Filomat 30:15 (2016), 4111–4124
DOI 10.2298/FIL1615111C

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Dynamic Box Office Forecasting Based on Microblog Data

Runyu Chen[a], Wei Xu[a], Xinghan Zhang[a]

[a]*School of Information, Renmin University of China, Beijing, 100872, P.R. China*

Abstract. Movies, as one of the most rapidly developing industries´ outcomes, have gained much attention these years. Especially in China, the world´ s second largest film market with a rapid growing speed, many film companies intend to foresee the future box office in advance to better arrange their income and expenditure. Unlike some traditional forecasting model based on several movie-related features, this paper comprehensively utilizes the real-time social media, microblog, to realize a more accurate weekly box office forecasting model. The features weekly extracted from microblogs can be divided into count based features and context based features, along with the existing box office and the screen arrangements, to predict the box office in next week. For count based features, not only the total volume of related microblogs and the diffusion effect considers the number of followers, several unnoticed features like authentication users, gender ratio and mobile-users ratio are also introduced into the original predicting model. For content based features, a duplicate semantic analysis method is proposed. The number of tweets which can indeed influence others´ purchase decision, along with the number of tweets with positive and negative influence is the results of the analysis system. On this basis, guided effect for each influential tweets are identified by the praise, comment and retweet times. Some machine learning models are then adopted after using genetic algorithm (GA) for feature selection. The empirical study shows that our research can dynamic forecast box office with a sustainable good performance.

1. Introduction

Microblog, as a considerable source of electronic word of mouth (e-WOM), has rapidly developed since Twitter was launched. As a leading example, the number of active users per month in Twitter has beyond 200 million in 2014. In China, Sina Weibo is the most influential microblogging platform which already has more than 500 million registered users in 2013. Similar to Twitter, the message in Sina Weibo is limited to 140 characters. Because of the excellent readability, tweets can be widely shared in a short time so that greatly enriched people´ s daily life. In order to extract various information from such a social media, many research work have been done in recent years. For example, Hong et al. [1] predicted the popularity of tweets which is measured by the number of retweets and the information propagation. Gupta & Kumaraguru [2] analyzed the credibility of information in a tweet based on the crucial content and source based features. Guille & Hacid [3] established a predictive model for information diffusion

2010 *Mathematics Subject Classification*. Primary 62P20; Secondary 90B50
Keywords. Box Office Forecasting, Microblog, Text Analytics, Genetic Algorithm, Machine Learning
Received: 14 July 2015; Accepted: 12 April 2016
Communicated by Dr. Alex Maritz and Dr. Charles Xie
Corresponding author: Wei Xu
Research supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (16XNH118).

Email addresses: ry.chen@ruc.edu.cn (Runyu Chen), weixu@ruc.edu.cn (Wei Xu), yaoyaozxh@sina.com (Xinghan Zhang)

in social broadcasting network (SBN). Besides these, there are still much mining work about social media themselves [4, 5].

Online WOM is also a vital factor for the consumers´ purchase decision in recent years. Zendesk [6] indicated that eighty-eight percent of consumers refer to e-WOM before purchasing. Therefore, several studies have paid attention to the effect of e-WOM on product sells. For example, Dhar & Chang [7] spotted that the volume of the blogs related to a music album has a great impact on the future sales of that album. Bollen et al. [8] used the mood information extracted from Twitter to predict the stock market. Skoric et al. [9] attempted to predict electron result with twitter, although the predictive power is not as strong as they expected. As the outcomes of one of the most rapidly developing industries, movies, as an experience product, has also gained much attention. The predictive power of online WOM on box office has been identified by many researchers [10–13]. Among these researches, although both the volume and valence of the WOM have been embodied in models, the unique characteristics of microblog tweets are weakly considered. Relatively precise, Du et al. [14] proposed a predicting framework that not only includes count based features but also content based features. They preprocessed the garbage microblogs first and then used semantic classification method instead of traditional sentiment analysis to improve the prediction accuracy. Compared to their work, we introduce much more attributes of microblog data and also a more suitable semantic analysis method for tweets.

The present research targets to dynamically forecast box office based on microblog data. The reasons why microblog data can be used for the predicting are easy to be explained. First, as intangible products for entertainment, other consumers´ evaluations are particularly important for movies. Meanwhile, microblog is the most popular real-time media to gather such information. Second, movies are a sort of low-cost products that consumers can easily turn consciousness into purchase decision [15]. In our study, a comparatively rounded analysis is used for microblog data. For count based features, not only the total volume of related microblogs and the diffusion effect considers the number of followers, we also introduce several unnoticed features like authentication users, gender ratio and mobile-users ratio into the original predicting model. For content based features, we propose a duplicate semantic analysis method for the noisy and short content in microblogs. The number of tweets which can indeed influence others purchase decision, along with the number of tweets with positive and negative influence is the results of the analysis system. On this basis, guided effect for each influential tweets are identified by the praise, comment and retweet times. To utilize the real-time performance of microblogs, we build a weekly predicting model in view of both the latest and microblog data already there. After feature selection, some machine learning models like support vector machine (SVM) and neural network (NN) are applied for prediction instead of linear models. The results of our experiments show that the proposed method can indeed improve the precision of box office predicting.

The rest of this paper is organized as follows. Section 2 discusses previous studies related to box office forecasting models. Then, the proposed methodology for box office forecasting based on microblog data is illustrated in Section 3. In Section 4, the study results and discussion of our experiments are reported. In the last section, concluding remarks and the future directions of our research work are presented.

