Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 280–291
Expert Systems with Applications
www.elsevier.com/locate/eswa

Bayesian belief network for box-office performance: A case study on Korean movies

Kyung Jae Lee [a], Woojin Chang [b,*]

[a] Graduate Program in Technology and Management, Seoul National University, Sillim-dong, Gwanak-gu, Seoul 151-742, Republic of Korea
[b] Department of Industrial Engineering, Seoul National University, Sillim-dong, Gwanak-gu, Seoul 151-742, Republic of Korea

Abstract

Due to their definition as experience goods with short product lifetime cycles, it is difficult to forecast the demand for motion pictures. Nevertheless, producers and distributors of new movies need to forecast box-office results in an attempt to reduce the uncertainty in the motion picture business. Previous research demonstrated the ability of certain movie attributes such as early box-office data and release season to forecast box-office revenues. However, no previous research has focused on the causal relationship among various movie attributes, which have the potential to increase the accuracy of box-office predictions. In this paper a Bayesian belief network (BBN), which is known as a causal belief network, is constructed to investigate the causal relationship among various movie attributes in the performance prediction of box-office success. Subsequently, sensitivity analysis is conducted to determine those attributes most critically related to box-office performance. Finally, the probability of a movie's box-office success is computed using the BBN model based on the domain knowledge from the value chain of theoretical motion pictures. The results confirm the improved forecasting accuracy of the BBN model compared to artificial neural network and decision tree.

Keywords: Bayesian belief network; Casual belief network; Box-office performance; Domain knowledge

1. Introduction

The demand for movies, a ubiquitous cultural product along with books and music, has been significantly increasing due to the increment of personal desire for cultural life. As most cultural products are defined as experience goods with short product lifetime cycles, except books, it is difficult to forecast their demands (Chang & Ki, 2005). However, a producer who has to produce new movies continuously, and a distributor who is responsible for selecting promising movies need to be able to predict the demand to reduce the uncertainty in the market (Eliashberg, Elberse, & Leenders, 2006).

Exploring the factors influencing box-office performance is the first priority when analyzing the motion picture industry, because it is a basic foundation for movie-related policy establishment (Yoo, 2002). In Korea, Korean movies have maintained a market share above 50% in the past 3 years, confirming the local dominance of the Korean motion picture industry over foreign movies (KOFIC., 2005). It is important for the continual success of the industry to analyze the attributes affecting box-office success, not only for Korean moviemakers but also for foreign producers or distributors who release their movies in Korea.

Previous research concerning box-office performance can be divided into psychological (or behavioral) and economic (or quantitative) perspectives (Eliashberg et al., 2006; Sharda & Delen, 2006). The former approach focused on individual decisions to attend a movie among the vast array of entertainment options, and more critically, to choose a particular movie. Researchers adopting this approach aim to relate such variables as opinions, needs, values, attitudes and personality traits to consumers' decision-making processes (Eliashberg, Jonker, Sawhney,

* Corresponding author. Tel.: +82 2 880 8335; fax: +82 2 889 8560.
E-mail address: changw@snu.ac.kr (W. Chang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.09.042

& Wierenga, 2000; Eliashberg & Sawhney, 1994; Sawhney, Mohanbir, & Eliahberg, 1996; Zufryden, 1996). Such studies generally use data collected by surveying individual consumers. On the other hand, studies within the economic approach explored factors that influence collective movie attendance decisions. The economic approach sought to explore the attributes that influence the financial performance of motion pictures (Elberse & Eliashberg, 2003; Jedidi, Krider, & Weinberg, 1998; Litman & Ahn, 1998; Litman & Kohl, 1989; Neelamegham & Chintagunta, 1999; Ravid, 1999; Scohay, 1994). These studies typically used aggregate data on movie-going behavior collected by industry trade sources. In this paper, we focus on the economic approach with the aim of determining the casual relationship between attributes related to box-office performance.

There are some cases where promising movies show poor box-office performance, and inconspicuous movies show unexpectedly good results. This implies that box-office success is not solely influenced by individual movie attributes such as director, actor, distributor, release season, screen number etc., but is also dependent on the causal relationship among all attributes related to movies. Previous research (Ainslie, Dreze, & Zufryden, 2005; Eliashberg et al., 2000; Jedidi et al., 1998) attempted to determine which attributes affect box-office success through multivariate analysis without considering the interactions among these attributes. Therefore, such previous research experienced difficulty in excluding the probable multicollinearity among attributes, which decreased the explanation power. These endogeneity problems have not been fully solved yet despite the importance of predicting a movie's financial success. Elberse and Eliashberg (2003) demonstrated the effectiveness of 3SLS (3-stage least squares) in resolving an endogeneity problem between box-office revenue of the demand side and number of screens of the supply side. Although a unique study on the endogeneity problem, it remained far from an ultimate solution in that the variables (attributes) for the demand and supply equation were chosen arbitrarily and the causality among other variables, except screen number, was not considered.

Artificial neural network (ANN) and decision tree (DT) have been suggested as suitable forecasting methods to consider the endogeneity problem. ANN can consider interactions among input nodes indirectly by calculating the relative weight between inputs by hidden nodes. For this reason, ANN is a proper method to solve the endogeneity problem. However, ANN also has a limitation in specifying explicit relationships due to its 'black-box' nature, which means the causal relationships between variables are not identified. On the other hand, although DT can consider the direct influence between variables since the variables in DT are connected directly, it has fallen from favor due to its lower forecasting performance than ANN in many studies (Sharda & Delen, 2006).

The novelty of this paper is that the casual relationship between movie attributes is analyzed to predict box-office success using Bayesian belief network (BBN) based on the domain knowledge from the value chain of theatrical motion pictures preceding their viewing by movie-going audiences. BBN is becoming the most powerful data-mining tool because it can easily reflect domain knowledge and can simply notice cause and effect (Antal, Fannes, Timmerman, Moreau, & De Moor, 2004; Aronsky & Haug, 2000). In this paper, we compare the forecasting accuracy of BBN, ANN and DT for box-office success.

The remainder of this paper is organized as follows. Section 2 introduces a BBN model to predict box-office success and the data and variables used in the forecasting model. Section 3 presents the test results on the independence between attributes, sensitivity analysis, and BNN performance evaluation. Finally, the overall contribution of this study, along with its limitations and further research directions, is discussed in Section 4.

