7th International Conference on Modelling, Identification and Control (ICMIC 2015)
Sousse, Tunisia - December 18-20, 2015

Modelling on Movie Box-Office Prediction Based on LFM Algorithm

Dong-ruRuan[1], Tao Liu[1] and Kai Gao[1*]

[1] *School of Information Science & Engineering, Hebei University of Science and Technology, China, 050000*

ruandr@hebust.edu.cn
435768300@qq.com
gaokai@hebust.edu.cn

***Abstract*—Concerning the limitations that the accuracy of predication is lower in the traditional model of movie box-office prediction and classification, this paper proposes a novel model of movie box-office revenue prediction. The model is based on the relationship of movie box-office and the user behaviors. The algorithm could be summarized as follows. Firstly, this paper uses the Latent Factor Model (LFM) to classify the movies. Secondly, for different categories, this paper constructs a series of linear movie box-office prediction models. And these models use total box-office as the dependent variable, while the independent variables are user reviews and user ratings which are the major manifestation in vertical media. Finally, this paper uses the experimental results to adjust the independent variables in different models. The experimental results demonstrate that the model performs better on prediction than the traditional methods.**

***Index Terms*—Vertical media, Box-office, Linear prediction models, LFM**

I. INTRODUCTION

With the formation of the new subject of film science, movie which used to be considered as art form has already become an art commodity. So, the film investment institutions pay more attention on how to get more box-office. At present, the recommendation and prediction technology have been well applied in the field. The usual approaches on prediction always use the vertical media and the information of items to build the corresponding prediction models. This paper proposes a novel model of movie box-office revenue prediction. The model takes the advantage of LFM in classification and takes full account of the user's behavior in vertical media. The experimental result draws a conclusion that the LFM has better performance than the traditional algorithms in movie box-office revenue prediction.

The rest of this paper is organized as follows. Section 2 describes the exiting methods on this domain. Section 3 describes the classification based on LFM and the linear movie box-office prediction models. Section 4 shows the experimental results and analysis. Section 5 describes the conclusions and future works.

II. RELATED WORK

Prediction is an important task in recommendation analysis. Generally, the commonly used methods are based on user's registration information and historical behaviors. Reference [1] uses a deep learning approach to map users and items to a latent space and the similarity between users and their preferred items is maximized. To improve the sensitivity of user engagement metrics, reference [2] uses an approach by utilizing prediction of the future behavior of an individual user. Reference [3] uses an extended reinforced poisson process model with time mapping process to model the retweeting dynamics and predict the future popularity. Reference [4] improves the home-screen apps' usage experience through a prediction mechanism that allows to show to users which app he/she is going to use in the immediate future. And the prediction technique used in reference [4] is based on a set of features representing the real-time spatiotemporal contexts sensed by the home-screen apps. Reference [5] embarks on the challenges to investigate the trust prediction problem with the homophily effect. Inspired by recent successes of offline evaluation techniques for recommender systems, reference [6] uses the historical search logs to reliably predict online click-based metrics of a new ranking function, without actually running it on live users. From external knowledge sources such as online social networks, reference [7] starts a project that aims at studying the extent to which links between buyers and sellers, i.e. trading interactions in online trading platforms, can be predicted. Reference [8] describes and evaluates methods for learning to forecast forthcoming events of interest from a corpus containing 22 years of news stories. Reference [9] proposes the basic model and method that predicting box office. Reference [10] publishes that box office and the corresponding numbers of search queries are positively related. Google uses a liner regression model to predict the box office. The independent variables in the model include the numbers of search queries and the number of movie adventure hits. According to the historical information of the movie, reference [11] uses simply liner regression model to predict cumulative number during the movie opened. Reference [12] proposes a new model to predict box-revenue of movie, and it is based on the neural network. By studying the relationship between the film and the social media user behavior, reference [13] reveals the powder of word-of-mouth on the sales performance. Reference [14]

978-0-9567157-5-3 © 2014 IEEE

specifically analyzes both positive and negative sentiment on social media.

