Jurnal Teknologi

Full paper

# Forecasting Box-Office Revenue by Considering Social Network Services in the Korean Market

Taegu Kim[a], Jungsik Hong[b*], Hoonyoung Koo[c]

[a]*Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea*
[b]*Department of Industrial and Information System Engineering, Seoul National University of Science and Technology, 232 Gongneung-ro, Nowon-gu, Seoul 139-743, Korea*
[c]*School of Business, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 305-764*

*Corresponding author: hong@seoultech.ac.kr

**Article history**

Received :4 April 2013
Received in revised form :
25 July 2013
Accepted :15 October 2013

**Abstract**

The Korean movie market is an extremely dynamic market. Social network services are also rapidly growing in Korea and comments on movies in social network services (SNS) are increasingly influencing the movie industry. In this paper, we address the issue of forecasting box-office revenue by considering the comments on movies in SNS. We analyze the data in the Korean movie market by using regression analysis and the Bass diffusion model. Our results show that the number of screens is the only significant variable before release, whereas positive and negative mentions on SNS are also essential after release. In addition, the hybrid method provides the idea of employing SNS data into diffusion models for obtaining effective forecasting results.

*Keywords*: Social network service; bass; regression



## 1.0 INTRODUCTION

The Korean movie market is one of the fastest-growing movie markets with a series of significant international successes. Recently, the Hallyu (Korean Wave) phenomenon has expanded the coverage of the Korean entertainment industry in both Western and Asian countries. In particular, movies play an important role in the Hallyu phenomenon. Reflecting the importance of the Korean movie market, diverse research has been conducted on the determinants of success of this market.[1-4] These previous works deal with general success factors such as genre, actors, directors, investment, number of screens, and online scores.

A movie is an experimental product. As a result, a number of studies have attempted to utilize the word of mouth (WOM) effect when forecasting the success of a movie. These studies were based on various sources of WOM, such as interviews and random dialing[5], simulated data[6], and online message boards.[7]

Recently, some researchers have attempted to utilize social network service (SNS) data, which are regarded as the most direct and genuine source of WOM, to forecast movies' success.[8-11] The following two models have been employed for predicting movie success using SNS data: the regression model[12,13] and the Bass diffusion model.[14-16]

Although Abel *et al*.[17] and Lică and Tuţă[18,19] showed that the number of positive mentions on SNS is a more effective factor for forecasting movie success than the total number of mentions on SNS, Liu[13] insisted that the total number of mentions on SNS is the most effective factor. Dellarocas *et al*.[15] showed that movie user reviews have more explanatory power than expert movie reviews. Chakravarty *et al*.[20] insisted that the influence of reviews varies with the movie watching frequency.

Regarding the Korean movie market, research on movie success using SNS data has seldom been conducted despite the high usage of SNS and deep penetration of high-speed internet access in this country. Besides the explosive growth of SNS, SNS analysis websites are emerging to enable researchers to utilize SNS data easily.

In this paper, we propose a new method for forecasting Korean movie success using SNS data and a hybrid of the regression and Bass diffusion models. This is the first attempt to use SNS data for Korean movie forecasting. We verify the utility of SNS data and validate our new hybrid method of the regression and the Bass models with actual Korean movie data.

## 2.0 MATERIALS AND METHOD

### 2.1 Data

We analyzed full-length films released in Korea between October 2011 and August 2012. For the SNS analysis, we selected films that had been made for the Korean market and in Korean. Afterwards, movies with titles that were inappropriate

for a search keyword (e.g., "Mother") were excluded. We also considered both the number of viewers and the length of the screening period. Movies with an audience of less than 10,000 or that had a run shorter than two weeks were ruled out. As a result, we chose 47 of the 103 films for our analysis. The data for box-office information, number of screens, and number of seats were collected from the Korean Box Office Information System of the Korean Film Council (http://www.kobis.or.kr)[21]. The SNS data were collected from a Korean SNS analysis website (pulseK.com)[22]. The SNS data contained keyword mentions, positive mentions, and negative mentions on Twitter, Facebook, and other forms of SNS.

