

北京交通大学

硕士学位论文

电商网红社交行为商业影响的分析与预测

Analyzing and Predicting Promotion Effects
of E-commerce Celebrities' Social Behavior

作者：盛烨

导师：郭宇春

北京交通大学

2019年6月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：



签字日期：2019年5月31日

导师签名：



签字日期：2019年5月31日

学校代码：10004

密级：公开

北京交通大学

硕士学位论文

电商网红社交行为商业影响的分析与预测

Analyzing and Predicting Promotion Effects
of E-commerce Celebrities' Social Behavior

作者姓名：盛烨

学 号：16120116

导师姓名：郭宇春

职 称：教授

学位类别：工学

学位级别：硕士

学科专业：通信与信息系统

研究方向：信息网络

北京交通大学

2019年6月

致谢

本论文的研究工作是在我的导师郭宇春教授的悉心指导下完成的。郭宇春教授科学的工作方法、开阔的视野以及渊博的知识给了我极大的帮助。郭宇春教授严谨的治学态度、精益求精的学术风范、恪尽职守的工作作风以及豁达的人生理念，深深的感染和激励着我不断进取，对我以后的工作和学习有着很大的影响。在此衷心感谢多年来郭宇春教授对我的悉心指导和关怀。

同样感谢陈一帅老师对本次毕设的帮助和指导，在读研期间，多次和陈老师一起修改代码、讨论论文，每次交流都能产生新的想法。陈老师在工作生活中饱满的热情以及负责的态度将会一直是我学习的榜样。

感谢实验室里的所有老师。衷心感谢赵永祥老师、李纯喜老师、郑宏云和张立军老师在我研究生学习阶段对我的无私帮助和关怀，我所有的科研成果都凝结着各位老师的辛勤汗水。在此向各位老师表示诚挚的谢意。

另外，在实验室工作和撰写论文期间，王一师姐、李梦月师兄、李俊峰师兄、刘翔、陈滨、唐伟康、宋云鹏、韩致远等同学对我的研究工作给予了热心帮助，在此向他们表示我的感谢之意。还要诚挚的感谢国家自然科学基金(No. 61572071, 61271199, 61301082)的资助。

最后，特别感谢一直无微不至的关心、支持我的父母和爱人，正是他们热情的鼓励和默默的奉献，才使得我顺利的完成学业，成为社会的有用之才。

摘要

电商网红是指使用电商平台进行经营活动，并且借助社交平台开展营销活动的“网络红人”。基于电商网红的商业模式的经济规模在 2016 年已经达到了 580 亿，超过了当年的中国电影票房。发现和研究电商网红在社交平台上的典型行为模式，挖掘电商平台与社交平台的潜在联系，以深入理解这种新的跨平台的经济现象，从而有助于评估电商网红自身的商业价值、改进社交营销方式。

目前，对电商网红的研究仍停留在商业案例和定性概念层面讨论上，缺乏基于规模化的实际网络测量的量化分析。为此，本文以新浪微博和淘宝电商平台为具体测量对象，进行了跨平台地测量和数据融合，建立了电商网红原始数据集；基于该数据集，分析了电商网红在社交平台和电子商务平台上的社交商业行为和经营活动，建立电商网红商业价值模型，评估和预测电商网红社交行为商业价值。

本文的工作主要有如下几个方面：

(1) 编写网络爬虫对微博和淘宝平台进行测量，形成电商网红社交平台数据集和相应电商平台销量数据集。该数据集包括 108 位网红在微博平台共 8 年零 4 个月的社交数据及其 2018 年 4 月的淘宝销量数据，和被网红点赞和转发过的 46470 位微博用户的基本信息数据。

(2) 首次构建了刻画网红营销行为的特征工程。基于电商网红社交数据集提取并分析了电商网红的三种典型营销行为，包括广告行为、促销行为和口碑营销行为。挖掘三种行为的统计规律，基于行为分析构建了电商网红营销行为特征 35 个、日常行为特征 4 个，提取了基本信息特征两个，共计 41 个社交特征。

(3) 首次构建了基于社交特征的电商网红销量水平评估模型。在上述特征工程的基础上，使用随机森林、逻辑回归、kNN 等分类算法，构建电商网红销量模型，发现了最能影响电商网红销量的 10 个特征；实验表明，模型最高精确率可达 0.83。

关键词：社交商业化；电商网红；行为分析；预测；机器学习

ABSTRACT

E-commerce celebrities refers to the "internet celebrities" who use e-commerce platform for business activities and carries out marketing with the help of social network platform. The economic scale of the business model based on e-commerce celebrities has reached 58 billion yuan in 2016, surpassing the box office of Chinese movies in that year. To discover and study the typical behavioral patterns of e-commerce celebrities on social network platforms, tap the potential links between e-commerce platforms and social network platforms, so as to deeply understand this new cross-platform economic phenomenon, thus helping to evaluate the commercial value of e-commerce celebrities and improve its social marketing.

At present, the research on e-commerce celebrities is still at the level of business cases and qualitative concepts, lacking of quantitative analysis based on scale of actual network measurement. For this reason, this paper takes Sina Weibo and Taobao as specific measurement objects, carries out cross-platform measurement and data fusion, and establishes the original data set of e-commerce celebrities. Based on this data set, it analyses the social business behavior and business activities of e-commerce celebrities on social network platforms and e-commerce platforms, establishes the business value model of e-commerce celebrities, evaluates and predicts the business value of e-commerce celebrities' social behavior.

The main work of this paper is as follows:

(1) Write web crawler to measure micro-blog and Taobao platform, and form e-commerce celebrities' data sets of social network platform and sales data sets of e-commerce platform. The data set includes 108 e-commerce celebrities' social data on Weibo for 8 years and 4 months, Taobao sales data in April 2018, and the basic information data of 46,470 Weibo users liked and reposted by e-commerce celebrities.

(2) This is the first time to construct a characteristic project to depict the marketing behavior of e-commerce celebrities. Based on the social data set of e-commerce celebrities, this paper extracts and analyses the marketing behavior of e-commerce celebrities, including advertising behavior, promotion behavior and word-of-mouth marketing behavior. Mining the statistical rules of the three kinds of behavior, based on the behavior analysis, 35 marketing behavior characteristics and 4 daily behavior characteristics of e-commerce red net are constructed, and two basic information characteristics are extracted, a total of 41 social characteristics.

