오피니언 마이닝기반 방송-소비 영향 모델링

김진아*, 신윤미*, 문남미**
*호서대학교 컴퓨터공학과
**호서대학교 컴퓨터정보공학부
e-mail:jina9406@gmail.com

Opinion Mining based Broadcasting-Consumption Impact Modeling

Jinah Kim*, Yoonmi Shin*, Nammee Moon**
*Dept of Computer Engineering, Hoseo University
**Division of Computer and Information Engineering, Hoseo University

요 약

소비자의 행동 예측을 하는 데 있어 기존의 소비 행동과 더불어 외부 환경 요인 중 하나인 방송 미디어에 대한 영향 반영이 요구되며, 이 때, '스낵컬처' 시대에 알맞은 분석이 요구된다. 본 논문에서는 네이버 TV에서의 국내 방송 영상 콘텐츠를 활용하여 방송이 소비에 끼치는 영향에 대한 모델링을 진행하였다. 월별 선호도가 높은 방송들을 대상으로 텍스트 마이닝을 통해 방송 영상 콘텐츠의 제목, 내용, 태그, 댓글을 활용하여 주요 키워드를 추출하였으며, 이를 바탕으로 SO-PMI 기반의 오피니언 마이닝을 통해 소비 성향 키워드를 필터링하여 소비 감성 지수를 계산하였다. 이때, 소비 선호를 파악 가능한 소비 감성 사전을 새로 구축하여 활용하였다. 최종적으로, 소비자의 연령과 성별을 분류하여 방송 콘텐츠의 조회수 및 좋아요수를 반영한 방송 선호율과 소비 감성지수를 바탕으로 방송-소비 영향 모델링을 설계 및 구현하였다.

1. 서론

소비자가 소비 행동을 하는 데 있어 소속된 준거집단의 특성이나 거주 지역에 따른 문화 등 다양한 외부 환경 요인이 존재하지만, 그중에서 방송과 미디어로부터 오는 영향이 매우 크다. 때문에, 홈쇼핑이나 영화 혹은 TV 방송 프로그램의 PPL(Product PLacement) 노출에 대해서 효과 극대화를 위한 상품 배치나, 노출 시간 등 다양한 연구가 진행 중이다[1]. 그러나 이로 인한 소비 예측에 관한 연구는 부족한 실정이다. 최근, '스낵컬처'의 확산으로 방송 프로그램을 시청하는 방식이 TV를 찾기보다는 모바일 기기로 짧은 시간의 영상 콘텐츠를 즐기는 추세로 바뀌고 있다. 이러한 문화 현상을 반영할 수 있는 소비 예측에 관한 연구가 필요하다.

소비 예측을 하는 데 있어 가장 중요한 것은 선호도 파악이다. SNS(Social Network Service)나 온라인 쇼핑몰, 방송 혹은 영화 관련 페이지의 등급이나 리뷰를 활용하여 오피니언 마이닝을 통해 긍부정을 분석함으로써 선호도를 파악하는 연구가 많이 이뤄지고 있다[2-4].

본 논문에서는 국내에서 가장 점유율이 높은 네이버 TV에서 선호도가 높은 방송에 대한 영상 콘텐츠에 대한 데이터 크롤링을 통해, 소비 예측이 가능한 방송-소비 영향 모델링을 제안하고자 한다. 이를 위해 소비 감성 사전을 새로 구축함으로써 소비 선호도를 파악하고자 하였으며 이를 모델링에 반영하여 연령과 성별로 소비 영향 정도를 계산하고자 하였다.

2. 관련 연구

오피니언 마이닝은 긍정(Positive)와 부정(Negative)로 구분하여 선호도를 판별하는 기술로, 주로 SNS나 리뷰 등의 비정형 데이터인 텍스트를 분석함으로써 감정을 파악한다.

