KIPS Tr. Software and Data Eng.
Vol.3, No.7 pp.263~270 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.7.263

Predicting Movie Success based on Machine Learning Using Twitter

Junyeob Yim† · Byung-Yeon Hwang††

ABSTRACT

This paper suggests a method for predicting a box-office success of the film. Lately, as the growth of the film industry, a variety of studies for the prediction of market demand is being performed. The product life cycle of film is relatively short cultural goods. Therefore, in order to produce stable profits, marketing costs before opening as well as the number of screen after opening need a plan. To fulfill this plan, the demand for the product and the calculation of economic profit scale should be preceded. The cases of existing researches, as a variable for predicting, primarily use the factors of competition of the market or the properties of the film. However, the proportion of the potential audiences who purchase the goods is relatively insufficient. Therefore, in this paper, in order to consider people's perception of a movie, Twitter was utilized as one of the survey samples. The existing variables and the information extracted from Twitter are defined as off-line and on-line element, and applied those two elements in machine learning by combining. Through the experiment, the proposed predictive techniques are validated, and the results of the experiment predicted the chance of successful film with about 95% of accuracy.

Keywords : SNS, Twitter, Machine Learning, Predicting Movie Success

트위터를 이용한 기계학습 기반의 영화흥행 예측

임 준 엽† · 황 병 연††

요 약

본 논문에서는 영화의 흥행을 예측하기 위한 방법을 제안한다. 최근 영화시장이 성장함에 따라 시장의 수요를 예측하기 위한 다양한 연구들이 수행되고 있다. 영화는 비교적 수명주기가 짧은 문화상품이다. 따라서 안정적인 수익을 창출하기 위해 개봉 전 마케팅비용 및 개봉 후 스크린 수 등에 대한 설계가 필요하다. 이를 위해서는 상품의 수요와 경제적인 수익규모에 대한 계산이 선행되어야 한다. 기존 관련 연구들의 경우 예측을 위한 변수로서 주로 영화 자체의 속성들이나 시장에서의 경쟁요인 등을 이용한다. 그러나 정작 상품을 구매하는 주체인 잠재관객들에 대한 비중은 비교적 미비하다. 따라서 본 논문에서는 사람들이 가진 영화에 대한 인지도를 고려하기 위해 트위터를 하나의 설문표본으로서 활용했다. 기존에 사용된 변수들과 트위터에서 추출한 정보를 오프라인 요소와 온라인 요소로 정의하고, 두 요소를 취합하여 기계학습을 적용했다. 실험을 통해 본 논문에서 제시하는 예측기법을 검증했으며, 실험결과 약 95%의 정확도로 영화의 흥행을 예측했다.

키워드 : SNS, 트위터, 기계학습, 영화흥행 예측

1. 서 론

소셜 네트워크 서비스(Social Network Service: SNS)는 온라인상에서 지인들과의 관계를 구축하고 서로간의 소통을 돕는 새로운 커뮤니티 공간이다. SNS를 이용하는 사용자들은 사이버 공간상에서 이미 알고 있거나 같은 관심사를 가진 사용자들끼리 서로 독자적인 관계를 형성한다. 또한 기존의 온라인 소통방식보다는 빠르고 쉽게 개인의 감정이나 정보를 서로에게 공유할 수 있다[1]. 이러한 특징은 정보의 새로운 확산공간을 만들었으며, 최근 스마트 기기의 발전으로 인한 인터넷 접근성의 확대와 함께 SNS 이용자의 급격한 증가를 초래했다. 그 중 트위터(Twitter)는 2006년 3월에 서비스를 시작하여 꾸준히 사용자가 증가하고 있는 마이크로블로깅 서비스(Microbloging Service)이다. 특히, 2014년 1월을 기준으로 트위터 계정을 보유한 사용자는 약 6억 4천 5백만 명 이상이었으며, 트위터 내에서 매일 약 5천 8백만 개가량의 트윗(Tweet)이 생성되고 있다[2].

트위터의 주된 특징은 실시간성과 정보의 빠른 확산력이

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업(No. 2011-0009407)의 연구비 지원으로 수행되었음.
† 준 회 원 : 가톨릭대학교 컴퓨터공학과 석사과정
†† 종신회원 : 가톨릭대학교 컴퓨터정보공학부 교수
Manuscript Received : February 13, 2014
First Revision : April 29, 2014; Second Revision : May 23, 2014
Accepted : May 26, 2014
* Corresponding Author : Byung-Yeon Hwang(byhwang@catholic.ac.kr)

다. 트위터는 트윗이라는 개념의 140자로 제한되는 단문텍스트 서비스를 제공하는데, 이는 사용자로 하여금 보다 손쉽고 간편한 정보의 생산을 유도한다. 또한, 다른 SNS와는 다르게 매우 개방적인 네트워크 구조를 지니고 있다. 트위터 내에서는 팔로워(Follower)-팔로위(Followee)라는 개념의 관계를 사용하여 다자 간 소통이 이루어진다. 이 관계는 한쪽의 일방적인 팔로우(Follow)만으로도 관계형성이 가능하다. 예를 들어 특정 사용자들이 사용자 A를 팔로우하게 되면 사용자 A의 타임라인에 게재된 트윗들을 다른 팔로워들이 볼 수 있다. 또한 리트윗(Retweet)기능이 있어서 사용자 A가 남긴 트윗을 다른 사용자 B가 개인의 타임라인에 게재하여 사용자 B를 팔로우한 또 다른 사용자들이 해당 트윗을 볼 수 있다.