This paper proposes a novel model to predict box-office of movie. The experiment result also shows that this liner regression model based on LFM classification has better performance than expert classification, especially for foreign movies released in China.

III. METHODOLOGY

The movie box-office revenue prediction proposed in this paper is based on movie classification, so the primary work is to classify the movies. After that, for each category, this paper uses users' reviews and the corresponding ratings to build the linear prediction models.

A. LFM Algorithm Based Classification

Movies can be divided into domestic and international by production country. According to the movie type, they can also be divided into action movies, science fiction movies, cartoon movies and so on. In 2014, the movies which released in China contained 36 types， and some types include relatively few films, for example, there are only two films ,i.e., "*Breakup Buddies"* and "*Continent"*, belong to the type of road movie. Generally, for linear regression model, small data set may get better fitting curve, but the estimation error would be large at the same time.

Latent Factor Model (LFM) is a kind of latent semantic analysis technology. Recently, it always applies in some recommendation system to cluster the items automatically. That is to say, it can find out the latent topics and categories. Through the formula (1), LFM calculates the user's score on the each item.

$$R_{UI} = P_U Q_I = \sum_{K=1}^{K} P_{U,K} Q_{K,I} (1)$$

In formula (1), R_{UI} is a real user-item matrix and the value in it represents user's score on the item. K stands for the number of categories, so LFM can learn about the category which item depends on from the matrix $Q_{K,I}$ .The number in matrix $Q_{K,I}$ represents the degree of correlation between the item and its category. It is not necessary to concern about classification angle and the granularity of classification, because the classification results are formed automatically according to the users' ratting behavior. By formula (2), i.e., the optimal loss function, we can obtain the matrix $Q_{K,I}$.

$$C = \sum_{(U,I)\in K} \left(R_{UI} - \sum_{K=1}^{K} P_{U,K} Q_{K,I}\right)^2 + \lambda \|P_U\|^2 + \lambda \|Q_I\|^2 (2)$$

$\lambda\|P_U\|^2 + \lambda\|Q_I\|^2$ in formula(2) is used to prevent over fitting. During the iterative process, $P_{Uk}$and Q_{KI} are updated by the formula(3) and formula(4) step by step. The meaning of αis the corresponding learning speed.

$$P_{Uk} = P_{Uk} + \alpha\left(\left(R_{UI} - \sum_{K=1}^{K} P_{U,K} Q_{K,I}\right) Q_{KI} - \lambda P_{UK}\right) (3)$$

$$Q_{KI} = P_{Uk} + \alpha\left(\left(R_{UI} - \sum_{K=1}^{K} P_{U,K} Q_{K,I}\right) P_{Uk} - \lambda Q_{KI}\right) (4)$$

The algorithm of the proposed approach is described as follows.

Algorithm : LFM Algorithm based Classification

```
Input:  $R_{UI}$, K, λ, α N, E;
Output:  $P_{U,K}$, $Q_{K,I}$;
Steps:
1.    [$P_{U,K}$, $Q_{K,I}$] = Init Model [$R_{UI}$, K].
2.    for step in range(0,N)
          for user, items in $R_{UI}$
            err = err  - Predict(user, item)
            for f in range(0, K)
            P[user][f] += α* (eui * Q[f][item] - λ* P[user][f])
            Q[f][item] += α* (eui * P[user][f] - λ* Q[f][item]
              End if err<E
End for
End for
```

B. Linear regression prediction models

Linear regression is an approach for modeling the relationship between a scalar dependent variable*y*and one or more explanatory variables (or independent variable) denoted *x*. It is the first type of regression analysisto be studied rigorously, because this models which depend linearly on their unknown parameters are easier to fit than the models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. From the figure 1, it can obviously find that there is one kind of linear relationship exists in the box-office and user reviews.

