공학박사학위논문

Popularity prediction of video content and clustering using networks: Movies, TV programs and Youtube channels

영상 콘텐츠에 대한 인기도 예측 및 네트워크를 활용한 군집화: 영화, TV 프로그램, 유튜브 채널

2021 년 8 월

서울대학교 대학원

산업공학과

안 용 대

Popularity prediction of video content and clustering using networks: Movies, TV programs and Youtube channels

영상 콘텐츠에 대한 인기도 예측 및 네트워크를 활용한 군집화: 영화, TV 프로그램, 유튜브 채널

지도교수 조 성 준

이 논문을 공학박사 학위논문으로 제출함

2021 년 7 월

서울대학교 대학원

산업공학과

안 용 대

안용대의 공학박사 학위논문을 인준함

2021 년 7 월

위 원 장 이 재 욱

부위원장 조 성 준

위 원 윤 명 환

위 원 이 영 훈

위 원 고 태 훈

Abstract

Popularity prediction of video content and clustering using networks: Movies, TV programs and Youtube channels

Yongdae An
Department of Industrial Engineering
The Graduate School
Seoul National University

The content market, including video content market, is a high-risk, high-return industry. Because the cost of copying and distributing the created video content is very low, large profit can be generated upon success. However, as content is an experience good, its quality cannot be judged before purchase. Hence, marketing has an important role in the content market because of the asymmetry of information between suppliers and consumers. Additionally, it has the characteristics of One Source Multi Use; if it is successful, additional profits can be created through various channels. Therefore, it is important for the content industry to correctly distinguish content with a high probability of success from the one without it and to conduct effective marketing activities to familiarize consumers with the product. Herein, we propose a methodology to assist in data-based decision-making using machine learning models and help in identifying problematic issues in video content markets such as movies, TV programs, and over-the-top (OTT) market.

In the film market, although marketing is very important, decisions are still made based on the sense of practitioners. We used the market research data collected through online and offline surveys to learn a model that can predict the number of audiences on the opening-week Saturday, and then use the learned model to propose a method for effective marketing activities. In the TV program market, programming is performed to improve the overall viewership by matching TV programs and viewer groups well. We learn a model that predicts the audience rating of a program using the characteristics of the program and the audience-rating information of the programs before, after, and at the same time, and use the resulting data to assist in decision-making to find the optimal programming scenario. The OTT market is facing a new problem of user's perception bias caused by the "recent recommendation" system. In the fields of politics and news particularly, if the user does not have access to different viewpoints because of the recommendation service, it may create and/or deepen a bias toward a specific political view without the user being aware of it. In order to compensate for this, it is important to use the recommended channel while the user is well aware of what kind of channel it is. We built a channel network in the news/political field using the data extracted from the comments left by users on the videos of each channel. In addition, we propose a method to compensate for the bias by classifying networks into conservative and progressive channel clusters and presenting the topography of the political tendencies of YouTube channels.

Keywords: Data mining, Machine learning, Artificial Intelligence, Decision support system, Recommendation system, Marketing, Prediction, Clustering, Box-office, Ratings, Broadcasting programming, Youtube, Channel network, Keyword extraction,

Filter bubble, Personalization

Student Number: 2015-21140

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The size and influence of the domestic video content market, including movies and TV programs, is continuously increasing. According to the Motion Picture Association of America (MPAA), the Korean film market was the 5th largest market worldwide in 2020. Of the global film market ($41.1 billion), the Korean film market accounts for $1.6 billion, followed by those of North America ($11.9 billion), China ($9 billion), Japan ($2 billion), and the United Kingdom ($1.7 billion). This is higher than India (7th), which is famous for its Bollywood film industry. Some of the major firms involved in the investment and distribution of Korean films are *CJ ENM, Next Entertainment World*, and *Showbox* and their films are sold worldwide every year. *The Handmaiden* (2016) directed by Chanwook Park was presold to 175 countries, breaking the record for the highest presales for a Korean film. Joon-ho Bong's global project *Snowpiercer* (2013) was also sold to 167 countries, demonstrating the power of Korean cinema [67]. In 2020, Bong's film *Parasite* won four Academy Awards, which further raised the status of Korean films. Korean TV programs such as dramas and entertainment shows are also attracting global attention. *Winter Sonata* and *Dae Jang Geum* were the protagonists of the first Korean wave by becoming hugely popular in the neighboring countries such as Japan, China, and Vietnam,

followed by recently screened *It's Okay to Not Be Okay* and *Itaewon Class* on the over-the-top (OTT) platform Netflix. *It's Okay to Not Be Okay* even topped Netflix viewing this month in seven Asian countries, including Hong Kong and Vietnam, and was 2nd in Japan, 4th in Peru, 5th in Australia, and 8th in Russia [63].

The content market, including the video content market, has three main characteristics. First, it is a high-risk, high-return industry. Second, as the content is considered the experience goods, it is difficult to understand its quality before purchase. Hence, marketing plays an important role in the content market because of the asymmetric availability of information between suppliers and consumers. Third, because content has the characteristics of One Source Multi Use, if it is successful, additional profits can be created through various channels. Hence, it is important to properly distinguish the content with a high probability of success from the one without it. It is also important to conduct effective marketing to make consumers aware about the content [21, 137].

Previously, many attempts have been made to apply Artificial Intelligence (AI), a technology that has recently been attracting everyone's attention, to the video content market. Many companies have begun to use AI to better differentiate the content types. For example, *Scriptbook*[1] provides AI-based scenario analysis services to predict the success of a scenario. *Cinelytic*[2] uses AI to determine the next blockbuster and scenario, analyze the market, and calculate efficient production costs. *Vault*[3] uses artificial neural network algorithms to suggest the optimal release time of a new movie.

[1] https://www.scriptbook.io/
[2] https://www.cinelytic.com/
[3] https://www.vault-ai.com/

However, the domestic academia of Korea still analyzes the factors affecting the box office. In addition, important industry players such as distributors and broadcasting stations are still making decisions based on the intuition of a few experts [2]. Even when deciding marketing expenses, which account for 30% of movie production costs and broadcast programming that can significantly impact the success of TV programs, decisions are made based on the experience and sense of practitioners instead of using data.

The contribution of this dissertation is threefold. First, it provides a solution to real-world problems of the video content market, such as marketing strategies for the movie market, programming strategies for the TV program market, and the filter-bubble phenomenon caused by the recommendation system of the OTT market. Second, it proposes a framework that can assist in data-based decision-making by both experts and users. Earlier, it was not possible to objectively evaluate the prediction, and therefore, it was not possible to analyze and resolve the problem. However, we can now quantify the evaluation of our predictions, add data, retrain the model, and so on. As a result, the performance can be continuously improved. Finally, for solving the problem, a framework with a high explanatory power was proposed both as a model and a method. Chapter 2 describes the use of a decision-tree model to directly identify the factors and rules affecting the number of movie audiences so that an effective marketing strategy could be executed. Chapter 3 identifies the factors with the highest influence on viewership ratings and discusses the use of the XGBoost model to predict the viewership ratings of TV programs. Chapter 4 describes the clustering of YouTube channels with a k-means model and demonstrates clustering by merging clusters based on keywords for postprocessing.

This dissertation is organized as follows. In Chapter 2, we propose a model that predicts the audience number on the opening Saturday using market research data obtained through online and offline surveys to help developing a marketing strategy. In Chapter 3, we propose a two-step framework for broadcasting programming strategy. The first step involves predicting the rating of a program with its program attributes, the ratings of programs assigned in the previous and following time slots, and the ratings of programs aired simultaneously in competing channels. The second step involves identifying the best airtime slot from the many candidates. In Chapter 4, we proposed a framework for constructing a channel network using comments on videos uploaded to YouTube's political/news channel, and calculating the degree of channel clustering and political bias for each channel. Finally, we discuss the contributions and future work of this dissertation in Chapter 5.

Chapter	**Market**	**Purpose**	**Proposed Method**
2	Movie	Support a marketing strategy	- Predict movie audience on first Saturday - Used method: Decision tree
3	TV Program	Support optimal broadcasting programming strategy	- Predict ratings of TV Program - Used method: XGboost
4	OTT	Support users avoid being politically biased	- Estimate user isolation score - Used method: k-means, Random walk with restart

Chapter 2

Prediction of Movie Audience on First Saturday with Decision Trees

2.1 Background

In Korea's motion picture industry, 'distribution' carries a significant meaning, as it encompasses the overall process of film release and showing. Efficient distribution helps to produce better film content, which may eventually lead Korean movies to gain a competitive advantage in the global market. In Korea, once the producers complete the production of a motion picture, the distributor estimates the proper number of screens to be secured for showing based on the production budget and the break-even point. This number is finalized after the coordination with multiple companies. Although not an absolute rule of thumb, the general belief in the market is that the revenue earned tends to increase with the number of screens secured. Hence, the ultimate goal of the distributors has been to secure as many screens as possible, so that the revenue exceeds the production cost in the shortest amount of time [112]. In the case of the foreign film industry, the investors and distributors are completely separate and independent entities, each playing their own roles. As for the Korean market, in contrast, the primary investors of the subject movie usually take on the role of distributors as well, since there exists only a small number of investing entities and distributors. In other words, a distributor in the Korean film industry

Table 2.1: The events and generated data that occur at each stage

Stage	Event	Data
Planning & development	- Scenario review - Investment decision - Pre-production	- Director/Actor - Production cost - Genre/Grade
Production	- Production - Post-production	
Marketing	- Monitoring preview - Survey	- List of the competitive Movies - Online/Offline survey

is, often times, also the investor of the subject movie. Therefore, the distributor tends to invest more actively in the marketing of the movie in order to retrieve the investment cost and maximize profit.

The process of film distribution is largely divided into the stages of planning and development, production, and marketing. Table 2.1 shows the generated events and data that occur at each stage. The marketing stage plays a key role in drawing people into the theater before the movie is actually released [116, 95, 49, 61]. Unless marketing is performed effectively, a new release will not win over other competing movies because it will not attract people's attention. In Korea, a box office score on the opening weekend also affects a theater's strategy. Based on weekend performance, the theater decides weekly how many screens and seats each movie will be allocated, in order to maximize sales. From the perspective of the distributor, if the movie does not attract large audiences during the weekend, the theater will give its screens and seats to others, and the movie is likely to lose the momentum to draw people in the future due to the low number of screens allocated to the movie. In order to avoid this situation, the distributor needs to consider a marketing strategy based on each movie's properties such as its director, actor, and competitive film information, as well as periodic information such as whether it is released in the summer peak

season or the holiday season. However, the prediction methods that rely on historical experience and the subjective intuition of the practitioners are not systematic or objective, and thus face several problems.

We cannot improve the accuracy of predictions using past decision cases. The reliability of predictions cannot be quantified because such predictions rely on the subjective feelings and experiences of practitioners. In addition, it is difficult to pass on an individual's experience. The experience and intuition of an individual is not something that can be clearly and logically explained because it is acquired while performing work over a long period of time. That is why it is very difficult to understand and communicate with colleagues. Furthermore, if the practitioner in charge of the business is replaced or leaves, the criteria for investment decisions in marketing that the company has maintained will also change. Owing to these problems, the current method used by film distributors is not suitable for use as a reference to determine marketing costs.

In order to solve these problems, we build a predictive model as a data-driven method that can forecast the audience numbers in a movie's first weekend utilizing market research data such as a simple awareness, effective awareness, preference, and effective preference of the movie, as well as basic attribute data such as nationality, distributor, rating, director, actor, and genre of the film. The constructed predictive model solves the three aforementioned problems. First, the released movie data and the weekly decision data are combined to consistently improve the prediction performance. Additionally, by applying a new algorithm to the accumulated database, the performance of the model can be gradually improved over time. Second, the reliability of predicted values can be evaluated. Conventionally, quantifying reliability

is difficult; however, the proposed model can facilitate this task. This is important in that the proposed model can potentially replace the current model in the future. Finally, we can organize a task in a form that can be delivered. Because prediction is based on the data-driven model rather than the intuition and experience of an individual, anyone can make predictions using the model regardless of his or her domain expertise. Therefore, even if the company changes the practitioners who make investment decisions, the company can maintain their investment standards using the proposed system.

To solve these three problems, we have created a standard for investing more strategically in marketing. Until now, investment was made without any clear plans because the decision standards for marketing expenses were not clear. Moreover, even if an investment was made, there was no standard on how the funds should be distributed. It is difficult to evaluate the exact effect investments that specific areas would have. However, the proposed method in this study can determine the most important variables and the extent to which the value of a variable should be increased to attain a certain audience number. For example, if the value of a variable in the market research data is less than the desired standard, the marketing capital could be injected into screening new trailers in order to achieve a desired number of audience participants on the first weekend. On the other hand, if the film is less known to the public, marketing capital can be expended to raise awareness by installing additional billboards and/or increasing the number of advertisements. Essentially, the company can invest effectively to maximize profit by considering the correlation between how much money is spent on marketing and the results achieved at the box office.

The rest of the chapter is organized as follows. Section 2.2 explains how previous studies are different from this one in terms of data, purpose, and methodological aspects of the research. Section 2.3 explains the data collected for the research and the trained predictive model. Section 2.4 shows the accuracy of the trained predictive models. In order to effectively communicate our findings, we used a Decision Tree as the main model and simultaneously compared it with the tree-based Gradient Boosting and Random Forest models. Section 2.5 summarizes this chapter by summarizing its findings.

2.2 Related work

The task of predicting movie popularity has been intensely studied both in Korea and abroad, of which related work so far is summarized in Table 2.2.

From the methodological perspective, the task of movie popularity prediction branches out into two mainstreams: a regression problem, which aims to predict the number of tickets sold directly, and a classification problem, where the success of the movie is categorized into a handful of classes as the target of prediction. These mainstream approaches can be categorized further depending on: (1) whether the "unit" of the movie popularity is defined as the number of seats sold or the revenue the movie has earned; and (2), whether the "temporal scale" of the popularity is limited to the first weekend following the release date or refers to the entire showing period. Studies focusing on the Korean market typically employ the number of seats sold as the measure for the movie popularity instead of total revenue, which is the usual convention in the U.S. market. This is due to the fact that the secondary venues for revenue, such as Blu-ray and DVD sales/rentals, are very weak in the

Table 2.2: Related studies by research purpose

(a) Regression

Target	**Attendance**	**Revenue**
Total	Park et al. [114] Kim and Hong [72] Marshall et al. [96] Hur et al. [52]	Liu [92] Delen et al. [31] Zhang et al. [146] Abel et al. [1] Qin [117] Wen and Yang [139] Lovallo et al. [94] Mestyán et al. [100] Rui et al. [121] Song and Han [132] Lepori [89] Kim et al. [70] Vujić and Zhang [135]
The first Saturday	Chang et al. [13]	Basuroy et al. [6] Dellarocas et al. [32] Joshi et al. [56] Moretti [101]

(b) Classification

Target	**Attendance**	**Revenue**
Total	Lee and Chang [83] Yim and Hwang [143] Lee et al. [84]	Sharda and Delen [124] Abel et al. [1] Bhave et al. [7] Lash et al. [77] Ghiassi et al. [42]
The first Saturday		Du et al. [34] Ghiassi et al. [42]

Korean industry. At the same time, the film industry in Korea comprises only a few entities that play the roles of the producer, distributor, and manager of movie theaters simultaneously, and the choice of the performance metric tends to align with the convenience of these entities who are the ultimate consumer of such metrics. Despite their methodological differences, most of the studies on the U.S. or Northern American market employ total revenue as the measurement of the success of the motion pictures. As for the prediction of attendance, the total number of tickets sold has mainly served as the popularity metric, and only a few studies approach the task as a classification problem. To our knowledge, we are the first in formulating the attendance prediction task as a classification problem, focusing on the performance during the first weekend following the release date.

Current and past research has expended much effort to identify the factors that affect the box office by using the data related to the essence of a movie, such as genre, released country, and grade. For this purpose, a number of models such as linear regression models have been used. Chang et al. [13] used 392 Korean and foreign films released in Korea between June 2007 and September 2008 using variables such as genre, rating, running time, and sequel flag, and based on the questionnaire data of 15,844 people aged 14 or older. They used the perception, preference, and intention of the audience as psychological variables to predict the opening weekend box office score, and studied the results through hierarchical regression analysis and path analysis. Park et al. [114] collected the data on 246 movies released in Korea to discover the determinants of the box office performance of the domestically screened films from 2009 to 2010, and trained a linear regression model with 202 movies (excluding 44 movies lacking production information). Additional variables included

rank, genre, star power of actors and directors, sequel flag, total production costs, number of screens, number of reviews, and online survey data. Chang et al. [13] and Park et al. [114] performed studies similar to our own in that they used survey data. However, their studies differed from the purpose of our study, which aims to systematize the marketing expenditure process through machine prediction. They are also different in that they collected small amounts of data in a short period and focused on the analysis of the factors affecting the box office score.

Current studies on the influence of the elements of a film have been conducted. Lash et al. [77] divided the elements of movies released in the United States from 1921 to 2014 into three categories: release time, movie content, and casting and production. Through this model, they analyzed quantitatively how each factor affected the box office performance of the movies. In this study, however, the revenue and audience number of each movie were not analyzed, as the study dealt with the problem of binary classification as to whether a movie achieved net profits. Their focus was on finding more accurate models rather than finding influential individual factors through Bayesian or neural network models. Lee and Chang [82] developed a Bayesian model to predict total ticket sales by using countries, directors, actors, genres, the number of reviews, ratings, the number of screens, seasons, audience share rates, and share rates of distributors. Zhang et al. [146] predicted the total revenue for 241 movies released in China from 2005 to 2006 using the released country, actor power, propaganda, running time, competition, and movie information. Neural networks were used for this prediction. These studies aim to improve the accuracy of the total revenue prediction models, yet there are few studies whose research objectives are to provide interpretable results that can directly benefit the

decision-making process of the active agents in the field.

Recently, attempts have been made to utilize more diverse data according to the development of online user–participation sites such as *Wikipedia* and social network services such as *Twitter* and *Facebook*. Kim et al. [70] developed three sequential forecasting models for predicting the non-cumulative and cumulative box office earnings: (1) prior to, (2) a week after, and (3) two weeks after release. 212 movies were selected among those released in Korea, and social network services data were collected by using a data collecting service called pulseK[1]. Asur and Huberman [3] studied 289 million *Twitter* data for 24 movies released in the U.S. from December 2009 to January 2010, including the number of tweets focused on directors and actors, and the number of tweets and retweets for movie marketing URLs to predict total weekend audience sales. Zimbra et al. [148] collected approximately 4 million tweets from a 9-weeks period on 29 movies released in the United States. Using iOS and Android platforms, the authors conducted a sentiment analysis in order to observe the differences in the users' responses in three separate periods of time—before, during, and after the release of the movie—and the possible connection between the types of platform and success of the movie. Mestyán et al. [100] trained a multivariate linear regression model with 312 movies to predict the total revenue of movies released in the U.S. in 2010, using activity records from *Wikipedia*. In the current study, we also considered using additional resources such as *Twitter* and *Wikipedia*, but we did not use them because they are not highly utilized by Koreans.

In addition, there have been studies to extract additional features by analyzing the scenarios of movie scenes and audience reviews. Eliahberg et al. [38] predicted

[1] http://eng.konantech.com/

box office performance of a movie at the point when only its script and estimated production budget were available. They extracted three levels of textual features from the scripts using screenwriting domain knowledge, human input, and natural language processing techniques. Du et al. [34] used microblog data to predict movie success. They extracted data from 120,413 microblogs written by 68,269 users for 24 movies by extracting frequently used words and emotions through text mining. They used 17 movies as training data and 7 movies as test data. However, there were two problems. One is that the number of movies was significantly lower than other studies. Moreover, when they split the data between the training data and test data, they did not consider the time flow of the power of the directors and actors, and instead simply used random sampling. In Korea, Lee et al. [84] used the reviews created before the release date of the top-ranked 375 movies of the *NAVER* movie section that opened in Korea from October 25, 2012 to December 31, 2014. They used ensemble models and predicted the final audience number. Hur et al. [52] predicted the popularity of 606 movies released in Korea from 2012 to 2016 through a machine learning method based on the independent subspace method. At this time, in addition to the in-film, external, and audience element variables, an emotional analysis of the review text was used. In the study by Hur et al. [52], they used the screen share as an external variable for each movie to predict the performance of a movie when its release is imminent. Our study is different because we excluded from our model the variables that can be obtained immediately before the release of a movie. Moreover, we trained a basic model at the planning stage, using only basic information about movies, while simultaneously developing a model for one week before release of the movies.

The principal difference of our study from the existing literature lies in the data we employ. We use results from online and offline surveys to gain deeper insights into the current market. At the same time, the size of the movie data we collected and analyzed in this study is greater than that of any previous work we consider. Finally, the interpretability of the results from our model may serve as guidelines for the marketing agents in the field during their decision-making process before the release of the movie.

2.3 Predictive model construction

2.3.1 Data

We obtained 7 years of online and offline survey data, from January 2010 to February 2017, from one of the major domestic distributors. We also collected data on the basic properties of movies from the Korean Film Council[2] and the website of *NAVER* movie section[3], the most popular portal site in Korea.

