Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 15(4): 217-226, Dec. 2017 **Regular paper**

Analyzing Box-Office Hit Factors Using Big Data: Focusing on Korean Films for the Last 5 Years

Youngmee Hwang[1*], Kwangsun Kim[2], Ohyoung Kwon[3], Ilyoung Moon[3], Gangho Shin[4], Jongho Ham[5], and Jintae Park[3], *Member, KIICE*

[1]School of General Education, Sookmyung Women's University, Seoul 04310, Korea
[2]Department of Mechatronics Engineering, Korea University of Technology and Education, Cheonan 31253, Korea
[3]Department of Computer Science & Engineering, Korea University of Technology and Education, Cheonan 31253, Korea
[4]Division of Theater and Cinema, Daejin University, Pocheon 11159, Korea
[5]College of Liberal Arts and Cross-Disciplinary Studies, University of Seoul, Seoul 02504, Korea

Abstract

Korea has the tenth largest film industry in the world; however, detailed analyses using the factors contributing to successful film commercialization have not been approached. Using big data, this paper analyzed both internal and external factors (including genre, release date, rating, and number of screenings) that contributed to the commercial success of Korea's top 10 ranking films in 2011–2015. The authors developed a WebCrawler to collect text data about each movie, implemented a Hadoop system for data storage, and classified the data using Map Reduce method. The results showed that the characteristic of "release date," followed closely by "rating" and "genre" were the most influential factors of success in the Korean film industry. The analysis in this study is considered groundwork for the development of software that can predict box-office performance.

Index Terms: Big data, Box office analysis model, Box office internal/external factors, Korean film analysis

I. INTRODUCTION

The Korean film industry is now the 10th largest in the world. According to statistical data from the Korean Film Commission, domestic film production is showing continuous growth, with 183 works produced in 2013. This is an almost 50% increase from the 118 works produced in 2008. The total revenue of film industry (including ticket sales, optional markets, and export) in 2013 reached a record-breaking KRW 1.8839 trillion. There have been nine Korean films and three foreign films ("Avengers", "Interstellar", and "Frozen") that have recorded over 10 million views in the last five years. Considering the population of Korea is 50 million, having several films with over 10 million audience members each year is an extraordinary feat. According to Box Office Mojo, revenues for the films "Frozen" and "Interstellar" have been recorded as the second highest in the world, only behind that of the United States.

The amount of capital and investment in the film industry has expanded in proportion to the thriving growth of the Korean film market. Considering the substantially unstable nature of the film industry to either dramatically fail or succeed, the necessity of an engineering approach to insuring investment stability has become increasingly clear.

It is crucial at this point to extract and analyze the various

Received 15 October 2017, **Revised** 28 October 2017, **Accepted** 08 November 2017
***Corresponding Author** Youngmee Hwang (E-mail: hym4322@sm.ac.kr, Tel: +82-2-710-9825)
School of General Education, Sookmyung Women's University, 100, Cheongpa-ro 47-gil, Yongsan-gu, Seoul 04310, Korea.

Open Access **https://doi.org/10.6109/jicce.2017.15.4.217** print ISSN: 2234-8255 online ISSN: 2234-8883

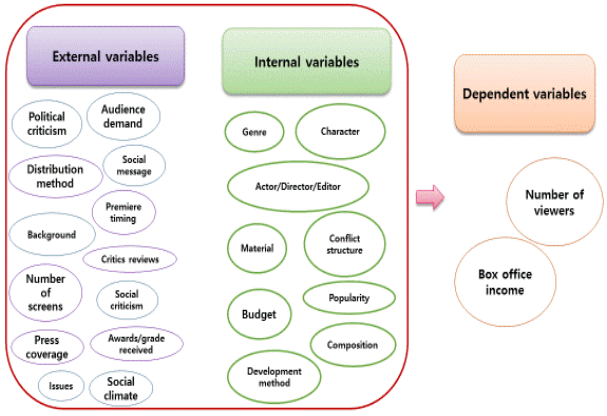

Fig. 1. External/internal variables.

factors that contribute to box office success, ultimately paving the way to predicting a film's success while still in its production stage. For this purpose, the authors analyze the objective indices of the internal and external textual factors that affect box-office hits. The phrase 'internal factors' is hereby used to characterize elements of the preproduction-postproduction stage of filmmaking. Whereas 'external factors' refers to circumstances that influence the film after the complete production, such as distribution, reviews from media and critics, advertisement, word-of-mouth, and the social ambience at the time of release (Fig. 1). Previous studies fail to address the influence of internal factors on the success of films, which makes their contributions to the study of film production and industrial development almost negligible. Since these studies have been performed sporadically on relatively few works and concentrated on only external factors, they have not fully analyzed the actual production processes within the film industry. However, to truly attain investment stability, it is logical and necessary to explore and analyze the internal factors of film production and their influence on success in the box office.

