영화 흥행에 영향을 미치는 요인 분석

이정원[1*] · 전병일[1] · 김세민[2] · 이규전[1] · 이충호[1**]

[1]한밭대학교 · [2]전주교육대학교

An Analysis of the Factors Affecting the Movie's Popularity

Jeongwon Lee[1*] · Byungil Jeon[1] · Semin Kim[2] · Gyujeon Lee[1] · Choong Ho Lee[1**]

[1]Hanbat National University · [2]Jeonju National University of Education

E-mail : mentor1023@daum.net / chlee@hanbat.ac.kr

요 약

본 연구는 한국영화진흥원의 박스오피스의 상위권 영화 상세정보 및 네이버의 영화 평점 데이터를 수집하여, 영화 흥행에 영향을 미치는 중요한 요인들을 영화 관람객 및 평점을 기준으로 분석하고자 한다.

ABSTRACT

The study aims to collect detailed movie information from box office of the Korea Film Council and data on Naver's movie ratings to analyze important factors affecting the movie's popularity based on movie audiences and ratings.

키워드

Bigdata, Analysis, Data Analysis, Bigdata Analysis, Movie's Popularity

Ⅰ. 서 론

한국 영화 흥행에 영향을 미치는 요인들이 무엇인지 도출하기 위하여 한국영화진흥원의 박스오피스 상위권 영화 상세정보 및 네이버의 영화 평점 데이터를 수집하여 영화 관람객 및 평점을 기준으로 연구하고자 한다.

Ⅱ. 분석 절차

본 연구의 분석을 위하여 매년 상위 10위권의 영화정보를 제공하고 있는 영화진흥위원회의 일별 박스오피스 데이터 및 상위 10위권에 대한 개봉작을 기준으로 네이버에서 제공하는 네티즌, 관람객 및 평론가에 의한 영화평점 데이터를 수집하여, 영화명, 제작사, 주연, 조연 등 영화 관련 요인들 간의 연관규칙을 찾아내는 연관성 분석 방법을 사용하였으며, R 프로그래밍 언어 및 RSudio 툴을 활용하여 분석을 수행하였다.

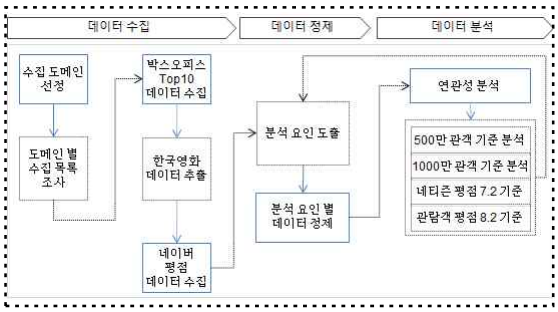

그림 1. 데이터 분석 흐름도

Ⅲ. 분석 방법

분석 방법은 Open API 및 웹크롤링 기술을 적용하여 원천 데이터를 수집하였으며, 영화별 요인 분석을 위하여 필요한 데이터를 추출 및 정제하여 연관성 분석을 수행하였다.

* speaker

** corresponding author

본 연구를 위해 수집할 원천 데이터 항목은 표1과 같다.

표 1. 데이터 수집 항목

구분	수집대상	수집 방법	수집기간
영화진흥위원회 박스오피스	일별 박스오피스 (매년 10위권 영화정보 제공)	API	2003.11.11. ~ 2018.10.31
	영화 상세정보		
네이버	네티즌 영화평점	Web Crawling	영화진흥위원회 박스오피스 영화 중 매년 10위권 내 개봉작
	관람객 영화평점		
	평론가 영화평점		

영화진흥위원회 박스오피스 원천데이터 수집은 Open API 데이터 제공 방식을 활용하여 R 프로그래밍으로 2018년 10월을 기준으로 과거 15년간 데이터를 그림2-1 및 그림2-2와 같이 수집 하였다.

```
# 영화 API 정보 저장
movie_list <- list()

for(i in 1:(as.numeric(end_date - start_date)+1)) {
  parameter_date <- str_replace_all(as.character(start_date+i-1), "-", "")
  url <- paste0('http://www.kobis.or.kr/kobisopenapi/webservice/rest/boxoffice/searchDailyBoxOfficeList.json?key=',
  print(url)
  movie_list[i] <- fromJSON(file = url)
 }
 return(movie_list)
}
#=============================================================================
## 오픈API를 이용한 데이터 수집 실행 : 기간을 정하여 데이터 수집
movie_list <- scraping_movie_info('2003-11-11', '2018-10-31')

## list 확인 - json 형태
listviewer::jsonedit(movie_list, mode="view")

#movie_list(영화진흥원 boxoffice 전체 data)는 따로 .RData 오브젝트로 저장 하지 못했음

## 리스트 영화정보를 데이터프레임으로 변환
movie_df <- tibble::tibble()
?tibble::tibble
tmp_df
```