In particular, to predict the audience number for an opening weekend, we collected box office score data on a daily, weekly, monthly, and yearly basis through the homepage of the Korean Film Council. This data is largely divided into three categories: box office, screen share rate, and seat occupancy rate, and additional data that is essential but not detailed enough in the Korean Film Council, such as information on the actors and crew, while storylines were collected from *NAVER* movie section. Through this process, we obtained more detailed information on directors, leading actors, supporting actors, etc., which we used to calculate the power of directors and actors. Finally, we conducted experiments on 325 Korean movies

[2]http://www.kofic.or.kr/kofic/business/main/main.do
[3]http://movie.naver.com/

and 296 U.S. movies, which had both total production cost information and survey information.

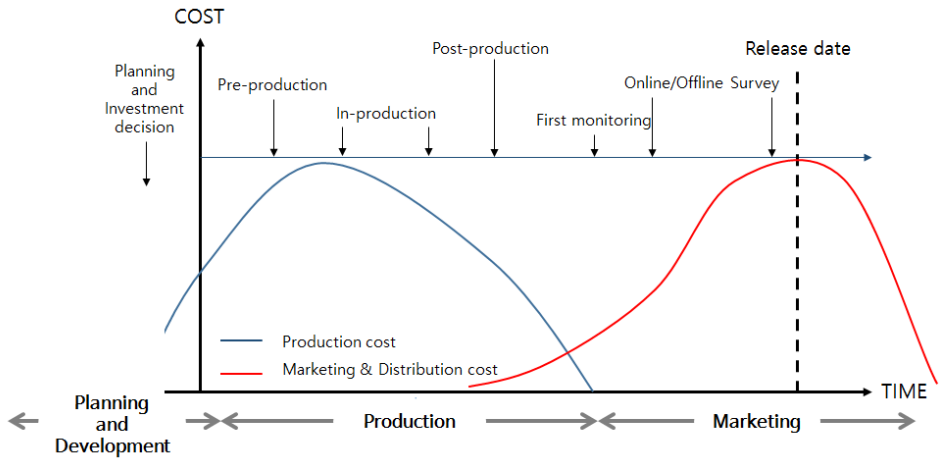

Figure 2.1: Movie production events based on the opening date [22]

Figure 2.1 shows the production cost and marketing cost based on the opening date, where the amount of data that can be used is different based on the opening date. In particular, in the case of market research data, as the survey is conducted weekly, the closer to the opening date, the larger the amount of survey data. Datasets are classified into four types, as shown in Table 2.3, according to available data for each prediction time. D is a dataset that adds director and actor power data based on past performances and basic movie attributes such as released country, distributor, rating, genre, running time, and total production cost after the director and actors have been selected. D_{-6}, D_{-3}, and D_{-1} are the datasets with the survey data added during the marketing stage. D_{-6} refers to the dataset that includes the market research data collected six weeks prior to opening, and D_{-1} refers to the dataset that includes the market research data collected from six weeks before opening to

Table 2.3: Dataset description

Dataset	Data		
	Basic Attributes[1]	**Director/Actor power**	**Survey (before opening)**
D	O	O	X
D_{-6}	O	O	6 weeks
D_{-3}	O	O	6 weeks $\sim$ 3 weeks
D_{-1}	O	O	6 weeks $\sim$ 1 week

[1] country, distributor, grade, genre, running time, total production cost, etc.

one week before opening.

2.3.2 Target variable

After labeling the number of tickets that each movie sold on a Saturday into three classes, we set the classes (audience number) as the target variable. The classes are used as the criterion by which the distributor empirically judges whether or not a movie can succeed, using the first weekend audience number. As it happens, the decision-making of many distributors in Korea during the marketing process of the subject movie relies on the arbitrary sectioning of the estimated seats sold and guessing which section the subject movie would fall into. This is to simplify the regression problem of predicting the exact number of seats sold into a classification problem, hence speeding up the decision-making process by establishing a different response plan for each class. In this study, we adopt the general standards conventionally used in the movie industry for binning number of seats sold. These classes correspond to real categories that are used in the real movie business world. Table 2.4 shows the criteria dividing the classes and movies belonging to each class. We further categorize movies at the domestic box office into Korean and U.S. This is because the total production cost varies enormously depending on the location of production. We discuss the production cost in greater detail in Section 2.3.3. Note

Table 2.4: Target class description

Class	Number of first Saturday audience	Korean		U.S.	
		Movie	Number(%)	Movie	Number(%)
1	More than 300,000	*New Trial* *The Priests* *Asura : The City of Madness*	64(10%)	*The Martian* *Star Wars : The Force Awakens* *Interstellar*	29(5%)
2	More than 150,000 and less than 300,000	*Familyhood* *The Tiger* *Coinlocker Girl*	88(14%)	*Pixels* *Mad Max: Fury Road* *The Maze Runner*	51(8%)
3	Less than 150,000	*Misbehavior* *Lucid Dream* *A single rider*	173(28%)	*The Big Short* *The Hateful Eight* *Burnt*	216(35%)
Total		325(52%)		296(48%)	

that the survey questionnaire and the actor/director importance data is collected only domestically, hence it is considered invariant to the production location.

The audience at the opening weekend showing is a very important criterion for determining how many screens the theater will allocate to the movie. If the audience number is too low on the opening weekend, the theater attempts to maximize sales by allocating the screens and seats of the movie to other movies. So, even if a movie may be successful in the long run, if it does not attract enough audience during the opening weekend, it will be in a disadvantageous position compared to competing movies in terms of the number of screens and seats. Therefore, the newly released movie must try to attract as many audiences as possible on the opening weekend, and predicting the audience number on the opening weekend is essential to maintaining the competitiveness of the movie.

In this study, we consider it important to improve the decision-making system so that the distributor can use the marketing budget as efficiently as possible by using our model. As shown in Figure 2.1, the marketing activity is still ongoing even after the opening, although the cost is lower at the time of opening. However, after opening the movie, it is difficult to evaluate the influence of marketing expenses on

the audience because the audience is influenced by reviews or posts by the audiences who have already watched the movie. So, in order to focus on how to use the survey data to make effective marketing decisions before opening, we predict the audience number on the first Saturday.

In addition, we formulate the popularity prediction task as a classification problem, where the class of the number of seats sold on Saturday of the release weekend is the target prediction variable. Such a restriction originates from the fact that: (1) the marketing agents in the field have traditionally relied on the number of audiences from the Saturday of the release weekend to make an educated guess of the movie's success; and (2), there exists a highly positive correlation between the number of audiences from the Saturday of the release weekend and the total number of audiences.

2.3.3 Predictor variable

We used basic attribute data of movies, such as released country, distributor, rating, genre, running time, total production cost, and director/actor power, as well as market research data obtained from online and offline surveys, such as awareness and preference as predictor variables. When considering the differences between the total production cost of Korean and U.S. movies, the total production cost is normalized by country, and we newly include director and actor power from previous performances by taking into account their history. In particular, in the case of market research data, we received online and offline survey data for 7 years, from January 2010 to February 2017, from one of the major domestic distributors in Korea.

The normalized total production cost

The total production cost is the most important variable of the basic properties of a movie. There is a very large gap in terms of production cost between Korean and U.S. movies. In the case of Korean movies, the average total production cost, including marketing expenses, is approximately 6 billion won. On the other hand, in the case of American movies released in Korea, there are many movies with much larger total production costs, where the average total production cost of the movies we analyzed was approximately 100 million dollars—roughly 20 times more than the Korean movies we analyzed. In order to take into account differences in production scale, the total production cost for each country is normalized by a z-score [47, 76]. Table 2.5 shows the total production cost of movies in Korea and the U.S. and the normalized total production cost. Even *The Admiral: Roaring Currents*, whose total production cost was very high among Korean movies, had a lower budget than *Burnt*, whose production cost was lower than that of the average U.S. movie.

Director and actor power

In the case of director and actor power among the basic attributes, much research has been conducted on whether or not a star actor can influence the box office score, and a strong influence by a star actor has been proven in various studies [90, 123, 131]. In addition, research on the Italian film industry by [5] has shown that not only the popularity of actors but also the popularity of directors influences attendance, albeit in a nonlinear way, and it has been proven that the coexistence of these two factors will have a greater effect on attendance [114]. In order to quantify the influence of directors and actors, we calculated the directors' and actors' power based on the

Table 2.5: Total cost and normalized total cost for each country

(a) Korean movie

Movie	**Total cost (Billion won)**	**Nomalized total cost**
The Outlaw	1	-0.95
The Client	3.9	-0.10
Roaring Currents	15	2.65
Asura: The City of Madness	9	1.19
Luck-key	4.1	0.00

(b) U.S. movie

Movie	**Total cost (Million $)**	**Nomalized total cost**
Unknown	70	-0.60
Captain America: The First Avenger	200	1.15
Ted	85	-0.40
The Hobbit: The Battle of the Five Armies	325	2.83
Burnt	20	-1.27

box office scores of previous movies that they directed and acted in, based on the opening date of the movie, as shown in Equation 2.1. As the influence of the director and actor(s) on the public decays naturally over time, to give a higher weight to the movie's performance close to the movie's opening, we applied a decay factor so that influence decreases with time. In this study, we set the decay factor to 0.5.

$$Power_A = \sum_{\forall i | D_i < D_A} \left[e^{-\frac{1}{3\times 365}(D_A - D_i)} \times \frac{Aud_i}{10,000,000} \times DF \times I_{P_A == P_i} \right],$$

where D_A : Opening date of the target movie,

D_i : Opening date of the movie i,

P_A : Director or actor of the target movie,

P_i : Director or actor of the movie i,

Aud_i : Total number of audience of movie i,

DF : Decaying factor

(2.1)

Table 2.6 shows the calculated director powers for the top 10 movies using Equation 2.1. Director Lee Joon-ik, who directed *King and the Clown* and *The Throne*, and director Choi Dong-hoon, who directed *Jeon Woo-chi: The Taoist Wizard* and *The Thieves*, are ranked at the top. Likewise, Michael Bay's latest work, *13 Hours*, had the greatest director power, which was produced after several other successful Bay movies such as *Transformers: Revenge of the Fallen*, *Transformers: Dark of the Moon*, and *Transformers: Age of Extinction*. Table 2.7 shows the total audience number and director power for each of Michael Bay's movies. As shown in Table 2.7, director powers differ depending on the opening date of the movie—even for the same director—because we calculated the director power via the box-office score of previous movies, based on the opening date of each movie.

In addition, as expected, the power of domestic actors such as Oh Dal Soo and Hwang Jung-min, who appeared in various big hit movies, was high. Among the actors who appeared in more than twelve movies released in Korea, there are

Table 2.6: Top 10 movies for director power

Rank	Movie	Opening date	Director	Director power
1	*13 Hours*	2016-03-03	Michael Bay	1.22
2	*DongJu; The Portrait of A Poet*	2016-02-17	Lee Joon-ik	1.19
3	*Assassination*	2015-07-22	Choi Dong-hoon	1.17
4	*Seoul Station*	2016-08-17	Yeon Sangho	1.14
5	*Transformers: Age of Extinction*	2014-06-25	Michael Bay	1.09
6	*Transformers: Dark Of The Moon*	2011-06-29	Michael Bay	1.02
7	*The Himalayas*	2015-12-16	Lee Seok-hoon	0.93
8	*Sunny*	2008-07-24	Lee Joon-ik	0.90
9	*Confidential Assignment*	2017-01-18	Kim Sung-hoon	0.90
10	*Interstellar*	2014-11-06	Christopher Nolan	0.89

Table 2.7: Director power of director Michael Bay for each movie

No	Movie	Opening date	Total Audience	Director power
1	*Transformers: Revenge Of The Fallen*	2009-06-24	7,387,680	0.69
2	*Transformers: Dark Of The Moon*	2011-06-29	7,749,860	1.02
3	*Transformers: Age of Extinction*	2014-06-25	5,295,021	1.09
4	*13 Hours*	2016-03-03	124,249	1.22

Ryan Reynolds, Liam Neeson, Michael Fassbender, Anne Hathaway, Gerard Butler, Hugh Jackman, and others. Figure 2.2 shows the trend of the actor power for each actor according to the movie opening date. In the case of Hugh Jackman, who is particularly loved in Korea, the total audience and actor power per movie released in Korea are shown in Table 2.8.

As with the director power, the actor power may also be influenced by the performance of the previous movies and the release date of the current movie, whose effect is discounted at a decaying rate. Under the assumption that the importance of the director(s) and actor(s) of the subject movie will be realized to a greater extent as the timing of the observation approaches the movie's release date, it is reasonable to allot greater weight to the success of more recent movies than to that of older movies. Hence, we employ a decaying factor and use it to discount the power variables depending on the distance of the variable's timing from the release date,

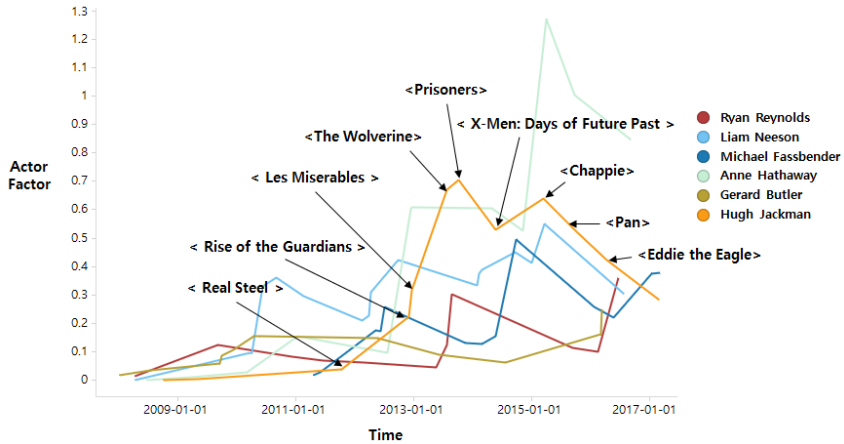

Figure 2.2: Actor power for each actor

instead of the conventional total number of movies released before the release of the subject movie, to account for the director/actor influence.

Market research data

We used two types of questionnaire data, namely that of offline and online surveys that were conducted every week. The biggest difference between online and offline surveys is that the former are conducted by online research firms, while the latter are conducted by offline exit survey firms through papers with surveys written in the field, such as at movie theaters. For this reason, the questions of offline surveys are much fewer than the online surveys, and are organized in a way to comprehensively examine several movies. On the other hand, since online surveys are conducted by participants who are willing to participate, we can ask more directly about how much they know and how much they want to watch each movie. In the offline survey, we divided the movies into two groups (6–3 weeks before opening and 2 weeks before opening to opening week) based on the time of the opening, and asked participants

Table 2.8: Actor power of Hugh Jackman for each movie

No	Movie	Release date	Total Audience	Actor power
1	*Deception*	2008-10-02	73,126	0.00
2	*X-Men Origins: Wolverine*	2009-04-30	1,303,452	0.00
3	*Real Steel*	2011-10-12	3,579,666	0.04
4	*Rise of the Guardians*	2012-11-29	1,081,096	0.22
5	*Les Miserables*	2012-12-19	5,920,520	0.32
6	*The Wolverine*	2013-07-25	1,075,333	0.67
7	*Prisoners*	2013-10-02	188,702	0.71
8	*X-Men: Days of Future Past*	2014-05-22	4,313,871	0.53
9	*Chappie*	2015-03-12	573,661	0.64
10	*Pan*	2015-10-08	186,174	0.52
11	*Eddie the Eagle*	2016-04-07	226,318	0.42
12	*Logan*	2017-03-01	2,168,633	0.28

to select all movies they knew in each group. Based on the survey data, we collected offline market research data such as the top-3 preferences and top preference.

Figure 2.3 shows a sample of an offline survey conducted in February 2013. In the online survey, we divided movies into six groups, from movies that lasted six weeks based on the opening date to movies which lasted one week, and a questionnaire was conducted. The questionnaire participants selected all the movies they knew from each group, choosing one of four ratings on how much they knew about the selected movies, and evaluating them based on seven ratings. Through this, we collected market research data such as simple awareness, effective awareness, preference, and effective preference data. Figure 2.4 is a sample of an online survey conducted in May 2013. In order to extract more information from collected survey data, derived variables such as preference, second & third preference, net simple awareness, net effective awareness, and net preference were additionally generated using the complement relation among the respective factors. Table 2.9 explains each variable, and Figure 2.5 shows the relationships among variables. Variables in the market research

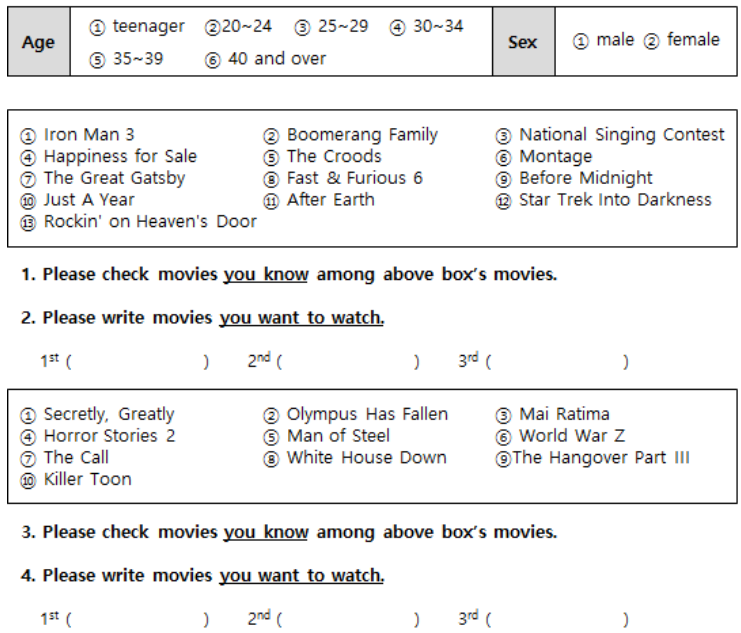

Figure 2.3: A sample of offline survey form

data from online/offline surveys were arbitrarily transformed for data privacy reasons. In order to distinguish among variables, we added the word "online" before the name of the transformed variable for the online survey, and added "offline" to the variable name for the offline survey. Depending on how many weeks remained before the release date, the variables were masked as a number between 1 and 6, and the details were converted using letters from the alphabet such as a, b, c, and d. If a variable is "online_a-6," it indicates a specific market research data variable obtained as a result of an online survey six weeks before the movie's opening date.

Seasonality

As mentioned earlier, the movie's first weekend performance is highly correlated with the total audience. This is because theaters allocate more screens to movies

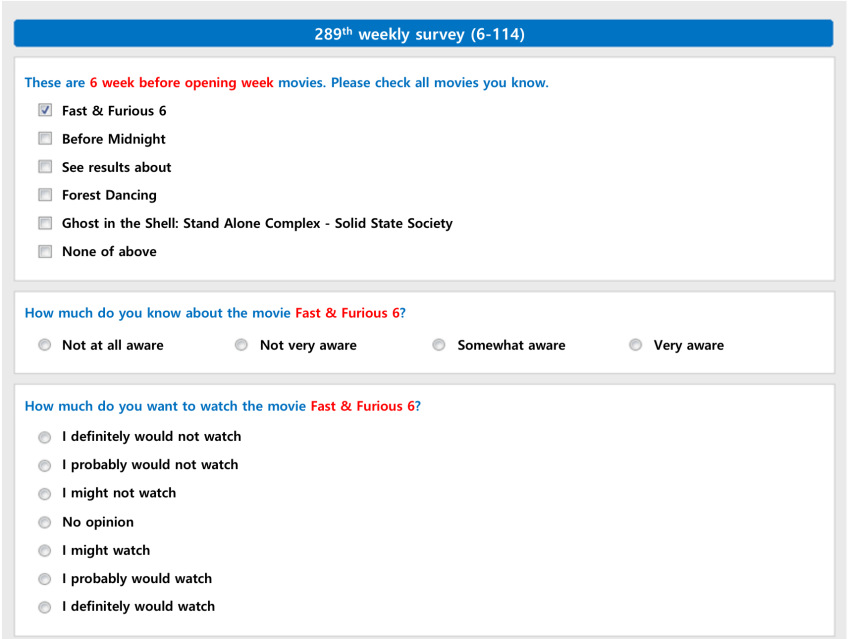

Figure 2.4: A sample of online survey form

that are more likely to hit the market in order to maximize their profits, as well as attracting more visitors due to the fact that many people watched and word-of-mouth effects. In addition, there are time-related factors such as when the film is released, which can greatly affect the film's audience outcome [37]. Even in the same movie and competition environment, a movie that is released when many people want to watch it will attract a greater audiences than when it is not. In fact, the total number of audience who watched the movie every month has been growing steadily since 2010, as shown in Figure 2.6. In January, July, August, and December, the number of visitors to theaters tends to be higher than in other months. In March, April, October and November, the total number of visitors tends to be relatively low compared to other months. This phenomenon mainly coincides with vacation periods of people between the ages of 10 and 29, not only because of their individual

Table 2.9: Online and Offline survey variables and their formula

Survey type	Variable	Formula
Offline	best 3 preference	People who chose the movie in their top 3 intending to watch list / Total respondents
	Highest preference	People who chose the movie as their first in their top3 intending to watch list / People who chose the movie in their top3 intending to watch list
	Not in third preference	1 - best 3 preference
	Within second and third preference	best 3 preference - Highest preference
Online	Simple awareness	People who know the movie / Total respondents
	Effective awareness	People who checked awareness of scale 3 and 4 / People know the movie
	Preference	People who checked preference of scale 6 and 7 / People know the movie
	Effective preference	People who checked awareness of scale 3 and 4 checked preference of scale 6 and 7 / People who checked awareness of scale 3 and 4
	NET simple awareness	Simple awareness - (Effective awareness ∩ Preference)
	NET effective awareness	Effective awareness - Effective preference
	NET preference	Preference - Effective preference

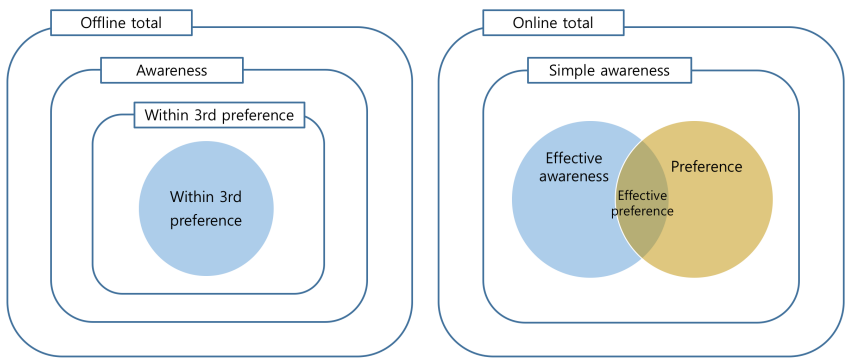

Figure 2.5: Diagram on relationships among survey variables

consumption, but also because of the intensive family visits. In order to reflect the effects of the release period of the movie, we divided months into “hot” (January, July, August, and December), “cold” (March, April, October, and November) and “normal” (February, May, June, and September) seasons to reflect seasonality.

