Quantifying Domestic Movie Revenues Using Online Resources in China

Jia Xiao[1*], Song-nan Yang[2], Xin Li[1], He Jin[1] and Shanzhi Chen[3]

[1.] State Key Laboratory of Networking and Switching
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road, Haidian District, Beijing 100876, China.
[2.] Beijing Guodiantong Network Technology Co. Ltd.
Beijing 100070, China.
[3.] State Key Laboratory of Wireless Mobile Communications
China Academy of Telecommunications Technology
Beijing 100083, China.

Abstract — **In this paper, we apply a machine learning method to predict a movie's box-office performance using data collected from Chinese mainstream search engines, video sites and WeChat. Domestic films are firstly classified into different types according to movie industry experts. And we make multiple linear regression with one year's samples of five main genres. Then these models are applied to predict the performance of the last film with the same type in 2014.The results show that four types have reached over 80 percent prediction accuracy except action movies. Then we prove that this type is more prone to be inaccurate than others.**

Keywords - multiple linear regression, box office revenue, pearson correlation coefficient, R-square.

I. INTRODUCTION

With the explosive speed of Chinese movie industry, both the quantity and the quality of domestic films have been improved drastically. According to data published by Chinese Flim & Television Industry Research Center(CFIRC)[1], it is found that the box office income of China has increased from more than 13 billion in 2011 to nearly 30 billion in 2014, with an average growth ratio of about 30% annually. More and more investors are willing to put their capital into this industry. Alibaba group even created online financial product promoted by Alipay platform to attract retail flows. However, one single movie could be the difference of box office smash or loss for investors in a given month. Not only is box office income the most important indicator of the success of a film, but also the profits of a group of investors. To mitigate risks and boost returns, the prediction of movie sales revenue in China is very worthy of study.

In Section 2, we review the previous work in box office prediction. Section 3 depicts the overall distribution of box office revenue with data proof. Section 4 addresses the question of useful features and online data sources of domestic movies in China. We describe the method of building models with multiple linear regression based on film industry expertise. And we predict the performance of the last movie in 2014 belonging to specific type. Given the prediction results, Section 5 analyzes the prediction deviation and discusses the most powerful reasons contributing to this phenomenon .

II. RELATED WORKS

In this area, researchers have never stopped the pace. There already exist a great number of works with various ethods in box office prediction. Generally, these works could be categorized into two types on the basis of the data source. One type is based on quantifying the elements extracted from the movie itself, the other is from the data people make public online or offline.

Litman, who was the first one in this direction, mentioned critic's rating in 1983[2].

Some papers make use of WOM, which is mainly collected from social media, trying to realize analysis because of two main reasons contributing to this phenomenon. Firstly, there are volumes of data about movies and related social media, which are easy to be accessed and collected. Secondly, WOM exert positive or negative influence on revenue depending on whether the film resonates with audiences and there could be a correlation between social media contents and movie box office.

Xiwang Yang, Yang Guo and Yong Liu(2011)[3] proposed a Baysian-inference based movie recommendation system which leverage the embedded social structure inside a social network to provid accurate and personalized recommendations.

HAO Yuan-yuan, LI Yi-jun, YE Qiang and ZOU Peng(2013)[4] revealed the variances of impacts of three information sources, including volume and valence of online reviews, critic reviews and movie production budgets on movie box office revenues.

Rui Yao and Jianhua Chen(2013)[5] applied sentiment analysis and machine learning methods to study the relationship between the online reviews for a movie and its box office revenue performance. It is indicated that online reviews can be a good indicator for predicting a movie's box office revenue.

Krushikanth R.Apala(2013)[6] identified several pattens according to the predictive models for the box office performance of the movies, on the basis of factors derived from Twitter, Youtube and the IMDb movie database.

Shyam Gopinath, Pradeep K.Chintagunta, Sriram Venkataraman(2013)[7] measured the effects of pre- and post- release blog volume, blog valence and advertising on the performance of 75 movies in 208 geographic markets of the U.S. It is found that release day performance is impacted most by pre-release blog volume and advertising, whereas post-release performance is influenced by blog valence, user rating valence and advertising.

Jaehoon Lee,Giseop Noh and Chong-kwon Kim(2014)[8] provided a visualization approach to find clearly hidden relations between movies and their evaluation. The word-of-mouth effects are proved through analyzing the patterns with reviews.

Apart from WOM(word-of-mouth),film is a miraculous work of art composed of a variety of innate elements, such as performer, director, genre, producer, script, etc. There is a partly correlation between some of these elements and box office revenue. Different researchers hold different points of view.

Timothy King(2007)[9] pointed out that there is no correlation between critical ratings for movies and box office revenue.

Zhang Yusong(2009)[10] revealed that the box office revenue has positive correlation with capital of the film, has negative correlation with piracy and has almost nothing to do with valence of the film, which means large investment equals to high income.

He Ping(2011)[11] insisted that creation, time length and advertising promotion are the key elements of a successful film with high income. She also deemed that the investment determines the revenue.

HU Xiao-li, LI Bo and WU Zhengpeng(2013)[12] researched on the factors by the multiple linear regression and analyzed the extent of several factors' impact. They indicated that the combination of actor and director is the most influential factor among all the elements and director contributes more to revenue than actor.

Jehoshua Eliashberg,Sam K.Hui,Z.John Zhang (2014)[13] developed a methodology, which make use of textual features from scripts based on screenwriting domain knowledge, human input, and natural language processing techniques, to predict box office performance of a movie at the point of green-lighting.