2. Literature Review

There are many researches work about box office forecasting, which mainly innovates in one of these two parts: explanatory variables and forecasting models. According to different backgrounds and theories, new explanatory variables can be introduced as features into the forecasting models to improve the forecasting precision. Meanwhile, commonly used forecasting models can be classified as statistical models and machine learning models. Attached below is the code for table.

2.1. Explanatory Variables

As the inherent attributes, some movie-related characteristics have been considered as explanatory variables. Brewer et al. [16] identified that both movie genres and MPAA ratings are valuable to predict box-office revenue. Many researches support their conclusions [10, 17], whereas others doubt [12, 18]. The powers of actors and directors are also widely considered. Some studies consult the list of famous actors to judge the power [16, 18], while others use the actor´ s historical box offices instead [17, 18]. For evaluating directors, professional comments have also been used by Wen & Yang [20].In addition, as measurable outcomes during the film production process, some studies tried to concern the budget and advertising cost which both meaningful predictors [21, 22].

Screen arrangements and the release time are also two commonly used explanatory variables. In our proposed method, the number of screens arranged for the predicting week are also employed, similar to what some of the related work have done [19]. Besides, Asur & Huberman [23] used the number of cinemas but not the screens to experiment. Some studies have taken account of the movie´ s release time. For example, Duan et al. [24] regarded weekends as a special period. Gong et al. [17] proved that a movie can get a better box office earnings if the release time is during summer holidays.

Some variables are extracted from online WOM, which can basically divide into two categories: volume and valence. The review data collected from some professional movie web sites like Yahoo! Movies are widely used [22, 24, 25]. Duan et al. [24] revealed that the volume matters, whereas the valence did not. However, Chintagunta et al. [22] found the valence can influence the box office earnings by analyzing national online reviews on movies in specific geographic areas. Although the importance of different variables varies, the predicting power of online WOM on box office are indeed identified [11]. In comparison, the contents of microblog WOM are short and unstructured without numeric ratings. Therefore, Henning-Thurau et al. [26] utilized text mining to classify tweets and found that both the valance and volume make a difference on box office. Rui et al. [27] divided the text into intention tweets and non-intention tweets, and then analyzed them separately. Besides these content based features, Du et al. [14] also considered other features especially for microblogs, such as the number of followers, comments and retweets.

In our study, some unnoticed features related to microblogs are considered. For count based features, not only the total volume of related microblogs and the diffusion effect considers the number of followers, authentication users, gender ratio and mobile-users ratio are all introduced into the original predicting model. For content based features, we propose a duplicate semantic analysis method for the noisy and short content in microblogs. The number of tweets which can indeed influence others´ purchase decision, along with the number of tweets with positive and negative influence is the results of the analysis system. On this basis, guided effect features for each influential tweets are identified by the praise, comment and retweet times.