The data collection horizon ranged from three weeks before release to five weeks after release. The screening of almost all movies tends to end four weeks after their release.

### 2.2 Methodology

Two target variables, total box-office revenue and weekly box-office revenue, are used by this paper for forecasting. The regression model is used to forecast the total box-office revenue. Since the weekly box-office revenue is a time series variable and usually shows a truncated S-shaped curve, the Bass diffusion model has been used to forecast the S-shaped curve in recent papers.[14-16]

The Bass diffusion model[23] comprises three parameters: market potential (m), innovation coefficient (p), and imitation coefficient (q). The limitation of the Bass model is that these parameters are estimated by solely using daily box-office revenue data, because this model, unlike the regression model, cannot include several factors that affect movie success simultaneously.

In order to utilize the relative advantages of the two models, we propose a hybrid method. The total box-office revenue, estimated by a regression model using box-office information, number of screens, number of seats, and positive/negative mentions on SNS, is used as the market potential parameter in the Bass model to enable the stable estimation of the other two parameters more easily using daily box-office data. If the market potential can be estimated accurately using the regression model, then it is relatively easier to estimate the other two parameters accurately. Moreover, in cases where limited data is available, the market potential tends to be underestimated and it is difficult to estimate this parameter accurately in most cases.[24] The hybrid method is designed to overcome this problem and provide more accurate estimation by including information other than the time series data.

Our hybrid method can be realized by using the following procedure:

(i) Estimate the total box-office revenue by a regression model of equation (1):

$$Box_{tot} = c_0 + \sum_{i=1}^{n} c_i x_i \,, \qquad (1)$$

where $Box_{tot}$ is the total box-office revenue and $x_i$ is the i-th independent variable corresponding to a movie's success factor.

(ii) The total box-office revenue of a movie, estimated using equation (1), is used as the market potential of the movie. The other two parameters of the Bass model can be obtained using equation (2) as follows:

$$\min_{p,q} \sum_{t=1}^{T} \left[ x_t - \left( N(p,q,t;m) - N(p,q,t-1;m) \right) \right]^2 , \qquad (2)$$

where $N(p,q,t;m)$ is the cumulative adopters (box-office revenue) at time $t$ given the market potential $m$ and $x_t$ is the net adopters at time $t$.

## 3.0 RESULTS AND DISCUSSION

### 3.1 Regression Analysis

Regression models are used for analyzing the major factors affecting box-office revenue. Owing to the different conditions of the available information, the regression models are separately developed for before and after release situations. We developed two equations with total box-office revenue and weekly box-office revenue as target variables for each situation. The before release model contains the number of screens on the day of release, number of seats on the day of release, cumulative number of mentions on SNS, cumulative number of positive mentions on SNS, and cumulative number of negative mentions on SNS.

Target variables for the before release model are total box-office revenue and first week box-office revenue (see Table 1.)

The regression analysis indicates that only the number of screens and the number of seats on the day of release are significant variables for total box-office revenue and first week box-office revenue after release (see Table 2.)

The regression model using only the number of screens on the day of release day as an independent variable has the most explanatory power, as shown by the following equations:

$Box_{tot}$ = -14562444268.5 + 73124889.0153* $sc$ ,

$Box$ = -4216667998.66 + 23911077.7203* $sc$ .

The variables included as potential independent variables in the after release analysis of the major factors affecting box-office revenue are slightly different from those included in the before release analysis. Target variables for the after release model are total box-office revenue and next-week box-office revenue (see Table 3.)

For the after release analysis, all the potential independent variables were found to significantly affect box-office revenue. The most explanatory power was obtained by using the following regression models:

$Box_{tot}$ = -332879757.283 + 0.97696244097* $Box_{sum}$ + 1.61007044823* $Box$ + 1858206.95175* $pos_1$ - 1014127.56871* $pos_{-1}$ - 2417884.22347* $neg_1$ .