(3) For the first time, an evaluation model of e-commerce celebrities bonus sales level based on social features is constructed. On the basis of the above-mentioned feature engineering, a red sales model of e-commerce network is constructed by using random forest, logistic regression, kNN and other classification algorithms, and 10 features that can most affect the sales of e-commerce celebrities are found. The experiment shows that the highest accuracy of the model can reach 0.83.

KEYWORDS: social commercialization; e-commerce celebrity; behavior analysis; prediction; machine learning

目录

摘要	v
ABSTRACT.....	vii
1 引言	1
1.1 研究背景及意义	1
1.2 国内外研究现状	2
1.2.1 社交媒体与社交媒体营销	2
1.2.2 “社交网络+电子商务”模式.....	3
1.2.3 社交网络用户行为分析	4
1.3 本论文的主要研究内容	5
1.4 本论文的主要贡献	5
1.5 本论文的组织结构	6
2 技术背景	7
2.1 在线社交网络——微博	7
2.2 电子商务平台——淘宝	7
2.3 机器学习介绍	8
2.3.1 分类和回归	8
2.3.2 多分类问题的拆分	8
2.3.3 多分类问题机器学习性能评估方法	9
2.4 机器学习算法	11
2.4.1 随机森林算法	11
2.4.2 逻辑回归算法	12
2.4.3 k 近邻算法	13
2.5 开发平台	13
2.5.1 Anaconda 简介	13
2.5.2 Scikit-Learn 库	14
2.6 爬虫技术	14
2.7 本章小结	15
3 电商网红社交与销量数据的获取	17
3.1 问题描述	17

3.2	电商网红名单获取	17
3.3	基于 Python 爬虫的社交数据获取	19
3.3.1	自动获取游客 cookie	19
3.3.2	用户基本信息获取	22
3.3.3	用户原创微博与转发微博获取	24
3.3.4	用户点赞微博获取	27
3.4	基于 Web Scraper 的销量数据获取	28
3.5	采集成果与数据字段介绍	31
3.6	数据预处理	31
3.7	本章小结	32
4	电商网红营销行为测量分析	33
4.1	问题描述	33
4.2	电商网红广告行为分析	33
4.2.1	上新微博内容分析	34
4.2.2	上新微博比例分析	35
4.2.3	上新间隔分析	36
4.3	电商网红促销行为分析	39
4.3.1	抽奖微博内容分析	39
4.3.2	抽奖偏好分析	40
4.4	电商网红口碑营销行为分析	41
4.4.1	转发点赞内容分析	41
4.4.2	转发点赞行为倾向分析	42
4.5	本章小结	46
5	电商网红销量评估模型	47
5.1	问题描述	47
5.2	电商网红有效社交网络定义	47
5.3	营销行为特征量的构建	48
5.3.1	上新行为特征量	48
5.3.2	抽奖行为特征量	48
5.3.3	转发点赞行为特征量	49
5.4	日常行为特征量的构建	52
5.5	社交特征量分析	52

5.5.1	基于聚类的电商网红模式分析	53
5.5.1	社交特征量与销量的相关性分析	55
5.6	电商网红淘宝销量预测模型	58
5.6.1	模型构建	58
5.6.2	模型性能评估	59
5.7	本章小结	62
6	结论	65
6.1	本文工作总结	65
6.1.1	电商网红社交与销量数据的获取	65
6.1.2	电商网红营销行为测量与分析	66
6.1.3	电商网红销量评估模型	66
6.2	未来工作展望	67
	参考文献	69

1 引言

1.1 研究背景及意义

当今社会，在线社交网络与人们的生活息息相关，成为人们获取信息、与他人沟通的主要方式之一。随着社交技术的发展，社交网络平台已经形成巨大的用户规模，这使得社交网络内的信息能以很低的成本进行快速、广泛的传播。这一优势使得很多企业发现社交网络平台可以成为一种新的营销工具，他们在社交网络平台上发布与企业、产品或服务相关的各种信息，以求得到更好的推广效果。同时，社交网络自然而然地将相同兴趣的用户聚集在一起，方便了企业进行精准营销。诸如此类的将社交网络优势转化为商业利润的探索是社交网络研究中的新兴方向，在线社交网络在市场驱动下，开始向商业化进行改革。近十年，电子商务在中国消费市场中迅速崛起，消费者的消费习惯从线下消费迅速转变为在线消费，社交网络开始与电子商务结合，催生了社交网络与电子商务互相融合的模式的出现。2013年4月29日，中国最大的社交网络平台新浪微博与中国最大的电子商务平台阿里巴巴宣布签署战略合作协议，两大平台将在数据交换、网络营销等领域合作，旨在建立更活跃的社交网络平台，为消费者带来能够产生有效互动的社会化电商模式^[1]。

电商网红正是“社交网络+电子商务”模式探索中的既得益者。一部分淘宝个体商户从企业的社交媒体营销中获得经验，在社交网络平台上开展私人店铺营销，以个人魅力与店铺口碑吸引社交网络平台粉丝，进而转化为店铺粉丝、扩大店铺规模。另一方面，一些名人利用已有粉丝规模，通过电商平台形成粉丝量向商业利润的转换。这两类人群都在“社交网络+电子商务”模式多年的探索中创造了巨大的经济利益，根据2016年CBNData的统计，电商网红产值预估达580亿，超过2015年中国电影总票房。2017年3月23日，微博宣布推出网红电商平台，促成网红、企业和服务商之间的资源共享，并加大对网红和企业的运营扶持^[1]。电商网红的成功使得投资人、电商平台和社交平台更加看好网红群体。

然而，电商网红只是社交商业化进程中的一个阶段性产物，他们的成功预示着社交商业化将对各行各业开始潜移默化地改革。只有深入了解电商网红的商业模式，才能将这种模式在其他行业场景下进行复制甚至改进。电商网红的商业模式由其社交平台 and 电商平台的商业行为共同决定，尤其是社交平台上的行为具有很高的营销价值。所以观测和分析电商网红的社交行为，对社交商业化的普及有着深远的意义。

对于电商网红或者其他企业来说，与消费者的互动、沟通等社交活动是一套复

杂的行为组合，社交活动与商业利润的具体关系目前难以衡量。想要确定社交活动带来的商业效果并优化社交环节，就必须分析社交过程中每个行为特征与经济效益的关系。这需要研究者对电商网红的社交行为进行深入地量化分析。