2.3.4 Decision Tree and ensemble prediction models

The purpose of this study is to predict the opening audience numbers by using the trained model, and then to establish a marketing strategy to maximize the marketing effect. Therefore, interpretability is also an important factor, along with the accuracy of the trained model. We selected the Decision Tree [107], which has

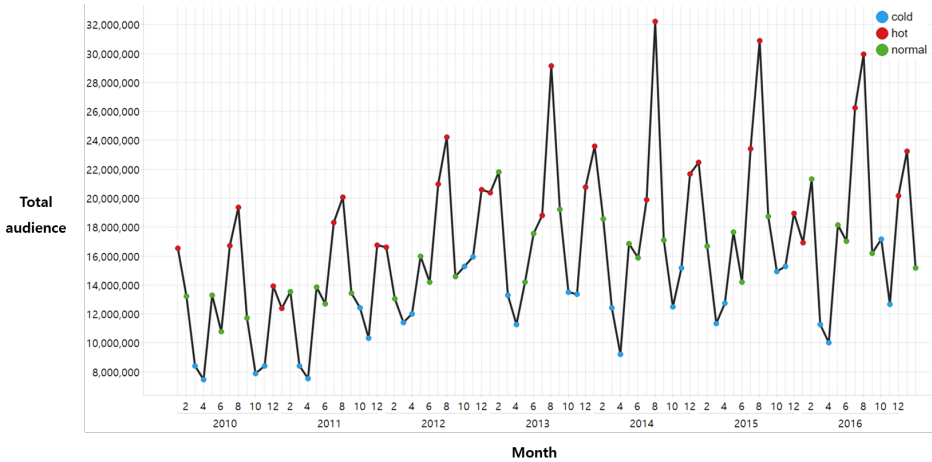

Figure 2.6: Monthly total audience

high interpretability for the trained results and which is robust against outliers. In order to compare and evaluate the performance of the trained model, the Gradient Boosting [40] and Random Forest [10] were trained at the same time. We excluded black box models such as neural networks, which have relatively poor interpretability compared with tree based models. The candidate parameter values for each model are shown in Table 2.10, and were learned through the validation dataset.

We do not horse-race our performance with other existing literature, since there are no other studies that employ the same level of data we used, nor are there any that formulate the same prediction problem as we did.

2.4 Prediction model evaluation

As shown in Table 2.3, the datasets were divided into weekly time spans and the models were repeatedly trained accordingly based on the prediction times. The total data consists of 325 Korean and 296 American movies, for which the total production cost and survey information was obtained. To divide the data for training, validation,

Table 2.10: The candidate parameter values for each model

Models	Parameters and values
Decision Tree	Minimum number of samples for splitting : 5, 15 Minimum number of samples in leaf nodes : 5, 15 Maximum depth of estimator : 3, 4, 5, 6
Gradient Boosting	Number of estimators: 500 learning rate: 0.01, 0.1 Maximum depth of estimator : 3, 4, 5, 6 Minimum number of samples for splitting : 5, 15
Random Forest	Number of estimators : 500 Minimum number of samples for splitting : 5, 15 Maximum depth of estimator : 3, 4, 5, 6

and testing, the datasets were once again separated into training sets, validation sets, and test sets, divided by the release year of movies in 2014 and 2015. As a result, we trained models using 386 movies released from January 2010 to December 2013 and validated the performance of the models presented in Section 2.3.4 with 73 movies released from January 2014 to December 2014, measuring their accuracy against 162 films released from January 2015 to February 2017. We considered using K-Fold cross validation due to the data size. However, since the director and actor power used as predictor variables are inherent in past box office performances, we did not use the K-Fold cross validation to consider time flow.

The accuracy is the proportion of the actual class of the audience on the opening Saturday among the total number of movies, as shown in Equation 2.2. Generally, the larger the number of classes is, the more difficult it is to construct a model with high accuracy.

Table 2.11: Accuracy for each model and dataset

Dataset	Accuracy		
	Decision tree	**Gradient boosting**	**Random forest**
D	0.605	0.630	**0.667**
D_{-6}	0.636	0.673	**0.716**
D_{-3}	0.728	0.698	**0.735**
D_{-1}	**0.778**	0.728	0.765

$$
\begin{aligned}
Accuracy = \frac{\mid \mathrm{x} : \mathrm{x} \in D_t \wedge \hat{y}(\mathrm{x}) = y(\mathrm{x}) \mid}{\mid \mathrm{x} : \mathrm{x} \in D_t \mid} \\
\text{where } D_t : \text{Test datset,} \\
\hat{y}(\mathrm{x}) : \text{Predected class,} \\
y(\mathrm{x}) : \text{Actual class}
\end{aligned}
\tag{2.2}
$$

In this study, because the interpretability of the model is as important as the accuracy, we trained the Decision Tree, Gradient Boosting, and Random Forest tree-based models, which can extract important variables with high prediction performance. Decision Tree is excellent in the interpretability because it can find rules directly through learning. Gradient Boosting and Random Forest can extract important variables through variable importance, so we can measure the influence of variables with these algorithms.

Table 2.11 shows the performance of the models trained with the datasets, and Table 2.12 shows the variable importance of each dataset.

The model performance using the D_{-6}, D_{-3}, and D_{-1} datasets with market research data improves as the time of prediction approaches the release date, and data collected at a point closer to the opening date is more effective in predicting the opening audience number. The confusion matrix of the Decision Tree for each

Table 2.12: Variable importance of the Decision Tree

(a) D

Variable	Importance
Normalized_Total_Production_Cost	0.608
Actor_power_0.5	0.194
Grade(All)	0.054
Genre(Romance)	0.051
Country(Korea)	0.048
Season(Hot)	0.038
Season(Cold)	0.003
Genre(Horror)	0.003

(b) D_{-6}

Variable	Importance
Online_h-6	0.636
Actor_power_0.5	0.075
Online_l-6	0.071
Normalized_Total_Production_Cost	0.064
Season(Hot)	0.042
Online_g-6	0.037
Online_o-6	0.036
Online_d-6	0.022
Online_j-6	0.016
Season(Normal)	0.001

(c) D_{-3}

Variable	Importance
Online_g-3	0.540
Normalized_Total_Production_Cost	0.137
Online_h-3	0.190
Actor_power_0.5	0.076
Season(Hot)	0.040
Online_k-3	0.012
Online_j-3	0.005

(d) D_{-1}

Variable	Importance
Online_g-1	0.574
Normalized_Total_Production_Cost	0.189
Online_n-1	0.137
Online_h-1	0.046
Season(Hot)	0.031
Genre(Drama)	0.015
Director_power_0.5	0.009

Table 2.13: Confusion matrix of the Decision Tree for each dataset

(a) D

		Actual		
		Class 1	**Class 2**	**Class 3**
Predicted	Class 1	24	6	0
	Class 2	12	6	12
	Class 3	13	21	68

(b) D_{-6}

		Actual		
		Class 1	**Class 2**	**Class 3**
Predicted	Class 1	12	16	2
	Class 2	0	18	12
	Class 3	2	27	73

(c) D_{-3}

		Actual		
		Class 1	**Class 2**	**Class 3**
Predicted	Class 1	23	3	4
	Class 2	6	13	11
	Class 3	5	15	82

(d) D_{-1}

		Actual		
		Class 1	**Class 2**	**Class 3**
Predicted	Class 1	18	6	6
	Class 2	3	15	12
	Class 3	1	8	93

dataset is shown in Table 2.13. In particular, the f1-scores for each class of dataset D_{-1}, which is the most accurate and closest to the opening date, are 0.69, 0.51, and 0.87 for class 1 ($<$ 150,000 audience), class 2 (150,000 to 300,000 audience), and class 3 ($>$ 300,000 audience), respectively. While class 1 and class 3 fit well, class 2 has a weaker tendency to fit.

Specifically, let us compare the Decision Trees among the different prediction times. The information that appears in the nodes of the Decision Trees is as follows. The color of the node varies according to the rule used in the branch, the Gini coefficient of the branched data, the distribution of data by the class of node, the class information of prediction, and the class distribution of data allocated to each node. Gini impurity is a measure of misclassification given to a partitioned subset. It measures how a randomly sampled item from the subset would be misclassified if it were labeled randomly according to the distribution of labels in the subset. As shown in Equation 2.3, the Gini impurity can be computed by summing the probability p_i of an item with the label i being chosen times the probability $1 - p_i$ of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

$$I_G(p) = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2$$
$$= 1 - (\sum_{i=1}^{J} p_i \sum_{i=1}^{J} p_i - \sum_{i \neq k} p_i p_k) = \sum_{i \neq k} p_i p_k \tag{2.3}$$

where J : classes,

$i \in \{1, 2, ..., J\}$,

p_i : the fraction of items labeled with class i in the set

If there are many movies for the class of 150,000 audience or fewer, the color of the node is orange. If there are many data in the class of between 150,000 and 300,000 audience, the color of the node is green. If there are many data for the class of 300,000 or more audience, the color of the node is purple.

From Table 2.12, we know that unlike the Decision Tree for dataset D, where the variable of the normalized total cost is the most important, various market research data also influence the Decision Tree for datasets D_{-6}, D_{-3}, and D_{-1}. Moreover, the effect of the normalized total production cost is reduced. This indicates that the market research data is more effective in predicting the number of spectators because it implicitly contains the information of the normalized total production cost.

When we compare the Decision Tree models for datasets D_{-6}, D_{-3}, and D_{-1}, only dataset D_{-6} has access to market research data six weeks before release date, but datasets D_{-3} and D_{-1} also use market research data previously surveyed. Because using the data surveyed in a given week is more effective in predicting the audience number, among the variables of the model, the variables most closely re-

Table 2.14: The prediction result of the Decision tree for each dataset

Movie	Opening date	Total audience	1st Sat. audience	Actual class	Predicted class			
					D	D_{-6}	D_{-3}	D_{-1}
Night at the Museum: Secret of the Tomb	2015-01-14	1,076,461	177,554	2	3	2	2	2
Mad Max: Fury Road	2015-05-14	3,830,353	290,442	2	2	2	2	2
Terminator Genisys	2015-07-02	3,229,800	518,316	1	1	2	1	1
Inside Out	2015-07-09	4,941,734	279,092	2	3	3	3	2
Minions	2015-07-29	2,612,040	239,145	2	3	2	2	2
Mission: Impossible - Rogue Nation	2015-07-30	6,103,081	766,502	1	1	1	1	1
The Intern	2015-09-24	3,604,184	115,835	3	3	2	2	2
The Good Dinosaur	2016-01-07	1,321,864	197,495	2	3	3	3	3
The Age of Shadows	2016-09-07	7,483,039	661,311	1	1	1	1	1
The Map Against the World	2016-09-07	963,742	88,353	3	1	2	2	2

lated to the week before the opening date are given higher priority. In other words, the closer to the release date, the more informative market research data becomes.

Table 2.14 shows 10 movies selected from the learned Decision Tree model. Figure 2.7 shows the results of class prediction for each dataset. For *The Age of Shadows*, *Mission: Impossible - Rogue Nation*, and *Mad Max: Fury Road*, all models accurately predicted the actual class from the datasets. In the case of the Korean movie *The Age of Shadows*, which was directed by the famous director Kim Jee-woon and starred such famous actors as Song Gang-ho, Gong Yoo, and Han Ji-min, the director and actor power was high. The expectations of the audience were revealed in the survey data. As a result, the movie was predicted as class 1. Similarly, in the case of the foreign movie *Mission: Impossible - Rogue Nation*, involving Christopher McQuarrie, who previously succeeded with *Mission: Impossible 6*, and Tom Cruise, who is highly regarded in Korea, class 1 was predicted from the first to the final prediction. *Mad Max: Fury Road* was predicted to be class 2, owing to relatively low director and actor power, as director George Miller and actor Tom Hardy were not well known in Korea before the movie. Marketing activities could not influence the audience's recognition and preference, and the movie finally received a class 2 as a result.

On the other hand, in the case of *The Map Against the World*, similar to *The Age*

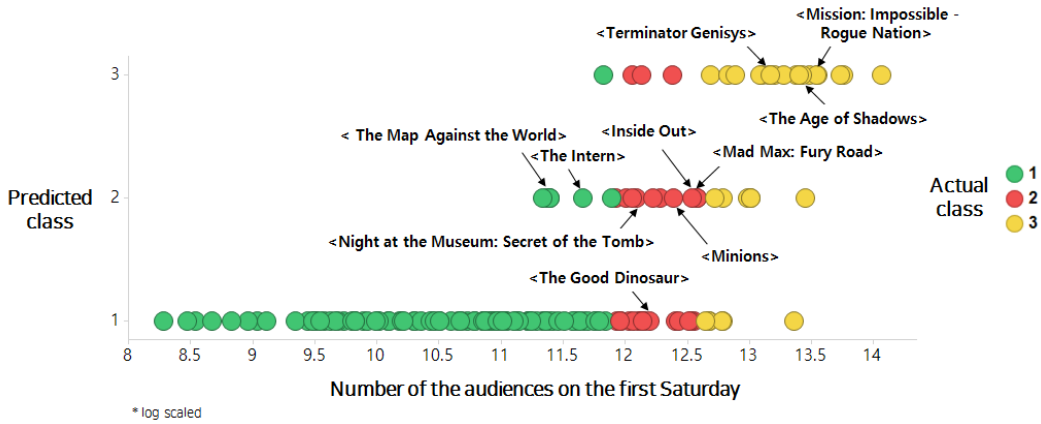

Figure 2.7: The prediction result for dataset D_{-1}

of Shadows, the famous Korean director Kang Woo-suk directed the movie, in which the famous actors Cha Seung-won and Yoo Jun-sang also appeared. Although the movie was predicted as class 1 for dataset D without market research data, the class was predicted to be lower as it approached the opening date because of its relatively lower audience preference (compared to *The Age of Shadows*). Through this analysis, it was confirmed that the relative superiority of recognition and preference for movies released at the same time is important.

In the case of the movies *Inside Out*, *Minions*, and *The Good Dinosaur*, they were predicted as class 2 despite the domestic environment in which animated movies generally do not succeed. For animated movies in general, there were many cases predicted as class 3 for dataset D, because famous actors generally do not participate in those movies. However, when *Inside Out* and *Minions* were close to their opening week, the prediction changed to class 2, the actual number of the Saturday's audience numbers for each movie. This was interesting given that both films were screened at the same time as big hit movies such as *Mission: Impossible - Rogue Nation*,

Pixels, and *Terminator Genisys*. We assumed this was because they steadily raised public awareness and preference through active marketing such as movie trailers and bus advertisements. On the other hand, in the case of *The Good Dinosaur*, it finally recorded as class 2, but it was predicted as class 3 until the week prior to the release date. Despite the lack of awareness and preference, we believe the reason why it ultimately reached class 2 is because the movies that screened at the same time were relatively weak movies such as *Steve Jobs*, *Mood of the Day*, *Goosebumps*, *Remember You*, *Point Break*, and *The Hateful Eight*. So, we assume that relative awareness and preference are important in some cases.

The Intern, *Terminator Genisys*, and *Night at the Museum: Secret of the Tomb* were predicted as a lower class initially, but as they moved closer to the opening date, a more accurate class was predicted through datasets with market research data. In these cases, the movies were actively promoted in Korea through movie trailers and bus advertisements.

2.5 Summary

Marketing activity by distributors is a significant factor in attracting audiences to the theater before a movie is actually released. In particular, the audience number on the opening weekend is important, because it is highly affected by marketing activities before the release and because it determines how many screens will be allocated to the movie. Therefore, the distributor predicts the audience number on the opening weekend and develops a marketing strategy in order to gain a competitive advantage over other films screened at the same time. However, as distributors make a prediction based on practitioners' experiences and intuitions, it is difficult to

quantify the reliability of predicted values and deliver the correct marketing strategy.

In this study, we propose a method that predicts the audience number on the opening Saturday using market research data obtained through online and offline surveys and basic attributes of films extracted from 325 Korean and 296 U.S. movies since 2010.

The test was carried out using four types of datasets divided by prediction time points, and the Decision Tree algorithm proved excellent not only for providing an explanation of the prediction but also in learning performance. As the opening date approaches, the accuracy of the prediction increases and the variables associated with the time close to the opening date are selected as key variables. The market research data obtained from online surveys was more influential than the that from offline surveys, as more online market research variables were included in the top-ranked list of influential variables. Furthermore, closer to the opening date, online market research data significantly impacts the audience number forecast for the opening Saturday. In the absence of survey data, the normalized total production cost, actor's power, grade, and seasonality were selected as the main factors. However, in datasets where survey data was included, market research data such as preference and awareness were selected as the main factors so that public interest in directors and actors was reflected in the survey data.

The significance of our study is as follows. First, we defined the derived variable to reflect the influence of the director and actor. This method provides more reasonable explanation than previous studies, which rely only on the performances of previous movies. In addition, the power of director and actor is selected as the most important factor in dataset D, which, similar to other research, has no survey

data. Second, we establish a model for analyzing factors that have an effect on box office attendance. The data used in the model involved online and offline data from a 7-year period, obtained from distributors on films released in Korea, and we built a database directly collected from the Korean Film Council, *NAVER* movies. Based on this database, further studies could be easily performed. Third, the proposed model can change the decision-making process regarding the distributor's investment. With regard to marketing, through a trained model based on the time-to-release, a strategy could proceed with the goal of increasing the lower-than-standard variable if the predicted class is lower than the target class. Therefore, the distributor can implement an efficient budget considering the relationship between marketing cost and the predicted box office result. In terms of local film distribution, before production in the case of Korean movies, it is possible to determine whether the film will be successful or not by using a variable related to the power of directors and actors. In the distribution of movies produced abroad, it is possible to broadly predict the degree of success at the local box office. Previously, these types of decisions have been determined by the experience and intuition of practitioners, but with our model it is possible to make objective decisions through various simulations. Finally, the three problems mentioned in the introduction can also be solved. This study showed that the accuracy of a prediction could be gradually improved because a decision from a trained model is not only objective but also becomes more accurate as data are accumulated. In addition, it is possible to perform additional analyses on the accuracy of the prediction and organize the marketing strategy systematically.

Chapter 3

Prediction of TV program ratings with Decision Trees

3.1 Background

Ratings, recorded by people meters attached to TV sets, serve as an essential standard measure for television producers and broadcasters [59]. The essentiality of ratings stems from the following two reasons: first, ratings are proportionally related to the program sales because, from the perspective of the advertisers, audience size reflects the range of exposure as their product of interest airs through the TV program. Second, ratings play a significant role in determining the price for exporting programs. Program exports take primarily two forms: (1) an ex-ante export sale, which takes place before the first airing of the program on television, or (2) an ex-post export sale after the program finale. In ex-post export sales, a TV program with higher ratings is perceived as a domestic success, resulting in a higher export sales price. On the contrary, low ratings may can down the export sales prices for the respective TV programs [66].