To attain this comprehensive analysis, the authors turned to the technology of big data. Beyond the abilities of pre-existing database management tools to collect, save, manage, and analyze data, big data technology extracts value from massive datasets and analyzes the results. Big data technology is expected to make accurate predictions about diverse societies and provide customized information to individual researchers. The significance of big data is being emphasized because big data comprises information provided from all areas of study, including social, economic, cultural, scientific, and technological fields. Many researchers have used big data in the past in their construction of prediction models. A case example is by Salehan and Kim [1], which utilizes big data to analyze consumer patterns, predict demand, and proceeded with research on schemes that can be utilized for marketing strategies. Through the utilization of big data, the authors aim to draw more accurate, integrated results that reflect all possible influential factors. As for the analysis of the data, MapReduce, and data mining are the representative techniques that are used for analysis of big data. The authors too, used these techniques for analyzing the pattern of data and utilizing it as a material for analysis.

Another distinctive aspect of this paper is the target market. Previous international studies have examined various success factors. However, whether these findings hold true or can be applied in the domestic market of South Korean cinema is now being explored. Therefore, this paper seeks to verify previous international findings and further analyze the contributing factors. The subjects of this research are Korean movies that have ranked among the top 10 positions from 2011–2015. The authors have focused their research on this specific time period because of the significant changes to the Korean film industry brought on by the development of multiplexes. Because of this significant shift, it was logical that the research should focus on the film revenues after the change. Although the result of this research is intended to contribute to the production and investment, it can also be a crucial reference for foreign films seeking distribution in Korean theaters.

The analysis model provided in this study is established by the preparation of basic data that will enable meaningful evaluation throughout the film production environment. This study aims to prioritize internal factors, examining which characteristics hold more importance in determining the potential success of a film. This comparison can be a practical contribution to decision-making in the production stage, offering strict data that indicates the gravity of each variable. Furthermore, the analysis of all internal and external factors will form a database for an integrated system, ultimately allowing for the development of a program with which the authors can draw an accurate prediction of box office success. This study is the groundwork for the establishment of such program.

II. LITERARY REVIEW

The most basic process in analyzing the film industry is investigation of the factors that influence film success. Research on the structural factors that determine box office hits has been conducted since the 1970s in the fields of economics, business administration, and journalism. Wang et al. [2] pointed out that as the audiences' consumption attitudes, consumption patterns, and consumer groups are all in the midst of great changes, it is necessary to improve the film's revenue potential by excellent script selection, accurate market positioning, effective product marketing,

and accurate forecasting of box office trends. Therefore, many studies are under way to exploit box office prediction. Recent studies that examine contributing factors of film success are as follows.

Fetscherin [3] analyzed success-prediction factors in Indian Bollywood films, including production, brand, distribution, and consumer factors. Genre and rating were included in production factors; star and director power were included in brand factors; season of theatrical release, number of screens, and distribution power were included in distribution factors; and consumer evaluation was included in consumer factors. According to the research, the level of influence of each factor was ranked from most important to least as distribution, production, brand, and consumer factors. The authors aim to verify whether these findings also hold true for the Korean market.

According to Treme [4] in his study, "The exposure of the star actor of the film in media on the film's marketing success", exposure of the leading actor in media does not have an influence on sales. Although this is a notable conclusion, it is still necessary to study the correlation of box office success with other factors. However, the exposure of leading actors in media has not much influence on reducing the investment risk of films. This study contemplates ways to reduce risk in the investment of films, which is a more profound issue in the industry.

Mestyan et al. [5] conducted research on a predictive model that showed "the popularity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia, the well-known online encyclopedia." In another study, Liu [6] used the data from the Yahoo Movies website to explain the relationship between online word-of-mouth and box office revenue. In this study, the total volume and valence were investigated with results indicating that the volume of word-of-mouth recommendations most correlates to box office results. Similarly, Duan et al. [7] constructed a dynamic equation system with the effect of online word-of-mouth being the precursor, and retail sales the outcome. The results were exactly the same with word-of-mouth volume being a critical indicator of box office success. The distinction to be made with the authors' research with regard to these studies is that they revolve around online word-of-mouth, which is an element that takes effect after the distribution. The authors' study aims to research internal and external factors of film success in order to reduce the production risk of films before they enter distribution and inform decisions regarding the amount of safe investment along with other decisions that must made prior to distribution.

Simonton [8] evaluated the definition of success for films, with the conclusion that three key indicators, namely, critical evaluation, financial performance, and movie awards, are responsible for the determination of whether a film has been successful or not. Simonton further explained the predictors of film success and how they were related to each of the success indicators. The predictors were categorized as either "production" or "distribution." The former includes budget, screenplay (sequels, remakes, adaptations, true stories), genre, Motion Picture Association of America (MPAA) rating, run time, and personnel (producer, director, actor, composer). The latter includes the release date, number of screenings, major distributor, marketing expenditures, and market competition. The limitation of his study lies in his failure to explain the list factors that were identified as influential. Our study aims to compare the gravity of influential factors to aid practical investment decisions.