另一方面，网红经济公司也可能是社交商业化时代的受益者。对他们来说，批量培养网红并获得持续成功的难度非常大，目前还没有一个完整的挖掘、培养、孵化网红体系。如何对网红商业价值进行衡量、如何将资源进行合理分配、如何培养网红，这些都成为网红经济公司目前必须解决的难题。为了解决这个问题，研究者需要建立一个社交数据与网红商业价值之间的模型。

针对上述问题，本文研究电商网红社交特征对其销量的影响。第一个研究内容是测量和获取电商网红的社交网络数据与商业数据。第二个研究内容是提取和分析电商网红在社交平台上的广告、促销和口碑营销等行为，挖掘这些营销行为规律。第三个研究内容是基于行为分析，从网红的社交行为数据中构建行为特征，结合网红的个人基本信息特征，共同建立一个模型来分析网红的社交特征与网红销量水平的关系。这些研究不仅有助于各行各业理解电商网红模式，而且有助于电商网红和网红经济公司对自身进行精准定位，为社交优化提供指导性建议。

1.2 国内外研究现状

本研究讨论基于社交网络平台微博和电子商务平台淘宝联合运营模式下的电商网红，她们在淘宝上定期售卖商品，在微博上开展营销活动，吸引新客户的同时并维护老客户。淘宝和微博分别是中国最著名的电子商务平台和社交平台，所以本节首先总结社交媒体营销的相关研究现状与发展方向，接着对“社交网络+电子商务”模式的产生与演进进行详细叙述，最后回顾社交网络用户行为分析的各类研究成果，总结现有研究的不足，并提出本文可改进的方向。

1.2.1 社交媒体与社交媒体营销

美国公共关系协会将社交媒体（Social media）定义为：支持互联网的相关服务和技术。文献[2]中提出，社交媒体是一种工具和平台的统称，人们通过使用社交媒体来感知世界，并按照主观意愿分享各种信息。

信息化时代下，在线社交媒体为信息传播做出了巨大的贡献，越来越多的学者意识到社交媒体的巨大优势，投身于社交媒体营销的研究中。文献[3]中提出了一个营销核心理论：社交媒体有效连接了企业与用户。文献[4]对品牌口碑与社交媒体营销的相关影响进行了研究。最终发现同时拥有营销者和消费者两种角色的用

户对口碑营销的效果影响非常突出。研究[5-7]指出, 社交媒体用户的社交网络属性与网络中的节点关系会影响社会口碑传播的感知价值。在文献[8]的研究中论述了品牌与社交媒体的关系需要很多个步骤才能稳固建立。文献[9]中提到, 企业原创内容对社交媒体用户的购买倾向影响更大。但是, 一些学者还指出, 社交传播目前仍有不足。文献[10]中指出, 社交传播比大众传播有影响力, 但是并不如传统的媒体有效果。文献[11]还发现社交媒体在吸引新用户和推动品牌的推广方面, 不如传统营销有效。文献[12]给出了关于这个问题的一个解释: 消费者偏爱于时尚类品牌互动, 较少与消费品品牌互动, 这导致了社交媒体影响力的局限性。另一方面, 文献[13]还发现无论社交媒体营销效果的好坏, 消费者都仍然受到其影响。这意味着消费者社交网络有很大的价值, 它正是企业品牌的产品信息网络的一个补充。

在我国, 近几年社交媒体出现了爆发式增长, 吸引了不少学者的注意。其中, 文献[14]基于社交、认知和享乐三个目的划分社交媒体营销, 并以品牌认知、品牌溢价和品牌市场划分了企业资产, 对社交媒体营销与企业资产的关系提出假设。文献[15]通过实验解释了消费者参与“微信红包”活动时, 参与欲望和抗拒感共同作用的特殊机制, 给企业的类似社交活动实践提供了参考。文献[16-21]都研究了微信营销对微信用户主力军——中学生及大学生的影响。

上述研究成果的出现, 意味着企业已经意识到微博和微信这样的社交媒体都可以成为营销的有效工具。品牌的社交营销理论研究和实证研究都比较充分, 基于社交的商业化已经渐渐渗透到每一个社交媒体中。

1.2.2 “社交网络+电子商务”模式

2005年 Yahoo!公司最早提出了“社交化电子商务”这一概念, 随后几年国内外学者们开始了社交网络与电子商务相互融合这一模式的研究。一部分学者认为电子商务是该模式的主导, 即社交化电子商务。例如文献[22]认为社会化电商是以社交网络为传播途径, 社交行为位营销的辅助的电商模式。文献[23-24]认为消费者对产品的在线评价反馈是社交化电商的核心。而另一些学者则更加重视社交媒体, 例如文献[25]认为社交技术使得用户与商家产生了更好的沟通, 用户体验的提高促进了电商的发展。文献[26-27]通过不同的切入点解析社交网络用户的行为, 从而分析出用户社交网络传播动机的类型, 将用户资源网与电商资源网结合, 优化了电商营销成本分配。

一些学者对我国社交网络与电子商务相互融合的模式的发展现状进行分类。文献[28-29]将国内的“社交网络+电商模式”下的所有网站根据业务的侧重点不同, 划分为以下三种: (1)第三方独立导购类网站; (2) 基于社交化媒体的社交化电子商

务，电子商务平台自行通过社交媒体展开营销；(3) 基于社交化媒体的社交化电子商务，社交网络平台为电子商务平台赋能，使平台上的企业享受到高效的营销服务。

上述研究成果从理论层面，通过分析个例的方式考察了社交网络+电子商务模式的现状与未来发展。目前，还没有研究从数据科学的角度观测电商平台的社交化或是社交平台为电商平台的赋能，将是本研究的研究重点。

1.2.3 社交网络用户行为分析

复杂网络一直是近十年来国内外学者关注的热点问题之一。复杂网络可以分为拓扑研究和用户行为研究。其中，社交网络与复杂网络相似，在线社交网络更有其特点。国内外对于在线社交网络上的用户行为研究很多，下面本文将对不同的研究方向一一叙述。

第一类是基于时间序列的用户行为特征建模研究。文献[30-31]指出用户浏览网站和邮件通信的时间间隔是非泊松性的。在此基础上，文献[32-34]又提出了截止时间的用户行为动力学模型和用户行为模式实证。

