多摄像头系统的时间同步

（申请清华大学工学硕士学位论文）

培 养 单 位：计 算 机 科 学 与 技 术 系
学　　科：计 算 机 科 学 与 技 术
研　究　生：丁　旭
指 导 教 师：陶　品　副　教　授

二〇一七年四月

the Synchronization of Multicamera System

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Engineering

by

Ding Xu

(Computer Science and Technology)

Thesis Supervisor : Associate Professor Tao Pin

April, 2017

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

随着摄像头技术的日益发展，为了获得更高性能、实现更多功能，利用多个摄像头构建多摄像头系统成为了一个重要的发展方向，关于多摄像头系统的研究逐渐深入，应用范围也逐渐广泛起来。为了满足系统应用的各项需求，需要精确控制多摄像头系统内的各个摄像头在特点时间进行拍摄，实现拍摄时间的同步。

为了实现多摄像头系统的拍摄时间同步，本文提出了一种利用 LED 点阵进行拍摄时间检测，根据时间差调整拍摄时间的同步方法。该方法主要分为摄像头拍摄时间检测和多摄像头系统时间同步控制两部分。

在进行摄像头拍摄时间检测时，主要利用 FPGA 和 LED 点阵组成检测系统，由 FPGA 按照一定的编码规律控制点阵中各个 LED 灯不断变化，组成不同的 LED 点阵状态形成状态序列。根据不同的编码方法，可以判断各个点阵状态在序列中所处的位置。当利用摄像头对 LED 点阵进行拍摄时，根据拍摄到的 LED 点阵的状态，可以确定该状态在状态序列中所处的位置，根据每个状态的持续时长，即可检测摄像头的拍摄时间。对于不同的编码方法，能够得到不同的检测精度，对摄像头也有不同的性能、参数要求，本文共设计了五种编码方法，并通过实验对其检测效果进行了比较。利用该方法对摄像头的拍摄时间进行检测，检测系统无需与摄像头进行数据通信，对摄像头没有特殊的硬件接口要求，能够适用于大多数的摄像头，适用范围较广。同时，该方法能够在摄像头正常工作的过程当中进行检测，无需暂停摄像或对摄像头进行调整。

在获得摄像头的拍摄时间之后，就可以对多摄像头系统进行同步调整。首先将各个摄像头拍摄到的 LED 点阵图像，由服务器统一进行图像识别，检测摄像头的拍摄时间，计算各个摄像头之间的时间间隔，选定基准摄像头，通过暂停摄像或者调整帧率的方法改变摄像头的拍摄时间，从而控制系统内各个摄像头能够在同一时刻进行拍摄，并同步多次迭代验证不断提高同步精度，实现系统同步。该同步方法基于拍摄时间检测方法实现，能够获得较高的同步精度。同时，由服务器进行图像处理和同步控制，对于摄像头的计算性能要求较低，通过参数调整能够实现不同的同步精度，方法灵活性较高。

本文还利用树莓派电脑控制摄像头进行拍摄，图像处理服务器进行系统控制，搭建了一个具有可扩展性的多摄像头系统，对上述检测和同步方法进行了验证。

关键词：多摄像头系统；时间同步；拍摄时间检测；LED 点阵

Abstract

With the development of camera technology, in order to achieve higher performance and more functions, multi-camera system has become an important research direction. There are more and more researches on the multi-camera system. In order to meet the needs of the system application, the system must achieve the shooting time synchronization, to accurately control cameras to shoot at a certain moment.

In order to realize the time synchronization of multi-camera system, this paper proposes a high precision method to detect the synchronization accuracy by using a FPGA based LED matrix as a detector. The method is divided into two parts: camera shooting time detection and multi-camera system time synchronization control.

This paper use a detection system composed by FPGA and LED dot matrix to detect the camera shooting time. The LED lights in the LED matrix continue to change forming a state sequence. Depending on the encoding method, it is possible to determine the position of each matrix state in the sequence. When shooting the LED dot matrix using the camera, the position of the state in the state sequence can be determined, and the shooting time of the camera can be detected. For different encoding methods, different detection accuracy can be obtained, the camera also has different performance and parameter requirements. In this paper, five encoding methods are designed, and the results are compared according to experiments. With this method, the detection system needs no data communication with the camera, so the camera does not have to equip special hardware interface. This method can be applied to most of the camera. At the same time, the method can be applied during the normal working process of the camera, without pausing or adjusting the camera.

After the camera's shooting time is detected, the multi-camera system can be adjusted synchronously. First of all, the server identifies images of the LED dot matrices captured by cameras. So the server can get cameras' shooting time and calculate the time intervals between the various cameras. According the reference camera, the server then can change the cameras' shooting time by pausing cameras or adjusting the frame rate. So that the cameras in the system can shoot at the same time. And this method can continuously improve the synchronization accuracy by multiple iterative verification. This synchronization method can achieve a high synchronization accuracy for the use of the shooting time detection method. At the same time, the cameras' computing performance

requirements are lower and the method flexibility is higher.

This paper also builds a scalable multi-camera system, using the Raspberry to to control the camera and the image processing server to control the system. With this system, the above detection and synchronization method has been verified.

Key words: multicamera system; synchronization; shooting time detection; LED matrix

目 录

第 1 章 引言 1
1.1 研究背景 1
1.2 主要研究内容和挑战 5
1.3 主要贡献和组织结构 5
第 2 章 研究现状和相关工作 6
2.1 本章引言 6
2.2 影视剧与社交网络 6
2.3 因果分析法 8
2.4 基于倾向值匹配的因果分析 11
2.5 本章小结 13
第 3 章 数据库与测量分析 15
3.1 引言 15
3.2 数据集 15
3.2.1 数据库 15
3.2.2 特征提取 18
3.3 测量分析 19
3.3.1 微博影响力与话题热度 19
3.3.2 推广模式 21
第 4 章 演员社交推广行为的影响力模型 25
4.1 引言 25
4.2 倾向值匹配算法 25
4.2.1 混淆变量 25
4.2.2 推广策略 25
4.2.3 算法步骤 26
4.3 结果分析 27
4.4 模型显著性及平衡性检验 29
第 5 章 基于话题演化的演员社交推广行为影响力模型 30
5.1 引言 30
5.2 电视剧话题演化规律 30

5.3 倾向值匹配算法 31
5.3.1 混淆变量及推广策略 31
5.4 结果分析 32
5.5 模型显著性及平衡性检验 33
第 6 章 总结与展望 35
6.1 工作总结 35
6.2 未来展望 36
参考文献 37
致 谢 41
声 明 41
个人简历、在学期间发表的学术论文与研究成果 43

第 1 章　引言

1.1　研究背景

随着互联网技术的不断发展，社交网络已经成为人们日常生活中不可或缺的一部分，人们可以通过社交网络获取新闻，与他人进行交流互动，发布个人信息等等。对于大多数中国互联网用户来说，QQ、微博、微信、陌陌等社交应用是其日常上网的主要使用应用。根据中国互联网络信息中心（CNNIC）的调查结果显示[1]，每日的上网时长在两个小时以上的用户达到 79.5%，而其中 77.0% 的时间是用在社交应用上。社交网络使得人们不再受报刊、电视等信息来源的限制，拥有了更多获取信息的渠道。

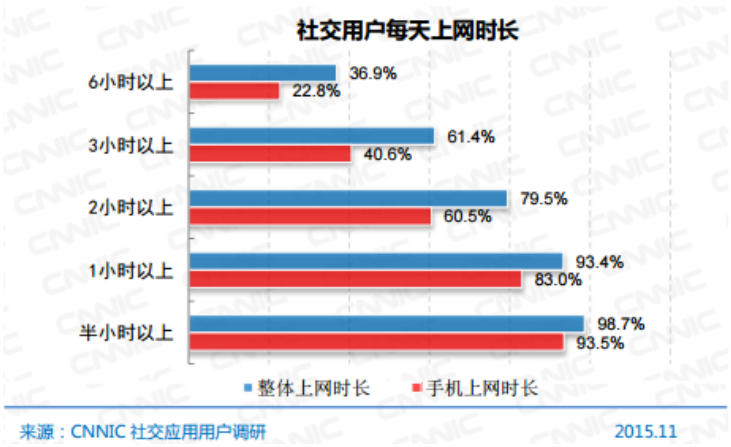


图 1.1　社交用户每日上网时长

对于信息发布者来说，社交网络给他们提供了一个新的推广和营销平台。由于大量用户花费大量时间在社交网络中，传统信息发布渠道日渐衰落，而社交网络凭借其传播范围广、扩散速度快、受众数量大、推广形式多等特点，逐渐成为一种全新的、重要的、高效的推广方式。Shama Hyder 在《网络社交媒体营销》一书中提到社交媒体必将取代传统媒介成为营销推广的主要手段[2]。

利用社交网络进行推广营销，能够在短时间内将信息传递给目标群体。在微博、微信公众号等社交平台上，流行用户往往拥有几百万甚至几千万的关注量，其发布的每一条消息都有大量的用户会关注到。且传播速度快，实时发送接收，还能够根据受众的不同发布不同的信息，使得推广行为更加具有针对性。同时，利用社交网络进行推广，形式上更加灵活多样，文字、图片、语言、视频，能够更加充分进行展示。优秀的推广内容甚至能够通过用户自发进行传播，呈现爆发式营销

增长。文献[3],[4],[5],[6],[7] 均对基于社交网络的营销行为进行了深入分析。著名社交网络公司 Facebook，在 2016 年广告业务的总收入达到 268 个亿美元[8]。而微博在 2016 年 Q4 季度中，其广告收入也达到了 1.879 亿美元[9]。而这仅仅是这两家社交网络公司自身的广告营销收入，其最大的贡献是为其用户提供了一个社交网络的营销平台，由此产生的推广营销效益更加巨大。

由于社交网络推广方式的灵活性，包括商品、活动、人物等众多对象均可以进行推广，而本文主要研究的是影视剧在社交网络上的推广。由于影视剧的关注人群广泛，且娱乐性、话题性强，能够更好地适用于社交网络上的多种推广模式。在文献[10],[11],[12],[13] 中均介绍了基于社交网络数据对影视剧进行推广的研究方法。

随着大数据、机器学习等技术的不断进步，社交网络在影视剧推广方面的重要作用越来越不容忽视。以著名的国产电影《失恋 33 天》的社交网络营销为例，该部电影在 2011 年上映，凭借成果的网络营销手段，取得的 3.5 亿人民币的优秀票房，是愿预估票房的十倍以上[14]。在 2011 年的 11 月 16 日，通过百度搜索引擎搜索“失恋 33 天”关键字，能够找到相关结果约 394 万个，在谷歌搜索“失恋 33 天”，能够找到约 5900 万个相关结果。而在社交网站微博上搜索“失恋 33 天”能够找到约 670 万条消息，而在腾讯微博上，则能够搜索到 330 万条消息[14]。

图 1.2　《失恋 33 天》百度搜索结果

这部电影以新浪微博、腾讯微博等社交网站为重要推广平台，打造了拥有近 10 万粉丝量的官方微博，并创建了众多热门话题，通过宣传短片、制造话题等众多手段，针对网络用户的特点精准投放营销策略，并引导用户参与话题讨论，通过转发、评论等手段，将用户自身的社交网络融入到宣传网络当中，进一步扩大了宣传效果。在日后的各类总结点评中，《失恋 33 天》都被认为是电影营销手段的一次最成功的创新，利用社交网络对影视剧进行推广的方法从此日益发展。

通过对现有社交网络推广案例进行分析即可发现，目前针对影视剧的网络营销手段层出不穷，而且未来还存在着继续增多的趋势。例如宣传方会在微博、QQ、微信等社交媒体上创建官方账号，发布关于影片的文字、图片消息，演员消息，宣传短片等。这样的宣传方式会在影片上映前后持续较长时间，吸引大量粉丝关注，提高影片关注度。例如最近上映的电影《嫌疑人 X 的献身》，其在微博上的官方账号共发布了 744 条微博，拥有 62 万粉丝关注，发布的每条微博有数千条的转发、评论、点赞[①]。