Due to the aforementioned reasons, ratings are a highly significant measure of popularity, and broadcasters strive to achieve higher ratings by attracting a greater number of viewers. There are primarily two different strategies broadcasters may take to increase audience size: (1) observing and unveiling viewer preferences to pro-

vide new programs that satisfy viewers' tastes or (2) switching airtimes among a select group of TV programs to provide the target viewers with more appropriately matched programs. According to Shim [125], broadcasters attempt to increase audience size by switching the airtime slots of their TV productions, assuming that viewers may land on a program and watch it not entirely by choice but also by coincidence. Given that a considerable portion of program views results from coincidence, broadcasters schedule programs to maximize exposure to potential viewers [99]. As witnessed in numerous cases, a mere rescheduling of TV program airtimes may increase the audience size without incurring additional production costs. In response to market needs, past literature on ratings has branched into two major streams: research regarding the ratings themselves and studies focusing on the strategies to maximize ratings. In research on the ratings themselves, topics such as viewing data collection methodologies or factor analysis via ratings prediction models have been discussed in various ways. Broadcasting programming strategies, however, have rarely been studied, and only limited discussion has occurred regarding the genre distribution apparent in currently adopted airtime schedules.

It has been conventional in the industry that the decision-maker chooses a broadcasting programming scenario out of a few scenarios built according to the broadcaster's internal scheduling strategies, based on their experience and intuition. In this study, we suggest a scientific model through which one may evaluate a set of given broadcasting programming scenarios and choose one with the highest expected ratings, hence leading to an objective and robust selection of an optimal strategy. Moreover, by continuously updating the data and upgrading the suggested model, one may be expected to construct a more advanced broadcasting programming de-

cision system. Our research utilizes ratings log data ranging from July 2016 through October 2017 (16 months in total) provided by Nielsen Korea to build a ratings prediction model with competitive accuracy. Furthermore, we suggest a framework to help a decision-maker choose an optimal scenario out of a set of given broadcasting programming scenarios by exploiting the result of our prediction model to assure the maximum audience size.

The rest of the chapter is organized as follows. In Section 3.2, we introduce past literature on the topic and point out our academic contributions. Section 3.3 elaborates on the primary factors of analysis and the models employed in this study. In Section 3.4, we discuss the results of the experiments, reporting the ratings prediction performance and analyzing the effect of airtime changes on the audience size. Section 3.5 summarizes this chapter by summarizing its findings.

3.2 Related work

Past literature primarily involved approaches to directly examine the ratings themselves, while another major group of research focused on optimal broadcasting programming strategies.

3.2.1 Research on the ratings themselves

A vast volume of studies was devoted to methodologies for collecting viewer counts. Cho [19] criticized the current circumstances and problems regarding the data collection strategies for viewer counts while studying the approaches taken in countries of interest to measure audience size for a given TV program. Cho et al. [20] further extended his research in the scope of constructing a viewer count verification system based on in-depth interviews with a group of industry professionals. Hwang

[53] gave an overview of emerging viewing habits and future data collection methods while examining the main issues concerning the quantization of domestic viewing habits. There exists a myriad of other similar studies that discuss the viewing count collection system or the legal or political issues involved, yet few have successfully proposed a new metric to replace the existing ratings.

On the other hand, researchers have analyzed factors influencing viewing counts using prediction models. Regression or decision tree methods were employed to ensure interpretability [26, 4, 86, 87, 68, 85, 23, 102]. Choi [26], in particular, relied on regression analysis as he discovered that when a program was located between a set of specific programs, these back-to-back programs had a significant adjacency effect on its viewing count. Research has predominantly focused on drama because the popularity or the social effect a particular drama conveys on the public plays a central role in increasing the competitiveness of the subject broadcaster [86, 145, 23, 4]. At the same time, it can be easily inferred that drama has a notably higher chance than other genres to secure loyal audiences due to contextual continuity or viewing habits. Researchers have reported that "star" actors in the cast [86], the number of episodes aired, connections among human factors, or a "star" screenwriter are factors that influence the ratings [145]. Furthermore, Lee [85] showed that Internet protocol televisions (IPTVs) had a negative effect on terrestrial broadcasting real-time view counts. Myung et al. [102] used view count data from web portal live coverage and a decision-tree approach to analyze factors leading to watching Korean soccer league game broadcasts. This work contributed to the literature by considering in-game factors such as the total score, the difference in the final score, and ball possession ratios, which may be compared to the contents of a television program.

Most of the existing literature includes factors cumulatively based on results from prior studies in their model to reflect the basic and contextual attributes of the programs when examining the significance of the influential factors. Our work, in contrast, assumes that the program's basic and contextual characteristics are already reflected in the initial ratings of the program. With this underlying assumption, we effectively utilize environmental variables relevant to the program airtime scheduling as we directly predict future ratings based on the prior viewing counts.

3.2.2 Research on broadcasting programming

Unlike research on ratings, research on broadcasting programming has not been as active. Only a few studies have proposed candidates for airtime scheduling policies based on the genre distribution of select broadcasters' programming. Most of these studies made proposals for broadcasting programming of domestic media shows produced by public broadcasters using airtime scheduling policies taken from global broadcasting companies as the basic reference group [79, 60, 46]. Media with advanced broadcasting systems such as BBC or NHK were the usual choice for the baseline models, from which suggestions for potential improvement were derived in light of the comparative analysis of broadcasting programming, the number of programs by program type, airtime patterns by time slot, and other characteristics. Research on broadcasting programming problems, once limited to public broadcasters, extended to the generalist channels. Cho [18], Kang & Eun [58], and Shim [125] have delved into the broadcasting programming strategies of terrestrial television, while Choi [25] focused on the generalist channels' airtime scheduling strategies as well as the diversity of program genres. Although the scope of the research has expanded from public broadcasting to the generalist channels, discussions of broadcasting pro-

gramming of public broadcasters' programs were limited mostly to ex-post analysis [105].

On the contrary, our work contributes to the existing literature by assisting hands-on workers with model-based airtime scheduling, hence adding objectivity and consistency during the decision-making process, which was not done in the past when broadcasting programming was based on personal experiences and intuition.

3.3 Predictive model construction

3.3.1 Target variable

Nielsen publishes the number of ratings calculated as the average number of viewers flowing in and out of one minute through Arianna, a self-developed audience rating analysis program, but it is not disclosed. Therefore, in this study, after preprocessing the ratings log data, the sum of the weights of the viewers for each program was calculated to replace the ratings. The weights are values obtained by assigning statistical weights to members of each panel by selecting panels by region, number of TV units, number of families, monthly income, subscription by platform, gender, and age based on the basic surveys. In the end, the ratings are the sum of these weights divided by the total viewer population. To understand the correlation between the sum of the weights through preprocessing and the actual audience rating on Arianna, one of the major domestic broadcasters provided data on 6,414 program viewers calculated through Arianna for 330 programs aired from January to October 2017. The correlation between them was 0.993, confirming that they had a very strong positive linear relationship.

In this study, we solve a multiclass classification problem in which the model

predicts the ratings class, binned by 0.2 million viewers in each of seven classes, for each episode of the given television program. Each episode's class serves as the target variable of our model, weighted by the demographic information of the viewers of the relevant program.

The range of 0.2 million was chosen as advised by the current broadcasting producers. These domain experts indicated that the broadcasting staff in charge of scheduling program airtimes finds it easier to work with bins rather than the exact number of viewing counts, and it responds more quickly. Moreover, such an approach enables computation of the probability of a given program being in a particular class relative to other classes, which can be used to deduce the level of confidence in judgment during the decision-making process. Table 3.1 reports class descriptions, the number of television programs by class, and the corresponding proportion. From Table 3.1, it can be easily seen that most programs fall in Class 0. Such an imbalance in data can cause the model to learn to predict the major class only. To address this class imbalance issue, we use the synthetic minority oversampling technique (SMOTE) approach suggested by Chawla et al. [15] before learning the model, where we oversample data for each respective minority class to even out the class ratio in the resulting data.

3.3.2 Predictor variable

As illustrated in Figure 3.1, the predictor variables consist of three types of attributes of a given television program: basic information, ratings by gender and age group, and the television-viewing environment.

First, a given program's basic information includes general characteristics such as genre, day of broadcasting, a flag for terrestrial broadcasting, or other relevant

Table 3.1: Target class description

Class	Class Description	Count	Rate
0	x<200,000	260,786	0.91
1	200,000<=x<400,000	11,590	0.04
2	400,000<=x<600,000	5,033	0.02
3	600,000<=x<800,000	3,350	0.01
4	800,000<=x<1,000,000	2,609	0.01
5	1,000,000<=x<1,200,000	2,127	0.01
6	1,200,000<=x	1,340	0.00
Total		**286,835**	**1**

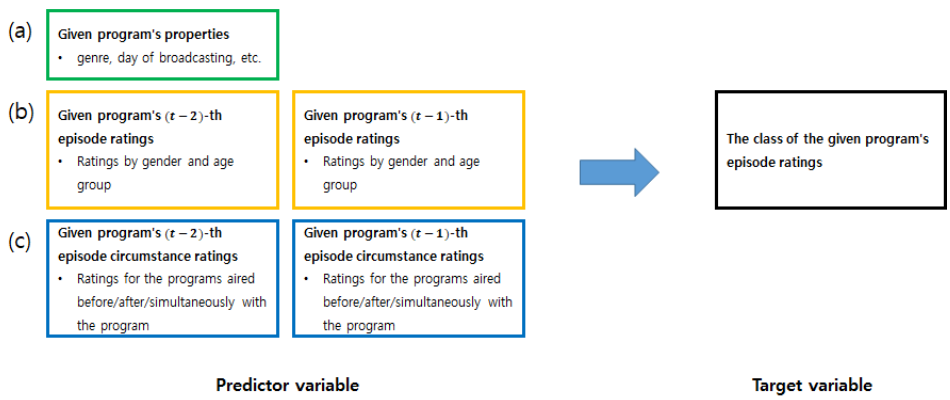

Figure 3.1: Predictor variables and target variable (a) properties (green box), (b) given episode ratings information by gender and age group (yellow boxes), and (c) given episode ratings information for the programs aired before, simultaneously with, and after the airtime (blue boxes)

information. Ratings differ depending on whether the program was aired on a weekday or during the weekend, and if aired during the week, whether it was on Friday or another weekday. The target audience can differ vastly by genre or by terrestrial versus Internet broadcasting; hence, these types of information were also considered basic attributes of a program. Second, we use view counts by gender and age group from the past two episodes as the subject program's ratings attribute. Given that the objective of this study is to build a decision-making system for broadcasting

programming, we aim to predict the following week's ratings at the program level by using the past two episodes' view count information. Here, we assume that the information about the director(s), actors, and screenwriter(s) is already reflected in the past view counts because the ratings already contain basic attribute information such as the airtime and broadcast channel [29, 39]. Third, we include view counts for the programs broadcast before, simultaneously with, and after the airtime of a given program to reflect the effects of adjacency or competition arising from broadcasting programming [126]. The adjacency effect is an intra-channel factor that represents the impact the programs broadcast immediately before and after on the same channel have on the subject program. That is, through the adjacency effect, one may take into account viewers who watch a program, do not exit, and stay on the same channel to watch the following program. In contrast, the competition effect is an inter-channel factor that reflects the information about the competition for a given program. For example, if a program is scheduled to be broadcast simultaneously with a program with high view counts and a large base of loyal viewers, such as *Running Man*, and it competes for the same target audience, then its ratings are very unlikely to be high.

We provide the complete list of predictor variables in Table 3.2.

3.3.3 Prediction Model

The objective of this study is to learn a meticulous model to predict a program's ratings next week based on the past view count information, the results of which may assist hands-on staff with the decision-making process of broadcasting programming to maximize the audience size. Therefore, it is essential to ensure that the model predicts the following week's view count as accurately as possible. Therefore, we

Table 3.2: Predictor variable description

No.	Variable description
1	ratings of Male of $(t-1)$-th episode
2	ratings of the people between the ages of 0 and 9 of $(t-1)$-th episode
3	ratings of the people between the ages of 10 and 19 of $(t-1)$-th episode
4	ratings of the people between the ages of 20 and 29 of $(t-1)$-th episode
5	ratings of the people between the ages of 30 and 39 of $(t-1)$-th episode
6	ratings of the people between the ages of 40 and 49 of $(t-1)$-th episode
7	ratings of the people between the ages of 50 and 59 of $(t-1)$-th episode
8	ratings of the people between the ages of 60 and 69 of $(t-1)$-th episode
9	ratings of the people between the ages of 70 and 79 of $(t-1)$-th episode
10	ratings of the people between the ages of 80 and 89 of $(t-1)$-th episode
11	ratings of the TV program being aired before the program of $(t-1)$-th episode
12	ratings of the TV program being aired after the program of $(t-1)$-th episode
13	the first highest ratings of the TV program being aired simultaneously with the program of $(t-1)$-th episode
14	the second highest ratings of the TV program being aired simultaneously with the program of $(t-1)$-th episode
15	the third highest ratings of the TV program being aired simultaneously with the program of $(t-1)$-th episode
16	the fourth highest ratings of the TV program being aired simultaneously with the program of $(t-1)$-th episode
17	the fifth highest ratings of the TV program being aired simultaneously with the program of $(t-1)$-th episode
18	ratings of Male of $(t-2)$-th episode
19	ratings of the people between the ages of 0 and 9 of $(t-2)$-th episode
20	ratings of the people between the ages of 10 and 19 of $(t-2)$-th episode
21	ratings of the people between the ages of 20 and 29 of $(t-2)$-th episode
22	ratings of the people between the ages of 30 and 39 of $(t-2)$-th episode
23	ratings of the people between the ages of 40 and 49 of $(t-2)$-th episode
24	ratings of the people between the ages of 50 and 59 of $(t-2)$-th episode
25	ratings of the people between the ages of 60 and 69 of $(t-2)$-th episode
26	ratings of the people between the ages of 70 and 79 of $(t-2)$-th episode
27	ratings of the people between the ages of 80 and 89 of $(t-2)$-thepisode
28	ratings of the TV program being aired before the program of $(t-2)$-th episode
29	ratings of the TV program being aired after the program of $(t-2)$-th episode
30	the first highest ratings of the TV program being aired simultaneously with the program of $(t-2)$-th episode
31	the second highest ratings of the TV program being aired simultaneously with the program of $(t-2)$-th episode
32	the third highest ratings of the TV program being aired simultaneously with the program of $(t-2)$-th episode
33	the fourth highest ratings of the TV program being aired simultaneously with the program of $(t-2)$-th episode
34	the fifth highest ratings of the TV program being aired simultaneously with the program of $(t-2)$-th episode
35	whether broadcasting on Monday or not
36	whether broadcasting on Tuesday or not
37	whether broadcasting on Wednesday or not
38	whether broadcasting on Thursday or not
39	whether broadcasting on Friday or not
40	whether broadcasting on Saturday or not
41	whether broadcasting on Sunday or not
42	whether the genre of the program is education or not
43	whether the genre of the program is etc. or not
44	whether the genre of the program is drama or not
45	whether the genre of the program is sports or not
46	whether the genre of the program is kids or not
47	whether the genre of the program is entertainment or not
48	whether the channel airing program is major or not

let our model learn from various techniques, such as multiple logistic regression, XgBoost [16], random forest [10], and support vector machine [30], and we chose the one with the greatest accuracy to build the final learning system. Hyperparameters for each model were determined by selecting those with the best performance through 5-fold cross-validation.

3.4 Prediction model evaluation

3.4.1 Data

Our data, provided by Neilsen Korea, includes historical view count records for the 16 months from July 2016 through October 2017. Neilsen Korea uses the people meter approach to collect ratings data, beginning by selecting sample households based on the number of household members, number of televisions owned, and gender and age distribution. A device called the "people meter" is then installed on the television sets of the selected sample households. Each member of the sample household is assigned a unique identifier, and as household members record their viewing behavior on the people meter via a remote control, actions and attributes such as switching channels or changing viewers are labeled with the relevant household member's identifier. In addition to a unique identifier, every sample household member is given a population-representative weight. For example, Table 3 shows that member ID ad of household ID A has a weight of 10424.2, which is equivalent to stating that ad represents 10424.2 viewers of the same age and occupation group. Table 3.3 presents sample ratings data collected by the people meter. We replaced the household ID with an arbitrary combination of capitalized characters to protect personal information.

Because this study aims to predict the classes of TV program view counts, it

Table 3.3: Sample ratings data

Row ID	52263	6398	11245	50930	10439
Date	20160701	20160701	20160701	20160701	20160701
Household ID	A	B	C	D	E
Member ID	ad	aa	ad	ab	ac
Weight	10424.2	1920.6	7684.4	40187	2520.9
Gender	1	1	1	2	1
Age	14	49	5	39	34
Job	7	1	9	6	2
Education	4	9	1	9	9
Income	7	8	8	4	8
Channel	3	3	4	3	18
Watching Start Time	140	140	180	120	180
Watching End Time	139	159	199	139	219
Watching Duration	2126	60	780	1962	5575
Program Title	My Mind's Flower Rain	Live Info Box	MBC NEWSDESK	My Mind's Flower Rain	Show me the money season 5
Program Start Time	156	141	170	156	190
Program End Time	121	143	219	121	204
Genre	099099099	099099099	099099099	099099099	099099099

is necessary to collapse the raw data, which are collected at the individual level, to the program level. The final collected variables are shown in Table 3.2 in Section 3.3.2. Moreover, while some viewers may intend to watch a program, it is also quite common for viewers without any clear viewing objective to land on the most appealing program after exploring a handful of others. To exclude such cases from training, we discarded data if the viewing time was less than 50% of the entire program runtime. The final ratings per program were then computed by summing across the population-representative age- and gender-weighted viewing counts for each corresponding program. Furthermore, we used one-hot encoding for discrete variables such as weekdays and channels.

3.4.2 Experimental results

We randomly split our data 2:1 to create the training and test datasets, respectively. As can easily be seen from Table 3.1, the data are extremely imbalanced toward class 0, so we employed the SMOTE method to the training set by oversampling the data of the minority classes to balance the class ratios. The resulting training set was then used to train several methods, including multiple logistic regression, XgBoost,

random forest, and support vector machine, with various hyperparameter settings. We relied on 5-fold cross-validation to choose the hyperparameters with the best performance. Table 3.4 reports the prediction accuracies by model. From the four reported, we chose XgBoost, the one with the highest accuracy, as the learning vehicle in our broadcasting programming analysis. The confusion matrix resulting from the test data set using XgBoost is reported in Table 3.5.

Table 3.4: Accuracy of each model

Model	**Accuracy**
Multiple Logistic Regression	0.788
XgBoost	**0.961**
Random Forest	0.960
AdaBoost	0.942

Because SMOTE was used to alleviate the imbalance between classes, considering that the model was to solve a multiclass classification problem with seven classes, the 96.1% accuracy implies that the model was not biased toward the majority class, hence being well generalized. At the same time, its high accuracy plays in favor of our model as the appropriate choice for broadcasting programming strategy development. Table 3.6 reports the prediction results of some of the TV programs actually aired in Korea.

Table 3.5: Confusion matrix

		Actual Class							**Total**
		0	**1**	**2**	**3**	**4**	**5**	**6**	
Predicted Class	**0**	**86,137**	1,243	15	3	1	0	0	87,399
	1	442	**2,529**	438	15	2	1	1	3,428
	2	6	329	**1,059**	202	22	2	0	1,620
	3	0	26	289	**468**	137	10	2	932
	4	1	3	26	157	**240**	56	18	501
	5	0	0	5	14	82	**82**	56	239
	6	0	0	0	4	21	51	**461**	537
Total		86,586	4,130	1,832	863	505	202	538	94,656

Table 3.6: Prediction results for each program

Date	Program Title	Genre	Start Time	Class	
				Actual	Predicted
20161110	*Viewers Column*	Documentary	13:04:50	0	1
20161204	*At Sunday Night*	Entertainment	18:41:42	5	5
20170119	*MBC NEWS(05:00)*	News	05:07:25	0	0
20170222	*MBC NEWSDESK*	News	20:50:42	1	1
20170410	*Fire Engine, Ray*	Animation	09:28:19	0	0
20170428	*Screening Humanity*	Documentary	08:23:45	4	4
20170605	*Cultures*	Documentary	13:47:00	0	0
20170724	*MBC NEWSDESK*	News	20:50:12	2	2
20170805	*Sister is Alive*	Drama	23:05:23	6	6
20170912	*Teacher Oh Soon Nam*	Drama	08:24:42	4	4

In addition, the interpretability of the trained XgBoost model is excellent because the importance of each variable can be checked, as this is a tree-based model. For the XgBoost model, assuming the predicted value of the i-th instance in the t-th iteration is $\hat{y}_i^{(t-1)}$, the objective is as follows:

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \\ &\simeq \sum_{i=1}^{n} \left[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \\ \text{where } \Omega(f) &= \gamma T + \frac{1}{2}\lambda \left\| w \right\|^2, \\ g_i &= \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), \\ h_i &= \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \end{aligned} \tag{3.1}$$

If the constant term is removed from Equation 3.1 and $I_j = \{i | q(\mathbf{x}_i) = j\}$, we can rewrite Equation 3.1 by expanding Ω as follows:

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t(\mathbf{x}_i) + \frac{1}{2}h_i f_t^2(\mathbf{x}_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \\ &= \sum_{i=1}^{T}\left[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2\right] + \gamma T\end{aligned} \tag{3.2}$$

The gain generated for instance i when splitting due to a specific variable in the tree through the weight w_j^* of the fixed structure $q(\mathbf{x})$ is as follows:

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{3.3}$$

The importance of each variable is the average of the L_{split} generated by each variable in all trees. The top 10 variables in the trained model are shown in Table 3.7. It was reported that ratings of one or two past episodes, collected immediately before and after the episodes aired, played a significant role. This supports the findings from past literature that one of the most important factors on the view count is the ratings of the TV programs being aired before, after, and simultaneously with the program of interest. The result also shows that when predicting the next episode's ratings, measuring from one episode back is more significant than two episodes.

