영화 흥행 예측을 위한 영화 관객 수와 관련 트윗간의 상관관계 분석

임준엽, 황병연
가톨릭대학교 컴퓨터공학과
e-mail:junyeob1205@catholic.ac.kr

An Analysis of Corelation between Movie Attendance and Related Tweets for Predicting Box Office

Junyeob Yim, Byung-Yeon Hwang
Dept. of Computer Science and Engineering, The Catholic University of Korea

요 약

최근 들어 영화에 대한 수요가 증가하면서 국내 영화시장규모는 지속적으로 성장하고 있다. 이와 관련하여 여러 가지 위험요소를 제거하고 시장에서의 성공을 위해 영화의 흥행을 예측하기 위한 다양한 연구들이 진행되고 있다. 그러나 그러한 예측을 위한 관련 요소들 간의 상관관계를 정확한 수치로 표현하는 일은 매우 어려우며 관련연구 또한 아직 미흡하다. 본 논문에서는 트위터에서 발생되는 트윗을 설문 표본으로 삼고 영화 관련 트윗과 영화의 흥행을 의미하는 관객 수와의 상관관계를 분석하여 상관계수를 도출하였다. 실험 결과 실험에 사용된 영화 10편의 관객 수에 대한 데이터 모두 관련 트윗의 발생비율과 양의 상관관계를 가짐을 알 수 있었으며 이를 통해 트위터를 이용한 영화의 흥행 여부 예측에 대한 가능성을 제시했다.

1. 서론

영화는 사람들의 생활수준 향상과 이에 따른 문화생활 욕구의 증대로 인해 점점 그 수요가 증가되고 있다. 2013년 영화진흥위원회가 전국 영화관의 발권 데이터를 집계한 자료에 따르면 2012년 국내 영화 시장의 총 매출액은 8,316억 원, 총 관객 수는 11,461만 명인 것으로 나타났다. 이는 통계가 시작된 2004년 이후 2,391억 원, 3,774만 명에서 해마다 꾸준히 증가해온 것으로 국내 영화시장규모의 성장을 의미한다[1]. 그러나 배급사의 입장에서 보면, 영화는 하나의 문화 상품으로서 시장에서의 상품 가치를 판단하는 기준이 모호하고 상품의 생명주기가 짧아 관객의 수요를 예측하기가 어렵다. 이러한 문제를 해결하기 위해 다양한 연구들이 진행되어 오고 있으나 정작 상품 수요의 주체인 관객에 대한 연구는 상대적으로 부족한 실정이다.

이를 위해 사람들의 현재 생각이 실시간으로 반영이 되며 그에 관한 데이터 수집이 자유로운 트위터(Twitter)를 이용할 수 있다. 트위터는 140글자로 제한된 트윗(Tweet)이라는 단문 텍스트를 지원하는 SNS(Social Network Service)로서 그 구조적 특성상 개방적인 네트워크를 가지고 있다. 또한 다른 SNS에 비해 사용자들이 생산하는 정보의 주관성이 뚜렷하며 빠른 정보의 확산성을 지닌다. 이러한 트위터를 분석하면 특정 상품에 대한 수요자들의 생각을 추출하는 것이 가능하기 때문에 유용한 설문 표본이 될 수 있다.

본 논문은 영화의 흥행을 예측하기 위해 이를 결정짓는 요소 중 하나인 관객 수와 트위터 내에서 발생되는 관련 트윗간의 상관관계를 밝힌다. 본 논문의 구성은 다음과 같다. 2장에서 영화의 흥행에 영향을 미치는 SNS의 구전효과와 영화의 흥행을 예측하는 선행 연구를 살펴보고, 3장에서 데이터 수집방법과 분석방법을 설명한다. 그리고 4장의 실험 결과를 통해 5장에서 본 연구의 결론과 향후 연구 계획에 대해 소개한다.

2. 관련연구

최근 영화의 흥행에 관해 다양한 연구들이 진행되어오고 있다. 특히 박선영의 연구[2]에서는 영화의 흥행에 영향을 미치는 요소로서 SNS를 통한 구전효과를 주목했다. 이 연구에서는 영화 <써니>의 사례를 통해 영화의 제작 및 상영 과정에서의 SNS 활동들을 분석하였다. 이를 위해 개봉 전, 개봉 초기, 성숙기로 시간적 구간을 나누었으며 각 구간 동안 SNS의 기여도를 분석하였다. 분석 결과 사례로 제시한 영화의 성공요인을 SNS 활동으로 지목하였다. 또한 영화 <시라노 : 연애조작단>과 영화 <레지던트 이블 4 : 끝나지 않는 전쟁 3D>에서의 관련 트윗 개수와 흥행 결과로 영화의 흥행과 SNS가 밀접한 관계가 있음을 강조하였다.