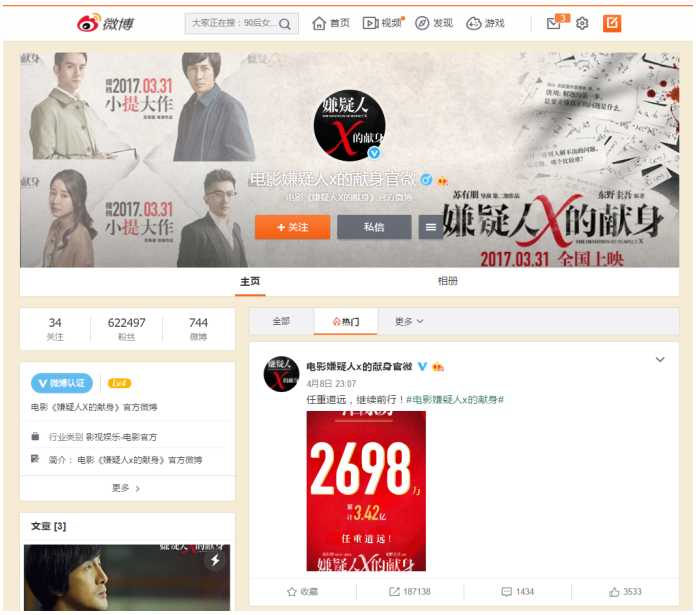


图 1.3　《嫌疑人 X 的献身》官微截图

同时，宣传方还会在社交网络上制造影视剧相关话题，并不断推高话题热度，扩大影片影响力。例如电影《致青春》在网络营销中通过推出“回忆青春”、推出主题曲、制造“有一种感情叫赵薇黄晓明”等流行语等形式制造多个热点话题，激发用户自动传播，最终实现病毒式传播效果。另外宣传方还可以通过各类影评、新闻报道提升观众的观看热情。例如最近热播的电视剧《人民的名义》在豆瓣网可以找到近 4 万篇影评[②]，短短一种之内在微信公众号内即出现了 55 篇与其有关的阅读量超过 10 万的文章，这样的推广方式能够为影视剧打造良好的口碑，塑造良好形象，吸引更多观众观看。

而网络推广的各种营销手段，往往是通过一些关键节点传播和扩散给广大用户，这些节点一般是由影视剧的官方账号、演员或者经过社交平台验证的“大 V”等用户组成。这些用户拥有巨大的粉丝量，能够将各类宣传消息推送给众多的用

① http://weibo.com/p/1002065746403567
② https://movie.douban.com/subject/26727273/

户，起到宣传媒体的作用[15]。

影视剧演员往往拥有众多粉丝，关于演员的新闻报道和演员自身发布的消息都能够获得极高的关注度，因此演员在社交网络中也就具有着较高的影响力[16]，是影视剧宣传推广的重要节点。影视剧演员可以通过自身的影响力在社交网络上制造话题，通过发布各种类型的消息，包括文字、图片、音视频等等，或者利用其它用户发布关于自己的新闻报道，即可以引起关注自己的粉丝的广泛反响，引发相关的热烈讨论。然后根据网络话题演化的规律[17]，在适当时刻不断推高话题热度，使其成为热点，获得更多受众。因此，在对影视剧进行宣传的过程中，演员就可以利用其巨大的影响力，制造并推动关于影视剧的相关话题，引起其在社交网络中的广泛关注，对影视节目的宣传推广起到重要的促进作用。

但是在进行推广时，不同的推广方式会受到不同的推广效果，同样的推广方式不同的用户使用也会得到不同的结果。如图 1.4所示，演员邓超在对其主演的电影《乘风破浪》进行宣传时，发布了两条微博，但是其转发和点赞数量却有很大的不同。可想而知，这两条微博能够获得的推广效果也是截然不同的。

(a) 转发:33441 评论:24750 点赞:209791　　(b) 转发:9327 评论:21477 点赞:560685

图 1.4　不同微博的推广效果比较

另外，对于不同的推广对象也需要采取不同的推广策略。例如，对于电影的推广来说，由于其属于一次性消费，只需要吸引观众进入影院观看即可。而对于电视剧的推广来说，由于其播放周期较长，需要对用户存在长时间的粘滞力，在吸引新用户进行观看的同时，还有保持老用户的持续关注。因此，对于电影和电视剧的推广手段就需要有不同的侧重点，需要有针对性地采用不同的推广方式。目前关于电影推广的研究较为常见，由于电影的高投入高产出性，宣传方也会投入更多精力关注电影推广。而关于电视剧有针对性的推广研究较少，推广方式也较为有限，需要给予更多的关注。

因此，为了能够获得更好的推广效果，最大限度地发挥社交网络的推广作用，需要更加具有针对性地为宣传者设计适合其自身特点的推广方式，本文选取演员在社交网络上的对于电视剧的推广行为作为研究对象，针对不同的推广模式进行

比较，根据演员自身特点，拍摄混淆变量的干扰，选取最优的推广模式，为演员获得最佳的推广效果提供帮助。

1.2　主要研究内容和挑战

本文所述研究主要存在如下几方面挑战：

a)

1.3　主要贡献和组织结构

针对以上提出的问题与挑战，本文设计了一种摄像头拍摄时间的检测方法和对多摄像头系统的同步方法，本文的主要贡献如下：

a)

本文的组织结构如下：

第二章主要介绍

第六章是对全文的总结，以及对下一步工作的展望。

第 2 章　研究现状和相关工作

2.1　本章引言

本章首先介绍了影视剧产业在社交网络上的发展与应用，包括利用社交网络数据对影视剧的票房进行预测，针对观众喜欢为其进行影视剧推荐。由于在社交网络上对影视剧进行宣传推广时，需要在众多的推广方式当中选取推广效果最好的方式，因此需要分析各种推广方式与推广结果之间的因果关系，本章主要对工具变量法、断点回归设计、格兰杰因果关系检验这三种因果分析方法进行了介绍，简要介绍了其分析原理以及现阶段的应用情况。最后针对演员在微博上推广电视剧的行为特点，本章选取倾向值匹配算法作为因果分析方法，并对其进行了介绍。

2.2　影视剧与社交网络

影视剧属于传统产业，诞生与发展由来已久，而社交网络属于新兴事物，近几年才蓬勃发展起来。但是随着社交网络的不断进步，影视剧与社交网络的结合越来越紧密，人们发现利用社交网络对影视剧的制作、宣传都能够起到积极作用。目前在社交网络领域关于影视剧的研究主要集中在两个方面，一是利用社交网络数据对影视剧进行预测，二是针对影视剧的推荐。

利用社交网络数据对影视剧进行预测，主要集中在大数据和机器学习领域。由于网络数据量的日益增大，通过大数据分析其内在关系，利用机器学习找出其中的统计学关系，更好地指导、预测影视剧的发展。其中最典型的应用之一即为票房预测，通过挖掘社交网络当中影视剧的相关信息，来预测未来影视剧票房收视率的发展趋势。例如王伟在文献[18] 中利用多元线性回归模型、BP 神经网络模型、支持向量机模型等预测模型，对微博上关于电影的微博数量、情感、营销等特征进行分析，预测电影票房情况。王晓耘等人在文献[19] 中对关于影视剧微博的情感倾向进行了分析，通过 SVM 算法对其进行分类，识别出观众观看意愿的强烈程度，再利用 BP 神经网络模型进行票房预测，能够准确地预测出电影上映首周的票房情况。王烁等人在文献[20] 中分析了网络搜索量与电影票房之间存在的联系，由于大多数观众习惯在观影之前通过网络搜索相关评论来确定是否要购票观看，因此，网络搜索量与票房之间也就存在一定的正相关关系，通过最小二乘法等方法可以通过搜索量的变化预测电影票房的变化。

同样，国外学者在票房预测领域也有着深入的研究。Sharda 等人在文献[21] 中

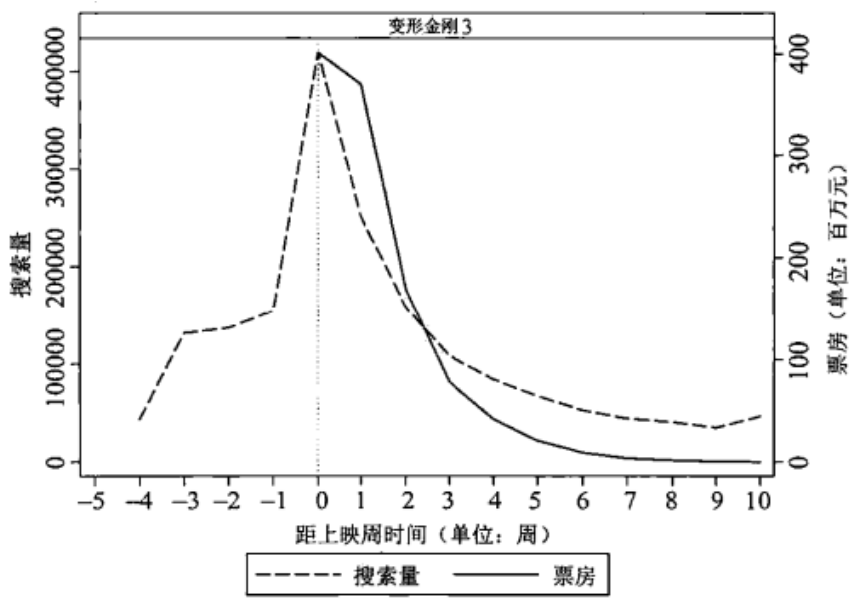


图 2.1　网络搜索量与票房之间的关系

将影视剧的票房预测问题转化为了票房收益的分类问题，避免预测具体收益值而是转为预测收益区间，这样的预测结果更加准确合理，能够为各层次用户提供参考依据。作者通过神经网络模型进行预测，并与其它算法进行了 10 倍交叉验证比较，证明其取得了更好的预测效果。

Asur 等人在文献[22] 中利用社交网站 Twitter 上关于电影的评论来预测票房。作者通过收集评论信息并对其进行情感分类，通过分析积极和消极评论对观众的观影需求的影响评判票房走势。由于评价数据直接来自观众，能够充分反映出观众对电影的真实感受，因此通过此方法的预测更加准确。

Sawhney 等人在文献[23] 中提出了一种简单模型，基于早期的票房数据来预测未来电影票房的总体走势。作者用一个排队理论框架，将消费者的观影过程分为两个步骤，即决定观影，以及采取行动观看，利用 BOXMOD-I 模型和前三周的票房数据即可预测未来票房情况。

第二类影视剧与社交网络的结合是利用社交网络数据向用户进行影视剧推荐，基于内容、协同过滤、基于规则等推荐算法种类较多，发展也较为成熟，将其应用到影视剧领域，能够根据用户和影片的特点，有针对性地进行推荐，获得更好的观影体验和票房收益。

Golbeck 等人在文献[13] 中设计了 FilmTrust 网站，将基于语义的社交网络进行整合，提高了电影预测检验的可信度。该推荐方法使用信任评级作为计算相似度的依据，利用信任网络推理算法 TidalTrust 作为用户个性化的预测评估的基础。

Bogers 等人在文献[24] 中提出了一种 ContextWalk 推荐算法，该算法包含了不

同类型的上下文信息，避免了单一使用影视剧评级信息，通过在上下文图上进行随机游走来模拟电影数据库网站上用户的浏览过程，并通过自我转换进行调整，以产生用户看不见的电影的概率分布，从而将更多的上下文合并到推荐过程中，提高推荐准确性。

Mukherjee 等人在文献[25] 中开发了一个电影推荐系统，使得用户可以通过该系统实现无约束，受限或基于实例的电影查询。该系统通过主动的信息收集获取用户的推荐反馈来学习用户模型，使用户交获得更好的交互效果。

2.3　因果分析法

无论是对影视剧的预测还是推广，都充分利用的社交网络大数据提供的海量信息，从中挖掘出相互关联，从而获得更大的收益。对于社交网络这个巨大的平台来说，利用其对影视剧进行宣传推广，也是一项重要的应用。但是对于众多的推广策略来说，宣传方需要知道哪种策略能够获得最好的推广效果，而且这就策略并非是千篇一律具有普适性的，需要根据宣传对象、宣传内容、受众群体等因素的差异具有更高的针对性。因此，就需要对于已知的各种推广策略，根据其在社交网络上现有的推广效果挖掘其中的因果关系，分析每一项策略对推广效果的独立影响，才能拍出其他因素对结果的影响，使得分析更加准确。