Lash and Zhao [9] proposed a decision-support system that could contribute to the process of film investment. Their system used historical data of films through an 11-year period, categorizing them with respect to "who" is related to a film, "what" a film is about, and "when" a film will be released. It analyzed more than one category under "hybrid," investigating how the chemistry of two major factors correlates to box office success. The research identified these relationships through a regression analysis. While the study of Lash and Zhao has a solid categorization of box office success elements, the authors of this study wish to improve objectivity through the utilization of big data.

The box office prediction researches that focus on the Korean cinema market are as follows. Lee and Chang [10] adopted a Bayesian belief network to construct an equation that factored in many of the same variables as mentioned in Simonton's text to investigate the casual relationship between these film attributes and box office performance. Here, the variables were inserted without taking into account whether they are those of preproduction, production, or post production. Instead, this study categorized the variables to the stages of production in which they take place. It is the difference that will allow the authors of the presented study to endow accuracy and efficiency in reducing production risks. Furthermore, the study was conducted in 2009, and as mentioned in the introduction, the Korean cinema market has changed greatly since that period. For a more accurate result, our study focuses on recent films, ranking in the top 10 audience between the years 2011 and 2015.

Park and Song [11] analyzed the relationship between production costs and box-office performance. According to Park, film rating, film genre, the star power of actors, and directors, premiere screen scale, distributor power, critical reviews, and online audience evaluation vary according to production costs, which revealed that production cost is an important variable in box office success. Yoo [12] and Kim [13] also consider these variables in their studies which

contribute to the domestic research on the influence of factors on box office performance. In addition, Park and Lee [14] analyzed the influence of newspaper coverage of premiered films on box office success. Park et al. [15] is also a worthwhile reference because it explains that the significance of data technology as a tool that can be used to refine and analyze a large amount of data in real time is substantially increasing. Users who have significant influence on social networks can be made more recognizable and this knowledge can be used to propose an application plan for long-term marketing. Cho et al. [16] stated that existing analytic research that considers job, age, and time period is necessary in conjunction with psychological pattern analysis and analyses with big query software combined with atypical data from opinion mining analysis and open source processing. When predicting a box office hit, they believe that an error range of less than ±7% (as compared with existing marketing analysis) can be derived and could be a significant standard for use as strategic promotion and marketing data for film producers or marketers. In addition, they showed the increasing influence of social networking technology. After performing online-review mining during the premiere week of 209 Korean films that were screened in 2013, the analyzed results were used to propose a box office hit record-predicting model.

In addition to the provided studies, the authors, Hwang et al. [17] conducted an analysis with one of the internal factors of the film, namely plot development, in their previous study. Specifically, the authors targeted instances that triggered the emotional responses of laughter and tears that appeared throughout the story. A plot has five stages: exposition, rising action, crisis, climax, and ending. The authors attempted to analyze successful films by genre, and ascertain in which part of the scenario laughter and tear inducing factors were focused. The results of the analysis would contribute to the construction of a comprehensive database for box office predictions in future scenarios. For this specific case, the authors adopted a web crawler program to gather relevant data. First, the authors collected terms relevant to laughter and tears. Next, the authors schematized the frequency of laughter and tear inducing factors as they could be found within the five plot stages of box office hit films by genre. The result of this small-scale study found that most commercially successful films, especially historical films, have laughter and tear inducing factors focused at the “rising action” phase of the plot.

III. METHODS

Our ultimate goal is to develop a prediction evaluation program for box office success by using internal and external variables relating to existing box office hits (e.g., films that have ranked among the top 10 in audience viewership from 2011 to 2015). The specific goal of this study is to provide a database for this program. Several independent variables must be examined in order to establish this analysis system. The analysis of these variables is represented as a final target value that can evaluate box office performance.

Analyzed factors (number of theaters, actual number of viewers, screening date) and evaluation by the public are collected using Open API. In addition, statistical data of the film industry from the Korean Film Commission is used to collect contextual and formal aspect variables. The big data Hadoop system in the Department of Computer Engineering at the Korea University of Technology and Education was utilized in this study to ensure efficient and stable data management.

A. The Method of the Data Collection and Categorization

To collect the necessary data for this study, the authors sorted the subject films by highest audience record to lowest. The information on each film was collected through the film promotion committee and local portal sites. For this study, the authors have collected the internal and external textual elements. The internal elements under consideration included movie directors, actors, and cinematic genre. The external elements rendered for analysis concerned the number of screenings for each film, profits, and release date.

In addition, the authors designed and developed a Web Crawler using Python and Java to collect data through web content mining.

The most important role of the Web Crawler is that it extracts unstructured data from the web through keywords, such as film titles, and saves them in a database. However, there is a discrepancy between web pages in terms of their structures and access levels. To overcome this problem, we switched the HTML Document Tree to regular expression that improved the reliability of the search.