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0009407).

영화의 흥행 예측을 시도한 연구도 있다. 이경재 등의 연구에서는 영화의 성공여부를 결정하는 요인으로서 영화 속성 관련 요인과 영화 외적 요인을 분류하여 설명하였다[3]. 또한 추정 모수의 효과에 대한 잘못된 가정으로 인한 기존 연구들의 한계점을 지적하면서, 새로운 베이지안 선택모형을 제안하였다. 이를 각각의 변수를 두고 인공 신경망 모형과 비교하였으며 2005년과 2006년 상반기에 상영된 영화들을 토대로 베이지안 선택모형이 더 우수하다는 결과를 보였다.

위의 두 연구들은 영화 흥행에 관해 미치는 여러 요소들을 분석하였다. 그러나 영화를 보는 관람객들의 생각을 수치적으로 계량화하고 서로간의 상관관계를 찾는 것에서는 한계를 보인다.

3. 실험 방법

3.1 데이터 수집

본 논문에서는 영화를 관람하는 관객들의 영화에 대한 관심을 가시적으로 수치화해 보고자 전체 트윗 중 영화 관련 트윗의 비율을 분석하였다.

우선 트윗을 수집하기위해 트위터에서 제공하는 Twitter Streaming API[4]를 이용해 2013년 04월 01일부터 2013년 09월 30일까지 6개월 동안 트윗을 수집하였다. 트위터의 경우 트위터와 별도의 파이어호스(Firehose) 계약이 없다면 트위터 내부에 접근하여 수집할 수 있는 트윗의 비율은 전체 발생되는 트윗의 1%에 불과하다. 하지만 본 연구에서 도출하고자하는 목표는 트윗의 수집량 보다 수집된 트윗에 포함된 영화 관련 트윗의 비율과 관계되어 있기 때문에 지속적으로 트윗을 수집할 수 있다면 수집량과 관계없이 분석이 가능하다.

수집된 트윗은 이범석의 연구[5]에서 밝힌 데로 SNS를 사용자들의 시간대별 사용패턴을 고려하여 하루 단위로 클러스터링 하였다. 이후 수집된 트윗을 영화 제목의 포함여부로 필터링(Filtering)하여 트윗에 대한 데이터 셋(Data Set)을 완성하였다.

실제 영화를 관람한 관객 수에 대한 데이터 수집은 영화진흥위원회에서 제공하는 박스오피스 통계자료를 이용하였다. 트윗 수집기간 동안 개봉한 영화를 임의로 10편을 선택하여 트위터와 마찬가지로 하루 단위로 관객 수에 대한 데이터를 분류하였다.

3.2 상관관계 분석

수집된 트윗과 영화 관객 수의 상관관계를 분석하기 위해 (식 1)과 같이 상관계수 r을 구하는 표준화된 공식을 사용하였다. 이를 이용하면 같은 시간대의 영화 관련 트윗 비율과 영화 관람객 수 사이의 상관정도를 수치화해서 표현할 수 있다. 수치가 1에 가까울수록 강한 양의 상관관계, -1에 가까울수록 강한 음의 상관관계를 가진다.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{(식 1)}$$

단, (식 1)을 사용하기 위해서는 비교할 대상들의 데이터양이 일치해야 한다. 따라서 이를 위해 통계에 반영되지 않은 상영기간 외의 관람객 수는 각 구간마다 0명으로 정규화 하였다.

