SNS를 활용한 영화 흥행 분석 및 예측

장은지* 현민지* 정예지* 전유원** 강희석***
*서울시립대학교 컴퓨터과학부
**서울시립대학교 통계학과
***아시아나 IDT
smeil123@naver.com, mgus09@naver.com, yejijung97@gmail.com, yuwon0325@naver.com, kanghs1@asianaidt.com

Analysis and Prediction of Movie Box-Office using SNS

Eunji Jang*, Minji Hyun*, Yeji Jung*, Yuwon Jeon**, Huiseok Kang***
*Department of Computer Science, University of Seoul
**Department of Statistics, University of Seoul
***ASIANA IDT

요 약

다양한 SNS에서 관객 반응을 수집하고, 그 중 비정형 데이터인 관람후기는 긍부정 반응을 분석해 수치화한다. 전처리한 데이터를 사용해 다중선형회귀모델과 인공 신경망 모델을 통해 분석하여 실제 영화 관객수와의 관계를 분석한다. 또한 이를 기반으로 통계분석을 통해 흥행 예측모델을 구성한다.

1. 서 론

최근 HW의 고사양화와 인공지능, 빅데이터 등 컴퓨팅기술의 발전, 그리고 활발한 소셜 네트워킹(Social Networking)을 기반으로 예측하기 어려웠던 소비자 니즈에 대한 분석이 활발하다. 특히 이러한 분석은 잠재적 니즈를 정확히 파악함으로서 보다 효과적인 타켓마케팅이 가능하게 하였다. 이러한 트렌드는 문화산업에서도 활발한데 특히 영화산업에서 특정 영화의 흥행 결정요인에 대한 연구와 함께 상업적 측면에서 흥행예측에 대한 관심이 증대되고 있다. 영화산업에서 관객의 수요를 예측하는 것은 매우 중요한 문제이다. 최근 다양한 SNS가 활성화되면서 관객의 관람후기는 해당 영화의 흥행에 결정적 요소가 되고 있다. 본 논문에서는 SNS에서 수집할 수 있는 소비자 반응에 집중해서 영화의 흥행을 예측한다. 즉, SNS를 중심으로 관객의 영화에 대한 관심과 긍부정 반응을 판별해 영화의 실제흥행과 어떤 관계가 있는지 분석하고 예측한다. 본 논문의 구성은 다음과 같다. 2장에서는 트위터와 페이스북의 영화 관람 후기데이터, 네이버와 다음의 관람 후기데이터를 Python 및 R을 활용해 개봉 전, 후 데이터를 수집하는 과정을 설명한다. 또한, 비정형 데이터인 게시물의 긍부정 반응을 판별하고 정제하는 과정을 설명한다. 3장에서는 Python과 R을 기반으로 통계분석을 실시해 영화 흥행에 영향을 주는 변수를 선정하고, 이를 토대로 다중선형회귀 분석 모델과 인공신경망 분석 모델에 적용하는 과정을 설명한다. 4장에서는 본 논문의 결과와 향후 과제를 제시한다.

2. 데이터 수집 및 전처리

2.1 SNS의 영화 리뷰 데이터 수집

일반적으로 SNS에서 데이터를 가져오는 방법은 API를 이용하는 방법과 웹 크롤링(Web Crawling)을 하는 방법이 있다. 본 연구에서는 특정 영화에 대한 언급 데이터만을 얻어오기 위해 웹 크롤링 기법을 활용하였다. 웹 크롤링이란 웹 크롤러(web crawler)가 하는 작업을 일컫는 말로, 웹크롤러는 자동화된 방법으로 여러 웹 사이트에 있는 특정 부분의 정보를 수집하는 프로그램이다. 크롤링은 python에서 Selenium과 Beautifulsoup 프레임워크를 사용하였다.

트위터의 경우 고급 검색 기능을 사용하여 영화 제목과 부제를 OR 기능을 사용하고, 기간을 설정한 페이지를 URL을 통해서 접근한다. 해당 페이지의 HTML 구조를 파싱하여 트위터 게시글 내용, 답글 수, 리트윗 수, 좋아요 수를 수집하였다.

페이스북은 먼저 로그인을 한 뒤, 영화명을 추가한 URL로 접근한다. 이 후 총 게시글 수, 게시날짜, 게시글 내용, 좋아요 수, 공유 수, 조회 수, 댓글 수의 데이터를 파싱한 후 전처리하여 저장한다. 게시날짜와 좋아요 수 등의 형식이 정형화되어 있지 않기 때문에 전처리 과정은 반드시 필요하다.

