Prediction Model of Movie Box Office Based on Social Media Data

Lili Yuan[1,a], Wen Yu[2,b,*]

[1] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China

[2] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China

[a]yuanlili1114@126.com, [b]yuwen@bupt.edu.cn

*Lili Yuan

Keywords: Box office prediction, Popularity of actor, Popularity of tweets, LDA.

Abstract. In the current movie box office forecasting research, most of the research is based on the basic features in the movie mining, ignoring the influence of actors in the social media platform on movie box office. This paper proposes to crawl the Sina Weibo and comment data, analyzing the popularity of actors and tweets. In addition, we propose the LDA topic model based on popularity of tweets, which as the input of the LDA model. Then the topic distributions based on popularity of actors are obtained. Combined with the above two features, a variety of regression algorithms are used to construct the prediction model. The experimental result shows that the feature extraction method can improve the forecasting effect of the box office to a certain extent.

基于社交网络数据的电影票房预测模型

袁丽莉[1, a]，余文[2,b,*]

[1]北京邮电大学，智能通信软件与多媒体北京市重点实验室，北京，中国

[2]北京邮电大学，智能通信软件与多媒体北京市重点实验室，北京，中国

[a]yuanlili1114@126.com, [b]yuwen@bupt.edu.cn

[*]袁丽莉

关键词:票房预测; 演员热度; 微博热度; LDA

中文摘要. 当前电影票房预测中，大部分研究是挖掘电影基本特征，忽略了社交网络平台上演员对电影票房的影响。基于上述问题，本文提出爬取新浪微博平台上与电影演员有关的微博及评论数据，分析演员热度特征，同时提出基于微博热度的LDA主题模型，通过计算微博热度作为LDA模型的输入，从而得到演员的热度主题分布特征。结合以上两种特征，使用多种回归算法构建预测模型。实验结果表明，该特征提取方法在一定程度上能提升票房的预测效果。

1. 引言

随着社交网络的普及，越来越多用户通过社交网络平台获取各种各样的资讯。同时，用户也会将自己的评论、想法发表到社交网络平台上。如何从海量的数据中挖掘出有价值的信息已经成为各行业争相研究的内容。其中电影是一个高风险的文化产业，我国目前只有少数电影投资是盈利的，70%的国产电影基本都难以回收成本。鉴于电影票房收入是衡量电影成功与否的重要指标[1]，对电影准确预测能更好地降低市场风险，促进电影市场的发展[2]。

目前，对电影票房预测的方法主要是提取电影的基本信息作为特征。Eliashberg[3,4]等通过使用网络数据，研究电影的类型与电影票房的关系，实验结果显示，电影的类型对票房有影响，但分类的类型过少，无法满足目前多种电影类型；Nelson[5]等研究了电影票房与电影续集的关系，结果表明续集会提高票房的收益，这是由于父电影口碑会给子电影带来一定影响，但没有定量进行票房预测；Goetzman[6]等发现在电影上映前后，电影评论会对票房产生影响，但只考虑了比较简单的电影属性，预测效果并不理想；Sharda[1]等使用神经网络作为预测模型，探究电影时长、上映日期、季节等因素与票房的关系，但提取电影票房的特征不够全面，且数据量过少，预测效果还有提升空间。Liu[7]等通过研究在社交网络上用户的购买意图、对用户评论进行情感分析来预测电影票房，但挖掘特征的方法比较简单，并没有考虑电影主演在社交网络上的影响力。

综合上述问题，本文主要研究电影演员微博以及热门评论数据，分析演员热度。同时，利用基于热度的LDA模型得到电影演员的热度主题分布特征。通过提取以上特征，结合多种机器学习的回归算法，对电影票房进行有效的预测。

2．特征构建

有研究表明，演员的人气和电影票房具有正相关关系。本文主要从两个方面进行特征提取：演员热度和演员热度主题分布特征。通过分析演员在微博上的热门度，提取演员热度特征，此特征可以展示出各个电影演员的人气。不同演员的合作也是影响电影票房的一个重要因素之一，本文根据演员的微博热度，使用LDA模型，挖掘演员热度主题分布特征，展示出不同演员组合对电影票房的影响。

