TF-IDF 와 연관 규칙 분석 기반 인플루언서 선별 기법

박정련 [1], 김민우 [2], 박지원 [1], 오하영*
[1] 아주대학교 영어영문학과
[2] 아주대학교 심리학과
*아주대학교 다산학부대학
e-mail : wjdfus0219@naver.com, ski1222@ajou.ac.kr, 1208jw@naver.com, hyoh@ajou.ac.kr

A Study on Tools for Agent System Development

JeongRyeon Park[1], Minwoo Kim[2], Jiwon Park[1], Hayoung Oh*
Dept. of English Language and Literature[1], Dept. of Psychology[2], DASAN University College*
Ajou University

요 약

소셜네트워크서비스(SNS)의 정치, 경제, 사회, 문화 전반에 걸친 영향력이 점점 더 커지고 있는 현실에서 가장 발빠르게 이들 매체를 전략적인 PR 도구로서 이용하고자 노력하는 조직들은 아마도 기업일 것이다. 본 연구에서는 TF-IDF 와 연관 규칙 기반 유투브 인플루언서 선별방안을 제안하여 기업 마케팅의 초석을 제공한다.

1. 서론

소셜네트워크서비스(SNS)의 정치, 경제, 사회, 문화 전반에 걸친 영향력이 점점 더 커지고 있는 현실에서 가장 발빠르게 이들 매체를 전략적인 PR 도구로서 이용하고자 노력하는 조직들은 아마도 기업일 것이다[1]. 특히 대중적인 명성을 지니고 있는 특정 개인를 이용하여 상품 및 서비스, 더 나아가 회사를 홍보하는 celebrity marketing 은 90 년 대 초부터 사용되어 왔으며 소셜미디어의 등장과 함께 인플루언서 마케팅으로 발전되어왔다[2]. 또한 현재 소비자의 모바일 사용 증가는 기존의 광고 형태를 벗어나 소셜미디어를 통한 바이럴 마케팅의 중요성을 대두시켰다[3]. 최근에는 유튜브, 아프리카 TV 를 비롯한 온라인 동영상 콘텐츠 시장이 활성화되면서 1 인 미디어의 크리에이터가 영향력 있는 일반인 인플루언서로 부상하고 있다[4]. 특히 많은 영상 콘텐츠 플랫폼인 '유튜브'는 이용자들이 자체 제작한 동영상을 업로드하고 서로 공유할 수 있는 동영상 공유 웹사이트로 2005 년 시작되었고 2006 년에 Google Inc 에게 인수된 이후에 광고 및 이용자 수익을 증가시키기 위해 더욱 적극적으로 콘텐츠를 수습하고 있다[5]. 하지만 대중적 및 일괄적으로 노출되던 매스미디어와 달리 1 인 미디어의 경우 특정 성향을 가진 사람들에게 노출이 된다. 따라서 기업은 효과적인 광고를 위해서 이전보다는 많은 노력을 요구한다. 유튜브의 경우 자체적으로 구독자 수, 조회수, 좋아요 수 등 다양한 수치 데이터를 제공하지만 이것만으로는 인플루언서를 선정하는데 어려움이 있다. 따라서 본 연구는 보다 효과적인 인플루언서 선정을 위한 방법으로 TF-IDF 를 제안하며 이를 증명하는 것을 목적으로 한다.

2. 실험 단계 및 관련 지식

본 연구에서는 유투브 인플루언서 선별방안을 제안하기 위해 아래와 같은 과정으로 데이터를 수집하고 분석을 수행했다. 유튜브 랭킹 시스템을 제공하는 social blade 에서 각각 "English study", "English education", "English learn"을 검색했을 때 B 등급을 받은 유튜브 채널을 선별 후, 채널 별 20 개의 영상을 선택했다. 유투브 채널 별 특징을 검출하기 위해 상대적 빈도수를 구할 수 있는 TF-IDF (Term Frequency inverse Document Frequency)를 고려했다. TF(Term Frequency)는 특정 단어 t 가 문서 d 에 등장하는 빈도수를 의미하며, tf(t,d)와 같이 함수로써 표현할 수 있다. 즉, TF 가 높은 단어일수록 특정 문서에서 해당 단어가 많이 사용되었다는 것을 의미한다. 따라서 TF 가 높을수록 의미 있는 단어일 확률이 높아진다. 이와 상응하는 개념으로 DF(문서 빈도, Document Frequency)가 있다. 이는 특정 단어 t 가 얼마나 많은 문서 등장하는지를 나타내는 수치이다. 즉, DF 가 높은 단어일수록 많은 문서에서 사용되었기 때문에 흔한 단어가 되고 이 단어로는 문서들 간의 차이를 확인하기 힘들어진다. 즉, DF 가 낮을수록 핵심어가 될 확률이 높아지는데, 이 점을 이용하여 DF 의 역수를 사용한다. 이를 IDF(역문서 빈도, Inverse Document Frequency) 라고 하며 IDF 의 수식은 식(1)과 같다.