이와 같은 이유로 트위터를 하나의 설문표본으로서 활용하면 새로운 지식을 생산해내는 것이 가능하다. 또한 이는 경제적인 가치로도 평가될 수 있기 때문에 많은 분야에서 다양한 연구들이 진행되어왔다. 특히 주식시장이나 정당구도 등의 동향예측에 대한 연구결과는 트위터가 다양한 예측의 도구로서 활용될 수 있음을 의미한다. 이러한 측면에서 트위터를 이용하면 영화의 흥행도 예측할 수 있다. 영화의 경우 수익 구조의 특성상 수명주기가 비교적 짧은 문화상품이다. 또한 긴 제작기간에 비해 비교적 단기간의 수익만으로 상품의 전체적인 경제적 가치가 평가되기 때문에 정교한 판매 전략이 필요하다. 이를 위해서는 우선 해당 영화에 대한 면밀한 수요예측과 경제적인 수익규모를 예측해 볼 필요가 있으며, 이는 영화의 흥행과 밀접한 연관이 있다.

영화흥행 예측을 시도하는 연구들은 이미 오래전부터 있어왔다. 대부분의 연구들은 주로 감독, 배우, 배급사 등 영화 자체의 속성에 의존하여 영화흥행 예측을 시도했다[3]. 그러나 마케팅과 관련된 배급전략이나 실제 수요에 직접적인 영향을 미치는 잠재관객에 대한 조사 및 활용은 미비하여 명확한 한계를 가지고 있다. 따라서 본 논문에서는 이를 개선하기 위해 트위터를 하나의 설문표본으로 활용하여 잠재관객들이 가지고 있는 영화에 대한 인지도를 반영했다. 또한 제안하는 시스템의 이해를 돕고자 영화의 내적요인들과 잠재관객의 인지도를 오프라인 요소와 온라인 요소로 정의했다. 이후 이 두 요소를 기반으로 기계학습을 적용하여 영화흥행을 예측하기 위한 모델을 제시했다.

논문의 구성은 다음과 같다. 2장에서는 트위터를 이용하여 예측 및 분류하는 분석기법에 관한 연구와 기존 영화의 흥행을 예측한 관련 연구들에 대해 다룬다. 3장에서는 제안하는 예측모델의 구조와 방법을 설명한다. 4장에서는 실험을 위해 오프라인 요소와 온라인 요소에 대한 데이터를 수집하는 방법과 이를 이용한 모델의 성능을 검증한다. 마지막 5장에서 본 논문의 결론과 향후 연구를 기술한다.

2. 관련 연구

트위터는 구조적 특성상 개방적인 네트워크를 가지고 있으며 다양한 API를 지원하기 때문에 이미 많은 연구에서 이용되고 있다. [4]에서는 행동경제학에서 개인의 감정이 행동과 결정에 영향을 준다는 점을 기반으로 군중의 심리가 경제지표에 영향을 줄 수 있다는 것을 증명했다. 이를 위한 실험에서 군중의 심리를 알기 위해 트위터 피드를 이용했다. 트위터 피드란 일종의 블로그에서 쓰이는 리트윗과 같은 기능이다. 트위터 피드에서 추출한 사용자들의 감정을 수치화했고 실험 결과 86.7%의 정확도로 다우존스 산업평균지수를 예측했다. 이처럼 트위터의 패턴을 분석하면 사회현상을 예측하는 것이 가능하다.

이 외에도 트위터를 이용하여 사람들의 정치성향에 대한 예측 및 분류를 시도한 연구들이 있다. [5]에서 트위터에는 사용자들의 정치성향이나 민족성이 드러난다는 점을 강조하였고, 이러한 특징을 이용하여 기계학습 기반의 트위터 사용자 분류 시스템을 제안했다. [6,7]에서는 트위터를 이용하여 정치적 지지율을 예측하고 분석하는 데 성공하였다.

한편 문화상품의 특징을 지닌 영화는 수익의 생명주기가 다른 상품들에 비해 비교적 짧기 때문에 영화의 수익과 직결되는 영화의 흥행을 예측하는 것이 필수적이다. 따라서 이와 관련하여 이미 다방면의 연구가 진행되어왔다. 영화흥행을 예측하는 대다수의 연구들은 주로 흥행을 예측하기 위한 변수들로 영화 내적인 요소와 영화 외적인 요소로 구분하여 진행된다. 영화 내적인 요소를 중점으로 다룬 연구[8]에서는 관객들이 관람할 영화를 선택할 때 이전 작품을 기반으로 한다는 특징을 이용하여 영화배우와 감독, 작가들이 영화의 상업적인 성공과 연관이 있다는 것을 밝혔다. [9]에서는 영화 내적인 요소를 기준으로 어떠한 요소가 경제적인 수익과 직접적인 연관이 있는지를 밝혔다. 실험을 위해 액션영화, 아동용 영화, 속편여부, 수상여부, 배급사의 예산 등을 속성으로 이용했으며 이들 간의 상관관계를 밝혔다.

