J. Soc. Korea Ind. Syst. Eng Vol. 39, No. 3 : 90-99, September 2016
http://dx.doi.org/10.11627/jkise.2016.39.3.090
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Clustering Analysis of Films on Box Office Performance : Based on Web Crawling

Jai-Ill Lee · Young-Ho Chun · Chunghun Ha†

School of Information & Computer Engineering, Hongik University

영화 흥행과 관련된 영화별 특성에 대한 군집분석 : 웹 크롤링 활용

이재일 · 전영호 · 하정훈†

홍익대학교 정보컴퓨터공학부 산업공학전공

Forecasting of box office performance after a film release is very important, from the viewpoint of increase profitability by reducing the production cost and the marketing cost. Analysis of psychological factors such as word-of-mouth and expert assessment is essential, but hard to perform due to the difficulties of data collection. Information technology such as web crawling and text mining can help to overcome this situation. For effective text mining, categorization of objects is required. In this perspective, the objective of this study is to provide a framework for classifying films according to their characteristics. Data including psychological factors are collected from Web sites using the web crawling. A clustering analysis is conducted to classify films and a series of one-way ANOVA analysis are conducted to statistically verify the differences of characteristics among groups. The result of the cluster analysis based on the review and revenues shows that the films can be categorized into four distinct groups and the differences of characteristics are statistically significant. The first group is high sales of the box office and the number of clicks on reviews is higher than other groups. The characteristic of the second group is similar with the 1st group, while the length of review is longer and the box office sales are not good. The third group's audiences prefer to documentaries and animations and the number of comments and interests are significantly lower than other groups. The last group prefer to criminal, thriller and suspense genre. Correspondence analysis is also conducted to match the groups and intrinsic characteristics of films such as genre, movie rating and nation.

Keywords : Characteristic of Films, Box Office Performance, Clustering Analysis, Web Crawling

1. 서 론

수요예측은 미래를 대비하고 현재를 정비할 수 있는 매우 중요한 기업 활동 중 하나이다. 영화산업에서도 정확한 영화 흥행의 예측은 제작비와 마케팅비의 절감, 그리고 이를 통한 수익성 증대 관점에서 매우 중요하다[5]. 이에 따라 국내·외에서 영화 흥행의 요인을 분석하고 이를 예측에 활용하고자 하는 시도가 지속적으로 이루어져 왔다[9, 10, 18, 23].

영화 흥행에 대한 예측은 우선 흥행에 영향을 미치는 변인을 도출하고, 해당하는 자료를 수집한 후, 회귀분석을 통하여 이에 따른 흥행의 정도를 추정하고 검증하는 방법을 주로 사용하였다[13, 14, 15, 18, 22].

영화의 흥행에 영향을 미치는 요인은 매우 다양한 데, 기존 연구에서는 주요 변인으로서 영화관람등급, 상영시

Received 29 June 2016; Finally Revised 24 August 2016;
Accepted 1 September 2016
† Corresponding Author : chunghun.ha@hongik.ac.kr

간, 장르, 속편 여부, 개봉 점유율, 제작비 규모, 주연배우 및 감독 파워, 리뷰, 제작국가, 구전 등을 고려하였다[1, 5]. 이러한 요인들은 그 특성에 따라 경제적 변인과 심리적 변인으로 구분할 수 있다[5]. 경제적 변인은 크게 장르와 상영시간과 같이 영화의 내재적 특성요소와 상영관 수 등 마케팅적 요소로 구성되며, 심리적 변인은 영화에 대한 관심, 의견, 전문가 평가 등으로 구성된다.

기존의 연구는 대부분 자료수집이 용이한 경제적 변인의 활용에 집중되어 왔다. 그러나 이러한 경제적 변인만으로는 실질 수요, 즉, 잠재수요를 정확히 추정하는 데에 한계가 있다[5]. 영화는 제품과 서비스가 결합된 특성을 가지므로 영화의 흥행은 심리적 변인의 영향력이 크다. 특히, 영화 흥행은 개봉 초기에는 영화평론가의 평가 그리고 후기에는 구전(Word of Mouth : WOM)의 영향을 많이 받는 특성이 있다[3].

영화는 가격이 고정되어 수익은 수요에 전적으로 의존하므로 영화 개봉 전에 수요예측을 통하여 적절한 전략을 수립하여야 한다. 그러나, 영화는 개봉 이후 그 가치를 확인할 수 있는 경험재의 특성을 가지므로 사전 예측의 정확도가 낮다. 영화 상영을 통한 이익을 극대화하기 위해서는 개봉 초기에 흥행에 영향을 미치는 다양한 심리적 변인을 분석하여 예측 수요를 갱신하고, 그 수준에 따라 상영관수의 조정이나 광고의 투입 등 마케팅 전략을 수정하고 보완하여야 한다. 그러나, 심리적 변인의 경우 고비용과 긴 조사시간을 요하는 설문조사와 같은 방법을 제외하고는 직접적인 측정이 어려울 뿐만 아니라, 자료가 수집되더라도 개인의 심리적인 판단에 근거하므로 높은 신뢰성을 확보하기 어렵다. 특히, 구전은 관람객이 영화를 직접 관람한 이 후에만 자료를 획득할 수 있으며, 시간에 따라 가치가 변화하는 동적 특성을 가지므로 주기적인 갱신이 필요하다. 이러한 상황에서 영화의 개봉 후 가급적 빠른 시기에 측정이 어려운 심리적 변인에 대한 자료의 수집과 분석, 그리고 이를 반영한 예측 수요의 갱신은 필수적인 활동이다. 최근 주목받고 있는 웹 크롤링(web crawling)과 텍스트 마이닝(text mining)은 이를 저비용으로 달성할 수 있는 유용한 도구이다. 웹 크롤링은 웹 크롤러(web crawler)라는 컴퓨터 소프트웨어를 이용하여 자동화된 방법으로 웹사이트상의 필요한 정보를 검색하고 추출하는 방법이고, 텍스트 마이닝은 자연언어처리(Natural Language Processing) 기술에 기반하여 비정형 텍스트 자료로부터 가치 있는 정보를 추출하는 기술이다.