From the papers listed above, we could conclude that most researchers could not achieve the objective of prediction because of three main reasons. First, some of these papers only summarize rules based on historical data without noticing industry law. Second, some could not publish result in advance because using lagging elements such as WOM. Third, some paper provide one solution to all movies, which would definitely cause to high error rate. Last but not the least, all the paper listed above failed to prove validity in making revenue prediction on Chinese domestic movies. We take a different approach. The objective of this paper is firstly, to find and collect available online data of domestic movies. Secondly, to construct reasonable models through analyzing film industry law. Lastly, to publish prediction results with acceptable accuracy pre-release and discuss possible reasons contributing to deviation.

III. SURVEY OF THE OVERALL DISTRIBUTION

Google had released a white paper in 2013, which revealed its box office prediction model, on the basis of following parameters: search query volume, search ad click volume, theater count, franchise status and audience score. Google announced that the degree of correspondence between the revenue predicted by this model and the actual value is 94%.

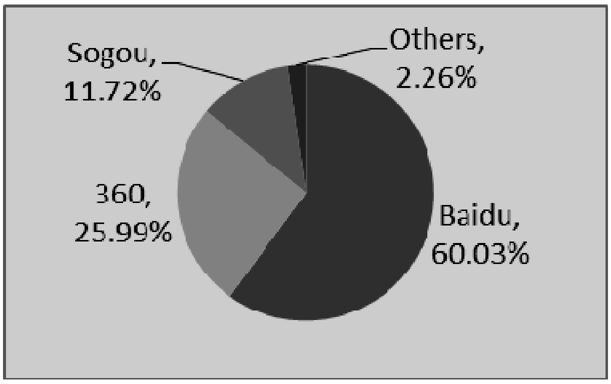

Fig. (1). Comparison of 2012 Box Office Index and Film-Related Search Index[14].

Nevertheless, the conclusion mentioned above are all from historical data. Google has never highlighted the magic when confronting with new films. Whether this method can be applied into Chinese market is also questionable. First of all, although it dominates in American search-engine market, Google may not have much more of a future in China. In Fig.2, according to the data published by CNZZ in 2014, Baidu,360 and sogou are the top 3 search engines in the Chinese market. Secondly, the online behavior of Chinese movie audience is not the same with that of American film viewers. As Chinese population of mobile internet users grows, query volume and paid click volume provided by search engines may not represent network patterns of the potential moviegoers. Last but not the least, recommendation from video websites remains one of the most influential sources throughout the decision process in choosing a film to watch. Despite it lists second next only to Google in terms of trailor search, youtube is incapable of signifying the searching pattern of Chinese film previews.

Alternatively, domestic major video sites are youku, douban and so on.

Besides computer technicians, we do research in this project with some senior film industry practitioners. We believe that movie business has its own inherent characteristics and laws. It is imperative to start this research abiding by industry rules. These seasoned experts are capable of enlighten us with sharing their insights during the cooperation process.

We gathered box office revenue from the website http://www.m1905.com with web crawler named jsoup. The data is published by week basis, supported by the state administration of radio, film and television. All the movies in m1905 have corresponding movie IDs. We extracted the weekly box office by passing the movie ID to the code and added them up to obtain the total box office of the movie.

The box office revenue released by this site contained data from four areas, mainland, Japan, Hongkong and North America. We collected data over the past four years from May 2010 to May 2014.Then the profit distribution curve is drawn to display the overall trend of box office in different regions.

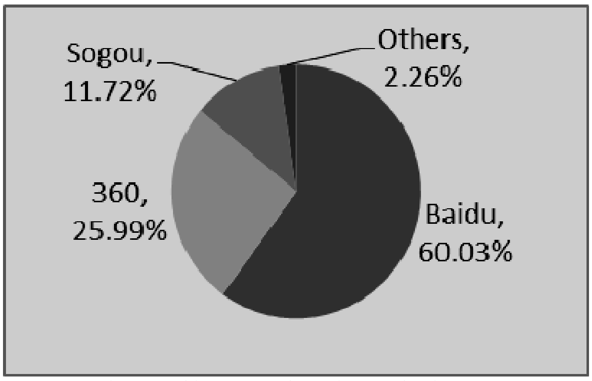

Fig. (2). Chinese Search Engine Market in 2014

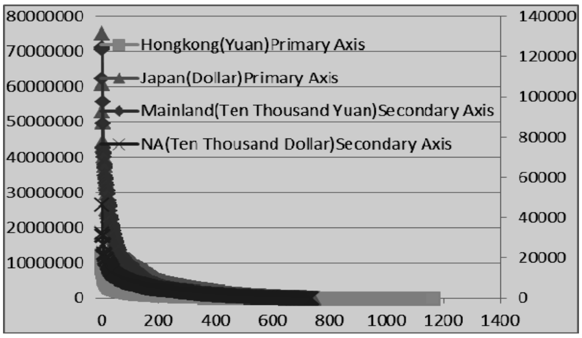

Fig. (3). Box Office Revenue Distribution Curve in Four Regions from 2010 to 201

As is shown in the curve graph, more than five sixths of the total movies' revenue are extremely low. In contrast, less than one sixth occupy a large cut of the market. That means if we predict that the profit of any movie not released yet is poor, the accuracy rate is higher than 80%. Obviously,the probability of a box-office hit is less than 20% vice versa. This phenomenon follows Zipf's law.