2. Model

We specify the causality and verify BBNs performance as a prediction model of box-office success. Our research flow is shown in Fig. 1. First, raw data are collected from several information sources, and are measured as variables (attributes). After discretizing each continuous variable using a classification and regression tree (CART) algorithm, a BBN model based on the domain knowledge from movie's value chain is constructed by structural learning process. Then, we investigate the critical variables for box-office performance through sensitivity analysis, and explore the optimized combination of variables to increase the success probability. Finally, the BBN model based on domain knowledge is applied to the prediction of box-office success and is compared with ANN, DT, and BBN without domain knowledge in terms of prediction accuracy. In the case of ANN and DT, evaluations are performed using each of two data sets: the original data set consisting of continuous and discrete variables, and the discretized data set. We used Bnsoftware (http://www.cs.ualberta.ca/~jcheng/bnsoft.htm) and Netica (http://www.norsys.com) to construct BBN for forecasting and perform sensitivity analysis in the visualized network, respectively. Neuroshell 2.0 was used for ANN, and SPSS AnswerTree 3.0 was used to determine the splitting rule of continuous variables. Computation of DT was performed by SAS E-Miner.

2.1. Bayesian belief network (BBN)

BBN represents a causal relationship between variables using DAG (Directed Acyclic Graph), as depicted in Fig. 2, and has the following advantages: (1) it is a powerful method to treat missing value problem, (2) it is good at prediction due to the knowledge on causal relationship between variables, and (3) it allows the easy use of prior knowledge or domain knowledge (Jensen, 1996). In a BBN framework, the attributes considered in a movie,

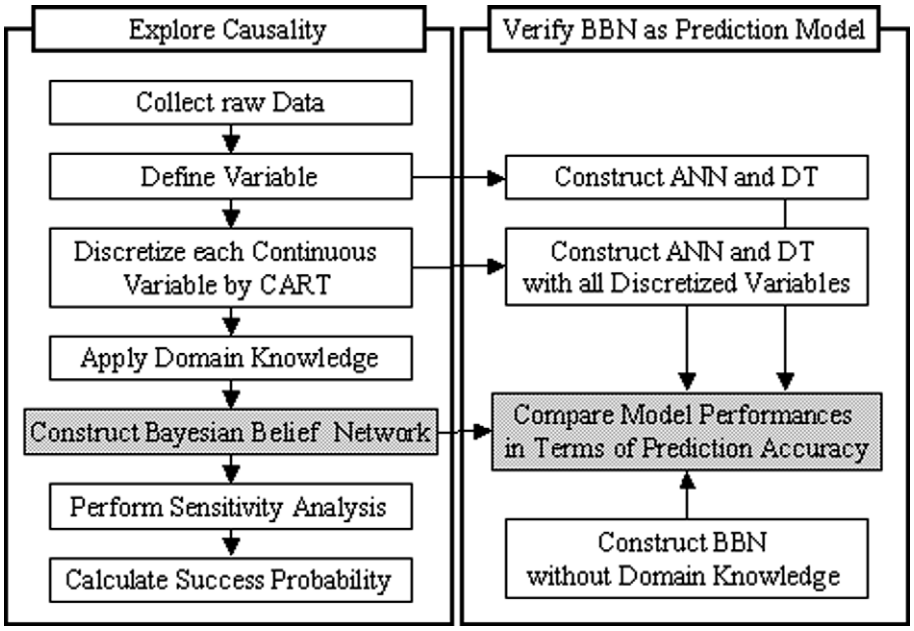

Fig. 1. Overall research flow.

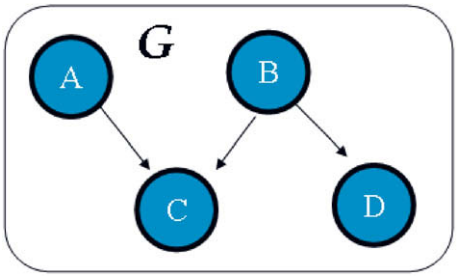

Fig. 2. Directed acyclic graph.

including the director, actor and genre, are converted to the case $X = \{X_1, \ldots, X_n\}$, where X_i's denote the attributes split into certain categories. The marginal distribution of a case is the multiplication of the conditional probability of each attribute having parent nodes, $Pa(X_i)$, which are the direct cause of the child node, X_i, as follows:

$$P(X) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$$

2.1.1. Structure learning

The network structure and the parameters of the local distributions must be determined from the data. The learning algorithm used in this paper is the three-phase construction algorithm, originally suggested by Cheng and Greiner (2001), comprising drafting, thickening, and thinning. In the first-phase, this algorithm computes the mutual information of each pair of nodes as a measure of closeness and creates a draft based on this information. In the second-phase, the algorithm adds edges when the pairs of nodes cannot be d-separated. In the third-phase, each edge of the current graph is examined using CI tests and will be removed if the two nodes of the edge can be d-separated. An edge orientation procedure is also conducted if the node ordering is not known.

2.1.2. Bayesian network classifiers

BBN is essentially a statistical model that makes it feasible to compute the (joint) posterior probability distribution of any subset of unobserved stochastic variables, given that the variables in the complementary subset are observed. This functionality makes it possible to use BBN as a statistical classifier by applying the winner-takes-all rule to the posterior probability distribution for the (unobserved) class node (Baesens, Egmont-Petersen, Castelo, & Vanthienen, 2002).

2.2. Data and variables

The data on the audience number, weekly screen number, and release date were collected from the Korean Film Council (www.kofic.or.kr), and the information about directors, actors, genre, Motion Picture Association of Korea (MPAK) rating, producer, distributor, budget, related article, and critics of moviegoer were obtained from the internet portal site, Naver (www.naver.com). The expert critics were collected through movie magazines such as Cine21 (www.cine21.co.kr), and the information about the age groups who purchase movie tickets in advance was gained from the web site for movie booking, Maxmovie (www.maxmovie.com). One hundred movies, which were released from January, 2005 to October, 2006 were analyzed.

In our three forecasting models for box-office performance, BBN, ANN, and DT, the audience number is used as the dependent variable while the other data are used as explanatory variables.

2.2.1. Dependent variable

Box-office gross revenue was the most frequently used dependent variable in the previous literature. However, the exact data on box-office gross revenue for Korean movies are not publicly available. Therefore, we define the total number of moviegoers of each movie as a dependent variable since the ticket price (list price) of Korean movies is fixed except for a few cases and box-office gross revenue is proportional to the total audience number. We discretized this continuous dependent variable into 2 and 3 groups for BBN. The dependent variable of 2-group, *Success2*, was classified into 'inferior' and 'superior' groups by the median audience value, and *Success3* split the dependent variable into 'bad', 'standard', and 'excellent' groups according to the one-third and two-third quantiles of the audience number. Thus, movies belonging to both 'superior' of 2-group and 'excellent' of 3-group are the most successful.

2.2.2. Explanatory variables

We consider the following variables as attributes affecting the audience number and box-office success.