최근 텍스트 마이닝(Text mining)기법의 발전과 다양한 프로그래밍 언어로 작성된 라이브러리가 풍부하게 제공됨에 따라 온라인상에 산재하여 있는 자료를 수집 및 분석하기 용이해 졌다. 특히 텍스트 마이닝이 기존의 단순 문서 분류에서 벗어나 오피니언 마이닝(opinion mining) 및 감성분석(sentiment analysis)으로 발전하고 있어 정성적인 데이터를 분석함에 있어 다른 분석기법에 비해 보다 유용해지고 있다고 할 수 있다. 이는 2003년도에 발표된 토픽 모델링 기법 중 하나인 LDA(Latent Dirichlet Allocation)알고리즘 이후에 가속화되고 있는데 이러한 추세를 반영하듯 2011년도에는 LDA를 변형한 HDP(Hierarchical Dirichlet Process)알고리즘이 발표되기도 하였다.

텍스트 마이닝의 성능은 알고리즘뿐만 아니라 텍스트 자료의 범위 선정 및 정제 과정에도 영향을 받는다. 이는 텍스트 마이닝 과정 중 필수과정인 단어 사전 만들기에서 성능이 크게 좌우되기 때문이다. 특히, 본 연구에서 진행한 영화를 예를 들면 액션 장르에 속한 리뷰와 다큐멘터리 장르에 속한 리뷰의 문장은 단어선정 및 감성 어휘의 사용부터 달라지게 된다. 이는 텍스트 마이닝을 수행함에 있어 영화라는 하나의 큰 카테고리로 볼 것이 아니라 비슷한 영화들 또는 비슷한 장르로 그룹화 하여 텍스트 마이닝을 수행하여야 효과적인 결과를 도출할 수 있음을 의미한다. 하지만 현재 영화를 그룹화 하는데 있어 가장 많이 쓰이는 장르의 경우 국내 유명 포털 사이트 기준으로 18가지가 등록되어 있어 그대로 적용하기에는 지나치게 세분화되어 있다.

본 연구의 궁극적인 목적은 영화 흥행의 수요 예측의 정확도 향상을 위한 텍스트 마이닝을 활용한 신뢰도 높은 심리적 변인의 측정 및 분석이다. 그러나, 앞에서 언급한 바와 같이 텍스트 마이닝의 성능은 자료의 범위 선정과 정제과정에 영향을 받는다. 따라서 본 논문에서는 텍스트 마이닝을 수행하기에 앞서 진행하는 선행연구로서 영화의 흥행을 기준으로 유사한 장르 또는 영화들을 묶어 유사그룹을 생성하는 기준을 제시하고자 한다. 이를 위하여 자료수집이 용이한 경제적 변인 외에 리뷰 글자 수, 영화평점, 리뷰 클릭 수와 같은 정량적인 심리적 변인도 고려하였다. 이를 위해 웹 크롤링을 활용하여 자료를 수집하고 이를 정제하여 사용하였다. 영화를 분류하기 한 방법으로는 K-means를 이용한 군집분석(Clustering Analysis)을 실시하였고, 그룹별 차이를 검정하기 위하여 일원배치 분산분석(One-way ANOVA Analysis)을 실시하였다. 또한 각 그룹별 영화의 내재적 특성을 분석하기 위하여 상응분석을 실시하였다.

본 논문의 구성은 다음과 같다. 우선 제 2장에서는 본 연구에서 사용한 영화의 흥행에 영향을 미치는 경제적·심리적 변인을 정의하고 이에 관련한 선행연구를 정리한다. 제 3장에서는 본 논문에서 영화 흥행의 변인 분석을 위해 사용한 분석절차와 관련 자료의 수집방법에 대한 설명을 한다. 제 4장에서는 수집된 자료를 바탕으로 군집분석을 실시하여 그룹을 분리하고, 각 그룹별 차이와 특성에 대한 분석을 위하여 일원배치 분산분석과 상응분석을 수행한다. 마지막으로 제 5장에서는 연구결과를 정리하고 향후 연구방향에 대하여 논한다.

2. 영화 흥행과 관련된 변인과 관련 선행연구

2.1 영화관람등급

영화관람등급은 영화의 내용에 따라 관람할 수 있는 연령대를 구분하는 기준으로 대부분의 국가는 이러한 관람등급에 대한 규정을 운영하고 있다. 우리나라는 '영상물등급위원회'라는 독립적 기관에서 이를 관리 및 운영하고 있다. 국내의 경우, 관람등급은 총 4가지로서 '전체관람가', '12세 이상 관람가', '15세 이상 관람가', '청소년 관람불가'로 구분된다. 영화 흥행의 관점에서 이러한 관람등급은 해당 영화를 관람할 수 있는 관객의 모집단 규모, 즉, 잠재시장의 규모를 결정하므로 영화 흥행에 크게 영향을 미친다[13, 14, 19, 20, 22]. 이는 Chang and Ki [4]의 미국 실증연구에서 확인할 수 있는데, 그들의 연구결과에 따르면, 성인만 관람이 가능한 R(Restricted)등급은 흥행에 부정적인 영향을 미치는 반면 부모 동반 시 관람이 가능한 PG(Parental Guidance Suggested)등급은 흥행에 긍정적인 영향을 미친다.

2.2 상영시간

영화의 상영시간은 평균적으로 1시간 40분 내외지만, 장르에 따라 편차가 큰 편이다. 영화 흥행에 대한 상영시간의 영향력을 실증적으로 분석한 Lee et al.[12]의 연구에 따르면, 상영시간이 길수록 영화 흥행의 정도가 큰 것으로 나타났다.

2.3 장르

영화는 내용적 속성에 따라 여러 가지 장르로 구분할 수 있다. 장르 또한 관람등급과 같이 잠재수요의 규모에 영향을 미친다. Litman and Kohl[13]의 실증연구 결과에 따르면, 미국에서는 코미디, SF, 공포 장르가 흥행에 긍정적인 미치는 것으로 나타났으며, 드라마 장르의 경우에는 부정적인 영향을 미치는 것으로 나타났다. 반면, 한국의 경우, Park and Jung[18]의 2006년부터 2008년간의 영화를 대상으로 한 실증연구에 따르면, 코미디가 흥행에 유의미한 영향을 미치는 장르로 분석되었다.