The curve of the box-office receipts in Mainland presents the characteristics of power-law distribution, which is depicted in Fig.4.Only 153 out of 1163 films gained over 100 million yuan in past four years. From the perspective of probability, we are more concerned with predicting the high box office receipt rather than the low one.

IV. Box Office Prediction

Although Google Flu Trends(GFT) have got amounts of criticism since published in Nature in 2009[15],it is advisable to absorb the essence of the modeling process. To improve the precision of mapping search queries to influenza-like illness, researchers dissected the data source from two dimensions, time and region.

A. First Dimension

Time is an extremely important dimension. The results of prediction will undergo radical changes over time, just as box office revenue. The number of cinema screens has experienced a surge each year. The data collected more than five years ago is more likely to be noise rather than signal to our model. According to industry insiders, China has begun to keep explicit, clear and accurate box office records since 2012.Consequently, the data prior to this year is not available for this project.

B. Second Dimension

Data about influenza changes greatly across different states in America and revenue also varies widely in different regions when mapped to film business. However, due to the goal of national box office prediction, the second dimension to our model is supposed to be changed, which is not region but genre according to industry experts. Genre summarizes the overall theme of a movie. It also could identify the potential target viewers. More importantly, the input and output of a movie vary drastically according to the difference of genre, which means it determines investment and revenue to some extent. Plus, the appetite of the market, especially the audience, for genre is changing with time and circumstance. Therefore, it is absolutely necessary to classify movies in accordance with genre.

There exists dozens of genre labels. And the classification of the same film has a certain differences among different video websites. For instance, Coming Home, directed by the famous director Zhang Yimou in 2014, is labeled as drama by youku, while marked as romance, drama, biography and history by m1905. Through discussion with movie industry experts in the project team, we consider eighteen genres here: romance, comedy, action, literary, horror, suspense, science-fiction, animation, adventure, child, war, family, biography, musical, fantasy, crime, history and erot. In view of the fact that a film may have more than one genre label, we only attach two labels to one single film, facilitating classification. The eighteen genres were divided into two groups, listed in Table 1. A film is supposed to attach one main label and the other auxiliary label or none. Taking an example mentioned before, coming home is marked as literary and romance.

C. Model

We originally planned to exploit two years' data to construct the prediction model. Three data sources are included: search engines, video websites and film industry expertise.

TABLE I. GENRE CLASSIFICATION

Classification	Genre
Main	Comedy, Action, Literary, Horror, Suspense, Adventure, War, Child, Erotic
Auxiliary	Romance, Sci-fi, Animation, Family, Biography, Musical, Fantasy, Crime, History

TABLE II. ELEMENTS OF MODEL

Series	Data	Source	Appendix
A	Movie Baidu Index	Baidu	Screening Date
B	Movie Average Baidu Index	Baidu	One Week before Showing
C	Movie 360 Index	360	Screening Date
D	Movie Average 360 Index	360	One Week before Showing
E	Movie Sougou Index	Sougou	Screening Date
F	Movie Average Sougou Index	Sougou	One Week before Showing
G	Director Baidu Index	Baidu	Screening Date
H	Director Average Baidu Index	Baidu	One Week before Showing
I	Director 360 Index	360	Screening Date
J	Director Average 360 Index	360	One Week before Showing
K	Director Sougou Index	Sougou	Screening Date
L	Director Average Sougou Index	Sougou	One Week before Showing
M	Two Leading Actors Baidu Index	Baidu	Screening Date
N	Two Leading Actors Average Baidu Index	Baidu	One Week before Showing
O	Two Leading Actors 360 Index	360	Screening Date
P	Two Leading Actors Average 360 Index	360	One Week before Showing
Q	Two Leading Actors Sougou Index	Sougou	Screening Date
R	Two Leading Actors Average Sougou Index	Sougou	One Week before Showing
S	Douban Score	Douban	Querying Date
T	Youku Score	Youku	Querying Date
U	M1905 Score	M1905	Querying Date
Y	Box Office Revenue	M1905	Querying Date

First *we* make use of search engines shown in Fig.2 to acquire the popularity of the director, two leading actors and the film itself. The three search engines' indices are collected respectively aiming at the same element listed in Table 2.

However, these indices are released with a one-day delay. We could only calculate the result on the second day of release with indices on the screening day.

Second the data of three video websites is collected:m1905,youku and douban. Firstly, the movies and box office data are provided by m1905. Secondly, youku possesses data-rich indices and it is market share first in China.

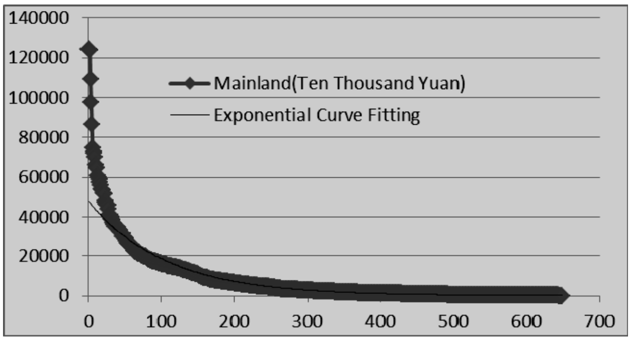

Fig. (4). Box Office Revenue Distribution Curve in Mainland.