영화 외적인 요소를 이용한 연구들도 다양하다. [10]에서는 영화에 대한 구전효과가 흥행에 영향을 미칠 수 있다는 점을 언급했다. 또한 이를 가시화하기 위해 야후에서 제공하는 영화정보 웹 사이트에 작성된 사용자들의 의견을 바탕으로 실험을 진행했다. 실험결과 사용자들에 의한 영화의 구전활동은 영화개봉 직전과 영화개봉 후 1주일 동안이 가장 활발했으며, 영화에 대한 기대도는 개봉 후 1주일을 기준으로 조금씩 낮아진다는 것을 밝혔다. 따라서 특정 영화에 대한 관객들의 구전활동의 규모를 연구한다면 영화의 흥행을 예측할 수 있다는 결론을 내릴 수 있다. 이와 더불어 구전의 내용에 대한 긍정 부정의 변화 추이는 영화의 흥행과 크게 연관이 없고 오직 규모만이 연관이 있었다고 밝혔다. 이에 본 논문에서도 온라인 구전활동의 긍정 부정 판별을 위한 트윗 내용의 감정추출은 수행하지 않았다.

해외의 사례 외에 국내에서도 영화의 흥행을 예측하고 특정 요소와의 상관관계를 밝힌 연구들이 있다. [11]에서는 국내에서 개봉한 영화 <써니>를 중심으로 영화 흥행에 대한 SNS의 영향력을 언급했다. 특히 영화의 생명주기를 시간순서로 나누어 각 시점별로 SNS가 마케팅으로서 어떻게 이용

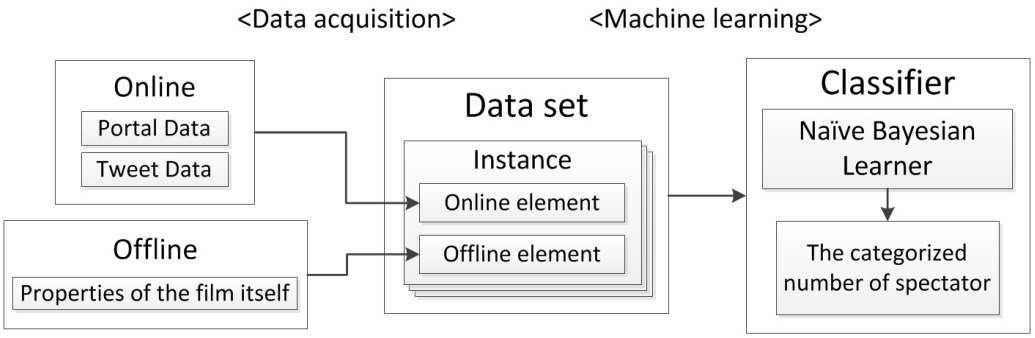


Fig. 1. The organization of predicting movie success model

될 수 있는지를 설명했다. 또한 SNS가 영화흥행의 결정요인이자 참고요인으로서 이용될 수 있다는 점을 강조했다. 이와 더불어 [12, 13]에서는 영화의 최대 관객 수와 SNS가 높은 상관관계를 가진다는 점을 강조했으며, 이를 이용하여 영화흥행 예측이 가능하다는 점을 부각하였다.

앞서 언급한 사례들에서는 영화의 내적인 요소보다 영화 관객 수에 더 큰 상관관계를 가지는 SNS를 적극적으로 활용하여 영화의 흥행을 예측하는 데에는 한계가 있었다. 따라서 본 연구에서는 기존 연구에서 영화 내적인 요소를 사용한 것과 더불어 SNS를 추가적으로 활용한 방법 간의 예측 정확도의 차이를 밝히고 최종 영화흥행 예측 모델을 제시한다.

3. 영화 흥행 예측기법

3.1 영화 흥행 예측 모델

본 연구에서 제안하는 영화 흥행 예측 모델은 크게 두 가지로 나뉘며 전체적인 구조는 Fig. 1과 같다. 예측을 위해 데이터를 수집하는 단계와 기계학습을 통해 예측을 수행하는 단계로 구성했다. 먼저 데이터 수집 단계에서는 특정 영화에 대한 온라인/오프라인 정보들을 수집한다. 데이터에 대한 자세한 설명은 3.2절에서 다루도록 한다. 수집된 데이터는 서로 결합되어 Table 1과 같이 한 편의 영화에 대해 하나의 인스턴스(Instance)를 생성한다. 생성된 인스턴스에는 영화정보에 대한 각각의 속성 값들이 수치화 또는 범주화되어 있다. 따라서 수집된 다수의 영화 편수만큼 동일 개수의 인스턴스가 생성된다. 이를 기반으로 기계학습을 적용했고 학습결과 최종 반환되는 클래스(Class)로 영화의 흥행을 예측한다. 예측은 시점에 따라 총 2번 수행된다. 영화의 개봉일과 개봉 후 1주일에 예측을 시도하며, 개봉 후 1주일에는 1주일간의 관람객 수에 대한 데이터가 각각의 인스턴스에 추가된다.