第二类是非时间序列的用户行为特征分析。文献[35-37]发现社交网络中普遍存在着幂律分布，文献[38]指出了人类活动的幂律分布与网络度的幂律分布的相关性。但是文献[39]认为行为周期性才是造成幂律时间间隔的根本原因。

第三类是基于用户特征的行为分类分析。这一类的研究成果非常多，具有代表性地主要有：文献[40-41]对拥有时间特性的用户行为进行了详细并且深入的研究，发现用户的行为特征规律。基于这些特征分析，文献[42-46]在微博和 Twitter 上对时间特性用户行为进行了建模。文献[47]对 YouTube 上的用户使用 K-Means 方法选取了 9 个特征，并将用户分为五个类别。文献[48]研究 Twitter 社交网络上不同用户的行为，地域和社交网络规模，最终将他们分为了三个类别。文献[49]运用了不同的模型预测 Twitter 用户的转发行为，特征分析在模型建立中起到关键性的作用。文献[50]中以实验证明分类算法可以用于微博转发行为的预测。文献[51-52]还给出了社交网络用户行为的实证分析。

综合以上，对于社交网络用户行为的研究中，几乎所有研究都是针对整体行为，然而，一些个体行为特征值得被研究却仍未受到关注。电商网红作为一种新的、社交化的电商营销模式下产生的新群体，有研究价值。所以本文对于电商网红这个局部群体进行行为分析。

1.3 本论文的主要研究内容

根据上一节的介绍可知，对电商网红的研究大多以理论分析为主，研究方法以采访、问卷、分析个例的方式为主，使用多种营销理论对电商网红商业模式进行分析，并没有给出可量化的规律。另一方面，许多研究电商网红商业价值的模型将网红的商业价值等价于网红的影响力与传播能力，这样的定义不够准确也不够直接，不能突出电商网红模式的特点。所以本论文的研究内容有以下三点：

一是测量和获取电商网红数据集。由于之前与电商网红相关的研究仍停留在案例研究和概念层面讨论上，缺乏基于实际网络测量数据的定量分析。所以本文综合各类数据报告与微博博主经验总结，尽可能多地列举符合研究条件的电商网红。再利用爬虫技术，获得电商网红的社交数据与淘宝销量数据，构建电商网红数据集。

二是可量化地探究电商网红的商业模式。想要了解电商网红的商业模式，有很多问题需要解决。例如电商网红有哪些商业行为？电商网红早期的行为与近期的行为相比有什么变化呢？他们如何与粉丝实现高效沟通与互动？如何调动粉丝、发挥粉丝的自媒体价值？如何激发消费者的购买动力？然而由于之前的研究几乎没有对大量电商网红进行过系统研究，所以本文基于采集到的电商网红数据集，用数据挖掘的方法从多个尺度分别定量分析了电商网红的商业行为规律。

三是如何预测电商网红的销量。建立模型前，还需要考虑网红的哪些社交行为特征更能影响其发展？网红的商业行为与非商业行为都有哪些价值？如何挖掘深层次社交行为特征来优化网红商业能力模型？使用什么样的模型能够达到最好的效果？由于数据集的稀缺，目前没有研究使用数据科学对这些问题进行系统地研究，所以本文通过机器学习与数据挖掘的方法，基于电商网红在社交网络上的社交行为数据，提取网红的商业行为特征与非商业行为特征预测电商网红的销量水平，探究商业行为与非商业行为对销量的影响。

1.4 本论文的主要贡献

上一节讨论了电商网红研究中存在的诸多缺陷，本文将通过以下方法解决问题并作出如下贡献：

(1) 本论文通过爬虫的方法测量和获取了电商网红数据集，该数据集包括 108 位电商网红的淘宝销量数据和 2010 年 1 月到 2018 年 4 月的微博社交数据、电商网红淘宝销量数据以及电商网红的有效社交网络中用户的基本信息数据。

(2) 本论文基于采集到的数据集，挖掘电商网红在社交平台上的营销行为，探索其商业模式。首先观察了电商网红的上新行为，上新行为指电商网红新商品开售

前的预热行为。发现所有网红的上新行为间隔围绕着 28 天上下小幅度波动，呈周期性。其次观察了电商网红的抽奖行为偏好，发现电商网红的抽奖行为往往伴随着上新行为，电商网红更偏向转发形式的抽奖，且转发抽奖与淘宝销量的相关性为 0.39。最后，本论文观察了电商网红的口碑营销行为，分析了网红转发点赞的倾向，发现电商网红更愿意大量转发点赞低粉丝量用户的微博，低粉丝量用户为网红提供的买家秀可以为网红创造更大的商业价值。

(3) 本论文基于电商网红社交行为数据，构建了电商网红的营销行为特征 35 个、日常行为特征 4 个，基于个人信息数据提取了个人基本信息特征两个，总计 41 个社交特征。通过聚类分析及特征量与电商网红淘宝销量相关性分析证实单一特征或少量建模不可行，必须使用大量特征建模。所以本文基于机器学习中的分类算法，使用 41 个社交特征对网红淘宝销量水平进行预测。本文发现使用全部特征的模型性能最优，精确率可达 0.76；使用近期的社交数据会比使用全局数据达到更好的模型性能，模型最高精确率最高可达 0.83。

(4) 本论文通过随机森林算法得到对销量影响最大的 10 个社交特征，并据此结果对电商网红社交营销优化提出建议。

1.5 本论文的组织结构

本文接下来的组织结构如下。

第二章将详细介绍本文所使用的各种技术的原理及应用方式，主要包括机器学习、爬虫算法、开发平台等。

第三章将通过基于 Python 和 Web Scrape 的爬虫方法，测量和获取了电商网红的微博社交数据集与淘宝销量数据集。

第四章将详细分析商网红商业行为特点。首先，对探索网红商业行为分析问题进行分析，包括从电商网红微博内容中提取电商网红的营销行为：上新行为、抽奖行为和转发点赞行为，分析这三种行为的规律，初步探索商业行为背后的商业价值。