그림 2-1. 영화진흥위원회 데이터 수집(Open API)

movieNm.x	novieNmEnovieNmO	showTm.x	prdtYear	openDt.x	rdtStatNr	typeNm	irectorNm	actorNm
내인 : 아내의 남자		71	2018	20180621	개봉	장편	임현태	
래반의 장미	Too Hot to	98	2018	20181018	개봉	장편	박진영	김인권
남자와 여자	Her husbar	69	2018	20180531	개봉	장편	주목	이상현
라한 장로		85	2017	20180709	개봉	장편	이진영	
동욱: 관음증-감독판		63	2018	20180529	개봉	장편	이민	
신비아파트: 금빛 도깨비와 비밀의 동	The Haunti	67	2018	20180725	개봉	장편	김병갑	조현정
마녀	The Witch	125	2018	20180627	개봉	장편	박혼정	김다미
안나 카레니나	ANNA KAF AHHa Kape	132	2018	20180629	개봉	장편	신예지	에카테리
가슴큰누나들		81	2018	20180612	개봉	장편	김경택	

그림 2-2. 영화진흥위원회 수집 데이터

네이버 영화평점 원천데이터 수집은 영화진흥위원회 박스오피스에서 수집한 매년 상위 10위권에 대한 영화를 기준으로, Web Crawling 방식의 R 프로그래밍을 통하여 그림3-1 및 그림3-2와 같이 수집 하였다.

```
naver_crawling_url = "https://search.naver.com/search.naver"

## 크롤링 data dataframe으로
movie_naver_crawling_info <- tibble::tibble()
#rm(movie_naver_crawling_info)

for(k in 1:(movieDfCnt)) {
 #영화코드
 movieCdValue = movie_df_detail_movieInfo_open$movieCd[k]
 #영화제목+감독이름 병합
 movieNmTxt1 <- movie_df_detail_movieInfo_open$movieNm[k]
 movieNmTxt2 <- movie_df_detail_movieInfo_open$directorNm[k]
 searchQueryTxt = paste(movieNmTxt1, movieNmTxt2, sep="+")
 print(paste(searchQueryTxt, k))
 #print(movieNmTxt)
 #query = URLencode(iconv(movieNmTxt, to="UTF-8"))
 movieTXT = if(Encoding(searchQueryTxt) == "unknown"){iconv(sear
 query = URLencode(movieTXT)
 query = str_c("?query=", query)
 temp_crawling = read_html(str_c(naver_crawling_url, query))
```

그림 3-1. 네이버 데이터 수집(Web Crawling)

movieNm	directorNm	movieGen	showTm	openDt	movieCtr	movieAdt	galPoint	galCnt
맨발의 청춘	김기덕	멜로/로맨스, 드라마	116분	1964. 개봉	한국	15세 관람	7.7	3
순간 그리고 순간: 디지털	렌역 라타나	드라마, 멜로/로맨스	108분	2007.01.04. 개봉	태국 외	15세 관람가		
마부	강대진	드라마	105분	1961. 개봉	한국		9.24	4
김약국의 딸들	유현목	1963. 개봉	한국	97분				
삼촉수	신상옥	드라마	110분	1961. 개봉	한국			
불간 마후라	신상옥	전쟁, 액션	100분	1964. 개봉	한국			
대석굴암	홍성기	드라마	한국	1965.01.01. 개봉				
자유만세	최인규	멜로/로맨스, 드라마	60분	1946.10.21. 개봉	한국	청소년 관람불가		
돈	김소동	드라마	125분	1958.03.09. 개봉	한국			
피맺힌	박남옥	멜로/로맨스, 드라마	90분	1955.12.10. 개봉	한국			
역튀	이강천	전쟁	84분	1956.04.04. 개봉	한국			
청춘 쌍곡선	한형모	코미디	100분	1956.04.10. 개봉	한국			
자유부인	한형모	드라마	한국	1956.06.09. 개봉	12세 관람			
시집 가는 날	이병일	가족, 드라마	78분	1956.11.27. 개봉	한국			