2.4 리뷰

영화의 경우 직접 관람하기 이전에는 그 속성을 평가하기 어렵다. 따라서 영화 흥행은 개인들 사이에서 특정 상품에 대한 정보를 주고받는 구전(Word of Mouth, WOM)에 영향을 많이 받는다[3]. 영화는 다른 산업과 달리 개봉이전에 광고와 홍보 활동 및 전문가 평가가 활발히 이루어지며, 이중 영화평론가의 평가가 영화 관람 의사결정에 영향을 준다고 알려져 있다[1, 2, 4]. 최근에는 전문가의 평가와 더불어 SNS 활동이 활발해짐에 따라 온라인 평가에 대한 연구도 활발히 진행되고 있다[21]. 그 중에서 Liu[16]는 온라인 영화평의 총량(Volumn)과 유발성(Valence)이 영화 흥행성과에 미치는 메커니즘에 관해 연구를 진행하였다. 그는 온라인 리뷰가 많아질수록 인지도의 정보효과(Informative Effect of Awareness)가 커지며, 긍정적인 평가가 많을수록 설득적 효과(Persuasive Effect on Attitude)가 증가한다고 하였다.

2.5 제작국가

제작 국가 역시 영화 흥행에 영향을 주는 변인으로 작용할 수 있다. 특정한 문화권에서 제작된 영화는 문화적 배경이 상이한 문화권에 수출될 경우 문화적 차이에 의해 매력도가 감소할 수 있다. 이러한 것을 문화적 할인(cultural discount)이라고 한다. Lee[11]는 그의 연구에서 문화적 할인율은 문화에 따라 모든 영화에 동일하게 적용되는 것은 아니고 장르와 복합적인 결합을 통하여 발현되는 것이라고 주장하였다.

3. 분석절차와 자료수집방법

3.1 분석절차

온라인에서 웹으로 서비스 되고 있는 자료를 수집하는 방법 중 하나인 웹 크롤링을 이용하여 영화 관련 자료를 수집하여 DB화 하였다. 그리고 분석하기에 앞서 수집된 자료 중 결측치와 분석대상이 아닌 자료를 제거하여 자료정제(Data Cleaning) 작업을 진행하였다. 분석자료로는 상영기간 전체를 대상으로 하는 자료와 요일특성을 없애기 위해 개봉일로부터 7일 단위로 취합한 자료를 활용하였다. 이렇게 생성된 분석자료를 이용하여 영화의 흥행과 온라인 리뷰와의 상관을 알아보기 위해 자료 탐색 차원에서 군집분석(clustering analysis)을 실시하였으며, 선행연구의 요인별로 일원배치 분산분석(one-way ANOVA analysis)를 실시하여 그룹별로 어떻게 차이가 나는지 알아보고자 하였다. 이와 더불어 장르와 관람등급 그리고 제작국가와 군집분석 후 생성된 그룹에 대하여 상응분석을 실시하였다.

3.2 변수의 조작적 정의와 측정

영화 흥행의 주요 변인은 관객 수와 매출액이다. 하지만 이 변수를 그대로 사용하면 초반에 배급사의 영향력에 의해 스크린 수를 많이 확보한 영화가 흥행 영화가

되기 때문에 분석에 왜곡이 발생할 수 있다. 따라서 본 연구에서는 스크린 당 매출액으로 두 변인을 통합하였다. 상영시간은 분 단위 기준으로 사용하였으며, 장르의 경우 네이버 무비에 등록되어 있는 정보를 활용하였다. 중복 장르가 있는 경우 처음 등록된 장르를 활용하였으며, 전체 빈도가 5% 미만인 장르는 유사장르로 통합하였다. 제작 국가는 해외와 국내로만 구분하였고, 영화평점은 네이버에 등록되어 있는 평점을 그대로 활용하였다. 이외에 리뷰의 경우 단순한 문구 작성인지 또는 의견을 작성한 것인지 정량화하기 위해 가장 단순한 리뷰 글자수를 계산하였으며, 이와 더불어 등록된 리뷰의 공감, 비공감 클릭 수를 이용하여 영화 별 전체 관련리뷰 클릭수를 계산하였다. 이를 정리하면 <Table 1>과 같다.

<Table 1> Operational Definition of Variables

Variables	Operational Definition	Ref.
Sales per Screen	Movie theater sales per screen.	[2, 15]
Movie Rating	Domestic standard is used.	[4, 13]
Running Time	Running time(minutes).	[2, 12]
Genre	The genre of 18 is used on the portal site. If multiple genres are registered, the first genre is selected.	[4, 13, 19, 20]
Nation	Categories divided into ‘overseas’ and ‘domestic.’	[9, 11, 12]
Score of Review	Movie Wanted score is registered on portal site during screening.	[16, 24]
Length of Review	Length of reviews is registered on portal site during screening.	[16]
Clicks of Review	Clicks of review are registered on portal site during screening.	[6, 7, 16]

3.3 자료수집의 대상과 방법

영화와 관계된 자료 중 매출액, 관객 수, 스크린 수와 같은 집계 자료는 영화진흥위원회의 자료를 이용하였다. 그리고 온라인 리뷰 및 평점 정보는 국내에서 가장 이용자가 많은 네이버 포털 사이트를 이용하였다. 데이터 크롤러는 java를 이용하여 구현하였으며, DBMS(Data Base Management System)로는 MySQL 5.7을 활용하였다. 그리고 데이터 마이닝 분석도구로는 R을 사용하였다.

영화진흥위원회 사이트에서 제공하는 자료 중 상영일자가 2013년 1월부터 2015년 12월까지 10,990건 자료가 수집 하였으며, 이를 다시 영화 별로 매출, 관객 수, 상영 스크린 수, 상영 횟수로 합산하여 644건의 자료를 취합할 수 있었다. 포털 영화 사이트에서도 위와 같은 조건을 적용하여 1,236,208건의 자료를 수집하였으며, 이를 영화별로 영화평점, 네티즌 리뷰, 네티즌 공감 및 비공감 횟수 등으로 평균값을 구하여 698건의 자료를 취합할 수 있었다. 이중 영화진흥위원회와 포털 영화 사이트에 동시에 존재하거나, 다시 상영되거나, 또는 수집 자료 중 누락 자료가 있는 자료를 제외하고 총 509편의 영화를 최종 분석에 사용하였다.

분석 자료는 자료 수집 기간 내에 개봉 일이 포함되어 있는 영화만을 대상으로 하였으며, 과거 상영되었다가 재개봉 된 영화는 제외처리 하였고, 중도에 상영을 중지했다가 재개한 기간은 포함하지 않고 개봉일로부터 연속적으로 상영한 기간만을 대상으로 하였다.