$$idf(t,D)=\log\left[\frac{|D|}{|\{d \in D : t \in d\}|}\right] \quad (1)$$

여기서 |D|는 전체 문서 집합 D 의 크기, 즉 전체 문서의 수를 의미하며, |{d ∈D:t ∈d}|는 단어 t 가 포함된 문서의 수를 의미한다. 특정 단어가 전체 문서

안에 존재하지 않을 경우 분모가 0 이 되기 때문에 이를 방지하기 위하여 보통 1+|{d ∈D:t ∈d}|를 사용하게 된다.

TF 와 IDF 를 곱한 통계적 수치로 이 값이 클수록 핵심어일 확률이 높아지기 때문에 이를 바탕으로 중요한 핵심어를 추출할 때 TF-IDF 를 활용할 수 있다. 또한, 동시에 IDF 를 고려하기 때문에 모든 문서에서 흔하게 나타나는 공통 단어를 쉽게 걸러낼 수 있게 된다. 식(2)는 최종적으로 TF-IDF 를 보여준다.

$$tf\text{-}idf(t,d,D)=tf(t,d)*idf(t,D) \quad (2)$$

위의 내용을 바탕으로 연관 규칙 X → Y 대한 평가 척도로는 지지도(support), 신뢰도(confidence) 및 향상도(lift)를 이용했다. 각 규칙에 대한 수식은 다음과 같다. 두 항목 X 와 Y 의 지지도는 전체 거래 건수 중에서 항목집합 X 와 Y 를 모두 포함하는 거래 건수의 비율을 말한다. 지지도는 좋은 규칙(빈도가 많은, 구성비가 높은)을 찾거나, 불필요한 연산을 줄일 때(prunning, 가지치기)의 기준으로 사용한다.

신뢰도(confidence)는 항목집합 X 를 포함하는 거래 중에서 항목집합 Y 도 포함하는 거래 비율 (조건부 확률)을 말한다. 신뢰도가 높을 수록 유용한 규칙일 가능성 높다고 할 수 있다.

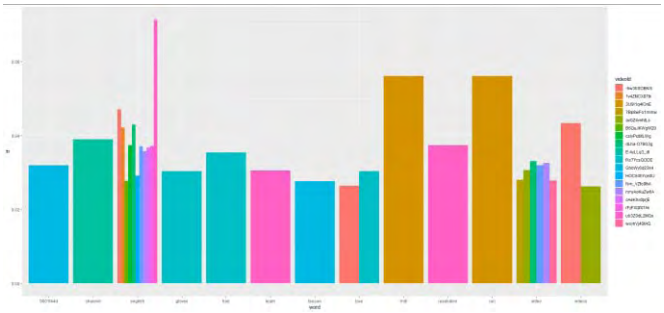
여기서 n(x)는 전체 transation 에서 항목집합 X 를 포함하는 transaction 수다.

향상도(lift)는 항목집합 X 가 주어지지 않았을 때의 항목집합 Y 의 확률 대비 항목집합 X 가 주어졌을 대 항목집합 Y 의 확률 증가 비율을 말한다. 즉, 향상도가 1 보다 크거나(+관계) 작다면(-관계) 우연적 기회(random chance)보다 우수함을 의미한다.

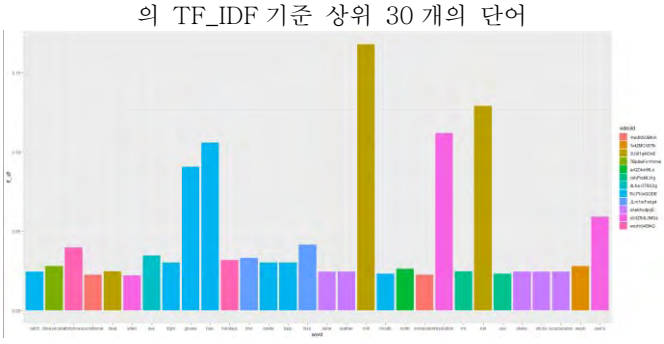
3. 본론

본론에서는 위의 개념을 토대로 실제 데이터를 시각화해 살펴보려 한다.