2.2.2.1. Power of director and actor. Most studies indicated that director power is not very significant and at times is much weaker than star power. Levin, Levin, and Heath (1997) showed in his survey that movie has independent brands such as directors and actors. Allbert (1998) conducted an empirical test demonstrating that the impression of the actors in the previous movie has an effect on the box-office performance of the current movie. De Vany and Walls (2004) showed that a movie with a star could reduce the uncertainty of theater revenues. We define two variables to measure the powers of director and star. First, dummy variables, *D_expert* and *A_expert*, are defined to indicate whether the current director and actor have played in at least one successful movie in the past, respectively. We include the variables *D_power* and *A_power*, indicating the box-office performance of the previous movie of the director and actor, respectively.

2.2.2.2. Genre. A number of studies (Litman, 1983; Litman & Kohl, 1989; Wyatt, 1991) have shown that comedy, fantasy, and horror are positively correlated with box-office success. Litman and Kohl (1989) determined that drama is negatively correlated with box-office success. To categorize movie genres, we divide them into 6 groups: comedy, action/crime, drama/human, horror/thriller, romance, and SF/fantasy. In this paper, the variable *Genre* is defined as the proportion of the average audience of each genre in the previous year.

2.2.2.3. Motion Picture Association of Korea (MPAK) rating. The content rating assigned by the Motion Picture Association of America (MPAA) has been considered as an important factor that influences the motion picture industry because the rating tends to determine the potential size of the audience (Litman & Ahn, 1998; Litman & Kohl, 1989; Prag & Casavant, 1994; Ravid, 1999). However, most previous research has failed to support this. We divided the rating into 4 groups by Motion Picture Association of Korea (MPAK) through the variable *Class*: over 12, over 15, over 18 (adult), and no rating.

2.2.2.4. Producer and distributor. Bagella and Becchetti (1999) showed that the experience of a producer is a considerable factor, which is as powerful as director power and star power, in Italy's motion picture industry. In addition, distributor power is regarded as a significant factor because the more power they have, the more screens they can obtain and sustain (Elberse & Eliashberg, 2003; Sawhney et al., 1996; Swami, Eliashberg, & Weingberg, 1999). Ainslie et al. (2005) calculated the power of each distributor and showed that among the big studios 20th Century Fox and New Line time their release more effectively. We defined the dummy variable *P_expert* representing whether the producer had made at least one successful movie in the past, and defined the variable *Dist* as the market share of each distributor in the previous year to measure the distributor's power.

2.2.2.5. Number of screens. In most previous studies, the screen number has been considered as the most significant factor determining movie performance (Ainslie et al., 2005; Chang & Ki, 2005; Elberse & Eliashberg, 2003; Sawhney et al., 1996; Swami et al., 1999). Especially, the screen number in the first week of release is the most important (Jones & Ritz, 1991; Neelamegham & Chintagunta, 1999). Elberse and Eliashberg (2003) determined that screen number is an endogenous variable because the screen number in two weeks from the release is affected by the audience number in the first two weeks. We restrict the variable *Screen* to the total screen number in for the first two weeks of showing.

2.2.2.6. Marketing. Some studies have investigated the external factors affecting box-office success such as the effects of marketing, word-of-mouth, or competition. Ravid (1999) showed that marketing expenditure is positively correlated with box-office success, and Zufryden (2000) suggested that the internet can be an effective marketing tool to extend the life cycle of a movie. Furthermore, the bigger distributor, the more positive effect it has on the success of the movie (Ainslie et al., 2005). In this paper, we add the variable *Online*, which represents the number of articles related to a movie on the web, to measure the marketing effects.

2.2.2.7. Word-of-mouth and Critique. The effect of word-of-mouth is indeed an important factor of box-office success for all cultural experience goods (Eliashberg & Shugan, 1997; Mahajan, Muller, & Kerin, 1984). Mahajan et al. (1984) developed a model explaining the delay of diffusion due to negative word-of-mouth. The critique and opinion of audiences or experts released to the pubic through the

mass media is positively related to box-office success, as much as word-of-mouth. Mahajan et al. (1984) showed that movie critique has an intimate relationship with word-of-mouth. Therefore, we also defined *V_critic* and *E_critic* as the average grade of moviegoers and experts' critiques, respectively, to determine the level of influence of movie critique on box-office performance.

2.2.2.8. Competition and seasonality. If a new movie is released simultaneously with another successful movie, the new movie has a lower chance of box-office success due to high competition. Many researchers have therefore reflected the market situation to include the competition effect and concluded that the competition is a significant factor (Elberse & Eliashberg, 2003; Jedidi et al., 1998). In this paper, we define *Compet1* and *Compet2* to express the competition on box-office success. The variables *Compet1* and *Compet2* indicate in the release time of a new movie whether any other current movie is showing good box-office performance and whether any foreign blockbuster movie is showing, respectively. We also consider the seasonality effect through the variable *Season*, which highlights movies played in summer vacation, Chuseok (Korean Thanksgiving day), and Christmas season.

2.2.2.9. Budget. Budget (production budget) has been regarded as an important factor for box-office success in previous research (Elberse & Eliashberg, 2003; Litman & Ahn, 1998; Litman & Kohl, 1989; Zufryden, 1996). The scope of budget determines whether a movie is a blockbuster or not. In this paper, the variable *budget* is defined as the pure production budget, excluding other activities like marketing.

2.2.2.10. Distribution and online. The variable on marketing expenditure, which was regarded as an important factor for box-office success in previous research, is excluded in this paper since accurate data on marketing expenditure are not publicly available in Korea. For alternatives to marketing expenditure we include *Dist* and *Online*, which represent the market share of each distributor in the previous year and the number of published articles related to the

Table 1
Description of variables

Category	Name	Description	Measurement
Box-office performance	*Success2*	Performance status based on the audience number	Superior: above median value Inferior: under median value
	Success3		Excellent: 0–33.3% in the ranking Standard: 33.3–66.6% Bad: 66.6–100%
Production	*Genre*	Attraction of each genre (comedy, action, drama, horror, romance, and SF/fantasy)	The proportion of the average audience of a movie in each genre in the previous year
	Class	MPAK viewing rate	4 categories (All: no rating, 12: over 12, 15: over 15, 18: over 18)
	Budget	Production budget	Log(production budget)
	A_expert	Actors' experience	1: casting a star with a box-office success 0: otherwise
	A_power	Star power affecting box-office performance	Ratio of the audience number of actor's previous movie to the average audience number of the top 10 movies in the corresponding year
	D_expert	Director's experience	1: directing at least one top 10 movies in the past years 0: otherwise
	D_power	Director power affecting box-office performance	Ratio of the audience number of director's previous movie to the average audience number of the top 10 movies in the corresponding year
	P_expert	Producer's experience	1: producing at least one top 10 movies in the past years 0: otherwise
Distribution	*Dist*	Distributor's power affecting box-office performance	Market share of each distributor in the previous year
	Online	Online effect	Log(the number of related articles on the web)
	Season	Seasonality effect	1: vacations, christmas season, Chuseok (Korean Thanksgiving day), 0: otherwise
	Compet1	Competition with other popular movies	Market share of top 5 movies in the previous week
	Compet2	Competition with foreign blockbuster movies	1: release of foreign blockbuster 0: otherwise
Exhibition	*Age*	Attraction to people over 30	Booking proportion of over 30
	E_critic	Expert's critique	Log(average grade from experts)
	V_critic	Word-of-mouth effect	Log(average grade from audiences)

movie, respectively, under the assumption that the marketing expenditure is positively correlated with distributors' market shares and the number of published articles.