第五章将详细介绍电商网红的社交数据与销量水平的建模问题。首先，对电商网红的销量水平建模问题进行描述。其次，从微博社交数据中构建了营销行为特征量、日常行为特征量，从微博个人信息数据中提取了电商网红个人基本信息特征量，通过聚类分析这些特征量整体的分布情况，通过相关性分析得到每个特征量与电商网红销量之间的关系。最后，基于机器学习中的分类算法，使用电商网红的特征预测其淘宝销量水平。

第六章将对整篇论文进行全面的总结和概括，列举本文的主要贡献，然后对未来的研究工作进行展望。

2 技术背景

本章将主要介绍本论文中用到的主要技术背景。首先，介绍在线社交平台微博与电子商务平台淘宝；然后，介绍机器学习的基本应用场景和性能评估方法；然后，对文中研究涉及的几种机器学习算法的原理进行简要介绍；最后介绍本研究的开发平台。

2.1 在线社交网络——微博

社交网络即互联网社交网络服务，旨在帮助公众建立社会性网络的互联网应用服务。2006年美国的 Twitter 出现在人们眼前，越来越多的人开始使用这种叫“微博”的信息分享和获取方式。微博是一种社交媒体也是一个网络平台，它将各种通信网络、终端设备整合再生。用户在这个平台上通过单向关注关系就可以获取他人发布的文字、图片、视频等多媒体信息，还可以转发，评论和点赞他人的信息。微博实现了信息的即时分享与传播互动。

2009年8月新浪推出“新浪微博”内测版，成为中国门户网站中第一家提供微博服务的网站。随着微博在网民中越来越火热，还出现了腾讯微博，网易微博，搜狐微博等。2013年新浪微博成为中国用户规模最大的微博，此后各门户网站微博陆续关闭。所以，如若没有特别说明，微博就是指新浪微博。

2.2 电子商务平台——淘宝

电子商务平台是一个为企业和个人提供在线交易的平台。电商平台充分利用了因特网无时空限制的优势，建立一个业务发展的框架系统，提供网络资源、安全保障、安全的网上支付，完成了资金流与物流，实现了资源共享。电商平台分为 B2B 平台，C2C 平台，O2O 平台等，其中中国主要的电商平台有淘宝、京东商城和拼多多等。

淘宝网是亚太地区较大的，包括 C2C、团购、分销、拍卖等多种商务模式在内的综合性零售商圈，由阿里巴巴集团于 2003 年 5 月创立。淘宝网是中国深受欢迎的网购零售平台，拥有注册用户数近 5 亿，每天有超过 6000 万的日均固定访问量，同时日均在线商品数也已经超过了 8 亿件，平均每分钟可以交易 4.8 万件商品。到 2011 年底，淘宝网单日交易额峰值达到 43.8 亿元^[53]，成为中国当之无愧的电商巨头。

2.3 机器学习介绍

近几年来,人工智能遍布人们的生活之中,手机、智能音箱、智能汽车等等产品都使用人工智能改变着人们的生活方式。人工智能通过计算机技术试图让机器模仿人类,作出与人类智能相似的判断与反应。随着大数据环境的形成与计算机计算能力的增强,人工智能越来越流行。人工智能技术主要有机器学习、数据挖掘、智能算法、模式识别和专家系统等,本文主要用到机器学习技术。机器学习是一门多领域交叉学科,涉及统计学、概率论和算法复杂度等多门学科,它利用大量数据所构成的经验模拟人类的学习行为,主要用于归纳和综合。机器学习可以分为分类问题、回归问题和结构化预测问题,下面将对本文中用到的分类问题做详细介绍。

2.3.1 分类和回归

机器学习中的回归问题指输入变量与输出变量为连续变量的预测问题,那么,输出变量为有限个离散变量的预测问题为分类问题。从输出是否连续不难看出,回归的目的是找到最优拟合,而分类的目的是寻找决策边界。本研究主要解决的是一个分类问题。

分类问题按照输出空间的多少,分为二元分类和多元分类。二元分类是指将一个样本划分到两个类别的集合中,非 A 类即是 B 类。这种二元分类在现实生活中应用十分广泛,例如判断用户是否为女性、用户的情感倾向是积极还是消极和明天是否是晴天等等。多元分类是二分类问题的一种扩展,一个样本最终会被划分成多个类别。这类问题在生活中也时常遇到,例如判断动物的品种、判断用户的年龄层等等。常见的机器学习分类算法有:逻辑回归、朴素贝叶斯、支持向量机、决策树、随机森林、k 近邻和神经网络等。其中随机森林、逻辑回归和 k 近邻在本文的研究中将会用到,下节将会做出详细介绍。

2.3.2 多分类问题的拆分

现实中遇到多分类学习任务,虽然有些二分类学习方法可以直接推广到多分类,但更多情况下需要使用一些策略,利用二分类器来解决多分类问题。考虑 N 个类别 $K_1, K_2, K_3, \dots, K_n$, 需要将这个多分类问题拆解为若干个分类任务,首先需要将问题进行拆分,接着为每个拆分出来的二分类任务分别训练分类器,最后对所有二分类器的预测结果进行集成,获得最终多分类的结果。该过程中有两个关键:第一是如何拆分多分类任务,第二是多个二分类器的结果如何集成。下面将主要介绍

多分类学习的三种拆分策略：“一对一”，“一对其他”和“多对多”^[54]。

“一对一”(OvO)将 N 个类别分别两两配对，从而产生 $N(N-1)/2$ 个二分类任务。例如OvO将为 K_2 类和 K_3 类训练分类器，那么这个分类器要把训练集中 K_2 类的样本作为正类， K_3 类的样本作为反类，其他类别同理。训练完所有的二分类器后，测试集的新样本将同时进入所有的二分类器，于是将得到 $N(N-1)/2$ 个分类结果，最终将结果集成，可以选择投票法^[55]，把样本被预测地最多的类别作为最终多分类的结果。

“一对其他”(OvR)只将一个类的样本作为正例、其余所有样本作为反例来训练，从而产生了 N 个分类器。训练完所有的二分类器后，测试集的新样本将同时进入 N 个二分类器中，若仅有一个分类器预测为正类，则对应的二分类器类别作为多分类的结果输出；若有多个分类器都预测为正类，则选择置信度最大的类别作为多分类的最终结果。不难看出，OvR需要训练的二分类器数量比OvO少，因此OvR的存储开销与测试时间开销比OvO少。但OvO的分类器在训练时只使用到了两类样本，所以在类别非常多的情况下OvO的训练时间更少。这两种拆分方法的预测性能相当，主要取决于样本数据分布情况。