3.5 Optimization strategy using the predictive model

To optimize the strategy for a program using the trained prediction model, the predicted values when the program is and is not changed are first compared. Then, if there is a high probability of belonging to a higher class when the program is changed, the characteristics of the program to be changed and the distribution of viewers in the desired period will generate synergy, so the program should be changed. In the past,

Table 3.7: Ten most important variables of the XgBoost model

Rank	Variable name	Importance
1	ratings of the TV program being aired after the program of $(t-1)$-th episode	0.040
2	ratings of the TV program being aired before the program of $(t-1)$-th episode	0.039
3	the first highest ratings of the TV program being aired simultaneously with the program of $(t-1)$-th episode	0.039
4	ratings of the TV program being aired after the program of $(t-2)$-th episode	0.037
5	ratings of the people between the ages of 10 and 19 of $(t-1)$-th episode	0.036
6	ratings of the TV program being aired before the program of $(t-2)$-th episode	0.036
7	the first highest ratings of the TV program being aired simultaneously with the program of $(t-2)$-th episode	0.034
8	ratings of the people between the ages of 50 and 59 of $(t-1)$-th episode	0.033
9	ratings of the people between the ages of 30 and 39 of $(t-1)$-th episode	0.032
10	ratings of the people between the ages of 40 and 49 of $(t-1)$-th episode	0.032

when changing the schedule, one scenario was selected after evaluating each scenario based on the experience and sense of the practitioners from among the schedule change scenarios prepared according to the internal strategy of each broadcaster. However, after evaluating each schedule change scenario using the trained model, we can optimize the schedule strategy by selecting the scenario with the greatest increase in viewer ratings.

Circumstances may vary, but there are primarily two different cases for changes in broadcasting programming: when a new program is being launched following the finale of the preceding program, and when the airtime of an existing program is being switched with another existing program in an attempt to increase the audience size for one or both programs. Indeed, these two cases may complement each other so that the airtimes of existing programs are shuffled as the new program is being launched. Because our model predicts future ratings using the past t times view count information, if the past data is unavailable at the time of learning, it is impossible to make a prediction. Hence, we narrow the focus of our study to assist decision-makers with the latter case where the airtimes of two existing programs are switched.

In Section 3.5.2, the proposed optimization strategy is applied to two cases in which the broadcasts were actually changed.

3.5.1 Broadcasting programming change process

The broadcasting programming change process takes place in anticipation of the synergistic effect between the characteristics of the program to be changed and the distribution of the potential viewers at the target airtime after the change. For example, if a particular program A comprises contents appealing to relatively younger audiences, and if the program's past view count data show that a younger age group makes up the largest proportion of the counts, then it may be effective to move the program to a time slot for which the expected ratio or count of younger viewers is higher. Ultimately, broadcasting programming is successful when the program's specific characteristics and the attributes peculiar to the given airtime are appropriately matched. To incorporate such a perspective in our prediction model, we modify the data processing and prediction model as illustrated in Figure 3.1 to include airtime information when predicting ratings, as depicted in Figure 3.2. Specifically, if we were to change program A to be broadcast at the airtime of program B, we train our model with the basic information and ratings information of program A as is, but we replace the adjacency and competition variables of program A with those of program B. If the ratings using the new data are predicted to be higher than the existing program schedule, then changing the airtime may be a fruitful choice.

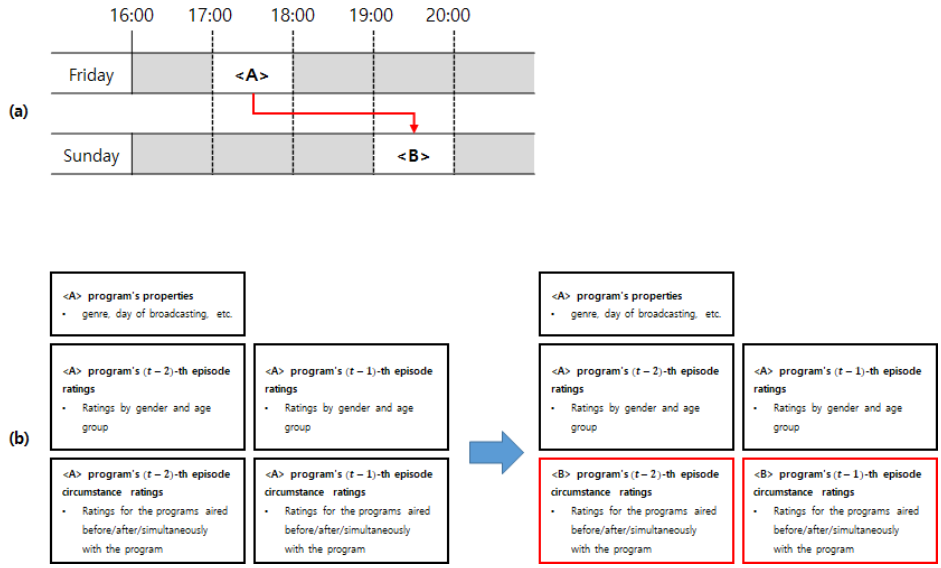

Figure 3.2: Data modification for broadcasting programming change process: (a) airtime slots (b) data

3.5.2 Case Study

All Broadcasting in the World* and *Wizard Of Nowhere

All Broadcasting in the World and *Wizard of Nowhere* are entertainment/reality shows aired on MBC, a public broadcasting company in Korea. As shown in Table 3.8, the airtime of *All Broadcasting in the World*, which used to be aired at 6:45 PM on Sundays, was swapped with *Wizard of Nowhere* to take over the 11:15 PM slot on Saturdays.

We computed the class probabilities of the expected ratings of the two programs before and after the schedule swap using the model as described in Section 3.4, of which the results are reported in Table 3.9. For *The Wizard of Nowhere*, whose airtime was moved from Saturday night to Sunday evening, the probability of reaching 0.4–0.6 million viewers increased. In contrast, our model predicted that for *All*

Table 3.8: Changed airtime case: *All Broadcasting in the World* and *Wizard of Nowhere*

Program Title	State	Airtime	
		Before	After
Wizard Of Nowhere	Changed	Saturday (23:15)	Sunday (18:45)
All Broadcasting in the World	Changed	Sunday (18:45)	Saturday (23:15)

Broadcasting in the World, which was swapped from Sunday evening to Saturday night, the probability of reaching 0.2–0.4 million viewers rather than 0.4–0.6 million would increase. In reality, the airtime swap did not result in much change in the ratings for *All Broadcasting in the World*, while the overall ratings dropped for *Wizard of Nowhere* after the reschedule. Consequently, the prediction and the reality altogether imply that the reschedule cannot be considered a wise decision.

Table 3.9: Class probabilities of *All Broadcasting in the World* and *Wizard of Nowhere*

(a) *Wizard Of Nowhere*

Class	Probability	
	before	after
Class 0	0.001	0.001
Class 1	0.149	0.015
Class 2	**0.662**	**0.976**
Class 3	0.178	0.007
Class 4	0.008	0.001
Class 5	0.001	0.000
Class 6	0.001	0.000

(b) *All Broadcasting In The World*

Class	Probability	
	before	after
Class 0	0.001	0.000
Class 1	0.049	**0.792**
Class 2	**0.749**	0.205
Class 3	0.192	0.003
Class 4	0.006	0.000
Class 5	0.001	0.000
Class 6	0.001	0.000

Baek Jong Won's Top3 Chef King* and *My Little Old Boy

Baek Jong-Won's Top 3 Chef Kings and *My Little Old Boy* are the two very popular entertainment/reality shows aired on SBS, another public broadcasting company in

Korea. *My Little Old Boy* took over the airtime slot of *The K-Pop Star: The Last Chance*, which aired from January through April of 2017, on its conclusion. *Baek Jong-Won's Top 3 Chef Kings* took over the airtime slot of My Little Old Boy, as indicated in Table 3.10.

Table 3.10: Changed airtime case of *Baek Jong-Won's Top 3 Chef King* and *My Little Old Boy*

Program Title	**State**	**Airtime**	
		Before	**After**
My Little Old Boy	Changed	Friday (23:20)	Sunday (21:15)
Baek Jong Won's Top3 Chef King	Changed	Saturday (18:10)	Friday (23:20)
The K-Pop Star: The Last Chance	Ended	Sunday (21:15)	

Again, we predicted the expected class probabilities of the ratings of the two programs before and after the airtime swaps, the results of which are reported in Table 3.11. In the case of *Baek Jong-Won's Top 3 Chef Kings*, whose airtime was switched from Saturday night to Friday night, the probability of attaining 0.4–0.6 million viewers was still predicted to be high, while that of reaching 0.2–0.4 million viewers increased sharply from 15.7% to 36.3% after the swap. For *My Little Old Boy*, which was moved from Friday night to Sunday night, our model predicted that the program would remain in Class 3 without the airtime change, but with the airtime swap, the probability of being in Class 4 increased from 33.9% to 76.3%. In reality, following the airtime switch, *Baek Jong-Won's Top 3 Chef Kings* recorded a drop in ratings to the class of 0.2–0.4 viewers, while *My Little Old Boy* showed an increase toward attaining more than 1.2 million viewers. However, despite the tiny

drop in the *Baek Jong-Won's Top 3 Chef Kings* ratings, the considerable increase in the *My Little Old Boy* ratings made the decision to switch the airtimes of these two programs potentially appropriate.

Table 3.11: Class probabilities of *Baek Jong Won's Top 3 Chef Kings* and *My Little Old Boy*

(a) *Baek Jong Won's Top 3 Chef Kings*

Class	**Probability**	
	before	**after**
Class 0	0.001	0.001
Class 1	0.157	0.363
Class 2	**0.832**	**0.629**
Class 3	0.008	0.005
Class 4	0.001	0.001
Class 5	0.001	0.001
Class 6	0.001	0.001

(b) *My Little Old Boy*

Class	**Probability**	
	before	**after**
Class 0	0.001	0.001
Class 1	0.001	0.001
Class 2	0.001	0.002
Class 3	**0.628**	0.224
Class 4	0.339	**0.763**
Class 5	0.029	0.009
Class 6	0.001	0.001

3.6 Summary

Ratings are an essential measure widely used in the market, one of the main functions of which is to serve as the standard metric for cutting advertisement expenses. At the same time, because ratings play an important role in the valuation of the subject program as it is being negotiated for media export, the broadcasters strive to attain view counts as high as possible. However, the bulk of the past literature focused on predicting the view counts themselves rather than developing a strategy to maximize the ratings. Moreover, the scope of the research has been limited to discussing or suggesting different distributions of the genres of the programs as the new potential programming policies.

Our study established a prediction model with a high accuracy given the view count data. Our model preprocesses the individual-level raw data to reflect the

following three types of the subject program's attributes: (1) basic information, (2) ratings by gender and age group, and (3) the television-viewing environment.

To attain interpretability in our model, we employed methodologies such as random forest and XgBoost, of which XgBoost was ultimately selected as the final prediction model due to its superior performance. Through the variable significance test, we showed that the adjacency and competition effects were most closely interlaced with the view counts. Furthermore, we conducted two case studies where we simulated broadcasting programming strategy scenarios with two actual program airtime swap cases, including (1) *Broadcasting in the World* and *Wizard of Nowhere* and (2) *Baek Jong-Won's Top 3 Chef Kings* and *My Little Old Boy*.

The industry convention has been that the involved personnel choose a broadcasting programming scenario based on their experience and intuition. Our work, on the other hand, suggests a scientific method to evaluate a set of given broadcasting programming scenarios, through which one may choose the one with the highest expected ratings, hence selecting an objective and robust optimal strategy. In addition, continuously updating the data and upgrading the suggested mode may lead to a more advanced broadcasting programming decision system.

Chapter 4

Relation detection of YouTube channels

4.1 Background

The production and distribution of professionalized information have involved centralized organizations composed of specialized agents with areas of expertise, who have, until recently, transmitted such information to the public through select channels such as televisions, newspapers, or radios [28]. The recent surge of smartphones alongside YouTube, a globally leading online content sharing platform, has overturned the ecology of information distribution. Nowadays, anyone could record a high-resolution video using the smartphone and share his/her creation quite easily through his/her YouTube channel. YouTube has expanded its market share by faithfully playing its role as a fast, convenient, easy-to-use and easy-to-access content distribution channel, and now people are querying and consuming information from YouTube on their smartphones at anywhere at any time needed [24].

Behind the rapid growth of YouTube stands the content recommender system which feeds a personalized selection of videos to its users. As Neal Mohan, the Chief Product Officer of YouTube, has stated that: "The company has said recommendations are responsible for about 70 percent of the total time users spend on the site [119]." YouTube's recommendation algorithm has been built by analyzing the con-

tent consumption patterns of its users. It is designed to raise user satisfaction and ensure they adhere to the platform by continuously feeding contents that matches the user's current interests.

YouTube's recommendation algorithm, originally anticipated to draw out positive effects from both the platform and the users, is now alleged to be one of the factors aggravating various polarization issues observed throughout the modern society [142]. Such a phenomenon is called the "filter bubbles". Here, the "filter" refers to the recommendation algorithm, which may result in confining the subject user in a "bubble" by recommending only the items matching the user's current interests without any exposure to the contents with different or opposite perspectives. That is, in other words, by the design of the recommendation algorithm, it selectively exposes the users to contents fit to their preferences, hence restricting them from accessing and experiencing diverse scopes of opinions, which, eventually, will lead to consolidation of the users' ideological inclinations in a very lopsided way [106].

The filter bubbles incurred by YouTube's recommender system has been raising various issues, globally; yet, it is a particularly more serious concern in Korea, owing largely to the behavioral patterns of content consumption of the domestic users. According to the reports published by the Reuters Institute for the Study of Journalism at University of Oxford, Korean users consume news via YouTube 40% of time, which is 14% higher than the average consumption rate, 26%, for the entire sample of 38 countries [51, 69]. Filter bubbles pose as a grave issue, especially within the scope of news recommendations which require exposure to a diverse set of social perspective. The effect of repeated filter bubbles on the continuous consolidation of parti pris does not only limit to the personal level; once aggregated to the

community level, such phenomena may potentially lead to political polarization to a threateningly grave degree [106].

In our study, we construct a YouTube politics/news channel network from the user comments left on the videos of the select channels, based on the underlying assumption that the users with similar political orientation will leave comments to the channels sharing closely related political stances. We run the k-means clustering algorithm on the resulting network to extract the conservative-oriented and the liberal-oriented groups, by using which we illustrate the landscape of Korean domestic politics/news channels in terms of their political stances. In addition, we define the concept of the isolation score as the difference in the given channel's distance from the centroid of the Conservative group and that from the centroid of the Liberal group. We use this metric to estimate the extent of isolation the users of each given channel face. Higher isolation score does not imply that the subject channel is biased towards a particular political orientation; rather, it implies that the users' consumption of the subject channel contents tend to lie near a particular political orientation.

There has been attempts made in the Korean academia to analyze YouTube's content recommendation algorithm [57, 65, 128, 55, 28, 88, 44, 106]. However, past literature has focused mainly on examining its influence on the public opinion within the scope of political polarization, fake news, or human memory [122].

Assuming that YouTube's recommendation algorithm drives its users into a certain direction by feeding them with candidate channels whose videos are to be watched next, then, the clustering information extracted from the channel network may allow the users an opportunity to evaluate themselves upon the extent of im-

partiality in the political information in consumption by informing them with which political stance the currently recommended channel is associated, as well as where it is located within the network.

The rest of the chapter is organized as follows. In Section 4.2, we introduce past research on YouTube analysis and shed light on the contributions our study make to the existing literature. Section 4.3 elaborates on the overall mechanism of our proposed model and the methodologies involved. We report the experiment results in Section 4.4, about which further discussions are detailed in Section 4.5. Finally, Section 4.6 summarizes the study.

4.2 Related work

A handful of studies has investigated YouTube channels by constructing a channel network and extracting insights from it. Recently in Korea, researches are analyzing YouTube channel networks in relation to political polarization, false information, and/or fake news. Chung et al. [28] examined the process of the spread of false information on YouTube and the pattern of its consumption by analyzing topics, wordclouds, network centrality measures, and the network of recommended channels. Lee et al. [78] chooses the infamous food-related fake news case of "onions inserted to the earhole mitigates the ear pain" to conduct the event study and show the who and how of the spread of fake news on YouTube by the means of social network and word co-occurrence network analysis.

On the other hands, attempts have been made to investigate fake news issues within the scope of a specific social problem. Kim et al.[71] and Han [45] have analyzed the primary channels and pipelines through which fake news is distributed by

conducting network analysis on fake videos published in regards to the Gwangju Uprisings. Similarly, Heo [50] has inspected the impartiality of YouTube and television channels on politics by constructing a network of 439 keywords relating to the issues about Corruption Investigation Office for High-Ranking Officials. Kim & Hong [62] compartments YouTube channels into two different classes, individuals and the press, and observed the differences in the channel effect, as well as the magnitude of the shock invoked by the message, as fake news spreads. His study has shown that the individual YouTube channels served as the main agent of the creation and the spread of fake news; in the meantime, the press channels, while screening out fake news by constant information authenticity checks, functions as the messenger broadcasting the fake news generated by individual channels.

Our study resembles past literature in that we conduct network analysis to investigate the inter-channel relations from the channel network. However, our work distinguishes itself from other studies by employing keyword analysis to build a YouTube channel map across the field of politics and news in general and to cluster channels according to their political stances.

Most of the network analysis researches has adopted YouTube channels or video networks in order to solve the task of community detection, topic detection, or graph partitioning. Local partitioning has been the popular choice for clustering a large-scale network; in this case, however, it has been found difficult to obtain clusters optimized for the entire network. Gargi (2001) addresses this issue by running additional pre- and post-processing before and after clustering so as to enhance performance on large-scale networks. Our study may resemble the work of Gargi et al. [41] in a sense that we set the number of clusters k to a large number to segment

data into many micro-clusters and run additional post-processing by relabeling via RWR. However, our post-processing focuses on the structural relations among channels extracted from the micro-clusters, while Gargi et al. [41], in contrast, relied on word occurrences observed from the videos in each of their micro-clusters.

There has been a few studies which exploits deep neural networks to learn and embed the structural information of the network so as to compute node representations. Perozzi et at. [115] and Grover & Leskovec [43] generate many random walk paths and substitutes them with word2vec sentences. The proposed method preserves the similarity information among nodes, as that among words are preserved in the word2vec method, as nodes are embed. On the other hand, Wang et al. [136] and Cao et al. [12] suggested utilizing autoencoder models in order to preserve the local and the global structures when embedding nodes. There exist other studies exploring various methods for embedding networks, yet a standalone model for general use applying to all types of networks has not yet been found, because of which one should search for methodologies serving as the best fit within the scope of research objectives. For this reason, we directly use the similarity matrix to conduct our analysis instead of computing representations via embedding methods.

4.3 Method

In this study, we analyze 23,382,992 user comments collected from 144,799 videos of 539 domestic politics and news channels starting from June 28th of 2007 through June 9th of 2020. From the collected user comments, we construct a YouTube channel network for domestic politics and news. We cluster the resulting network into the conservative channel and the liberal channel groups in order to examine the extent

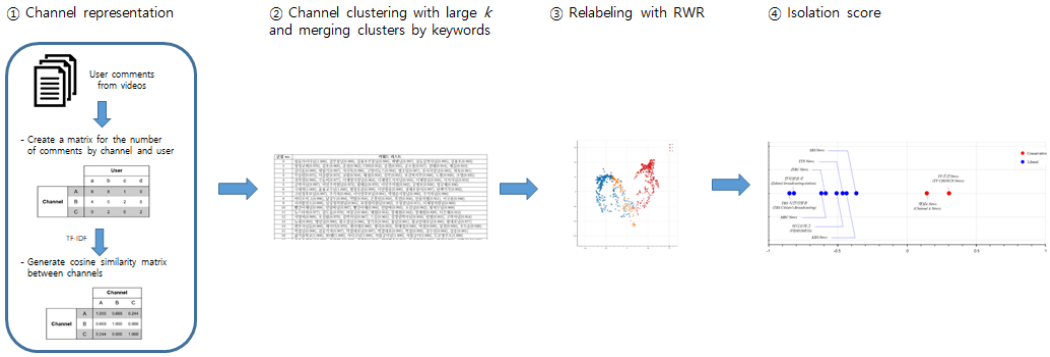

Figure 4.1: A Graphical Illustration of the Overall Process of Our Analysis

of isolation the users of each channel face.