We exclude some variables about sequels and movie awards since few sequel movies are released in Korea and it is impossible to know whether a new movie will receive an award or not before its release. The variables are defined and described in Table 1.

2.2.3. Discretization of variable

As BBN requires discrete variables as inputs, we apply the CART algorithm to find the splitting rule for the discretization of continuous variables. We split the 11 continuous variables into 3 levels, high, middle and low, according to the amount of impurity reduction. Fig. 3 shows the result of the splitting rule of *screens*; the other continuous variables are split in the same way. In order to simplify the comparison, we normalize the splitting values to 0–100 as shown in Table 2. The value classifying middle and high levels of variables in the exhibition stage: *V_critic, E_Critic*, and *Age* increases. The middle level ranges of *V_critic* and *E_critic* in particular are larger than those of the other variables.

Table 2
Splitting rules of continuous variables

Variable	Splitting rules		The length of middle level range
	Low vs. middle	Middle vs. high	
Genre	59.7	64.9	5.2
Budget	17.0	29.33	12.33
A_power	2.7	11.3	8.5
D_power	0.2	19.5	19.3
Screen	30.0	42.5	12.5
Dist	39.6	56.0	16.5
Online	61.6	72.8	11.2
Compet1	18.3	35.7	17.4
V_critic	56.9	94.6	37.7
E_critic	39.1	80.6	41.6
Age	75.9	92.0	16.1

2.3. Domain knowledge

Domain knowledge of the variables in Table 1 can be obtained from the value chain of theatrical motion pictures consisting of the three key stages – production, distribution, and exhibition – that precede their 'consumption' by movie-going audiences (Eliashberg et al., 2006). After we categorized the variables into these 3 groups according to the nature of each variable, we visualize the domain knowledge for BBN in Fig. 4.

(1) Variable level: variables within the same stage can interact with each other.
(2) Stage level: according to the passage of time, the variables in the production stage can affect the variables in both the distribution and exhibition stages. However, the variables in the exhibition stage, which is the last stage in the value chain, cannot affect any variable in the production and distribution stages.

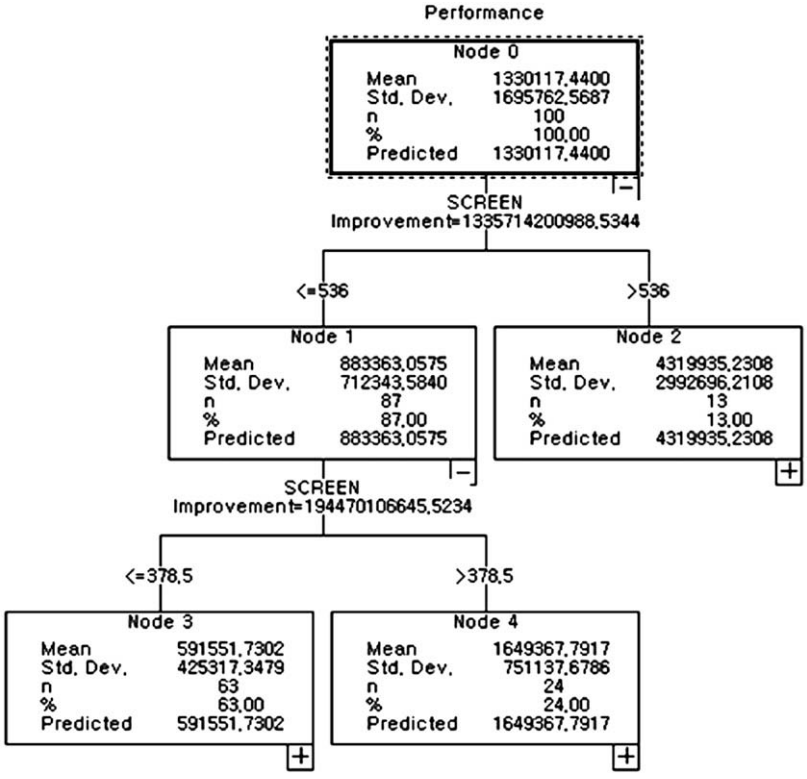

Fig. 3. Spliting rule of *Screen*.

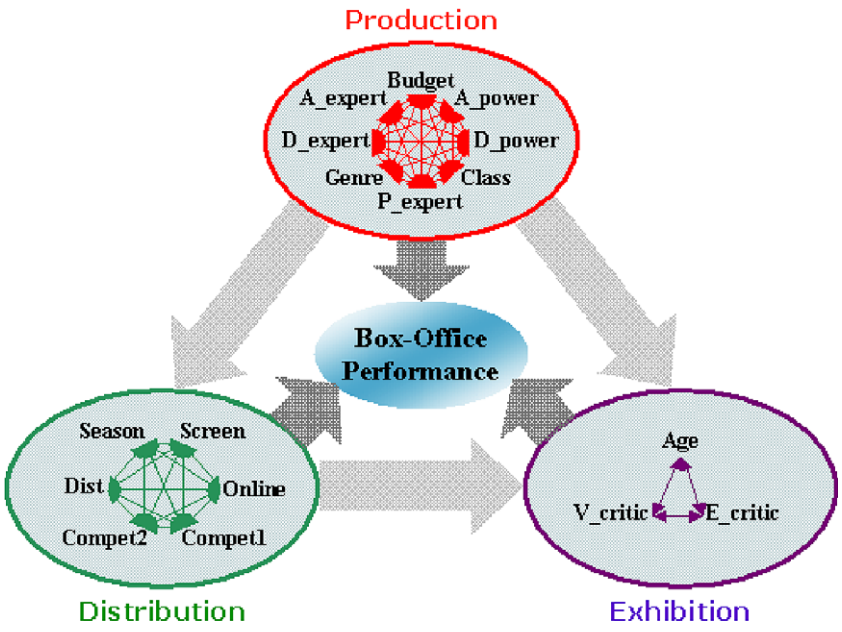

Fig. 4. Domain knowledge.