The Web Crawler plays two roles. The first role is to read the contents of a particular site into a text file and save it. The second role is to collect relevant contents based on a specified keyword. The use of web data was read by Python and Crawler 4, and stored on the server were Java and POI libraries. Fig. 2 illustrates the design of a web crawler.

The data collected through the Web Crawler was sorted by the number of moviegoers on a scale of 1 to 50. Through this process, this study compared and analyzed the elements of the film text both internal and external.

To allow for analysis, the authors categorized the internal and external factors such as the title, directors, actors, and release dates for films that have ranked the 10 films with the

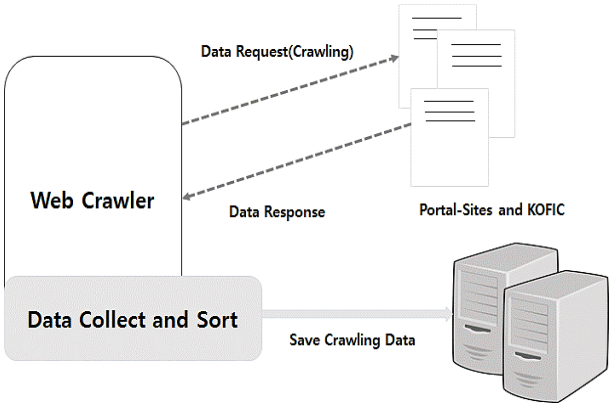

Fig. 2. Web Crawler design.

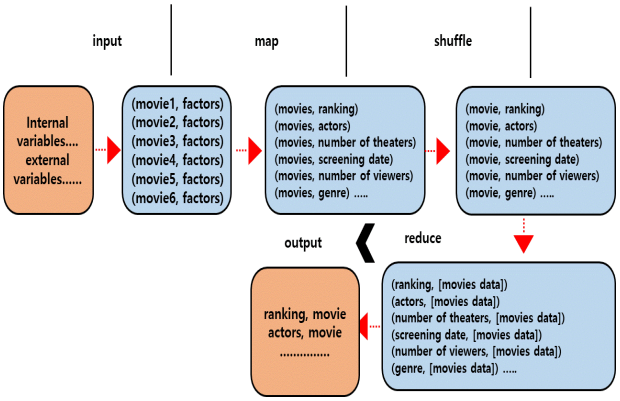

Fig. 3. MapReduce data processing.

highest viewership in the last 5 years.

B. The Method of the Data Analysis

We performed data mining of film element data for box office prediction. Olson and Wu [18] mentioned that data mining has proved worthy in almost all surveyed areas. Manjunath et al. [19] asserted that most cases of data mining for the purpose of data analysis is to extract values that are determined meaningful, or that can serve as a standard value. However, the data subjected for analysis in this research are the internal and external factors such as the title, directors, actors, and release dates for specified films. Therefore, most of the key data was comprised of text instead of numeric values. This led to several difficulties in the process of analysis and the determination of profundity in given texts and analyzed values.

The technique used in this study to analyze factors that influence box office performance from data constituted of text was the Term Frequency-Inverse Document Frequency (TF-IDF) method. This technique, analyzes the importance of data by examining the frequency of certain values or related figures within provided data, then uses the value TF and IDF in the analysis. According to Gupta et al. [20], Term Frequency is a numerated figure of the times that a certain value (or in this case, text) occurs in the given data, serving as an index of how frequently the value appears. Inverse Document Frequency, a reciprocal of TF, is used in data analysis to adjust meaningless data by heavily weighing occasionally appearing data, and lightly weighing frequent data, leading to the extraction of core data within provided data.

The multiplication of the function of these two values (TF, IDF) produces the final TF-IDF value. The value of TF-IDF rises as the target value appears more frequently in the selected data category, and as it appears less frequently in the total provided data. As this value can filter irrelevant data and extract key values, this study used the TF-IDF technique to analyze factors (values) that have heavily influenced the box office performance of films in the provided data.

Furthermore, we used Hadoop Framework to execute data mining based on the Web Crawling data, calculated TF, IDF and TF-IDF, and analyzed the performance of the data mining program.

The reason for employing the Hadoop Framework for this study is largely related to reliability, performance, and expandability. According to Jain and Bhatnagar [21], because the Hadoop Framework is based on a distributed process platform, parallel processing through several nodes makes it appropriate for handling large amounts of data, diminishing the need for advanced equipment. The framework provides a desirable environment for expanding additional data for analysis in the future.

In a typical MapReduce processing function, the master node receives a data Set from Crawler. The input data set is sent to each slave node according to the state of the sub-slave node. Each slave node classifies the received data based on the key, creates reducer through shuffle-reduce process, and stores the result in its own database. The used key means elements necessary for prediction of movie entertainment (text internal and external factors), and from the data of each movie, it is reclassified based on these elements. Also the physically connected slave node creates a copy of the generated reducer data and transfers it to another node for storage. When data loss occurs, it will automatically recover the text by using that copy. The logic for processing the data is represented in Fig 3.