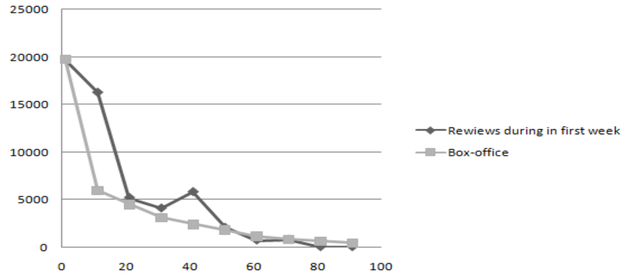

Fig.1. The relationship between box-office and reviews

In order to predict the movie box-office revenue, this paper proposes the five kinds of models. The meaning of the variables in each model is shown in table1.

$$Model1{:}\ R = \theta_1 + k_{11} FBO$$

$$Model2{:}\ R = \theta_2 + k_{21} TN$$

$$Model3{:}\ R = \theta_3 + k_{31} FIVES + k_{32} FOURS + k_{33} THREES + k_{34} TWOS + k_{35} ONES$$

$$Model4{:}\ R = \theta_4 + k_{41} TN + k_{42} FBO$$

$$Model5{:}\ R = \theta_5 + k_{51} FIVESS + k_{52} FOURS + k_{53} THREES + k_{54} TWOS + k_{55} ONES + k_{56} FBO$$

TABLE I
THE MEANING OF THE VARIABLES IN FIVE MODELS

Variable	The meaning
R	Box-office revenue
FBO	Daily box office of the first week
TN	The number of rating during the first week
FIVES	The number of5 stars during the first week
FOURS	The number of 4 stars during the first week
THREES	The number of 3 stars during the first week
TWOS	The number of 2 stars during the first week

978-0-9567157-5-3 © 2014 IEEE

ONES	The number of 1 stars during the first week

The θ and k in the linear regression models above are unknown parameters. In order to get the value of these unknown parameters, least square method is applied in this paper.

IV. EXPERIMENT RESULTS AND ANALYSIS

A. Movie box-office and the corresponding review dataset

With the rapid development of social media, more and more people share their opinion and rate the items in vertical media, such as Douban and IMDB. In this paper, the box-office and details of movies (type, production country, etc.) is provided by China Movie Box-Office Network. The rating date is from Douban which is the most popular movie review site in china. In detail, the movie dataset includes 100 movies which released in China and the box-office at top-100 in 2014.In addition, it consists of 37 foreign movies and 63 Chinese movies. The corresponding rating comment includes 619459 reviews that rated by 201748 users during the first week after the movie released. What's more, the 37 foreign movies contain 339290 reviews and 141710 users, while the 63 Chinese movies contain 280169 reviews and 132560 users. Whether Chinese movies or foreign movies, the relationship between the amount of movies that one user has rated and the corresponding amount of user belongs to a long-tailed distributions (see Figure2).The rating distribution in foreign movies matrix is basically the same, see figure2.