3. Results

The dependence between variables described in Section 2 is checked, and BBN is constructed using the variables so that the causal relationships between them can be visualized. Sensitivity analysis is conducted to find the variables which reduce the uncertainty of box-office performance and to estimate the probability of box-office success. Finally, the forecasting accuracy of BBN based on domain knowledge is compared to that of ANN, DT, and BBN without domain knowledge.

3.1. Correlation analysis and multicollinearity test

We previously mentioned the interactions of movie attributes affecting box-office performance. In this section, we justify the consideration of causal relationship of attributes for the prediction of box-office success by checking the independence of each variable via correlation analysis and multicollinearity test.

First, we perform the correlation analysis. The links in Fig. 5 represent the spearman correlation between variables. The bold lines and other lines are statistically significant at the 0.01 and the 0.05 level, respectively. Many variables are correlated with each other, with *Performance* and *Screen* being heavily correlated with the other variables. This suggests that the explanatory variables are interdependent.

The dependence between variables is confirmed by the multicollinearity of multiple regression. Table 3 presents 17 variables: 11 continuous and 6 discrete. Multicollinearity exists when variance inflation factor (VIF) is greater than 1.4. As shown in Table 3, many variables, except

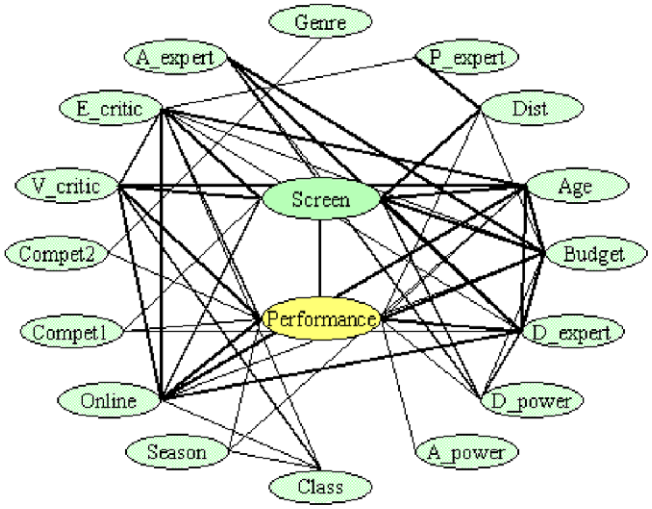

Fig. 5. Correlation between variables.

Genre, *P_expert*, *Season*, *Compet1*, and *Compet2*, have VIF > 1.4, indicating that these variables are correlated with other variables.

The presence of correlation and multicollinearity between variables implies that a wrong result could be obtained unless we consider the causal relationships between the variables.

3.2. Construction of BBN

We constructed BBN by structural learning based on domain knowledge, and the results of 2-group and 3-group are shown in Figs. 6 and 7, respectively. It is evident that many variables not only have an effect on box-office suc-

Table 3
Descriptive statistics and regression analysis

Name	Variable type	Descriptive statistics				Regression analysis		
		Average	SD	Min.	Max.	Coeff.	*P*-value	VIF
Performance	Continuous	1330117	1695763	107787	10405224			
Budget	Continuous	4.017	2.334	0.9	15.0	−0.077	0.384	2.036
Genre	Continuous	0.164	0.037	0.07	0.27	−0.049	0.486	1.271
Class12	Discrete	0.250	0.435	0	1	0.109	0.435	5.079
Class15	Discrete	0.460	0.501	0	1	0.140	0.368	6.311
Class18	Discrete	0.230	0.423	0	1	−0.024	0.859	4.645
A_expert	Discrete	0.300	0.461	0	1	−0.080	0.308	1.584
A_power	Continuous	0.251	0.326	0.01	2.35	0.069	0.304	2.036
D_expert	Discrete	0.090	0.288	0	1	0.001	0.987	1.893
D_power	Continuous	0.132	0.297	0	2.35	−0.044	0.573	1.608
P_expert	Discrete	0.340	0.476	0	1	0.000	0.998	1.334
Screen	Continuous	372.4	173.1	128	1260	0.629	0.000	2.813
Dist	Continuous	0.143	0.085	0.00	0.24	0.012	0.867	1.441
Season	Discrete	0.420	0.496	0	1	0.172	0.017	1.300
Online	Continuous	2.712	0.787	0	4.69	0.099	0.256	1.970
Compet1	Continuous	0.482	0.285	0.02	1.65	−0.064	0.339	1.168
Compet2	Discrete	0.190	0.394	0	1	−0.025	0.724	1.274
V_critic	Continuous	6.740	1.583	2.07	9.28	0.134	0.091	1.609
E_critic	Continuous	5.415	1.284	1	8	0.058	0.473	1.697
Age	Continuous	0.390	0.007	0.25	0.56	0.141	0.093	1.806

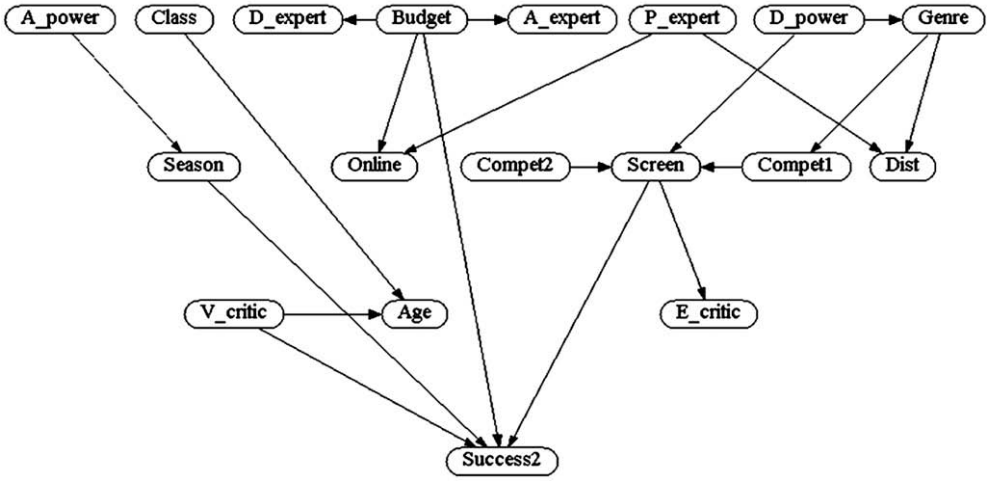

Fig. 6. Bayesian belief network for 2-group.