C. Importance of Analysis between Internal and External Textual Elements

As the data collected for this research was text-based, analysis of the TF, IDF, and TF-IDF values were compiled to digitize the frequency of text appearance. Standard TF and IDF values define a particular word as important when

it appears often in a document. This study added weight to the standard process of analysis in extracting analysis results.

All the elements were made into one matrix, and the respective TF, IDF, and TF-IDF were obtained. The importance was analyzed using the following expression for each numerical calculation.

$$TF = count(T)/Rn, \quad (1)$$

$$IDF = log(Rn/count(T)), \quad (2)$$

$$TF - IDF = TF * IDF. \quad (3)$$

A TF represents the ratio at which reference elements appear while IDF represents the percentage of movies in which reference elements appear. TF-IDF is a value obtained by multiplying TF by the IDF value. For each of these elements, the higher the value, the greater the importance, and TF-IDF is the final criterion for determining importance.

In the case of actors and directors as elements, there are no overlapping elements in the matrix. For TDM, values were obtained based on frequency in which elements appear in 50 whole movies, actors and directors as elements have values close to zero. This means that the effect of this factor on box office revenue is low. Moreover, when it comes to the number of screenings compared to the sales amount, the higher the ranking of these elements, the higher their numerical values; hence, this factor is not suitable for the box office prediction.

Therefore, the remaining elements, such as genre, opening time, and rating of importance, were evaluated, and the results are shown in Table 1.

Based on the analysis results, the average of TF-IDF for each element is as shown in Table 2. Based on the result of the analysis, the opening timing element is the most important factor, followed by rating and genre.

Table 1. Units for magnetic properties

Term (T)	No. 1 to No. 50	TF	IDF	TF-IDF
Ge_Crime	8	0.16	0.795880017	0.127640803
Ge_Action	14	0.28	0.552841969	0.154795751
Ge_Drama	36	0.72	0.142667504	0.102720603
Ge_Comedy	12	0.24	0.619788758	0.148749302
Ge_SF	1	0.02	1.698970004	0.0339794
Ge_Adventure	2	0.04	1.397940009	0.0559176
Ge_Romance	5	0.1	1	0.01
Ge_War	3	0.06	1.22184875	0.073310925
Ge_Horror	1	0.02	1.698970004	0.0339794
Ot_Spring	8	0.16	0.795880017	0.127340803
Ot_Summer	15	0.3	0.522878745	0.156863624
Ot_Fall	8	0.16	0.795880017	0.127340803
Ot_Winter	19	0.38	0.420216403	0.159682233
Ra_12	11	0.22	0.657577319	0.14466701
Ra_15	30	0.6	0.22184875	0.13310925
Ra_18	9	0.18	0.744747495	0.134050949

Table 2. Result of importance rating for box office

Element	Average of TF-IDF	Importance rating
Genre	0.09231042	3
Opening time	0.142806866	2
Rating	0.137275736	1

IV. RESULTS

A. Analysis of External Textual Elements

Several studies discussed earlier in this study commonly pointed to the number of screens as the biggest influence of a film's performance. The authors utilized big data to test this argument on the correlation between the number of screenings and the number of viewers.

This study also conducted an analysis on the external textual element (revenue, number of screens, audience rating, release date) of existing successful films (top 10 from 2011 to 2015), from a ranking list based on the number of audience (1st to 50th).

It was expected for the number of screens to have a strong positive relationship with the film ranking, but the analysis results showed that the ranking was dispersed evenly throughout the number of screens (see Fig. 4). In other words, there were cases in which a film ranked low despite having high screen numbers, and vice versa.

In the case of audience rating on film ranking, every film that ranked in the top 10 were rated PG-13, and the rest of the rankings were dispersed throughout PG and R ratings (see Fig. 5).

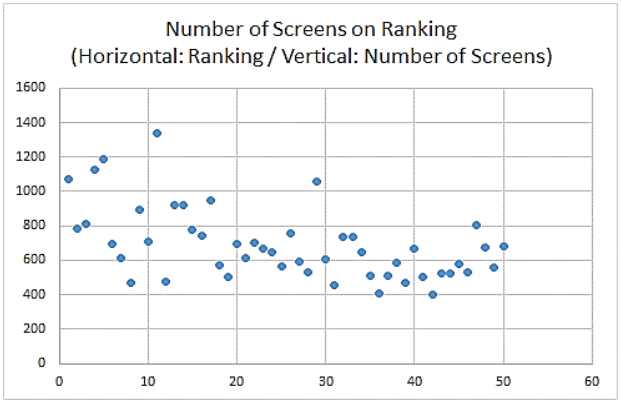

Fig. 4. Number of screens on ranking.