因果分析的目的是为了分析变量与结果之间存在的联系，且需要排除其他混淆变量对结果造成的影响。目前关于因果分析的应用多集中于社会科学领域，大多数分析过程采用定性分析的方法，还有少部分在小数据量的数据集上采用定性分析。在文献[26],[27],[28] 中，分别利用因果分析的方法对群体决策过程、汽车行业的用户忠诚度分析、物流管理等方面进行了分析研究。因果分析方法有很多种，其中主要有工具变量法、断点回归设计、格兰杰因果关系检验等方法。

工具变量法（instrumental variable）是指当在模型中某一变量与结果之间高度相关，且与其它随机误差变量无关，则可以利用此变量结合模型中的回归系数得到对结果的估计量，而这个变量就可以被称作工具变量[29]。该方法由 Philip G. Wrigh 在 20 世纪 20 年代提出[30]，其本质是为了解决变量的“内生性问题”，即此变量即影响问题的“因”，又影响问题的“果”[31]。对于一个典型的线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta X + \varepsilon \tag{2-1}$$

其中 y 为因变量，即待研究的“果”；x_1 为自变量，即待研究的“因”；X 为其

他变量；ε 为误差项。如果 x_1 与 X 不相关，就可以利用最小二乘法对方程进行无偏估计。如果变量之间相关，即上文所说的存在内生性问题，x_1 共同影响“因果”，则进行最小二乘法估计时就是有偏的。

为了解决这个问题，工具变量法引入了一个外生变量 z，使其与 X 不相关而仅与 x_1 相关，因此可以认为 z 通过影响 x_1 来影响 y，利用 z 与 x_1 之间的直接关系即可以估算 z 与 y 之间的间接关系[31]。

直观显示如图 2.2，自变量与其他变量之间互相影响，二者同时决定了因变量，而工具变量仅与自变量有关，通过工具变量能够产生对因变量的影响。

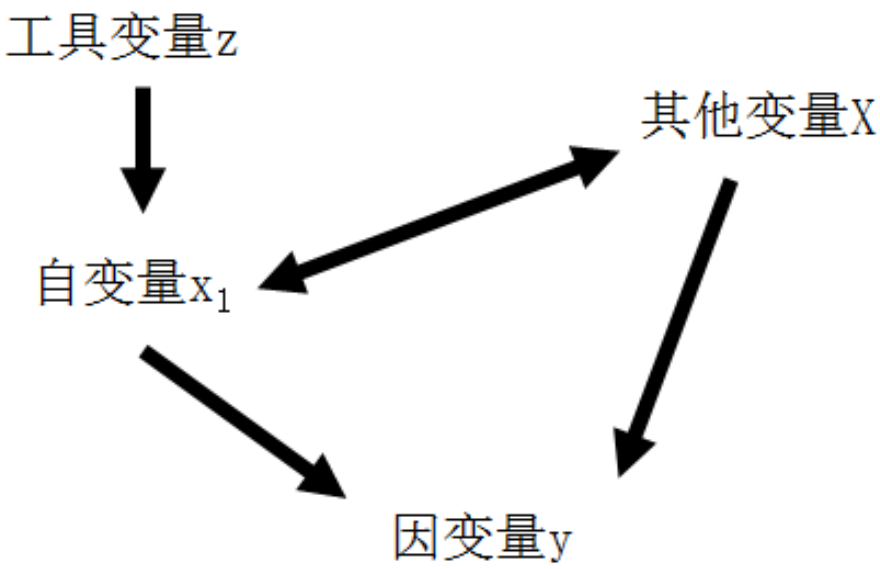


图 2.2　工具变量法示意图

文献[32],[33],[34] 通过数学推导验证了工具变量法的原理和可行性。在文献[35],[36] 和[37] 中，都利用工具变量法对经济社会体制、经济增长和贸易与学历等社会学问题进行了分析。

断点回归方法（regression discontinuity design）是一种拟随机实验，在进行因果分析时，其分析效果仅次于随机实验的一种方法，能够有效避免因果分析的内生性问题[38]。断点回归可以分为两类，一类是模糊断点回归（Fuzzy RD），可以分析协变量是确定型、不连续的情况，另一类是清晰断点回归 (Sharp RD)，可以分析协变量是概率函数的情况[39]。在利用断点回归模型进行因果检验时，往往还需要对检验结果进行稳健性检验，以验证检验结果的正确性。

该方法是由 Thistlethwaite 和 Campbell 在 1960 年首次提出[40]，在他们的研究中分析了学术获得荣誉奖励与学术成就之间的因果关系。由于学生是否能够获得奖励取决于其考试分数 x 是否能够超过阈值 c，如果高于阈值则授予奖励，否则不

授予。因此在考试成绩 c 出就产生了中断，c 就是这个断点。当通过分析发现学生的学术成就也发生了类似中断，例如成绩在 c 以下的学生的学术成就低于成绩在 c 以上的学生，那么就可以任务两者之间存在某种因果关系。如图 2.3所示，x 轴表示学习成绩，其在 c 处的中断对应了 y 轴学术成就的中断，因此可以看出，通过控制考试成绩就可以使得自变量是否取得荣誉奖励与因变量学生的学术成就完全独立，从而可以分析二者之间的因果关系。

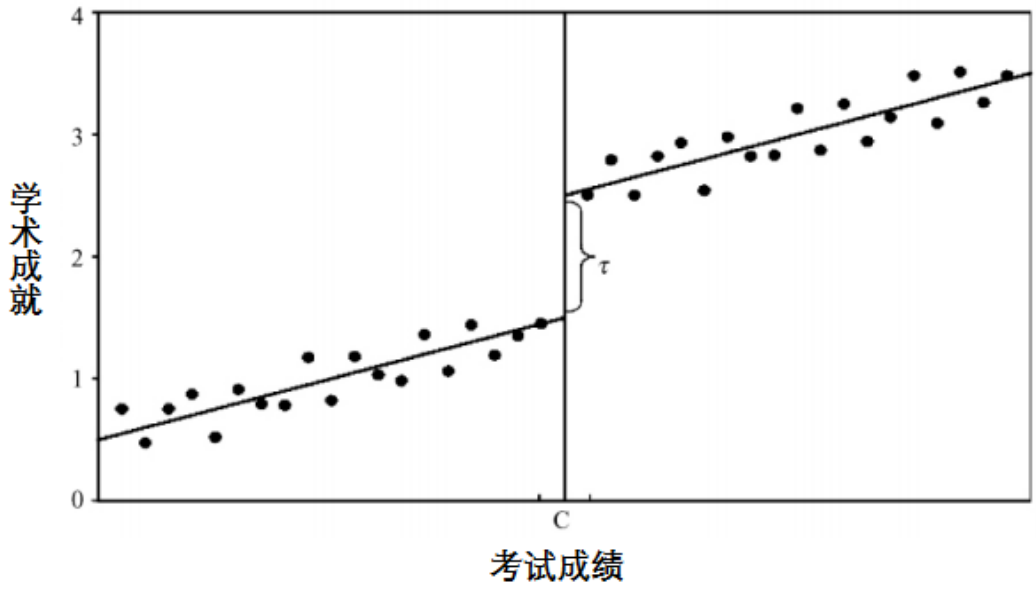


图 2.3　断点回归方法示意图[41]

王骏等人在文献[42] 中分析了在重点高中就读对学生学习成绩的影响。分析结果表明在重点高中读书对学生的学习成绩仅有轻微的正面影响，对于理科生来说高考成绩优于普通高中，但数值差异不大。而对于文科生来说在两种高中就读的成绩差异并不明显。

张川川等人在文献[43] 中分析了结婚年龄与婚姻稳定性之间的关系。通过断点回归方法，分析出婚年龄越大，离婚的概率越高的关联关系。而当利用传统最小二乘法进行估计时，结果显示没有明确的因果关系，这表明该方法的分析结果可能存在误差。

张建同等人在文献[44] 对上海市房地产价格指数和地方财政收入月度数据进行分析，研究了上海市房地产限购限贷政策同这两项指标之间的因果关系。分析结果表明该限购措施对房屋的销售价格具有统计意义上的负面影响，但经济意义上并不明显; 同时，对地方财政收入产生并没有显著影响。

席鹏辉等人在文献[45] 中利用多断点回归模型检验空气污染对地方环保投入的影响。分析结果显示较轻的空气污染会造成环保支出的减少，而当空气污染水平

较重时，就不会出现这种情况，二者之间不存在必然联系。

格兰杰因果关系检验（Granger Causal Relation Test）是著名经济学家 Clive W. J. Granger 提出的一种分析变量之间格兰杰因果关系的的方法，而这种关系是指“依赖于使用过去某些时点上所有信息的最佳最小二乘预测的方差”[46]。即在平稳的时间序列下，对于两个变量 XY，如果利用 X 和 Y 的过去信息对 Y 进行预测时，预测效果要由于单独利用 Y 的过去信息进行预测，则成 X 是 Y 的格兰杰原因[46]。对于格兰杰因果关系来说，其本质是统计意义上的一种因果性，并非实际意义上存在的客观关系，但虽然如此，通过格兰杰因果关系仍然能够揭示两个变量之间的关系。

Hiemstra 等人在文献[47] 中利用线性和非线性格兰杰因果关系检验每日道琼斯股票收益与纽约证券交易所交易量百分比变化之间的动态关系，从中能够发现股票收益与成交量之间存在着显著的双向非线性因果关系。

游和远等人在文献[48] 中同样利用格兰杰因果分析方法分析了地价变动与房价变动之间的因果关系，并通过对深圳市真实地价变动与房价变动分析，发现其中并不存在因果关系，否定了人们普遍存在的认为二者之间存在关联的直觉感受。

Haoyen Yang 等人在文献[49] 中通过使用 1954 至 1997 年期间更新的台湾数据，重新分析了能源消费与 GDP 之间的因果关系。还分析了台湾国内生产总值与总量之间的因果关系，利用格兰杰因果分析方法，作者还发现总能耗与 GDP 之间存在双向因果关系。

Soytas 等人在文献[50] 中研究了能源消耗与国内生产总值的时间序列特征，并对世界前十大新兴市场重新进行了检验。通过格兰杰因果分析发现能源消耗与 GDP 在阿根廷成双向因果关系，而在意大利和韩国能源消耗受 GDP 影响，在土耳其，法国，德国和日本则恰恰相反，即 GDP 受能源消耗的影响。

虽然格兰杰因果检验的检验效果及其应用方法存在一定争议[51],[52]，但利用其进行因果分析仍然起到一定的启示作用，揭示变量之间存在的一定关系。

2.4　基于倾向值匹配的因果分析

倾向值（propensity score）一词是由 Rosenbaum 和 Rubin 在 1983 年最早提出[53]，提出这一概念的主要目的还是为了为了排除混淆变量（confounding variables）对因变量的干扰，解决因果分析中的变量内生性的问题。倾向值的本质是一种条件概率，是在混淆变量被控制的条件下，研究因变量受自变量影响的一种条件概率。对于一些问题来说，是没办法利用随机试验来进行因果分析的。例如要研究大学教育对收入的影响，就无法随机找到被试，让一部分上大学而另一部分不是大学

对其进行判断。因此为了排除性别、年龄、地域差异等混淆变量对收入的影响，只观察是否上过大学这一条件，就需要通过倾向值匹配的方法来控制或消除这些差异的干扰。

简单来说，倾向值匹配的核心思想就是通过 Logistic 回归模型或 Probit 模型将所有混淆变量纳入模型当中，来计算因变量受到自变量影响的概率值，这个概率即为倾向值。然后将所有被试个体分为采用和没采用自变量的两组，对其中所有样本分别计算其倾向值，根据计算得到的倾向值的大小，对两组内的样本进行匹配，倾向值近似的样本匹配成为一对。这样匹配成功的两个样本，其倾向值近似，意味着各种混淆变量的近似相同，而唯一不同的就是是否采用了自变量。因此就可以通过倾向值近似样本的对比来观察自变量对因变量的影响结果[54]。

倾向值匹配算法一经提出，在经济学领域就引起了高度的重视，众多研究开始采用这个方法。

周振等人在文献[55] 中分析户籍差别对劳动力工资的影响，将户籍、性别、年龄、教育年限、职业性质、单位类型等因素作为混淆变量，利用倾向值匹配分析其中的因果关系。表 2.1显示了利用 Logistic 回归模型计算的各个样本的倾向值，表 2.2显示了对相近倾向值进行匹配之后的结果。