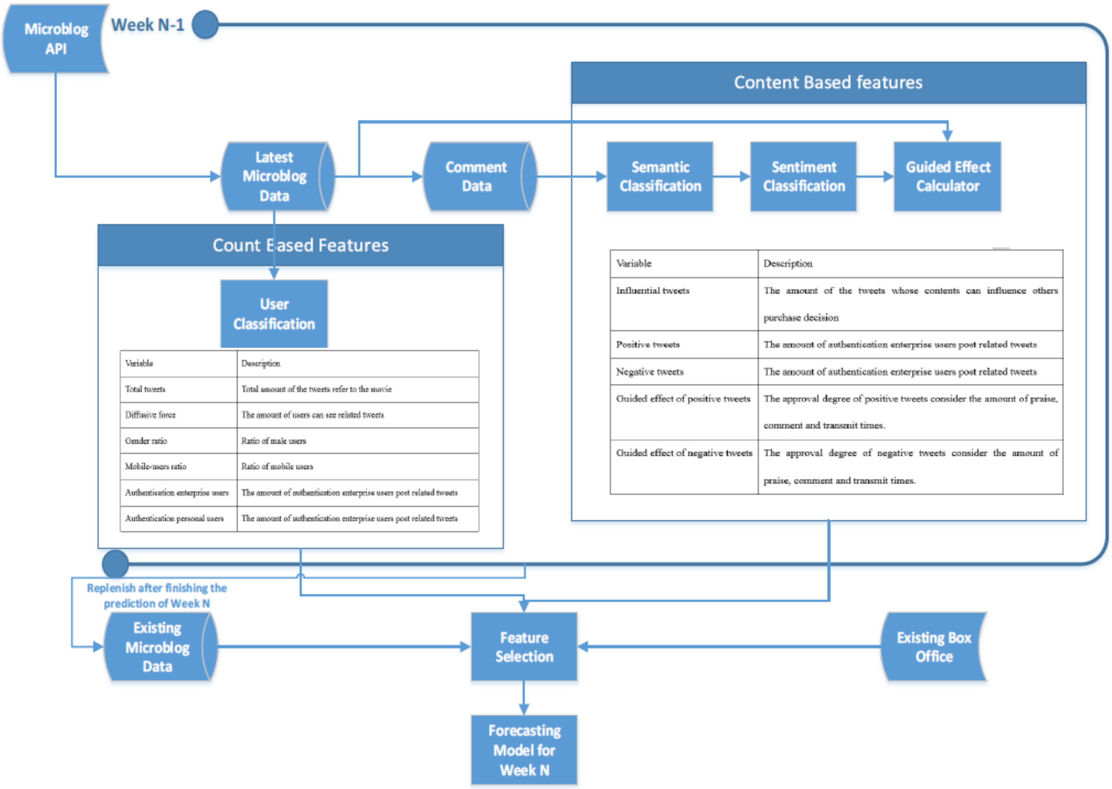
2.2. *Forecasting Models*

After selecting reasonable explanatory variables as the attributes, forecasting models are applied to complete predicting. Linear regression is the most commonly used statistical forecasting models [12, 18, 22, 24, 27].For example, [27] applied a dynamic panel data model to evaluate how Twitter WOM influence box offices. As an alternative to linear regression models, Sawhney & Eliashberg [28] proposed a forecasting model based on queuing theory which contains three box office patterns according to the probabilistic distribution. In addition, Dellarocas et al. [10] used Bass diffusion model to forecast the latter box office earnings utilizing the revenue of first three days.

Machine learning based models are widely applied in forecasting recent years. Compared to linear regression models, they suits the complex practical problem preferably. A large amount of these machine learning models formulated the forecasting as a classification problem. Zhang et al. [29] applied ANN for classifying the movies into six classes according to their box offices. Lee & Chang [30] adopted Bayesian belief network and classified movies into three categories depends on the audience number. Furthermore, some machine learning models are adopted in predicting not only classifying. Abel et al. [31] applied eight machine learning models to predict the box office revenues of movies which performed better than linear regression model. Similarly, Du et al. [14] used support vector machine (SVM) and neutral network as the forecasting model, while Kim et al. [32] combined support vector regression (SVR),Gaussian process regression (GPR) and k-nearest neighbors (k-NN) together to improve the predicting performance. In our study, we also applied some nonlinear machine learning forecasting models. Unlike previous studies, with the help of a relatively comprehensive description of the real-time microblog data, we build a weekly box office predicting model instead of predicting the total box office or just for the opening week. The forecasting model use both the latest and microblog data already there as original features, and then use genetic algorithm (GA) to select features which proved to perform better than a simple machine learning method.

3. A Methodology for Dynamic Box Office Forecasting

In this section, a methodology for dynamic box office forecasting based on latest microblog data is proposed. Along with the existing box office and the screen arrangements, we adopt a comprehensive feature description of microblog data, in which the features can be divided into count based features and context based features. To take full advantage of the characteristic of microblog as a real-time social media, we train different forecasting models for each of the release week adding latest data. Before choosing a machine learning forecasting model with relatively better performance, genetic algorithm (GA) is applied for feature selection. An overview of the proposed framework is illustrated in Fig. 1.



Variable	Description
Total tweets	Total amount of the tweets refer to the movie
Diffusive force	The amount of users can see related tweets
Gender ratio	Ratio of male users
Mobile-users ratio	Ratio of mobile users
Authentication enterprise users	The amount of authentication enterprise users post related tweets
Authentication personal users	The amount of authentication enterprise users post related tweets

Variable	Description
Influential tweets	The amount of the tweets whose contents can influence others purchase decision
Positive tweets	The amount of authentication enterprise users post related tweets
Negative tweets	The amount of authentication enterprise users post related tweets
Guided effect of positive tweets	The approval degree of positive tweets consider the amount of praise, comment and transmit times.
Guided effect of negative tweets	The approval degree of negative tweets consider the amount of praise, comment and transmit times.

Figure 1: The Methodology for Dynamic Box Office Forecasting

As can be seen from Fig. 1, we first obtained microblog data though API from the date one week earlier than the movies release time. For each tweet, there is a comprehensive feature description includes count based features and content based features. All of the features about microblog data are weekly collected, along with the existing box office and the screen arrangement, to predict the movies box office in next week. For predicting the box office of week 2-4, historical microblog data are also adopted as different input variables. Five machine learning models are adopted after using genetic algorithm (GA) for feature selection. The details will be discussed in the following subsections.

3.1. *Count Based Feature Sets*

In this paper, count based feature sets are constitute by the information we extracted from microblog data except comments. Compared with the previous work, we describe the microblog data much more comprehensive that some of the unnoticed features are also considered as shown in table 1.

Table 1: Count Based Features Extracted from Microblog Data

Variable	Description
Total tweets	Total amount of the tweets refer to the movie
Diffusive force	The amount of users can see related tweets
Gender ratio	Ratio of male users
Mobile-users ratio	Ratio of mobile users
Authentication enterprise users	The amount of authentication enterprise users post related tweets
Authentication personal users	The amount of authentication enterprise users post related tweets

Total tweets, similar to the volume of WOM in previous works [33] , have been proved to have a positive impact on box office revenues. Although Du et al. [14] presented that there are some garbage users should be filtered, we still adopt this primitive volume as a features for two reasons. First, advertisements related to the movie can also boost its exposure rates, which can even have a significant effect in a social media like microblog. Second, in Sina Weibo, once the number of tweets about a certain topic reach a high level, the topic can be selected as a hotpots which can be seen by every user. Meanwhile, different types of microblog contents are also considered in the content based feature sets.

The power of the tweets posted by different users varies. If there are a lot of users follow a person, his tweets can be observed more times. In our study, we raise a criterion named diffusion force (DF) by adding all the followers together.

$$DF = \sum_{i=1}^{M} followers_{useri} \tag{1}$$

where m is the amount of tweets mentioned the certain movie.

