第 43 卷 第 7 期 计 算 机 工 程 2017 年 7 月
Vol. 43 No. 7 Computer Engineering July 2017

·先进计算与数据处理· 文章编号:1000-3428(2017)07-0001-08 文献标志码:A 中图分类号:TP18

基于首播前搜索数据的电视剧流行度预测

朱寒婷,尹 敏,贺 樑

(华东师范大学 计算机科学与技术系,上海 200241)

摘 要: 现有对视频网站电视剧流行度预测的研究中考虑因素较少,并且极少能在电视剧首播前进行预测,这会使视频网站在做出版权购买、广告投放等决策时考虑不全面并且出现预测时间滞后的问题。为此,提出一种在首播前预测视频网站电视剧流行度的方法,综合考虑电视剧剧名和演员搜索数据,通过分析时间序列确定最早预测时间,使用多元线性回归模型实现电视剧流行度的预测。实验结果表明,该方法可利用首播前第 13—18 天的剧名和演员的百度搜索指数对 PPTV 和优酷 2014 年、2015 年上线的电视剧预测上线后 30 天的点播量,预测值与真实值之间的皮尔森相关系数分别达到 0.943 7 和 0.967 6,具有较好的预测效果。

关键词: 电视剧流行度;电视剧点播量排名;多元线性回归;特征融合;最早预测时间;百度搜索指数

中文引用格式: 朱寒婷,尹 敏,贺 樑. 基于首播前搜索数据的电视剧流行度预测[J]. 计算机工程,2017,43(7):1-8.

英文引用格式: Zhu Hanting, Yin Min, He Liang. TV Drama Popularity Prediction Based on Search Data Before the Premiere[J]. Computer Engineering, 2017, 43(7):1-8.

TV Drama Popularity Prediction Based on Search Data Before the Premiere

ZHU Hanting, YIN Min, HE Liang

(Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China)

【Abstract】 Existing methods for TV drama popularity prediction in video websites solely consider the single factor and most of them cannot predict before the premiere. This will lead to the video website making unreasonable decisions on the purchase of copyright and advertising with a few days' time lag. To solve this problem, this paper proposes a method which can predict the TV drama popularity in video websites before its premiere. This method first uses the search data of TV drama such as name and actor comprehensively. Then the method calculates the earliest prediction time through time series analysis. Finally, based on multiple linear regression model, it gets the optimal feature and predicts the popularity. Experiments result shows that this method can use the thirteenth to eighteenth days' Baidu search index of name and actor before the premiere to predict 30 days' on demand quantity ranking after premiere for TV dramas launched on PPTV and YOUKU in 2014 and 2015. The Spearman correlation coefficient between the prediction rank and real rank reaches 0.943 7 on PPTV and 0.967 6 on YOUKU. The result shows that this method has a good prediction effect.

【Key words】 TV drama popularity; TV drama on demand quantity ranking; Multiple Linear Regression (MLR); feature fusion; earliest prediction time; Baidu search index

DOI: 10.3969/j.issn.1000-3428.2017.07.001

0 概述

随着三网融合技术的持续推进和智能终端的迅速普及,网络视频行业呈现出蓬勃发展之势,成为用户最喜爱的观剧渠道之一[1]。以爱奇艺、优酷、土豆等视频网站为主的新媒体为电视剧的播出提供了全新的平台,各大视频网站着手于热播电视剧的版权引进,以提高网站访问量以及广告投放量。如果能在电视剧首播前一段时间预测出电视剧流行度,那么视频网站可以提前对热播电视剧进行版权购买、广告投放等决策。目前,对在视频网站上线的电视剧流行度预测方法考虑的因素较少,并且极少能够在电视剧首播之前进行预测。

本文以在首播前预测电视剧流行度为目标,综

基金项目: 国家科技支撑计划项目(2015BAH01F02);上海市科学技术委员会科研计划项目(16511102702);上海市经济和信息化委员会项目(150643)。

作者简介: 朱寒婷(1991—),女,硕士研究生,主研方向为数据挖掘;尹 敏,讲师;贺 樑,教授、博士、博士生导师。

收稿日期: 2016-06-14 **修回日期:** 2016-08-12 **E-mail:** htzhu@ica.stc.sh.cn

合考虑电视剧剧名和演员的搜索数据,提取并融合剧名和演员的搜索特征。通过时间序列分析确定能够在电视剧首播前预测的最早开始时间,并利用多元线性回归模型预测电视剧首播后一段时间的电视剧流行度。

1 相关工作

传统电视剧版权购买决策方法多数是通过传媒机构电视剧采购员依靠对剧本、导演和演员的知名度等因素进行经验评估[2],为获得决策结果需要的投入较高[3]。随着云时代的到来,大数据吸引了越来越多人的关注[4]。如果能通过数据挖掘与分析的方法来预测热播电视剧,那么对于视频网站而言,投入成本将大大降低。这也是视频网站电视剧版权交易探索的方向。