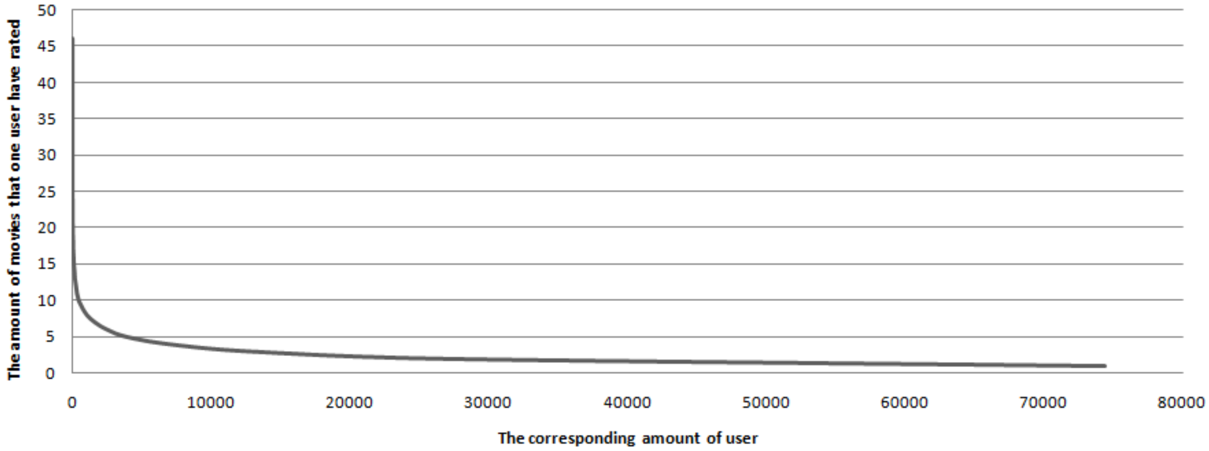

Fig.2.The rating distribution in Chinese movies matrix

B. Evaluation metrics

In formula (5), R^2 is used as the evaluation metrics. It represents the fitting degree of the linear regression model. The computation formula is shown as follows.

$$R^2 = \frac{\sum_{i=1}^{N}(X_i-\bar{X})^2\,(Y_i-\bar{Y})^2}{\sum_{i=1}^{N}(X_i-\bar{X})^2\sum_{i=1}^{N}(Y_i-\bar{Y})^2} \qquad (5)$$

C. Experiment results and analysis

The LFM is used to classify the movies, while the experts classification (by production country and type) is used to make a contrast test. Building the linear regression models is based on the above work.

1) Classification

According to the Chinese film box office market, the total dataset in this experiment is divided into two parts: foreign ones and Chinese.

For different dataset, the unknown parameters (λ, α) in formula 2 and formula 3 are not the same. A lot of experiments are need to get the unknown parameters in order to obtain the optimal function and efficient learning speed. Especially, the unknown parameterα in formula 3 is in a constant state of flux. At the beginning of learning process, in case of the huge gap between R_{UI} and $P_{U,K}Q_{K,I}$, the parameter (i.e., α) should be relatively larger. While, during the iterative process, the computing result of $P_{U,K}Q_{K,I}$ is reachingR_{UI}step by step. So, the learning speed controlled byα should be smaller at the next iterative. Then, after the iterative process, the max number in each column of $Q_{K,I}$ determines the category which the corresponding movie belongs to. The LFM classification result is shown in table 2.In addition, the *group* in table2 is different from the category (i.e., action and so on) in table3. The *groups* are generated by LFM classification. Even though each *group* has no specific meaning and it is only a movie collection, but the movies in each group have some latent meaning.

TABLE II
THE RESULT OF LFM CLASSIFICATION

λ	α	Production country	Category	Number
0.005	0.005	Foreign	Group1	10
		Foreign	Group2	13
		Foreign	Group3	14
0.01	0.01	Chinese	Group1	3
		Chinese	Group2	9
		Chinese	Group3	22
		Chinese	Group4	15
		Chinese	Group5	16

According to movie type published by distributor, the experts classification divides the Chinese movies into five categories (i.e., action, love story, comedy, thrille and cartoon story). Similar as the Chinese movies, the foreign movies are divided into three categories. The detailed result of expert classification is shown in table 3.

TABLE III
THE RESULT OF EXPERTS CLASSIFICATION

Production country	Category	Number
Foreign	action	16
Foreign	science fiction	13
Foreign	cartoon	8
Chinese	action	16
Chinese	love	22
Chinese	comedy	8
Chinese	thriller	4
Chinese	cartoon	13

2) *Linear prediction models experiment*

The results of the fitting degree in five models for each category are shown as follows.Table4 shows the results based on experts classification, while experiment results in table5 are based on LFM. From the results of table4 or table5, we can obviously find that the model5 performs better than the others. The results indicate the box-office correlated strongly with box office during the first week, and the number of users' reviews that include one star to five starts.

