영화 스크립트 텍스트 마이닝을 통한 흥행성과 예측

하현수, 황병연
가톨릭대학교 컴퓨터공학과
e-mail : {hss0924, byhwang}@catholic.ac.kr

Assessing Box Office Performance Using Movie Scripts Text Mining

Hyunsoo Ha, Byeong-Yeon Hwang
Dept. of Computer Science and Engineering, The Catholic University of Korea

요 약

영화 흥행 실패의 리스크를 줄이기 위해 객관적인 흥행 예측 지표가 요구된다. 본 논문에서는 영화 스크립트의 텍스트를 분석하여 흥행성과를 예측하는 기법을 제안한다. 객관적인 흥행 예측 지표는 누적 관객 수와 누적 매출액으로 설정하였다. 실험은 2010년 1월 1일부터 2016년 8월까지 개봉한 영화 중에서 누적 관객 수와 누적 매출액을 기준으로 상위 50위까지의 영화 스크립트를 분석하여 진행했다. 실험을 통해 영화 제작에 앞서 스크립트 분석만을 활용한 영화 흥행성과 예측이 가능함을 보였다.

1. 서론

2000년대 이후로 영화와 드라마의 제작비 규모는 점차 증가하는 추세이다. 제작비 규모가 커질수록 흥행 실패에 따른 리스크도 커지게 된다. 따라서 영화 제작과 개봉에 따른 이윤과 비용을 가늠할 수 있는 객관적인 지표가 필요하다. 영화 흥행 예측 알고리즘은 영화 스크립트의 텍스트를 장면 별로 분류하고 문장들의 의미를 파악한 뒤, 감정 단어에 가중치를 부여하여 성과를 예측하는 과정을 거친다.

논문의 구성은 다음과 같다. 2장에서는 본 논문의 관련연구로서 Eliashberg의 연구[1]와 김상호의 연구[2]를 다룬다. 3장의 실험방법에서는 영화 스크립트의 분석 및 예측 과정을 설명하고, 4장의 실험결과에서는 누적 매출액과 누적 관객 수에 대한 예측 정확도에 대해 기술한다. 마지막으로 5장에서는 본 연구에 대한 결론과 향후 연구계획을 소개한다.

2. 관련연구

[1]에서는 영화 스크립트 분석을 기반으로 영화의 흥행 실적을 예측하는 시스템을 개발하였다. 영화 스크립트 도메인 지식 및 자연 언어 처리를 사용하여 박스 오피스 결과를 도출한다. 영화 스크립트 분석 결과 값과 실제 박스 오피스의 결과를 유사도 및 상관관계 실험을 통해 입증하였다. 결론적으로 89%의 정확성을 나타내었으며, [1]에서 제안한 Kernel-I/II 기법을 거쳐 경제적으로 유의미함을 증명했다. 이 연구 결과는 영화 스크립트의 텍스트 마이닝을 통하여 실질적인 매출액까지 예측할 수 있다는 것을 의미한다.

[2]는 영화의 흥행성과에 어떤 요인들이 유의미한 영향력을 미치는지를 분석한 연구이며, 2012년 개봉한 한국 영화를 대상으로 실험하였다. 극장 관객 수를 종속변인으로 지정하고 요인으로는 개봉스크린 규모, 제작비, 관람등급, 영화 장르, 개봉시기, 속편, 배급사 파워, 전문가 평가, 온라인 평가로 지정하였다. 그러나 [2]는 작품 외적 요소에 관한 연구로 작품 내적인 요소들을 분석하지 않은 한계가 있다. 본 논문에서 제안하는 기법은 영화 스크립트를 대상으로 하여 작품 내적 요소를 분석하고 예측하는 기법으로 기존의 연구와의 차별점이 있다.

[3]은 감성분석에 의한 빅 데이터 분석을 실험한 연구이다. 텍스트에 포함된 감성 어휘를 파악하고 패턴을 발견하여 속성을 정의하였다. 강조 부사와 감성 어휘에 가중치를 두어 감성분석 알고리즘을 제안하였다. 또한 제품 특징별 어휘 사전을 구축하였다. 감성 어휘에 가중치를 부여하는 점을 일부 본 연구에 적용시킬 계획이다.