In addition, the number of authentication users is also considered. In Sina Weibo, an official personal or enterprise identification (ID) can be authenticated by submitting valid proofs. An authentication user can act as an expert, which will be paid much more attention by its followers. Therefore, we assume the number of authentication users can also have force on box office. In Sina Weibo, the authentication users can be classified as enterprise users and personal users.

In the business model that regards the audience demand as the guidance, we should study the market segments. In our study, we try to deeply mine the influence of microblog data on box offices so that the microblog user segments as a potential consumer ought to be considered. The gender radio is a concept from demography, which can be used to describe the characteristic of the consumer groups. Meanwhile, the ratio of mobile-users is also adopted by our work, which can be an indicator to focus on the consumer behavior model of behavioristics. Moreover, with the popularity of booking movie tickets on mobile client, the purchase process for mobile-users is more efficient.

3.2. *Content Based Feature Sets*

Content based feature sets are mainly extracted from the comments of microblogs by a duplicate semantic analysis method we proposed. As a real-time social network, the contents in microblog can be noisy that contains many unrelated topics like advertisements. For these tweets, they are filtered by the semantic classifier as whose contents cannot influence others. For influential tweets, the sentiment classifier is then used. Contrapose these short and unstructured tweets, unlike previous studies, both the sentiment words and emotional expressions are applied in our sentiment classifier to have a better understanding of the tweets. On this basis, guided effect for each influential tweets are identified by the praise, comment and retweet times. An overview of the duplicate semantic analysis framework for content based feature sets is shown in Fig. 2.

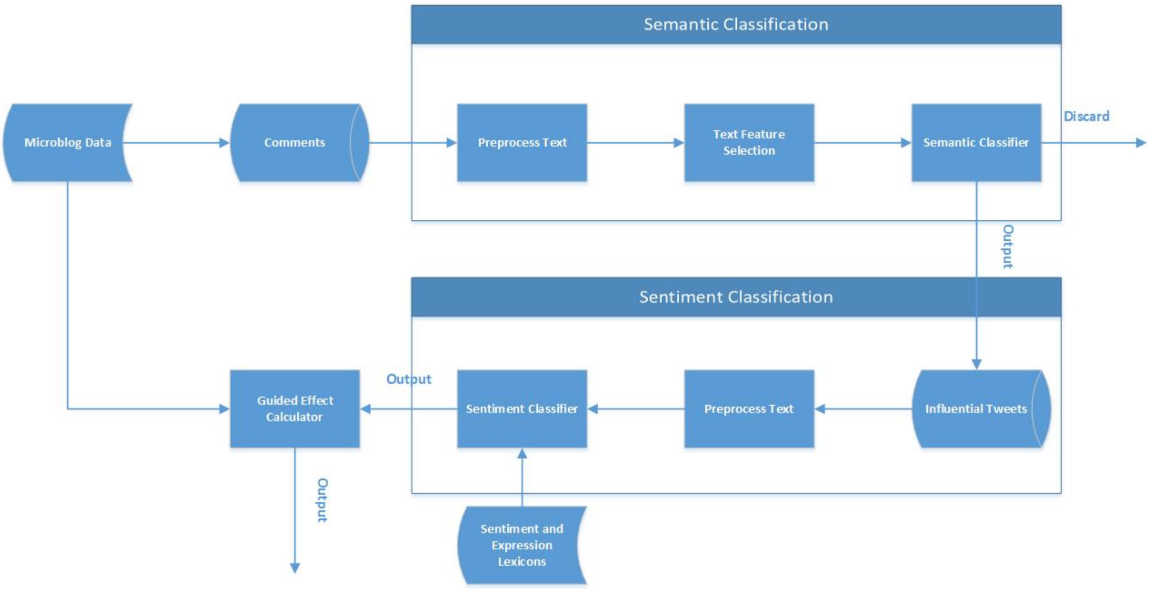


Figure 2: Duplicate Sematic Analysis Framework

As we can see in Fig. 2, some content based features are obtained though this framework. The brief description of these features is shown in Table. 2, while the details of the framework will be explained later.

Table 2: Content Based Features Extracted from Microblog Data

Variable	Description
Influential tweets	The amount of the tweets whose contents can influence others purchase decision
Positive tweets	The amount of authentication enterprise users post related tweets
Negative tweets	The amount of authentication enterprise users post related tweets
Guided effect of positive tweets	The approval degree of positive tweets consider the amount of praise, comment and retweet times
Guided effect of negative tweets	The approval degree of negative tweets consider the amount of praise, comment and retweet times

3.2.1. *Semantic Classification*

As a real-time social network, the contents in microblog can be noisy that contains many unrelated topics like advertisements. The main purpose of the semantic classification is to distinguish the tweets that can exactly influence other users and only the influential tweets are worthwhile for sentiment analysis later. To realize this classification, we first labeled 1000 typical influential tweets and uninfluential tweets to train the classifier. Word segmentations and stop words removal are applied for each tweets. And then TF-IDF (term frequencyCinverse document frequency) is used for representing the text features. After that, SVM (Support Vector Machine) is finally used as the classification model. The results of the semantic classification are as shown in Table 3.

Table 3: Content Based Features Extracted from Microblog Data

	Precision	Recall
Influential tweets	81%	79%
Uninfluential tweets	77%	85%

3.2.2. *Sentiment Classification*

After the semantic classification, tweets can indeed influence others are left. Most of the previous work did not finish the former step so that some unrelated tweets with a positive emotion could be regarded as a positive indicator to box office. In our content based feature sets, uninfluential tweets are just filtered by the first classifier.