OvO和OvR都是“多对多”(MvM)的特例。MvM每次将若干个类作为正类，若干个类作为反类，一共做 M 次划分。 M 个分类器分别对测试集进行预测，这些预测标记组成一个编码，将该编码与每个类别的编码分别比较，返回其中距离最小的类别作为最终预测结果。

2.3.3 多分类问题机器学习性能评估方法

本部分将主要介绍本文中用到的机器学习模型中分类模型的性能评估方法，主要介绍本文采用的预测模型的性能评估指标：精确率、召回率。

上一节中提到，多分类问题是二分类的拓展，那么多分类的性能评估指标也可以从二分类指标中引申。首先阐述二元分类问题的评估方法，假设正类样本为1，负类样本为0，那么将真实类别和预测类别的结果显示在下表中，就得到了该二分类预测问题的混淆矩阵，如表2-1所示。

根据上面的混淆矩阵，定义每个矩阵内每个单位的数据含义如下：

- (1) 真正例(True Positive, TP): 真实类别为1，也被预测成为1类的样本个数。
- (2) 假负例(False Negative, FN): 真实类别为1，却被预测成为0类的样本个数。
- (3) 假正例(False Positive, FP): 真实类别为0，却被预测成为1类的样本个数。
- (4) 真负例(True Negative, TN): 真实类别为0，也被预测成为0类的样本个数。

表 2-1 二元分类问题的混淆矩阵

Table 2-1 Confusion matrix for binary classification problem

	预测类别		
		1	0
真实类别			
	1	<i>TP</i>	<i>FN</i>
	0	<i>FP</i>	<i>TN</i>

通过以上几个概念，可以计算分类模型各个性能指标如下：

(1) 精确率(precision):

$$precision = TP / (TP + FP) \quad (2-1)$$

(2) 召回率(recall):

$$recall = TP / (TP + FN) \quad (2-2)$$

在多分类问题中，同样可以使用混淆矩阵来计算每个分类的精确率与召回率。假定有三个类别，分别标记为 0,1 和 2，那么本文使用到的三分类器的混淆矩阵如表 2-2 所示：

表 2-2 三分类器的混淆矩阵

Table 2-2 Confusion matrix for triple classification problem

	预测类别			
		0	1	2
真实类别				
	0	<i>A</i>	<i>B</i>	<i>C</i>
	1	<i>D</i>	<i>E</i>	<i>F</i>
	2	<i>G</i>	<i>H</i>	<i>I</i>

0 类的精确率与召回率如下：

$$precision_0 = A / (A + D + G) \quad (2-3)$$

$$recall_0 = A / (A + B + C) \quad (2-4)$$

同理，1 类的精确率与召回率如下：

$$precision_1 = E / (B + E + H) \quad (2-5)$$

$$recall_1 = E / (D + E + F) \quad (2-6)$$

最后，2 类的精确率与召回率如下：

$$precision_2 = I / (C + F + I) \quad (2-7)$$

$$recall_2 = I / (G + H + I) \quad (2-8)$$

2.4 机器学习算法

机器学习通过学习大量数据得到经验从而生成模型的人工智能科学，四种主要的学习方式：无监督式学习、半监督式学习、监督式学习和强化学习。

监督式学习是通过学习一部分有输入输出对应关系的数据，又称带标签的数据，从而生成一个函数。上节提到的分类和回归问题都属于监督式学习问题。常见的监督式学习算法有支持向量机算法、k 近邻算法和随机森林等算法。无监督式学习使用没有明确结果或标签的数据，通过学习过程来挖掘数据的内在规律，从而将数据分出类别。常见的无监督式学习算法有聚类算法，其通过距离相似度将相似的样本将会分为一类。半监督式学习则是使用一部分有标签、一部分无标签的数据集，在训练过程中还需要让模型学习数据的潜在结构，合理分配利用数据来进行预测。强化学习在监督式学习的基础上还加入了模型的反馈，它会根据输入数据和环境的交互反馈结果对下一步的行动和参数进行调整。

本文对电商网红的商业价值进行预测，属于监督式学习问题，文中使用到的监督式学习算法的具体原理如下文所述。

2.4.1 随机森林算法

决策树是一种树形结构，其内部节点代表一个属性测试，每个分支代表一次测试的输出，每个叶节点代表类别，常见的决策树算法有 C4.5、ID3 和 CART。随机森林(Random Forests, RF)算法是通过集成学习的思想将许多棵 CART 决策树组成起来的一种算法，假设在分类问题的场景下，每棵决策树都是一个分类器，那么 N 棵树就会有 N 个结果，与上文介绍的多分类器集成策略相似，随机森林将投票次数最多的类别制定为最终的输出。定义构建随机森林的迭代次数为 t，即累计需要的决策树个数， $t=1,2, \dots, T$ ，输入样本集为 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ 随机森林的构建过程主要有以下几个步骤：

- (1) 对训练集进行第 t 次随机采样，共采集 m 次，采样集 D_t 中得到 m 个样本；
- (2) 用 D_t 训练第 t 个决策树得到 $f_t(x)$ ，在训练决策树的节点时，注意只选择样本特征中的一部分进行训练，从中选出一个最优特征来做左右子树划分，增强模型的泛化能力；
- (3) 对于分类问题，T 个决策树投票数最多的类别作为最终类别。

随机森林算法优点很多：(1) 随机、有放回的挑选 N 个样本和随机挑选 m 个特征使得随机森林不容易陷入过拟合；(2) 在较大或高维度数据上表现很好且不需要降维处理；(3) 可以处理缺省值问题；(4) 可以快速训练同时计算特征参数的重