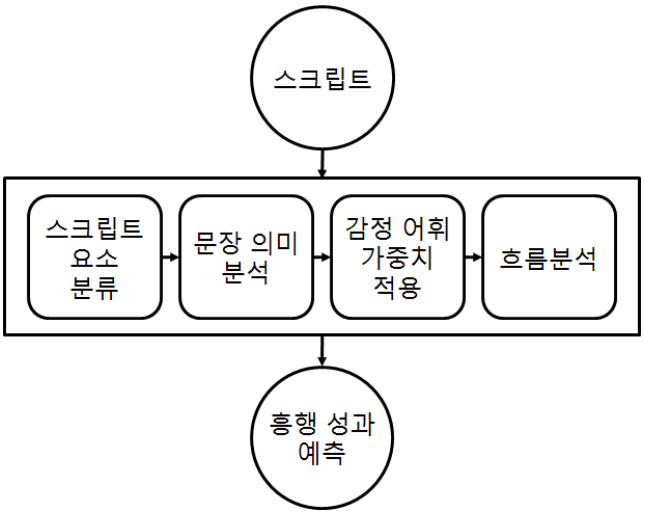
3. 실험방법

본 논문에서는 2010년 1월 1일부터 2016년 8월까지 개봉한 영화 중에서 누적 관객 수와 누적 매출액을 기준으로 상위 50위까지의 영화 스크립트를 분석하여 실험을 진행했다. (그림 1)은 영화 스크립트 텍스트 마이닝 알고리즘의 흐름도이다. 우선 영화 스크립트는 요소 별로 분류한다. 장면 별로 구분지어서 대사와 지문을 분리하여 스크립트 분석을 위한 스크립트 형태 변환 과정을 진행한다.

그 다음으로 문장 의미 분석 과정을 거친다. 의미 분석 과정은 스크립트 내의 단어들을 형태소 분석기[4]를 이용해 분리하는 과정을 의미한다. 이 과정을 거치면서 스크립트에 사용된 단어들의 의미와 빈도를 파악하여 수치화 시킨다. 그 후 감정 어휘 가중치 적용 과정을 진행한다. 이 과정은 스크립트 장면(scene)을 기준으로 즐거움, 액션, 분노, 감동, 로맨스, 스릴·공포 등의 장면으로 분류하여 단

어들과 연결시켜 가중치 수치를 적용하는 과정이다. 예시로 로맨스 분류 장면과 연결할 스크립트 내의 단어는 키스, 포옹 등을 들 수 있다.

마지막으로 흐름 분석 과정은 각 장면별 스크립트를 분석한 결과의 수치 값이다. 각 분류된 장면마다 키워드 빈도수와 사용된 감정 단어의 가중치를 계산하여 최종적인 결과 값인 흥미도를 측정한다. 총 100개의 데이터 영화 스크립트의 흥미도 값을 기계학습 시키고 테스트 데이터 10개의 스크립트를 적용시켜 실제 누적 매출액, 누적 관객 수와 예측된 누적 매출액, 누적 관객 수를 비교하였다.



(그림 1) 스크립트 텍스트 마이닝 흐름도

4. 실험결과

기계학습을 시켰던 100개의 영화들에 대한 실제 누적 관객 수와 누적 매출액은 영화진흥위원회[5, 6]의 데이터를 기준으로 실험을 진행하였다. 테스트 데이터로 활용될 스크립트는 총 10개의 영화 스크립트이며, 그중 5개는 흥행에 성공한 영화이고 5개는 흥행에 실패한 영화로 선별하였다.