In sentiment classification, according to the short and unstructured characteristics of microblog tweets, we build a movie-related sentiment lexicon. Realizing that nearly half of the tweets contains expressions, we also build an expression lexicons. Therefore, the sentiment words and the expressions are considered together in our sentiment analysis. The results are shown in Table 4.

Table 4: Precisions and Recalls of the Sentiment Classification

	Precision	Recall
Positive	79%	83%
Negative	76%	81%
Neutral	80%	75%

3.2.3. *Guided Effect Calculator*

In microblog, there are three actions users can take after reviewing others tweets: praise, comment and retweet. In our proposed method, we regard them as the approval of the tweet. Guided effect (GE) is defined as follows for each influential tweets that considers the amount of praise, comment and retweet times together.

$$DF = \sum_{i=1}^{m} praises_i + 2(comments_i + retweets_i) \tag{2}$$

where m is the amount of positive tweets and negative tweets respectively.

3.3. *Feature Selection Based on GA*

In our study, to obtain a better performance, historical microblog data are also brought in the forecasting models for later weeks. As a result, there are 46 feature candidates for the forecasting model for week 4. After data normalization, genetic algorithm (GA) is a random search method used the evolution of the biosphere for reference. The principle processes are as follows.

Step 1: Establishing the initial stage of a population. Feature selection is to get a feature vector collection through the process of dimension reduction. With the method of binary encoding, each binary coding bits corresponding to an original feature vector. To be specific, among the corresponding relationship between chromosomes and feature vectors, if some gene in the chromosomes values 1, the corresponding feature vector is selected. Initial population is established randomly that contains a string of 0/1.

Step 2: Calculating fitness function. Fitness function is a vital factor in GA that can directly influence the algorithm results. In our study, we use root mean square error (RMSE) as the fitness function to decide the error between the expected output and the forecasting output calculated by this feature vector sets.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{3}$$

Where n is the number of forecasting outputs, y_i is the expected output results, $\hat{y}_i$ is the real output calculated though the forecasting model.

Step 3: Selection operation.GA use the selection operation to choose relatively superior individuals. The individuals with high fitness will be more likely to inheritance to the next generation. The basic idea of the selection operator adopted in GA is that the probability of each individual to be selected is proportional to its value calculated from the fitness function.

$$p\left(x_i\right) = \frac{f\left(x_i\right)}{\sum_{j=1}^{N} f\left(x_j\right)} \tag{4}$$

Where N is the group sizes of individuals, $f\left(x_i\right)$ is the fitness of individual x_i.

Step 4: Crossover and mutation operation. Crossover operation refers to two pairs of chromosomes exchange some of its genes in some manner, thus generate two new individuals. As an auxiliary method for generating new individuals, mutation operation can change some of the genes in individual coding. The mutual cooperation of crossover operation and mutation operation completes both the global searching and local searching in the global searching space.

Step 5: Terminal condition. In our study, a certain number of iteration times is set as the terminal condition. Then the feature vector sets with the best performances is used as the results after feature selection.

3.4. Forecasting Models

After feature selection, five machine learning regression models are applied for each weeks box office forecasting, which are Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Support Vector Regression (SVR), Neural Network Regression (NNR).

Stochastic Gradient Descent (SGD) is a commonly used method to minimize the risk function and loss function. It is applied in Multiple Linear Regression (MLR) to determine the regression coefficients β, whose equation can be shown as follows.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_d x_{id} \tag{5}$$

Decision Tree (DT) is a classifier which can separate the current space into two during each division. Therefore, each leaf node is in a disjoint regions. For each predicting decision, according to the value of each dimensions, the results will be obtained in one of the N leaf node. The time complexity is low and the predicting process seems fast. However, DT is apt to be over-fitting even using pruning.

Random Forest (RF) can randomly build a forest model consists of some irrelevant decision trees. After obtaining the forest, all of the decision trees will judge on the new input sample and try to classify it. In the process of establishing a decision tree, sampling and completely splitting are two major factors. After twice sampling, the input sample for each decision tree is no longer the whole sample so that can effectively avoid over-fitting. Completely splitting is used for the data after sampling, which ensures a leaf node is unable to be split or that all samples in the node are in the same classification.

Support Vector Regression (SVR) is a method that can construct a linear decision function to realize linear regression in high dimensions after rising dimension. Like SVM, a kernel function is selected to map nonlinear data into a high dimension.

$$k\left(x, z\right) =< \phi_x \cdot \phi_z > \tag{6}$$

where ϕ is the map of x to inner vector space F. Generally used kernel functions are like polynomial kernel, Gaussian RBF kernel, index RBF kernel, hidden perception kernel and so on. Though some of the experiments show that different kernel functions do not have a great impact on the classification results, however, it usually influences a lot in the results of regression. Therefore, in our study, we use grid walk among all the kernel functions and adopt the one with best performance while applying GA.

Neural network (NN) is a mathematical model that imitates human brain system, which can be represented by the network topology, node characteristics and learning rules. The topological structures include input layer, hidden layer and output layer. A lot of input-output model mapping can be learned and stored without prior revealing the relationships because the weights and threshold of the model can be constantly adjusted.

4. Empirical Study and Results

4.1. Data Description and Evaluation Criteria

The datasets consist of two parts of data. Weekly box office and screen arrangements about 84 movies are provided by China Film Group Corporation. In addition, microblog data mentioned the movie from the date one week earlier than the release time to the end are collected from the Sina Weibo API. In our study, to dynamically use the latest microblog data, we build four forecasting model from the release week to the most usual end week (week 4). The summary statistics of input variables for each week respectively are shown in Table 5.