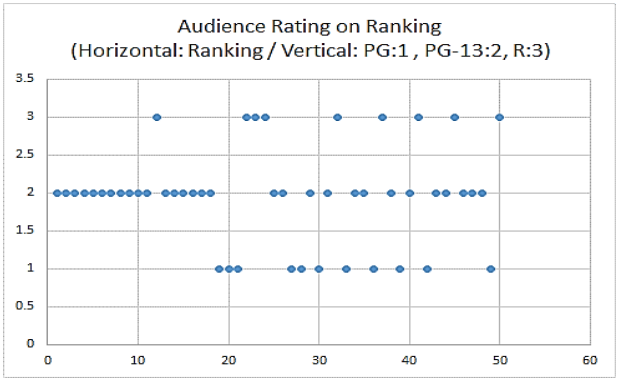

Fig. 5. Audience rating on ranking.

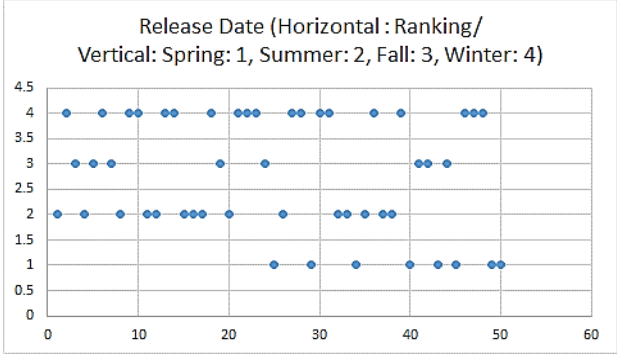

Fig. 6. Release date on ranking.

Regarding release date, films that were released in summer and winter mostly succeeded, whereas most of the films in the bottom ranks were released in spring and fall (Fig. 6).

As there are good reasons to believe that the genre of the film also has significant links to the success of film in relation to release date, further analysis that captures both of these elements is deemed desirable.

B. Analysis of Internal Textual Elements

Among various internal textual elements, this study used genre as a key factor in analyzing existing successful films (top 10 during the last 5 years), from a ranking list based on the number of audience (1st to 50th). This paper also examined the results with the release date, which is an external element, taken into consideration.

The results show that films in the genre of drama most frequently appear to be successful when compared with other genres (Fig. 7). Additionally, a cross analysis on an internal element (genre) and an external element (release date) revealed that although the drama genre produced the largest number of successful movies, the season in which they were released had more influence on their box office

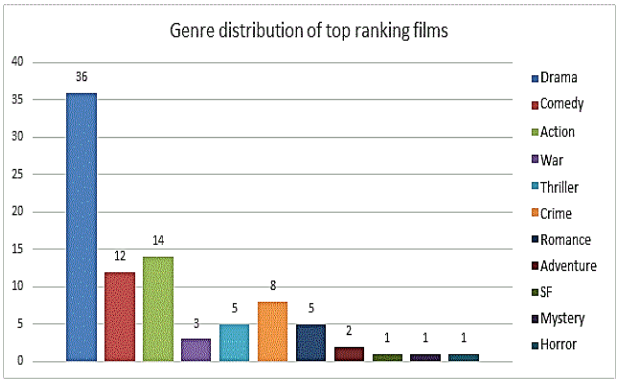

Fig. 7. Genre distribution of top ranking films.

performance. In the case of directors and actors, a lack of consistency as well as multiple overlapping elements made it impossible to extract meaningful data.

V. DISCUSSION AND CONCLUSION

Regarding the success of Korean films, this study will lay the groundwork for an opportunity to deviate from existing superficial and sporadic discussions. As asserted by Sowmya [22] big data analysis has already shown its potential to improve decision-making, risk minimization and to develop new products and services. By engineering research methodology using big data for use in film research and by addressing the internal and external factors of film production, this study deviates from the existing frame of research to arrange an integrated and a systematic research forum for films.

Using Web Crawler, the authors gathered information on film elements present online, constructed data sets, and performed data mining and analysis using Hadoop Framework and MapReduce method. The influence of internal and external factors on film production were not only verified, but also evaluated for their significance on film success. As each factor was established in text, it was necessary to quantify them for big data analysis. However, results showed that factors such as directors, genre, and actors were not suitable for quantification. To overcome this issue, each factors was subjected to analysis (TF, IDF, and TF-IDF) based on text frequency for meaningful data extraction. In the analysis method, we attempted to calculate TF, IDF and TF-IDF for each element, and to establish the extent of the performance based on the result.

When classifying box office hit elements into internal and external textual elements, the amount of data that can be obtained is vast. Therefore, an integrated management system for the enormous amount of data must be established, and the development of programs that can efficiently use

this data is required. A comprehensive judgment will be possible by correlating the internal elements found in film production and the external ones that are applied in distribution and marketing.