Table 1. An example of the instance

Movie title	Instance					
	Online element			Offline element		
	Screening grade (Age)	Running time (Minute)	…	Rated portal (Score)	Related tweets ratio (%)	…
Snowpiercer	15 over	125	…	7.99	0.12	…
…	…	…	…	…	…	…
Gravity	12 over	90	…	8.28	0.14	…

3.2 영화 흥행 예측 방법

영화의 흥행 정도는 해당 영화를 관람한 최대 관객 수로 수치화해서 표현할 수 있다. 그러나 0명 단위의 정확한 최대 관객 수를 예측하는 것은 불가능에 가까우며 무의미하다. 따라서 본 연구에서는 예측할 관객 수의 범위를 범주화하여 영화를 범주에 맞게 분류했다. 범주의 기준은 [12, 14]를 참고하였으며 Table 2와 같다. 범주화된 관객 수 범위는 기계학습을 통한 예측의 결과로서 하나의 클래스로 표현된다.

Table 2. Categorization of the number of spectator

Section	The number of spectator
1	less than 500,000
2	500,000~1,000,000
3	1,000,000~3,000,000
4	3,000,000~5,000,000
5	more than 5,000,000

본 연구에서는 나이브 베이지안 분류(Naive Bayes' Rule) 방법[15]을 영화 흥행 예측의 도구로서 활용했다. 나이브 베이지안 분류 방법은 하나 이상의 독립적인 속성들로부터 결과를 분류하는 확률적인 분류 방법이다. 각각의 독립적인 속성들이 결합되어 하나의 인스턴스를 이루며 여러 개의 인스턴스를 통해 분류할 결과에 대한 확률을 계산한다.

영화의 흥행을 예측하기 위해 사용된 각각의 속성들을 $x_1, x_2, \cdots, x_n$ 으로 표현하였다. 또한 분류 결과에 해당하는 관객 수 범위는 클래스 C로 표현하였고, C의 범주인 각 구간의 확률은 값 c로 표현하였다. 따라서 최종 계산해야 할 c는 (1)로 정의되며, 새로 입력될 영화에 대한 인스턴스는 (1)을 통해 가장 높은 c로 분류된다.

$$c = \underset{c_i \in C}{\arg\max} P(c_i | x_1, x_2, \cdots, x_n) \quad (1)$$

(1)은 베이지안 정리를 이용하면 (2)와 같이 변형할 수 있다. 이는 미리 계산된 확률들을 이용하여 예측할 관객 수 범위가 어느 구간에 포함되어야 할지를 결정한다는 것을 의미한다. 계산 결과 가장 확률이 높은 구간에 포함시킨다.

$$c = \underset{c_i \in C}{\arg\max} \frac{P(x_1, x_2, \cdots, x_n | c_i)\, P(c_i)}{P(x_1, x_2, \cdots, x_n)} \quad (2)$$

그러나 (2)의 경우 서로 다른 값 간의 대소만을 비교하는 것이므로 같은 값을 가지는 분모 $P(x_1, x_2, \cdots, x_n)$는 따로 계산할 필요가 없다. 따라서 예측을 위한 계산에 이용될 공식은 (3)으로 변환된다.

$$c = \underset{c_i \in C}{\arg\max} P(x_1, x_2, \cdots, x_n | c_i)\, P(c_i) \quad (3)$$

마지막으로 나이브 베이지안 분류 방법은 각 분류 조건인 $x_1, x_2, \cdots, x_n$가 서로 조건부 독립이다. 따라서 (3)을 (4)와 같이 단순화시킬 수 있다.

$$c = \underset{c_i \in C}{\arg\max} \Pi_j P(x_j | c_i)\, P(c_i) \quad (4)$$

최종 완성된 (4)를 이용하여 영화의 흥행을 예측한다. (4)에서 이용될 분류 조건들은 Table 3과 같다.

Table 3. Definition of total attribute

Attribute classification	Attribute name	Domain (numeric|nominal)
Offline element	Release day of the week	Day of the week (nominal)
	Screening grade	Character (nominal)
	Running time	Minute (numeric)
	Director effect	Integer (numeric)
	Actor effect	Integer (numeric)
	Country	Character (nominal)
	Distributer effect	Real number (numeric)
	After the opening week box-office results	Integer (numeric)
Online element	Rated portal	Real number (numeric)
	The number of portal survey participants	Integer (numeric)
	Incidence of related tweets during a week before release	Real number (numeric)
	Incidence of related tweets during a week after release	Real number (numeric)

Table 3에 제시된 속성은 본 논문에서 제안하는 영화 흥행 예측을 위한 관련 변수들로서 전체 예측 모델에서 수집 단계에 입력된다. 이는 오프라인 요소 수집과 온라인 요소 수집으로 나뉘는데, 오프라인 요소의 경우 기존의 영화의 흥행을 예측할 때 주로 사용되는 속성들을 이용한다. 반면에 온라인 요소의 경우 트위터와 포털 사이트의 자료를 이용하며 자세한 내용은 4.1절에서 다루도록 한다.