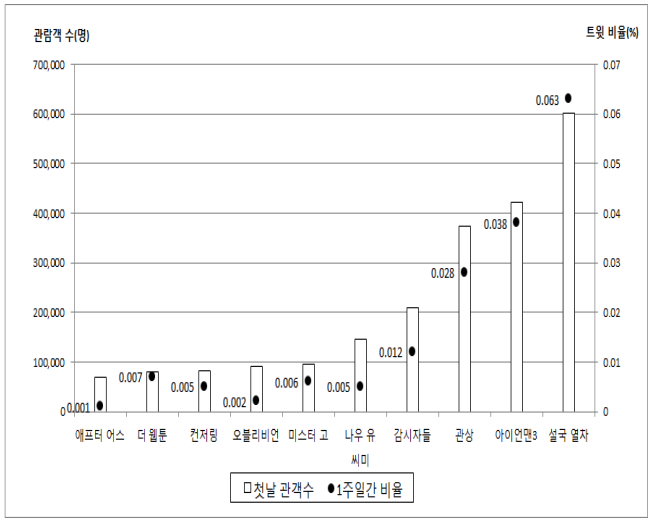
4. 실험 결과

트위터는 영화의 흥행에는 영향을 줄 수 있으나 직접적인 흥행의 결정요소로 보기는 어렵다. 그러나 흥행을 예측함에 있어 참고 요인이 될 수는 있다. <표 1>은 6개월 동안 수집된 트윗과 같은 기간 동안 영화 관람객 수를 비교하여 상관계수 r을 구한 결과이다. 실험 결과 모두 양의 상관관계를 보였으며 평균 r=0.681207696의 값을 가졌다. 이로 인해 영화관련 트윗의 발생빈도가 영화의 개봉과 선형적인 연관이 있음을 알 수 있다. 이는 관련 트윗의 발생비율을 분석하면 최종적으로 전체 영화 관객 수에 대한 예측이 가능하다는 것을 의미한다.

<표 1>영화 관람객 수와 관련 트윗간의 상관관계

영화 제목	상관계수
미스터 고	+ 0.402090004
더 웹툰	+ 0.432976833
오블리비언	+ 0.469640454
아이언맨3	+ 0.527274684
감시자들	+ 0.622045841
나우 유 씨미	+ 0.813679711
애프터 어스	+ 0.852117769
관상	+ 0.872152947
설국 열차	+ 0.885959347
컨저링	+ 0.934139371

이경재 등의 연구에 의하면 개봉 첫 주의 스크린 수가 영화 흥행에 있어 가장 높은 연관이 있다는 점을 알 수 있다. 따라서 최종 누적 관객 수 예측의 선행연구 차원에서 개봉 첫날에 관람객 수 예측에 대한 접근을 해볼 수 있다. (그림 1)은 영화 개봉 첫날의 관람객 수와 개봉 전 1주일간 발생한 전체 트윗 중 관련 트윗의 비율을 나타낸 그래프이다.



(그림 1) 영화 개봉 첫날 관람객 수와
개봉 전 1주일간의 관련 트윗

우선 관람객 수와 1주일간 트윗 비율 사이에서도 상관관계를 따져볼 수 있는데, 이 둘 사이의 상관계수는 r=0.982261411 로 계산이 되었다. 따라서 이 또한 높은 양의 상관관계를 가지고 있고, 개봉 첫 주의 트윗 비율을 계산해 보면 개봉일의 관객 수를 예측할 수 있다는 결론을 내릴 수 있다.

5. 결론 및 향후 연구 계획

본 논문은 영화의 누적 관람객 수를 예측하기 위한 선행 연구로서 영화 관람객 수와 영화 관련 트윗간의 상관관계를 분석하였다. 최종적으로 정확한 전체 영화 관람객 수를 예측할 수 있다면 마케팅 비용절감, 보다 정확한 스크린 수 책정 등 다양한 측면에서의 이득을 볼 수 있다. 그러나 정확한 예측을 위해서는 단순히 트윗의 개수만이 아닌 트윗의 내용적인 측면에서도 접근하여 감정추출과 같은 문맥 기반의 연구도 이루어져야 한다. 따라서 이에 대한 부분은 향후 연구 계획으로 남겨두도록 하겠다.

참고문헌

[1] 영화진흥위원회, http://www.kobis.or.kr/, 2013

[2] 박선영, “SNS를 통한 구전 효과가 영화 흥행에 미치는 영향 - <써니>의 사례를 중심으로”, 한국콘텐츠학회논문지, 제12권, 제7호, pp. 40-53, 2012.

[3] 이경재, 장우진, “베이지안 선택 모형을 이용한 영화 흥행 예측”, 대한산업공학회 춘계학술대회 논문집, pp. 1428-1433, 2006.

[4] Twitter Streaming API, http://dev.twitter.com/docs/streaming-apis, 2013.

[5] 이범석, “블로거의 포스팅 습관을 반영한 블로그 검색엔진”, 가톨릭대학교 대학원, 박사학위논문, 2010.