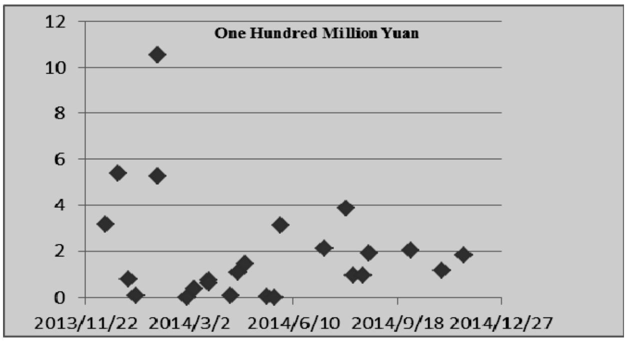

Fig. (5). Box Office Scatter Diagram of Action Films in 2013.

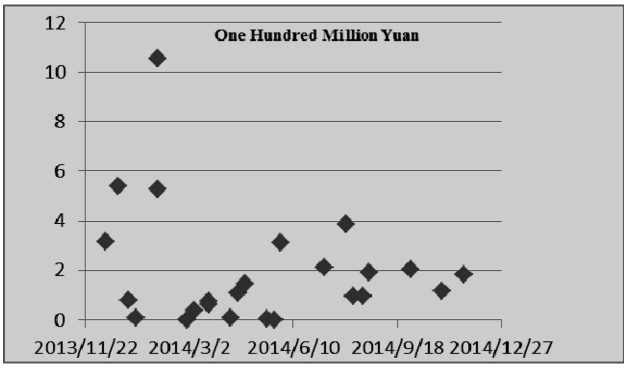

Fig. (6). Box Office Scatter Diagram of Action Films in 2014.

Third, the conversion ratio from douban score to box office revenue is considered relatively high [16].

Furthermore, youku provide the following eight sets of data: number of collection, number of comments, number of support, number of trample, movie whole network search index, movie whole network play index, movie youku play index and movie tudou play index. But time dimension is not included in these data and these data will keep on changing over time. In view of feedback interference, these data are abandoned in the process of modeling.

TABLE III. ANALYSIS OF VARIANCE OF (1)

Model		Sum of Squares	df	Mean Square	F	Sig
1	Regression	1320872	20	66044	399	.000b
	Residual	496.174	3	165		
	Total	1321368	23			

TABLE IV. REGRESSION EQUATION COEFFICIENT OF (1)

Model		Unstandardized Coefficients		Standardized Coefficients	Model	Sig
		B	Std. Error			
1	(Constant)	-15.686	87.613		-.179	.869
	VAR00001	.001	.000	.134	2.121	.124
	VAR00002	.000	.002	-.017	-.169	.876
	VAR00003	.002	.001	.625	3.981	.028
	VAR00004	.001	.002	.040	.405	.713
	VAR00005	.003	.004	.189	.765	.500
	VAR00006	-.002	.011	-.037	-.204	.851
	VAR00007	.145	.034	.982	4.263	.024
	VAR00008	-.255	.065	-1.041	-3.909	.030
	VAR00009	-.147	.064	-.803	-2.306	.104
	VAR00010	.644	.084	2.176	7.626	.005
	VAR00011	-.838	.128	-1.380	-6.524	.007
	VAR00013	-.001	.001	-.097	-.686	.542
	VAR00014	-.004	.002	-.463	-2.664	.076
	VAR00015	-.017	.003	-1.042	-6.125	.009
	VAR00016	.020	.003	1.099	5.921	.010
	VAR00017	.067	.006	1.993	11.536	.001
	VAR00018	-.053	.008	-1.415	-6.674	.007
	VAR00019	20.793	6.906	.116	3.011	.057
	VAR00020	20.616	9.844	.055	2.094	.127
	VAR00021	-34.749	9.337	-.127	-3.722	.034

As of 31st December 2014, a total of 25 action films were released this year. This number is 35 in 2013. Except the last one (The Taking of Tiger Mountain),the revenue scatter diagram of the remaining 24 films was drawn in Fig.6.There are 15 films whose box office is over 100 million yuan.

1. Action Movie Model

As of 31st December 2014, a total of 25 action films were released this year. This number is 35 in 2013. Except the last one (The Taking of Tiger Mountain),the revenue scatter diagram of the remaining 24 films was drawn in Fig.6.There are 15 films whose box office is over 100 million yuan. This number is same with that of 2013,which is shown in Fig.5.The revenue of the 25th one which was on shown had already been over 0.4 billion by the end of 2014.

Comparing Fig.5 and Fig.6, we can conclude that the number of action films is decreasing while the proportion of high box office films is increasing. In order to make the model reflect changes of market trend, we only adopt the data of 2014 to build the prediction model.

As shown in Table 2,there are twenty-two independent variables. We consider box office revenue as the dependent variable, and make a multiple linear regression model. The regression equation with enter method is as follow:

$$Y = -15.986 + 0.001 \times A - 318 \times 10^{-6} \times B - 0.002 \times C + 0.001 \times D + 0.003 \times E - 0.002 \times F + 0.145 \times G - 0.255 \times H - 0.147 \times I + 0.644 \times J - 0.838 \times K - 0.001 \times M - 0.004 \times N - 0.017 \times O + 0.02 \times P + 0.067 \times Q - 0.053 \times R + 20.793 \times S + 20.616 \times T - 34.749 \times U \tag{1}$$

and the analysis of variance is shown in the Table III.