The overall process of our analysis is graphically illustrated in Figure 4.1. The first stage of our analysis represents each YouTube channel as a vector, based on the user comments left on the videos of the respective channel. Then, we run keyword analysis on each of the initial k-clusters, based on the results of which are then agglomerated into a larger groups representing the liberals, the conservatives, and the unclearly defined. In addition, we further precise the clusters through relabeling using the RWR method. Finally, by referring to the conservative/liberal clusters' centroids, we compute the isolation scores channel-by-channel, from which we intend to determine the extent of isolation associated with the users of the recommended channel.

4.3.1 Channel representation

We first compute the channel-user matrix of the collected YouTube channels by vectorizing the channel information via item-based collaborative filtering method, one of the most classical approaches used for recommendation systems. The underlying assumption here is that the users with similar political orientation are more

likely to leave comments to the videos of channels sharing similar political tendencies/implications. We illustrate a toy example on how we vectorize the number of comments each user leaves for the respective channels in panel (a) of Figure 4.2. Here, we present a case in which the user a leaves 8, 4, and 0 comments while the user b leaves 8, 0, 2 comments and the user c and d each leave 1, 2, 0 and 0, 0, 2 comments for YouTube channels A, B, and C, respectively.

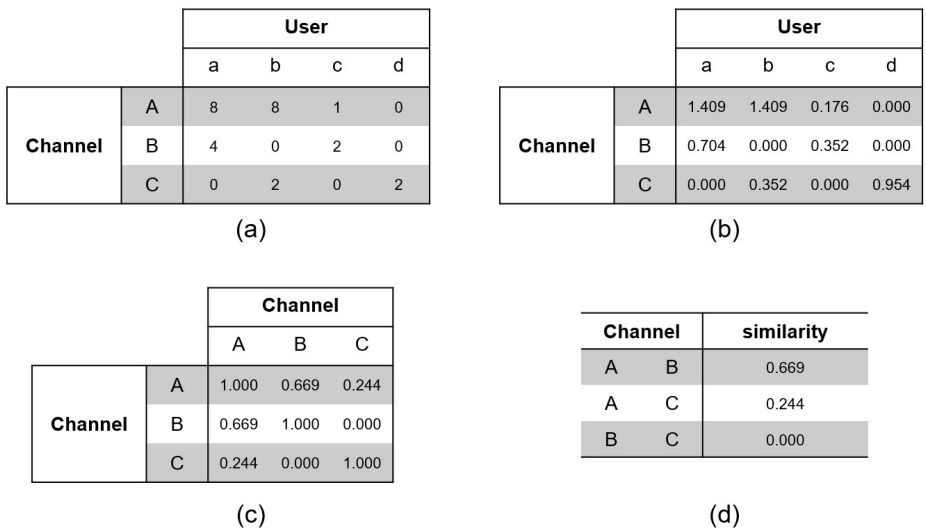

Figure 4.2: An Example of the Channel Similarity Computation, (a) the channel-user matrix; (b) TF-IDF adjusted channel-user matrix; (c) Inter-channel cosine similarity matrix; (d). similarities between a given pair of channels.

We also applied Term Frequency-Inverse Document Frequency (TF-IDF) method to the channel-user matrix. TF-IDF utilizes the term frequency and the inverse document frequency in order to discount the word's significance accordingly, and it is one of the frequent choices for pre-processing the document/word matrix when solving various text mining tasks. By processing the channel-user matrix via TF-IDF, we expect to reduce the importance of the users who leave comments to the

majority of the channels in analysis while placing more weights to the users who leave comments to a small number of a very specific set of channels, as TF-IDF usually does for words frequent throughout the most of the documents. We show an example of TF-IDF processing in Figure 4.2, where TF-IDF is applied to the initial channel-user matrix, as listed in panel (a), to result in the TF-IDF adjusted matrix, as shown in panel (b), reflecting the importance of user comment frequency across different channels.

With each channel representing a node, we define the weight of the YouTube channel network as the similarity between the given pair of channels. Here, we measure the similarity among channels by computing cosine similarity on the TF-IDF adjusted matrix. Cosine similarity puts heavier weight on the case where similar values occur in the same dimension of the given pair of vectors when measuring the distance between them, rather than the mere size of the vectors. Because our analysis assumes that the more similar the two given channels are, the greater the overlap between the pool of common users leaving comments to each channel, we rely on cosine similarity to represent the distance between the given pair of channels. Panel (c) of Figure 4.2 illustrates an example of the cosine similarity matrix calculated from the initial channel-user matrix as presented in panel (a) of the same figure. Equation 4.1 mathematically defines the cosine similarity we employ.

$$similarity = cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4.1}$$

4.3.2 Channel clustering with large k and merging clusters by keywords

We exploit the k-means method, one of the most classic clustering algorithms in practice, in order to cluster the YouTube channels. k-means algorithm clusters the given observations into k clusters based on the k given centroids, of which the mathematical definition is shown by Equation 4.2 [93].

$$argmin_C \sum_{i=1}^{k} \sum_{x \in C_i} |x - c_i|^2 = argmin_C \sum_{i=1}^{k} |C_i| \, Var(C_i)$$

C : cluster index

C_i : cluster i

c_i : the centorid of cluster i

x : a vector of a data point

k : a number of clusters (4.2)

Given k arbitrarily determined centroids, k-means algorithm groups data close to same centroid. Then, based on the clustering results, the positions of the centroids are updated accordingly. k-means algorithm repeats the two processes alternatively, updating the k centroids across all data points by minimizing the intra-cluster distances while maximizing the inter-cluster distances. The computational burden required for the k-means clustering is not as heavy as compared to the other clustering algorithms, and it is known to be quite scalable and to converge stably. Hence, k-means has served as one of the most preferred choices for clustering data.

One of the most crucial part of k-means clustering is to choose the appropriate number of k centroids, which is usually determined based on clustering metrics such

as the Silhouette Coefficient. At the same time, because the k-means algorithm clusters data into k groups in accordance to their distances from the k centroids, if the given data is sparse, it tends to drive data into particular clusters. In order to tackle this shortcoming, we first set a large number for k so that data is grouped into small-sized clusters and run keyword analysis on each of these micro-clusters. As for the keyword analysis, we exploit the proportion-based keyword extractor of the soynlp library[1], developed by [64]. The proportion-based keyword extractor considers the inter-cluster distinguishability as well as each cluster's representativity when selecting keywords. Mathematically, keywords are extracted based on the formula presented by Equation 4.3.

$$s(w, C_i) = \frac{p_i(w)}{p_i(w) + p_{-i}(w)} \tag{4.3}$$

Equation 4.3 computes the distinction score of the word w for the cluster C_i. Here, $p_i(w)$ represents the probability for the word w to appear in the cluster C_i; $p_{-i}(w)$, that for the word w to appear in the clusters of documents other that C_i. If $p_i(w)$ and $p_{-i}(w)$ are close to each other, then the relation of the word w to the cluster C_i is miniscule. If $p_i(w)$ is larger than $p_{-i}(w)$, then the word w is considered to have appeared distinctively in the cluster C_i. By definition, $s(w, C_i)$ ranges between $[0, 1]$, and as it gets close to 1, the word w is said to distinctively represent the cluster C_i. However, when the given word appears extremely rarely, then it is most likely that the word is found in one specific cluster, driving the associated distinction score to 1. In order to address this issue, we first screen out the top k_0 words, given $p_i(w)$. Among these, we select k_1 words by $s(w, C_i)$ as the final keywords. Such an

[1] https://github.com/lovit/soykeyword

approach assures that the selected keywords appear frequently, yet they are found only in the cluster C_i particularly [64].

Finally, we agglomerate the micro-clusters up to three macro-clusters, the Conservative, the Liberal, and the unclearly defined, by examining the micro-cluster keywords.

4.3.3 Relabeling with RWR

At this stage, we apply RWR to each channel in order to determine the number of the channels most significantly associated with given channel from each macro-cluster. The given channel is then re-assigned to the macro-group whose number of the significantly associated channels is the highest. For example, consider an arbitrary channel A, which is currently assigned to the unclearly defined group after the micro-cluster agglomeration. If the application of RWR to the channel A returns 20 channels from the Conservative, 4 from the Liberal, and 1 from the unclearly defined cluster, to be significantly associated, the channel A is then re-labeled to the Conservative cluster. The re-labeling process runs iteratively until no more movement is observed across different clusters (the model convergence). The pseudo-code for the relabeling process is presented in Algorithm 1.

Algorithm 1 Relabeling with RWR pseudo-code

```
1: while no more labels changed do
2:     for channel = 1, 2, ..., N_channels do
3:         similars ← find most similar top-k channels for the channel by RWR
4:         counts ← count similars by clusters
5:         label_new ← set new label with maximum value cluster label of counts
6:     end for
7: end while
```

4.3.4 Isolation score

Upon the completion of the relabeling process follows the computation of the isolation scores by channel. We refer to the isolation score as the extent of the isolation each channel's users face, estimated by the difference of the distance of the given channel from the Conservative cluster's centroid and that from the Liberal cluster's centroid. That is, in other words, the isolation score may be computed at the channel-level. The mathematical definition of the isolation score is given by Equation 4.4. If a given channel is positioned closer to the Conservative cluster, then its isolation score will be greater than 0; in contrast, if it is located closer to the Liberal cluster, then the isolation score will be smaller than 0. If the given channel found where its distance from the Conservative cluster centroid is the same as that from the Liberal cluster's centroid, then the corresponding score will equal zero.

$$
\begin{aligned}
score(x) = \parallel x - c_p \parallel - \parallel x - c_c \parallel \\
x : \text{a vector of a data point} \\
c_p : \text{the centorid of progressive cluster} \\
c_c : \text{the centorid of conservative cluster}
\end{aligned}
\tag{4.4}
$$

4.4 Result

4.4.1 Channel representation

In the experiment, we first vectorized the number of comments the users left for the videos uploaded to the select channels. Then, we further processed the channel-user matrix using via TF-IDF method and computed the cosine similarity on the resulting matrix. We collected videos uploaded to the channels in the area of Korean

politics/news, which can further be compartmented into three groups depending on the type of the channel creator/owner: (1) the press group, such as *헤럴드스토리 (HERALD), 채널A 뉴스 (Channel A News), YTN News, JTBC News, 서울의 소리* (The Voice of Seoul); (2) the government agency group, including *the Ministry of the Interior and Security, National Tax Service*, and *Defense Acquisition Program Administration*, and; (3) the individual or private agency groups, such as *이재정TV (Lee, Jaejung TV), 이봉규TV (Lee, Bonggyu TV), 안중규 TV (An, Joonggyu TV), 사람사는세상노무현재단 (Rohmoohyun Foundation), 가로세로연구소 (Hover Lab), 신의한수 (tlsdmlgkstn17).*

In order to validate that the similarity matrix appropriately represents the proximity among the collected channels, we selected two channels from each of the three macro-clusters and looked up the 5 closest channels and their corresponding similarity scores. Results are reported in Table 4.1. As for the press group channels such as *서울의 소리 (The Voice of Seoul)* and *채널A뉴스 (Channel A News)*, other press channels are detected to be the most similar channels. Similarly, government agency channels are found to be located near the Ministry of the Interior and Security's and National Tax Service's channels. In terms of the individual or private agency channels, *신의한수 (tlsdmlgkstn17)* is widely known for its Liberal orientation, and the cosine similarity matrix returns other Liberal-oriented private/individual channels. At the same time, as for *사람사는세상노무현재단 (Rohmoohyun Foundation)*, a very popular Conservative channel, a handful of other private/individual, Conservative-oriented channels are detected to be similar. These results imply that the similarity matrix proposed by our study well represents the similitude among the channels.

Table 4.1: The list of similar channels, by select channels from each macro-group

No.	Channel name	Group	List of channels
1	*채널A 뉴스 (Channel A News)*	Press	*채널A 뉴스TOP10(Channel A New Top10)*(0.489) , *SBS News*(0.448) , *YTN News*(0.438) , *TV조선 News (TV CHOSUN News)*(0.437) , *비디오머그(VIDEOMUG)*(0.383)
2	*서울의 소리 (The Voice of Seoul)*	Press	*팩트TV News (FactTV News)*(0.486) , *MediaVOP*(0.457) , *뉴스반장 (Head of news)*(0.436) , *시사타파TV (SISATAPA TV))* (0.421) , *알리미 황희두 (Information Deliverer Hwang, Heedoo)* (0.418)
3	*행정안전부* (Ministry of the Interior and Safety)	Government agency	*금융위원회 (Financial Services Commission)*(0.149) , *중소벤처기업부 (Ministry of SMEs and Startups)*(0.148) , *소상공인방송 (Small business TV)*(0.126) , *충청북도 (Chungcheongbuk-do)*(0.122) , *국세청 (National Tax Service)*(0.117)
4	*국세청 (National Tax Service)*	Government agency	*금융위원회 (Financial Services Commission)*(0.281) , *중소벤처기업부 (Ministry of SMEs and Startups)*(0.273) , *소상공인방송 (Small business TV)*(0.271) , *대한민국 통계청 (Statistics Korea)*(0.234) , *충청북도 (Chungcheongbuk-do)*(0.226)
5	*신의한수 (tlsdmlgkstn17)*	Individual or private agency	*이봉규TV (Lee, Bonggyu TV)*(0.562) , *공병호TV (Gong, Byeongho TV)*(0.492) , *[정광용 TV]레지스탕스TV (Jung, Kwangyong's Resistance TV)*(0.478) , *배승희 변호사 (Attorney Bae, Seunghee)*(0.476) , *전옥현 안보정론TV (Jeon, Okhyeon's Security Policy TV)*(0.440)
6	*사람사는세상노무현재단 (Rohmoohyun Foundation)*	Individual or private agency	*딴지방송국 (Ddanzi broadcasting station)(0.526)* , *팩트TV News (FactTV News)*(0.440) , *TBS 시민의방송 (TBS Citizen's Broadcasting)*(0.432) , *저널리즘 토크쇼 J (Journalism talk show J)*(0.427) , *김용민TV (Kim, Yongmin TV)*(0.414)

4.4.2 Channel clustering with large k and merging clusters by keywords

When we conducted k-means clustering on the similarity matrix, we set k to 31 in order to assure that the resulting micro-clusters are as homogeneously grouped as possible in terms of the political stance.

In order to characterize each of micro-clusters, we relied on the proportion-based keyword extractor to determine keywords which incorporates the cluster distinguishability and the representativity. Due to the limitations posed inevitably by the unsupervised nature of soynlp library's tokenizer, keywords that are not nouns for certain, such as *"저딴"* (which may be translated to "like that"), *"저걸"* (similarly, "right that"), are discarded from analysis.

The fastest way to characterize each cluster's political orientation is to look up the political figures from the cluster keyword list. For example, if the keyword list

includes figures on the liberal side, such as Byeongho Gong, Kyujae Jung, Gabje Cho, or Congressmen of the liberal party such as Kyungwon Na, Unju Lee, then the subject cluster is highly likely to be liberal. The same approach applies to the characterization of the Conservative-oriented clusters. In addition, we have observed that the liberal clusters are mainly associated with keywords negatively portraying President Moon and the left-wing, and the frequent keywords include "rigged election", "Communism", "divided people" or word relating to above. On the other hand, in case of the Conservative clusters, frequently appearing are the negative keywords regarding Seokyeol Yoon, the ex-Public Prosecutor General, alongside with words relating to the "press reform", "prosecution reform", "cleaning up old evils", etc. In terms of the clusters within the unclearly defined or heterogenous political orientation group, unlike the previous examples from the Conservative or the Liberal groups, terms relating to social issues such as Islam, displaced people, COVID-19, feminism, as well as life-related topics including slowtech communities, vaccines, and masks, are observed primarily. In fact, we found that such unclearly defined or heterogenous political group is comprised of the integrated news channels such as *KBS Gwangju*, *시사TVCHOSUN (TV CHOSUN on Current Events)*, *YTN News*, and *KBS News*. Based on such characteristics observed from the cluster keywords, we agglomerated the micro-clusters in to the three macro-clusters depending on their political orientation.

4.4.3 Relabeling with RWR

We applied RWR to all channels so as to extract the 20 most significantly associated channels from each of the three macro-clusters. for every given channel on the YouTube channel network. Every given channel was, then, relabeled to the most rel-

evant macro-cluster, based on the results from RWR. In case of the damping factor for RWR, we closely follow the work of [48] and set the parameter at 0.85. Inter-channel movements occurred for three rounds, and all of the movements involved relocation from the unclearly defined group to either the Conservative or to the Liberal group, except for the three channels which moved from the Liberal to the unclearly defined group. Those found in the unclearly defined group usually involved the integrated news channels touching upon various issues across diverse topics, and they were eventually relabeled either to the Conservative or to the Liberal group according to the characteristics of the topics majorly handled.

Figure 4.3 plots the clustering results from the relabeling process via RWR on a 2-dimensional space through T-SNE [134]. T-SNE is a popular choice for embedding the higher-space vectors on to a lower dimensional space. T-SNE attempts to preserve the neighbor proximity among data in higher dimensions, while converting the distance between the given data to stochastic probability for embedding, which leads to better stability as compared to other embedding algorithms for dimensionality reduction. Figure 3 plots the Conservative cluster in red, the Liberal in blue, and the unclearly defined in orange.

As compared to the case when the data are clustered promptly into three groups, we have found it to produce more meaningful clusters when dissecting data groups into k small (in this case, 31) clusters and then agglomerating them to a higher level. In addition, we have observed that the relabeling with RWR approach further refines the clustering results for channels with very specific characteristics such as *채널A 뉴스 (Channel A News)*, *시사TVCHOSUN (TV CHOSUN on Current Events)*, and *뉴데일리TV (New Daily TV)*.

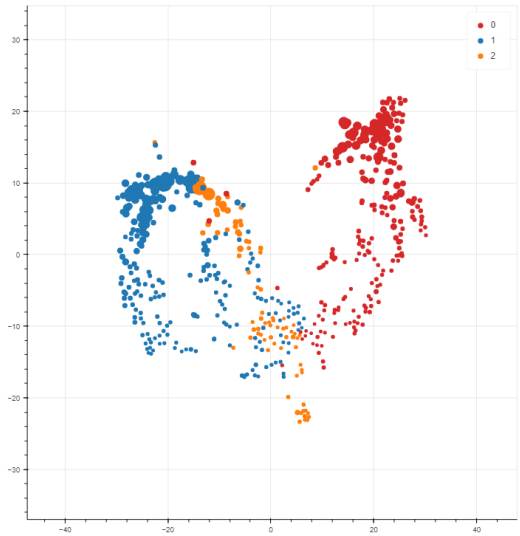

Figure 4.3: Clustering Results from the Relabeling Process Using RWR

Finally, Table 4.2 reports the list of channels by the type of the channel creator/owners – the individual or the private agencies, the press, or the government agencies – and the macro-clusters – the Conservative, the Liberal, or the unclearly defined group.

4.4.4 Isolation score

Figure 4.4 visualizes the isolation scores, as defined by Equation 4.4, computed for the Conservative- and the Liberal-oriented channels by the type of channel creator/owners. Please note that each panel of Figure 4.4 visualizes the top 10 channels with the largest size of subscribers by July, 2021 for each type of channel creator/owners, respectively. As for the government agencies, because almost all of the channels belong to the unclearly defined group, we only present the two channels *대한민국청와대 (The Blue House)* and *법무부TV (Department of Justice TV)* in Figure 4.4. We would like to note at this point that a higher isolation score for a

Table 4.2: List of channels by the type of channel creator/owners and the macro-clusters.

Group	Macro-cluster	List of channels
Individual or private agency	Conservative	*이봉규TV (Lee, Bonggyu TV), 대한민국 청아대 (blueguys1004)* *, 게릴라TV (Guerrilla TV), 안중규TV (An, Joonggyu TV)* *, 일산TV (24LIVE NEWS)*, etc.
	Liberal	*이재정TV (Lee, Jaejung TV), 사람사는세상노무현재단 (Rohmoohyun Foundation)* *, 김용민TV (Kim, Yongmin TV), 언론 알아야 바꾼다 (The media needs to change)* *, [공식] 새날 ([Official] SaeNal)*, etc.
	Unclearly defined	*킴킴변호사 (Attorney KimKim), 하태경TV (Ha, Taekyung TV)* *, 수고요 (Sugoyo), 우리가 알고 싶은 세계 (The world we want to know)* *, 우리동네 우리방송 (Our town our broadcasting)*, etc.
Press	Conservative	*AM1230우리방송 (AM1230 Woori Broadcasting), 채널A News (Channel A News)* *, TV조선 News (TV CHOSUN News), 채널A 뉴스TOP10 (Channel A News Top10)* *, 뉴스타운TV (Newstown TV)*, etc.
	Liberal	*서울의 소리 (The Voice of Seoul), YTN News* *, JTBC News, TBS 시민의방송 (TBS Citizen's Broadcasting)* *, SBS News*, etc.
	Unclearly defined	*헤럴드스토리 (HERALD), 서울신문 (TheSeoulShinmun)* *, 스브스뉴스 (SUBUSU News), 엠빅뉴스 (Mbig News)* *, 14F 일사에프 (14F)*, etc.
Government agency	Conservative	
	Liberal	*경기신용보증재단 (Gyeonggi Credit Guarantee Foundation)* *, 법무부TV (Department of Justice TV)* *, 대한민국청와대 (The Blue House)*
	Unclearly defined	*대한민국 정부 (The government of South Korea)* *, 대한민국 보건복지부 (Ministry of Health and Welfare)* *, 국방부(Ministry of National Defense)* *, 행정안전부 (Ministry of the Interior and Safety)* *, 청와대국민청원 (The Blue House's public petition)*, etc.

given channel does not necessarily indicate that the contents of the corresponding channel are biased towards a particular political orientation. Isolation score simply measures the difference of the distances from different cluster centroids, the Conservative and the Liberal centroids, namely, which potentially imply that the users of the subject channel tend to consume the channel's contents in a particular fashion in terms of political orientation.