现有在电视剧、电影等影视领域的流行度预测研究中主要有 2 类方法:基于历史数据的预测和基于外部数据的预测。

1.1 基于历史数据的预测

文献[5]对视频网站频道流行度的时间序列进行了相关的研究。文献[6]通过对视频点播的时间序列曲线进行平滑、平移、缩放并计算曲线之间的相似度对视频资源的点播量进行短期预测。这类方法可以预测视频网站中已有的视频资源点播量,但是使用此类方法的前提条件是有一定数量的历史数据,并且无法在视频资源上线前对其进行预测。

1.2 基于外部数据的预测

这类方法一般借助 2 类外部数据:社交网络数据以及搜索引擎数据。

1.2.1 社交网络数据

文献[7]通过分析电影上映前 Twitter 与电影相关 Tweet 增长的平均速率对上映周电影的票房进行预测;文献[8]通过相关性分析和聚类分析,在 Twitter 中找到了与电影票房显著相关的特征,结果表明,电影上映前几周被提到的次数与电影票房之间的相关性达到 0.86;文献[9]通过分析与电影相关的博客数对电影票房进行预测。不同社交平台有不同特点,用户情况也各不相同,社交网络数据较复杂,需要考虑的因素较多。

1.2.2 搜索引擎数据

文献[10]利用电影上映前 4 周发布的预告片标题相关搜索量、季节性因素和特许经营状态对电影票房进行了预测,相关性接近 0.94。基于搜索数据的预测方法对电影票房的预测效果显著,但是该方法是否同样适用于电视剧点播量的预测仍需进一步研究分析。

1.2.3 社交网络和搜索引擎的融合数据

文献[11]融合社交网络和搜索引擎数据对电视剧点播量排名进行预测,研究社交网络中与电视剧点播量显著相关的特征、首播前的新浪微博数据以及首播后的百度搜索数据,定义了电视剧生命周期,利用多元线性回归模型进行点播量排名预测视频点播系统中电视剧生命周期 30 天的集均点播量排名。该方法对电视剧点播量排名进行了很好预测,但是用到了首播后 3 天的搜索数据,无法在电视剧上线前进行预测,对于视频网站决策时间仍有一定的滞后性。

电影、电视剧虽同为视频资源,但是两者的预测方法区别较大。由于电影票房的预测是线下行为,因此大多借助外部数据进行预测,不会出现预测时间滞后的问题。而在视频网站上线的电视剧点播量预测是在线行为,目前已提出一些基于历史数据以及基于外部数据的预测方法,但是现有方法均存在预测时间滞后的问题,无法在电视剧上线前进行预测。针对以上预测方法的缺点和不足,本文提出利用上线前电视剧相关的搜索引擎数据来预测电视剧上线后一段时间的流行度,解决预测时间滞后的问题。本文提出方法和现有预测方法性能情况对比如表 1 所示。电视剧在电视台首播后并被视频网站购买时才会在视频网站上线,首播时间可以提前确定,而上线时间无法提前确定。由于首播时间早于上线时间,因此本文假设首播时间和上线时间为同一个时间。本文以预测在视频网站上线的电视剧在上线后生命周期30 天[11]内每天的集均点播量为目标,提出基于首播前搜索引擎数据的多元线性回归模型进行预测的方法。

表 1 本文方法与现有预测方法的性能对比

类型	方法	基于历史点播数据	基于社交网络数据	基于搜索引擎数据	能否提前预测	预测时间是否滞后
电影	文献[7]方法	×	√	×	√	×
	文献[8-9]方法	×	√	×	√	×
	文献[10]方法	×	×	√	√	×
电视	文献[5]方法	√	×	×	×	√
	文献[6]方法	√	×	×	√	√
	文献[11]方法	×	√	√	√	√
	本文方法	×	×	√	√	×

本文主要的贡献有:1)研究电视剧首播前可以进行预测的最早开始时间。2)研究电视剧剧名、演员等电视剧搜索特征与电视剧流行度的关系。3)研究不同的特征融合方式,在电视剧搜索特征集合中找到与电视剧点播量排名预测显著相关的最优特征。

2　视频网站电视剧流行度预测方法

本文以在首播前预测电视剧流行度为目标,从预测问题定义、预测方法的整体思路以及预测模型 3 个方面对视频网站电视剧流行度预测方法进行详细阐述。

2.1　预测问题定义

本文主要关注的问题为:

1)是否能够在首播前一段时间预测电视剧流行度,确定可以预测的最早开始时间,即最早预测时间。

2)电视剧首播前的相关搜索数据和电视剧流行度的关系。

为更好地说明预测问题及方法,做如下定义:

1)电视剧集合(T):$T=\{T_1,T_2,\cdots,T_k,\cdots,T_n\}$,电视剧 k 用 T_k 表示,n 为电视剧个数。

2)电视剧 T_k 集数为 E_k,首播时间 D_k,主要演员 A_k。

3)电视剧 T_k 剧名搜索特征 TI_k,演员搜索特征 AI_k。

4)电视剧 T_k 的每日点播量(CR^k):$CR^k=\{CR_1^k,CR_2^k,\cdots,CR_m^k\}$,$CR_m^k$ 为电视剧 T_k 上线后第 m 天的点播量,m 为预测天数 30 天。

5)电视剧 T_k 在预测天数 m 内的点播总量(CTR_k):$CTR_k=\sum_j^m CR_j^k$,$1\leqslant j\leqslant m$,集均点播量($ACTR_k$):$ACTR_k=CTR_k/E_k$。