表 2.1　城镇居民和务工农民的倾向值

组别	城镇居民			务工农民		
	均值	最大值	最小值	均值	最大值	最小值
1	0.437721	0.607368	0.062054	0.183598	0.272155	0.020464
2	0.723871	0.811386	0.60783	0.336588	0.395526	0.272155
3	0.870943	0.923928	0.811634	0.462265	0.530409	0.395977
4	0.941486	0.963302	0.913953	0.607857	0.70278	0.530694
5	0.977176	0.996481	0.963386	0.82286	0.989905	0.704851
total	0.790239	0.996481	0.062054	0.482652	0.989905	0.020464

从表中可以看出，对于不同的实验对象能够计算得到不同的倾向值，而在进行倾向值匹配是，选取不同的匹配半径能够获得不同的匹配结果，这样也就使得工资差别的计算结果不同。但是从结果能够看出，倾向值匹配法有效地排除了不相关的混淆变量对工资的影响，分析出户籍与工资之间的因果关系。

侯珂等人在文献[56] 中分析了父母外出打工对于农村留守儿童在未来压力感知、抑郁、师生关系和幸福感等方面的影响。通过对是否采用倾向值匹配算法分析得到的两种结果进行了对比，验证了倾向值匹配对于因果分析的重要性。

张紫薇等人在文献[57] 中通过对毕业生调查数据，分析了本土留学教育对个人

表 2.2　倾向值匹配结果

匹配半径	样本	城镇居民工资	务工农民工资	工资差别
	unmatched	31059.23	20438.69	10620.54
ε=0.001	ATT	29398.23	22381.91	7016.32
	ATU	20819.65	24493.87	3674.22
	ATE			5892.70
ε=0.0001	ATT	28056.43	21803.06	6253.37
	ATU	20525.62	25543.25	5017.63
	ATE			5710.42
ε=0.00003	ATT	26822.60	21699.87	5122.71
	ATU	21077.86	25512.25	4434.40
	ATE			4792.32

经济收入的影响。利用倾向值匹配算法排除了个人性别、专业意向、家庭背景等因素对结果的影响，得出了学本土留学教育存在显著经济收益的结论。

Jalan 等人在文献[58] 中研究了阿根廷福利计划中与工人收入净增长之间的关系。通过利用倾向值匹配算法，排除了其他因素的异质性影响，发现福利计划的参与者的平均收益增长是总工资的一半以上。同时还验证了该算法的鲁棒性。

Seeger 等人在文献[59] 中利用大量数据资料，分析了他汀类药物治疗对心肌梗死造成的影响，通过使用倾向值匹配算法滤除了许多这种关联的潜在混杂因素，分组归类了他汀类药物的引发剂组和非引发剂组，揭示了其余心肌梗死之间的因果关系。

2.5　本章小结

本章主要对社交网络上的影视剧预测和推荐进行了介绍，由于社交网络的用户群体庞大，产生的数据量也随之增大，利用大数据能够准确地挖掘其中隐含的内在联系，从而对影视剧的宣传推广进行指导。而对于影视剧来说，其所特有的营销需求就是要求广大的受众群体和频繁的曝光度，社交网络正好满足了这一需求，通过在社交网络上的宣传，能够使更多的用户接收到宣传信息，并且引导用户自发扩散信息，实现病毒式传播，病毒式营销。

但是目前现有的各种推广手段所能获得的推广效果不尽相同，对于不同类型的影视剧，不同的宣传方，不同的受众群体，不同的推广手段就会导致不同的推广效果。那么为了能够挖掘出其中的内在联系，寻找最优的推广手段，就需要通过因果分析法，判断推广手段与推广效果之间的因果关系。目前应用比较广泛的因果分析方法有很多种，各自拥有不同的应用场景和分析特性，本章对其中的三种

因果分析方法进行了简要介绍。但是可以看出，目前因果分析的各种方法还主要是应用在经济学、社会学等人文社科领域，应用的数据量较小，在计算机科学领域应用较少，尤其是社交网络、影视剧分析等方面更是少有人涉足。

本文选取了其中一种因果分析方法：倾向值匹配法对电视剧的推广模式进行分析。该方法相对于其他因果分析方法来说，能够将研究对象相关的各种变量因素包含到分析过程当中，避免了利用工具变量方法时需要寻找与其他变量完全无关的工具变量的过程，而且有时候这个变量是无法找到的；也避免了利用断点回归方法时用来区分实验组和对照组的断点的寻找。因此，倾向值匹配算法能够获得更广泛的适用性，得到更加严谨的分析结果。

第 3 章　数据库与测量分析

3.1　引言

由于本文主要研究的是演员在社交媒体上对电视剧的推广作用，涉及到电视剧、演员及微博信息。本章首先对电视剧及演员相关数据的数据库进行说明，然后对研究过程中需要的演员、电视剧、微博的特征及其提取过程进行介绍。之后对数据进行测量和分析，包括对微博影响力、话题热度、推广模式和他们之间的关系的研究和分析。

3.2　数据集

3.2.1　数据库

数据库系统包括演员数据库和电视剧数据库两部分，如图 3.1。数据获取来源是爱奇艺[①]、微博[②]和豆瓣[③]。通过爬虫模拟登陆、模拟搜索的形式获取。

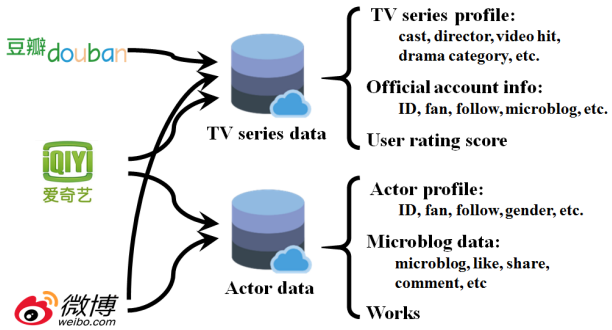


图 3.1　Data of the database

爱奇艺是国内领先的视频网站，提供海量、优质的电视剧、电影、网络视频服务，网罗了中国最广大的年轻用户群体。爱奇艺打造涵盖电影、电视剧、综艺、动漫在内的十余种类型的中国最大正版视频内容库，也是中国付费用户规模最大的视频网站[?]。爱奇艺数据爬虫获取爱奇艺电视剧页面每天的电视剧列表，然后访问每个电视剧的详情页，如图 3.2，获得电视剧的相关信息，包括电视剧名称、主演、导演、类型、集数、播放量、简介等信息。

① http://www.iqiyi.com/
② http://www.weibo.com/
③ http://www.douban.com/

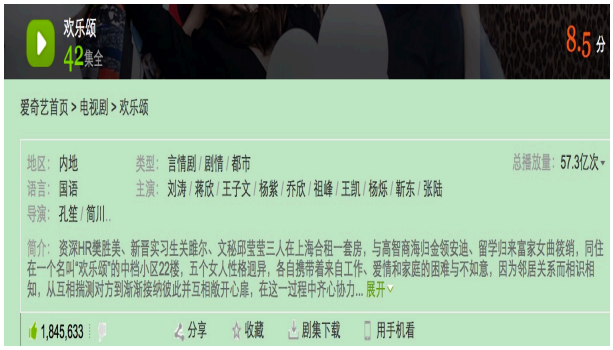


图 3.2　电视剧欢乐颂详情页

豆瓣是一个社区网站，提供关于书籍、电影、音乐等作品的信息，无论描述还是评论都由用户提供（User-generated content, UGC），是中国 Web 2.0 网站中具有特色的一个网站[?]。对于电影和电视剧的豆瓣评分能反应当前广大群众对其质量、水平的认同情况。豆瓣爬虫根据爱奇艺爬虫获得的电视剧列表，从豆瓣搜索入口模拟搜索，获得该电视剧的豆瓣评分，以反应该电视剧的质量，如图 3.3。

图 3.3　欢乐颂豆瓣主页

微博是由新浪公司推出的提供微博客的服务网站。用户可以在微博网页、客户端发布 140 汉字（280 字符）以内的信息，并可上传图片和链接视频，实现即时分享，同时，可以提供提供评论、转发、点赞功能。新浪微博是一个基于用户关系的信息分享、传播以及获取信息的平台，它占据中国微博用户总量的 57%，以及中国微博活动总量的 87%，是中国大陆访问量最大的网站之一[?]。微博爬虫包括电视剧官方微博爬虫、电视剧话题爬虫和演员信息爬虫。

大部分电视剧在微博上都有官方微博，如图 3.4，用来发布电视剧、演员相关的信息，比如电视剧预告、演员拍摄花絮、电视剧相关或者演员相关的推广信息等等，使得观众获得更多的关于电视剧和主演的信息和动态。电视剧官方微博爬虫根据爱奇艺爬虫获得的电视剧名称，获取电视剧在微博上官方微博的 id，然后模拟访问主页，获得该电视剧官方微博的基本信息，包括 id、粉丝数等，和发布的

所有微博信息。

同时，大部分电视剧在微博上会有一到多个话题，以“# 话题名 #”的形式存在，如图 3.5 中欢乐颂话题主页，当用户在微博中发布微博讨论电视剧时会带上相关话题，使话题活跃。电视剧话题爬虫根据从电视剧官方微博中出现及微博搜索入口搜索电视剧名字，可以得到电视剧的话题名称，进而获得该电视剧话题的相关信息，包括话题阅读量、话题讨论量、话题粉丝量等。

图 3.4　电视剧欢乐颂官方微博

图 3.5　电视剧欢乐颂话题主页

爱奇艺爬虫获取电视剧的主演后可以建成一个演员库，演员信息爬虫根据演员名字在微博上模拟搜索和百度搜索的形式获得演员微博上对应的 id，然后根据 id 获得演员的基本信息和发布的所有微博信息，如图 3.6。基本信息包括演员 id、昵称、性别、描述、注册时间、粉丝数、关注数、微博数等等。微博信息包括发布的微博内容、时间及其获得的点赞数、转发数、评论数。

图 3.6　刘涛微博主页

我们的爬虫运行在服务器后台，爱奇艺爬虫每天对电视剧信息进行爬取，微博爬虫每隔一段时间（3 个月）从微博上对演员信息增量爬取。电视剧数据获取时间从 2013 年 1 月到 2016 年 12 月 31 日，微博爬虫全量获取演员信息和发布的微博信息。

3.2.2　特征提取

为研究演员所发微博的影响力，我们以微博作为研究对象，因此需要获得微博的基本特征。推广电视剧的微博有三方面的特征，一是来自发微博的演员的特征，包括演员粉丝数、性别和作品数。二是来自该微博推广的电视剧的特征，包括电视剧的豆瓣评分和演员阵容。三是来自微博自身的特征，包括该微博的发布时间、日期和内容。发布微博的演员的特征中，粉丝数代表该演员在微博中粉丝量也就是影响力，当两个不同粉丝量的演员发布同一条微博时，粉丝量大的演员发布的微博会被更多的人看到，也就更容易起到更好的推广作用；男演员和女演员的行为本身有差异，其粉丝男女构成也有差异，相应的粉丝行为和对推广微博的行为也会有差异；演员的作品数代表该演员在影视界和群众视听中的活跃程度，演员出演的电视剧越多，越容易被观众记住也越有名。微博推广的电视剧的特征中，电视剧的豆瓣评分代表了看过的群众对该电视剧质量评价；电视剧的演员阵容用该电视剧所有主演的粉丝数的总和来表示，代表了这个电视剧整体的粉丝量和可能的受关注度，越多名人参与到同一部电视剧中，这个电视剧越容易火。微博自身的特征是由发布微博的演员决定的，包括时间、日期和内容，这是我们研究的主要内容。

对微博爬虫获取的微博，需要提取每一项微博的特征，因为微博自身的特征不需要从其他数据库或者爬虫获得，这里介绍前两方面特征的获取，如图 3.7。对微博所推广电视剧的豆瓣评分从豆瓣爬虫爬取或者已爬取的数据库中搜索可得。该微博推广电视剧的微博的演员阵容，先从爱奇艺爬虫获取的电视剧信息中获得该