要性。

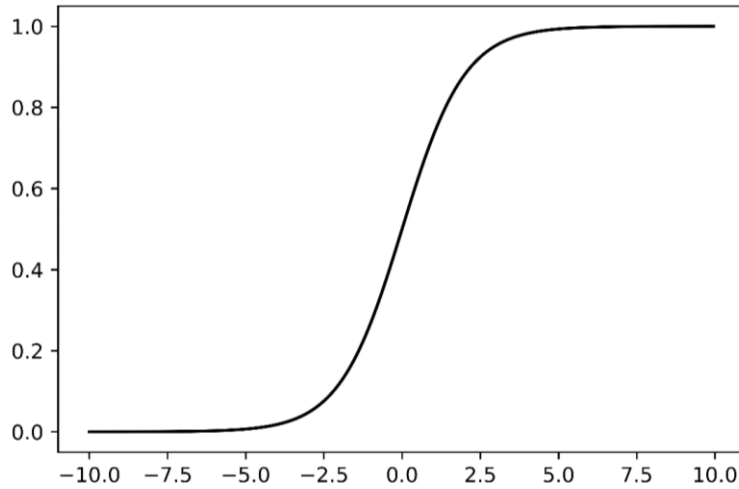


图 2-1 Sigmoid 函数曲线图

Figure 2-1 The curve of Sigmoid function

2.4.2 逻辑回归算法

逻辑回归(Logistic Regression, LR)并不是回归算法,它是一种分类算法。逻辑回归从线性回归演进而来,假设现在有一个连续的因变量 y 和一组自变量 x_1, x_2, \dots, x_n , 可以拟合出一个线性方程 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$, 其中各项系数用最小二乘法估计。但如果因变量变为非连续的二元变量,取值为 0 或者 1, 而方程的右侧是一个取值范围为 $(-\infty, +\infty)$ 的连续值,左右两边无法对应,此时线性回归方程难以解决问题。所以统计学家使用 logistic 函数将自变量取值变换映射到 $(0,1)$ 上, logistic 函数 $y = 1/1 + e^{-x}$ 的图像如图 2-1 所示。这是一个值域为 $(0,1)$ 的 S 型函数,具有无限阶可导的属性。接着,将线性回归方程改写如下:

$$y = 1/1 + e^{-z} \quad (2-9)$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (2-10)$$

下一步对公式 2-10 进行 Logic 变换,得到:

$$\ln(y/1 - y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (2-11)$$

再将 y 看作是 y 取值为 1 的概率,那么 $1 - y$ 就是 y 取值为 0 的概率,所以公式 2-9、公式 2-10 还可以改写为:

$$p(y = 0) = 1/1 + e^z \quad (2-12)$$

$$p(y = 1) = e^z/1 + e^z \quad (2-13)$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2-14)$$

这时使用“最大似然法”估计各个系数。

2.4.3 k 近邻算法

k 近邻(k-Nearest Neighbor, kNN)算法是一种基于特征值间距离的有监督机器学习分类方法，距离越近越容易被分在一类。它的基本思路是：输入测试数据，将测试数据的特征与训练集中对应的特征进行比较，分别计算它们之间的距离，找到最相似也就是距离最近的前 k 个数据，k 小于等于 20，则该测试数据的最终类别为 k 个数据中出现次数最多的那个类别。

kNN 算法通过计算距离来作为相似性指标，避免了对象之间的匹配问题，并且根据 k 个对象中最优的类别进行决策，这两点是 kNN 的优势所在。kNN 算法简单，易于实现，适合用于多分类问题。但 kNN 算法在测试集上计算量大，可解释性差。

2.5 开发平台

本节主要介绍本文研究中使用的开发平台，包括 Anaconda，Scikit-Learn 机器学习库及本研究中使用到的机器学习模块调用方式的介绍和 Python 爬虫。

2.5.1 Anaconda 简介

在使用 Python 时，需要在一个 Python 环境中安装解释器与包集合，在开发中，使用不同版本的 Python 需要复杂的环境变量设置和包管理。Anaconda 是一个包含包管理器和环境管理器的 Python 数据科学发行版本，它的出现轻松解决了上面的各种问题。它集合了 1500 多个开源软件包，在数据可视化、机器学习、深度学习等多方面都有涉及。Anaconda 支持各种操作系统，可以自动配置环境变量，不同版本环境也可以轻松管理。安装包中集成了大部分常用软件包，直接导入便可使用。没有内置在安装包中的其他软件包，可以用 conda 指令或者 pip 指令进行安装，再导入使用^[56]。安装 Anaconda 后菜单栏中出现如图 2-2 所示的界面，本文中的所有实验都是在 Anaconda3.4 环境下完成的。

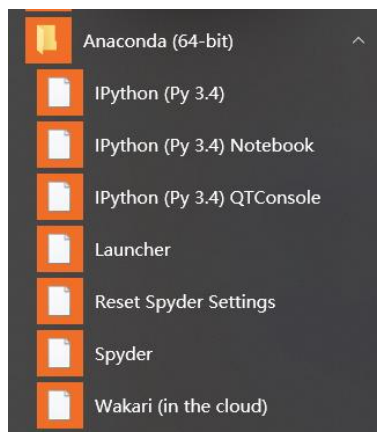


图 2-2 Anaconda 菜单项

Figure 2-2 Menu item of Anaconda

2.5.2 Scikit-Learn 库

Scikit-Learn 是 Python 语言实现的基于 NumPy、SciPy 和 matplotlib 机器学习算法库。它可以实现数据预处理、模型评估与选择、分类、回归、降维等常用机器学习算法^[57]，这些模块都被高度抽象化，用户只需要 3 到 5 行代码就可以完成一个分类器，增加了模型训练的效率。本研究使用 Scikit-Learn 中的数据标准化(StandardScaler())、归一化(Normalizer())、缺失值计算(Imputer())等方法对数据进行预处理。主要使用的 Scikit-Learn 模型有：逻辑回归模型、随机森林模型和 kNN 模型。逻辑回归模型使用 `sklearn.linear_model.LogisticRegression()` 实现，需要调整的参数主要有：惩罚参数(`penalty`)、正则化强度(`C`)、类权重(`class_weight`)、优化算法(`solver`)等。随机森林模型使用 `sklearn.ensemble.RandomForestClassifier()` 实现，需要调整的参数主要有：森林决策树的个数 (`n_estimators`)，划分节点所需最少的样本数 (`min_samples_split`)，树的最大深度 (`max_depth`) 等。kNN 算法模型使用 `sklearn.neighbors.KNeighborsClassifier()` 实现，需要调整的参数主要有：选取的邻居个数(`n_neighbors`)，计算近邻的算法(`algorithm`)、距离表示方式(`p`)、树叶大小 (`leaf_size`)。除此之外，本文还使用了测试集数据分隔函数 `train_test_split()` 和交叉验证函数 `cross_val_score()` 来训练和选择模型。