<그림. 1> channel Learn English With TV Series 의 댓글의 TF 기준 상위 30 개의 단어

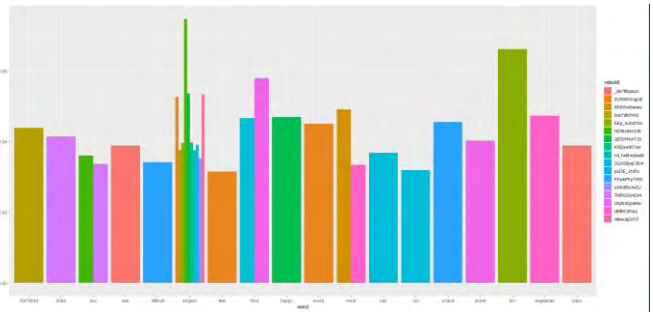


< 그림. 2> channel Learn English With TV Series 의 댓글의 TF_IDF 기준 상위 30 개의 단어

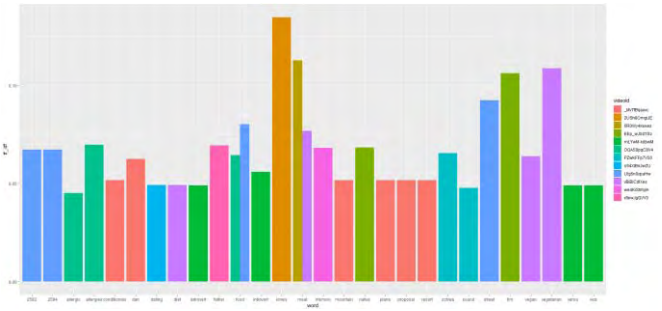


위의 두 표는 채널 Learn English With TV Series 에 대해 TF 와 TF-IDF 를 기준으로 상위 30 개의 단어를 선별한 것이다. TF 를 통해 선별한 단어 같은 경우 20 개의 영상 중 18 개의 영상에 분포해 있다. 해당 그래프에서 가장 높은 TF 값을 갖는 것은 english 로 0.0713 의 값을 가진다. 하지만, 이 단어의 경우 채널에 특성을 지정하기 하기는 너무 광범위하며 여러 영상에서 일반적으로 반복되고 있다. 또한 그외의 단어 역시 “learn”,” lesson”과 같은 일반적인 학습에 관련된 단어나 “love”, “video”,” channel”등 내용과 상관 없는 단어들도 많은 비중을 차지한다. 반면 TF-IDF 을 이용할 경우 특정 영상에서 중복적으로 반복되는 단어를 제외하고 분석이 가능하다. 또한 “vocabulary”, “laine”와 같이 보다 다양한 단어가 등장한다. 본 연구에서 진행한 표본이 채널을 특징을 일반화시켜서 담는 것은 한계가 있지만 stopwords 가 아닌 보다 의미있는 단어가 선별되는 것은 확인할 수 있다. 또한 TF-IDF 와 TF 에서 동시에 높은 수치를 보이는 “mill”, “run”등과 같은 단어에 대한 분석도 가능하다.

< 그림. 3> channel BBC Learning English 의 댓글의 TF 기준 상위 30 개의 단어

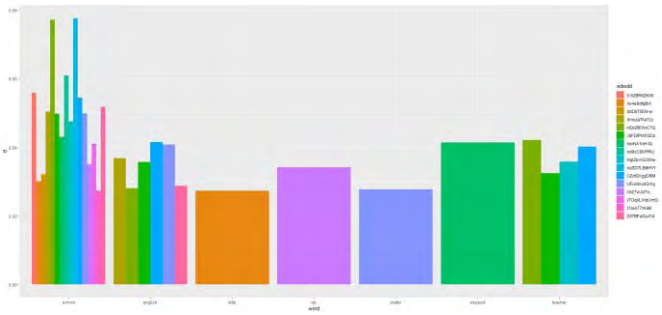


< 그림. 4> channel BBC Learning English 의 댓글의 TF_IDF 기준 상위 30 개의 단어

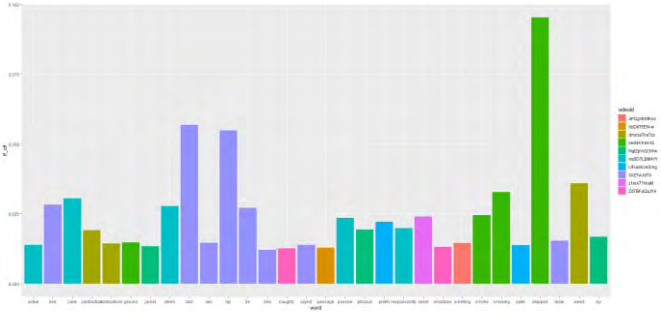


다른 채널에 대해서 데이터분석을 시행한 그림.3 과 4 의 경우도 TF-IDF 의 그래프에서는 “resort”, ” plan”, ” proposal”에서는 여행을 주제로 “vegan”, “vegetarian”에서는 채식주의를 이야기하고 있음을 예상해 볼 수 있다. 또한 이를 통해 해당 유튜브가 영어 교육 방송을 특정 “주제”를 선별해 교육하고 있는 것을 유추할 수 있다.