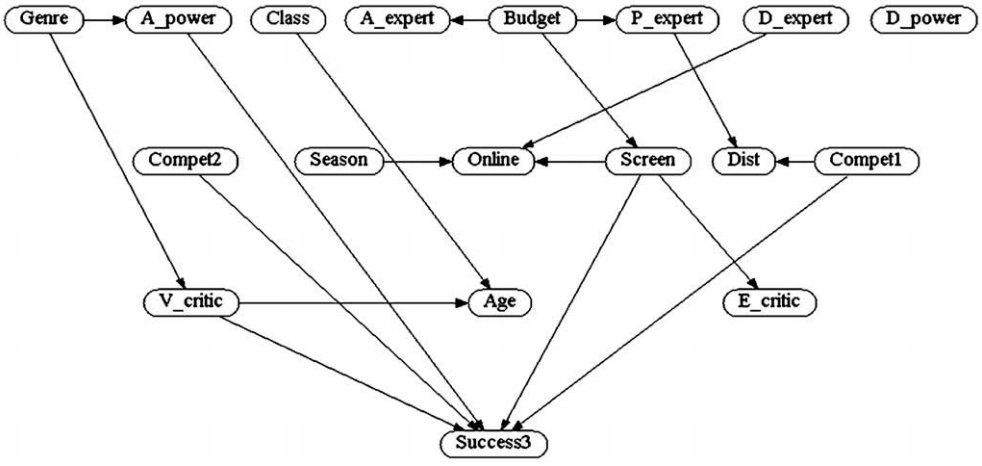

Fig. 7. Bayesian belief network for 3-group.

cess, such as *Success2* and *Success3*, but also on other variables directly or indirectly.

In 2-group case (see Fig. 5), *Screen*, *Season*, *V_critic*, and *Budget* directly affect *Success2*. *A_power* indirectly affects *Success2* via *Screen*. *D_power*, *Compte1*, and *Compet2* indirectly affect *Success2* via *Season*. *Genre*, influenced by *D_power*, also indirectly affects *Success2* through *Compet1* and *Screen* successively.

In 3-group case (see Fig. 6), *Screen*, *A_power*, *V_critic*, *Compet1* and *Compet2* directly affect *Success3*. *Genre* and *Budget* indirectly affect *Success3* via *V_critic* and *Screen*, respectively. However, *D_power* has no relationship with other attributes.

3.3. Sensitivity Analysis

We perform sensitivity analysis based on BBN to find how much it reduces the uncertainty of *Success2* node if we know the information about other nodes. The results, as depicted in Fig. 8, show that *Screen* is the most important variable to reduce entropy, followed by *Season*, *V_critic*, and *Budget*, which are directly connected to *Success2*.

Fig. 9 shows the sensitivity analysis result when box-office performance was divided into 3 groups, excellent, standard and bad, based on the audience number. *Screen* has the biggest effect on *Success3* as shown in 2-group,

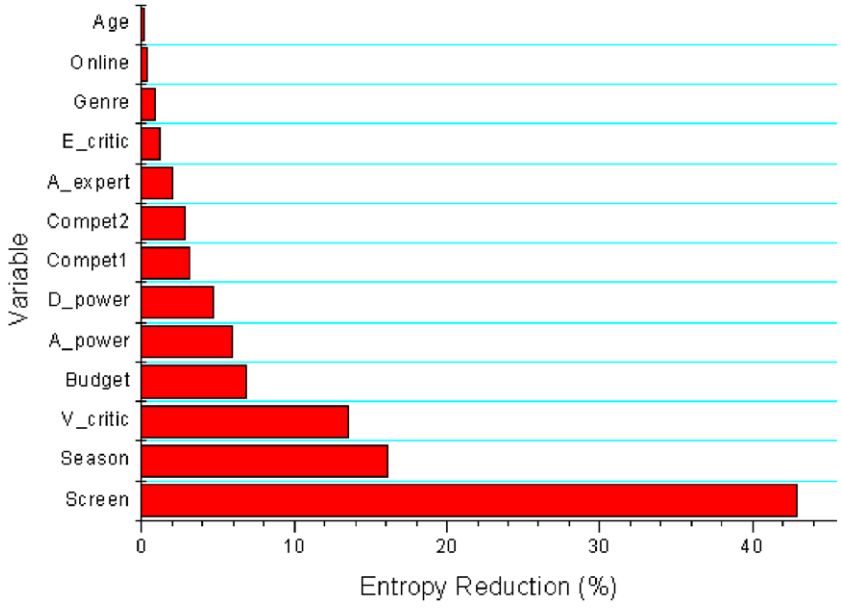

Fig. 8. Sensitivity analysis of *Success2*.

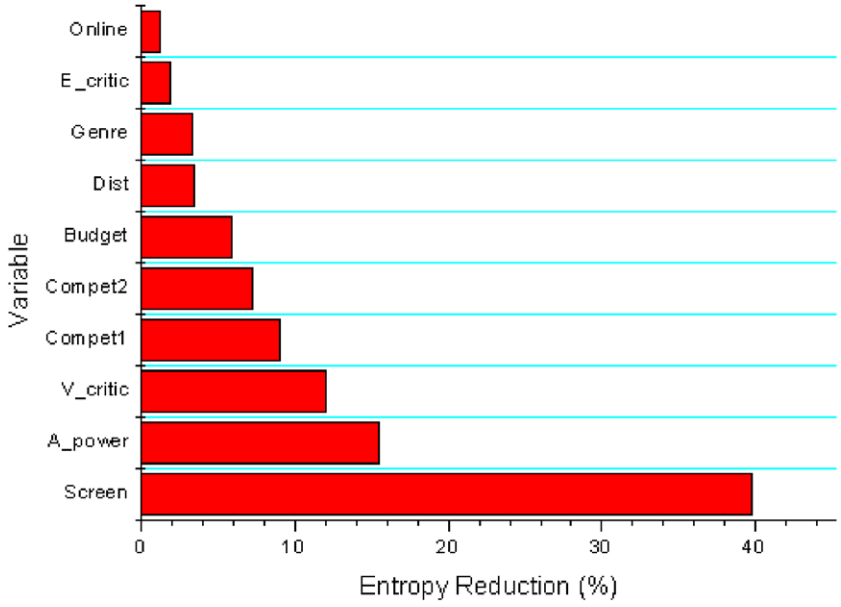

Fig. 9. Sensitivity analysis of *Success3*.

while *A_power*, *V_critic*, *Compet1*, and *Compet2*, which are directly connected to*Success3*, also affect *Success3*. It is notable that *Season*, which is the second most influential variable in reducing the entropy of *Success2*, does not affect *Success3*.

3.4. Success probability

Table 4 shows the probability of *Success2*, which is the chance that the total audience number of a movie is larger than the median number of Korean movies, based on 4 variables, *Screen, Season, V_critic*, and *Budget*, which are directly connected to *Success2*.