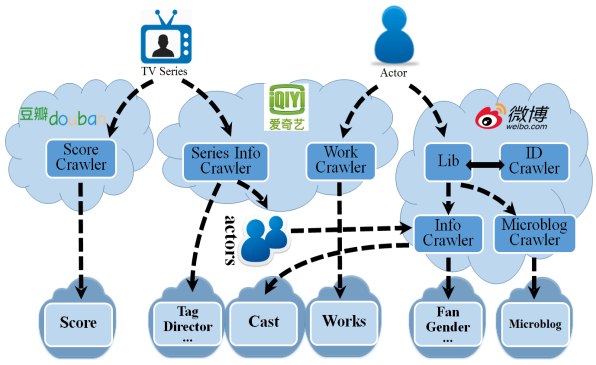


图 3.7　Feature Extraction Module

电视剧的主演，然后从微博爬虫或者已爬取数据中获得这些演员的粉丝数目，求和得到阵容。发布微博的演员的作品数为统计爱奇艺爬取的所有电视剧中由该演员主演的数目。演员的性别由微博爬虫或已爬取的演员基本信息中获得。

3.3　测量分析

3.3.1　微博影响力与话题热度

(1) 微博影响力。为了衡量一条微博的影响力, 我们采用粉丝参与度作为评判标准, 包括这条微博的转发数、评论数和点赞数，因为这些数据能反应看到并参与到电视剧推广中的用户数量。经过统计，所有演员的所有微博的转发数、评论数、点赞数的比例是 1 : 1.86 : 4.66, 为了使这三项评判指标具有相同重要程度, 因此赋予其权重比为 4.66 : 2.51 : 1，也能反应出转发、评论和点赞能带来的不同影响力。转发推广微博能使转发人的粉丝也看到，用户的参与感非常强，转发的人多了就具有滚雪球效应，能扩大宣传效果，因此其权重最大。用户通过评论推广微博参与到推广中，评论越多也越容易使该演员上微博热门，但评论并没有扩大用户群，所以用户参与感和影响力不如转发大。相比于转发和评论，用户点赞既不会扩大用户群也不需要输入文字，其用户参与感最弱，所以权重最小。定义微博影响力公式如下:

$$P_i^j = 4.66 * W_p^i + 2.51 * W_v^i + W_a^i \tag{3-1}$$

其中，P_i^j 表示与电视剧 j 相关的微博 i 的影响力，W_p^i, W_v^i 和 W_a^i 分别表示微博 i 的粉丝转发数、评论数和点赞数。

那么对一部电视剧带来的微博影响力是它的所有主演的发布的推广微博的影

响力之和：

$$T_j = \sum_{i=1}^{n} P_i^j \tag{3-2}$$

其中 T_j 表示电视剧 j 所有主演的推广微博影响力之和，n 为电视剧主演发布微博的数量。

(2) 电视剧微博话题热度。我们研究的电视剧，每部在微博中都有对应的微博话题。每个微博话题都有相应的阅读量和讨论量, 其中阅读量表示该话题的用户阅读数目, 讨论量表示用户发布的带该话题名的微博数目。电视剧上映前和上映中进行宣传时，宣传方都会在发布微博时带话题发布，即带“# 话题名 #”的形式发布微博。话题讨论量和阅读量越多，表明越多的用户参与到该话题讨论中，越容易成为热门话题，进而达到更好的宣传效果。用户带话题名发布微博，会增加微博话题讨论量，使其粉丝看到该微博，提高微博话题阅读量。也就是说微博话题阅读量是话题讨论和宣传效果的体现，代表了该电视剧微博话题覆盖到的人数，是其热度的体现。因此我们选用电视剧的微博话题阅读量来衡量电视剧在微博中的话题热度:

$$H_j = R_j \tag{3-3}$$

其中 H_j 表示电视剧 j 的微博话题热度,R_j 表示微博话题的阅读量。

(3) 电视剧微博影响力与微博话题热度。为检验电视剧微博影响力与话题热度的相关性, 我们选用皮尔森相关系数 (Pearson correlation coefficient) 和最大信息系数 (Maximal Information Coefficient, MIC) 来检验。皮尔森相关系数可以反映出两个变量之间的线性相关程度，系数的变化范围为-1 到 1。系数值为 1 意味着变量 X 和 Y 呈完全正相关关系，Y 随 X 增加而增加。系数值为-1 意味着 X 和 Y 呈完全负相关关系，Y 随 X 增大而减小。系数值为 0 意味着两个变量间没有线性关系。MIC 是专门用于快速探索多维数据集的双变量依赖关系的度量。MIC 是基于信息的非参数探索（MINE）统计方法系列的一部分，MINE 不仅可用于识别数据集中的重要关系，还可用于表征数据集。MIC 可以分析变量之间的广泛关系, 而不限定于特定的函数类型[60]。

表 3.1　Correlation between influence of microbloggings and topic hotness of TV series

variable x	varible y	PCC	MIC
influence	hotness	0.697	0.356

经过统计, 电视剧的微博影响力与最终话题热度高度相关, 两者之间相关性如表 1 所示。因此, 可以将微博影响力作为微博话题热度的评判标准, 即在判断推广微博带来的营销效果时，推广电视剧的微博影响力大则说明该电视剧话题热度高。电视剧话题演化的和微博影响力的动态数据也验证了二者之间的相关性，以欢乐颂为例。电视剧微博话题热度与微博影响力每日变化如图，从图中可以看出，二者趋势和形状几乎一致。

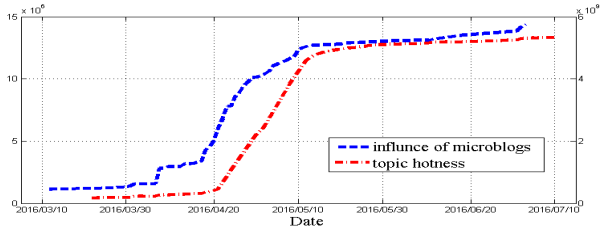


图 3.8　Everyday data of influence of microblogs and topic hotness

3.3.2　推广模式

(1) 推广周期。演员对电视剧的推广周期主要可分为三个阶段, 一是电视剧的筹备、拍摄阶段, 二是电视剧首播阶段即第一集播出前后, 三是首播之后的阶段，以电视剧“杉杉来了”为例，如图 3.8。演员在电视剧拍摄期间会分享一些电视剧拍摄地、拍摄进度、角色和剧情相关信息等内容，这些微博不但可以让粉丝对演员的动态有所了解，还可以让观众提前了解电视剧相关信息, 引起粉丝的兴趣和期待，起到提前宣传的作用。在第一集播放的前后几天是演员和电视剧官方宣传的高峰期，演员会发布微博进行大力度推广，我们将电视剧首集首播日前后 5 天定义为电视剧的首播阶段。统计发现虽然有些演员在微博中不活跃, 但是仍然会在首播前后, 发布微博进行推广宣传。首播阶段是影响收视率的关键时间，在这个时期演员推广可以提醒和号召粉丝看这个电视剧，促进提高话题热度和收视率。首播之后, 演员还会继续进行微博推广, 发布微博传递信息或者与粉丝互动，能在维持现有观众继续观看后续剧集的同时，吸引更多观众参与话题讨论并提高电视剧收视率。除此之外，当电视剧登陆其他电视台开播时，演员也会发布微博进行推广。因此, 从推广周期来看，演员推广微博时期分为三类, 分别为筹备阶段、首播阶段和首播后阶段。

(2) 推广时间。在一天中不同时间发布推广微博，由于用户使用习惯的差异，被用户看到的时长和概率不同，也就会达到不同的推广效果。比如同一条推广微博，在用户刷微博高峰和上班时间发布，肯定会获得不同的参与度。在一天中，午饭前、晚饭前及晚上通常都是用户刷微博的高峰时段。图 3-8 显示了演员发布微博

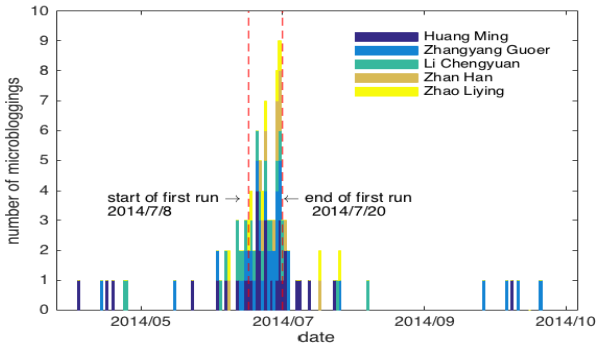


图 3.9　Microbloggings distribution of different promotion period.

时间的数目分布情况，同时也包括演员在电视剧上映前和上映后的数目差异。在本文中，我们想知道演员在不同时间段发布微博，哪个时间段会带来更好的推广效果，因此, 我们将演员发布的微博根据发布时间分成三类, 分别是上午 (1 时至 12 时), 下午 (12 时至 18 时), 晚上 (18 时至次日 1 时)。

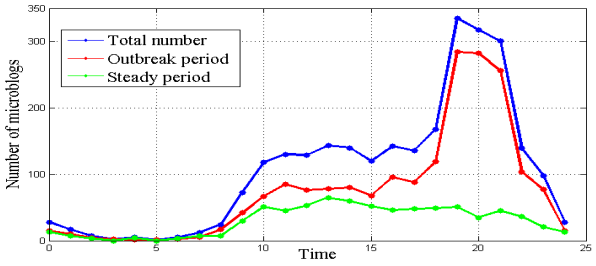


图 3.10　Distribution of actors' microblogs along the time

(3) 互动模式。在微博上推广电视剧过程中，演员会与其他主演、粉丝、电视剧官方微博等有很多交互行为来增加用户的参与度，促进电视剧的推广。图 3-9 展示了电视剧欢乐颂推广过程中出现的互动关系。图中每个节点表示一个演员或者官方微博账号，节点大小代表了与其他账号互动的数目，发过的微博中与其他账号互动越多，节点大小越大。有向箭头表明了互动的方向，边的粗细代表互动的频率，频率越高边越粗。一条微博可以使演员间、演员与官方微博间、演员与粉丝间建立互动关系，如图 3-9 中为一条微博的传播情况，图中黄色点是该电视剧的主演，每个点是一个微博用户，每条边是一个转发关系。

演员发布推广微博时有四种互动模式：a. 与电视剧其他主演互动。演员会转发其他主演微博, 或者发布微博 “@” 其他主演, 在微博中与其他主要进行交流。主演间互动一方面会增加微博话题性, 另一方面也会结合多方粉丝, 扩大推广效应。b. 与官方微博互动。经过统计发现, 一种常见的推广模式是, 官方微博作为源头, 发布推广微博, 并在微博中 “@” 主演, 主演将会转发这条微博, 提升官微推广的影响力。我们用累计分布图来分析演员响应官方微博的时间概率分布。以电视剧” 欢乐颂”

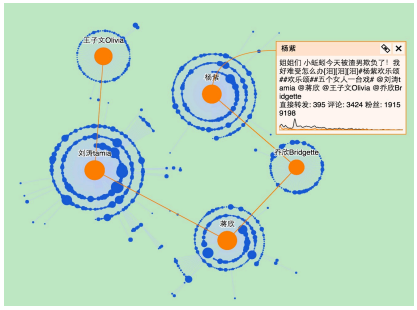


图 3.11　Interaction relationships of one microblog for TV series ”*Huanlesong*”

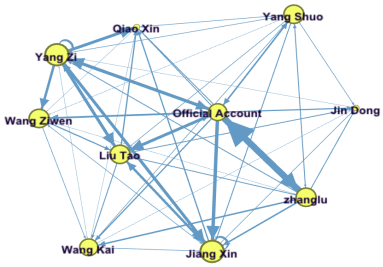
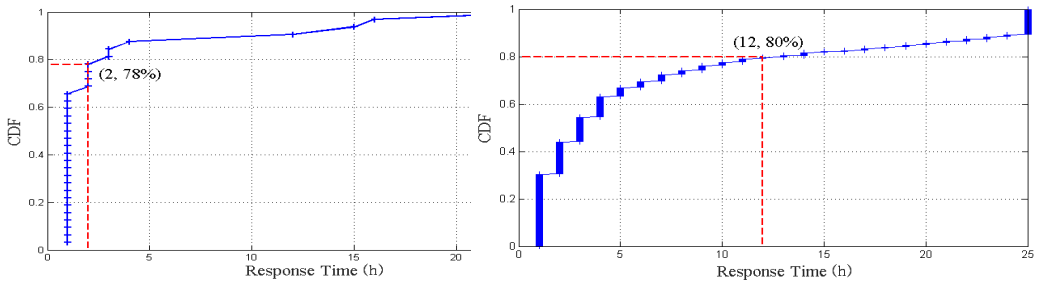


图 3.12　Interaction relationships of TV series ”*Huanlesong*”

