 영상간 유사도를 고려해서 추천시스템을 구현한다. 이후 댓글을 중복적으로

작성한 유저 정보를 수집해 영상 간 유사도를 입증하고 해당 사용자를 대상으로는 보다 개인 맞춤형 추천시스템을 구현하는 것을 목적으로 했다.

3.2. 제안하는 기법

제안하는 기법의 대표적인 그림은 그림 5와 같다. 데이터 크롤링, 수치 데이터 분석 (양적 분석) 및 유튜브 댓글 (텍스트 질적 분석)만 수행하여 인지도 높은 영상을 추천해주는 기존연구들과 달리 제안하는 기법에서는 은닉 데이터 분석을 활용하여 인플루언서 유튜브 채널 및 영상을 추천하는 것을 목표로 한다. 이를 위해, 영상들 및 댓글들의 내용의 상관관계가 평균적으로 높은 유튜브 채널 분석 후, 해당 결과를 TF-IDF(Term Frequency - Inverse Document Frequency))로 검증하여 일치 도가 높을 수도 높은 가중치를 부여했다. 이는 사용자가 단순히 영상을 일정시간 이상 보고 관심도가 높음은 물론 해당 영상 및 채널의 유일한 특징 및 속성으로도 일치하기 때문에 다른 사용자들에게도 인플루언서 유튜브 채널 추천 시 활용되면 더욱 좋다는 것을 의미한다. 특히, 채널 별 TF-IDF(Term Frequency - Inverse Document Frequency) 분석 결과에 최종적으로 LDA(Latent Dirichlet Allocation) 기반유사도 기반 타픽 모데링을 수행하여 최종적으로 인플루언서 채널을 추천해주는 기법을 제안한다.

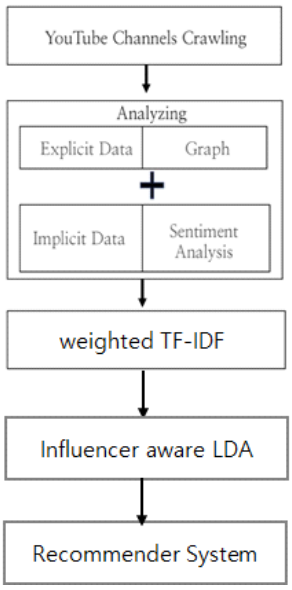


Fig. 5 Proposed scheme

Ⅳ. 성능평가 및 기대효과

해당 실험의 경우 데이터 수집 및 시각화는 해당 기능에 특화된 R언어를, TF-IDF 및 LDA 분석의 경우 분석의 편의를 위해 추가적으로 파이썬 작업이 병행되었다.

제안하는 기법을 바탕으로 실험은 크게 2 단계로 나뉜다. 첫번째는 유튜브에서 제공하는 수치데이터를 통한 변수 간의 관계분석으로 PCA의 경우 scree와 적재그림을 통해 시각화를 진행했다. 다음으로는 변수(수치데이터)별로 상관 계수를 구해 수치데이터를 통해 유의미한 결론을 도출 한다.

두번째의 경우 TF-IDF분석을 통해 채널 별 주요 단어를 추출했으며 해당 단어를 바탕으로 LDA 토픽 모델링을 진행하고 이를 바탕으로 채널 추천 시스템을 구축한다. 유튜브 추천시스템에 대한 평점 정보(rating information)는 없기 때문에 해당 성능 평가의 경우 채널 주제 별로 영상에 영상 내용과 평균적으로 상관관계가 높은 댓글을 작성한 사람은 높은 평점을, 그 반대일수록 낮은 평점을 매겼다고 가정하고 그림 6과 같이 평점 매트릭스를 생성한다. 예를 들어, $User_m$이 유튜브$_1$채널에 등록된 동영상 별 시청 후 영상 내용과 연관성이 높은 댓글을 달았다면 해당 평균값의 정규화(normalization) 값을 $User_m$이 유튜브$_1$채널에 매긴 평점으로 가정한다. 유튜브에서 사용자들은 대부분 댓글을 달지 않는 경향이 더 강하지만 본 연구에서는 어느 정도 인지도 높은 인플루언서 영상들 중 평균 임계치 이상의 댓글들이 달린 채널과 영상들만 고려했기 때문에 추천시스템에서 문제가 되는 sparse matrix problem은 없다고 가정한다. 만약, $User_m$이 유튜브$_1$채널의 여러 동영상에 댓글을 달았다면 평균 정규화 값을 평점으로 활용한다. 결과, 제안하는 기법은 단순히 인지도 높은 채널과 영상을 추천해주는 단계를 뛰어넘어 인플루언서 속성을 가진 채널 및 영상들 중에서 주제별 사용자에게 개인 맞춤형으로 추천해줄 수 있다는 장점을 지닌다. 또한, 일반 사용자들의 만족도는 물론 새로운 아이템에 대해서 인플루언서 기반 유튜브 마케팅을 하고자 하는 업체들에게도 인지도 높은 인플루언서 속성들을 추천해줌으로써 마케팅 효과의 만족도를 극대화할 수 있다.