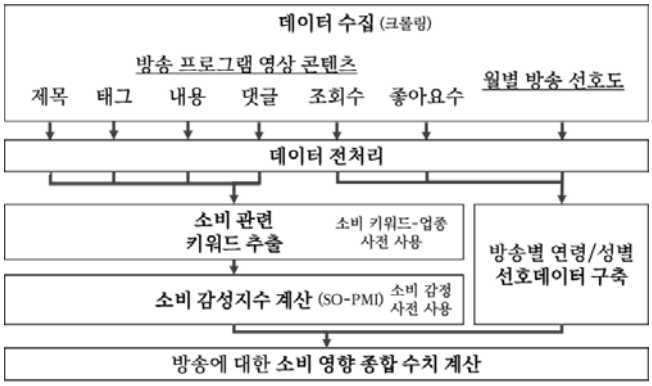
오피니언 마이닝은 PMI(Point-wise Mutual Information) 기반의 극성 분류 방식을 주로 활용한다. 이는 확률론에 기초한 방법으로 극성이 유사한 단어들이 같은 문서에 나올 확률이 높다고 가정한 것이며, 출현 빈도를 파악하기 때문에사용 및 계산이 단순할 뿐만 아니라 결과 예측도 정확하다. 그러나 단어 선택에 따라 결과가 크게 바뀌는 단점이 있다. 그리하여 이를 개선한 SO-PMI(Semantic Orientation from PMI)가 있는데, 다수의 기준 언어를 설정해 보완한 방식이다. 최근에는 SO-PMI 기반의 감성분석에 대한 연구가 주로 이루어지고 있으며, 평점과 리뷰를 동시에 사용하거나 SVM(Support Vector Machine)이나 나이브베이즈(Naive Bayes)등의 기계 학습 법과 결합하는 등 다양한 오피니언 마이닝 연구가 이뤄지고 있다[5-7].

본 연구에서는 방송 프로그램의 영상 콘텐츠의 댓글을 활용하여 극성을 분류하고자 하였으며, 댓글의 길이가 짧아 출현 빈도를 기반으로 하는 SO-PMI 방식을 통해 소비 선호도를 파악하고자 하였다. 그리고 이때 이 소비 선호도를 방송-소비 영향 모델링에 반영하고자 하였다.

3. 방송-소비 영향 모델링 개요

본 연구에서 제안하는 방송-소비 영향 모델링의 목적은 매월 선호도가 높은 국내 TV 방송 프로그램을 10개 선정하고, 이 프로그램이 소비에 끼치는 영향을 성별(남/여)과 연령별(20대/30대/40대/50대/60대)로 구분하여 방송 선호도에 따라 소비 영향 정도를 파악해 향후 예측하는 데 있다.

전체 프로세스는 다음 (그림 1)과 같다. 먼저, 영상 콘텐츠에 대한 데이터 크롤링을 통해 데이터를 수집한 후 전처리 과정을 거친다. 그리고 소비와 관련된 키워드를 추출하는 작업을 진행하는데 이때, 이 키워드를 중심으로 오피니언 마이닝을 진행하여 키워드별 소비 감성 지수를 얻는다. 최종적으로, 방송 프로그램의 소비 감성 지수와 방송 프로그램의 선호도를 종합하여 소비 영향 정도를 계산한다.



(그림 1) 모델링 프로세스

4. 방송-소비 영향 모델링 과정

3-1. 데이터 수집 및 전처리

(그림 2) 수집된 영상 콘텐츠에 대한 데이터

본 논문에서의 데이터 수집은 (그림 2)와 같이 '네이버 TV'의 HTML 형태를 파악해 파이썬(Python)기반의 크롤러를 통해 선호도 높은 방송 프로그램에 대한 영상 콘텐츠의 제목, 태그, 댓글을 크롤링하였다. 이때, 선호도 높은 프로그램의 기준은 매월 약 1,000명을 대상으로 전화 조사 인터뷰를 진행하는 한국 갤럽 리포트 '요즘 가장 좋아하는 TV 프로그램은?'의 결과 데이터를 활용하였다.

데이터 전처리를 위해서 오픈소스 R의 tm 패키지를 활용하였다. 문장부호나 숫자 등의 무의미한 데이터들이나 띄어쓰기가 없이 너무 긴 단어들은 정보 파악에 어려움이 있으므로 제거하였다.

3-2. 소비 감성 지수 계산

소비 감성 지수 계산을 위하여 소비 관련 키워드 추출이 진행되는데, 이때 3차례의 키워드 추출 과정을 거친다. 1차로, 영상 콘텐츠와 관련한 모든 내용에 대해 한글 형태소 분석을 통하여 보통명사와 고유명사에 해당하는 단어를 추출한다. 2차는 추출된 단어들을 대상으로 소비 관련된 키워드를 추출하는 과정으로, <표 1>과 같이 미리 구축된 소비 키워드-업종 사전을 활용한다. 이는 추후 소비 키워드에 따라 영향받는 업종을 파악하기 위함이다. 이때, 영상과 관련 높은 키워드를 추출하기 위하여 영상 콘텐츠의 제목, 태그, 내용에 대해 키워드를 추출한 것을 우선시하며, 댓글에 대해서는 관련이 없는 단어가 추출될 수 있으므로 빈도수가 높은 것만을 택한다.