Table 5: Summary statistics of input variables for all movies

	Week 0	Week 1	Week 2	Week 3
Screen arrangement	0	4518942	5702655	4031131
Historical box office	0	4536568199	9770547986	13157620229
Total tweets	956426	3095439	3116656	2123339
Diffusive force	85411	214429	155210	104293
Authentication enterprise users	258188	881066	912162	622256
Authentication personal users	18107039619	35370125765	32859738344	27625307213
Male users	535469	1603309	1721689	1169201
Female users	420957	1492130	1394967	954138
Mobile users	299303	736488	772944	534596
PC users	657123	2358951	2343712	1588743
Influential tweets	303902	1233132	1392218	994128
Positive tweets	247634	884947	958292	702063
Negative tweets	56268	348185	433926	292065
Guided effect of positive tweets	23016040132	1.87052E+11	2.73951E+11	2.11957E+11
Guided effect of negative tweets	1220670840	52728435294	76568582126	40781197050

After data normalization, 56 movies are randomly selected as the training data for each weeks forecasting model and others used for testing. To evaluate the predicting performance, three common criteria are adopted: R^2(Coefficient of determination), MAE (Mean absolute error) and RMSE (Root mean squared error).

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

4.2. Experimental Results

4.2.1. Experimental results without GA

In our study, 4 forecasting models are built for predicting the box offices in different weeks. During the experiments without using GA to select features, there are 11, 24, 35 and 46 input invariables for the forecasting models from week 1 to 4. In each weekly forecasting model, five methods are applied. The results are shown in Figure 3 to Figure 6.

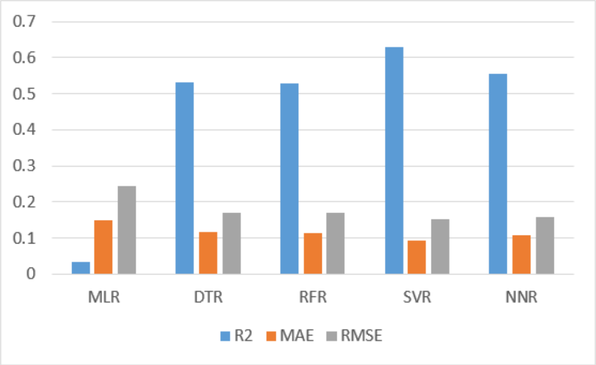


Figure 3: Forecasting Results for Week 1 without Feature Selection

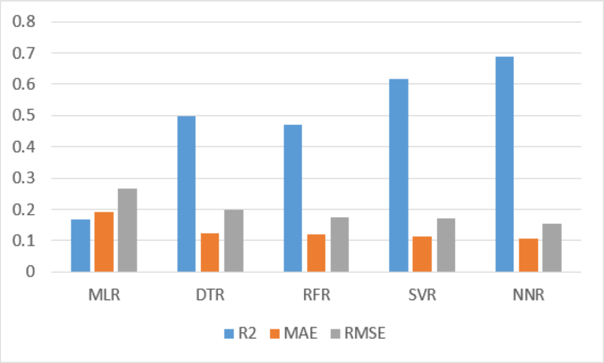


Figure 4: Forecasting Results for Week 2 without Feature Selection

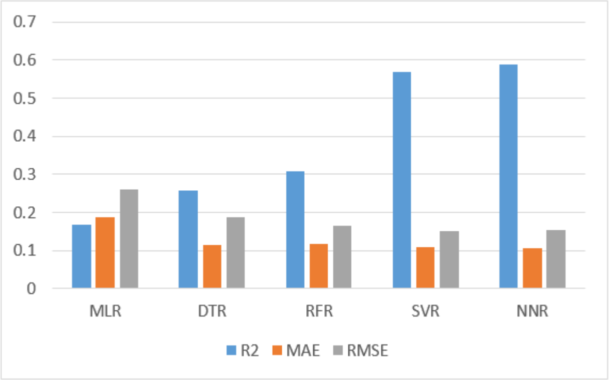


Figure 5: Forecasting Results for Week 3 without Feature Selection

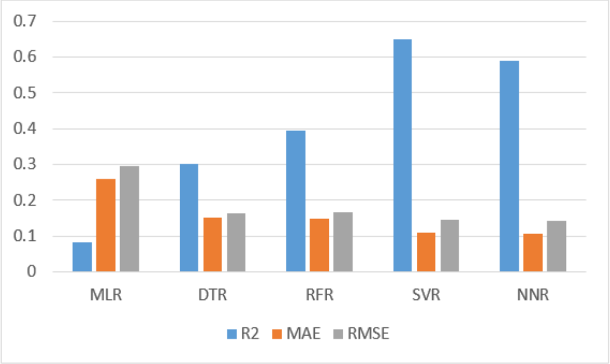


Figure 6: Forecasting Results for Week 4 without Feature Selection

As we can see in the experiment results, four nonlinear machine learning models all perform better than MLR. In comparison, SVR and NN gained better forecasting results than DTR and RFR in most times.

4.2.2. Experimental results with GA

Generic algorithm (GA) is used for feature selection that aims to improve the forecasting performance. Compared to the results without GA, the forecasting results demonstrated great advantages after feature selection. As we can see in Figure 7 to Figure 9, the contrast results of some of the forecasting models in week 4 are shown.