	$YouTube_1$	$YouTube_2$	$YouTube_3$	...	$YouTube_n$
$User_1$	0	3	4	...	5
$User_2$	1	1	2	...	?
:	:	:	:	...	:
$User_m$	4	?	5	...	4

Fig. 6 Implicit information of YouTube aware recommender system

그림 7, 8 및 9는 제안하는 기법의 pseudo code 및 성능평가 결과를 보여준다. 성능평가를 수행함에 있어 고려할 수 있는 다양한 요소 중에 제안하는 기법에서 고려한 요소들이 4개의 추축(PC1, PC2, PC3, PC4)이 전체 분산의 88% 이상을 차지하기 때문에 인플루언서 탐지 및 이를 고려한 추천시스템에 활용 시 큰 의미가 있음을 보인다. 특히, 제안하는 기법에서 추천한 영상들과 유튜브에서 개인 별 랭킹하는 영상들의 양적 및 질적 정보를 동일한 기준으로 일정 시간 관찰 시, 제안하는 기법에서 제안된 영상들의 인지도의 기울기가 훨씬 높음을 알 수 있었다. 해당 연구 결과를 바탕으로 아이템 홍보의 성공률을 이와 연관성 높은 댓글들을 분석하여 비교해본 결과 제안하는 기법에서 추천하는 영상들의 상관관계가 훨씬 높음을 알 수 있었다.

<textAnalyzer pseudocode>

```
REQUIRE:
    sets of influencer data I
    stopwords S
PROCESS:
    RAED I

    Sort the comments attribute in order lot's of likes
    Select only influencer names and comments from raw data

    FOR comments IN I
      token <- tokenize(comments)
      eliminate token including S

    WRITE tokenized comment datas with frequency

Output: sets of tokenized comment data with frequency D
```

Fig. 7 TextAnalyzer pseudocode of the proposed scheme

<TF-IDF & LDA pseudocode>

```
REQUIRE:
    sets of tokenized comment data with frequency D

PROCESS:
    READ D
    vector <- TfidfVectorizer(D)
    extract 10 topics about vector using LDA
    visualize each 10 topics
```