4. 실증 분석

4.1 표본의 특성

본 연구에 사용된 영화는 총 509편으로 국내 제작이 32.6%(166편), 해외 제작이 67.4%(343편)로 나타났다. 관람등급을 보면 가장 많이 등급은 15세 관람가로서 35.2%(179편)로 나타났으며 나머지 등급은 20% 초반대의 비율을 보였다. 장르별로 보면 드라마가 23.6%(120편), 액션 17.7%(90편), 애니메이션이 17.9%(91편)로 나타났으며 코미디를 비롯한 나머지 장르는 10% 미만으로 나타났다.

수집된 자료의 영화 별 총 매출액은 평균 731,708만원으로 나타났으며, 총동원 관객 수는 평균 95만 명으로 집계되었다. 전국기준 하루 평균 상영 횟수는 996회로 나타났으며 상영기간은 평균 3주 정도인 것으로 나타났다. 온라인 리뷰 관련하여 영화 별 평점은 10점 만점에 평균 7.729점이었으며, 네티즌 리뷰로 평균 43자를 작성하는 것으로 나타났다. 또한 영화 관련 리뷰의 클릭 수는 평균 16,965건으로 나타났다. 이를 정리하면 <Table 2>과 같다.

<Table 2> Descriptive Statistics of Collected Data

Variables	Mean	Min	Max	SD
Sales(Ten thousand Korea won)	731,708	766	11,040,521	1,448,021
Audiences(Ten thousand)	95	0.100	1,419	186
Average daily number of screening	996	7	4,164	747
Score of Review	7.729	1.167	9.324	1.072
Length of Review	43	27	66	6
Clicks of Review	16,965	16	228,992	27,131
Running Time(minutes)	109	46	180	19
Running time period(week)	3	1	12	2

4.2 군집분석을 이용하여 그룹화

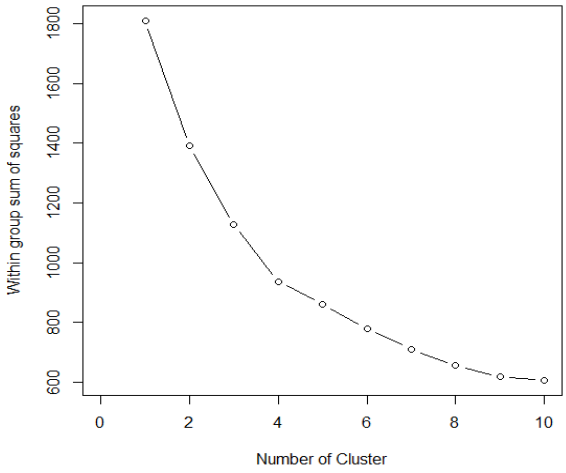
군집분석(cluster analysis)은 개인 또는 여러 개체 중에서 유사한 속성을 지닌 대상을 몇 개의 집단으로 그룹화한 다음, 각 집단의 성격을 파악함으로써 자료 전체의 구

조에 대해 이해하고자 하는 탐색적인 분석방법이다.

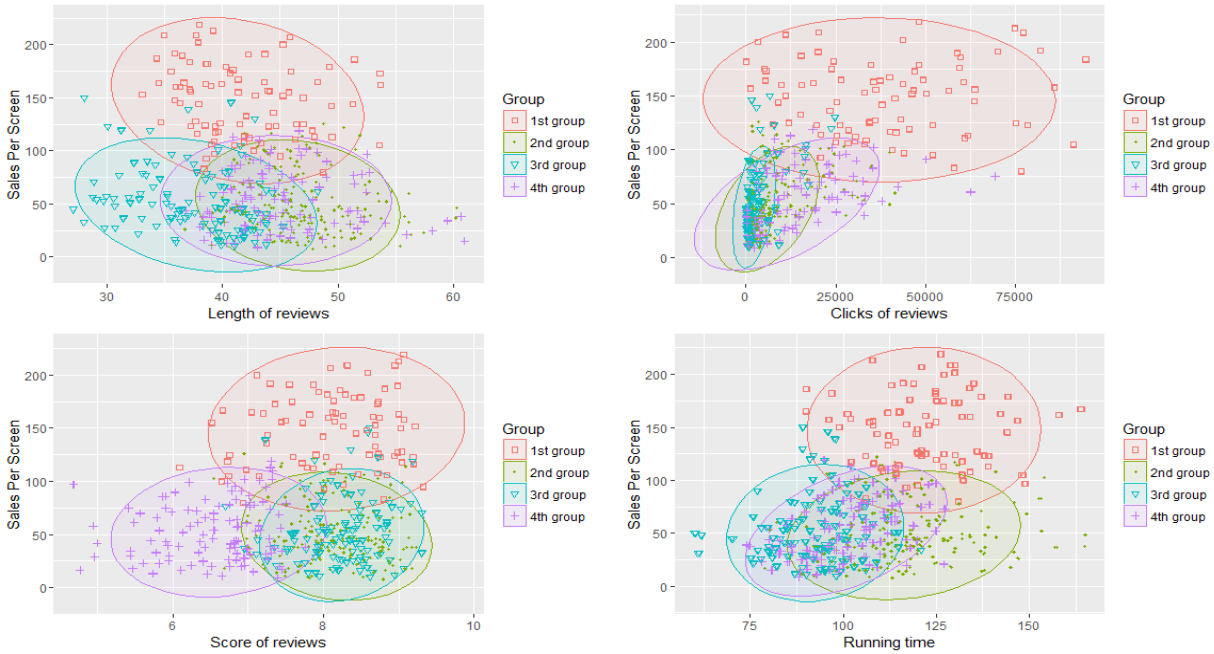
본 연구에서는 비계층적 군집분석의 대표적인 방법인 K-means 군집분석을 이용하였다. k-means 분석은 군집분석뿐만 아니라 다방면으로 쓰이고 있는 분석기법 중 하나이다. 실제 k-means분석의 경우 군집의 개수에 해당하는 k가 각 군집 내에서 중심위치에 해당하기 때문에 Moon과 Park은 TPS(Traveling Salesman Problem) 최적화 문제에서 K-means를 활용하기도 하였다[8]. 우선 군집분석을 위해 수치형 자료인 스크린 당 매출액, 상영시간, 리뷰 글자수, 영화평점, 리뷰 클릭 수를 변수로 사용하였으며, 분석 시 각 변수 별 단위의 영향을 제거하기 위하여 표준화된 Z값을 사용하였다. 그리고 표준화된 Z의 절대값이 3을 넘을 경우 이상치로 예외 처리하여 총 509건의 데이터 중 478건을 활용하였다.