TABLE IV
THE RESULT OF R^2 BASED ON EXPERTS CLASSIFICATION

Production country	Category	R^2				
		Model1	**Model2**	**Model3**	**Model4**	**Model5**
Foreign	action	0.86	0.425	0.741	**0.862**	**0.862**
	science fiction	0.935	0.256	0.938	0.958	**0.981**
	cartoon	0.83	0.489	0.955	0.879	**0.987**
Chinese	action	0.622	0.596	0.818	0.78	**0.84**
	love	0.879	0.423	0.684	0.889	**0.924**
	comedy	0.929	0.601	**1**	0.999	**1**
	thriller	0.917	0.011	0.144	0.917	**0.94**
	cartoon	0.854	0.318	0.866	0.869	**0.998**

TABLE V
THE RESULT OF R^2 BASED ON LFM CLASSIFICATION

Production country	Category	R^2				
		Model1	**Model2**	**Model3**	**Model4**	**Model5**
Foreign	Group1	0.973	0.196	0.916	0.975	**0.995**
	Group2	0.918	0.759	0.851	0.967	**0.995**
	Group3	0.834	0.773	0.8	0.96	**0.985**
Chinese	Group1	0.898	0.879	**1**	**1**	**1**
	Group2	0.891	0.806	0.955	0.925	**0.987**
	Group3	0.933	0.415	0.674	0.934	**0.971**
	Group4	0.762	0.834	0.906	0.935	**0.96**
	Group5	0.628	0.131	0.564	0.644	**0.718**

In figure 3, the X-axis represents the average fitting degree about all the movies include the Chinese and the foreign movies in each model .While, the result shown in figure4 and figure5 are respectively about the Chinese and the foreign movies. From the figure 4, the average $R^2$of prediction model5 based on LFM is as high as 95.1%. Meanwhile, for all the five models, classification based on LFM performs better than the manual results proposed by the experts. Especially, it makes the fitting degree improved 29.9% compared with model2. The results in the figure4 and figure5 are almost the same. While, in figure 5, the average R^2 of model5 reduces 1% which based on LFM algorithm.