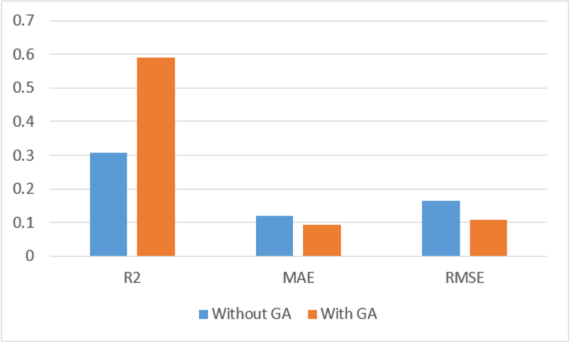


Figure 7: Forecasting Results based on RFR for Week 4 with Feature Selection

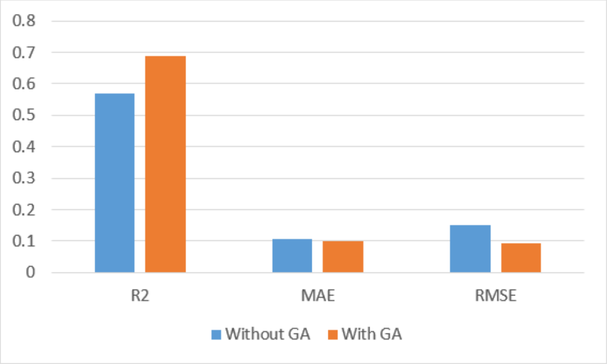


Figure 8: Forecasting Results based on SVR for Week 4 with Feature Selection

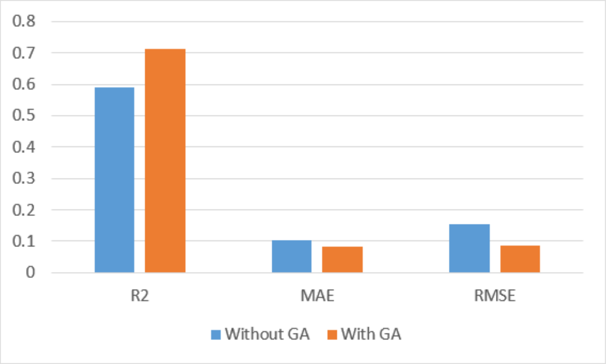


Figure 9: Forecasting Results based on NNR for Week 4 with Feature Selection

4.2.3. *Comparison and discussion*

In our forecasting models, SVR performs the best in week 1 while NNR do best in the rest weeks as shown in Table 6. Therefore, we choose NNR as the machine learning model to do the comparisons with some previous works. The arithmetic mean value of RMSE in four weeks are used to be compared.

Table 6: RMSE Results for Different Forecasting Models

	Week1	Week2	Week3	Week4
MLR	0.229	0.246	0.187	0.201
DTR	0.144	0.165	0.169	0.143
RFR	0.130	0.147	0.125	0.108
SVR	0.079	0.125	0.109	0.095
NNR	0.083	0.113	0.098	0.085

In our proposed methods, both the count based features and content based features are in deeper consideration. First we compare our results with the study that only considers the microblog volume as the count based features, meanwhile simple sentiment polarities as the content based features in experiment C1. Furthermore, we compare our results with the study with simple count based features but the same content based features as ours in C2 and the one with simple content based features but the same count based features in C3. The results can be seen in Table 7 which proved the effectiveness of our proposed method.

Table 7: RMSE Results for Different Forecasting Models

	Our Study	C1	C2	C3
Average RMSE	0.095	0.142	0.115	0.134

5. Conclusions and future work

This paper proposed a methodology for dynamic box office forecasting based on microblog data. Different from the traditional forecasting methods only use movie-related features, we utilize the real-time social media, microblog, to realize a more accurate weekly forecasting model. Compared to the previous work which has taken the microblog into consideration, we adopt a more comprehensive feature description of microblog data. The features we use can be divided into count based features and context based features. For count based features, not only the total volume of related microblogs and the diffusion effect considers the number of followers, we also introduce several unnoticed features like authentication users, gender ratio and mobile-users ratio into the original predicting model. For content based features, we propose a duplicate semantic analysis method for the noisy and short content in microblogs. The number of tweets which can indeed influence others purchase decision, along with the number of tweets with positive and negative influence is the results of the analysis system. On this basis, guided effect for each influential tweets are identified by the praise, comment and retweet times.

All of the features about microblog data are weekly collected, along with the existing box office and the screen arrangement, to predict the movies box office in next week. For predicting the box office of week 2-4, historical microblog data are also adopted as different input variables. Five machine learning models are adopted after using genetic algorithm (GA) for feature selection and the experimental results show that our method perform well.

In addition, our study still leaves some potential issues for further consideration. First of all, how to bring in other movie-related features and how to determine the relatively importance between these features and microblog-based features can be further investigated. Secondly, in our study, in view of both the real-time characteristic of microblog and the time delay in turning conscious into purchase decision, we use one week as the time period for dynamic forecasting. However, it may not be the best time period to predict an accurate result in time. Whats more, the machine learning based forecasting models are all in existence, we can also focus on a better forecasting model that may further improve the predicting results. We will pay more attention on these problems in our future researches to enhance experiments performance.

6. Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (16XNH118).

References

[1] L. Hong, O. Dan, & B. D. Davison, Predicting popular messages in twitter, In Proceedings of the 20th international conference companion on World Wide Web (2011, March) 57–58.

[2] A. Gupta, & P. Kumaraguru, Credibility ranking of tweets during high impact events, In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (2012, April) 2–8.