6)电视剧集合(T)的点播量(CTR):$CTR=\{CTR_1,CTR_2,\cdots,CTR_n\}$,集均点播量($ACTR$):$ACTR=\{ACTR_1,ACTR_2,\cdots,ACTR_n\}$。

7)电视剧 T_k 的流行度 $RANK_k$:$RANK_k$ 为 $ACTR_k$ 在 $ACTR$ 中的排名。

8)在电视剧集合(T)首播前一段时间的时间序列中可以预测的最早开始时间,即最早预测时间为 d;描述时间数据的最小单位为时间粒度 g。时间粒度是指日、周、旬、月、季、年等时间跨度,是描述时间数据的最小单位,表示时间点之间的离散化程度。

视频网站衡量电视剧流行度的量化指标是网络点播量,反映了观众对该剧的认可程度以及电视剧的热播程度,而集均点播量更能代表电视剧的平均水平。因此,本文对电视剧流行度的预测通过预测电视剧集均点播量排名实现。

2.2　基于搜索数据的电视剧点播量排名预测方法

本文提出基于外部数据的电视剧点播量排名预测方法,利用首播电视剧剧名和演员的搜索引擎数据来预测电视剧在视频网站上线后 30 天的集均点播量排名。本文首先确定电视剧首播日期前 4 周的剧名和演员的搜索特征,通过时间序列分析以及特征融合进行特征提取,利用多元线性回归模型预测电视剧在视频网站上线后 30 天的集均点播量,对不同特征集合的集均点播量结果进行准确率测评得到最优特征,根据最优特征预测电视剧集均点播量并进行点播量排名,如图 1 所示。

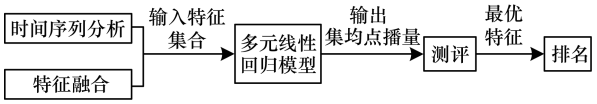

图 1　基于搜索引擎数据的电视剧点播量排名预测流程

2.2.1　电视剧搜索特征确定

观众对于未播出的电视剧的关注程度可以体现在与电视剧相关信息的搜索次数上。电视剧是否流行在于观众对其的关注度。为了能够提高观众对电视剧的关注度,提高电视剧的热播程度,电视剧制作方把更多的关注放在请名演员等方面;电视剧的播出方则动用各种手段做足播出前的宣传工作[12]。任何一部电视剧播放之前都会进行各种各样的广告宣传。在这期间,电视剧主要演员原有的知名度及其在观众中的影响力通常能起到很大的作用[13]。因此,除了电视剧剧名外,电视剧参演的演员是否具有影响力,也是吸引观众去欣赏该影视作品的推动因素之一。

本文利用百度指数统计平台统计电视剧剧名和演员的百度搜索指数并确定剧名和演员的搜索特征,百度搜索指数是以网民在百度的搜索量为数据基础,以关键词为统计对象,科学分析并计算出各个关键词在百度网页搜索中搜索频次的加权总和[14],具体步骤如下:

1)剧名百度搜索指数时间序列

本文通过对电视剧剧名搜索数据进行时间序列分析,确定最早预测时间以及时间粒度。电视剧 T_k 自首播时间 D_k 起前 4 周剧名每天的百度搜索指数为 $SDT_k=\{SDT_{k28},SDT_{k27},\cdots,SDT_{k1}\}$,不同时间粒度下百度搜索指数总和为 SST_k。

2)剧名搜索特征

若最早预测时间为 d,时间粒度为 g,电视剧 T_k 在首播时间 D_k 前 d—$d+g$ 天之间剧名的百度搜索指数为 $SDT_k=\{SDT_{kd+g},SDT_{kd+g-1},\cdots,SDT_{kd}\}$,前 d—$d+g$ 天之间剧名百度搜索指数总和为 SST_k

$=\sum_{i}^{d+g} SDT_{ki}, d \leqslant i \leqslant d+g$。剧名搜索特征为 $TI_k = SST_k$。

3)演员搜索特征

若最早预测时间为 d,时间粒度为 g,电视剧 T_k 在首播时间 D_k 前 d—$d+g$ 天之间演员每天的百度搜索指数为 $SDA_k = \{SDA_{kd+g}, SDA_{kd+g-1}, \cdots, SDA_{kd}\}$,前 d—$d+g$ 天之间演员百度搜索指数总和为 $SSA_k = \sum_{i}^{d+g} SDA_{ki}, d \leqslant i \leqslant d+g$;演员搜索特征为 $AI_k = SSA_k$。

2.2.2 特征融合和预测

特征融合和预测步骤具体如下:

1)根据时间粒度 g 以及最早预测时间 d,确定电视剧剧名搜索特征 TI 以及演员搜索特征 AI。

2)对比不同搜索特征对电视剧点播量预测的影响,对 TI,AI 进行特征融合,具体为:

(1)特征值相加的特征融合方式。将 TI,AI 的特征值相加,构成新特征 $TI+AI$。

(2)特征数增加的特征融合方式。将 TI,AI 特征进行组合,构成新特征 $\{TI, AI\}$。

(3)确定电视剧集合 T 的特性特征集合 $X = \{TI, AI, TI+AI, \{TI, AI\}\}$。

(4)在特征集合 X 中选出最优特征 X^b,将最优特征 X^b 作为多元线性回归模型输入参数预测电视剧集均点播量。

(5)根据集均点播量预测值进行排名,得到电视剧流行度预测的最终结果。

2.3 预测模型

本文利用多元线性回归模型对电视剧搜索特征集合 X 进行预测并比较预测效果,找出与 $ACTR$ 最相关的特征 X^b,得到预测结果,电视剧集均点播量预测模型如式(1)所示:

$$ACTR_k = \sum_{i=1}^{p} \beta_i^* X_{ki} + \varepsilon_k, p = 1, 2, \cdots, L \tag{1}$$

其中,ε_k 代表随机误差,用来衡量其他不可观测的因素对 $ACTR_k$ 的影响;L 为输入的特征集合 X 中每个特征的特征数。

在式(1)中,由于 β_i 对公式的求解没有影响,因此令:

$$\boldsymbol{B} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2}$$

将式(1)化简为:

$$ACTR = X\boldsymbol{B} + \boldsymbol{\varepsilon} \tag{3}$$

在多元线性回归模型中一般用最小二乘估计法估计参数向量 $\boldsymbol{B}$,求得 $\boldsymbol{B}$ 的估计值 $\hat{\boldsymbol{B}}$:

$$\hat{\boldsymbol{B}} = (X^t \cdot X)^{-1} \cdot X^t \cdot ACTR \tag{4}$$

得到电视剧集均点播量预测值:

$$\widehat{ACTR} = \hat{\boldsymbol{B}} \cdot X \tag{5}$$

由式(5)可以得到电视剧集均点播量的预测排名:

$$\widehat{RANK}_k = RANK(\widehat{ACTR}) = RANK(\sum_{i=1}^{p} \hat{\beta}^* X_{ki}) \tag{6}$$

3 实验数据准备

在确定预测方法后,本文对实验数据进行采集和预处理。

1)根据“电视剧台综合指数榜”电视剧列表,本文获取了 153 部在 2014 年 10 月 1 日—2015 年 6 月 15 日之间首播并在 PPTV 上线的电视剧以及 54 部在 2015 年 7 月 1 日—2015 年 12 月 31 日之间首播并在优酷上线的电视剧,分别获取剧名、主要演员、首播日期以及集数。

2)根据每部电视剧的首播日期以及集数,在 PPTV 历史点播数据中统计电视剧首播日后 30 天的集均点播量;在“中国网络视频指数”中获取了优酷电视剧 30 天的集均点播量。

3)以剧名、主要演员作为关键词,在百度指数平台中获取电视剧首播前 4 周每天的百度搜索指数。

4 实验与结果分析

4.1 评价指标

为衡量本文预测方法是否有效,本文采用皮尔森相关系数(r)。r 是一种线性相关系数,用来描述电视剧点播量预测值和真实值线性相关强弱的程度。当 $r>0$ 且接近 1 时,表明点播量预测值和真实值正相关,预测准确率高;当 $r<0$ 且接近 -1 时,则相反。计算公式具体如下:

$$r = \frac{\sum_{k}^{n} (\widehat{ACTR}_k - \overline{\widehat{ACTR}_k})(ACTR_k - \overline{\widehat{ACTR}_k})}{\sqrt{\sum_{k}^{n} (\widehat{ACTR}_k - \overline{\widehat{ACTR}_k})^2 \sum_{k}^{n} (ACTR_k - \overline{\widehat{ACTR}_k})^2}} \tag{7}$$

4.2 时间序列分析

本文对 153 部在 PPTV 上线的电视剧进行划分,其中 143 部电视剧作为训练分析数据;对 54 部在优酷上线的电视剧进行划分,其中 45 部电视剧作为训练分析数据。

4.2.1 剧名百度搜索指数的时间序列趋势分析

为分析首播前电视剧剧名百度搜索指数与在不同视频网站上线是否相关,分别对在 PPTV 以及优酷上线的电视剧进行以单天为时间粒度的时间序列趋势分析。所有在 PPTV 和优酷上线的电视剧首播

前 4 周($pd_1 \sim pd_{28}$)剧名平均百度搜索指数的时间序列趋势如图 2 所示。从图 2 中可以看出在 PPTV 和优酷上线的电视剧均有以下现象:随着时间点逐渐接近首播日,百度搜索指数逐渐增长,涨幅越来越大。这表明电视剧剧名搜索指数的时间变化趋势与电视剧上线视频网站无关,和电视剧的宣传期有关,电视剧首播前 1 周左右为宣传最为集中的阶段,因此搜索量明显增多。由此可见,在不同视频网站上线的电视剧在首播前的剧名百度搜索指数的时间序列趋势变化一致。