Fig. 8 TF-IDF & LDA pseudocode of the proposed scheme

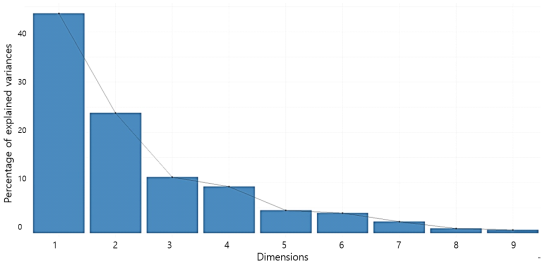


Fig. 9 Performance evaluation of the proposed scheme

Ⅴ. 결 론

일반 사용자들의 만족도는 물론 사업자들의 성공적인 마케팅 전략을 위해 인플루언서 기반 유튜브 채널 및 영상 추천시스템을 제안했다. 이를 위해, 유튜브 54개의 영상에 대한 수치 데이터(조회수, 좋아요 수, 싫어요 수, 댓글 수, 채널의 영상 수, 채널 조회수, 채널 구독자 수, 채널 영상 수, 채널 평균 조회수)를 분석했다. 1차적 결과를 토대로 싫어요 수와 댓글의 수의 상관계수가 높은 것으로 부정적인 댓글이 많을 것이라고 예상했지만 실제 유튜브의 댓글을 분석해보면 그렇지만은 않은 것을 알 수 있었다. 또한 유튜브의 경우 각 채널, 영상 별로 상이한 특성을 지니기 때문에 단순히 수치데이터를 통한 인플루언서 추천은 한계가 있으며 효과적이지 못하다는 한계가 존재한다. 따라서 본 논문에서 영상을 보는 인플루언서 마케팅의 소비자의 반응인 댓글에 대한 분석을 질적으로 수행하고 TF-IDF(Term Frequency - Inverse Document Frequency) 및 LDA(Latent Dirichlet Allocation)을 활용하여 추천시스템을 구현했다.

제안하는 기법에서 추천된 영상들과 유튜브 자체에서 추천하는 영상들을 일정시간동안 양적 및 질적 데이터 분석 비교 시, 제안하는 기법에서 추천해주는 영상들이 사용자 만족도는 물론 마케팅 홍보 성공율 상관관계 측면에서 훨씬 높음을 알 수 있었다. 향후 연구로는 제안하는 기법을 실제 기업체에서 초석으로 활용할 수 있는 방안을 모색하고 인플루언서의 영향력을 다 각도에서 분석해보는 것이다.

ACKNOWLEDGEMENT

This research was supported by the MIST (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute for Information & communications Technology Promotion)" (20150009080031001), Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2017R1D1 A1B03035557) and Ajou University research fund.

REFERENCES

[1] Y. I. Chang, and Y. S. Jung, “A Study on YouTube Product Review Channel Subscribers’ Product Attitude Formation Process,” *The e-Business Studies*, vol. 20, no. 2, pp. 77-97, Apr. 2019.

[2] J. Davidson, B. Liebald, J. Liu, P. Nandy, and T. Van Vleet, “The YouTube Video Recommendation System” in *Proceedings of the fourth ACM conference on Recommender systems*, pp. 293-296, Apr. 2010.

[3] M. Kim, J. R. Park, J. W. Park, and H. Y. Oh, “Channel Attribute Analysis Scheme for Trustworthy Youtube Influencer Detection”, in *Proceeding of The 29th Joint Conference on Communications and Information on Big Data and Social Network*, Apr. 2019.

[4] M. Kim, J. R. Park, J. W. Park, and H. Y. Oh, “Influencer Attribute Decision-Making based on Principal Component Analysis” in *Proceeding of KIPS Conference on Web Science*, pp. 672-674, May. 2019.

[5] J. W. Park, M. Kim, J. R. Park, and H. Y. Oh, “Stable Influencer Selection Criteria Scheme through Youtube Analysis of Hourly Comments” in *Proceeding of Korea Institute of Next Generation Computing*, Apr. 2019.

[6] P. Schultes, V. Dorner, and F. Lehner, “Leave a Comment! An In-Depth Analysis of User Comments on YouTube” in *Proceeding Wirtschaftsinformatik Proceedings*, pp. 659-673, Apr. 2013.

[7] J. R. Park, M. Kim, J. W. Park, and H. Y. Oh “A Study on Tools for Agent System Development” in *Proceeding of KIPS Conference on information system*, pp. 293-295, Apr. 2019.

박정련(Jeong-Ryeon Park)

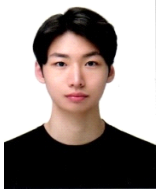
아주대학교 영어영문학과 재학

※관심분야 : 소셜정보망 및 데이터 마이닝

박지원(PARK JIWON)

아주대학교 영어영문학과 재학

※관심분야 : Human-Computer Interaction, Data Science, Systems for Machine Learning, Network 보안



김민우(Minwoo Kim)

아주대학교 디지털미디어학과 재학

※관심분야 : HCI, 소셜 및 금융데이터 분석, 머신러닝 및 딥러닝 적용

오하영(Hayoung Oh)

서울대학교 컴퓨터 공학박사

※관심분야 : 소셜정보망 및 데이터분석