2.1 演员热度

演员热度特征是挖掘演员在微博上的人气度。分析演员每条微博的转发数、评论数、点赞数、以及发微博的频率等数据可以直观的展现出该演员在微博平台的热门程度，从而判断用户对其作品的兴趣度。每个演员热度特征计算方法如表1：

表1　演员热度特征计算方法

特征	计算公式	备注
微博数	$WEIBO\_NUM(i)=wn$	wn表示演员i所发的微博总数。
关注数	$FOLLOW\_NUM(i)=fol$	fol表示演员i的关注数。
粉丝数	$FANS\_NUM(i)=fans$	fans表示演员i的粉丝数。
平均转发数	$AVG\_TRAN(i)=\frac{\sum_{j=1}^{wn}t_j}{wn}$	t_j 表示第j条微博的转发数。
最大转发数	$MAX\_TRAN(i)=\max(t_j),\ j\in[1,wn]$	
平均评论数	$AVG\_COM(i)=\frac{\sum_{j=1}^{wn}c_j}{wn}$	c_j 表示第j条微博的评论数。
最大评论数	$MAX\_COM(i)=\max(c_j),\ j\in[1,wn]$	
平均点赞数	$AVG\_LIKE(i)=\frac{\sum_{j=1}^{wn}l_j}{wn}$	l_j 表示第j条微博的评论
最大点赞数	$MAX\_LIKE(i)=\max(l_j),\ j\in[1,wn]$	
平均转发率	$AVG\_PR(i)=\frac{\sum_{j=1}^{d}p_j}{d}$	p_j 表示第j天所发微博数，d表示某段时间内的天数。
最大转发率	$MAX\_PR(i)=\max(p_j),\ j\in[1,d]$	

2.2 演员热度主题分布

演员组合是影响票房的重要因素之一，选择合适的演员搭配往往能比单个演员带来更大的话题度，吸引观众去影院观看电影。因此，本文引入微博热度计算方法，通过挖掘演员的热度主题分布特征进行票房预测。

2.2.1 LDA模型

LDA[8]是一个包括文档、主题、单词的3层贝叶斯模型，使用概率的方法对模型进行推导，从而找出文本集的语义结构，挖掘文本的主题。其基本原理是每个文本都可用一系列主题的混合分布来表示，记为 $P(z)$，同时每个主题是词表中全部单词上的概率分布，记为 $P(w|z)$。因此，一个文本中每个单词的概率分布如式（1）所示：

$$P(w_i)=\sum_{i=1}^{T}P(w_i|z_i=j)P(z_i=j) \tag{1}$$

2.2.2 基于热度的LDA主题模型

本文利用电影演员的微博热度，提取出演员热度主题分布特征。微博热度可以很好地反映出演员所产生的微博影响力，能扩散到多少人看到这条微博，获取该条微博所传达的信息。对于微博热度计算是根据信息论中自信息量的定义：一个随机事件发生某一结果所带来的信息量，式（2）表示事件A所包含的信息量：

$$I(A)=-\log P(A) \tag{2}$$

假设微博m 的评论数为c，转发数为r，点赞数为l，则此微博的热度计算方式如式（3），一个含有N个单词的微博，每个单词的热度计算方法如式（4）：

$$Heat(m)=-\log\frac{1}{c+r+l+1} \tag{3}$$

$$Heat(w)=\frac{Heat(m)}{N} \tag{4}$$

在引入微博热度的定义后，基于热度的LDA模型的分布中，$z'=Heat(z)$ 表示某个主题的热度，$w'=Heat(w)$ 表示某个单词的热度，总体服从的函数分布不变，式（5）（6）为主题热度概率 $p(z'=j)$ 和词的热度概率 $p(w_i'|z'=j)$ 的计算方式：