为例，78% 的主演的响应时间在 2 小时以下。统计发现，对所有电视剧而言见图 3-9，80% 的主演的响应时间都会在 2 小时以下。c. 原创非互动微博。此类微博由演员原创, 虽然不包含与其他人的互动, 但通过表达针对电视剧剧情或者人物的感想和看法, 更好的表达演员的情感和想法, 更能够吸引粉丝的注意力。d. 其他互动模式。除了以上三种互动的其他模式称为其他互动模式，包括演员转发视频网站官方微博的微博、转发粉丝微博、转发宣传媒体微博等。

图 3-9 显示了所有演员微博中以上四种模式所占比例，其中也包括了在电视剧上映前后的比例差异。



(a) TV series: ”*Huanlesong*”　　(b) All TV series

图 3.13　Response time of actors to official accounts’ references

(4) 推广模式与话题热度。为验证各种推广模式与话题热度的相关性, 我们用皮尔森相关系数和最大信息系数来检验。通过两种检测方法可以表明各个推广模

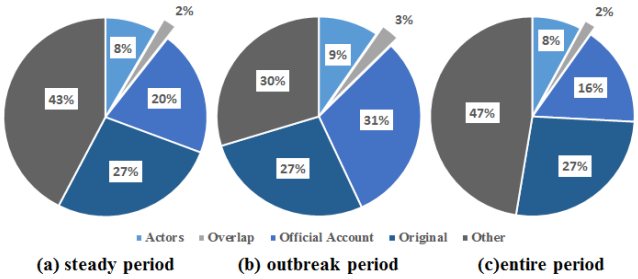


图 3.14　Proportion of interaction modes

式与话题阅读量具有较强的相关性，因此接下来我们就可以检验各个推广模式对话题热度的因果性和有效程度。

表 3.2　The correlation between promotion patterns and topic hotness

promotion patterns		PCC	MIC
promotion period	preparation	0.389	0.284
	premiere	0.515	0.342
	post-premiere	0.813	0.399
promotion time	morning	0.239	0.306
	afternoon	0.542	0.407
	evening	0.856	0.340
interaction mode	actors	0.390	0.292
	official accounts	0.745	0.356
	original	0.600	0.297
	other	0.630	0.349

第 4 章　演员社交推广行为的影响力模型

4.1　引言

理想情况下，检验策略有效性的黄金标准是基于实验的方法，比如做 A/B 测试，不同的策略可以随机分配给用户，但是完全随机对照实验存在很多限制，比如成本非常高，操作困难，在实际环境中不可行。而采用非随机对照实验则容易出现组间基线不齐，数据偏倚。因此提供一个统计方法而不是实验方法来检测策略的有效性是非常重要和有必要的。倾向值匹配算法解决了以上问题，它将混淆变量纳入 Logistic 模型, 归纳成一个倾向值, 在进行倾向值匹配时, 混淆变量得以控制, 使得对最终结果的影响只能归因于自变量 [11]。比如, 将微博发布时间作为自变量, 控制其他混淆变量近似相等, 如选择同一个演员, 在电视剧首播期间, 发布内容相似的原创微博, 因此最终这些微博影响力的差异只能归因于发布时间。

4.2　倾向值匹配算法

4.2.1　混淆变量

在倾向值匹配算法中，混淆变量是指，除了研究策略外的其他可能会影响策略有效性的变量。如演员 A 和演员 B 发布的微博带来的影响力的不同，除了策略的不同，还可能是因为二者的粉丝量不同带来的差异。这些混淆变量对推广效果的影响就是选择性误差，在倾向值匹配算法中，通过控制混淆变量来遏制选择性误差对结果的影响。本章中，选用微博的与演员相关的特征粉丝量作为混淆变量，记做 X。为压缩粉丝数数据尺度，对粉丝数取 10 的对数。

4.2.2　推广策略

为研究演员社交行为对电视剧的推广作用, 将演员发布的推广微博作为研究对象, 自变量是三类推广模式下的 10 项推广策略，如表 4-2。我们定义每一项推广策略都是一项策略 t，其值为 T。在我们的研究中，$T = 0, 1$，其中“1”代表采用这项策略，“0”代表不采用这项策略。同时，我们定义与采用干预 t 的微博的影响力作为输入 $Y(t)$。当研究一项推广策略时，另外的推广模式下的推广策略会作为混淆变量加入变量 X。

表 4.1　Promotion patterns and promotion strategies

Promotion Pattern	Strategy
Promotion period	preparation Premiere period Post-premiere period
Promotion time	morning afternoon evening
Interaction mode	actors official accounts original microblog others

4.2.3　算法步骤

如果两条微博唯一的差异是采用或没采用一项策略，那么我们可以合理将这两条微博最终影响力的差异完全归因于这项策略的有效程度。这也告诉我们，可以通过控制其他变量都相同或者类似来计算策略的有效性。倾向值匹配算法便是通过控制倾向值来达到控制和匹配的目的。在基于倾向值的匹配过程中混淆变量被控制起来了，那两组微博影响力的差异就只能归因于是否采用了策略，进而能评估策略的有效性。具体的，算法流程如表 4-3。

步骤一：计算倾向值。我们将混淆变量纳入逻辑斯地回归模型来计算倾向值，倾向值指的是计算研究个体在控制可观测到的混淆变量的情况下[?]，受到某种自变量影响的条件概率。本文中指的是这条微博在其特征固定的情况下，采用某项策略的条件概率：

$$e_i = P(T_i = 1|X_i) \tag{4-1}$$

步骤二：基于倾向值进行匹配。根据是否采用了某项策略，微博整体可分为两个样本集合。对两个样本集合中的微博根据倾向值采用匹配算法进行匹配。匹配算法有多种，从匹配数量角度有一对一匹配、多对一匹配和多对多匹配，从匹配方法角度有贪心匹配、最近邻匹配、基于卡尺的最近邻匹配等等[?]。本文中采用一对多匹配和基于卡尺的最近邻匹配。在基于卡尺的最近邻匹配中，卡尺距离通常为倾向值对数的标准差的 0.2 倍，这个距离被证明在多种参数设置中会产生最优的风险差异估计。通过计算，在本章的实验中，卡尺距离为 0.09。

步骤三：评估策略效果并检验有效性。在基于倾向值对两组微博进行配对后，

通过计算所有配对的微博的影响力差异，便能得到策略带来的有效水平。在我们的数据中，因为我们关注的是最终采用了某项策略的微博，因此我们用平均干预效果 ATT(Average Treatment effect for the Treated) 来衡量策略效果。然后用 t 检验（t-test）来检验干预效果的显著性水平，即检验两组微博中的微博影响力水平是否有显著差异。

$$ATT = E[Y(1) - Y(0)|T = 1] \tag{4-2}$$

步骤四：平衡性诊断。倾向值匹配的目的是通过控制倾向值近似来达到混淆变量类似的目的，因此检验配对微博在混淆变量上是否有显著差异是必不可少的，它代表配对的微博在混淆变量上的相似程度。平衡性检验能评估倾向值模型是否充分恰当的进行了应用。标准差用于量化采用和没采用策略的两组微博混淆变量平均值的差异[?]，其对连续变量和二值变量的公式定义分别如式 4-1 和 4-2。最后，累积密度图和分位数位图（qq 图）用于比较采用策略和没采用策略两组混淆变量的分布情况。

$$d = \frac{(\overline{x}_{treated} - \overline{x}_{untreated})}{\sqrt{\frac{s^2_{treated} + s^2_{untreated}}{2}}} \tag{4-3}$$

其中，$\overline{x}_{treated}$ 和 $\overline{x}_{untreated}$ 分别对应采用策略和没采用策略组样本的连续混淆变量的均值, $s^2_{treated}$ 和 $s^2_{untreated}$ 分别对应采用策略和没采用策略组样本的连续混淆变量的方差.

$$d = \frac{(\hat{p}_{treated} - \hat{p}_{untreated})}{\sqrt{\frac{\hat{p}_{treated}(1-\hat{p}_{treated}) + \hat{p}_{untreated}(1-\hat{p}_{untreated})}{2}}} \tag{4-4}$$

其中，$\hat{p}_{treated}$ 和 $\hat{p}_{untreated}$ 分别对应采用策略和没采用策略组样本的二值混淆变量的均值。

4.3　结果分析

通过倾向值匹配算法, 得到针对不同推广策略的干预效果, 可以更好的指导演员进行推广, 提高电视剧微博热度。

(1) 推广周期。按推广周期计算筹备阶段、首播阶段和首播后阶段的平均干预效果 ATT 如表 5 所示。可以看到, 在筹备阶段发布的微博的平均干预效果为负值, 且 t 检验结果显著，说明筹备阶段发布的推广微博影响力不如非筹备阶段, 即首播阶段和首播后阶段。而首播阶段发布微博的影响力效果虽然为负值，但是经过 t 检验，与非首播阶段发布微博的影响力效果没有显著差异。与二者相比, 首播后阶段发布微博的影响力效果与筹备阶段和首播阶段相比有显著差异，会显著提高。分析原因可能是粉丝们在筹备阶段和首播阶段只能看到关于电视剧的少量信息和花絮, 参与感较低, 导致关注度不高; 而当用户看过电视剧后, 在演员利用微博进行电视剧推广时, 粉丝们可以针对剧情、人物等积极参与讨论, 对微博内容也可以有更多的态度和观点, 此时进行推广, 效果会更好。

表 4.2　Comparison of each promotion period

Promotion Period	ATT	Significance
preparation period	-674.204	0.000
premiere period	-182.854	0.260
post-premiere period	757.065	0.000

(2) 推广时间。计算各个推广时间干预效果 ATT 和 t 检验的结果如表 6 所示。与想象中不同的是，早晨发布推广微博获得的推广效果要显著好于中午和晚上发布。原因可能是，推广信息具有时效性，对当天来说，早晨发布的微博一天中被看到的时长和概率是最大的，到了第二天，消息的时效性大大降低，导致粉丝的参与度大大降低。因此建议演员更多的在早晨发布推广微博。

表 4.3　Comparison of each promotion time

Promotion Time	ATT	Significance
morning	879.083	0.000
afternoon	-138.270	0.036
night	-314.376	0.000

(3) 互动模式。按互动模式将演员推广行为分为与其他主演互动, 与官微互动, 原创非互动和其他互动 4 种方式。经过倾向值匹配算法, 得到结果如表 7 所示。可以看到演员与其他主演互动、与官微互动时, 平均干预效果都有所增加, 都对推广有显著效果。因为主演和主演互动本身就在制造话题性, 与官微互动, 会转发一些电视剧情节、花絮及与演员相关的信息, 信息量较大, 会增加粉丝参与。而原创非互动模式与非原创相比, 平均干预效果显著提高, 推广效果提高的最大。分析演员

的原始微博可知, 原创微博更能表达演员的情感和对粉丝参与的情感呼吁, 因而更能得到粉丝的支持和参与, 推广效果也会更好。与之对比, 转发其他视频网站、粉丝的微博等其他互动方式的推广效果明显不如另外 3 种互动方式, 不管从携带信息量还是从演员情感角度看, 其他互动方式的宣传效果都会较差。

表 4.4　Comparison of each interaction mode

Interaction Mode	ATT	Significance
actors	759.315	0.000
official accounts	3413.222	0.000
original	12007.880	0.000
other	-7195.497	0.000

OBJ OBJ OBJ OBJ OBJ OBJ

综上可知, 在演员进行推广时, 建议从宣传周期上, 选择首播后阶段发布更多的微博进行推广; 从发布时间上, 多选择上午发布, 增加用户参与度; 在微博互动方式上, 在时间允许的情况下, 多发原创微博, 多与官微和演员互动, 能更好的推广电视剧, 提高微博话题热度, 增加电视剧点击量。

4.4　模型显著性及平衡性检验

平衡性检验用来评估倾向值匹配模型是否恰当的进行了应用。对每项策略计算标准差。结果显示标准差的取值范围为 0 到 0.049，意味着在采用策略组和没采用策略组，配对的连续变量和非连续变量的平均值非常类似。总之，上述分析显示倾向值匹配算法在我们的数据中被恰当的进行了应用。因此在我们研究的策略和利用的混淆变量的基础上，对观测数据得到的结论是在电视剧首播后期、早晨发布原创或与主演和官微互动的微博，能达到更好的宣传效果，提高粉丝的参与度和话题热度。

