International Journal of Electronic Commerce Studies
Vol.5, No.1, pp. 91-96, 2014
doi: 10.7903/ijecs.1027

USING TEXT MINING TECHNIQUES TO ANALYZE HOW MOVIE FORUMS AFFECT THE BOX OFFICE

I-ping Chiang
National Taipei University
151, University Rd., San Shia District, New Taipei City, 23741 Taiwan
ipchiang@mail.ntpu.edu.tw

Yean-Fu Wen
National Taipei University
151, University Rd., San Shia District, New Taipei City, 23741 Taiwan
yeanfu@mail.ntpu.edu.tw

Yu-Chun Luo
National Taipei University
151, University Rd., San Shia District, New Taipei City, 23741 Taiwan
andyro0408@gmail.com

Ming-Chien Li
National Taipei University
151, University Rd., San Shia District, New Taipei City, 23741 Taiwan
alliechian@gmail.com

Chiao-Ying Hsu
National Taipei University
151, University Rd., San Shia District, New Taipei City, 23741 Taiwan
jocelin.hsu@gmail.com

ABSTRACT

As a forecasting tool, audience movie reviews provide a guide for film companies. This study uses a text mining technique to analyze the American film market. It explores movie reviews including word of mouth (WOM) factors (i.e., movie content, positive, negative, and promotion) and related factors (i.e., time, rating, and the number of ratings) for the box office. According to the relationship between the keyword clusters, the major factors that affect the box office are determined. The findings provide reference for movie producers to manipulate WOMs.

Keywords: Box Office, Film Industry, Forecast, Text Mining, WOM

1. INTRODUCTION

A number of various factors and combinations of these factors have an indirect impact on the film industry. In the proliferation of new acting, the word of mouth (WOM) effect is considered to be more credible and persuasive publicity[1]. Past scholars have explored various effects at the box office, including the factors of studios, exhibition, budget, a sequel, and critical reflection[2]. However, few have studied the impact of implicit information from audience comments.

This study uses text mining techniques to uncover the correlation between comment keywords. This work identifies the most suitable cluster classification and the process of keyword filter based on the frequency of word appearance. Frequency analysis, including the regression prediction model and time series analysis, observes past experiences and predicts development trends.

Traditionally, film critic requires expert background to issue a comment. Today, comments can be issued anonymously on the Internet. Wu *et al*. found that anonymous critics have equal chance to participate and express commentaries and suggestions[3]. The sensory experiences of reviewers are important[4] for products or services. The audience criteria usually rely on other channels or objectivity of review information. Gershoff *et al*. state that the critic is the first to receive the product message for reviewers and then spreads the role of WOM[5]. Reinstein and Snyder studied the impact of the movie review for the box office. They found that the review can effectively predict the box office[6].

Electronic WOM influences the behavior of audience through the reputation effect[4]. Basuroy and Chatterjee *et al*. found a significant impact on the critic for the weekly movie box office that decreases over time. Positive and negative features have no significant relationship with the weekly box office. However, they pointed out that negative information leads to higher box office than positive WOMs[7]. On the other hand, no matter how many negative or positive comments are, the higher frequency of discussions the higher box office so that eWOM is an important marketing tool.

Text mining, also known as text data mining or text analytics, adopts the word frequency and the number of articles to obtain a file library for

information search, event correlation, etc. Data mining and text mining have several differences:

[11] Data mining is used to analyze structure data from a relational database, while the text mining is used to mine the data from an unstructured data field[8].

[12] The meaning of the data column is given for data mining, but the meanings of data must be analyzed from the document structure, even though the forum data is generated from a relational database.

[13] Data mining explorers the new findings from the relation between columns, but text mining analyzes the contents within a field. A field of an article may contain varying keywords for factor analysis.

2. RESEARCH PROCEDURES

The research framework is shown in Figure 1. The box office is related to the comments, rating, and weekly classification. The middle court of the figure shows four cluster names: positive, negative, content and promotion. The rating is the evaluation of the movie review. We adopt the STATISTICA text mining tool to mine the IMDB, a popular American movie online box office discussion site, and Mojo Box Office, which provides the annual movie box office earnings. We retrieve the top grossing 29 movies between January and October 2011 from the Mojo Box Office through Gooseeker web crawler. The web pages are divided into four categories (science fiction, action adventure, comedy and animation) and into eight weeks based on the release date. The page files are transformed into an XML file based on the related tags and keywords. The advantage of XML format is that the interested field is independent from other contents and can easily be imported into a mining tool. Then, we configure the related functions to generate interesting information, such as keyword frequency and rank.

We initially filter 1,000 keywords based on the keyword frequency. The *k*-mean algorithm is a method of cluster analysis which aims to partition the observation keywords into exactly *k* clusters, where each observation belongs to the cluster with the nearest mean. The algorithm is simple so that we can easily control the number of clusters. Hence, the *k*-means method is adopted to generate four clusters based on word frequency distance after the redundancy words are discarded.