4.5 Discussion

4.5.1 On the Representativeness of the Channel Preferences of the Users from Their Comments

In this study, we have established inter-channel similarities by using the number of the comment counts each user left on the videos uploaded to the YouTube channels.

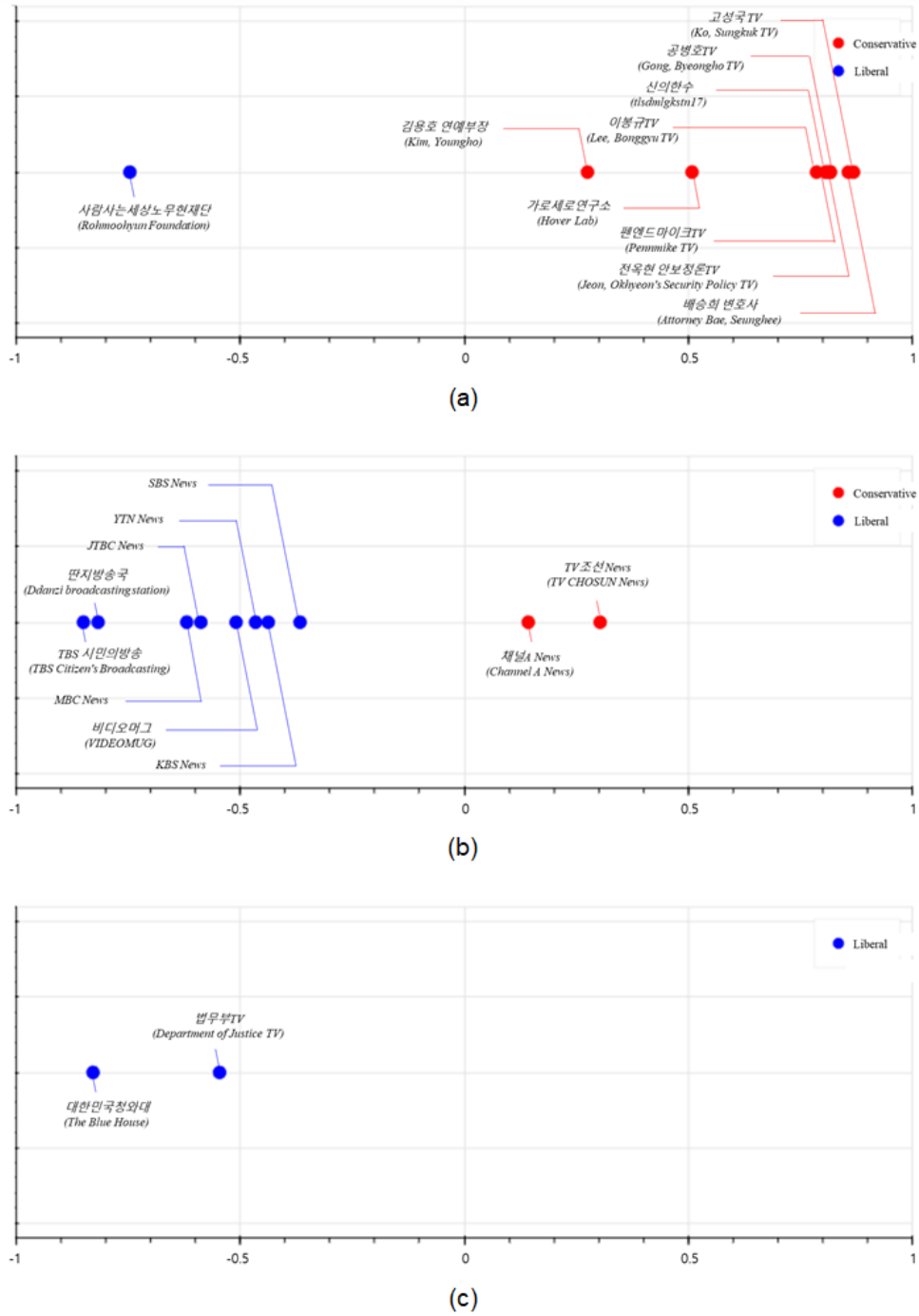

Figure 4.4: Isolation scores for each channel by Macro-cluster, (a) Individual or private agency, (b) Press, (c) Government agency

However, a user may leave comments not only to the channels he/she likes but also to those in dislike. Such a case may result in noisy signals, hence producing misleading relations in terms of channel similarity. We feel very fortunate to be able to report that, after the scrutinizing review of the similar channels for each given channel in the similarity matrix, we could successfully observe that the liberal-side channels share greater similarities with other liberal-oriented channels, while the conservative channels share greater similarities with other conservative channels, from which, altogether, we have concluded that similarity was well represented in the similarity matrix.

4.5.2 On Relabeling with RWR

Post-processing via Relabeling with RWR works well given that the data has been clustered to a decent level, from which we may appropriately characterize the resulting micro-clusters and agglomerate them into a more coarsely grained groups. For example, suppose that a handful of the Liberal micro-clusters were mistakenly agglomerated into the unclearly defined macro-cluster. In this case, running RWR will result in relabeling the conservative-oriented channels in the Conservative group as the unclearly defined group instead of re-assigning the liberal-oriented channels to the Liberal group. In other words, if the micro-clusters are erroneously agglomerated in the first place, running RWR relabeling will not help improve the clustering performance. In this study, however, we rely on the cluster-wise keyword analysis to agglomerate micro-clusters into a cohesive and comprehensible groups, we were able to fully enjoy the advantage of relabeling channels via the means of RWR.

4.6 Summary

This study collects 23,382,992 user comments left on 144,799 videos of 539 domestic politics and news channels uploaded from June 28, 2007 to June 9, 2020 via GCP API for analysis. Given the assumption that users with similar political orientation will tend to leave comments to the channels sharing similar political implications, we construct a YouTube politics/news channel network by using the user comments left on the videos of each channel, from which the structural relationship among the given channels are examined. Furthermore, we have defined the concept of the "isolation score" by measuring the difference of each channel's distances from the Liberal cluster and the Conservative cluster centroids. Finally, we compute the isolation score on our YouTube channel matrix in order to estimate the extent of isolation each channel's users face.

As for the topics in politics/news, if the recommender system fails to introduce the diverse social perspectives to the users, they may be led to take on a particular stance with very little exposure to or understanding of other point of views. In order to cope with such cases, it is important for the users themselves to be able to characterize the political stance conveyed by the channels recommended. The cluster information from the YouTube channel network we constructed may serve as the vehicle through which the users may evaluate themselves upon the extent of impartiality in the political information at exposure as the result of the system's recommendation services.

Moreover, we tackle the limitations posed by the conventional k-means clustering approaches, where they fail to properly cluster scarce information, by incorporating keyword analysis and relabeling via RWR. Our approach shows more coherent and

comprehensible clustering results from the YouTube channel network. On the other hand, our proposed method enables measurement of the level of each channel user's isolation by introducing the concept of the channel-wise isolation scores.

Chapter 5

Conclusion

5.1 Contribution

The contribution of this dissertation is threefold. First, it provides a solution to real-world problems of the video content market, such as marketing strategies for the movie market, programming strategies for the TV program market, and the filter-bubble phenomenon caused by the recommendation system of the OTT market. Second, it proposes a framework that can assist in data-based decision-making by both experts and users. Earlier, it was not possible to objectively evaluate the prediction, and therefore, it was not possible to analyze and resolve the problem. However, we can now quantify the evaluation of our predictions, add data, retrain the model, and so on. As a result, the performance can be continuously improved. Finally, for solving the problem, a framework with a high explanatory power was proposed both as a model and a method. Chapter 2 describes the use of a decision-tree model to directly identify the factors and rules affecting the number of movie audiences so that an effective marketing strategy could be executed. Chapter 3 identifies the factors with the highest influence on viewership ratings and discusses the use of the XGBoost model to predict the viewership ratings of TV programs. Chapter 4 describes the clustering of YouTube channels with a k-means model and

demonstrates clustering by merging clusters based on keywords for postprocessing.

In Chapter 2, we propose a model that predicts the audience number on the opening Saturday using market research data obtained through online and offline surveys to help developing a marketing strategy. Marketing activity by distributors is a significant factor attracting audiences to the theater before a movie is actually released. Importantly, the audience number on the opening weekend is highly affected by marketing activities before the release and also determines how many screens will be allocated to the movie. Therefore, the distributor predicts the audience number on the opening weekend and develops a marketing strategy in order to gain a competitive advantage over other films being screened at the same time. However, as distributors make a prediction based on practitioners' experiences and intuitions, it is difficult to quantify the reliability of predicted values and deliver the correct marketing strategy.

In Chapter 3, we propose a two-step framework for it in this paper. The first step involves predicting the rating of a program with its program attributes, the ratings of programs assigned in the previous and following time slots, and the ratings of programs aired simultaneously in competing channels. The second step involves identifying the best airtime slot from the many candidates. Experiments were performed with actual Korea Nielsen ratings log data from July 2016 to October 2017. Broadcasters strive for high ratings to secure a high advertising revenue and a high export price. Two different strategies are used: (1) creating programs that more viewers find attractive, or (2) finding the most appropriate airtime slot with suitable target viewers. The latter strategy has received less attention in the literature.

In Chapter 4, we construct a YouTube politics/news channel network from the

user comments left on the videos of the select channels, based on the underlying assumption that the users with similar political orientation will leave comments to the channels sharing closely related political stances. We clustered the nodes of the resulting network to extract the conservative-oriented and the liberal-oriented groups, by using which we illustrate the landscape of Korean domestic politics/news channels in terms of their political stances. In addition, we define the concept of the isolation score as the difference in the given channel's distance from the centroid of the Conservative group and that from the centroid of the Liberal group. We use this metric to estimate the extent of isolation the users of each given channel face. Higher isolation score does not imply that the subject channel is biased towards a particular political orientation; rather, it implies that the users' consumption of the subject channel contents tend to lie near a particular political orientation.

5.2 Future Direction

In this dissertation, we proposed methods to assist relevant actors in decision-making in various fields of the domestic media industry. Within the assumed setting, our proposed methods showed satisfactory results, but there are still areas for further research and improvement.

In Chapter 2, we can extend our study in at least three directions. The first is to quantify the director and actor power by analyzing and verifying their validity. Existing studies quantify the power of directors and actors based on the performances of their previous movies, whereas this study involved a time decay factor, which assumes that the impact of previous movies will be halved in three years in order to substantiate the power of directors and actors. This assumption must be proved

by quantitative analysis such as cross-validation. The second direction is to improve the method for conducting online and offline surveys. In this st udy, we highlight how market research data from surveys plays an important role in predicting movie audiences. The prediction performance of the model will be improved if we develop additional questionnaire items, besides awareness and preference, to more exactly represent the intent of the audience. In conclusion, in terms of the third direction, further research should be conducted to predict the total box office score. We set a limited forecasting time period to predict the audience number on the first weekend, which is affected by the marketing strategy before release. However, the audience number on the first weekend is used as a criterion to determine how many screens will be allocated to a movie, which influences the overall box office score. Thus, this study could be used to forecast the total box office score before release.

In Chapter 3, we can develop our study further in two different ways. First, we can improve the results of our experiment by collecting more information about the television programs in our sample. In this study, we assumed that the view counts already incorporated some program-specific information, such as the airtime slot and the attributes of the channel airing the program, while the historical ratings were assumed to reflect the demographic attributes of the relevant program, including the director(s), cast, and screenwriter(s). However, even for the same television program, the guest stars differ from episode to episode. This poses a potential problem because when "star" guests appear, the view count may increase, corresponding to the publicity effect, while in contrast, episodes starring less popular guests are usually associated with lower view counts. The range of influential factors is not limited to the guest stars but also includes the star producers, such as *Young-seok Na* or

Byeong-Uk Kim or the announced screenwriter *Eun-sook Kim* when attracting an audience. If we were to collect more detailed data on such demographic information of the subject programs, we might be able to build a model with higher prediction accuracy. One may also, based on a view count prediction model trained on given data from the past, take the path of establishing a simulation model as a new program is being launched. In this study, we predicted the next episode's ratings class given the view count information from previous episodes, the structure of which makes it impossible for the model to make predictions without past data at the given timestamp t. However, a new program always takes off following the finale of an existing program, which calls for further research suited for such circumstances.

In Chapter 4, there are three different directions in which further research may stem from our study. Firstly, one may improve the quality of the channel representations by differentiating the positive comments from the negative ones, as briefly touched upon in subsection 4.5.1. Our proposed approach now assumes that the size of the overlap between the users for the given pair of channels translates simply and directly to the magnitude of their similarity. Our results did not show any sign of degradation due to such a simple setting. It could yet be the case, for example, that the same pool of users leave positive comment to an arbitrary channel A while leaving negative comments to another arbitrary channel B. In this case, the number of the common users leaving comments to both channels may be the same, yet the similarity between the channels is most likely to be very low. If positive comments can be differentiated from the negative comments when building the channel representations, it will certainly help construct a more refined channel network. Secondly, one may expand the scope of analysis encompass all the topics in the domestic range.

For now, our study has focused on the domestic politics/news channels. Our reason behind here is that intensive and graphic conflicts are easily invoked and actively witnessed in the realm of the selected topic. There does exist, however, other forms of conflicts than just politics, such as sexism or generation gaps, which could ultimately benefit from the content analysis of YouTube-like online platforms. Finally, one may conduct time-series analysis on the YouTube channels to capture the chronological changes in their political stances. For instance, our analysis, which takes the entire time range altogether, *YTN News*, *JTBC News*, *채널A뉴스 (Channel A News)*, *SBS News*, and *KBS News* are located around the liberal-oriented clusters. However, it is highly likely that some of the channels make shift their political stance as time progresses. It will be a quite meaningful research to dissect data by presidency and observe the shifts in the political stances the politics/news channels take from one presidency to another.

Bibliography

[1] F. Abel, E. Diaz-Aviles, N. Henze, D. Krause, and P. Siehndel, *Analyzing the blogosphere for predicting the success of music and movie products*, in 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2010, pp. 276–280.

[2] K. C. AGENCY, *Breaking new ground in audio media platform*, 2021.

[3] S. Asur and B. A. Huberman, *Predicting the future with social media*, in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, vol. 1, IEEE, 2010, pp. 492–499.

[4] J. Bae, *An Analysis on the Factors in Drama Ratings - Focusing on the Drama Attributes and Audience Factors*, Korean Journal of Broadcasting and Telecommunication Studies, 19 (2005), pp. 270–309.

[5] M. Bagella and L. Becchetti, *The determinants of motion picture box office performance: Evidence from movies produced in italy*, Journal of Cultural economics, 23 (1999), pp. 237–256.

[6] S. Basuroy, S. Chatterjee, and S. A. Ravid, *How critical are critical reviews? the box office effects of film critics, star power, and budgets*, Journal of marketing, 67 (2003), pp. 103–117.

[7] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar, *Role of different factors in predicting movie success*, in Pervasive Computing (ICPC), 2015 International Conference on, IEEE, 2015, pp. 1–4.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of statistical mechanics: theory and experiment, 2008 (2008), p. P10008.

[9] R. Brath and D. Jonker, *Graph analysis and visualization: discovering business opportunity in linked data*, John Wiley & Sons, 2015.

[10] L. Breiman, *Random forests*, Machine learning, 45 (2001), pp. 5–32.

[11] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*, CRC press, 1984.

[12] S. Cao, W. Lu, and Q. Xu, *Deep neural networks for learning graph representations*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.

[13] B. Chang, Y. Lee, B. Kim, and S. Nam, *Elaborating movie performance forecast through psychological variables : Focusing on the first week performance*, Korean Journal of Journalism & Communication Studies, 53 (2009), pp. 346–371.

[14] J. Chang, *An experimental evaluation of box office revenue prediction through social bigdata analysis and machine learning*, The Journal of The Institute of Internet, Broadcasting and Communication, 17 (2017), pp. 167–173.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *Smote: synthetic minority over-sampling technique*, Journal of artificial intelligence research, 16 (2002), pp. 321–357.

[16] T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[17] X. Cheng, C. Dale, and J. Liu, *Statistics and social network of youtube videos*, in 2008 16th Interntional Workshop on Quality of Service, IEEE, 2008, pp. 229–238.

[18] S. Cho, *Analysis of Programming Strategies of Television Broadcasting Networks: Focused on Rescheduled Programs, 1990-1999*, Korean Journal of Broadcasting and Telecommunication Studies, 14 (2000), pp. 387–428.

[19] S. Cho, *Current status and improvement plan of people meter ratings*, Korean Association for Broadcasting & Telecommunication, (2008), pp. 6–19.

[20] S. Cho, I. Song, and J. Park, *Planning for the Establishment of Verification System of TV Audience Measurement Focus Group Interview with Experts*, Advertising Research, (2012), pp. 542–585.

[21] E. Choi, *Analysis of Future Growth in Korea Movie Industry*, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 8 (2008), pp. 134–143.

[22] G. Choi, *Korean movie industry*, Hyungseul Publishing Networks, 2012.

[23] H. Choi, Y. Park, S. Jung, and H. Kim, *A Study on a Model of Predicting the Ratings for the First Installment of Terrestrial Television Soap*

Operas through Data Mining, The Journal of Korean Institute of Information Technology, 15 (2017), pp. 1–10.

[24] M. Choi, *A Study on the Displace Effect of Smartphone Broadcasting and Video Service for Watching TV*, Korean Journal of Broadcasting and Telecommunication Studies, 27 (2013), pp. 172–205.

[25] S. Choi, *The Audience Ratings, Genre Diversity & Programming Strategy in Comprehensive Programming Channel*, Studies of Broadcasting Culture, 24 (2012), pp. 75–109.

[26] Y. Choi, *A Study on the Adjacent Effects as the Determinants of the Ratings*, Studies of Broadcasting Culture, 4 (1992), pp. 245–255.

[27] D. Choudhery and C. K. Leung, *Social media mining: prediction of box office revenue*, in Proceedings of the 21st International Database Engineering & Applications Symposium, ACM, 2017, pp. 20–29.

[28] J. Chung, M. Kim, and H. Park, *Big Data Analysis and Modeling of Disinformation Consumption and Diffusion on YouTube*, Discourse and Policy in Social Science, 12 (2019), pp. 105–138.

[29] R. Cooper, *An expanded, integrated model for determining audience exposure to television*, Journal of Broadcasting & Electronic Media, 37 (1993), pp. 401–418.

[30] C. Cortes and V. Vapnik, *Support vector machine*, Machine learning, 20 (1995), pp. 273–297.

[31] D. Delen, R. Sharda, and P. Kumar, *Movie forecast guru: A web-based dss for hollywood managers*, Decision Support Systems, 43 (2007), pp. 1151–1170.

[32] C. Dellarocas, N. Awad, and X. Zhang, *Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning*, ICIS 2004 Proceedings, (2004), p. 30.

[33] Z. Di, J. Xiu, J. Lin, and Y. Qian, *Research on movie-box prediction model and algorithm based on neural network*, in Cloud Computing and Intelligence Systems (CCIS), 2016 4th International Conference on, IEEE, 2016, pp. 224–228.

[34] J. Du, H. Xu, and X. Huang, *Box office prediction based on microblog*, Expert Systems with Applications, 41 (2014), pp. 1680–1689.

[35] A. Dvir, A. K. Marnerides, R. Dubin, and N. Golan, *Clustering the unknown-the youtube case*, in 2019 International Conference on Computing, Networking and Communications (ICNC), IEEE, 2019, pp. 402–407.

[36] L. Einav, *Seasonality and competition in time: An empirical analysis of release date decisions in the us motion picture industry*, Working Paper, Harvard University, (2001).

[37] L. Einav, *Seasonality in the us motion picture industry*, The Rand journal of economics, 38 (2007), pp. 127–145.

[38] J. Eliashberg, S. K. Hui, and Z. J. Zhang, *Assessing box office performance using movie scripts: A kernel-based approach*, IEEE Transactions on Knowledge and Data Engineering, 26 (2014), pp. 2639–2648.

[39] J. E. Fletcher, *The syndication marketplace*, na, 1993.

[40] J. H. Friedman, *Greedy function approximation: a gradient boosting machine*, Annals of statistics, (2001), pp. 1189–1232.

[41] U. Gargi, W. Lu, V. Mirrokni, and S. Yoon, *Large-scale community detection on youtube for topic discovery and exploration*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, 2011.

[42] M. Ghiassi, D. Lio, and B. Moon, *Pre-production forecasting of movie revenues with a dynamic artificial neural network*, Expert Systems with Applications, 42 (2015), pp. 3176–3193.

[43] A. Grover and J. Leskovec, *node2vec: Scalable feature learning for networks*, in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.