그림 3-2. 네이버 수집 데이터

수집한 영화진흥위원회 박스오피스 및 네이버 평점 원천 데이터에서 영화 별 배우·제작사·감독·장르·관람등급 등 요인 분석을 위한 데이터 정제를 그림4-1~그림4-6과 같이 R 프로그래밍을 수행 하였다.

```
#영화 별 배우 정보 data.frame
movie_df_detail_actors <- tibble::tibble()
## movieInfo$movieCd movie_df_detail_movieinfo

for(i in seq_along(movie_list_detail)) {
  tmp_df <- map_df(movie_list_detail[[i]]$movieInfo$actors, bind_rows)
  tmp_df <- tmp_df %>% mutate(movieCd = movie_list_detail[[i]]$movieInfo$movieCd)
  movie_df_detail_actors <- bind_rows(movie_df_detail_actors, tmp_df)
}
```

그림 4-1. 데이터 추출·정제(영화별 배우)

```
#영화 별 제작사 정보 data.frame
movie_df_detail_companys <- tibble::tibble()
## movieInfo$movieCd movie_df_detail_movieinfo
for(i in seq_along(movie_list_detail)) {
  tmp_df <- map_df(movie_list_detail[[i]]$movieInfo$companys, bind_rows)
  tmp_df <- tmp_df %>% mutate(movieCd = movie_list_detail[[i]]$movieInfo$movieCd
  movie_df_detail_companys <- bind_rows(movie_df_detail_companys, tmp_df)
}
```

그림 4-2. 데이터 추출·정제(영화별 제작사)

```
#영화 별 감독 정보 data.frame
movie_df_detail_directors <- tibble::tibble()
## movieInfo$movieCd movie_df_detail_movieinfo
for(i in seq_along(movie_list_detail)) {
  tmp_df <- map_df(movie_list_detail[[i]]$movieInfo$directors, bind_rows)
  tmp_df <- tmp_df %>% mutate(movieCd = movie_list_detail[[i]]$movieInfo$movieCd
  movie_df_detail_directors <- bind_rows(movie_df_detail_directors, tmp_df)
}
```

그림 4-3. 데이터 추출·정제(영화별 감독)

```
#영화 별 장르 정보 data.frame
movie_df_detail_genres <- tibble::tibble()
## movieInfo$movieCd movie_df_detail_movieinfo
for(i in seq_along(movie_list_detail)) {
  tmp_df <- map_df(movie_list_detail[[i]]$movieInfo$genres, bind_rows)
  tmp_df <- tmp_df %>% mutate(movieCd = movie_list_detail[[i]]$movieInfo$movieCd
  movie_df_detail_genres <- bind_rows(movie_df_detail_genres, tmp_df)
}
```

그림 4-4. 데이터 추출·정제(영화별 장르)

```
#영화 별 관람등급 정보 data.frame 하나만
movie_df_detail_audits <- tibble::tibble()
## movieInfo$movieCd movie_df_detail_movieinfo
for(i in seq_along(movie_list_detail)) {
 tmp_df <- map_df(movie_list_detail[[i]]$movieInfo$audits[1], bind_rows)
 tmp_df <- tmp_df %>% mutate(movieCd = movie_list_detail[[i]]$movieInfo$movieCd)
 movie_df_detail_audits <- bind_rows(movie_df_detail_audits, tmp_df)
}
```

그림 4-5. 데이터 추출·정제(영화별 관람등급)

Ⅳ. 분석 결과

데이터 분석은 영화 관객수 500만명과 1000만명 이상으로 구분하여, 지지도(support)의 변화에 따른 연관성(apriori) 분석을 그림5 및 그림7과 같이 R 프로그램을 수행하여 그림6 및 그림8과 같은 결과를 얻었다

```
movie_rule<-apriori(movie.trans_500,
                    parameter = list(support=0.001,confidence = 1,minlen = 2, maxlen=4),
                    appearance = list(rhs=c("관객_500만이상"),
                                      default="lhs"))
```

그림 5. 관람객 500만명 기준 흥행 요소별 연관성 분석(support=0.001기준)

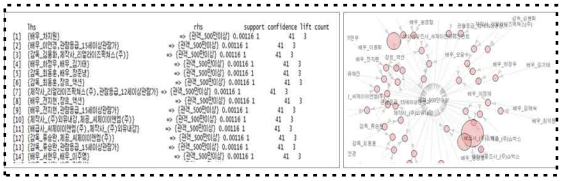

그림 6. 관람객 500만명 기준 흥행 요소별 연관성 분석 결과(support=0.001기준)

```
movie_rule<-apriori(movie.trans_1000,
                    parameter = list(support=0.00078,confidence = 1,minlen = 2, maxlen=4),
                    appearance = list(rhs=c("관객_1000만이상"),
                                      default="lhs"))
```