SNS데이터의 경우 ,영화명으로만 검색해서 가져온 데이터이기 때문에 흔한 이름의 영화라면 영화와 관련 없는 게시글도 검색이 된다. 그렇기 때문에 오차를 줄이기 위해서 저장한 데이터의 게시 날짜를 이용해 영화개봉일 전 후 일주일만 사용했다. 이렇게 하면 해당 영화와 관련성이 높은 글만 수집이 가능하다.

2.2 포털 사이트의 영화 리뷰 데이터 수집

크롤링 대상 포털사이트로는 점유율이 높은 국내 포털사이트

중 다음, 네이버를 선정하였다. 크롤링 도구로는 R의 RCurl, rvest 등의 패키지를 활용하였고 주요 수집 데이터는 영화의 평점, 리뷰 등의 데이터를 저장하였다.

2.3 데이터 전처리

수집한 텍스트의 긍부정 판별에 앞서, 한글을 제외한 모든 문자를 제거해 데이터를 정제한다. 그 후에 긍부정 판별 단계

는 텍스트를 형태소 단위로 쪼개는 형태소 분석 과정과 형태소의 긍부정을 판단하는 긍부정 분석 과정으로 나누어진다. 형태소 분석 과정에서는 자연어 처리 파이썬 패키지인 koNLPy를 사용하여 텍스트를 이루고 있는 형태소의 리스트를 얻어냈다. 긍부정 사전으로는 서울대 컴퓨터 언어학과에서 만든 KOSAC 기반의 Korean Sentiment Lexicon을 사용하여 문장의 긍부정 반응을 분석하고, 추가로 텍스트 출현빈도 수가 100 이상이지만 사전에는 없는 형태소일 경우 긍부정 사전에 직접 추가하여 보다 더 정확한 분석이 이루어질 수 있도록 보완하였다. 긍부정 판별 후 긍정비율을 변수로 생성해 이를 예측 모델링에 활용하였다.

$$\text{긍정비율} = \frac{\text{긍정형태소개수}}{\text{긍정형태소개수} + \text{부정형태소개수}}$$

(그림 1) 긍정비율 변수 생성에 대한 산식

데이터 수집과 전처리 과정을 거친 후 만들어진 변수들은 (표 1)과 같다. 각 변수들은 개봉 전, 후로 나눠서 사용한다.

(표 1) 전처리 후 데이터의 전체 변수

변수 이름	의미
DAUM_POS_RATIO	다음 긍정 비율
DAUM_RATE	다음 별점
FACEBOOK_CNT	페이스북 게시글 수
FACEBOOK_COMMENTS	페이스북 댓글 수
FACEBOOK_LIKES	페이스북 좋아요 수
FACEBOOK_POS_RATIO	페이스북 긍정 비율
FACEBOOK_SHARES	페이스북 공유 수
FACEBOOK_SHOWS	페이스북 조회수
NAVER_POS_RATIO	네이버 긍정 비율
NAVER_RATE	네이버 별점
TWITTER_CNT	트위터 게시글 수
TWITTER_COMMENTS	트위터 답글 수
TWITTER_LIKES	트위터 좋아요 수
TWITTER_POS_RATIO	트위터 긍정 비율
TWITTER_SHARES	트위터 공유 수

3. 특성 변수 선택 및 예측모델링

3.1 특성 변수 선택

수집된 데이터를 기반으로 예측모델을 만들기에 앞서 변수선택을 진행하였다. 종속변수는 영화의 흥행을 판단하기에 가장 대중적인 척도라고 판단한 영화별 관객 수로 선택하였고 독립변수 선택은 변수별 상관관계 비교, 단계적 회귀 방법을 사용하여 결정하였다. 단계적 회귀 분석은 전진 선택법과 후진 제거법을 번갈아 진행하여 독립변수들과 종속변수와의 관계에서 영향력이 큰 독립변수들을 밝혀내는 방법으로서 변수들의 추가, 제거가 반복되면서 AIC값이 가장 작은 독립변수들의 부분집합이 선택된다. 분석 결과 AIC값이 2939.82에서 2857.2가 되어 10개의 변수가 선택되었다. 이 회귀식의 경우 p-value값이 1.12e-13으로 0.05보다 작기 때문에 통계적으로 의미가 있다고 볼 수 있으며 Adjusted R-squared값이 0.6014로 종속변수의 60.14%를 설명할 수 있는 것으로 나타났다.

단계적 회귀와 영향력이 높다고 나타난 변수들 중 상관관계가 높은 변수 일부를 제거 후 최종 모델에 적용할 변수를 선택하였으며 최종 선택된 변수는 (표 2)와 같다.