$Box_{next}$ = -146213272.748 + 0.700490898335* $Box$ + 610124.891919* $pos_1$ - 418897.039602* $pos_{-1}$ - 494332.898897* $neg_1$ .

The potential independent variables are all meaningful at the 5% significance level. Table 4 shows the adjusted R-square values of the above two regressions, along with the normalized coefficients and variance inflation factors of the independent variables.

As generally expected, realized box-office information is the most important factor for forecasting box-office revenue. Besides box-office information, the number of positive mentions on SNS is the most important factor. Naturally, the coefficient of negative mentions on SNS has a negative value. The negative value of the coefficient of the positive mentions during last week may be caused by the gap between word-of-mouth expectations and the actual experience of watching a movie with higher expectations. For both equations, all VIF values are lower than 5, which implies that they do not suffer from multicollinearity.

The most noticeable difference between before release and after release is the significance of the variables related to mentions on SNS. The result shows that mentions on SNS before a movie release do not significantly affect future box-office revenue. Instead, the number of screens, which indicates the marketing and distribution power, is the only factor for determining future box-office revenue. In the case of after release, SNS reflecting the response of movie-watchers provides meaningful information for future box-office revenue. In contrast to the findings by Liu[13], the number of positive mentions on SNS is a more significant factor than the total number of mentions on SNS. Moreover, we found that besides historical box-office data, number of positive mentions on SNS is the most accurate factor for forecasting box-office revenue. This is consistent with the results of Abel *et al*.[9,17] and Lică and Tuță.[19]

**Table 1** Variables used in before release models (t = 1, 2, 3)

| Variables | Descriptions |
|---|---|
| $Box$ | Box-office revenue for the first week |
| $Box_{tot}$ | Total box-office revenue |
| $sc$ | Number of screens on the day of release |
| $st$ | Number of seats on the day of release |
| $rcg_t$ | Weekly number of mentions on SNS |
| $emo_t$ | Weekly number of positive or negative mentions on SNS |
| $pos_t$ | Weekly number of positive mentions on SNS |
| $neg_t$ | Weekly number of negative mentions on SNS |

**Table 2** Result of regression analysis (values are adjusted R squares)

| Significant Independent Variables | Dependent Variables | |
|---|---|---|
| | $Box_{tot}$ | $Box$ |
| $sc$ | 0.545116 | 0.680093 |
| $st$ | 0.537275 | 0.648547 |

**Table 3** Variables used in after release models

| Variables | Descriptions |
|---|---|
| $Box_{tot}$ | Total box-office revenue |
| $Box_{next}$ | Next-week box-office revenue |
| $Box$ | This-week box-office revenue |
| $Box_{sum}$ . | Cumulative box-office revenue until this week |
| $sc$ | Predetermined number of screens on the first day of next week |
| $st$ | Predetermined number of seats on the first day of next week |
| $rcg_t$ | Weekly number of mentions on SNS (if t = 1, then this week and if t = -1, then last week) |
| $emo_t$ | Weekly number of positive or negative mentions on SNS |
| $pos_t$ | Weekly number of positive mentions on SNS |
| $neg_t$ | Weekly number of negative mentions on SNS |

**Table 4** Result of after release model

| Independent Variables | $Box_{tot}$ | | $Box_{next}$ | |
|---|---|---|---|---|
| | Coefficient | VIF | Coefficient | VIF |
| $Box_{sum}$ | 0.6615 | 2.40 | – | – |
| $Box$ | 0.4007 | 4.00 | 0.8723 | 2.94 |
| $pos_1$ | 0.2057 | 4.84 | 0.3380 | 4.19 |
| $pos_{-1}$ | -0.1107 | 2.53 | -0.2288 | 1.90 |
| $neg_1$ | -0.0969 | 2.26 | -0.0991 | 2.22 |
| *Adjusted* $R^2$ | 0.957788 | | 0.92763 | |