< 그림. 5> channel Learn English with Emma 의 댓글의 TF_IDF 기준 상위 30 개의 단어



< 그림. 6> channel Learn English with Emma 의 댓글의 TF_IDF 기준 상위 30 개의 단어



다른 채널에 대해서 데이터분석을 시행한 그림.5 와 6 의 경우도 TF 의 경우 채널 이름에 대한 언급이 대부분이지만 TF-IDF 그래프의 경우 영상별로 연관성이 있는 단어 “stopped”, “smoke” 가 추출되었으며 이를 이용해 gerund 에 대한 설명을 하고 있다는 점을 유추할 수 있다.

< 그림. 7> channel BBC Learning English 연관 규칙 분석

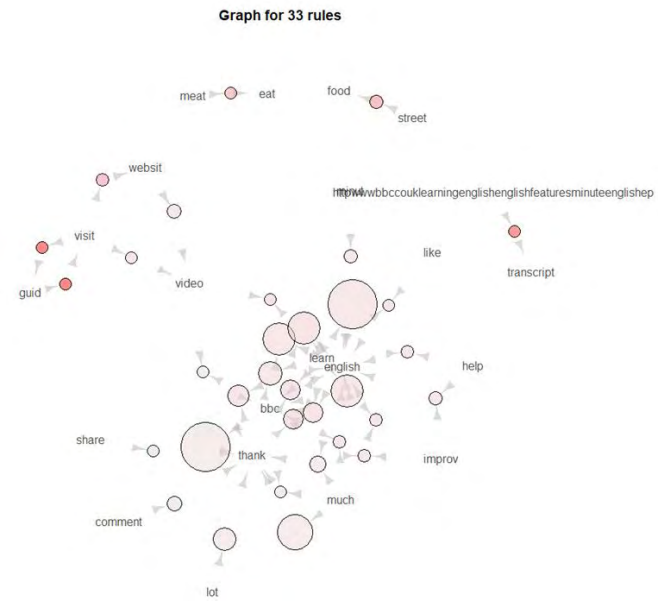


그림. 7 의 경우 channel BBC Learning English 에 대해 연관 분석 규칙을 시각화한 것이다. 그래프의 원의 크기는 Support, 색상 진하기가 Lift 를 의미하며 최소 support 는 0.01, 최소는 confidence 는 0.6 으로 설정한 그래프다.

일반적으로 TF 가 높은 단어가 규칙으로 생성되는 것을 볼 수 있다.

4. 사사어구

This work 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)" (2015-0-00908).

5. 결론

본론의 그림 1~6 을 살펴보면 TF 보다는 TF-IDF 에서 보다 다양한고 연관성이 있는 단어가 도출되는 것을 살펴보았다. 하지만 TF-IDF 만 이용할 경우 보다 단어간의 연관성을 휴리스틱하게 판단해야하는 한계가 존재한다.

단어 연관 규칙의 분석의 경우에는 일반적으로 단어들의 빈도수에 기인해 규칙을 생성한다. 따라서 해당 분석의 경우 관련도를 분석할 수 있지만 유튜브 환경내에서는 큰 의미 없이 반복되는 단어의 규칙이 시각화된다. 따라서 이 두 연구를 접목한 분석을 할 경우 인플루언서 선정에 보다 큰 도움이 될 것을 발견했기에 해당 연구의 초석을 제공할 수 있다고 판단된다.

참고문헌

[1] Hwang Seonguk “How Korean Top 100 companies Use Social Network Services: An Analysis of Relationship Cultivation Strategies, Message Topics, and Posting Types” 방송문화연구 제25권 제1호, 2013.6, 235-273

[2] Fril-jeaperson.C(2017) Celebrity endorser’s credibility: effect on consumers’ attitude toward advertisement: Factors influencing vloggers credibility among viewers and their relation with attitude toward advertisement, Luleå University of Technology, Department of Business Administration, Technology and Social Sciences.p.62

[3] Choe Jiyun, Jeong Yunjae “뷰티 인플루언서 마케팅 활용 전략: 매스미디어와 소셜미디어의 비교를 중심으로*” The Korean Journal of Advertising, Vol.28. No.4(2017). pp.47~72|ISSN 1225-0554

[4] Choi, Seung Woo· Park, Bo Ram* “A Study of Advertisement Avoidance by the Type of Digital Video In-stream Ad” Journal of the Korean Society of Design Culture Vol.21, No.3, 2015.9