(표 2) 최종 선택된 변수 목록

FACEBOOK_LIKES_AFTER
NAVER_POS_RATIO_AFTER
TWITTER_CNT_AFTER
FACEBOOK_SHARES_AFTER
FACEBOOK_COMMENTS_AFTER
TWITTER_LIKES_BEFORE
DAUM_POS_RATIO_BEFORE
NAVER_RATE_AFTER
TWITTER_COMMENTS_BEFORE
DAUM_RATE_AFTER

3.2 예측 모델링

예측모델링을 위한 알고리즘으로서 다중선형회귀모델과 인공신경망 모델을 선택하였다. 선형회귀 모델을 세워 모델과 변수가 통계적으로 유의미한 결과를 갖는지 확인하였다. 인공 신경망은 입력을 반복하며 가중치(함수의 parameter)를 갱신해나간다. 최종적으로 원하는 결과를 도출해내는 함수의 parameter를 유추해낸다.

3.2.1 다중선형회귀 모델

다중선형회귀 모델에는 앞서 선택된 10개의 변수를 적용하여 회귀식을 만들었다. train 데이터와 test 데이터는 9:1의 비율로 만들었으며, train 데이터를 적용한 결과 p-value값이 1.174e-11로 0.05보다 작게 나와 통계적으로 의미가 있다고 할 수 있다.

$$H(X) = WX^T + b$$

$$W = (w_1 \quad \dots \quad w_n),\ X = (x_1 \quad \dots \quad x_n)$$

(그림 3) 다중선형회귀 공식

Adjusted R-squared값은 0.5341로 train 데이터 중 53.41%를 설명하고 있음을 보였다. test 데이터를 적용한 결과 예측도는 92.54%로 높게 나타났다. 사용된 독립변수들을 각각 살펴보면 변수들 중 5개의 변수만이 p-value값이 0.05보다 작게 나와 유의미한 것으로 판별할 수 있었다.

$W = (\,23.66\ \ 15450000\ \ 2574\ \ 182.2\ \ -22.09\ \ -19.8\ \ 3629000\ \ -199800\ \ 684.2\ \ -176900\,)$
$b = -8131000$
x_1: FACEBOOK_LIKES_AFTER, x_2: NAVER_POS_RATIO_AFTER
x_3: TWITTER_CNT_AFTER, x_4: FACEBOOK_SHARE_AFTER
x_5: FACEBOOK_COMMENTS_AFTER, x_6: TWITTER_LIKES_BEFORE
x_7: DAUM_POS_RATIO_BEFORE, x_8: NAVER_RATE_AFTER
x_9: TWITTER_COMMENTS_BEFORE, x_{10}: DAUM_RATE_AFTER

(그림 4) 변수 10개 다중선형회귀 식 대입 변수

이를 바탕으로 회귀식을 다시 만들어보았다.

$W = (\,15.73\ \ 13750000\ \ 2148\ \ 166.4\ \ 3675000\,)$
$b = -9895000$
x_1: FACEBOOK_LIKES_AFTER, x_2: NAVER_POS_RATIO_AFTER
x_3: TWITTER_CNT_AFTER, x_4: FACEBOOK_SHARE_AFTER, x_5: DAUM_POS_RATIO_BEFORE

(그림 4) 변수 5개 다중선형회귀 식 대입 변수

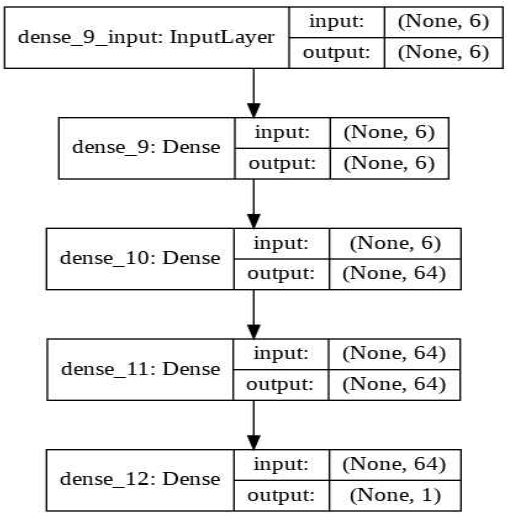
위 5개의 변수만 사용한 회귀식은 각 변수 모두 유의미했으며 전체 모델 또한 p-value값이 3.196e-13으로 통계적으로 유의하다고 할 수 있었다. Adjusted R-squared값은 0.5201로 앞선 회귀식보다는 0.14정도 감소하였으나 test 데이터의 예측도는 92.87%로 소폭 상승하였다. (예측도는 소수 셋째 자리에서 반올림) 또한, 변수를 모두 선택했을 때보다, 변수선택시에 예측도가 높은 것으로 확인되었다. 결론적으로 다중선형회귀 모델의 예측도는 비교적 높은 편이고, Adjusted R-squred는 50%이상으로 예측에 대해 신뢰할 수 있는 수준으로 판단되었다.