군집분석에서 군집의 개수를 결정하는 것은 중요한의사결정 요소이다. 다양한 방법이 존재하지만 본 논문에서는 비계층적 방법에 적용이 가능한 스크리 그림(scree plot)을 사용하였다. <Figure 1>은 군집의 개수에 따른 군집 변동을 보여주고 있다. 4개의 군집 이후로는 기울기가 완만해지고 있음을 확인할 수 있다. 4개의 군집에 대한 전체 변동대비 군집 내 변동은 48.2%(between_SS/total_SS = 859/1781)로 나타나 군집에 대한 설명력은 높지 않으나, 군집 수 증가에 따른 개선폭이 작고, 심리적 변인인 리뷰 글자 수, 영화평점, 리뷰 클릭 수의 영향력을 고려하여 4개로 결정하였다.

4개의 군집에 대한 특징을 분석하기 위하여 상영시간, 리뷰 글자 수, 영화평점, 리뷰 클릭 수에 따른 스크린 당 매출액을 산포도로 표현하고 그룹별로 분류하면 <Figure 2>와 같다. <Figure 2>의 산포도를 통하여 다음과 같은 그룹별 특징을 도출할 수 있다. 첫 번째 그룹(1st group)은 다른 그룹에 비해 매출이 높고 리뷰 클릭 수가 많은 흥행에 성공한 그룹으로 나타났다. 두 번째 그룹(2nd group)은 다른 그룹에 비해 리뷰 글자 수가 많은 것으로 나타났다. 세 번째 그룹(3rd group)은 상영시간이 짧고, 리뷰 글자 수가 적었으며 또한 리뷰의 클릭 수도 적은 비관심 그룹으로 나타났다. 네 번째 그룹(4th group)은 다른 집단에 비해 온라인 영화 평점이 낮은 그룹으로 나타났다. 이에 대해 통계적으로도 유의한 차이가 있는지를 알아보기 위해 각 그룹별로 분산분석을 진행하였다.



<Figure 1> Scree plot of Cluster



<Figure 2> Scatter plot by groups

4.3 분산분석을 이용한 각 그룹별 특성 검증

K-means 군집분석을 통하여 분리한 총 네 개의 그룹에 대해 일원배치 분산분석(One-way ANOVA)을 이용하여 스크린 당 매출액, 상영시간, 리뷰 글자 수, 영화평점, 리뷰 클릭 수로 5개의 변수에 대해 평균의 차이가 유의한지 검정하였다. 차이가 유의할 시 어느 그룹에서 차이가 나는지를 알아보기 위해 사후검정으로 Tukey HSD Tests를 실시하였다.

4.3.1 그룹별 스크린 당 매출액 평균 비교

군집분석 결과의 1그룹은 다른 그룹은 비해 영화 흥행에 성공한 그룹으로 <Figure 2> K-means 분석을 통한 그룹별 산포도를 통해 스크린 당 매출액이 높음을 관측 할 수 있었다. one-way ANOVA 분석 결과, 1그룹은 스크린 당 매출액이 평균 148.16만 원으로 나타났으며, 2그룹은 50.88만 원, 3그룹은 53.30만 원, 4그룹은 53.24만 원으로 각각 나타났다. 또한 p-value가 0.001 미만으로 그룹 별로 스크린 당 매출액이 유의하게 차이가 나는 그룹이 있는 것으로 나타났다(F = 239.276, p-value < 0.001). 구체적으로 어떠한 그룹들에서 차이가 있는 지를 알아보기 위해 Tukey HSD Tests를 실시한 결과, 유의수준 0.05에서 1그룹의 매출액이 다른 2, 3, 4그룹에 비해 유의하게 높게 나타났다. 즉, 군집분석 결과의 1그룹은 다른 그룹에 비해 스크린 당 매출액이 높은 그룹으로 흥행에 성공한 그룹임을 확인 할 수 있었다. 이를 정리하면 <Table 3>과 같다.

<Table 3> Result of Tukey HSD Tests on Movie theater Sales per Screen

Dependent variable	Group				*F*	*p*-value
	1st	2st	3st	4st		
Movie theater sales per screen	148.16				239.276	< 0.001
		50.88	53.30	53.24		

4.3.2 그룹별 상영시간 평균 비교

군집분석 결과의 3그룹은 다른 그룹에 비해 상영시간, 리뷰 글자 수, 리뷰 클릭수가 낮아 대중으로부터 관심이 낮은 그룹이었다. 이중 영화 상영 시간이 다른 그룹에 비해 유의하게 낮은 지 알아보았다. one-way ANOVA 분석 결과 1그룹의 상영시간은 평균 122.23분, 2그룹은 117.17분, 3그룹은 92.42분, 4그룹은 100.96분으로 나타났으며, 각 그룹별로 상영시간이 유의하게 차이가 나는 그룹이 있는 것으로 나타났다(F = 114.946, p-value < 0.001).

사후검정으로 Tukey HSD Tests를 실시한 결과, 유의수준 0.05에서 모든 그룹마다 상영시간의 차이가 있는 것으로 나타났으며, 특히 3그룹의 경우 92.42분으로 다른 집단에 비해 낮게 나타나는 것을 확인 할 수 있었다. 이를 정리하면 <Table 4>와 같다.

<Table 4> Result of Tukey HSD Tests on Running time

Dependent variable	Group				*F*	*p*-value
	1st	2st	3st	4st		
Movie Running time	122.23				114.946	< 0.001
		117.17				
				100.96		
			92.42			

4.3.3 그룹별 리뷰 글자 수 평균 비교

군집분석 결과 2그룹은 리뷰 글자 수가 다른 그룹에 비해 높은 그룹이었다. one-way ANOVA 분석 결과 1그룹의 경우 리뷰 글자 수가 평균 41.69자, 2그룹은 46.79자, 3그룹은 37.45자, 4그룹은 45.20자로 나타났으며 각 그룹별로 리뷰 글자 수가 유의하게 차이가 나는 그룹이 있는 것으로 나타났다(F = 101.446, p-value < 0.001).