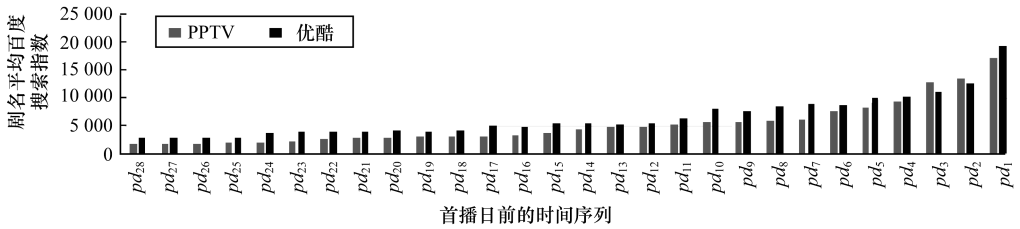

图 2　PPTV 和优酷电视剧首播前的剧名平均百度搜索指数

4.2.2　剧名百度搜索指数和点播量相关性分析

为进一步分析在不同视频网站上线的电视剧首播前剧名百度搜索指数和集均点播量在时间序列上是否相关,利用统计产品与服务解决方案(Statistical Product and Service Solutions,SPSS)中的相关分析模块,在以天为时间粒度的时间序列上,分别将 PPTV 和优酷上线的电视剧 *SDT* 作为自变量同因变量 *ACTR* 进行相关性分析,其相关系数 *R* 如图 3 所示。从图 3 中可以看出,在 PPTV 和优酷上线的电视剧剧名百度搜索指数和点播量相关系数变化趋势差别较大,两者相关系数曲线均呈现不规则波动,峰值区间不一致。对产生的原因进行分析:一方面用户对于视频网站的选择具有偏好性,在不同视频网站上线的电视剧点播量的数量级别可能不同;另一方面以单天为最小时间单位,时间粒度较小,数据变化频率较快,产生的随机波动较大。

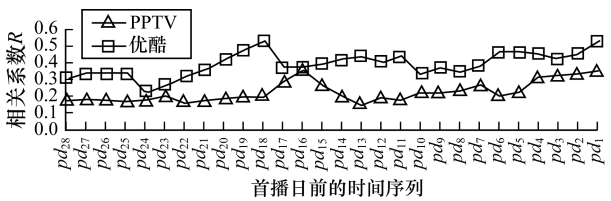

图 3　PPTV 和优酷电视剧首播前剧名百度搜索指数与 *ACTR* 的相关性

4.2.3　时间粒度分析

在时间序列分析中,其中有一个重要问题就是时间的描述和划分,常用的时间粒度是日、月和年,时间粒度对关联规则的有效性、周期长度以及序列模式都有许多影响[15-16]。在首播前剧名百度搜索指数和点播量时间序列分析中的时间粒度为单天,由于分析粒度较细,数据波动剧烈,并且在不同视频网站上线的电视剧没有序列共性,因此为了消除在不同视频网站上线的电视剧剧名百度搜索指数和点播量的相关性差异以及平滑单天为时间粒度导致的剧烈波动,需要选择适度的时间粒度,对几天或几周同时考虑。

本文将时间粒度从 2 天开始逐天递增至 10 天,计算 *SST*,分别将在 PPTV 和优酷上线的电视剧 *SST* 作为自变量同因变量 *ACTR* 进行相关性分析,得到相关系数 *R* 以及最高相关系数 max(*R*)。不同时间粒度下的 max(*R*)如图 4 所示,随着时间粒度逐渐增大,max(*R*)逐渐变小。时间粒度越小,数据变化频率越快,不可控因素越多;时间粒度越大,虽然实现了相对平衡,但是会弱化数据间的差异。为了确定时间粒度阈值,利用相邻时间粒度之间 max(*R*)的差值进行变化分析,差值结果如图 5 所示,横坐标 1 表示相邻时间粒度 2 天与 1 天之间 max(*R*)的差值,2 表示相邻时间粒度 3 天与 2 天之间 max(*R*)的差值,依次类推,从图 5 可以看出相邻时间粒度为 6 时,相邻时间粒度 7 天与 6 天之间 max(*R*)的差值趋于平缓,即时间粒度 7 天为差值变化陡峭和平缓的分界点,7 天之前下降的幅度较大,7 天之后下降的幅度较小,趋于平稳,基本不变。因此,通过差值分析,本文将时间粒度的阈值设定为 7。