[3] A. Guille, & H. Hacid, A predictive model for the temporal dynamics of information diffusion in online social networks, In Proceedings of the 21st international conference companion on World Wide Web (2012, April) 1145–1152.

[4] K. Lerman, & T. Hogg, Using stochastic models to describe and predict social dynamics of web users, ACM Transactions on Intelligent Systems and Technology (TIST) (2012) 3(4) 62.

[5] S. Huang, M. Chen, B. Luo, & D. Lee, Predicting aggregate social activities using continuous-time stochastic process, In Proceedings of the 21st ACM international conference on Information and knowledge management (2012, October) 982–991.
[6] Zendesk, Available at¡http://www.zendesk.com/resources/customer-service¿ (2013).
[7] V. Dhar, & E. A. Chang, Does chatter matter? The impact of user-generated content on music sales, Journal of Interactive Marketing (2009) 23(4) 300–307.
[8] J. Bollen, H. Mao, & X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science (2011) 2(1) 1–8.
[9] M. Skoric, N. Poor, P. Achananuparp, E. P. Lim, & J. Jiang, Tweets and votes: A study of the 2011 singapore general election, In System Science (HICSS), 45th Hawaii International Conference on (2012, January) 2583–2591.
[10] C. Dellarocas, X. M. Zhang, & N. F. Awad, Exploring the value of online product reviews in forecasting sales: The case of motion pictures, Journal of Interactive marketing (2007) 21(4) 23–45.
[11] R. Sharda, & D. Delen, Predicting box-office success of motion pictures with neural networks, Expert Systems with Applications (2006) 30(2) 243–254.
[12] L. Qin, Word of blog for movies: a predictor and an outcome of box office revenue?, Journal of Electronic Commerce Research (2011) 12(3) 187–198.
[13] H. Rui, Y. Liu, & A. Whinston, Whose and what chatter matters? The effect of tweets on movie sales, Decision Support Systems (2013) 55(4) 863–870.
[14] J. Du, H. Xu, & X. Huang, Box office prediction based on microblog, Expert Systems with Applications (2014) 41(4) 1680–1689.
[15] J. Villanueva, S. Yoo, & D. M. Hanssens, The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth, Journal of marketing Research (2008) 45(1) 48–59.
[16] S. M. Brewer, J. M. Kelley, & J. J. Jozefowicz, A blueprint for success in the US film industry, Applied Economics (2009) 41(5) 589–606.
[17] J. J. Gong, W. A. Van der Stede, & S. Mark Young, Real Options in the Motion Picture Industry: Evidence from Film Marketing and Sequels, Contemporary accounting research (2011) 28(5) 1438–1466.
[18] D. Lovallo, C. Clarke, & C. Camerer, Robust analogizing and the outside view: two empirical tests of case - based decision making, Strategic Management Journal (2012) 33(5) 496–512.
[19] A. Elberse, & J. Eliashberg, Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures, Marketing Science (2003) 22(3) 329–354.
[20] K. Wen, & C. Yang, Determinants of the Box Office Performance of Motion Picture in China-Indication for Chinese Motion Picture Market by Adapting Determinants of the Box Office (Part II), Journal of Science and Innovation (2011) 1(4) 17–26.
[21] J. Eliashberg, J. J. Jonker, M. S. Sawhney, & B. Wierenga, MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures, Marketing Science (2000) 19(3) 226–243.
[22] P. K. Chintagunta, S. Gopinath, & S. Venkataraman, The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets, Marketing Science (2010) 29(5) 944-957.
[23] S. Asur, & B. Huberman, Predicting the future with social media, In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (2010, August) 492–499.
[24] W. Duan, B. Gu, & A. B. Whinston, The dynamics of online word-of-mouth and product salesłAn empirical investigation of the movie industry, Journal of retailing (2008) 84(2) 233–242.
[25] Y. Liu, Word of mouth for movies: Its dynamics and impact on box office revenue, Journal of marketing (2006) 70(3) 74–89.
[26] T. Hennig-Thurau, C. Wiertz, & F. Feldhaus, Exploring the Twitter Effect: An investigation of the impact of microblogging word of mouth on consumers´ early adoption of new products, (2012) Available at SSRN 2016548.
[27] H. Rui, Y. Liu, & A. Whinston, Whose and what chatter matters? The effect of tweets on movie sales, Decision Support Systems (2013) 55(4) 863–870.
[28] M. S. Sawhney, & J. Eliashberg, A parsimonious model for forecasting gross box-office revenues of motion pictures, Marketing Science (1996) 15(2) 113–131.
[29] L. Zhang, J. Luo, & S. Yang, Forecasting box office revenue of movies with BP neural network, Expert Systems with Applications (2009) 36(3) 6580–6587.
[30] K. J. Lee, & W. Chang, Bayesian belief network for box-office performance: A case study on Korean movies, Expert Systems with Applications (2009) 36(1) 280–291.
[31] F. Abel, E. Diaz-Aviles, N. Henze, D. Krause, & P. Siehndel, Analyzing the blogosphere for predicting the success of music and movie products, In Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on (2010, August) 276–280.
[32] T. Kim, J. Hong, & P. Kang, Box office forecasting using machine learning algorithms based on SNS data, International Journal of Forecasting (2015) 31(2) 364–390.
[33] V. Dhar, & E. A. Chang, Does chatter matter? The impact of user-generated content on music sales, Journal of Interactive Marketing (2009) 23(4) 300–307.