Table 3에서 오프라인 요소에 해당하는 감독 효과와 배우 효과의 경우 인물들의 이름만으로 예측 모델에 학습을 시키기는 어렵다. 따라서 해당 감독과 배우가 이전 작품에서 가장 높은 흥행을 기록했을 때의 관객 수를 값으로 지정한다. 여기서 하나의 영화에 여러 명의 감독과 배우가 있을 수 있는데, 이때는 가장 높은 기록을 보유한 감독과 배우를 대표인물로 지정한다. 단, 배우의 경우 주연으로 출연한 작품만을 고려한다. 예를 들어 영화배우 송강호는 영화 <괴물>에서 약 1,300만 명의 관객 수를 동원했으며 이후 출연작 중에 더 높은 흥행 기록은 없었다. 따라서 영화 <괴물>이 개봉한 시점 이후로 영화배우 송강호가 주연으로 출연했고 해당 영화에 함께 주연으로 출연한 다른 배우 중에 더 높은 흥행실적을 가진 배우가 없을 경우 해당 영화의 배우 효과는 1,300만이 된다. 이와 더불어 배급사효과는 작년도 개봉작들의 평균 관객동원 수로 지정하였다. 감독과 배우의 경우 영화제작에 참여한다는 사실만으로도 흥행에 영향을 줄 수 있다. 그러나 배급사는 앞서 언급한 변수와는 달리 제작비의 규모나 마케팅능력과 같이 영화 외적인 부분에 영향을 미칠 수 있는 요소이다. 따라서 변수의 특징에 맞게 배급사효과는 1년간의 흥행실적을 말해주는 관객 수를 속성으로 반영한다.

한편 나이브 베이지안 분류 기법은 속성에 포함된 특정 범주에 대해 각각이 일어날 확률을 기반으로 클래스에 해당하는 결과를 분류해내는 방법이다. 따라서 Table 3에서 수치화되어 표현된 연속적인 값을 가지는 속성은 그대로 적용할 수 없다. 이를 위해 (5)와 같이 이미 잘 알려진 가우지안 함수를 이용했다. (5)를 이용하면 해당 속성에 대한 독립적인 확률 값을 계산하여 (4)에 적용할 수 있다. 여기서 μ는 하나의 속성에 대해 연속적인 값들의 평균을 의미하며 σ은 표준편차를 의미한다.

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

4. 실험 및 결과

4.1 데이터 수집

온라인 요소에 대한 데이터를 수집하여 영화의 잠재고객들의 수요를 수치화하기 위해 포털 자료와 트위터를 이용했

다. 우선 포털 자료에 대한 데이터는 국내에서 가장 많은 사용자를 보유한 네이버에서 제공하는 네이버 영화[16]의 설문 자료를 이용하였다. 관련 자료는 2013년 11월 27일에 일괄적으로 수집했다. 포털 자료에는 포털 평점과 포털 평점 설문에 참여한 참여자 수가 있다. 포털 평점은 네이버 사용자들의 설문을 통해 사용자 개인이 느끼는 영화에 대한 평가를 0점부터 10점까지 입력한 자료이다. 한편 포털 평점 설문에 참여한 참여자 수는 사용자들의 영화에 대한 인지도를 고려하기 위해 속성에 추가하였다.

트위터에 대한 데이터를 수집하기 위해 트위터에서 제공하는 스트리밍(Streaming) API[17]를 이용하였다. 트위터는 트위터 사(社)와 별도의 파이어호스(Firehose)계약이 없다면 일반 개발자에게는 전체 트윗 발생량의 무작위 1%만을 제공한다. 그러나 본 연구의 경우 전체 트윗 내에서 특정 영화와 관련된 트윗 비율을 예측에 이용할 데이터로서 이용한다. 따라서 지속적으로 트윗을 수집하고 저장할 수 있다면 트윗의 수집량과는 관계없이 예측이 가능하다.

수집한 트윗 코퍼스(Corpus)에서 특정 영화에 대한 트윗들을 추출하기 위해 2013년 4월 1일부터 8개월간 발생한 트윗을 수집했다. 트윗 데이터의 수집과정은 Fig. 2와 같다. 우선 트윗의 내용을 분석하여 특정 영화제목이 포함된 트윗을 해당 영화에 대한 트윗으로 구분했다. 이때 띄어쓰기는 고려하지 않았다. 예를 들어 실험 데이터로 쓰인 트윗 중 “휴..은밀하게위대하게 두번째본다.. 다 죽는다고ㅠㅠ”의 경우 띄어쓰기를 제대로 하지 않았으나 의미상 영화 제목을 포함하고 있으므로 영화 <은밀하게 위대하게>에 대한 트윗으로 구분했다.

다음으로 [10]에서 언급한 바와 같이 트윗은 개봉 전 1주일간의 트윗과 개봉 후 1주일간의 트윗으로 구분했다. 또한,

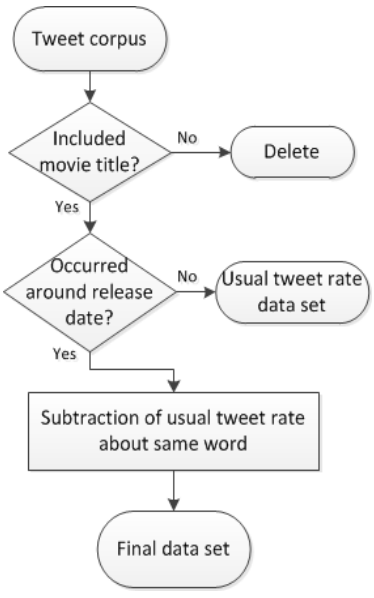


Fig. 2. Collecting process of tweet data

데이터로 이용될 영화들에 대해 각각 개봉일이 다르므로 영화별로 기간을 달리하여 트윗 데이터를 구분했다. 결과적으로 개봉 전과 후에 트위터 내에서 각각의 영화마다 해당 영화가 언급된 트윗의 비율이 두 개씩 반환된다.