The level of *V_critic* shows high correlation with the probability of *Success2*. The low level of *V_critic* is critical to decrease the probability of *Success2*. Although *Screen*, *Season*, and *Budget* are high levels, the probability of *Success2* is only 31% when *V_critic* is low level, which may explain the failures of blockbuster movies. Meanwhile, the high level of *V_critic* is a necessary condition for the high probability of *Success2*. *Screen* is also a necessary condition for the high probability of *Success2*. When a movie has a potential for success, a large screen number is necessary to maximize the success. *Budget* is not correlated with the probability of *Success2*: even a high level of *Budget* induces the lowest probability of *Success2* when *Screen, Season*, and *V_critic* are low levels. Although *Screen, Season, V_critic*, and *Budget* are high levels, the probability of *Success2* is just 50%. However, when *Screen, Season*, and *V_critic* are high levels and *Budget* is middle level, the probability of *Success2* becomes the highest value. The failure of some previous blockbuster movies may explain this.

Table 5 represents the probability of *Success3*, which is the chance that the total audience number of a movie is larger than the top one-third quantile number of Korean movies, based on 5 variables, *Screen, A_power, V_critic*, *Compte1*, and *Compete2*, which are directly connected to *Success3*. *Screen*, *A_power*, and *V_critic* are positively correlated with the probability of *Success3* in general. *Compet1* and *Compete2* are negatively correlated with the probability of *Success3* in general since the high levels of *Compet1* and *Compet2* imply an intensively competitive situation. When *Screen* is low level, and *Compet1*, *Compet2*, and *V_critic* are high level (see cases 1 and 3), a high level of *A_power* has an adverse effect on the probability of *Success3*, which is contrary to our expectations. As expected, when *Screen*, *A_power*, and *V_critic* are high levels, and *Compet1* and *Compet2* are low levels, the probability of *Success3* results in the highest value.

Table 4
Conditional probability table of *Success2*

Case #	Screen	Season	*V_critic*	Budget	Success probability
1	Low	Low	Low	High	0.25
2	High	High	Low	High	0.31
3	Low	Low	Low	Low	0.33
4	Low	Low	High	Low	0.42
5	Low	High	Low	Low	0.50
6	High	Low	Low	Low	0.50
7	Low	High	High	Low	0.50
8	High	High	High	High	0.50
9	Middle	Low	Middle	Middle	0.58
10	High	Low	High	Low	0.64
11	High	High	High	Low	0.68
12	High	High	High	Middle	0.71

Table 5
Conditional probability table of *Success3*

Case #	*Screen*	*A_power*	*V_critic*	*Compet1*	*Compet2*	Success probability
1	Low	High	Low	High	High	0.21
2	Low	Low	High	Low	Low	0.27
3	Low	Low	Low	High	High	0.27
4	Low	Low	Low	Low	Low	0.33
5	Low	High	Low	Low	Low	0.33
6	High	Low	Low	Low	Low	0.33
7	High	High	High	High	High	0.33
8	High	High	Low	Low	Low	0.44
9	Middle	High	High	Low	Low	0.51
10	High	High	High	Low	Low	0.79

As a result, the success formula of Korean movies can be stated as follows. The conditions required to belong to the 'superior' group (upper level of 2-group for box-office success) are large screen number, opening in a high-demand season, high critic ratings from moviegoers, and a moderate budget. The conditions required to belong to the 'excellent' group (the highest level of 3-group for box-office success), which has a clearer meaning of box-office success than the 'superior' group mentioned above, are large screen number, star cast, high critic ratings from moviegoers, and the absence of competitive movies such as currently popular movies and foreign blockbusters.

3.5. Model performance comparison

We compare the forecasting performance of BBN based on domain knowledge with that of ANN, DT and BBN without domain knowledge. For ANN, we use the multilayer perceptron (MLP) neural network architecture that is known to be a strong function approximator for the prediction and classification of problems. ANN and DT are conducted twice according to the original and discrete data sets, where the former consists of both continuous and discrete variables. We also compare the prediction power of BBN between using and not using the domain knowledge. We performed 10-fold cross validation for the reliable results. The forecasting accuracies for 2-group and 3-group are represented in Tables 6 and 7, respectively. The following results are presented.

- The prediction power of BBN using the domain knowledge is better than that not using the domain knowledge.
- The performance of BBN is superior to that of ANN and DT, regardless of the data when the domain knowledge is applied.

Table 6
Model comparison of 2-group

	BBN domain knowledge		Benchmarking models			
			ANN		DT	
	Yes	No	Original	Discrete	Original	Discrete
Set 1	89.1	87.0	78.3	87.0	73.9	76.1
Set 2	89.6	87.5	83.3	89.6	79.2	85.4
Set 3	91.3	84.8	80.4	82.6	65.2	78.3
Set 4	88.1	78.6	69.0	81.0	73.8	83.3
Set 5	82.4	76.5	74.5	84.3	70.6	82.4
Set 6	93.2	86.4	61.4	75.0	75.0	77.3
Set 7	82.1	76.9	79.5	84.6	76.9	79.5
Set 8	85.1	85.1	80.9	95.7	74.5	78.7
Set 9	86.4	84.1	65.9	88.6	79.5	72.7
Set 10	89.4	83.0	72.3	87.2	63.8	78.7
Average	87.7	83.0	74.6	85.6	73.2	79.2

Table 7
Model comparison of 3-group

	BBN domain knowledge		Benchmarking models			
			ANN		DT	
	Yes	No	Original	Discrete	Original	Discrete
Set 1	78.3	72.7	64.0	66.0	66.0	76.0
Set 2	89.6	85.1	50.0	76.1	75.9	76.1
Set 3	76.1	74.2	66.7	78.4	62.7	66.7
Set 4	83.3	82.0	39.2	68.6	43.1	54.9
Set 5	82.4	79.1	43.4	71.7	60.4	71.7
Set 6	88.6	80.9	45.2	71.4	66.7	73.8
Set 7	89.7	87.7	48.0	70.0	56.0	66.0
Set 8	78.7	77.4	42.2	71.1	57.8	70.9
Set 9	84.1	82.8	42.3	80.8	61.5	68.7
Set 10	78.7	74.2	38.0	74.0	48.0	64.0
Average	83.0	79.6	47.9	72.8	59.8	68.9

- When comparing the forecasting accuracy results in 2-group with those in 3-group, the BBN results slightly decrease in 3-group, while those of ANN and DT become much poorer in 3-group.
- ANN and DT with discrete data show better performance than ANN and DT with the original data set.