<표 1> 누적 관객 수와 누적 매출액 등급 기준

등급	누적 관객 수(명)	누적 매출액(원)
1	100만 이하	100억 이하
2	100만 ~ 200만	100억 ~ 200억
3	200만 ~ 300만	200억 ~ 300억
4	300만 ~ 400만	300억 ~ 400억
5	400만 ~ 500만	400억 ~ 500억
6	500만 ~ 600만	500억 ~ 600억
7	600만 ~ 700만	600억 ~ 700억
8	700만 ~ 800만	700억 ~ 800억
9	800만 ~ 1000만	800억 ~ 1000억
10	1000만 이상	1000억 이상

누적 매출액과 누적 관객 수는 정확한 수치를 예측하기에 어려움이 따르기 때문에 구간별 등급으로 지정하였다. 총 10개의 등급으로 나누었으며 자세한 기준은 <표 1>과 같다. 스크립트 분석의 결과 수치 값이 낮은 스크립트는 흥행 실패라는 결과가 도출되도록 하였다. 흥행 실패의 정확한 기준은 손익분기점을 넘지 못한 영화로 설정하였다.

<표 2> 실제 영화 실적과 스크립트 예측 실적 비교

영화	실제 누적 관객 수	예측 누적 관객 수	실제 누적 매출액	예측 누적 매출액
1	9	8	10	10
2	7	10	9	10
3	5	6	6	8
4	10	9	10	9
5	9	6	10	7
6	2	흥행실패	1	흥행실패
7	2	4	3	흥행실패
8	3	흥행실패	3	5
9	1	흥행실패	1	흥행실패
10	1	흥행실패	2	흥행실패

<표 2>는 영화들의 실제 누적 관객 수와 누적 매출액을 본 논문에서 제안한 기법을 적용하여 스크립트 예측 결과에 따른 예측 누적 관객 수와 누적 매출액을 비교한 표이다. 영화 1부터 5는 흥행한 영화이고 6부터 10은 흥행에 실패한 영화이다. 흥행한 영화의 누적 관객 수 등급 예측의 오차율은 23.46% 이며, 누적 매출액 등급 예측에 대한 오차율은 16.89%이다. 비교적 낮은 오차율을 보이고 있으며, 흥행에 실패한 영화들에 대해서는 대부분 흥행실패라는 제대로 된 결과 값을 도출하였다. 따라서 흥행에 성공할 가능성이 있는지에 대하여 예측할 수 있는 정도의 성과를 보였다.

그러나 영화 7의 누적 관객 수 등급 예측과 영화 8의 누적 매출액 등급 예측에서는 오류가 있었다. 영화 7과 8의 스크립트 내의 감정 어휘의 빈도수가 높아 비교적 높은 가중치를 적용시킨 점이 결과에 영향을 끼쳤다.

5. 결론 및 향후 연구과제

본 논문에서는 영화의 스크립트만을 분석하여 누적 관객 수와 누적 매출액을 예측하는 기법을 제안하였다. 텍스트 마이닝을 이용하여 스크립트를 분석하는 과정에 대하여 알아보았다. 또한 실제 영화 흥행 성과와 예측한 흥행성과를 비교하여 스크립트 분석만을 활용한 영화 흥행성과 예측이 가능함을 보였다.

본 연구의 한계점으로서 흥행에 성공한다면 성공하는 요인이 무엇인지를 파악하지 못하는 점을 들 수 있다. 또한 스타파워나 대형 제작, 배급사의 요인을 적용하지 않았다는 점이 본 연구의 한계점이다. 이러한 한계점은 외부적인 흥행 요인을 추가적으로 적용시키고, 흥행 성공 요인을 파악하여 더 세부적인 예측 결과를 도출하는 것을 목표로 정하여 개선 할 것이다.

참고문헌

[1] J. Eliashberg, S. K. Hui, Z. J. Zhang. “Assessing Box Office Performance Using Movie Scripts : A Kernel-Based Approach”, IEEE Computer Society, 26(11), pp.2639-2648. 2014.

[2] 김상호, 한진만. “한국 영화의 흥행성과 결정요인 분석”, 사회과학연구, 53(1), pp.191-214. 2014.

[3] 서정렬, 고찬. “감성분석에 의한 Big Data 분석”, 융복합지식학회논문지, 2(1), pp.15-21. 2014.

[4] Apache Lucene Korean Analyzer and dictionary, http://sourceforge.net/projects/lucenekorean, 2013.

[5] 영화진흥위원회, http://www.kobis.or.kr/kobis/business/main/main.do

[6] 영화진흥위원회 OpenAPI, http://www.kobis.or.kr/kobisopenapi/homepg/main/main.do