$$p(z'=j)=\frac{Heat(z'=j)}{\sum_{i=1}^{T}(z'=i)} \tag{5}$$

$$p(w_i'|z'=j)=\frac{Heat(w_i,z'=j)}{Heat(z'=j)} \tag{6}$$

根据分布的定义代入Dirichlet分布函数可得 $p(z'),p(w'|z')$：

$$p(z')=\left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D\prod_{d=1}^{D}\frac{\prod_j\Gamma\left(h_j^d+\alpha\right)}{\Gamma\left(h_*^d+T\alpha\right)} \tag{7}$$

$$p(w'|z')=\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T\prod_{j=1}^{T}\frac{\prod_w\Gamma\left(h_j^w+\beta\right)}{\Gamma\left(h_j^*+W\beta\right)} \tag{8}$$

由于上式不可直接计算，在采用Gibbs采样时，通过构造马尔可夫链进行迭代后使Gibbs采样接近于目标分布[9]，此后的Gibbs采样便可以用来表示目标分布的样本值，用来计算后验概率 $p(z_i|z_{-i},w_i)$：

$$p(z_i=j|z_{-i},w_i)\propto\frac{h_{-i,j}^{w_i}+\beta}{h_{-i,j}^{*}+W\beta}\frac{h_{-i,j}^{d_i}+\alpha}{h_{-i,*}^{d_i}+T\alpha} \tag{9}$$

由此可知，$(h_{-i,j}^{w_i}+\beta)/(h_{-i,j}^{*}+W\beta)$代表$w_j$在主题j的热度分布，$(h_{-i,j}^{d_i}+\alpha)/(h_{-i,*}^{d_i}+T\alpha)$代表主题j在微博$d_j$中的热度分布。

在进行Gibbs采样时，取得一部分后验分布值$p(z|w)$，可以计算θ和φ的样本估计值：

$$\hat{\theta}_j^d=\frac{h_j^d+\alpha}{h_*^d+T\alpha} \tag{10}$$

$$\hat{\varphi}_j^w=\frac{h_j^w+\beta}{h_j^*+W\beta} \tag{11}$$

由于计算量是热度，因此在Gibbs采样过程中需要稍作改进。Gibbs采样通过对LDA中的两个关键的矩阵——文档主题矩阵、主题词矩阵与$p(z_i|z_{i-1},w_i)$之间的循环迭代计算，当数值变化收敛时便终止，此时矩阵和各参数的值便是最终结果。改进之处在于矩阵和$p(z_i|z_{i-1},w_i)$之间的循环迭代时，矩阵之中每次变化的是主题或词的热度。

2.2.3 演员热度主题分布特征计算方法

基于热度的LDA主题分布模型算法流程如图1所示：

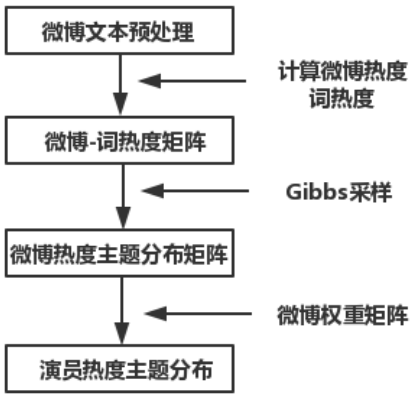

图1　演员热度主题分布特征流程图

a)对微博文本进行预处理；

b)通过公式（3）（4）计算每条微博的热度及词的热度，得到微博-词热度矩阵$\boldsymbol{H}$:

$$\boldsymbol{H}=\begin{bmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \vdots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{bmatrix} \tag{12}$$

其中m表示所有微博词表的大小，n表示演员在某一段时间内所发微博的数目，h_{ij}表示第i条微博中，第j个词的词热度，若微博中没有出现该词，则$h_{ij}=0$。

c)通过Gibbs采样后，得到微博热度主题分布矩阵$\boldsymbol{TH}$；

$$\boldsymbol{TH}=\begin{bmatrix} th_{11} & \cdots & th_{1n} \\ \vdots & \vdots & \vdots \\ th_{p1} & \cdots & th_{pn} \end{bmatrix}=\left[\boldsymbol{t_1},\boldsymbol{t_1},\ldots,\boldsymbol{t_n}\right] \tag{13}$$