TABLE V.MODEL SUMMERY

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.931a	0.866	.860	8.95749
2	.958b	0.918	.910	7.200916

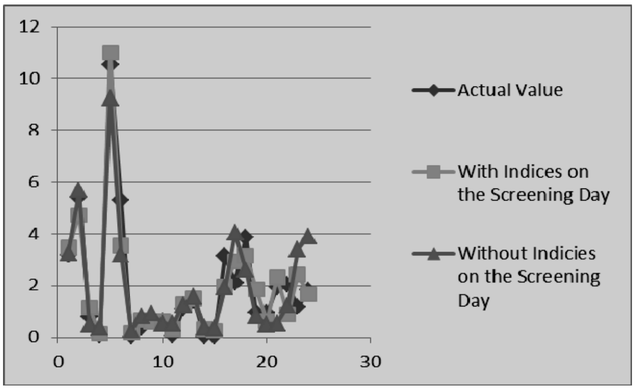

Fig. (7). Comparison of the Three Values Concerning Action Movies.

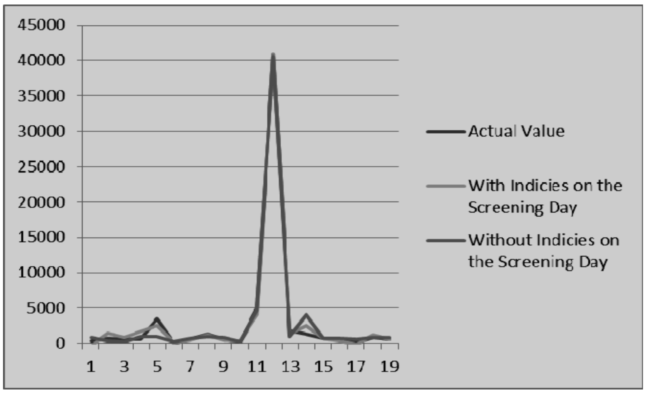

Fig. (8). Comparison of the Three Values Concerning Horrible Movies.

$$Y = 2.0074 + 0.001918 \times D \qquad (2)$$

The Pearson correlation coefficient of (5) between results of (3) and actual values is 0.958.The one is 0.897 when applied to (4).

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \square \sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad (3)$$

The comparison of actual values, results of (3) and (4) is depicted in Fig.7.

According to (3) and (4), the prediction result of the last action film in 2014 is 0.21 billion and 0.11 billion respectively, which are significantly lower than the actual value(0.8866 billion).We will analyze the prediction deviation in the next section.

2. Horrible Movie Model

There are twenty horrible movies screened in 2014 according to our classification method. We gathered the following elements listed in Table 2.

The prediction model based on elements without indicies on screening day is as follow:

$$Y = \begin{cases} 121.043 + 0.176 \times C \\ 65.292 + 0.337 \times C - 0.273 \times B \end{cases} \qquad (4)$$

R square values of (6) are 0.994 and 0.996,so the lower half part is chosen.

$$Y = 65.292 + 0.337 \times C - 0.273 \times B \qquad (5)$$

The prediction model based on elements without indices on screening day is as follow:

$$Y = 192.889 + 0.297 \times B \qquad (6)$$

The comparison chart of actual value, results of (7) and (8) is shown in Fig.8.

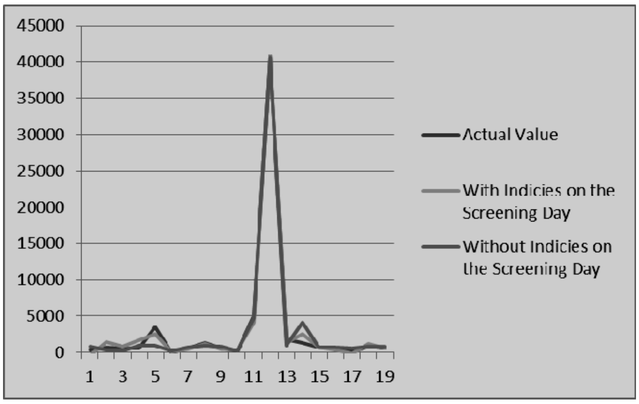

Fig. (8). Comparison of the Three Values Concerning Horrible Movies.

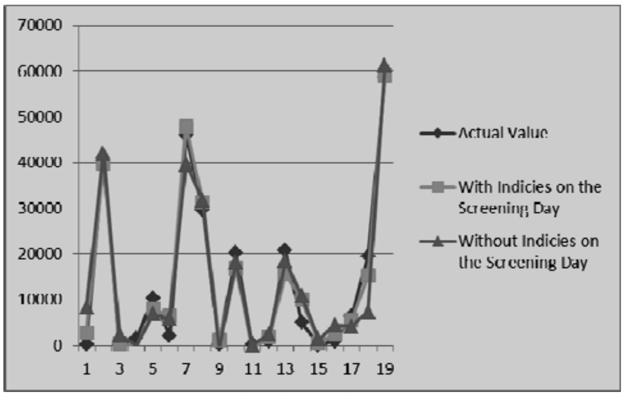

Fig. (9). Comparison of the Three Values Concerning Literary and Romance Movies.

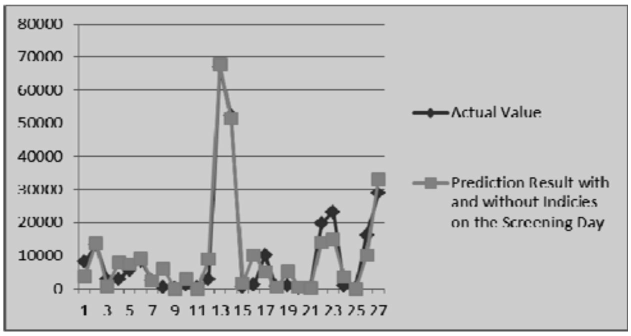

Fig. (10). Comparison of the Three Values Concerning Child and Animation Movies.