사후검정으로 Tukey HSD Tests를 실시 결과, 유의수준 0.05에서 모든 그룹마다 리뷰 글자 수가 유의하게 차이가 있는 것으로 나타났으며, 2그룹이 경우 46.79자로 다른 집단에 비해 높게 나타나는 것을 확인 할 수 있었다. 이와 더불어 대중으로부터 관심이 적은 3그룹의 경우 리뷰 글자 수가 적은 것을 확인 할 수 있었다. 이를 정리하면 <Table 5>과 같다.

<Table 5> Result of Tukey HSD Tests on Length of Reviews

Dependent variable	Group				*F*	*p*-value
	1st	2st	3st	4st		
Length of review		46.79			101.446	< 0.001
				45.20		
	41.69					
			37.45			

4.3.4 그룹별 리뷰 점수 평균 비교

군집분석 결과 4그룹은 다른 그룹에 비해 리뷰 평점이 낮은 그룹이었다. one-way ANOVA 분석 결과 10점 만점 기준에서 1그룹의 경우 리뷰 점수의 평균 8.12점, 2그룹은 8.15점, 3그룹은 8.23점, 4그룹은 6.56점으로 나타났으며, 각 그룹별로 리뷰 점수가 유의하게 차이가 나는 그룹이 있는 것으로 나타났다(F = 191.744, p-value < 0.001).

사후검정으로 Tukey HSD Tests를 실시한 결과, 유의수준 0.05에서 4그룹이 6.56점으로 다른 1, 2, 3그룹보다 특히나 낮음이 관측되었다. 즉, 군집분석 결과의 4그룹은 다른 그룹에 비해 리뷰 평점이 낮은 그룹임을 확인할 수 있었다. 이를 정리하면 <Table 6>과 같다.

<Table 6> Result of Tukey HSD Tests on Score of Reviews

Dependent variable	Group				*F*	*p*-value
	1st	2st	3st	4st		
Score of review	8.12	8.15	8.23		191.744	< 0.001
				6.56		

<Table 7> Result of Tukey HSD Tests on Clicks of Reviews

Dependent variable	Group				*F*	*p*-value
	1st	2st	3st	4st		
Clicks of review	38,817				142.963	< 0.001
				13,472		
		7,926				
			3,528			

4.3.5 그룹별 영화 리뷰 클릭 수 평균 비교

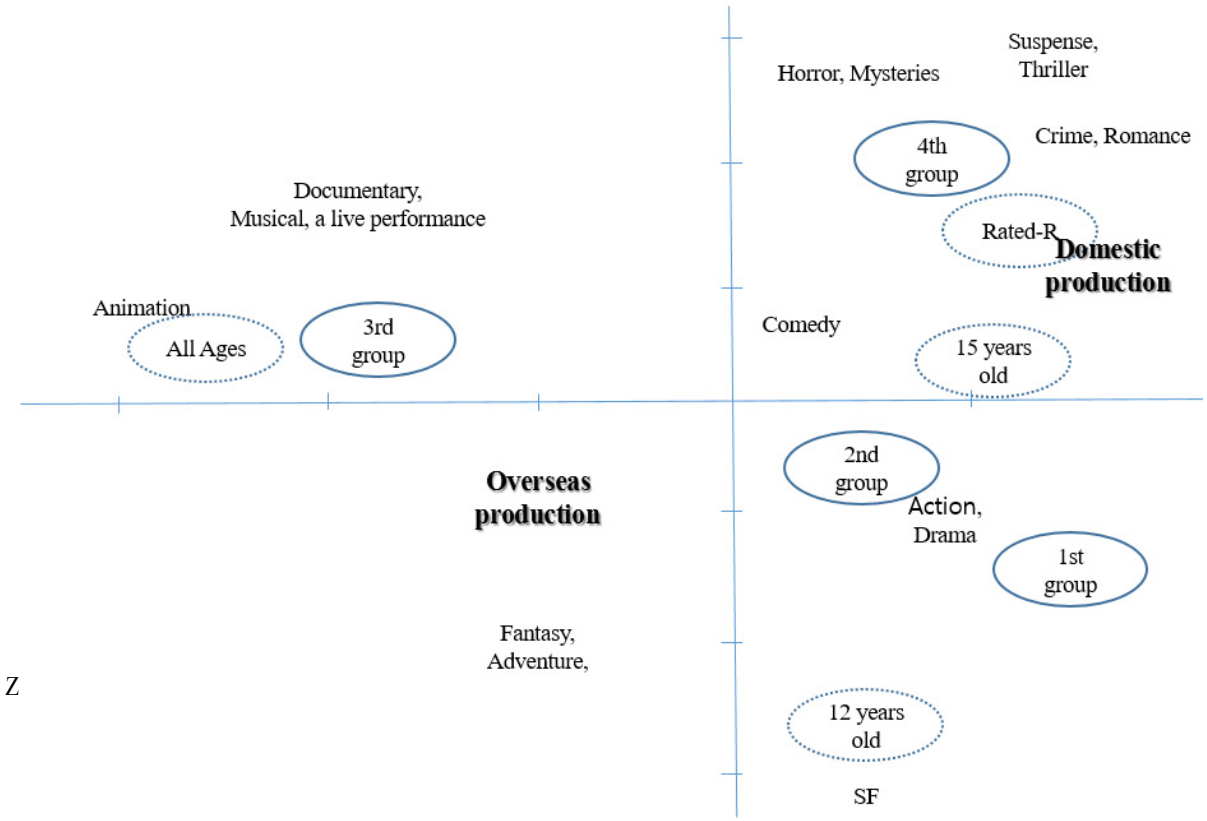
그룹별 온라인 영화 리뷰 클릭 수가 그룹별로 차이가 있는지 알아보았다. one-way ANOVA 분석 결과 1그룹의 경우 리뷰 클릭 수는 평균 38,817회, 2그룹은 7,926회, 3그룹은 3,528회, 4그룹은 13,472회로 나타났으며 각 그룹별로 온라인 영화 리뷰 클릭 수가 유의하게 차이가 나는 그룹이 있는 것으로 나타났다(F = 142.963, p-value < 0.001).

사후검정으로 Tukey HSD Tests를 실시한 결과, 유의수준 0.05에서 흥행에 성공한 1그룹이 평균 리뷰 클릭 수가 38,817회로 다른 모든 그룹에 비해 가장 유의하게 높았고, 대중으로부터 관심이 적은 3그룹이 3,528회로 가장 낮음을 확인 할 수 있었다. 의외로 리뷰 평점이 낮은 그룹은 4그룹이 13,472회로 흥행 그룹 다음으로 평균 리뷰 클릭 수가 많은 것으로 나타났다. 이를 정리하면 <Table 7>과 같다.