그림 7. 관람객 1,000만명 기준 흥행 요소별 연관성 분석(support=0.00078기준)

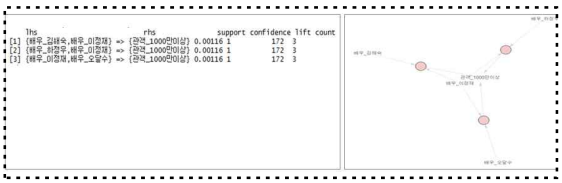

그림 8. 관람객 1,000만명 기준 흥행 요소별 연관성 분석 결과(support=0.00078기준)

Ⅴ. 결 론

영화 관람객 500만명을 기준으로 흥행 요소별 연관성 분석결과 가장 중요한 요소는 배우가 가장 중요한 요소로 분석되었으며, 다음으로 배급사로 많이 상영이 되어야 관객이 많이 볼 수 있다는 기본적인 조건을 형성하는 요소로 분석 되었다.

500만명을 기준으로 가장 영향도 있는 배우는 송강호, 가장 많은 영화를 배급한 배급사로는 CJ Entertainment는 20편, 쇼박스는 12편으로 전체 63편 중 50% 이상을 차지하는 등 대형 배급사의 영향력이 거대함을 보여주고 있다.

영화 관람객 1,000만명을 기준으로 흥행 요소별 연관성 분석결과 가장 중요한 요소는 배우가 가장 중요한 요소로 분석되었으며, 다음으로 감독으로 관람객이 증가 할 수록 영화 감독의 영향을 크게 받는 것으로 분석 되었다.

1,000만명 이상의 관람객을 유치한 영화의 경우 영화 감독이 많은 영화를 제작하였으며, 같은 배우와 제작을한 동일한 패턴을 보이고 있는 연관성을 나타내고 있다.

네티즌 평점 및 실제 관람객의 평점 평균을 기준으로한 흥행 요소별 연관성 분석 결과 실제 관람을 하지 않은 데이터가 포함된 네티즌 평점의 경우 평점에 영향을 주는 요소가 다양하게 나타났으며, 네티즌들의 감성적 영향을 많이 받는 등 주관적인 평점이 반영되는데 반하여, 실제 영화를 관람한 관람객 기준으로 분석을 한 결과 주관적인 평가보다는 영화의 질적인 객관적 평가가 진행된 것으로 분석되었으며, 상업영화보다 다큐멘터리 및 가족영화 등이 상대적으로 흥행에 높은 연관성을 나타낸 것으로 분석되었다.

References

[1] S. Y. Kim, S. H. Im and Y. S. Jung, “A Comparison Study of the Determinants of Performance of Motion Pictures : Art Film vs. Commercial Film,” *The Journal of the Korea Contents Society*, Vol. 10, No. 2, pp. 381-393, Oct. 2010.

[2] Y. H. Kim, “A study on the service quality and customer loyalty in regional cultural festival,” *The Journal of the Korean data & Information Science Society*, pp. 437-446, 2010.

[3] S. J. Kwon, “Factors influencing Cinema Success: using News and Online Rates,” in *Proceeding of the 38th Annual International Symposium on Computer Architecture*, Korea, pp. 35-55, 2014.

[4] S. H. Phak, H. J. Song, “The Determinants of Motion Picture Box Office Performance: Evidence from Korean Movies Released in

2011," Chonbuk National University, Korea, pp. 45-79, 2012.

[5] E. H. Choi, T. N. Pyo, Y. J. Park, J. G. Yong, "Integrated Study in Critical Components that Lead Korean Movies to Success," Journal of the Korean Data Analysis Society, Korea, pp. 2773-2784, 2009.

[6] B. S. Kim, "Comparison of Factors Predicting Theatrical Movie Success : Focusing on the Classification by the Release Type and the Length of Run," Korean Journal of Journalism & Communication Studies, Korea, pp. 257-287, 2009.

[7] J. W. Kim, "A study on big-data"s effect for predicting financial success of a film," korea entertainment industry society, Korea, pp. 78-81, 2014.

[8] O. J. Lee, S. B. Park, D. U. Chung, E. S. You, "Movie Box-office Analysis using Social Big Data," The Korea Contents Society, Korea, pp. 527-538, 2014.

[9] Y. H. Kim, J. H. Hong, "A Study for the Drivers of Movie Box-office Performance," The Korean Journal of Applied Statistics, Korea, pp. 441-452, 2013.