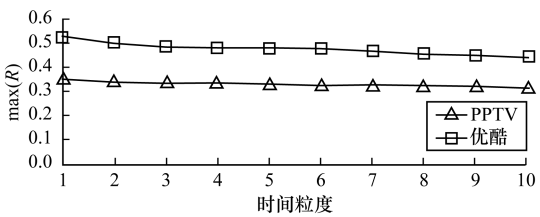

图 4　不同时间粒度的最高 *R* 值

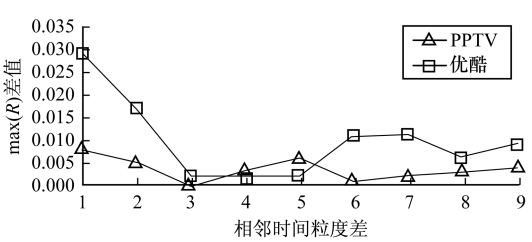

图 5　相邻时间粒度的最高 *R* 差值

4.2.4 候选时间粒度和最早预测时间分析

为确定最佳时间粒度和最早预测时间,本文通过对时间粒度阈值范围内 *SST* 与 *ACTR* 的相关系数曲线进行分析,相关系数 *R* 如图 6 所示。

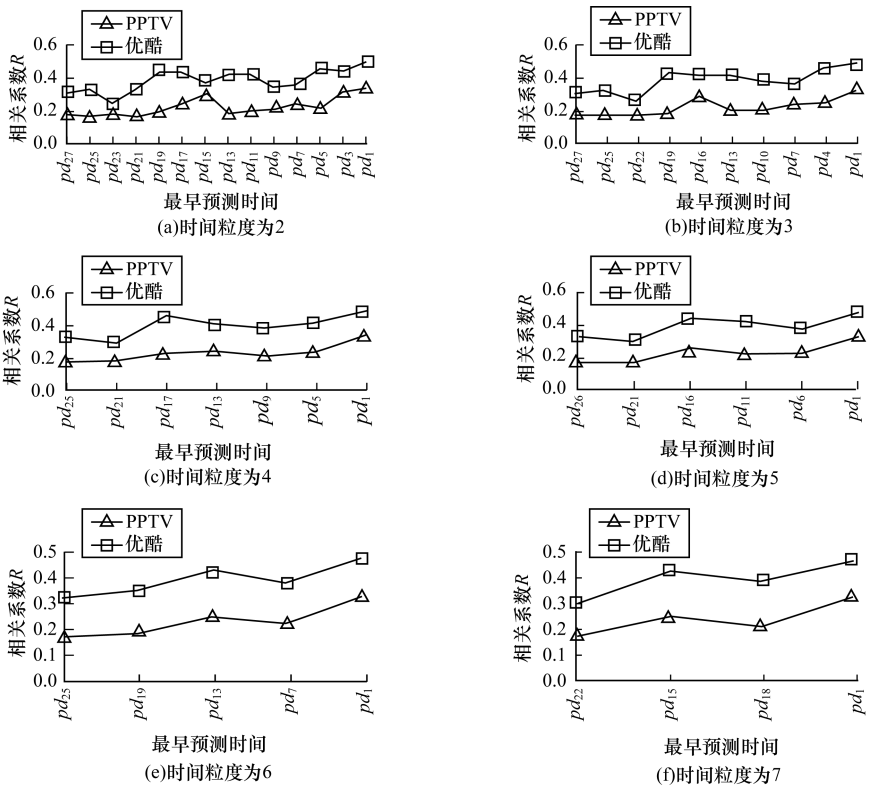

图 6 PPTV 和优酷电视剧在不同时间粒度下首播前剧名百度搜索指数与 *ACTR* 的相关性

从图 6 中可以看出随着时间粒度的增大,在 PPTV 和优酷上线的电视剧相关系数曲线逐渐呈现规则波动,变化趋势逐渐趋于一致,时间粒度为 5,6,7 时完全一致,均有 2 个明显的峰值,即最优的最早预测时间。由于粒度为 5,6,7 的相关系数 *R* 比较接近,无法确定最优时间粒度和最早预测时间,因此根据粒度 5,6,7 各自的 2 个峰值,确定候选时间粒度以及最早预测时间。

1)时间粒度 $g=5$,最早预测时间 $d=pd_1$ 以及 $d=pd_{16}$。

2)时间粒度 $g=6$,最早预测时间 $d=pd_1$ 以及 $d=pd_{13}$。

3)时间粒度 $g=7$,最早预测时间 $d=pd_1$ 以及 $d=pd_{15}$。

4.3 特征融合

根据时间序列分析得到的候选时间粒度以及最早预测时间分别进行 3 组对比实验:单个搜索特征,剧名搜索特征和演员搜索特征特征值相加以及特征数增加的特征融合方式,分别对 10 部在 PPTV 以及 9 部在优酷上线的电视剧进行测试,确定最优特征并预测点播量排名。

1)时间粒度 $g=5$,最早预测时间 $d=pd_1$ 以及 $d=pd_{16}$。预测结果评价指标如表 2 所示。当最早预测时间为 pd_1 时,在 PPTV 上线的电视剧数据集最优特征为 *AI*,*r* 最高为 0.953 4,在优酷上线的电视剧数据集最优特征为{*TI*,*AI*},*r* 最高为 0.954 4;最早预测时间为 pd_{16} 时,两者的最优特征均为 *TI* + *AI*,*r* 分别是 0.936 8 和 0.947 9。

表 2 时间粒度 $g=5$ 时的预测结果

实验组	特征	最早预测时间 pd_1		最早预测时间 pd_{16}	
		PPTV	优酷	PPTV	优酷
单个特征	*TI*	0.897 7	0.888 0	0.890 3	0.164 5
单个特征	*AI*	**0.953 4**	0.909 9	0.805 8	0.872 3
特征值相加	*TI* + *AI*	0.917 2	0.943 4	**0.936 8**	**0.947 9**
特征数增加	{*TI*,*AI*}	0.928 4	**0.953 9**	0.914 5	0.878 0

2)时间粒度 $g=6$,最早预测时间 $d=pd_1$ 以及 $d=pd_{13}$。预测结果评价指标如表 3 所示。当最早预测时间为 pd_1 时,在 PPTV 上线的电视剧数据集最优特征为 *AI*,*r* 最高为 0.962 3,在优酷上线的电视剧数据集最优特征为{*TI*,*AI*},*r* 最高为 0.952 3;当最早预测时间为 pd_{13} 时,两者的最优特征均为 *TI* + *AI*,*r* 分别是 0.943 7 和 0.967 6。