4.4 상응분석을 이용한 그룹별 포지셔닝 분석

상응분석은 서로 연관성이 높고, 적어도 2개 이상의 범주형 값들을 가지고 있는 명목변수들 간의 빈도분할표 자료를 이용하여 변수 값들 간의 구체적인 연관관계를 일종의 시각적인 포지셔닝 맵으로 나타낼 수 있는 다변량 통계기법이다. 본 논문에서는 제 4.2절과 제 4.3절을 통하여 도출된 그룹별 특성과 영화의 장르, 관람등급, 제작국가와 같은 영화의 내재적 특성 간의 관계를 분석하기 위하여 상응분석을 실시하였다. 상응분석을 통한 그룹별 포지셔닝 맵은 <Figure 3>에서 확인할 수 있다.

<Figure 3>을 보면, 흥행 그룹에 속하는 1그룹 및 리뷰 글자 수가 많은 2그룹의 경우, 액션, 드라마, 판타지,



<Figure 3> Symmetric Correspondence Analysis Map

<Table 8> Characteristic Table with Groups

Group	Rating	Genre	Nation	Description
1st Group	12 years old, 15 years old	Action, Fantasy, Adventure, Comedy	Overseas	This group is in high sales of box office. Especially the number of clicks on review is higher than other groups.
2nd Group				This group is similar to the 1st group in running time, genre and score of review. Only difference is that the length of review is longer than the 1st group even though box office sales are low.
3rd Group	All ages	Documentary, Musical, A live performance, Animation		This group's audience prefers to documentary and animation genre and the number of comments and interests is significantly lower than other groups.
4th Group	Rated-R	Horror, Mysteries, Suspense, Thriller, Crime, Romance	Domestic	Audience of the 4th group prefer to criminal, thriller and suspense genre which deal with provocative topics. Even though the 4th group is not the highest in box office, it is verified that this group is receiving more attention than other groups except the 1st group.

모험, SF와 같은 장르에 속했으며, 관람등급도 12세 관람 등급 또는 15세 관람 등급의 영화가 속하는 것으로 나타났다. 즉, 흥행그룹 및 리뷰 글자 수가 많은 그룹은 다른 장르나 또는 관람등급에 있어 대중의 접근이 쉬운 그룹으로 나타났다. 반면 평점이 낮았던 4그룹의 경우 공포, 미스터리, 서스펜스, 스릴러, 멜로, 로맨스. 범죄와 같은 장르의 영화가 속했으며, 관람등급 면에 있어서도 청소년 관람불가 등급이 속하는 것으로 나타났다. 또한, 이러한 장르 및 관람등급의 영화는 해외 영화보다는 국내 영화가 연관이 많은 것으로 나타났다. 마지막으로 대중으로부터 관심이 적었던 3그룹의 경우 다큐멘터리, 뮤지컬, 공연실황, 애니메이션으로 나타났으며, 관람등급은 전체 관람 등급에 속하는 것으로 나타났다.

4.5 분석 결과

군집분석과 상응분석을 통한 결과는 다음과 같이 정리할 수 있다.

첫째, 1그룹에 속했던 흥행그룹의 경우 장르 측면에서 보면 액션, 드라마, 판타지, 모험, SF, 코미디가 상관이 높은 것으로 나타났으며, 이 중에서도 액션과 드라마가 특히 높은 것을 확인 할 수 있었다. 기존 미국의 실증 연구는 코미디, SF, 공포 장르가 흥행성적에 긍정적인 영향을 미치고, 드라마 장르의 경우 부정적인 영향을 미친다고 하였다[13]. 하지만 국내의 경우 이와는 다르게 액션, 드라마가 인기가 높은 것으로 밝혀졌다. 이러한 현상은 문화적 할인의 영향이라고 볼 수 있다. 관람등급을 보면 12세 관람가 및 15세 관람가가 상관이 높게 나타났으며, 이는 등급이 해당영화를 볼 수 있는 잠재고객의 규모를 결정하기 때문에 관람등급이 낮으면 흥행에 정적인 영향을 미친다는 연구 결과와 일치한다[4].

둘째, 2그룹의 경우 장르 및 관람등급에서는 1그룹과 일치하였으나, 흥행에는 실패한 그룹으로 1그룹과 비슷한 특성을 가지나 리뷰에 작성된 글자 수가 다른 그룹에 비해 많은 것으로 나타났다.

셋째, 3그룹의 경우는 대중으로부터 관심이 낮은 그룹으로 애니메이션, 다큐멘터리, 뮤지컬, 공연실황이 해당되었으며, 리뷰 클릭 수, 리뷰의 글자 수 작고 상영시간도 짧은 것으로 나타났다. 기존 연구[14]에서도 상영시간이 길수록 흥행의 정도가 큰 것으로 나타났는데, 동일한 결과가 나온 것은 대중으로부터 관심이 낮은 장르인 다큐멘터리, 뮤지컬, 공영실황의 장르가 상영시간이 평균 92분으로 짧기 때문인 것으로 추정할 수 있다. 이 그룹에서는 관람등급은 전체 관람가와 상관관계가 높았다.

넷째, 4그룹의 경우 공포, 미스터리, 서스펜스, 스릴러, 멜로, 로맨스, 범죄의 장르가 속했으며, 청소년 관람불가의 작품이 많은 것으로 나타났다. 미국의 경우에는 코미디, SF, 공포 장르가 흥행성적에 정적인 요소로 작용하였지만[13], 국내 영화산업에서는 공포, 미스터리, 서스펜스, 스릴러와 같은 장르의 경우 흥행그룹에 속하지는 않았다. 특이점으로는 리뷰평점이 다른 그룹은 10점 만점 중 평균 8.0 이상으로 나타났으나, 4그룹에서는 평균 6.56점으로 나타났다. 반면 리뷰 클릭 수는 흥행 그룹 다음으로 많은 것으로 나타났다. 즉 4그룹은 다소 자극적인 장르의 영화가 속한 그룹으로 리뷰 평점은 낮으나 리뷰 사이트에서의 관심은 높은 것으로 나타났다.