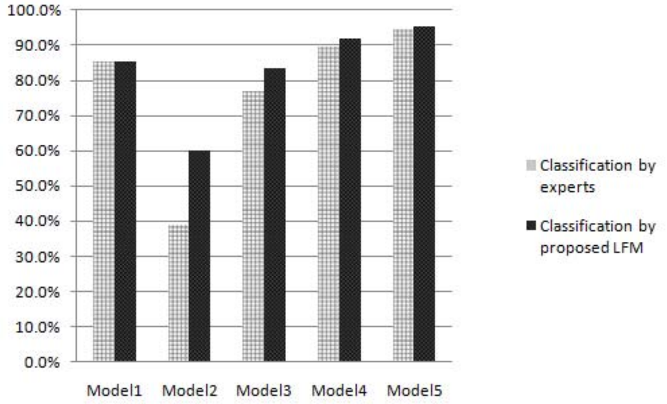

Fig.3.The R^2in the models include all 100 movies

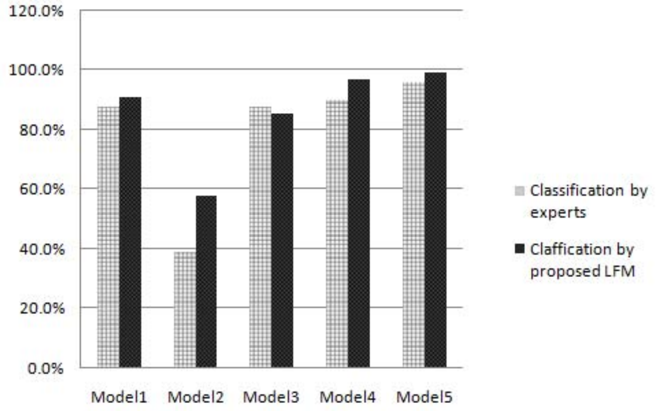

Fig.4.The R^2 in the models include 37 foreign movies

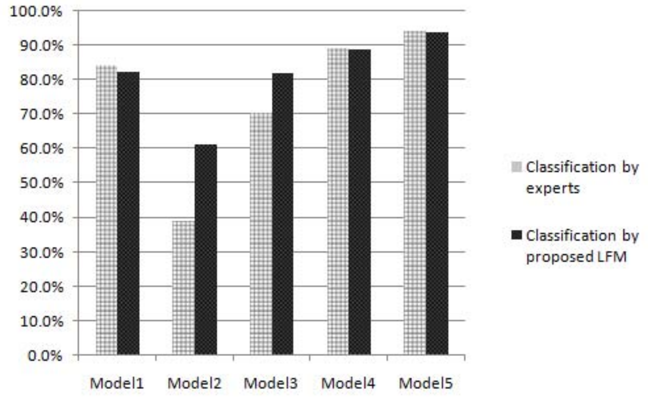

Fig.5.The R^2 in the models include 63 Chinese movies

V. CONCLUSIONS AND FUTURE WORKS

Item classification and model building are essential in prediction analysis. This paper applies both the manual results and the LFM classification to divide the movie set. Then, for each movie set, the linear regression is used to build the box-office revenue prediction models. Experiment results show the LFM classification has better performance than the experts ones, especially on the date set of foreign movies released in China. As for the future works, it is necessary to optimize the LFM algorithm to reduce the time complexity. It is also necessary to add some more details about the movies to improve the performance of the model such as the film director's influence and so on. With the help of LFM, more detailed variables will be combined to improve the performance.

ACKNOWLEDGMENTS

This work is sponsored by National Science Foundation of Hebei Province (No. F2013208105) and National Science Foundation of China (No.61272362). It is also supported by Key Research Project for University of Hebei Province (No. ZD2014029) and Society Science Development Research of Hebei Province (No.2015030344)

REFERENCES

[1] Ali. Elkahky, Y. Song, X. D. He. "A Multi-View Deep Learning Approach for Cross DomainUser Modeling in Recommendation Systems",*In Proceedings of the 24nd international World Wide Web Conferences Steering Committee*, pp.278-288, 2015.

[2] A. Drutsa, G. Gusev, P. Serdyukov. "Future User Engagement Prediction and Its Application toImprove the Sensitivity of Online Experiments",*In Proceedings of the 24nd international World Wide Web Conferences Steering Committee*, pp.256-266, 2015.

[3]S. Gao, J. Ma, Z. M. Chen. "Modeling and Predicting Retweeting DynamicsonMicroblogging Platforms",*In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* pp.107-116, 2015.

[4] R. Baeza-Yates, D. Jiang, F. Silvestri, et al."Predicting The Next App That You Are Going To Use",*In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* pp.285-294, 2015.

[5] J. Tang, H. Gao, X. Hu."Exploiting homophily effect for trust prediction ",*In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining,* pp.53-62, 2013

[6] L. Li, JY. Kim, I. Zitouni."Toward Predicting the Outcome of an A/B Experimentfor Search Relevance",*In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* pp.37-46, 2015

[7] C. Trattner, D. Parra, L. Eberhard, et al."Who will Trade with Whom?Predicting Buyer-Seller Interactions in Online TradingPlatforms through Social Networks",*In Proceedings of the 24nd international World Wide Web Conferences Steering Committee,* pp.387-388, 2014

[8] K. Radinsky, E. Horvitz."Mining the Web to Predict Future Events",*In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining,* pp.255-264, 2013

[9] B. R. Litman, L. S. Kohl."Predicting financial success of motion pictures: The'80s experience[J]",*Journal of Media Economics,* pp.35-50, 1989.

[10] R. Panaligan,A. Chen. " Quantifying movie magic with google search[J]",*Google Whitepaper—Industry Perspectives+ User Insights,*2013.

[11] P. Marshall, M. Dockendorff, S. Lbáñez."A forecasting system for movie attendance",*Journal of Business Research*, 66(10): pp.1800-1806, 2013.

[12] J. Zheng, S. Zhou."Modeling on box-office revenue prediction of movie based on neural network",*Journal of Computer Applications*, 3:030, 2013.

[13]M.S. Zhou,D. M. Han."Prediction model for film box office based on social media user comments and user attention",*Microcomputer & Its Applications,* 33(18): pp.73-75, 2014.

[14] W. Wang, W. Ning, H. Wang."Online negative public sentiment does not matter?—Empirical evidence from social media and movie industry",*Computing, Networking and Communications (ICNC), 2015 International Conference on. IEEE,* pp.1122-1126, 2015