第 5 章　基于话题演化的演员社交推广行为影响力模型

5.1　引言

本章是对第四章应用的优化应用，使倾向值模型更合理的应用在演员推广行为的分析中。不但结合话题演化规律，将话题根据电视剧上映时间分成平稳期和爆发期，分别在两个时期分析演员的推广行为和推广策略，而且将更多会影响电视剧话题热度的因素考虑在内，使分析结果更具有可靠性。

5.2　电视剧话题演化规律

研究发现，微博热点话题在传播过程中，不仅关键词高度集中、受外界环境影响显著，而且传播具有周期性。话题周期一般包括话题发生、话题发展、话题高潮和话题消退四个阶段[61]。在话题发展的过程，起引导作用的是大量“意见领袖”的参与。意见领袖是指在人际传播网络中的“活跃分子”，他们经常为他人提供信息、施加影响，在大众传播效果的形成过程中起着重要的中介或过滤的作用，他们将信息扩散给受众，形成信息传递的两级传播[?]。

统计发现，电视剧微博话题演化符合话题演化的基本特征，具有一定的周期特性，如图 5-1，电视剧“柠檬初上”的话题发展正是经历了话题发生、话题发展、话题高潮和话题消退的过程。在电视剧话题演化的过程中，演员充当了意见领袖的作用，他们不仅拥有很庞大的粉丝群体，还有很大的社交影响力。他们发布的微博信息往往能受到高度关注，在电视剧话题发展过程中，演员发挥了核心作用。在电视剧话题发展过程中有三个重要时间节点导致了电视剧话题向下一阶段的发展，分别是电视剧上映时间确定、首集播出时间、首轮播出结束。在电视剧上映时间确定之前，演员会不时发些电视剧相关信息包括电视剧拍摄进度、拍摄期间趣事、角色定妆照等信息的微博，提前宣传使电视剧话题在微博产生。当电视剧上映时间确定后，主演通常都会发微博宣传电视剧，推动话题发展和酝酿，吸引粉丝和普通用户注意上映时间，吸引第一拨群众。当电视剧首轮播放第一集之后，演员发布微博来吸引观看用户参与电视剧剧情、角色相关的讨论，用户也因为看过电视剧，在微博上更愿意参与到话题讨论中去，带来话题热度的大发展。演员和用户的这种行为会持续到电视剧首轮播放结束后几天，之后话题因为其时效性，最终走向消退。

电视剧话题出现和发展期相对于话题爆发和消退期存在本质上的差别，即电

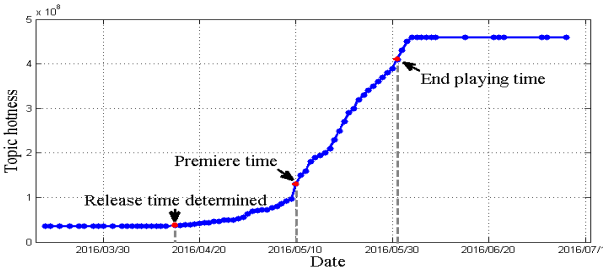


图 5.1　Topic evolution of "*Ningmeng Chushang*"

视剧是否已经上映，带来用户对电视剧的内容了解的差异和参与方式的不同，进而导致话题热度发展不同。因此，本文中将电视剧话题根据电视剧是否上映分成两个主要时期，上映前为平稳期，上映后为爆发期。在电视剧话题的整个发展过程中，因为电视剧所处宣传期不同，演员在微博上不同时期的推广方式也会有差异。例如，对发布推广微博的时间，如图 3-1，在平稳期夜晚发布微博很少，白天不同时间发布微博数量比较平稳。但是在爆发期，在电视剧上映前后演员会有一个发布微博的高峰。在此时间发布微博不但能维持既有用户群观看电视剧、参与到电视剧话题讨论，还能吸引和号召新用户观看当天剧情。在不同时期演员发布推广微博的互动模式也有差异，从图 3-2 可以看出，与平稳期相比，爆发期演员会更多的与官方微博互动。因为当电视剧上映后，官方微博会发布更多的微博，包括剧情讨论、角色讨论、剧照、下集预告等等。因此，研究演员在不同时期推广策略的有效性是非常有必要的。

5.3　倾向值匹配算法

5.3.1　混淆变量及推广策略

在倾向值匹配算法中，应该控制混淆变量使得混淆变量在策略组和非策略组分布相似，这样才能更好的分析策略的效果。考虑到影响一条微博在微博中热度的因素有很多，除了演员的推广策略本身，还受演员本身属性及电视剧属性等的影响，因此，本章中我们选用如表 5-1 中包括演员和电视剧相关属性的信息作为混淆变量，使分析结果更具有可靠性。

因为话题演化的周期特性，以及在电视剧宣传的不同时期演员都需要在微博中进行推广，所以在平稳期和爆发期分别分析演员推广的有效策略更为合理。在算法应用中，将演员行为数据按照电视剧上映时间分成两部分，对两个时期的数据分别应用倾向值匹配算法，其中，推广策略如表 5-2。同样地，我们定义每一项推广策略都是一项策略 t，其值为 T。在我们的研究中，$T = 0, 1$，其中"1"代表采用这项策略，"0"代表不采用这项策略。同时定义与采用干预 t 的微博的影响力

表 5.1　Baseline Covariates

Type	Characteristics
Actors	number of fans gender number of works
TV series	cast score

作为输入 $Y(t)$，混淆变量为 X，当研究一项推广策略时，另外的推广模式下的推广策略会作为混淆变量加入 X。

5.4　结果分析

将倾向值匹配算法分别应用在演员推广的平稳期和爆发期，结果如下：1）平稳期推广策略效果分析。推广策略有效性和 t 检验结果如表 5-3 所示。从表中可以看出，与在早晨和中午发布微博相比，在晚上发布微博的效果更好。用户晚上通常会花较多的时间在社交网络上，对他们而言，晚上发布的推广微博有更大的概率被他们看到。对互动模式而言，发布原创微博带来的推广效果要明显远远好于其他互动模式。与之相反，与官微互动和与其他人互动带来的推广效果较差。因为这种微博不能很好的表达演员自己的情感和想法，而且此时电视剧并未上映，粉丝参与电视剧推广微博更多的是因为演员，而不是因为剧情，因此对能更好表达演员自己想法和情感的原创微博会带来更好的粉丝参与度。与此同时，与其他主演互动的模式没有明显好的推广效果。

表 5.2　Effects of promotion strategies in steady period and outbreak period

Steady Period			Outbreak Period		
Strategy	ATT	T-test significance	Strategy	ATT	T-test significance
morning	-0.141	0.267	morning	0.222	0.410
afternoon	0.023	0.837	afternoon	0.211	0.008
evening	0.359	0.024	evening	-0.176	0.004
actors	-0.088	0.561	actors	-0.078	0.302
official accounts	-0.318	0.002	official accounts	-0.001	0.989
original microblog	2.633	0.000	original microblog	1.860	0.000
others	-0.743	0.000	others	-0.616	0.000

2）爆发期推广策略效果分析。根据图 5-3 可以看出，在爆发期，下午发布的

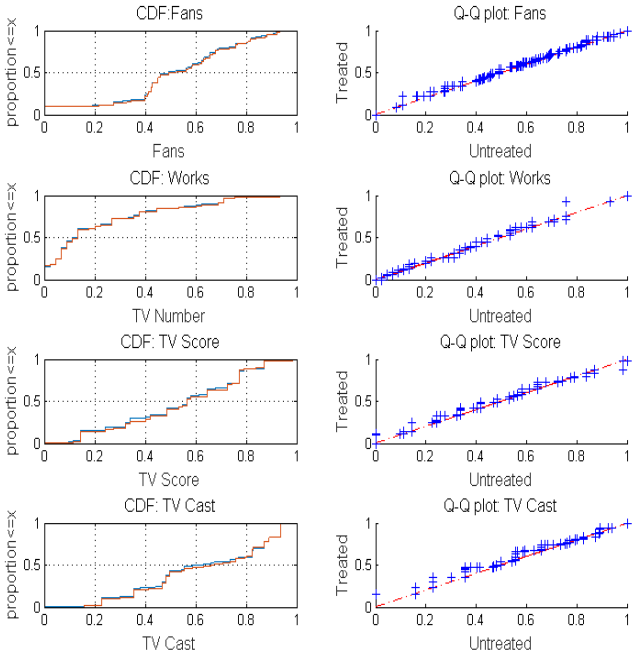
推广微博会带来更好的推广效果。在爆发期，演员通常在晚上会集中发很多微博，导致平均每条微博的粉丝参与度相对不高。从互动模式来看，发原创微博仍然是最有效的推广策略，原创微博需要演员投入更多的精力编写，同时更好的表达了他的想法和对粉丝参与的期望。而与其他人互动仍然效果最差，与官方微博和其他主演互动的推广效果并不显著。因此可知，在爆发期，演员在下午发布原创微博能获得更好的宣传效果。

5.5　模型显著性及平衡性检验

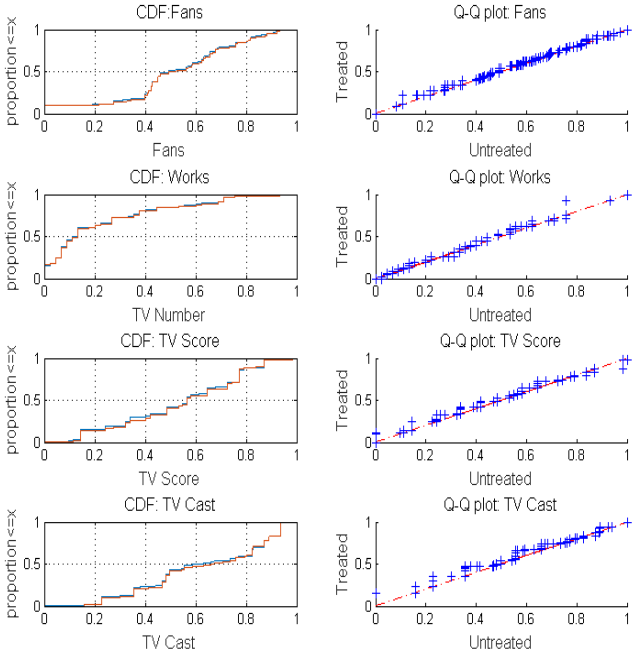
平衡性检验用来评估倾向值匹配模型是否恰当的进行了应用，包括分布计算各个策略下，所有混淆变量标准差在策略组和非策略组的平均值是否有显著差异，作累计分布图和分位数-分位数图（qq 图）来看混淆变量在策略组和非策略组是否分布一致。表 5-4 显示了两个时期，在各个时间策略下混淆变量的标准偏差。平稳期的标准偏差值在 0.001 到 0.073 之间，爆发期的标准偏差值在 0.002 到 0.037 之间，表明了策略组和非策略组的匹配样本的平均值分布一致。图 5-6 以是否采用原创微博策略下微博所属演员粉丝量、所属演员作品数、所属电视剧评分、所属电视剧阵容的累计分布图和分位数-分位数图来刻画在采用策略组和非策略组的混淆变量的分布情况，从图中可以看出，平稳期和爆发期的策略组和非策略组的这几个连续混淆变量的分布都非常类似。因此，倾向值匹配算法在我们的数据中得到了适当的应用，结果准确。

表 5.3　Standardized difference of the mean of steady period and outbreak period

Steady Period				Outbreak Period			
	morning	afternoon	night		morning	afternoon	night
actors	0.012	0.073	0.058	actors	0.028	0.027	0.008
official accounts	0.029	0.007	0.001	official accounts	0.007	0.011	0.003
original microblog	0.053	0.006	0.034	original microblog	0.008	0.003	0.008
other	0.019	0.009	0.047	other	0.018	0.006	0.006
#works	0.006	0.023	0.007	#works	0.006	0.013	0.007
gender	0.012	0.003	0.054	gender	0.023	0.047	0.010
#fans	0.008	0.033	0.003	#fans	0.005	0.037	0.002
cast	0.012	0.033	0.032	cast	0.002	0.023	0.003
score	0.030	0.049	0.043	score	0.002	0.034	0.018



(a) steady period



(b) outbreak period

图 5.2　Comparing distribution of continuous covariate in propensity-score matched sample