We apply (7) and (8) to predict box office revenue of the last horrible movie named Bloody Doll in 2014.The prediction result are 19 million and 13.40 million. The actual value is 23.5 million. Therefore, the deviation rate are 19.14% and 42.8% respectively. It means that if we release prediction on the second day rather than the screening day, the result is more accurate.

3. Literary Romance Movie Model

According to the theory and modeling method discussed above, the prediction model based on elements with indices on screening day of 19 films is as follow:

$$Y = \begin{cases} 966.716 + 1.693 \times E \\ -1593.634 + 1.563 \times E + 1.761 \times G \\ -983.812 + 1.038 \times E + 1.469 \times G + 1.531 \times F \\ -705.896 + 0.615 \times E + 1.946 \times G + 5.344 \times F - 0.311 \times B \\ 17676.827 + 0.795 \times E + 2.357 \times G + 5.792 \times F - 0.369 \times B - 2954.315 \times U \end{cases} \qquad (7)$$

R square values of (9) are 0.814,0.884,0.925,0.965 and 0.978,so the last part is chosen. The prediction model based on elements without indices on screening day is as follow:

$$Y = \begin{cases} 4023.463 + 3.785 \times F \\ 3809.249 + 7.12 \times F - 0.331 \times B \\ 1465.213 + 7.524 \times F - 0.391 \times B + 7.176 \times J \end{cases} \qquad (8)$$

R square values of (10) are 0.78, 0.85 and 0.937, also the last part is chosen. Therefore, the final model is:

$$Y = \begin{cases} 17676.827 + 0.795 \times E + 2.357 \times G + 5.792 \times F - 0.369 \times B - 2954.315 \times U \\ 1465.213 + 7.524 \times F - 0.391 \times B + 7.176 \times J \end{cases} \tag{9}$$

We apply this model to predict the profit of the last literary and romance movie in 2014,whose name is Fleet of Time. The result are 591 and 612 million. The actual value is 584 million. The deviation ration are only 1.2% and 4.8%.The result based on elements with indices on the screening day are more precise, which is the same with the previous model. The comparison chart is shown below.

Certainly, if prediction result is less than zero, it means that the box office would probably be very terrible.

4. Child Animation Movie Model

There are 31 films of this type released in 2014.To avoid the statistical noise, five films whose box office revenue are not provided by m1905 were not taken into account during modeling(Their performance is also very poor).Same as above, the prediction model based on elements with indices on screening day of 25 films is as follow:

$$Y = \begin{cases} 1898.837 + 0.284 \times B \\ 1470.665 + 0.213 \times B + 0.062 \times D \\ 2026.165 + 0.163 \times B + 0.175 \times D - 0.12 \times E \end{cases} \tag{10}$$

R square values of (12) are 0.961,0.98 and 0.986,so the last part is chosen. The prediction model based on elements without indices on screening day is the first and second part of (12), the final model is:

$$Y = \begin{cases} 1470.665 + 0.213 \times B + 0.062 \times D \\ 2026.165 + 0.163 \times B + 0.175 \times D - 0.12 \times E \end{cases} \tag{11}$$

We use this model to predict the revenue of the last child and animation movie in 2014, named Kuiba III. The result are 26.55 and 20.10 million. The actual value is 23.50 million. The deviation rate are 11.5% and 16.9%.Ditto the latter is better than the former. The comparison chart is shown as follow.

5. Comedy Romance Movie Model

In 2014, 38 movies fall into this category and eleven of them are discarded because their box office profits are dropped (also too poor) by M1905.We plan to build the model with data with 26 films. The prediction model based on elements with indices on screening day is the same as the one without those.

$$Y = \begin{cases} 1995.262 + 0.175 \times N \\ 865.785 + 0.135 \times N + 0.233 \times B \\ -9150.984 + 0.131 \times N + 0.230 \times B + 2265.678 \times T \end{cases} \tag{14}$$

The square values of R are 0.859, 0.917 and 0.940.Therefore, the last part is chosen.

$$Y = -9150.984 + 0.131 \times N + 0.230 \times B + 2265.678 \times T \tag{12}$$

We use (15) to predict the revenue of the last comedy and romance movie in 2014, named Love On The Cloud. The result is 0.3315 billion. The actual value is 0.2869 billion. The deviation ratio is 15.55%. The comparison chart is shown as follow.

V. Prediction Error Analysis

We apply R-square and correlation coefficient(CC) value to test the fit of the model. R-square can be calculated as follows:

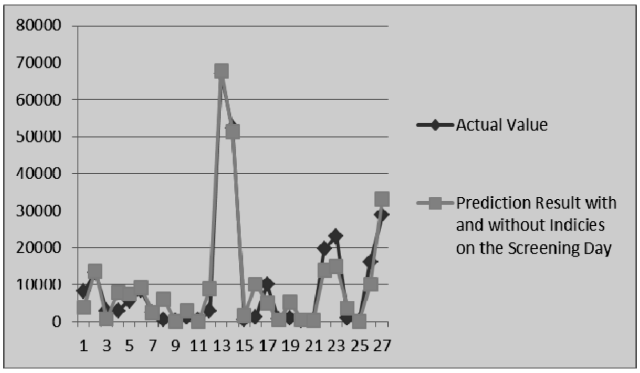

Fig. (11). Comparison of the Three Values Concerning Comedy and Romance Movies.