表 3　时间粒度 $g=6$ 时的预测结果

实验组	特征	最早预测时间 pd_1		最早预测时间 pd_{13}	
		PPTV	优酷	PPTV	优酷
单个特征	TI	0.899 1	0.859 7	0.903 7	0.156 7
单个特征	AI	**0.962 3**	0.912 0	0.787 7	0.897 8
特征值相加	$TI+AI$	0.919 2	0.944 6	**0.943 7**	**0.967 6**
特征数增加	$\{TI,AI\}$	0.928 9	**0.952 3**	0.934 7	0.897 2

3)时间粒度 $g=7$,最早预测时间 $d=pd_1$ 以及 $d=pd_{15}$。预测结果评价指标如表 4 所示。当最早预测时间为 pd_1 时,在 PPTV 上线的电视剧数据集最优特征为 AI,r 值最高为 0.976 5,在优酷上线的电视剧数据集最优特征为 $\{TI,AI\}$,r 最高为 0.951 3;最早预测时间为 pd_{15} 时,两者的最优特征均为 $TI+AI$,r 分别是 0.941 6 和 0.950 8。

表 4　时间粒度 $g=7$ 时的预测结果

实验组	特征	最早预测时间 pd_1		最早预测时间 pd_{15}	
		PPTV	优酷	PPTV	优酷
单个特征	TI	0.900 0	0.839 1	0.896 0	0.171 0
单个特征	AI	**0.976 5**	0.912 1	0.804 9	0.874 6
特征值相加	$TI+AI$	0.919 8	0.944 5	**0.941 6**	**0.950 8**
特征数增加	$\{TI,AI\}$	0.925 6	**0.951 3**	0.920 4	0.899 3

4.4　结果对比分析

实验结果分析如下:

1)不同视频网站对电视剧点播量预测的影响。由于不同视频网站用户不同,偏好不同。不论时间粒度是多少,最早预测时间为 pd_1 时,PPTV 和优酷最优特征没有共性,而最早预测时间为 pd_{13},pd_{15} 以及 pd_{16} 时,PPTV 和优酷的最优特征均为 $TI+AI$。

2)不同特征融合方式对电视剧点播量预测的影响。根据最早预测时间为 pd_{13},pd_{15} 以及 pd_{16} 实验组的结果,在 TI,AI 单个特征中 r 最高的均为 TI;在特征值相加的特征融合方式中 r 均高于特征数增加的特征融合方式,其中 r 最高为最早预测时间为 pd_{13} 的实验组,PPTV 和优酷 r 分别为 0.943 7 和 0.967 6。

实验结果表明,剧名搜索特征和演员搜索特征对电视剧点播量预测均相关,并且在剧名搜索特征的基础上融合演员搜索特征预测效果更好。最优特征的时间粒度为 6,最早预测时间为 pd_{13} 实验组中的 $TI+AI$,即电视剧首播前第 13—18 天剧名和演员的百度搜索指数总和。

4.5　点播量排名预测

将最优特征作为多元回归模型参数对电视剧点播量进行预测,其中 10 部在 PPTV 上线的电视剧点播量及排名预测结果如表 5 所示,9 部在优酷上线电视剧点排名播量及排名预测结果如表 6 所示。对于排名 2和排名 3 的电视剧虽然没有预测出正确的顺序,但是排名前 5 的电视剧最终预测出的结果依然都排在前 5。对预测排名较低或者预测错误的电视剧进行原因分析,这些电视剧均是军旅题材的电视剧,一方面这类题材的电视剧受众面较窄,而在视频网站进行点播的观众较多都是年轻的观众,年轻观众对军旅题材的电视剧关注度较低;另一方面,这些电视剧主要演员知名度均较低。

表 5　PPTV 电视剧排名预测结果

电视剧名称	集均点播量真实值	点播量排名真实值	集均点播量预测值	点播量排名预测值
花千骨	196 287	1	167 592	1
两生花	103 278	2	49 748	3
你是我的姐妹	78 859	3	81 759	2
偏偏喜欢你	74 296	4	45 682	4
后海不是海	28 014	5	24 627	5
怒放	20 735	6	19 074	10
劫中劫	19 991	7	24 508	6
香火	8 801	8	20 858	7
剧场	5 273	9	19 590	9
金玉瑶	4 243	10	20 378	8

表 6　优酷电视剧排名预测结果 1

电视剧名称	集均点播量真实值	点播量排名真实值	集均点播量预测值	点播量排名预测值
长在面包树上的女人	10 034 356	1	14 926 264	1
嫂子嫂子	3 936 170	2	8 096 018	3
老婆大人是 80 后	3 620 282	3	8 246 194	2
爱情碟中谍	1 920 613	4	6 613 800	4
地雷英雄传	1 723 960	5	5 280 255	5
我是赵传奇	1 697 028	6	3 337 934	8
继父回家	1 067 758	7	4 804 248	6
鄂尔多斯风暴	299 704	8	2 490 839	9
爸爸是条龙	200 322	9	4 133 994	7