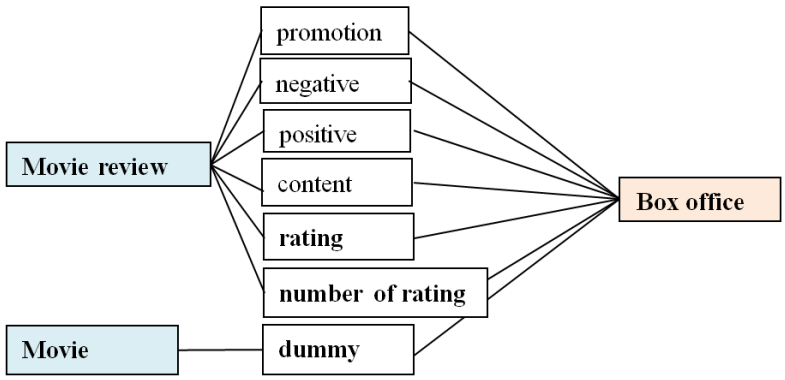

Figure 1. Research framework

For example, content cluster includes story scene (170.05), and gun (77.05); promotion includes potential (2526.60) and introduce (1297.60); negative WOM includes hate (17.50), horrible (0.50), and terrible (49.50); and positive WOMs includes pretty (308.67), fantast (252.67), and beauty (243.67). This process has been executed many times with different initial values to converge and enhance the quality of the classified results.

The regression of Eq. (1) considers box office (unit: 10 million), which denotes Y as affected by the following factors: (i) X_1, the frequency of content keywords; (ii) X_2, the frequency of negative WOM keywords; (iii) X_3, the frequency of positive WOM keywords; (iv) X_4, the frequency of promotion keywords; (v) X_5, the box office rating; (vi) X_6, the number of rating; and (vii) X_7, the dummy variable.

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 \quad (1)$$

Table 1. Stepwise multiple regression analysis

					Change the statistical volume					
	R	*R* square	Adjusted *R* square	The estimated standard deviation	*R* square change value	*F* change	df_1	df_2	Significant *F* change	Durbin-Watson test
Eq. (1)	.750[a]	.563	.444	6.576E7	.563	4.721	6	22	.003	2.083
Eq. (2)	.793[a]	.630	.506	6.196E7	.630	5.098	7	21	.002	1.994

Table 2. Verification of individual variables

Mode	Standardized coefficients		
	Beta distribution	*t* value	Significant
(Constant)		-.837	.412
Content (X_1)	.083	.151	.881
negative WOMs (X_2)	-.344	-.860	.399
positive WOMs (X_3)	-.599	-1.215	.238
promotion (X_4)	1.754	2.432	.024
rating (X_5)	.140	.960	.348
the number of rating (X_6)	-.051	-.094	.926
dummy (X_7)	.386	1.944	.065

Table 1 shows the multiple regression results, where the adjusted *R*-square is 44.4%. The adjusted *R*-square is changed to 50.6% to improve the explanatory power, as shown in Eq. (2). Table 3 shows the results of the *t*-test, which indicates that the set of promotion keywords significantly affect the box office. Thus, the use of promotional marketing practices, merchandise, celebrity endorsements, as well as the time and place of promotion influence the reviewers and box office both before and after the movie launch.

$$Y = 0.083X_1 - 0.344X_2 - 0.599X_3+1.75X_4+0.14X_5 - 0.051X_6+0.386X_7 \quad (2)$$

3. SUMMARY

Movie producers can post review message on the forum based on the eWOM effect of the keywords to third party reviewers for discussion. Despite the factors of pre-sale promotion and forecasts (or advertising and marketing practices), the factors of content cluster are the most important factor on the box office earning among these four clusters. Thus, movie producers can narrow their scope to plan their budgets and resource allocation on the target classification. This analysis shows that the use of promotion keywords is the most important for marketing of the film. The more attention a film receives from an audience and the larger the keywords and film promotion with the less unnecessary costs. As a result, the more customers get attention and the more the interests of remuneration, the greater the box office results.

4. REFERENCES

[1] T.-C. Lin, The dynamic effect of word of mouth - Take American movies as examples. *The master thesis of the Department of International Business*, National Taiwan University, 2008.

[2] S. Basuroy, S. Chatterjee, and S.A. Ravid, How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing*, 67(4), p103-117, 2003. http://dx.doi.org/10.1509/jmkg.67.4.103.18692.

[3] J.R. Wu, Word-of-mouth life cycle–movie reviews for example. *The master thesis of the Department of Business Administration*, National Taiwan University of Science and Technology, 2007.

[4] Y. Liu, Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3), p74-89, 2011. http://dx.doi.org/10.1509/jmkg.70.3.74.

[5] A.D. Gershoff, A. Mukherjee, and A. Mukhopadhyay, Consumer acceptance of online agent advice: Extremity and positivity effects. *Journal of Consumer Psychology*, 13(1). p161-170, 2003. http://dx.doi.org/10.1207/S15327663JCP13-1&2_14.

[6] D.A. Reinstein, and C.M. Snyder, The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The journal of industrial economics*, 53(1), p27-51, 2005. http://dx.doi.org/10.1111/j.0022-1821.2005.00244.x.

[7] M.H. Burzynski, and D.J. Bayer, The effect of positive and negative prior information on motion picture appreciation. *The Journal of Social Psychology*, 101(2), p215-218, 1977. http://dx.doi.org/10.1080/00224545.1977.9924009.

[8] R. Feldman, and J. Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York, NY: Cambridge University Press.

[9] R. Goldman, and J. Widom, Dataguides: Enabling query formulation and optimization in semistructured databases. *Paper presented at the 23rd International Conference on Very Large Data Bases*, Athens, Greece, August 25-29, 1997.