<표 1> 소비 키워드-업종 사전 예시

키워드	업종	키워드	업종
삼겹살	한식	컴퓨터	전자기기
곱창	한식	블라우스	의류
한우	한식	원피스	의류
초밥	일식	팩트	화장품
짜장면	중식	립스틱	화장품

소비 키워드 추출의 마지막 3차 과정은 소비 키워드 중에서 소비 성향이 높은 키워드를 추출하는 과정이다. 소비 성향을 판단하기 위하여 SO-PMI 기반의 오피니언 마이닝을 활용해 해당 키워드에 대한 긍부정 인식을 판별한다. 이는 방송 콘텐츠의 소비 키워드가 포함된 댓글만을 필터링하여 진행되며 소비 감정 점수를 구한다. 소비 감정 점수는 긍정일 경우는 +1을, 부정일 경우에는 -1을 부여하여 키워드별로 점수를 종합하여 계산한다. 이때 극성 분류를 하기 위해 미리 구축된 소비 감성 사전이 기준이 된다. 이는 기존의 감성 사전과 달리 영상 콘텐츠 중심의 소비 선호를 파악할 수 있는 감성 사전이며, <표 2>와 같이 소비와 관련된 긍부정의 형용사 및 동사로 구성된다.

<표 2> 소비 감정 사전 예시

긍정 단어		부정 단어	
맛있겠다	먹고싶다	별로다	맛없겠다
가보고싶다	기대된다	가기싫다	기대이하다
해보고싶다	사고싶다	하기싫다	사기싫다
예쁘다	갖고싶다	안예쁘다	갖기싫다
맘에든다	궁금하다	맘에안든다	기대안된다

최종적으로, 소비 관련 키워드는 방송 이름, 회차(날짜 포함), 소비 성향 키워드, 소비감정 점수로 구성된다. 각 방송 회차별로 소비 감정 점수를 합산하여 최종적인 소비 감성 지수(*Consumption Emotion, CE*)를 구한다.

3-3. 소비 선호도 계산 및 방송-소비 영향 모델링

소비 선호도 계산은 한국 갤럽 리포트의 선호 프로그램 분석 결과 데이터를 활용해 월별 선호도가 높은 방송 프로그램을 선정하였으며, 방송별 선호율을 계산하였다. 또한, 조회수는 영상 콘텐츠의 관심도를 반영하며 노출 횟수를 나타내는반면, 좋아요수는 호감도를 반영하여 선호도를 확인하는데 직접적인 요소가 된다. 그리하여 다음 식(1)과 같이 방송 프로그램 p에 대한 w주차의 선호율(PP)을 바탕으로 좋아요수($Like$)를 조회수($View$)로 나눈 값을 가중치로 부여함으로써 방송별 소비 선호도(*Consumption Preference, CP*)를 계산하였다. 이때, 선호율(PP)은 월별 데이터이므로 각 주차별로 선호도를 같게 부여한다.

$$CP_{(p,w)} = \left(\frac{\sum_{w=1}^{n} Like_{(p,w)}}{\sum_{w=1}^{n} View_{(p,w)}} \right) \times PP_{(p,w)} \qquad 식(1)$$

방송별 소비 선호도와 '3-2.소비 감성 지수 계산'에서 구한 방송별 소비 감성 지수를 활용하여 최종적으로 방송-소비 영향 모델링을 진행한다. 이때, 각 방송의 회차마다 소비에 영향을 끼치는 기간은 방송한 주를 포함한 2주로 설정하였다. 이는 방송된 후 시간이 많이 소요될수록 관심이 감소하는 것과 일반적으로 재방송이 일주일 후에 진행되므로 영향이 확장되는 것을 반영한 것이다.