为进一步验证最优特征的预测效果,随机获取 7 部2016 年 1 月—2016 年 2 月期间首播并在优酷上线的电视剧,获取其上线后 30 天的集均点播量及首播前第 13—18 天的剧名和演员的百度搜索指数和,预测准确率为 0.956 3,具体结果如表 7 所示。可以发现,排名在前的《女医明妃传》《远得要命的爱情》《搭错车》等热播电视剧都得到了准确的预测。虽然本文提出的预测方法对排名较后的电视剧未给出正

确的排名,但是对于热播电视剧的预测具有一定的参考价值。

表 7 优酷电视剧排名预测结果 2

电视剧名称	集均点播量真实值	点播量排名真实值	集均点播量预测值	点播量排名预测值
女医明妃传	10 328 950	1	14 926 264	1
远得要命的爱情	5 262 084	2	5 703 311	2
搭错车	2 991 783	3	3 620 313	3
生死翻盘	1 397 925	4	2 546 906	6
因为爱	1 194 849	5	2 818 486	5
天伦	936 376	6	2 490 839	7
新萧十一郎	292 124	7	3 475 575	4

5 结束语

本文提出的电视剧流行度预测方法主要利用首播前的电视剧剧名百度搜索指数确定可以预测的最早开始时间和时间粒度,根据最早预测时间和时间粒度分别得到剧名和演员搜索特征,融合剧名和演员搜索特征并利用多元线性回归模型以及预测视频网站中的电视剧点播量的排名来预测电视剧流行度,得出以下结论:1)在研究电视剧搜索特征对预测的影响时,剧名特征和演员特征与预测显著相关,并且融合 2 个搜索特征的效果优于单个搜索特征。2)特征值相加的特征融合方式对预测的提升效果明显。3)完全使用电视剧首播前的数据,可以在电视剧首播前第 12 天预测出热播电视剧,避免广告投放等决策的滞后性。实验结果表明,对在 PPTV 和优酷上线的电视剧点播量进行预测,真实值和预测值之间的相关系数分别是 0.943 7 和 0.967 6,并且对热播电视剧预测的效果较好。因此,本文提出的方法能够辅助视频网站的决策。

虽然本文能够对电视剧流行度进行预测,但是除了排名靠前的电视剧预测效果较好之外,排名靠后的准确度还有待提高。另外,除了剧名、演员等电视剧搜索数据,未考虑其他电视剧相关搜索信息。因此,下一步将考虑更多的电视剧信息,如预告片的播放量、题材、档期等,利用新浪微博等社交数据进一步提高预测准确度。

参考文献

[1] 西安电视剧版权交易中心.2014 年电视剧网络新媒体播出平台分析[EB/OL].(2015-05-29).https://sanwen8.cn/p/ebaZZN.html.

[2] 吴玉玲,高 铭.电视剧版权交易评估指标体系的建构[J].当代传播,2014(2):105-107.

[3] 洪皓轶.电视剧收视率预估的市场化操作模式构建探析[J].电视研究,2013(2):71-73.

[4] 欧阳柏成.大数据时代的数据挖掘技术探究[J].电脑知识与技术,2015,11(15):3-4.

[5] Qiu Tongqing,Ge Zihui,Lee S,et al.Modeling Channel Popularity Ynamics in a Large IPTV System[J].ACM SIGMETRICS Performance Evaluation Review,2009,37(1):275-286.

[6] Chen H,Hu Q,He L.Clairvoyant:An Early Prediction System For Video Hits[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management.New York,USA:ACM Press,2014:2054-2056.

[7] Asur S,Huberman B A.Predicting the Future with Social Media[C]//Proceedings of Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.Washington D.C.,USA:IEEE Press,2010:492-499.

[8] Sadikov E,Parameswaran A G,Venetis P.Blogs as Predictors of Movie Success[C]//Proceedings of International Conference on Weblogs & Social Media.Washington D.C.,USA:IEEE Press,2009:304-308.

[9] Mishne G,Glance N.Predicting Movie Sales from Blogger Sentiment[C]//Proceedings of Computational Approaches to Analyzing Weblogs Conference.Menlo Park,USA:AAAI Press,2006:301-304.

[10] Panaligan R.Quantifying Movie Magic with Google Search[EB/OL].(2013-05-18).http://www.tuicool.com/articles/mei2Qf.

[11] 徐晓枫,贺 樑,杨 静.融合社交与搜索数据的电视剧点播排名预测研究[J].计算机工程,2015,41(8):6-12,17.

[12] 戈小燕.电视剧观众的收视心理探析[J].视听纵横,2012(4):101-103.

[13] 倪祥保,杨 秋.略论剧本与角色、演员及观众——以电视剧《人间正道是沧桑》为例[J].中国电视,2010(10):12-15.

[14] 百度.百度指数[EB/OL].[2016-12-18].http://index.baidu.com.

[15] Bettini C,Wang X,Jajodia S,et al.Discovering Frequent Event Patterns with Multiple Granularities in Time Sequences[J].IEEE Transactions on Knowledge & Data Engineering,1998,10(2):222-237.

[16] 唐常杰,于中华,游志胜,等.基于时态数据库的 Web 数据周期规律的采掘[J].计算机学报,2000,23(1):52-59.

编辑 陆燕菲