4. Discussion and conclusion

We suggested the application of a BBN using domain knowledge to predict box-office success. Four research results can be stated. First, the attributes related to box-office performance are interdependent with multiple causal relationships. Second, the prediction power of BBN using the domain knowledge is better than that of ANN and DT, thereby confirming the suitability of BBN to forecast box-office performance. Third, the continuous variable is discretized by the CART algorithm so that BBN can efficiently handle the information from the continuous variables. Finally, the screen number is the most important factor for box-office success in the Korean market, but alone it cannot always assure box-office success because of the causal relationships among the movie attributes.

The critical attributes affecting box-office success are different between *Success2* and *Success3* except for *Screen* and *V_critic*. *Success2* and *Success3* indicate the cases where the movie audience number is ranked in the top 50% and top 33.3% among the movies in the corresponding year, respectively. Thus, we can regard the probabilities of *success2* and *success3* as indices for the success in Korea of non-blockbuster and blockbuster movies, respectively. These results indicate that the necessary conditions for the success of a non-blockbuster movie are the reservation of a large screen number, peak season release such as summer vacation and holidays, good word-of-mouth around the movie, and a production budget no bigger than the upper boundary of the middle level, 29.33 million dollars. Meanwhile, the necessary conditions for the success of a blockbuster movie are the reservation of a large screen number, star casting, good word-of-mouth around the movie, and a low level of competition, i.e., a movie showing without any other popular or blockbuster movies.

This research is meaningful in that we have forecasted box-office performance with consideration for the interdependency between movie attributes and thereby obtained more accurate results than available through any other method. However, the absence of an explanatory variable for marketing expenditure, which is undoubtedly an important factor for box-office success, could have weakened the prediction power of BBN based on domain knowledge, despite of the inclusion of *Dist* and *Online* as alternatives.

The present study only focused on the box-office success of Korean movies. The inclusion in future study of foreign movies' box-office performance will provide a clearer understanding of box-office success in Korea.

Acknowledgement

This work was funded by the Korean Film Council.

References

Ainslie, A., Dreze, X., & Zufryden, F. (2005). Modeling movie life cycles and market share. *Marketing Science, 24*(3), 508–517.

Allbert, S. (1998). Movie stars and the distribution of financially successful films in the motion picture industry. *Journal of Cultural Economics, 22*, 249–270.

Antal, P., Fannes, G., Timmerman, D., Moreau, Y., & De Moor, B. P. (2004). Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine, 30*, 257–281.

Aronsky, D., & Haug, P. J. (2000). Automatic identification of patients eligible for a pneumonia guideline. In *Proceedings of AMIA Annual Symposium* (pp. 12–16).

Baesens, B., Egmont-Petersen, M., Castelo, R., & Vanthienen, J. (2002). Learning Bayesian network classifiers for credit scoring using Markov chain Monte Carlo search. In *Proceedings of the 16th International Conference on Pattern Recognition* (Vol. 3), (pp. 49–52).

Bagella, M., & Becchetti, L. (1999). The determinants of motion picture box-office performance: Evidence from movies produced in Italy. *Journal of Cultural Economics, 23*, 237–256.

Chang, B. H., & Ki, E. J. (2005). Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics, 18*(4), 247–269.

Cheng, J., & Greiner, R. (2001). Learning Bayesian belief network classifiers: Algorithms and system. *Lecture Notes in Computer Science, 2056*, 141–151.

De Vany, A. S., & Walls, W. D. (2004). Motion picture profit, the stable paretian hypothesis, and the curse of the superstar. *Journal of Economic Dynamics & Control, 28*, 1035–1057.

Elberse, A., & Eliashberg, J. (2003). Demand and supply dynamics for sequentially related products in international markets: The case of motion pictures. *Marketing Science, 22*(3), 329–354.

Eliashberg, J., Elberse, A., & Leenders, M. (2006). The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Science, 25*(6), 638–661.

Eliashberg, J., Jonker, J.-J., Sawhney, M. S., & Wierenga, B. (2000). MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science, 19*(3), 226–243.

Eliashberg, J., & Sawhney, M. S. (1994). Modeling goes to Hollywood: Predicting individual differences in movie enjoyment. *Management Science, 40*(9), 1151–1173.

Eliashberg, J., & Shugan, S. M. (1997). Film critics: Influencers or predictors? *Journal of Marketing, 61*, 68–78.

Jedidi, K., Krider, R. E., & Weinberg, C. B. (1998). Clustering at the movies. *Marketing Letters, 9*(4), 393–405.

Jensen, F. (1996). *An Introduction to Bayesian Networks*. Heidelberg, Germany: Springer-Verlag.

Jones, J. M., & Ritz, C. J. (1991). Incorporating distribution into new product diffusion models. *International Journal of Research in Marketing, 8*, 91–112.

KOFIC. (2005). Korean Motion Picture Industry Statistics of the Year, Korean Film Council.

Levin, A. M., Levin, I. P., & Heath, C. E. (1997). Movie stars and authors as brand names: Measuring brand equity in experiential products. *Advances in Consumer Research, 24*, 175–181.

Litman, B. R. (1983). Predicting success of theatrical movies: An empirical study. *Journal of Popular Culture, 16*, 159–175.

Litman, B. R., & Ahn, H. (1998). Predicting financial success of motion pictures: The early '90s experience. *Motion Picture Mega-Industry*, 172–197.

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

Mahajan, V., Muller, E., & Kerin, R. A. (1984). Introduction strategy for new products with positive and negative word-of-mouth. *Management Science, 30*(12), 1389–1404.

Neelamegham, R., & Chintagunta, P. (1999). A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Science, 18*(2), 115–136.

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

Ravid, S. A. (1999). Information, blockbusters, and stars: A study of the film industry. *Journal of Business, 72*(4), 463–492.

Sawhney, M. S., Mohanbir, S., & Eliahberg, J. (1996). A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science, 15*(2), 113–131.

Scohay, S. (1994). Predicting performance of motion pictures. *Journal of Media Economics, 7*, 1–20.

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

Swami, S., Eliashberg, J., & Weingberg, C. B. (1999). SilverScreener: A modeling approach to movie screens management. *Marketing Science, 18*(3), 352–372.

Wyatt, J. (1991). High concept, product differentiation, and the contemporary US film Industry. *Current research in Film: Audiences, Economics, and Law, 5*, 86–105.

Yoo, H. S. (2002). The determinants of motion pictures box-office performances: For movies produced in Korea between 1988 and 1999. *Korean Society for Journalism and Communication Studies, 46*(3), 183–213.

Zufryden, F. S. (1996). Linking advertising to box-office performance of new film release: A marketing planning model. *Journal of Advertising Research*, 29–41.

Zufryden, F. S. (2000). New film website promotion and box-office performance. *Journal of Advertising Research*, 55–64.