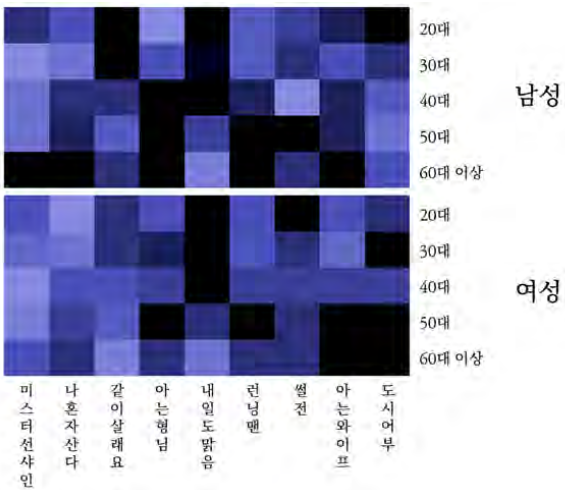
방송 프로그램 p에 대한 w주차의 소비 선호도(CP)와 소비 감성지수(CE)를 반영한 소비 영향 수치(*Consumption, Figure, CF*)는 식(2)와 같이 구한다. 이 소비 영향 수치는 1주일에 1-2번 방영하기 때문에 1주일 단위로 계산된다. 그리하여 식(3)과 같이 이전 주에 방영한 프로그램의 소비 영향 수치와 이번 주에 방영한 프로그램의 소비 영향 수치를 합산하여 최종 소비 영향 종합 수치(*Final Consumption Figure, FCF*)를 구하였다. 이는 <표 3>과 같이 연령 및 성별로 분류되어 계산되며 (그림 4)와 같이 HeatMap으로 표현될 수 있다.

$$CF_{(p,w)} = CP_{(p,w)} \times CE_{(p,w)} \qquad 식(2)$$

$$FCF_{(p,w)} = CF_{(p,w-1)} + CF_{(p,w)} \qquad 식(3)$$

<표 3> 최종 소비 영향 종합 수치 예시

성별	연령	소비영향 종합수치
남	20대	19.465
	30대	11.243
	40대	9.651
	50대	3.423
	60대	1.215
여	20대	22.549
	30대	18.345
	40대	5.845
	50대	2.421
	60대	1.211



(그림 3) 방송별 소비영향 종합 수치 HeatMap

5. 결론 및 기대효과

본 논문에서는 방송 프로그램의 짧은 영상 콘텐츠를 활용하여 연령 및 성별 선호도를 반영한 방송-소비 영향 모델링을 제안하였다. 소비 영향 정도를 파악하기 위하여 SO-PMI 기반의 오피니언 마이닝을 통해 영상 콘텐츠의 소비성향 키워드를 추출하여 소비 감성 지수를 계산하였다. 이때, 소비 선호를 파악할 수 있는 소비 감성 사전을 새로이 구축하여 활용하였다. 최종적으로, 방송에 대한 소비 선호도와 소비 감성 지수를 가지고 소비 영향 모델링을 진행하였다. 본 논문에서 제안하는 방법이 '스낵컬처'를 반영한 소비 예측이라는 점에서 의의가 있으며, 단순한 방송별 소비 예측을 넘어 PPL 효과 예측 등으로 발전될 수 있을 뿐만 아니라 방송 이외의 영화 등의 다양한 영상 콘텐츠들로 확장되어 적용 가능할 것으로 기대한다.

ACKNOWLEDGEMENT

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1A2B4008886).

참고문헌

[1] 오창우 "PPL 효과 측정을 위한 방법론적 탐색 MCQ 척도에 의한 제품의 재인, 회상, 선택의도의 측정", 한국광고홍보학보 16(2), pp.261-302, 2014.04

[2] Hu, Ya-Han, Yen-Liang Chen, and Hui-Ling Chou "Opinion mining from online hotel reviews - A text summarization approach", Information Processing & Management 53(2), pp.436-449, 2017

[3] Unnisa, Muqtar, Ayesha Ameen, and Syed Raziuddin "Opinion mining on Twitter data using unsupervised learning technique", International Journal of Computer Applications 148(12), pp.12-19, 2016

[4] 김유영, 송민 "영화 리뷰 감성 분석을 위한 텍스트 마

이닝 기반 감성 분류기 구축", 지능정보연구 22(3), pp.71-89, 2016.09
[5] Sun, Xiao, et al "Fine-grained emotion analysis based on mixed model for product review", Int. J. Netw. Distrib. Comput 5(1), pp.1-11, 2017
[6] 안현우, 문남미 "오피니언 마이닝과 머신러닝을 이용한 페이스북 인기 게시물 예측 시스템", 한국방송미디어공학회 학술발표대회 논문집, pp.70-73. 2017.11
[7] 이상훈, 최정, 김종우 "영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석", 지능정보연구 22(2), pp.97-113, 2016.06