Among the various internal variables, the authors were able to draw a conclusion on the influence of genre on box office success in Korea. Other variables including actors, directors, and editors had shown little consistency. From the external variables, the authors examined the number of screenings, release dates, and rating. Although the number of screens at the period of release was pointed out as an influential variable from existing studies, this study proved that it does not have much influence on Korean audience viewership.

Regarding the release date, the Korean cinema market showed similarities with the international study results, as most films with high audience records were released in either summer or winter. Finally, this paper discovered that when the authors arranged every film that has ranked in the top 10 from 2011 to 2015, the 10 most viewed films were from the PG-13 group. Additionally, the authors attempted to weigh the significance of two different variables by comparing the influence of an internal element (genre) and an external element (release date). Although the films of the drama genre had been found to be the most popular in the Korean cinema market, the release date of the films had more influence on their success. As this analysis was extracted from Korean film data, this may not hold true for any other markets. For instance, according to Lash and Zhao [9] the genre of comedy is the most popular when examining the American film industry. The differing results between existing foreign research and Korean research can be explained by the consumption pattern of middle and high school students. Middle and high school students have significant 'ticket power' in Korea. This may explain the overwhelming trend of popularity among PG-13 films as well as he relationship of school vacations to the popular release dates, because these students have more time for leisure activities during summer and winter vacations, the biggest indicator for film success in Korea is the release date within the summer and winter seasons.

By examining, collecting, organizing, and processing these substantial Korean film case studies, a systematic data conservation and its efficient use in production processes can be promoted. In addition, precise analysis and database establishment of film internal and external elements can be provided sufficiently and can serve as practical data in film production. Through the findings of this study, the Korean film industry may gain insight into future growth. The efficiency of film production can be maximized through existing film data analysis, which will contribute to the minimization of risk.

REFERENCES

[1] M. Salehan and D. Kim, "Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics," *Decision Support Systems*, vol. 81, pp. 30–40, 2016.

[2] X. Wang, Y. Wang, J. Chai, X. Feng, and Z. Liu, "The big data applications in film industry Chain," *International Journal of Computer Applications*, vol. 9, no. 12, pp. 1–8, 2016.

[3] M. Fetscherin, "The main determinants of Bollywood movie box office sales," *Journal of Global Marketing*, vol. 23, no. 5, pp. 461–476, 2010.

[4] J. Treme, "Effects of celebrity media exposure on box-office performance," *Journal of Media Economics*, vol. 23, no. 1, pp. 5–16, 2010.

[5] M. Mestyan, T. Yasseri, and J. Kertesz, "Early prediction of movie box office success based on Wikipedia activity big data," *PLOS One*, vol. 8, no. 8, article no. e71226, 2013.

[6] Y. Liu, "Word of mouth for movies: its dynamics and impact on box office revenue," *Journal of Marketing*, vol. 70, no. 3, pp. 74–89, 2006.

[7] W. Duan, B. Gu, and A. Whinston, "The dynamics of online word-of-mouth and product sales: an empirical investigation of the movie industry," *Journal of Retailing*, vol. 84, no. 2, pp. 233–242, 2008.

[8] D. K. Simonton, "Cinematic success criteria and their predictors: the art and business of the film industry," *Psychology & Marketing*, vol. 26, no. 5, pp. 400–420, 2009.

[9] M. T. Lash and K. Zhao, "Early predictions of movie success: the who, what, and when of profitability," *Journal of Management Information Systems*, vol. 33, no. 3, pp. 874–903, 2016.

[10] K. J. Lee and W. J. Chang, "Bayesian belief network for box-office performance: a case study on Korean movies," *Expert Systems with Applications*, vol. 36, no. 1, pp. 280–291, 2009.

[11] S. H. Park and H. J. Song, "The determinants of motion picture box office performance: evidence from Korean movies released in 2011," *Journal of Social Sciences*, vol. 51, no. 1, pp. 45–79, 2012.

[12] H. S. Yoo, "The determinants of motion pictures box office performances: for movies produced in Korea between 1988 and 1999," *Korean Journal of Journalism & Communication Studies*, vol. 46, no. 3, pp. 183–213, 2002.

[13] E. M. Kim, "The determinants of motion picture box office performance: evidence from movies exhibited in Korea," *Korean Journal of Journalism & Communication Studies*, vol. 47, no. 2, pp. 190–220, 2003.

[14] J. W. Park and G. O. Lee, "A study on the effects of the newspaper coverage of motion pictures on box office performance," *Korean Journal of Journalism & Communication Studies*, vol. 48, no. 6, pp. 62–83, 2004.

[15] H. S. Park, Y. H. Lee, and Y. H. Song, "Research on influencer of social networks for long-tail marketing," in *Proceedings of the Korea Society of Management Information System Spring Conference*, pp. 774–787, 2014.

[16] S. Y. Cho, H. K. Kim, B. S. Kim, and H. W. Kim, "Research on box office hit and prediction through online review mining," in *Proceedings of the Korea Society of Management Information System Spring Conference*, pp. 671–684, 2014.