TABLE VI. R-Square Values of the Five Types of Films

R-square Model	With Indices On the Screening Day	Without Indices On the Screening Day
Action	0.918	0.804
Horrible	0.996	0.989
Literary and Romance	0.978	0.937
Child and Animation	0.986	0.980
Comedy and Romance	0.939	0.939

TABLE VII. Correlation Coefficient Values of the Five Types of Films

CC Model	With Indices On the Screening Day	Without Indices On the Screening Day
Action	0.958	0.897
Horrible	0.998	0.997
Literary and Romance	0.989	0.967
Child and Animation	0.987	0.983
Comedy and Romance	0.969	0.969

TABLE VIII. Pearson Correlation Coefficients among Schedule, Revenue and BSQV.

Movie Title	Correlation Coefficient Between Schedule And Daily Revenue	Correlation Coefficient Between Schedule And Daily BSQV
The Taking of Tiger Mountain	81.33%	91.10%
Gone With The Bullets	90.47%	91.93%

We could analyze from (15) that box office revenue of this type is related to the film's attention, the stars' popularity and douban score. That's a clear sign that douban and performers could play as an indicator of success to this type. It is worth mentioning that we achieved over 90 percent accuracy with this method when predicting revenue of the film named The Breakup Guru in June 2014,which is nearly 0.7 billion. At that time, its rival is the box office Champion(Transformers IV) and its score exceeded most people's expectation, except our team.

We could analyze from (15) that box office revenue of this type is related to the film's attention, the stars' popularity and douban score. That's a clear sign that douban and performers could play as an indicator of success to this type. It is worth mentioning that we achieved over 90 percent accuracy with this method when predicting revenue of the film named The Breakup Guru in June 2014,which is nearly 0.7 billion. At that time, its rival is the box office Champion(Transformers IV) and its score exceeded most people's expectation, except our team.

$$R - square = \frac{\sum_{i=1}^{n} \omega_i (\hat{y}_i - \overline{y_i})^2}{\sum_{i=1}^{n} \omega_i (y_i - \overline{y_i})^2} \tag{13}$$

R-square values of the five models are shown in Tab.6.The experiments suggest that prediction result calculated from indices on the day of release is more precise than the average value of one week before screening. Meanwhile, the prediction accuracy of action film achieves the lowest among the five types.

Correlation coefficient values of the five models are shown in Tab.7.The result is consistent with what is manifested in Table 6. We could derive that the lower of R-square and correlation coefficient values, the higher of deviation rate.

Based on the above analysis, we could conclude that the action model is more likely to be inaccurate than the other four models. Therefore, it failed in our prediction test, far from 80 percent accuracy rate.

There are many factors contributing to prediction deviation, which are obvious or subtle. Among them, the most powerful elements are WOM and schedule, which are governed by film viewers and cinema managers respectively. WOM is supposed to have increasing impact on consumer purchase decision. The key elements of WOM are volume and valence. Volume reflects the number of people who become to know or focus on the film, while valence is representative of consumer's experience about it. These data could be collected from social networks, fan and video websites, but containing much noise because of internet water army. They are also ought to be normalized before processing. However, the persuasive effect of WOM could be embodied by search query volume. Because search is a sign of interest no matter where the information is from (online or offline), whether it is real or not (true or fake), what nature of it (good or bad).

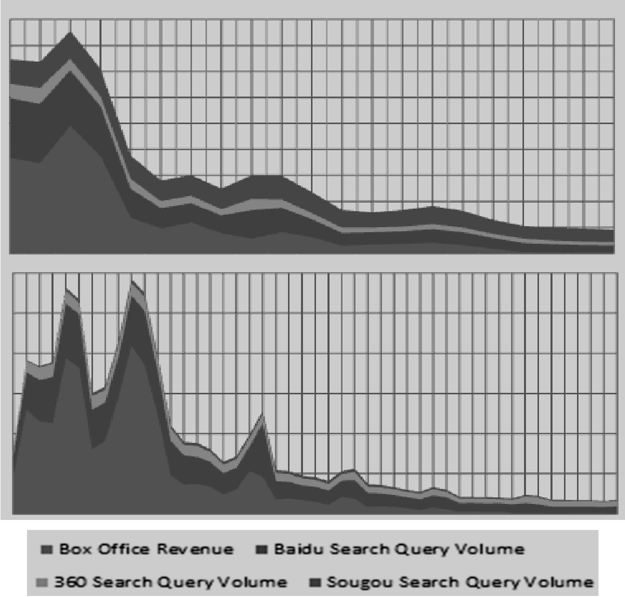

Fig. (12). Comparison of Box Office and the Three Search Query Volume during Screening.

Cinema manager is definitely powerful enough to exert influence on box office revenue through screening schedule. The percentage of schedule would be decreased to stop loss with a small audience, while it would be increased to raise income with a large audience by cinema manager. Furthermore, the number of screening days is uncertain to each film in different theatre chain because of different factors. Not to mention that SARFT(State Administration of Radio, Film and Television) has reigned supreme in China. It forced the box office champion Transformers 4 to be taken out of theatres at the 32nd day since releasing, making the revenue below 2 billion.

Taking two extremes in 2014 as an example, one is Gone With The Bullets with bad rap, the other is The Taking Of Tiger Mountain with good reputation. We crawled the daily box office revenue from WeChat official account.