한편 영화 <관상>과 같이 영화제목 자체가 평소에도 일반적으로 쓰이는 단어일 수 있다. 이 경우 해당 트윗이 영화 <관상>만을 뜻하는 내용인지 명확히 알 수 없다. 따라서 평소에 해당 영화 제목을 포함했던 트윗의 비율을 따로 저장하여 영화 개봉일 전후 1주일간 증가한 만큼의 트윗 비율을 기준으로 삼았다.

오프라인 요소에는 영화 내적인 요소가 포함된다. 영화는 2013년 4월에서 10월 사이에 개봉한 영화들 중 무작위로 60편의 영화를 선택하여 데이터 셋(Data Set)을 구성했다. 영화에 대한 정보는 영화진흥위원회[18]에서 제공하는 자료를 기준으로 수집했다.

4.2 실험 결과

실험에 들어가기에 앞서 실험에 사용된 인스턴스들과 본 논문에서 제안한 속성들이 각각 얼마나 높은 정확도를 가지는지 알아보았다. Fig. 3은 범주별로 각 데이터가 어떻게 분포하는지를 나타낸 그래프이다. 총 5개의 범주에 대해 총 60편의 영화가 실험에 이용되었으며 평균 181만 명의 관객이 동원되었다. Fig. 4는 60편의 영화에 대해 하나의 속성만으로 공통된 기계학습을 적용한 결과 도출된 영화 흥행 예측에 대한 정확도이다. 예를 들어 속성으로 사용된 개봉요일(Release day of the week)이 36.7%라는 것의 의미는 해당 속성만을 사용해서 영화의 흥행을 예측할 경우의 정확도를 뜻한다. 실험 결과 온라인 요소에 해당하는 속성들이 비교적 높은 정확도를 보였다. 한편 배급사의 경우 오프라인 요소들 중 가장 높은 예측 정확도를 보였다. 이는 경제적 힘이 있는 배급사가 상대적으로 많은 제작비를 투자하여 그에 비례하는 관객 수를 확보했다는 것으로 해석된다. 또한 온라인 요소에 해당하는 포탈 평점은 매우 낮은 예측 정확도를 보였으며, 포털 평점 설문 참여자수는

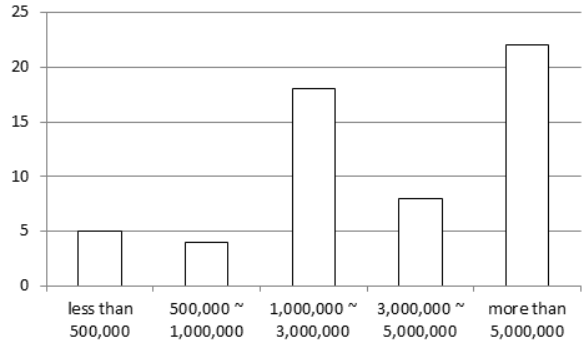


Fig. 3. The number of instances by each section

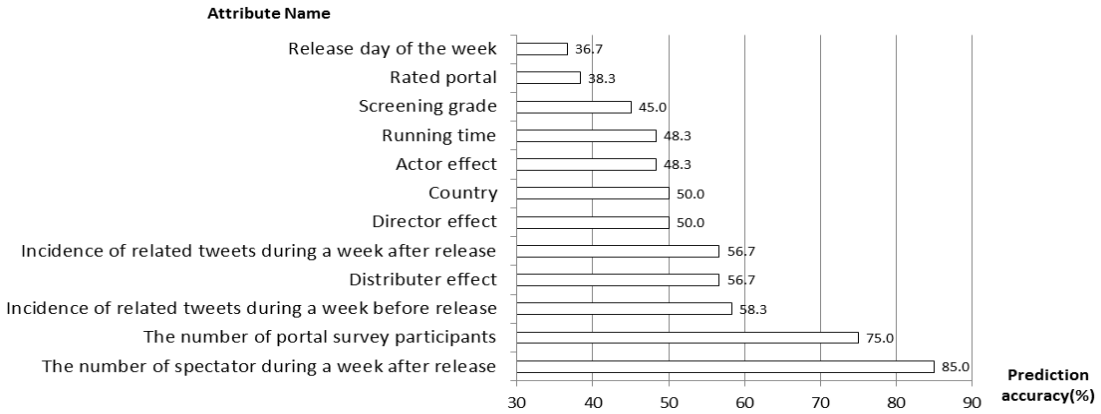


Fig. 4. Prediction accuracy by attributes

높은 예측 정확도를 보였다. 이는 [10]에서처럼 영화의 흥행이 영화를 관람한 관객들의 평가보다는 특정 영화를 얼마나 많은 사람들이 인지하고 있는지와 연관이 있다는 것으로 해석된다.