其中，th_{ij}表示微博j在主题i下的分布概率，t_j表示微博j的主题分布。

d)计算每条微博的权重，假设微博i的评论数为c，转发数为r，点赞数为l，微博发表时间与电影上映时间的天数差值为d，则微博i的权重$weight_i$计算方法如式（14），演员j微博权重向量$\boldsymbol{w_j}$定义式（15）：

$$weight_i = \log\frac{c+r+l}{d+1} \quad (14)$$

$$\boldsymbol{w_j} = [weight_1, weight_2, \ldots, weight_n]^T \quad (15)$$

e）每个演员可以看作是由多条微博构成的，因此演员i热度主题分布hd_i计算方法如下：

$$hd_i = \boldsymbol{TH} \cdot \boldsymbol{w} \quad (16)$$

3．实验过程及结果

3.1 实验数据

本实验基于Scrapy框架编写爬虫程序，爬取豆瓣电影、电影网、微博等相关信息。从豆瓣电影网站中，爬取2016年国产电影261部；从电影网中，爬取电影每周票房数据，共1035条。从新浪微博中，爬取电影演员的微博资料、微博信息、微博评论，共310万条数据。其中新浪微博上的数据只爬取电影上映前一个月和上映后一个月内所发表的微博和评论。详细内容如表2所示：

表2 实验数据

来源	内容	数量（条）	备注
豆瓣电影	电影的基本信息	261	2016年上映的国产电影
电影网	电影每周票房信息	1035	
新浪微博	微博用户的基本信息	986	微博用户指的是电影主演、导演和官方微博
	微博信息	9.4万	只提取电影上映前一个月和上映后一个月内的数据
	每条微博下的评论信息	304.1万	提取微博下的热门评论

3.2 实验步骤

实验步骤如图2所示，首先对数据进行预处理，包括分词，去停用词，去url，去用户名，去转发内容。然后，根据第2节中提到的特征提取方法，对每个电影演员提取演员热度特征、演员热度主题分布特征。接下来，由于每部电影是与多个演员有关的，因此需要根据演员在电影中的重要程度，得到演员的权重矩阵，将多个特征向量整合为一个特征向量，并进行归一化。最后，使用SVR-rbf，Logistic Regression和Linear Regression回归模型对电影票房进行预测。

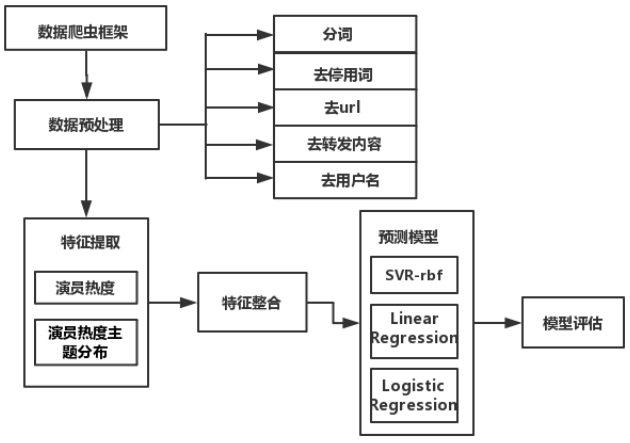

图2 票房预测流程图

3.3 评估方式

本实验采用三倍交叉验证的方式对实验结果进行验证，使用相对绝对误差（RAE）来作为预测模型的评估标准，计算方式如式（17）：

$$RAE = \frac{\sum_{i=1}^{n} | \hat{y}_i - y_i |}{\sum_{i=1}^{n} | y_i - \bar{y}_i |} \tag{17}$$