[17] Y. M. Hwang, J. T. Park, I. Y. Moon, K. S. Kim, and O. Y. Kwon, "The Box-office success factors of films utilizing big data-focus on laugh and tear of film factors," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 20, no. 6, pp. 1087–1095, 2016.

[18] D. L. Olson and D. Wu, "Predictive models and big data," in *Predictive Data Mining Models*. Singapore: Springer, pp. 95–97, 2016.

[19] R. Manjunath, R. K. Channabasva, and S. Balaji, "Big data MapReduce Hadoop distribution architecture for processing input splits to solve the small data problem," in *Proceedings of 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, India, pp. 480–487, 2017.

[20] P. Gupta, A. Sharma, and J. Grover, "Rating based mechanism to contrast abnormal posts on movies reviews using MapReduce paradigm," *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, Noida, India, pp. 262–266, 2016.

[21] A. Jain and V. Bhatnagar, "Movie analytics for effective recommendation system using Pig with Hadoop," *International Journal of Rough Sets and Data Analysis*, vol. 3, no. 2, pp. 82–100, 2016.

[22] T. S. R. Sowmya, "Cost minimization for big data processing in geo-distributed data centers," *Asia-Pacific Journal of Convergent Research Interchange*, vol. 2, no. 4, pp. 33–41, 2016.

Youngmee Hwang
received her B.S., M.S., and Ph.D. degrees in the Department of Korean Modern Literature from Sookmyung Women's University, Seoul, Korea, in 1980, 1995, and 1999, repectively. She is a professor of School of General Education, Sookmyung Women's University from 2002. She is also a film critic. She is presently the president of the Korean Society for Thinking and Communication and the director of the Korean Society for Engineering Education (KSEE). Her research focuses on education using films and interdisciplinary education. Also she studies the relation between factors in Film and forecasting box office hit utilizing big data. She is also the president of FIPRESCI KOREA (Korea association of International federation of film critics). She was a member of Jury of FIPRESCI Prize in Busan (2015), Cannes (2013) and Berlin (2012) International Film Festival.

Kwangsun Kim
received his B.S. degree from the Department of Mechanical Engineering, Hanyang University, Seoul, Korea, in 1978. He received his M.S. and Ph.D. degrees from the Department of Mechanical Engineering, University of Kansas, KS, USA in 1983 and 1986, repectively. He was design engineer of Gibbs & Hill Inc., Dravo Engineering Group, USA in 1986-1988. He was researcher as a research faculty, Yale University, CT, USA in 1988-1989. Since 1992, he has been researcher and teaching as professor, Korea University of Technology and Education, Cheonan, Korea. His research interest includes mechanical engineering, numerical simulation, semiconductor equipment, engineering education, energy system. He is a fellowship member of the ASME.

Ohyoung Kwon
received his B.S., M.S., and Ph.D. degrees from the Department of Computer Science, Yonsei University, Seoul, Korea, in 1990, 1992, and 1997, respectively. He is a professor of School of Computer Science and Engineering, Korea University of Technology and Education (KOREATECH). He currently serves as a dean of online lifelong institute of KOREATECH from March 2016. His research focuses on system software, embedded computing and high performance computing. He is trying to apply high performance computing in online education.

Ilyoung Moon
received his B.S., M.S., and Ph.D. degrees from the Department of Telecommunication & Computer Engineering, Korea Aerospace University, Seoul, Korea, in 1990, 1992, and 1997, respectively. He is currently a professor of School of Computer Science & Engineering, Korea University of Technology and Education (KOREATECH). His research interests are in the areas of web programming, wireless network. He is a member of Korean Institute of Communication Sciences, a member of Korea Institute of Electronics Engineers and a member of Korea Navigation Institute.

Gangho Shin
received his B.S., M.S., and Ph. D. degrees in the Department of Film Theory from Chung-Ang University, Seoul, Korea, in 1985, 1988, and 1996, repectively. He is a Professor of Department of Cinema, Daejin University. He was a President of the Association of Korean Film Critics, a President of the Korea Film Association.

Jangho Ham
received his B.S., M.S., and Ph.D. degrees in the Department of Poetry Literature from The University of Seoul, Seoul, Korea, in 1997, 2002, and 2009, repectively. He is a professor of College of Liberal Arts and Cross-Disciplinary Studies, The University of Seoul from 2009. He is presently the director of the Korean Society for Thinking and Communication. His research focuses on the image of poetry and Film. Also he studies the relation between factors in Film and forecasting box office hit utilizing big data.

Jintae Park
received his B.S. and M.S. degrees in the Department of Computer Science & Engineering from Korea University of Technology and Education. He is currently a Ph.D. course of School of Computer Science & Engineering, Korea University of Technology and Education (KOREATECH). His research interests are in the areas of web standardization & network optimization.