다음으로 기존의 연구들에 비해 본 논문에서 제시한 예측 모델이 어느 정도의 정확도 향상을 보였는지 알아보았다. Fig. 5는 온라인 요소의 포함여부와 예측하는 시점에 따라 예측 정확도가 달라질 수 있음을 나타낸 것이다. 기존 연구에서처럼 온라인 요소를 제거할 경우 개봉일과 개봉 1주일 후의 예측에서는 각각 73.3%와 88.3%의 정확도를 보였다. 반면에 본 연구에서 제안하는 온라인 요소를 포함할 경우 각각 78.3%와 95%로 두 시점에서 모두 정확도 향상이 있었다. 이는 영화의 흥행을 예측할 때 온라인 요소를 포함하면 보다 높은 정확도로 예측을 수행할 수 있다는 것을 의미한다. 또한 영화 개봉 1주일 후에 예측을 수행할 경우 관람객 수에 관한 속성이 추가됨으로써 잠재 관객들의 영화에 대한 관심이 온라인상에 반영된 것으로 해석할 수 있다. 따라서 온라인 요소를 포함하고 개봉 후 1주일간의 트윗을 함께 분석하는 것이 가장 높은 정확도를 보인다고 결론내릴 수 있다.

마지막으로 온라인 요소의 포함여부에 대해 데이터 개수가 다를 경우에도 예측 정확도의 차이를 보이는지 알아보기 위해 실험을 진행하였다. 실험을 위해 이전 실험에서 이용한 두 가지의 동일 인스턴스 60개를 10개씩 구분하였고, 각 클래스의 비율을 고려하여 무작위로 추출하였다. 실험 결과 Fig. 6과 같은 추이를 보였다. 우선 인스턴스의 개수가 적을 경우에는 예측 정확도가 매우 높았다. 이는 나이브 베이지안 분류기가 데이터에 기반한 분류 규칙을 생성하기 때문이다. 따라서 두 실험 결과를 비교하려면 전체적인 추이를 살펴보아야 한다. 기존 대다수의 연구에서처럼 온라인 요소를 배제하고 예측을 한 결과보다는 온라인 요소를 이용하여 예측을 한 결과가 전체적으로 더 높은 예측 정확도를 가지는 것을 확인할 수 있다. 또한 온라인 요소를 포함한 경우 가장 많은 데이터 셋이었던 60개의 데이터에서 95%의 예측 정확도를 얻었다. 따라서 이를 본 논문에서 제시하는 영화 흥행 예측 모듈의 최종 정확도로 정한다.

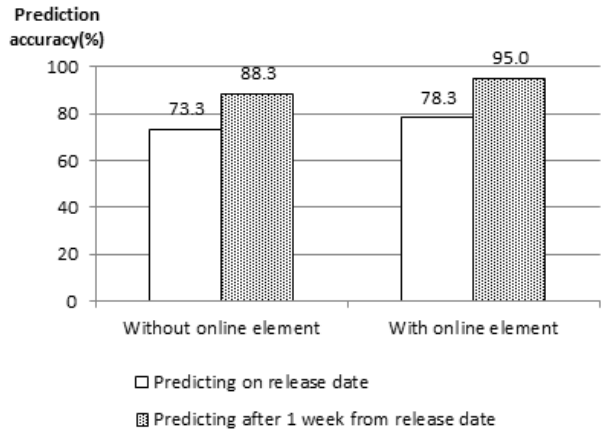


Fig. 5. Comparison of prediction accuracy

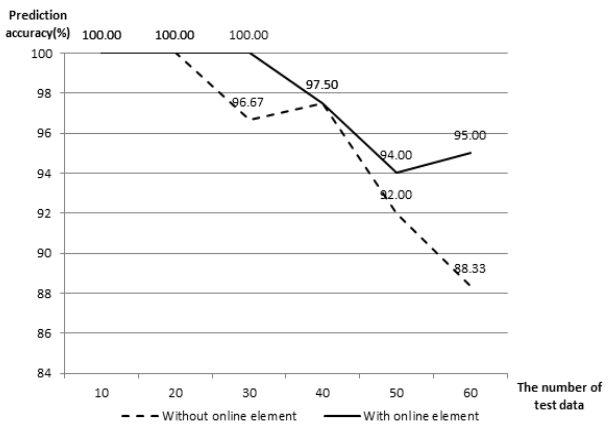


Fig. 6. Prediction accuracy by the number of data

5. 결론 및 향후 연구계획

본 연구에서는 영화의 흥행을 예측하기 위해 트위터를 이용했다. 기존 연구에서 영화 내적인 요소들만을 위주로 예측을 시도했던 것과는 달리 잠재 관객들의 영화에 대한 인지도를 고려하였으며, 이를 위해 트위터와 포털 자료를 이용했다. 실험을 통해 제안하는 예측 모델의 효율성을 검증하였으며, 최종 실험 결과 95%의 예측 정확도를 보였다.