第 6 章　总结与展望

6.1　工作总结

随着软、硬件技术的不断提升，摄像头的各项性能指标的逐步提高，但造价却越来越便宜。这就使得对于摄像头的大规模应用成为可能，关于包含多个相同或不同摄像头的多摄像头系统的研究也日趋增多，多摄像头系统的应用场景也越来越广泛。而对于多摄像头系统来说，由于系统内可能存在的各种误差、延时，如何将所有摄像头的拍摄时间进行同步，成为了一个系统研究的关键问题，而本文则提出了一个基于 LED 点阵检测系统的多摄像头系统的时间同步方法。

该方法利用 FPGA 与 LED 点阵组成检测系统，使得 LED 点阵不断变化各个 LED 灯的亮灭组合方式，形成一系列的 LED 点阵状态。当摄像头拍摄到 LED 点阵时，根据拍摄图像中的 LED 点阵状态，可以判断该状态在状态序列当中所处的位置，从而可以根据各个状态的持续时间检测出摄像头的拍摄时间。而根据 LED 点阵编码方法的不同，这个状态序列可以呈现出不同的显示规律，摄像头在进行拍摄时也能够得到不同的拍摄效果，导致不同的检测精度。通过对不同编码方法的实验对比，可以发现进位流水编码方法的状态序列长，检测精度高，而且具有叠加可识别性，同时可以适用于大多数摄像头，对摄像头没有过多的性能、参数要求。

在获得摄像头的拍摄时间后，可以根据该时间计算多摄像头系统内各个摄像头之间的时间差，即当前系统的同步误差。然后选取拍摄时间的中间值，控制各个摄像头进行拍摄时间的调整。对摄像头拍摄时间的调整主要有拍摄暂停和帧率变换两种方法，这两种方法都能够取得较好的调整效果，使得系统内所有摄像头最终实现拍摄时间的同步。

在实际应用中，本文利用一台图像处理服务器和四台树莓派电脑搭建了一个多摄像头系统。并利用该系统对上述各方法进行验证。在克服了卷帘快门摄像头果冻效应和多摄像头视野校准等问题之后，使用该系统对 LED 点阵的各种编码方法进行了比较，同时也对摄像头拍摄时间的检测方法进行了验证，还对该系统进行了时间同步实验，都取得了很好的实验结果。

6.2　未来展望

在后续工作中，为了提高检测精度，可以扩大 LED 点阵中包含的 LED 灯的数量。目前的检测精度受果冻效应的影响，当点阵中每行 LED 灯的数量增多时，可以使得每个 LED 灯的亮灭变化加快，从而提高检测精度。同时在进行系统同步的过程中，受系统内随机延时的影响需要多次迭代逐渐同步。而在后期系统优化过程中可以通过大量重复性试验，从随机延时中挖掘出一定规律，从而提高系统同步速度。

参考文献

[1] 中国互联网络信息中心. 2015 年中国社交应用用户行为研究报告 [EB/OL]. [2017-04-23]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/sqbg/201604/t20160408_53518.htm.

[2] Hyder S. The Zen of social media marketing: an easier way to build credibility, generate buzz, and increase revenue. BenBella Books, Inc., 2016.

[3] Tuten T L, Solomon M R. Social media marketing. Sage, 2014.

[4] Saravanakumar M, SuganthaLakshmi T. Social media marketing. Life Science Journal, 2012, 9(4):4444–4451.

[5] Heymann-Reder D. Social Media Marketing. Addison-Wesley Verlag, 2012.

[6] Zavišić S, Zavišić Ž. Social network marketing. CROMAR kongres (22; 2011), 2011.

[7] Bolotaeva V, Cata T. Marketing opportunities with social networks. Journal of Internet Social Networking and Virtual Communities, 2010, 2010:1–8.

[8] Facebook. Facebook reports fourth quarter and full year 2016 results[EB/OL]. [2017-04-23]. https://investor.fb.com/investor-news/press-release-details/2017/facebook-Reports-Fourth-Quarter-and-Full-Year-2016-Results/default.aspx.

[9] 微博. 微博 2016 年 q4 季报解读 [EB/OL]. [2017-04-23]. http://weibo.com/ttarticle/p/show?id=2309614081122958030309.

[10] 张琦. 新媒介环境下的中国电影营销策略研究 [D]. 福州: 福建师范大学, 2011.

[11] 于瑞华. 基于 web2. 0 的电影营销策略研究. 电影文学, 2012, (15):18–19.

[12] Ono C, Kurokawa M, Motomura Y, et al. A context-aware movie preference model using a bayesian network for recommendation and promotion. International Conference on User Modeling. Springer, 2007. 247–257.

[13] Golbeck J, Hendler J, et al. Filmtrust: Movie recommendations using trust in web-based social networks. Proceedings of the IEEE Consumer communications and networking conference, volume 96. Citeseer, 2006. 282–286.

[14] 熊莉, 冯利芳. 《失恋 33 天》电影新营销的胜利. 成功营销, 2012, (2):74–79.

[15] Kwak H, Lee C, Park H, et al. What is twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web. ACM, 2010. 591–600.

[16] Shafiq M Z, Ilyas M U, Liu A X, et al. Identifying leaders and followers in online social networks. IEEE Journal on Selected Areas in Communications, 2013, 31(9):618–628.

[17] He Q, Chen B, Pei J, et al. Detecting topic evolution in scientific literature: How can citations help? Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009. 957–966.

[18] 王伟. 基于微博数据的电影票房预测研究 [D]. 重庆大学, 2015.

[19] 王晓耘, 袁媛, 史玲玲. 基于微博的电影首映周票房预测建模. 现代图书情报技术, 2016, 32(4):31–39.

[20] 王炼, 贾建民. 基于网络搜索的票房预测模型——来自中国电影市场的证据. 系统工程理论与实践, 2014, 34(12):3079–3090.

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

[22] Asur S, Huberman B A. Predicting the future with social media. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, volume 1. IEEE, 2010. 492–499.

[23] Sawhney M S, Eliashberg J. A parsimonious model for forecasting gross box-office revenues of motion pictures. Marketing Science, 1996, 15(2):113–131.

[24] Bogers T. Movie recommendation using random walks over the contextual graph. Proc. of the 2nd Intl. Workshop on Context-Aware Recommender Systems, 2010.

[25] Mukherjee R, Sajja N, Sen S. A movie recommendation system–an application of voting theory in user modeling. User Modeling and User-Adapted Interaction, 2003, 13(1):5–33.

[26] Lin C J, Wu W W. A causal analytical method for group decision-making under fuzzy environment. Expert Systems with Applications, 2008, 34(1):205–213.

[27] Bauer H H, Huber F, Bräutigam F. Method supplied investigation of customer loyalty in the automotive industry—results of a causal analytical study. Customer retention in the automotive industry. Springer, 1997: 167–213.

[28] Gießmann M. Complexity management in logistics. causal analytical study on the influence of procurment complexity on the logistics success, 2010.

[29] baidu. 工具变量法 [EB/OL]. [2017-04-23]. http://baike.baidu.com/item/%E5%B7%A5%E5%85%B7%E5%8F%98%E9%87%8F%E6%B3%95.

[30] Stock J H, Trebbi F. Retrospectives: Who invented instrumental variable regression? The Journal of Economic Perspectives, 2003, 17(3):177–194.

[31] 陈云松. 逻辑, 想象和诠释: 工具变量在社会科学因果推断中的应用. 社会学研究, 2012, 6:192–216.

[32] Arellano M, Bover O. Another look at the instrumental variable estimation of error-components models. Journal of econometrics, 1995, 68(1):29–51.

[33] Nelson C, Startz R. Some further results on the exact small sample properties of the instrumental variable estimator, 1988.

[34] Cragg J G, Donald S G. Testing identifiability and specification in instrumental variable models. Econometric Theory, 1993, 9(02):222–240.

[35] 陈林, 朱卫平. 中国地区性行政垄断与区域经济绩效——基于工具变量法的实证研究. 经济社会体制比较, 2012, (4):195–204.

[36] 方颖, 赵扬. 寻找制度的工具变量: 估计产权保护对中国经济增长的贡献. 2011..

[37] 陈昊. 出口贸易与学历误配: 缓解还是加剧?——基于多工具变量法的经验研究. 财经研究, 2014, 3:42–51.

[38] 余静文, 王春超. 新“拟随机实验”方法的兴起——断点回归及其在经济学中的应用. 经济学动态, 2011, (2):125–131.

[39] Imbens G W, Lemieux T. Regression discontinuity designs: A guide to practice. Journal of econometrics, 2008, 142(2):615–635.

[40] Thistlethwaite D L, Campbell D T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational psychology, 1960, 51(6):309–317.

[41] Lee D S, Lemieuxa T. Regression discontinuity designs in economics. Journal of economic literature, 2010, 48(2):281–355.

[42] 王骏, 孙志军, et al. 重点高中能否提高学生的学业成绩–基于 f 县普通高中的一个断点回归设计研究. 北京大学教育评论, 2015, (2015 年 04):82–109.

[43] Zhang C. 结婚年龄与婚姻的稳定性: 来自断点回归的证据. 2012..

[44] 张建同, 方陈承, 何芳. 上海市房地产限购限贷政策评估: 基于断点回归设计的研究. 科学决策, 2015, (7):1–23.

[45] 席鹏辉, 梁若冰. 空气污染对地方环保投入的影响——基于多断点回归设计. 2015..

[46] baidu. 格兰杰因果关系检验 [EB/OL]. [2017-04-23]. http://baike.baidu.com/item/%E6%A0%BC%E5%85%B0%E6%9D%B0%E5%9B%A0%E6%9E%9C%E5%85%B3%E7%B3%BB%E6%A3%80%E9%AA%8C/2485970.

[47] Hiemstra C, Jones J D. Testing for linear and nonlinear granger causality in the stock price-volume relation. The Journal of Finance, 1994, 49(5):1639–1664.

[48] 游和远, 谭术魁, 林宁. 基于格兰杰因果关系检验模型的地价与房价关系分析 — 对深圳市的实证研究. 蘭州工業高等專科學校學報, 2007, 14(1):54–58.

[49] Yang H Y. A note on the causal relationship between energy and gdp in taiwan. Energy economics, 2000, 22(3):309–317.

[50] Soytas U, Sari R. Energy consumption and gdp: causality relationship in g-7 countries and emerging markets. Energy economics, 2003, 25(1):33–37.

[51] 庞皓, 陈述云. 格兰杰因果检验的有效性及其应用. 统计与决策, 1999, (9):17–19.

[52] 孔凡文, 才旭, 于淼. 格兰杰因果关系检验模型分析与应用. 瀋陽建築大學學報 (自然科學版), 2010, 26(2):405–408.

[53] Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects. Biometrika, 1983. 41–55.

[54] 胡安宁. 倾向值匹配与因果推论: 方法论述评. 社会学研究, 2012, 1:221–242.

[55] 周振, 牛立腾, 孔祥智. 户籍歧视与城乡劳动力工资差异——基于倾向值的匹配分析. 区域经济评论, 2014, (4):122–130.

[56] 侯珂, 刘艳, 屈智勇, et al. 留守对农村儿童青少年社会适应的影响: 倾向值匹配的比较分析. 心理发展与教育, 2014, 30(6):646–655.

[57] 张紫薇, 柯佑祥. 大学本土留学教育个人收益的计量分析: 基于倾向值匹配法的研究. 教育与经济, 2016, (1):33–38.

[58] Jalan J, Ravallion M. Estimating the benefit incidence of an antipoverty program by propensity-score matching. Journal of Business & Economic Statistics, 2003, 21(1):19–30.

[59] Seeger J D, Williams P L, Walker A M. An application of propensity score matching using claims data. Pharmacoepidemiology and drug safety, 2005, 14(7):465–476.

[60] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. science, 2011, 334(6062):1518–1524.

[61] 赵龙文, 公荣涛, 陈明艳, et al. 基于意见领袖参与行为的微博话题热度预测研究. 情报杂志, 2013, 32(12):42–46.

致 谢

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1989 年 8 月 12 日出生于辽宁省阜新市。

2008 年 9 月考入清华大学自动化系，2012 年 7 月本科毕业并获得工学学士学位。

2014 年 9 月免试进入清华大学计算机科学与技术系攻读硕士学位至今。

发表的学术论文

[1] Ding Xu, Tao Pin. Synchronization Detection of Multicamera System based on LED Matrix. International Conference on Embedded Software and Systems, 2016

[2] 丁旭, 陶品. 基于 LED 点阵的摄像机拍摄时间检测方法. Conference on Pervasive Computing of China, 2016