5. 결 론

영화의 흥행 예측을 위해서는 영화평이나 구전과 같은 심리적 변인에 대한 측정 및 분석이 수반되어야 한다. 그러나, 저비용으로 빠른 시간에 이에 대한 신뢰성 있는 결과를 도출하기 위해서는 웹 크롤링이나 텍스트 마이닝과 같은 정보기술의 활용이 필수적이다. 본 논문은 텍스트 마이닝을 효과적으로 수행하기 위한 선행연구로서 군집분석을 활용하여 영화 흥행을 기반으로 4개의 그룹을 분리하고, 상응분석을 통하여 각 그룹별 영화의 내재적

특성을 분석하였다. 본 연구는 향 후 텍스트 마이닝 수행에서 텍스트 자료의 범위 선정 및 정제 과정의 신뢰도를 향상시킬 것으로 예상한다. 본 연구는 심리적 변인에 대하여 클릭 수나 리뷰 수와 같이 정량적 요인의 반영에 제한되어 있으나, 향 후 연구에서는 텍스트 마이닝을 통하여 정성적 요인도 반영함으로써 심리적 변인에 대한 정밀한 분석이 가능할 것으로 기대한다.

References

[1] Ahn, S.-A. and Kim, T.-J., The Determinants of Opening Share and Decay Rate in Motion Pictures, *Korea marketing Review*, 2003, Vol. 18, No. 3, pp. 1-17.

[2] Basuroy, S., Chatterjee, S., and Ravid, S.A., How critical are critical review? The box office effects of film critics, star power, and budget, *Journal of Marketing*, 2003, Vol. 67, No. 4, pp. 103-117.

[3] Bayus, B., Word of mouth : The indirect effect of marketing efforts, *Journal of Advertising Research*, 1985, Vol. 25, No. 3, pp. 31-39.

[4] Chang, B. and Ki, E., Devising a practical model for predicting theatrical movie success : Focusing on the experience good property, *Journal of Media Economics*, 2005, Vol. 18, No. 4, pp. 247-269.

[5] Chang, B.-H., Lee, Y.-H., Kim, B.-S., and Nam, S.-H., Elaborating Movie Performance Forecast Through Psychological Variables : Focusing on the First Week Performance, *Korean Journal of Journalism and Communication Studies*, 2009, Vol. 53, No. 4, pp. 346-371.

[6] Dellarocas, C., Zhang, X., and Awad, N.F., Exploring the value of online product reviews in forecasting sales : the case of motion pictures, *Journal of Interactive Marketing*, 2007, Vol. 21, No. 4, pp. 23-45.

[7] Duan, W., Gu, B., and Whinston, A.B., Do online reviews matter? An empirical investigation of panel data, *Decision Support Systems*, 2008, Vol. 45, No. 4, pp. 1007-1016.

[8] Ha, J.-M. and Moon, G.-J., An application of k-Means Clustering to Vehicle Routing Problems, *Journal of the Society of Korea Industrial and Systems Engineering*, 2015, Vol. 38, No. 3, pp. 1-7.

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

[10] Ko, J.-M., Study on the Factors Affecting Box Office Performance of Korean Movies-Focused on Patriotism Factor, *Journal of Media economics and culture*, 2008, Vol. 6, No. 4, pp. 7-39.

[11] Lee, F., Audience taste divergence over time : An analysis of U.S. movies' box office in Hong Kong, 1989-2004, *Journalism and Mass Communication Quarterly*, 2006, Vol. 83, No. 4, pp. 883-900.

[12] Lee, Y.-H., Chang, B.-H., and Park, K.-W., An Exploratory Study for Comparing Factors Affecting Box Office Performances between Countries : Focusing on Performances of U.S. Movies in South Korea and U.S., Korea *Regional Communication Research Association*, 2007, Vol. 7, No. 1, pp. 185-222.

[13] Litman, B. and Kohl, L.S., Predicting financial success of motion pictures : The '80s experience, *Journal of Media Economics*, 1989, Vol. 2, No. 2, pp. 35-50.

[14] Litman, B.R. and Ahn, H., Predicting financial success of motion pictures : The '90s experience, the motion picture mega-industry, Allyn & Bacon, 1998, pp. 172-197.

[15] Litman, B.R., Predicting Success of Theatrical Movies : An Empirical Study, *Journal of Popular Culture*, 1983, Vol. 16, No. 4, pp. 159-175.

[16] Liu, Y., Word or mouth for movies : Its dynamics and impact on box office revenue, *Journal of Marketing*, 2006, Vol. 70, No. 3, pp. 74-89.

[17] Park, J.-P., Park, M., and Kim, H.-W., A Study on the Relationship between Operational Method and Performance of Web Sites-Effect of CSR on Employees' Organizational Commitment and Productive Behaviors, *Journal of the Korean Society for Quality Management*, 2015, Vol. 43, No. 1, pp. 67-84.

[18] Park, S.-H. and Jung, W.-K., The Determinants of Motion Picture Box Office Performance : Evidence from Movies Released in Korea, 2006-2008, *Korea Regional Communication Research Association*, 2009, Vol. 9, No. 4, pp. 243-276.

[19] Prag, J. and Casavant, J., An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry, *Journal of Cultural Economics*, 1994, Vol. 18, No. 3, pp. 217-235.

[20] Ravid, S.A., Information, blockbusters, and stars : A study of the film industry, *Journal of Business*, 1999, Vol. 72, No. 4, pp. 463-492.

[21] Sung, Y.-S., Park, J.-Y., and Park, E.-A., The Influence of On-line Word of Mouth Information On Viewing Intention toward Motion Picture, *Advertising Research*,

2002, Vol. 57, pp. 31-52.

[22] Wyatt, R.O., High concept, product differentiation, and the contemporary U.S. film industry. In B. Austin(Ed.), Current research in film : Audiences, economics, and law, 1991, Vol. 5, pp. 86-105.

[23] Yoo, H.-S., The Determinants of Motion Pictures Box Office Performances-For Movies Produced in Korea Between 1988 and 1999, *Korea Regional Communication Research Association*, 2002, Vol. 46, No. 3, pp. 183-213.

[24] Zufryden, F.C., Linking advertising to box office performance of new film releases : A marketing planning approach, *Journal of Advertising Research*, 1996, Vol. 36, No. 4, pp. 29-41.

ORCID

Jai-Ill Lee | http://orcid.org/0000-0002-2807-1148
Young-Ho Chun | http://orcid.org/0000-0002-8235-3955
Chunghun Ha | http://orcid.org/0000-0002-4222-2555