其中 $\hat{y}_i$ 为预测值， y_i 为真实值， $\bar{y}_i$ 为平均值。RAE值越小，说明票房预测效果越好。

3.4 实验结果

3.4.1 LDA主题数对票房的影响

本实验使用GibbsLDA++工具，各参数设置为Topic = 10，α=0.5，β= 0.1，循环迭代抽样的次数设为2000 次。从图3可以看出，LDA主题数过多或过少的时候，会导致RAE增大，预测效果不理想。当主题数在20-30时，三种预测模型的RAE值达到最小，预测模型效果达到最优。当主题数为25，预测模型为SVR-rbf时，预测模型达到最优效果。因此选择LDA主题数为25作为后续实验的参数。

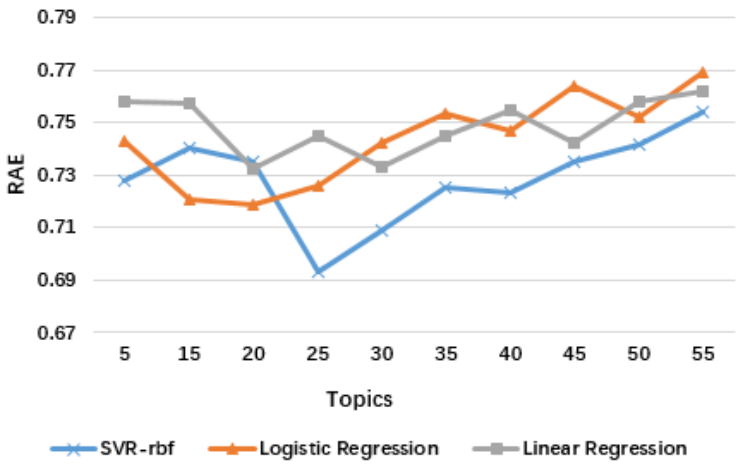

图3　LDA主题数与RAE的关系

3.4.2 传统LDA模型与基于微博热度LDA模型对比结果

实验结果如图4所示，基于微博热度的LDA模型的RAE值比传统LDA模型提取演员主题分布的特征要低，其中使用SVR-rbf作为回归模型的票房预测效果最好，说明本文提出的基于微博热度LDA模型提取特征能有效提高票房预测效果。这是因为传统的LDA模型只依赖于词频，而影响票房更重要的因素是演员的人气，因此使用微博词热度作为LDA模型的输入，得到的演员热度主题分布特征能更有效地预测票房。

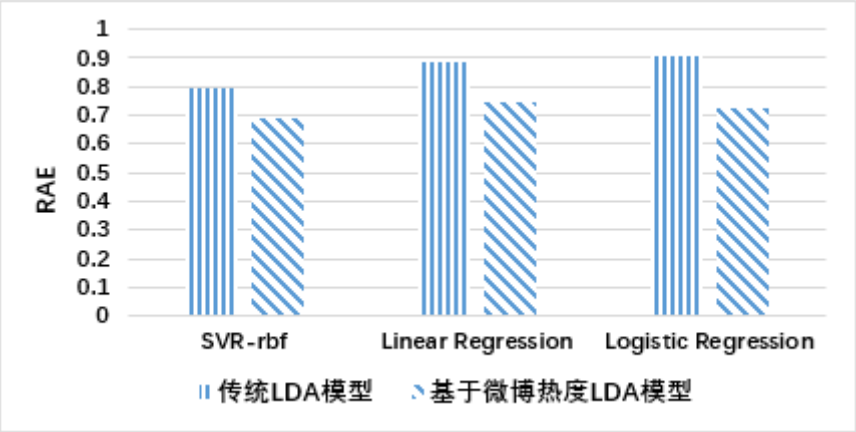

图4　传统LDA模型与基于微博热度LDA模型对比结果

3.4.3 不同特征组合对票房的影响

实验采用SVR-rbf作为预测模型。表3表示演员热度和演员热度主题分布特征在不同组合情况下的实验结果。从实验结果可以看出，仅仅使用单种特征作为预测模型的输入，RAE比较高，预测效果不理想。当使用两种特征组合时，RAE的值比使用单种特征要低，预测效果更好，这说明电影票房受演员人气影响，观众对电影喜好的往往也是由演员决定的，人气高的演员往往能带来更大的话题，引起更多观众的关注，从而带来更多的票房。