[44] M. Ham and S. Lee, *Effects of Selective Exposure to YouTube Political Videos on Attitude Polarization: Verifying Mediating Effects of Political Identification*, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 21 (2021), pp. 157–169.

[45] E. Han, *Extremist traits of 5 · 18 history Distortion: Focusing on YouTube 5 · 18 distortion contents*, Journal of Democracy and Human Rights, 20 (2020), pp. 87–135.

[46] J. Han, *Multichannel programming policy of public broadcasting*, Studies of Broadcasting Culture, 12 (2000), pp. 87–111.

[47] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[48] T. H. Haveliwala, *Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search*, IEEE transactions on knowledge and data engineering, 15 (2003), pp. 784–796.

[49] T. Hennig-Thurau, M. B. Houston, and S. Sridhar, *Can good marketing carry a bad product? evidence from the motion picture industry*, Marketing Letters, 17 (2006), pp. 205–219.

[50] M. Heo, *A Network Analysis of the Bias in the Youtube and TV Channels: Focusing on the 'Corruption Investigation Office' Issue*, Journal of Digital Contents Society, 21 (2020), pp. 1453–1464.

[51] S. Hölig, U. Hasebrink, and J. Behre, *Reuters Institute Digital News Report 2020: Ergebnisse für Deutschland*, vol. 50, DEU, 2020.

[52] M. Hur, P. Kang, and S. Cho, *Box-office forecasting based on sentiments of movie reviews and independent subspace method*, Information Sciences, 372 (2016), pp. 608–624.

[53] S. Hwang, *Total Audience Measurement : Change in the Broadcasting Environmental: Focus on Total Audience Measurement Issue and Meaning*, Studies of Broadcasting Culture, 26 (2014), pp. 63–84.

[54] S. Jang, *The Influence of Political Broadcasting Use in YouTube on Political Talks, Political Efficacy and Political Participation*, Journal of Political Communication, null (2020), pp. 89–132.

[55] S. Jo, G. Lee, G. Jun, S. Hur, H. Kim, and S. Kim, *Proposal for Improvement of Filter Bubble by YouTube Recommendation System*, The HCI Society of Korea, (2020), pp. 903–906.

[56] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, *Movie reviews and revenues: An experiment in text regression*, in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 293–296.

[57] J. Jun, S. Hwang, and Y. Yoon, *A Verification about the Formation Process of Filter Bubble with Personalization Algorithm*, Journal of Korea Multimedia Society, 21 (2018), pp. 369–381.

[58] I. Kang and H. Eun, *A study of audience duplication: Inherited effect and transition effect*, Korean Journal of Broadcasting and Telecommunication Studies, 17 (2003), pp. 121–160.

[59] S. Kang, J. Heejeong, J. Kim, and J. Song, *A study on domestic drama rating prediction*, The Korean journal of applied statistics, 28 (2015), pp. 933–949.

[60] T. Kang, *Organizing strategy of public broadcasting in the digital era*, Studies of Broadcasting Culture, 11 (1999), pp. 37–69.

[61] E. V. Karniouchina, *Impact of star and movie buzz on motion picture distribution and box office revenue*, International Journal of Research in Marketing, 28 (2011), pp. 62–74.

[62] C. Kim and J. Hong, *How Fake News Become 'Real' News on Youtube: A Case Study of Political Propagandization and Reaction Among Political Actors Related to the ⟨Gosung Wildfire⟩*, The Journal of Political Science and Communication, 23 (2020), pp. 403–439.

[63] H. Kim, *Even if you don't like korea, even if you don't know korea... 'k-drama sickness' around the world*, 2016.

[64] H. Kim, *Unsupervised Korean Tokenizer and Extractive Document Summarization to Solve Out-of-Vocabulary and Dearth of Data*, PhD thesis, Seoul National University, 2019.

[65] I. Kim and J. Kim, *Youtube Algorithm and Confirmation Bias*, The Journal of Korean Association of Computer Education, 25 (2021), pp. 71–74.

[66] S. Kim, *Why can not stations give up rating?*, 2013.

[67] S. Kim, *K-movies in hollywood... 200 million korean movie audience to the 5th largest market in the world*, 2019.

[68] S. Kim and K. Kim, *A Study on factors affecting the viewer rating of "My Little Television" : Focusing on SNS Big Data*, Journal of Digital Contents Society, 17 (2016), pp. 1–10.

[69] S. Kim and W. Kim, *Youtube's big leap ¡digital news report 2019¿ korea-related key results*, Media Issue, 5 (2019).

[70] T. Kim, J. Hong, and P. Kang, *Box office forecasting using machine learning algorithms based on sns data*, International Journal of Forecasting, 31 (2015), pp. 364–390.

[71] Y. Kim, J. Chae, and C. Joo, *A Study on the Distorted and Faked News on the 518 Democratization Movement: A Network Analysis*, Journal of Democracy and Human Rights, 21 (2021), pp. 5–40.

[72] Y. Kim and J. Hong, *A study for the development of motion picture box-office prediction model*, Communications for Statistical Applications and Methods, 18 (2011), pp. 859–869.

[73] T. N. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, arXiv preprint arXiv:1609.02907, (2016).

[74] J. Klausen, E. T. Barbieri, A. Reichlin-Melnick, and A. Y. Zelin, *The youtube jihadists: A social network analysis of al-muhajiroun's propaganda campaign*, Perspectives on Terrorism, 6 (2012), pp. 36–53.

[75] M. A. Koschat, *The impact of movie reviews on box office: Media portfolios and the intermediation of genre*, Journal of Media Economics, 25 (2012), pp. 35–53.

[76] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, *Data preprocessing for supervised leaning*, International Journal of Computer Science, 1 (2006), pp. 111–117.

[77] M. Lash, S. Fu, S. Wang, and K. Zhao, *Early predictions of movie success: the who, what and when of profitability*, in International Conference on

Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, 2015, pp. 345–349.

[78] G. Lee, S. Sohn, and E. Jeong, *Network Analysis on the Diffusion of Fake Medical Information on YouTube: A Case Study of a Fake News "Inserting an Onion in the Ear to Heal Earaches"*, Health Communication Research, 17 (2018), pp. 97–129.

[79] J. Lee, *Comparison analysis of public broadcasting programming pattern*, Studies of Broadcasting Culture, 10 (1998), pp. 47–72.

[80] J. Lee, *MS Windows NT kernel description*, 2016.

[81] J. Lee, *YouTube Journalism Phenomenon: Causes, Characteristics, and Effects*, Korean Journal of Broadcasting and Telecommunication Studies, (2019), pp. 193–193.

[82] K. Lee and W. Chang, *Predicting financial success of a movie using bayesian choice model*, in KIIE 2014 Spring Conference, Korean Institute of Industrial Engineers, 2006, pp. 1428–1433.

[83] K. Lee and W. Chang, *Bayesian belief network for box-office performance: A case study on korean movies*, Expert Systems with Applications, 36 (2009), pp. 280–291.

[84] K. Lee, J. Pakr, I. Kim, and Y. Choi, *Predicting movie success with machine learning techniques: ways to improve accuracy*, Information Systems Frontiers, (2016), pp. 1–12.

[85] S. Lee, *A Study on the Viewing Rate Trends of Digital Media Service Special Reference to Terrestrial Real Time Broadcasting of IPTV*, Journal of Digital Convergence, 15 (2017), pp. 471–477.

[86] W. Lee and S. Kim, *The Impact of Content Variables on Rating Performance in Television Dramas*, Korean Journal of Broadcasting and Telecommunication Studies, 21 (2007), pp. 492–535.

[87] W. Lee, N. Lee, and J. Kim, *An Empirical Study on Forecasting Model of Popularity Rating for Drama Programs*, Journal of Digital Contents Society, 13 (2012), pp. 325–334.

[88] Y. Lee and C. Lee, *What Do The Algorithms of The Online Video Platform Recommend: Focusing on Youtube K-pop Music Video*, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 20 (2020), pp. 1–13.

[89] G. M. Lepori, *Positive mood and investment decisions: Evidence from comedy movie attendance in the us*, Research in International Business and Finance, 34 (2015), pp. 142–163.

[90] B. R. Litman and L. S. Kohl, *Predicting financial success of motion pictures: The'80s experience*, Journal of Media Economics, 2 (1989), pp. 35–50.

[91] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, *Predicting movie box-office revenues by exploiting large-scale social media content*, Multimedia Tools and Applications, 75 (2016), pp. 1509–1528.

[92] Y. Liu, *Word of mouth for movies: Its dynamics and impact on box office revenue*, Journal of marketing, 70 (2006), pp. 74–89.

[93] S. Lloyd, *Least squares quantization in pcm*, IEEE transactions on information theory, 28 (1982), pp. 129–137.

[94] D. Lovallo, C. Clarke, and C. Camerer, *Robust analogizing and the outside view: two empirical tests of case-based decision making*, Strategic Management Journal, 33 (2012), pp. 496–512.

[95] T. Lukk, *Movie marketing: opening the picture and giving it legs*, Silman-James Press, 1997.

[96] P. Marshall, M. Dockendorff, and S. Ibáñez, *A forecasting system for movie attendance*, Journal of Business Research, 66 (2013), pp. 1800–1806.

[97] F. Martin and J. Hutchinson, *Deep data: Analyzing power and influence in social media networks*, Second International Handbook of Internet Research, (2020), pp. 857–877.

[98] W. S. McCulloch and W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, The bulletin of mathematical biophysics, 5 (1943), pp. 115–133.

[99] W. McDowell and J. Sutherland, *Choice versus chance: Using brand equity theory to explore tv audience lead-in effects, a case study*, The Journal of Media Economics, 13 (2000), pp. 233–247.

[100] M. Mestyán, T. Yasseri, and J. Kertész, *Early prediction of movie box office success based on wikipedia activity big data*, PloS one, 8 (2013), p. e71226.

[101] E. Moretti, *Social learning and peer effects in consumption: Evidence from movie sales*, The Review of Economic Studies, 78 (2011), pp. 356–393.

[102] W. Myung, Y. Won, and M. Lee, *An Analysis on Watch Determinants of K-League Broadcasts Using Data Mining Method : Focused on A Portal Site Broadcasts*, Journal of Sport and Leisure Studies, 64 (2016), pp. 195–209.

[103] M. E. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Physical review E, 69 (2004), p. 026113.

[104] J. No, *83% of koreans use youtube 2.5 times longer than kakaotalk*, 2020.

[105] H. Oh, *A Study of the Change and Distinctive Features of Programming Field in Korean Public Service Broadcasting : Focusing on KBS Programming Field Analysis from 2003 to 2012*, Korean Journal of Broadcasting and Telecommunication Studies, 29 (2015), pp. 185–219.

[106] S. Oh and H. Song, *Youtube recommendation algorithm and journalism*, Korea Press Foundation, 2019.

[107] L. Olshen, C. J. Stone, et al., *Classification and regression trees*, Wadsworth International Group, 93 (1984), p. 101.

[108] A. Omari and A. R. Figueiras-Vidal, *Post-aggregation of classifier ensembles*, Information Fusion, 26 (2015), pp. 96–102.

[109] J. C. Paolillo, *Structure and network in the youtube core*, in Proceedings of the 41st Annual Hawaii international conference on system sciences (HICSS 2008), IEEE, 2008, pp. 156–156.

[110] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, *Community detection in social media*, Data Mining and Knowledge Discovery, 24 (2012), pp. 515–554.

[111] D. Park, *Effect of YouTube Usage on the Audience's Attitude and Perception of Importance of Issue*, Journal of Digital Convergence, 18 (2020), pp. 411–416.

[112] J. Park, *Who is the movie 'distributors', we have not known well?*, 2016.

[113] S. Park, S. Kim, and S. Joung, *Effects of Politics Channels of YouTube on Political Socialization*, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 20 (2020), pp. 224–237.

[114] S. Park, H. Song, and W. Jung, *The determinants of motion picture box office performance : Evidence from korean movies released in 2009-2010*, Journal of Communication Science, 11 (2011), pp. 231–258.

[115] B. Perozzi, R. Al-Rfou, and S. Skiena, *Deepwalk: Online learning of social representations*, in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701–710.

[116] J. Prag and J. Casavant, *An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry*, Journal of Cultural Economics, 18 (1994), pp. 217–235.

[117] L. Qin, *Word-of-blog for movies: A predictor and an outcome of box office revenue?*, Journal of Electronic Commerce Research, 12 (2011), p. 187.

[118] B. Rieder, Ò. Coromina, and A. Matamoros-Fernández, *Mapping youtube*, First Monday, (2020).

[119] K. Roose, *Youtube's product chief on online radicalization and algorithmic rabbit holes*, 2019.

[120] P. J. Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of computational and applied mathematics, 20 (1987), pp. 53–65.

[121] H. Rui, Y. Liu, and A. Whinston, *Whose and what chatter matters? the effect of tweets on movie sales*, Decision Support Systems, 55 (2013), pp. 863–870.

[122] Y. Ryu, *An Exploratory Study on News Perception of YouTube Current Affairs and Political Channel Users*, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 21 (2021), pp. 628–644.

[123] J. Sedgwick and M. Pokorny, *Movie stars and the distribution of financially successful films in the motion picture industry*, Journal of Cultural Economics, 23 (1999), pp. 319–323.

[124] R. Sharda and D. Delen, *Predicting box-office success of motion pictures with neural networks*, Expert Systems with Applications, 30 (2006), pp. 243–254.

[125] M. Shim, *A study on programme genre segmentations of Terrestrial Broadcasters : Comparative study of Korea, USA, UK, Japan, Taiwan*, Korean Journal of Broadcasting and Telecommunication Studies, 17 (2003), pp. 37–75.

[126] M. Shim and J. Han, *Determinants of the prime-time television program choice*, Korean Journal of Journalism & Communication Studies, 46 (2002), pp. 177–216.

[127] Y. Shin, S. Chun, and T. Kwon, *Analyzing differences in news reports from political orientation in YouTube*, Proceedings of Symposium of the Korean Institute of communications and Information Sciences, (2020), pp. 1225–1227.

[128] Y. Shin and S. Lee, *An Analysis of Filter Bubble Phenomenon on YouTube Recommendation Algorithm Using Text Mining*, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 21 (2021), pp. 1–10.

[129] V. Simonet, *Classifying youtube channels: a practical system*, in Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 1295–1304.

[130] M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave, *Analyzing (social media) networks with nodexl*, in Proceedings of the fourth international conference on Communities and technologies, 2009, pp. 255–264.

[131] S. Sochay, *Predicting the performance of motion pictures*, Journal of Media Economics, 7 (1994), pp. 1–20.

[132] J. Song and S. Han, *Predicting gross box office revenue for domestic films*, Communications for Statistical Applications and Methods, 20 (2013), pp. 301–309.

[133] L. Tang and H. Liu, *Community detection and mining in social media*, Synthesis lectures on data mining and knowledge discovery, 2 (2010), pp. 1–137.

[134] L. Van der Maaten and G. Hinton, *Visualizing data using t-sne.*, Journal of machine learning research, 9 (2008).

[135] S. Vujić and X. Zhang, *Does twitter chatter matter? online reviews and box office revenues*, Applied Economics, 50 (2018), pp. 3702–3717.

[136] D. Wang, P. Cui, and W. Zhu, *Structural deep network embedding*, in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 1225–1234.

[137] J. Wasko, *How hollywood works*, Sage, 2003.

[138] M. Wattenhofer, R. Wattenhofer, and Z. Zhu, *The youtube social network*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 6, 2012.

[139] K. Wen and C. Yang, *Determinants of the box office performance of motion picture in china-indication for chinese motion picture market by adapting determinants of the box office (part ii)*, Journal of Science and Innovation, 1 (2011), pp. 17–26.

[140] S. Wernicke, *A faster algorithm for detecting network motifs*, in International Workshop on Algorithms in Bioinformatics, Springer, 2005, pp. 165–177.

[141] S. Wernicke and F. Rasche, *Fanmod: a tool for fast network motif detection*, Bioinformatics, 22 (2006), pp. 1152–1153.

[142] S. Yang, *Reality and Challenges of Traditional Journalism in the Era of YouTube Journalism*, Journal of Social Science, 31 (2020), pp. 245–262.

[143] J. Yim and B. Hwang, *Predicting movie success based on machine learning using twitter*, KIPS Transactions on Software and Data Engineering, 3 (2014), pp. 263–270.

[144] H. Yoganarasimhan, *Impact of social network structure on content propagation: A study using youtube data*, Quantitative Marketing and Economics, 10 (2012), pp. 111–150.

[145] S. Yu and S. Kim, *An Analysis of Factors which Affect the Rating of Drama ; Focusing on the Production Factors of Outsourcing Drama*, Journal of Media Economics & Culture, 8 (2010), pp. 7–48.

[146] L. Zhang, J. Luo, and S. Yang, *Forecasting box office revenue of movies with bp neural network*, Expert systems with applications, 36 (2009), pp. 6580–6587.

[147] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, *Topic oriented community detection through social objects and link analysis in social networks*, Knowledge-Based Systems, 26 (2012), pp. 164–173.

[148] D. Zimbra, K. R. Sarangee, and R. P. Jindal, *Movie aspects, tweet metrics, and movie revenues: The influence of ios vs. android*, Decision Support Systems, 102 (2017), pp. 98–109.

국문초록

영상 콘텐츠를 비롯한 콘텐츠 시장은 크게 세 가지 특징을 가지고 있다. 첫 번째는 고위험 고수익 시장으로 성공 가능성이 불확실하고 위험도가 높지만 만들어진 영상 콘텐츠를 복제하여 배급하는 비용은 매우 적기 때문에 성공 시에 고수익창출이 가능하다는 점이다. 두 번째는 경험재적 특성으로 구입 이전에 제품의 품질을 제대로 파악할 수 없고, 공급자와 수요자간의 정보의 비대칭적 존재로 인하여 마케팅 활동이 중요한 역할을 한다는 점이다. 세번째로는 One Source Multi Use 의 성격을 가지고 있기 때문에 성공하는 경우, 다양한 창구를 통해서 추가 이윤을 창출할 수 있다는 점이다. 이러한 특징들로 인해서, 성공 가능성이 높은 콘텐츠와 그렇지 않은 콘텐츠를 잘 구분하는 것과 수요자들이 제품에 대해서 잘 알 수 있도록 콘텐츠에 대해서 효과적인 마케팅 활동을 하는 것은 콘텐츠 시장에 중요한 문제다. 본 논문에서는 영화, TV 프로그램, OTT 서비스 등 다양한 영상 콘텐츠 시장에서 실제로 발생하고 있는 문제들에 대해서 데이터의 특성과 목적에 맞는 머신 러닝 모델을 활용하여 데이터 기반의 의사 결정을 보조하는 방법론을 제안하고자 한다.

첫번째로, 영화 시장에서 마케팅은 굉장히 중요한 부분이지만, 여전히 실무자들의 감에 의존한 의사 결정이 이루어지고 있다. 우리는 온라인과 오프라인 설문조사를 통해서 수집한 시장 조사 데이터를 활용해서 개봉주 토요일 관객수를 예측하는 모델을 학습하고, 학습한 모델을 활용하여 효율적인 마케팅 활동을 할 수 있는 방안을 제안한다. 두번째로, TV 프로그램 시장에서 TV 프로그램 편성은 TV 프로그램과 시청자 그룹을 잘 매칭시켜서 전체적인 시청률을 향상 시키기 위해서 이루어진다. 우리는 프로그램의 특성과 프로그램의 이전/이후/동시 프로그램들의 시청률 정보를 활용해서 프로그램의 시청률을 예측하는 모델을 학습하고, 이를 활용해서 다양한 편성 시나리오

중에서 최적의 편성 시나리오를 찾을 수 있도록 의사 결정을 보조하는 프로세스를 제안한다. 마지막으로, OTT 시장에서는 최근 추천 시스템으로 인해서 사용자의 인식이 편향되는데에 대한 문제가 대두되고 있다. 특히, 정치/뉴스 분야의 경우, 사회의 다양한 관점을 사용자가 추천 서비스로 인해서 접하지 못한다면, 사용자가 인지하지 못한채로 특정 정치적 성향에 대한 편향성이 더욱 심화되는 문제가 있을 수 있다. 이를 보완하기 위해서는 추천 받은 채널이 어떤 성향의 채널인지를 사용자 스스로 잘 인지하고 이용하는 것이 중요하다. 우리는 사용자가 각 채널의 영상에 남긴 댓글 데이터로 뉴스/정치 분야의 채널 네트워크를 구축했다. 그리고, 보수와 진보 성향의 채널 군집으로 네트워크를 구분하여 유튜브 채널들의 정치적 성향에 대한 지형을 제시함으로써 편향성을 보완하는 방안을 제안한다.

주요어: 데이터마이닝(data mining), 기계학습(machine learning), 인공지능(artificial intelligence), 의사결정 지원 시스템(decision support system), 추천 시스템(recommendation system), 마케팅(marketing), 예측(prediction), 군집화(clustering), 박스오피스(Box-office), 시청률(ratings), 방송 편성(broadcasting programming), 유튜브(youtube), 채널 네트워크(channel network), 키워드 추출(keyword extraction), 필터버블(filter bubble), 개인화(personalization)

학번: 2015-21140