(표 3) 다중선형회귀 모델 결과 비교

구분	변수 30개	변수 10개	변수 5개
p-value	4.493e-09	5.968e-12	3.196e-13
Adjusted R-squred	0.6174	0.5426	0.5201
정확도	62.75%	92.54 %	92.87 %

3.2.2 인공 신경망 모델

인공신경망 모델에서는 입력차원에 비해 학습 데이터의 수가 적으므로 일련의 전처리 과정이 필요하다. 특성변수 선택과정에서 최종 선정된 변수(표 2) 중 영향력이 높은 상위 5개의 변수만을 모델링에 사용했고, 적은 양의 학습 데이터를 보완하기 위해 Data Augmentation 과정을 거쳐 1000개의 데이터로 확장시켰다. 연구에 사용한 신경망 네트워크는 6개의 입력(특성 변수 5개와 상수항)을 받는 input layer, layer 당 6, 64, 64개의 perceptron으로 구성된 hidden layer 3개, 예측값을 출력하는 output layer로 구성되었다. (optimizer : rmsprop, cost function : Mean Squared Error, activation function : ReLU)



(그림 5) 인공 신경망 모델

990개의 데이터를 사용하여 batch 크기 64, epoch 2000으로 학습하여 예측 모델을 만들었고, 10개의 테스트 데이터를 수치 예측에 사용하였다. 예측 결과는 다음과 같다. (예측 관객 수와 상대오차는 각각 소수 첫째 자리, 둘째 자리에서 반올림)

(표 4) 관객 수 예측 결과와 상대오차

예측 관객 수(명)	실제 관객 수(명)	상대오차
3,248,625	1,335,193	143.3 %
310,006	388,301	20.2 %
264,977	240,148	10.3 %
5,415,578	11,211,627	51.7 %
3,248,625	1,335,193	143.3 %
123,787	111,541	11.0 %
2,541,335	2,614,601	2.8 %
3,294,026	3,637,122	9.4 %
3,048,389	3,279,296	7.0 %
736,734	903,195	18.4 %
평균 상대오차	41.6 %	

4. 결 론

본 연구에서는 2017년~2018년 개봉한 영화를 대상으로, SNS, 포털사이트의 반응을 통해 영화의 흥행을 예측해 보았다. 단계적 회귀 분석으로 30개 변수 중 흥행 예측에 유의미한 변수를 선정하였고, 다중선형회귀 모델과 인공신경망 모델로 예측을 수행하였다. 다중선형회귀 모델은 쉽게 표현할 수 있기 때문에 많이 사용되며 결론적으로 관객 수 예측에서는 변수를 모두 사용했을 때보다 특정 변수로 한정하는 것이 예측정확도를 높이는 방법이라고 판단할 수 있었다. 변수를 한정하여 차원을 줄였을 때 예측에 대한 신뢰도는 어느정도 긍정적인 결과를 도출해 낼 수 있었다. 반면 인공 신경망 모델의 경우는 예측값의 평균 상대오차가 크고 시행마다 값의 편차가 커서 신뢰성이 높지 않았다. 이는 부족한 학습 데이터의 양이 원인인 것으로 판단되고, 향후에는 더 많은 데이터 수집, Gaussian Noise를 이용한 Data Augmentation 등으로 좋은 데이터 셋을 구축하여 예측 모델을 보완할 필요가 있다고 판단된다. 또한, 시간의 흐름에 따른 SNS의 반응을 고려할 수 있는 시계열 분석을 더한다면 SNS 반응의 변화 흐름에 따른 관객 수의 추이를 보다 더 정확하게 예측할 수 있을 것으로 기대된다.

5. 참고문헌

[1] Sitaram Asur, Bernardo A. Huberman, *Predicting the Future with Sociail Media*(2010)

[2] L. Doshi, K.Krauss, S.Nan, and P.Gloor, *Predicting moive prices through dynamic social network analysis*(2010)

[3] 강지훈, 박찬희, 도형록, 김성범, *데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법*(2014)

[4] 허민회, 강필성, 조성준, *Predicting Box-Office with Opinion mining reviews*(2013)

[본 논문은 과학기술정보통신부 정보통신창의인재양성사업의
지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.]