表3　不同特征组合方式的RAE

预测模型	**AP**	**PLDA**	**AP+PLDA**
SVR-rbf	0.901	0.824	**0.693**
Linear Regression	0.943	0.887	0.745
Logistic Regression	0.964	0.861	0.726

3.4.4 本文与Baseline对比实验

Baseline实验选择Liu[7]等实验结果，该实验选用社交网络中的数据，通过分析用户购买意图以及情感分析来预测票房结果。本实验中，采用电影上映一个月前的数据来预测首周票房，使用全部数据来预测总票房。实验结果如表4所示，从总的来看，首周票房RAE比总票房RAE低，说明无论是哪种预测模型，预测首周票房效果更佳。这是因为首周票房主要与前期的宣传有关，而在微博上宣传主要依赖于演员，但随时上映时间增加，影响票房的因素增多，比如电影的口碑、排片量等，因此总票房的预测效果不如首周票房好。三种预测模型中，SVR-rbf的首周票房RAE和总票房RAE最小，分别为0.316和0.693，预测效果最优。与Baseline实验比较，SVR-rbf的预测效果更好，说明本文提出的特征提取方法能有效预测电影票房，这是因为本文主要从演员的角度分析，能一定程度上提升票房预测的效果。

表4　本文实验与Baseline对比

预测模型	**首周票房RAE**	**总票房RAE**
Baseline	0.340	0.730
SVR-rbf	**0.316**	**0.693**
Linear Regression	0.338	0.745
Logistic Regression	0.352	0.726

4. 结语

本文根据目前社交网络中影响电影票房因素，挖掘演员热度和基于微博热度的演员热度主题分布特征，使用多种回归算法，构建电影票房预测模型。与以往研究重点不同，本文主要挖掘电影演员的人气度对票房的影响力，实验结果表示，该方法能一定程度提升电影票房的预测效果。

本文还有很大的提升空间，在特征提取方面，还可以分析更多影响电影票房的因素，如对用户评论进行情感分析，判断用户对演员的喜爱程度。在微博热度计算方面，热门评论内容也是衡量微博热度的一个重要因素，可以考虑如何将热门评论融入到计算微博热度中，这些都是将来有价值的研究方向。

致谢

本文为国家自然科学基金项目（11272066和11472049）的阶段性成果之一。

References

[1] Sharda R, Delen D. Predicting box-office success of motion pictures with neural networks[J]. Expert Systems with Applications, 2006, 30(2):243-254.

[2] Zhang L, Luo J, Yang S. Forecasting box office revenue of movies with BP neural network[J]. Expert Systems with Applications, 2009, 36(3):6580-6587. S. K. Goyal, A joint economic-lot-size model for purchaser and vendor: A comment, *Decision Sciences*, vol.19, pp. 236-241, 1988.

[3] Eliashberg J, Hui S K, Zhang Z J. Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(11):2639-2648.

[4] Eliashberg J, Hui S K, Zhang Z J. From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts[J]. Management Science, 2007, 53(6):881-893.

[5] Nelson R A, Glotfelty R. Movie stars and box office revenues: an empirical analysis[J]. Journal of Cultural Economics, 2012, 36(2):141-166.

[6] Goetzmann W N, Ravid S A, Sverdlove R. The pricing of soft and hard information: economic lessons from screenplay sales[J]. Journal of Cultural Economics, 2013, 37(2):271-307.

[7] Liu T, Ding X, Chen Y, et al. Predicting movie Box-office revenues by exploiting large-scale social media content[J]. Multimedia Tools & Applications, 2016, 75(3):1509-1528.

[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[9] David M. Probabilistic topic models[J]. IEEE Signal Processing Magazine, 2010, 27(6):55-65.