향후 연구로는 보다 정확한 영화 관련 트윗 수집을 위해 노이즈 제거 및 자연어 처리에 대한 부분을 연구할 계획이다. 또한 영화 흥행 예측에 필요한 다양한 속성들에 대해 신경망(Neural Network) 분석과 같은 방법을 통해 각 속성별 예측 기여도를 분석할 것이다. 이를 통해 특정 속성을 추가하거나 제거한다면 보다 높은 예측 정확도를 얻을 수 있을 것으로 기대된다. 이와 더불어 제안하는 예측 방법은 예측 결과로써 반환되는 최종 결과가 범주화 된 형태이기 때문에 정밀한 성능평가가 어렵다. 따라서 이에 대한 개선 방안도 향후 연구할 계획이다. 마지막으로 결과에 대한 영향은 미비한 것으로 보이나 애니메이션 영화의 경우 배우효과에 관한 속성을 정확히 반영하지 못했다. 따라서 이와 관련된 해결방안도 찾아야 할 것이다.

Reference

[1] L. Barbosa and J. Feng, “Robust Sentiment Detection on Twitter from Biased and Noisy Data,” Proc. of the 23rd International Conference on Computational Linguistics, pp.36-44, 2010.

[2] Statistic Brain (2014, Jan. 01), Twitter Statistics [Online]. Available : http://www.statisticbrain.com/twitter-statistics/

[3] E.-M. Kim, “The Determinants of Motion Picture Box Office Performance: Evidence from Movies Exhibited in Korea,” Korean Journal of Journalism & Communication Studies, Vol.47, No.2, pp.190-220, 2003.

[4] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” Journal of Computational Science, Vol.2, No.1, pp.1-8, 2011.

[5] M. Pennacchiotti and A.-M. Popescu, “A Machine Learning Approach to Twitter User Classification,” Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, pp.281-288, 2011.

[6] E. T. K. Sang and J. Bos, “Predicting the 2011 Dutch Senate Election Results with Twitter,” Proc. of the SASN/the EACL Workshop on Semantic Analysis in Social Network, pp.53-60, 2012.

[7] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting Election with Twitter: What 140 Characters Reveal about Political Sentiment,” Proc. of the Fourth International AAAI Conference on Weblogs and Social Media, pp.178-185, 2010.

[8] S. Albert, “Movie Stars and the Distribution of Financially Successful Films in the Motion Picture Industry,” Journal of Cultural Economics, Vol.22, No.4, pp.249-270, 1998.

[9] N. Terry, J. W. Cooley, and M. Zachary, “The Determinants of Foreign Box Office Revenue for English Language Movies,” Journal of International Business and Cultural Studies, Vol.2, No.1, pp.1-12, 2010.

[10] Y. Liu, “Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue,” Journal of Marketing, Vol.70, No.3, pp.74-89, 2006.

[11] S.-Y. Park, “Influence of the Word-of-Mouth Effect through SNS on the Movie Performance-Focused on the Case of <Sunny>-,” Journal of the Korea Contents Association, Vol.12, No.7, pp.40-53, 2012.

[12] Y. H. Kim, J. H. Hong, “A Study for the Development of Motion Picture Box-office Prediction Model,” Communications of the Korean Statistical Society, Vol.18, No.6, pp.859-869, 2011.

[13] J. Yim, B.-Y. Hwang, “An Analysis of Correlation between Movie Attendance and Related Tweets for Predicting Box Office,” Proc. of the 40th Korea Information Processing Society Fall Conference, Vol.20, No.2, 2013.

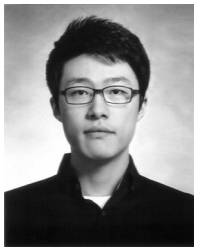
[14] G. Lee, U. Jang, “Predicting Financial Success of a Movie Using Bayesian Choice Model,” Proc. of the KIIE/KORMS Spring Conference, pp.1428-1433, 2006.

[15] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Ed., pp.85-99, Morgan Kaufmann Series in Data Management Systems, 2011.

[16] NAVER (2013, Nov. 27), Naver Movie [Online]. Available: http://movie.naver.com/

[17] Twitter (2012, Sep. 24), The Streaming APIs|Twitter Developers [Online]. Available: https://dev.twitter.com/docs/streaming-apis

[18] Korean Film Council (2013, Nov. 27), KOFIC Cinema Tickets Integrated Computer Network [Online]. Available: http://www.kobis.or.kr/



임 준 엽

e-mail : junyeob1205@naver.com
2013년 가톨릭대학교 컴퓨터공학과(학사)
2013년~현 재 가톨릭대학교 컴퓨터공학과 석사과정
관심분야 : 소셜네트워크분석, 데이터베이스, 데이터마이닝, 정보검색

황 병 연

e-mail : byhwang@catholic.ac.kr
1986년 서울대학교 컴퓨터공학과(학사)
1989년 KAIST 전산학과(석사)
1994년 KAIST 전산학과(박사)
1994년~현 재 가톨릭대학교 컴퓨터정보공학부 교수
1999년~2000년 (美) 미네소타대학교 방문교수
2007년~2008년 (美) 캘리포니아주립대학교 방문교수
관심분야 : 소셜네트워크분석, XML 데이터베이스, 정보검색, 데이터마이닝, 지리정보시스템