**Table 5** Forecasting total box-office revenue (100 million won)

| Weeks After release | Regression | | Bass | |
|---|---|---|---|---|
| | RMSE | MAE | RMSE | MAE |
| 1 | 58.95 | 40.77 | 83.06 | 157.83 |
| 2 | 34.34 | 20.46 | 30.10 | 69.19 |
| 3 | 22.25 | 13.33 | 16.43 | 44.37 |
| 4 | 23.49 | 12.15 | 9.05 | 22.36 |

### 3.2 Forecasting Comparison

In this section, we compare three different methods for forecasting box-office revenue after release. The regression model in section 3.1 and the Bass model are used for forecasting the total and next-week box-office revenues. The new hybrid model is also used for next-week box-office revenue forecasting. Daily box-office data are used by the Bass model and K-fold cross validation is adopted for comparing the performance of each model. We use the data for 46 out of 47 movies for the estimations for each model and the data from the remaining movie is used to compare the forecasting accuracies of each model.

Table 5 shows the results of forecasting the total box-office revenue using the regression and Bass models. The regression model exhibited greater accuracy than the Bass model, especially in forecasting the total box-office revenue for the first week after release. The difference in the accuracy between the two models may be caused by insufficient data for the Bass model and the use of SNS data for the regression model.

Consequently, when we consider forecasting box-office revenue, the regression model in section 3.1 could be recommended for total box-office revenue and our hybrid method could be more adequate for weekly box-office revenue at an early stage after release. In any case, the SNS data at t - 1 should be considered as a major factor for forecasting box-office revenue at "t."

## ■4.0 CONCLUSION

We analyzed data from the Korean movie market and SNS to identify the factors that influence box-office revenue before and after movie release. Besides the number of screens, the numbers of positive and negative mentions on SNS were found to be the most significant variables in forecasting box-office revenue after release, whereas none of the SNS data variables were found to have the explanatory power to forecast box-office revenue before release. Comparison results from different forecasting methods on after release weekly box-office revenue show that the hybrid method can improve the forecasts obtained by using the Bass model by including SNS data and can provide better forecasts than other methods, particularly one week after release.

Despite notable findings, this study is subject to some limitations, which suggest areas of possible future research. First, the analysis might suffer from a limitation in sampling because only Korean movies were selected for the SNS analysis. Hence, wider research that includes foreign films would contribute to a more general and robust conclusion. Second, only two models are employed in this study, namely simple regression and Bass diffusion models. Applying more sophisticated algorithms, as shown in Elena *et al.*[25], Bakhary *et al.*[26], and Samsudin *et al.*[27], would help to improve the forecasting accuracy. Finally, more consideration needs to be given to the characteristics of the data. For example, heteroskedasticity of time series data can be incorporated in the forecasting model to reflect aspects of the motion picture industry more accurately.

## Acknowledgement

The authors appreciate purseK.com for providing the social network data.

## References

[1] Park, S., and W. Chung. 2009. The Determinants of Motion Picture Box Office Performance: Evidence from Movies Released in Korea. *Journal of Communication Science*. 9(4): 243–276.

[2] Choi, B. and S. Choi. 2011. Determinants of Movie Survival Time in the Korean Movie Exhibition Market. *Journal of Economic Studies* 29(3): 139–160.