The Pearson correlation coefficients of the first film(Gone With The Bullets) between daily box office revenue and the three search query volume are 0.958(Baidu),0.824(360) and 0.827 (Sougou) respectively. The numbers of the second film(The Taking Of Tiger Mountain) are 0.911(Baidu),0.836(360) and 0.499 (Sougou) separately. According to these figures, Baidu is closest to WOM after releasing.

It is suggested in Table VIII that:

1) Schedule and revenue have shown a strong positive correlation during screening period.

2) Schedule and BSQV(Baidu Search Query Volume) have indicated a more strong positive correlation while BSQV implies WOM referring to the analysis above.

3) It can also be deduced that BSQV and revenue have a strong positive correlation based on 1) and 2),which we had deducted from Fig.12.

VI. CONCLUSION

In this paper, we propose a multiple linear regression prediction model that leverages online resources including search engines and video websites. We gathered information from these resources before and on the day of release to deter echo interference. Three mainstream search engines, three authoratitive media sites and one WeChat Official Account are selected owing to data dimension. To reflect changes in the market, movie samples in one year are collected. More importantly, we work with experts in tv and film business to classify domestic films and five main kinds are carefully chosen. Stepwise method of multiple linear regression is utilized to build the model instead of other methods. The experiments show that prediction result calculated from indices on the day of release is more precise than the average value of one week before screening. Our model could do prediction with beyond 80 percent accuracy to the other four film types except action movies and the best one even achieved over 98 percent. We calculate correlation coefficient and R-square value to evaluate the fitting degree of the five models. Then we focus on the analysis of WOM and schedule, which are the two strongest factors contributing to daily box office revenue. Baidu search index is proved to be an indicator of WOM and there is a positive correlation between it and schedule.

Nowadays, the essence of commercial films is to attract audiences. Moreover, it is better to win popularity before screening than gathering fans after releasing. Taking the famous reality show named Where Are We Going Dad? as an example, the film of the same name have gained 700 million revenue in 2014.It is easier for the film to gain both fame and wealth with big fans before producing.

Future work can proceed in several directions. Due to the limited time in advance, we are supposed to analyze the relationship between schedule and competition in order to shift the prediction time to an earlier date. Secondly, we aim to study those factors contributing to box office which cannot be solved by linear regression.

ACKNOWLEDGMENT

This work is supported by “the Fundamental Research Funds for the Central Universities(Grant No. 2014PTB-00-02)”， and the National Science Fund for Distinguished Young Scholars (Grant No. 61425012) in China.

REFERENCES

[1] Chinese Film Industry Research Report in 2013. Chinese Film Press. April,2013.

[2] Litman, Barry R. (1983). Predicting Success of Theatrical Movies: An Empirical Study. Journal of Popular Culture, 16 (spring), 159-175.

[3] Xiwang Yang,Yang Guo,Yong Liu. Bayesian-inference Base Recommendation in Online Socail Networks.

[4] HAO Yuan-yuan,LI Yi-jun,YE Qiang,ZOU Peng. Dynamic Impacts of Online Reviews and Other Information Sources on Sales in Panel Data Environment:Evidence from Movie Industry. International Conference on Management Science & Engineering, September 10-12,2008.

[5] Rui Yao, Jianhua Chen. Predicting Movie Sales Revenue using Online Reviews. IEEE International Conference on Granular Computing(GrC) in 2013.

[6] Krushikanth R.Apala,Merin Jose,Supreme Motnam,C.-C. Chan, Kathy J. Liszka, and Federico de Gregorio. Prediction of Movies Box Office Performance Using Social Media. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM), August 25-29,2013.

[7] Shyam Gopinath,Pradeep K.Chintagunta,Sriram Venkataraman. Blogs,advertising and local-market movie box-office performance. Social Science Research Network, February 8, 2013.

[8] Jaehoon Lee,Giseop Noh,Chong-kwon Kim. Analysis & Visualization on movie’s popularity and reviews. In International Conference on Big Data and Smart Computing(BIGCOMP) 2014,pp:189-190.

[9] King Timothy. Does Film Criticism Affect Box Office Earnings?Evidence from Movies Released in the U.S. in 2003. . Journal of Culutual Economics, 2007(31): 171-186.

[10] Zhang Yusong and Xin Zhang. Analysis of Factors that Influence Income of Movies. Economic Forum, 2009(4): 130-132.

[11] [11] He Ping. Analysis of Factors that Affect the Box Office Income of a Film. Chinese Film Market, 2011(11): 8-10.

[12] HU Xiao-li,LI Bo,WU Zheng-peng. The Analysis of the Factors Which Influence Film Box Office. Jounal of Communication University of China(Science and Technology). Vol.20,No.1,Feb,2013.

[13] Jehoshua Eliashberg,Sam K.Hui,Z.John Zhang. Assessing Box Office Performance Using Movie Scripts:A Kernel-based Approach. IEEE Transactions on Knowledge and Data Engineering, January 15,2014.

[14] Google Whitepaper,Industry Perspective+User Insights. Quanfying Movie Magic with Google Search. June 2013.

[15] Jeremy Ginsberg,Matthew H.Mohebbi1,Rajan S.Patel,Lynnette Brammer,Mark S.Smolinski & Larry Brilliant. Detecting Influenza Epidemics Using Search Engine Query Data. Nature, Vol 457,19 February 2009.

[16] Zhang Lin. Foreign Movies’ eWOM and Box Offices. Master thesis:Tsinghua University,2012.