[3] Kim, B. and T. Y. Pyo. 2001. Forecasting Model for Box-Office Revenue of Motion Pictures. *Seoul Journal of Business.* 36(1): 1–21.
[4] Kim, H. Y. 2011. Analysis of Spectator Mobilizing Power for 2000's Korea Movies Based on Construction of Network. *Journal of the Korea Contents Association*. 11(1): 429–437.
[5] Zufryden, F. S. 1996. Linking Advertising to Box Office Performance of New Film Releases: A Marketing Planning Model. *Journal of Advertising Research*. 36(4): 29–42.
[6] Eliashberg, J., J. J. Jonker, M. S. Sawhney, and B. Wierenga. 2000. MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures. *Marketing Science*. 19(3): 226–243. doi: 10.2307/193187.
[7] Chintagunta, P. K., S. Gopinath, and S. Venkataraman. 2010. The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets. *Marketing Science*. 29(5): 944–957.
[8] Dhar, V., and E. A. Chang. 2009. Does Chatter Matter? The Impact of User-Generated Content on Music Sales. *Journal of Interactive Marketing*. 23(4): 300–307.
[9] Abel, F., E. Diaz-Aviles., N. Henze., D. Krause., and P. Siehndel. 2010. Analyzing the Blogosphere for Predicting the Success of Music and Movie Products in Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on. 276–280.
[10] Sadikov, E., A. Parameswaran., and P. Venetis. 2009. Blogs as Predictors of Movie Success Stanford University, Technical Report 2009.
[11] Qin, L. 2011. Word-Of-Blog For Movies: A Predictor and an Outcome of Box Office Revenue? *Journal of Electronic Commerce Research*. 12(3): 187–198.
[12] Asur, S., and B. A. Huberman. 2010. Predicting the Future with Social Media. Arxiv preprint arXiv:1003.5699.
[13] Liu, Y. 2006. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. *Journal of Marketing*. 70(3): 74–89.
[14] Wang, F., R. Cai., and M. Huang. 2010. Forecasting Movie-Going Behavior Based on Online Pre-Release WOM and Opening Strength. in Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on. 1–4.
[15] Dellarocas, C., X. M. Zhang., and N. F. Awad. 2007. Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures. *Journal of Interactive Marketing*. 21(4) : 23–45.
[16] Toubia, O., J. Goldenberg., and R. Garcia. 2011. Improving Diffusion Forecasts Using Social Interactions Data. Working Papers.
[17] Abel, F., E. Diaz-Aviles., N. Henze., D. Krause., and P. Siehndel. 2010. *Exploiting the Blogosphere to Forecast Profit of Music and Movie Products*. Technical report, L3S Research Center2010.
[18] Lică, L., and M. Tuță. 2011. Using Data from Social Media for Making Predictions about Product Success and Improvement of Existing Economic Models. *International Journal of Research and Reviews in Applied Sciences*. 8(3): 301–306.
[19] Lică, L., and M. Tuță. 2011. Predicting Product Performance with Social Media. *Informatica Economica.* 15(2): 46–56.
[20] Chakravarty, A., Y. Liu., and T. Mazumdar. 2010. The Differential Effects of Online Word-of-Mouth and Critics' Reviews on Pre-release Movie Evaluation. *Journal of Interactive Marketing*. 24(3): 185–197.
[21] Korean Box Office Information System. Provided by Korean Film Council. Available: http://www.kobis.or.kr/kobis/business/mast/mvie/searchMovieList.dopulsek.com. Available: www.pulsek.com.
[22] Bass, F. M. 1969. A New Product Growth for Model Consumer Durables. *Marketing Science.* 15(5): 215–227.
[23] Van den Bulte, C., and G. L. Lilien. 1997. Bias and Systematic Change in the Parameter Estimates of Macro-Level Diffusion Models. *Marketing Science.* 16(4): 338–353.
[24] Elena, M., M. H. Lee, N. H. Abd Rahman, and N. A. Bazilah. 2012. Fuzzy Time Series and Sarima Model for Forecasting Tourist Arrivals to Bali. *Jurnal Teknologi.* 57(1).
[25] Bakhary, N., K. Yahya, and N. Ng Chin. 2004. Univariate Artificial Neural Network in Forecasting Demand of Low Cost House in Petaling Jaya. *Jurnal Teknologi B.* (40B): 67–75.
[26] Samsudin, R., P. Saad, and A. Shabri. 2012. Hybrid Neural Models for Rice Yields Times Forecasting. *Jurnal Teknologi*. 52(1): 135–147.

.