2.6 爬虫技术

本部分主要介绍爬虫的基本原理，并介绍利用 Python 语言及其软件包爬取网络内容的基本思路和关键方法。

由于社交平台对外开放的 API 需要付费，在不和社交平台进行合作的前提下，

基于网络爬虫的数据获取方式是更为常用的。爬虫技术一般可以分为四类：通用网络爬虫、主题网络爬虫、增量性爬虫和深层网络爬虫。网络爬虫通过 HTTP 协议对网页 URL 发送请求，从而获取网页中的各种信息。在社交网络用户行为数据的获取中，爬虫程序还需要对社交平台展示页面中的数据结构进行分析，根据页面功能的变化编写和调整抓取逻辑。

Python 的爬虫架构主要由五个部分组成，分别是调度器、URL 管理器、网页下载器、网页解析器和应用程序。调度器负责 URL 管理器、网页下载器和网页解析器之间的调度协调工作。URL 管理器中存储了已经爬取过的 URL 地址与还没被爬取过的 URL 地址，防止重复抓取。网页下载器通过传入 URL 地址来下载网页，常用的网页下载器有 `urllib2`、`lxml` 和 `requests`，其中 `urllib2` 是 Python 自带的工具包，`lxml` 和 `requests` 属于第三方包，需要下载安装。网页解析器将下载好的网页字符串进行解析，可以按照特殊要求来提取需要的信息，也可以根据 DOM 树来解析。常用的网页解析器有正则表达式、`html.parser`、`beautifulsoup` 等。其中 `html.parser` 为 Python 自带的工具包，`beautifulsoup` 为第三方工具包，功能更加强大。应用程序即将提取出的数据组合。

2.7 本章小结

本章主要介绍了论文所使用的关键技术背景。具体包括：(1)介绍了社交网络和微博；(2)介绍了电商平台与淘宝；(3)介绍了机器学习的概念及其算法，有决策树，随机森林、逻辑回归和 k 近邻；(4)对 Anaconda 和 Scikit-Learn 简要介绍。(5) Python 爬虫及其工具包。这些技术是本文研究工作的理论支持，对理解后文的研究内容具有重大的实际意义。

3 电商网红社交与销量数据的获取

本章主要介绍电商网红社交与销量数据的获取方法，首先获取了电商网红名单；接着介绍了基于 Python 的社交网络爬虫与基于 Web Scraper 的电商销量爬虫设计实现过程及关键问题，最后展示了数据采集结果并进行了数据预处理。

3.1 问题描述

第一章提到电商网红在数据科学领域的研究十分匮乏，是一个新兴方向，没有公开的电商网红数据集。必须先获取电商网红的社交数据和个人基本信息数据，才能深入分析电商网红的行为特点以及各种行为对其商业水平的影响。想要获取电商网红数据首先需要解决电商网红的识别与选取问题。本研究综合公开数据报告结果与微博网红店铺经验分享博主所公开的知名电商网红名单，制定了数据集所需的有效电商网红名单。

基于电商网红名单，本研究使用爬虫技术爬取名单内的电商网红在微博的社交数据与其淘宝销量数据。用户数据资源宝贵，免费数据获取十分困难。微博公开的商业接口中社交数据不全面，字段较少，且爬虫频次有限制，获取数据慢，必须编写爬虫程序以获取更全面的数据。然而微博与淘宝反爬机制越发完善，需要更加复杂的设计才能避开反爬机制。所以本研究对所需的微博网页进行详细分析，制定了一套完整的爬虫流程，避开微博的登陆、验证、封账号等反爬设计，获取更加完整的用户微博数据。

3.2 电商网红名单获取

电商网红多以个体商户形式在淘宝、微店等电商平台经营，在微博、小红书、抖音等社交平台或含社交功能的平台进行店铺的宣传与日常社交。其中，微博与淘宝于 2013 年 4 月建立了稳定的合作关系，微博成为淘宝最大的社交流量入口。微博与淘宝的数据理论上有很强相关性，所以本研究将在淘宝开店，主要在微博上进行社交的电商网红作为研究对象。根据 2016 年 CBNDData 中国电商红人大数据报告，电商网红店铺售卖的商品中女装占比最大，其次是女鞋和母婴，且女装红人的影响力正在大幅度像其他各行各业辐射。所以本研究锁定经营女装的电商红人。然而，知名女装电商红人究竟有多少是难以确认的。目前没有任何公开文献资料与行

表 3-1 电商网红名单
Table 3-1 The list of e-commerce celebrities

电商网红微博 ID			
ALU_U	DALU 大璐	Z_子晴	腻总 ninitalk
Lin 张林超	DDshadow	阿花花酱	朴瑟 seul
AM_FASHION	delicious 大金	白涩涩的茉莉	十元诗苑
MALInv	D-nana	蔡珍妮	时髦人绵羊 er
Tikilee	DoraPE	茶茶丷	孙嘉一 Zoe
yeswomen 小宜	Dreamy_梦梦	超级蓝大大	唐颖 Connie
ZY 喜哥	FAIRY_WANG	陈小末	滕雨佳 Amiu
阿希哥 VCRUAN	FEERIQUE_梵莉可	大兔子 PINZIKO	王火锅是火锅王
狼宝-LangBoom	FengFan_x	大喜庆	王幼宣幼乖
林珊珊_Sunny	fishdo	大小姐 77	魏妮妮 Yanni
卢洁云	Hera 是你的苗哥	大旭呀大旭	我是 miss 阮阮
美美 de 夏夏啊	Isabella 陈子仙	管阿姨	小饭噶
呛口小辣椒	JIN_阿金	龟酱 turtle	心蓝 grace
小刘小粒赵大喜	Lisa 兔牙	韩雨嘉 Yoga	凶猛熊猫 87911
小怡吖小怡	LOVE-小银子	郝静 sevi	许雯 May
雪梨 Cherie	MINI 猪七七	狠赵狠蛇蛇	杨泡泡 quan
于 momo 小饺子	MOSSMOSSMOSS	花花 ONGAHONG	杨小卷_NatureQ
张大奕 eve	Mr 九九_s	花花-Yumi	余人三日
13c13c	NanaStore 微博	加比 Gaby	造型师邹邹
onlyanna	oopsLUNA	金蘑菇菇	张思佳-SISI
_aaayuko	Saya 一	李雨桐 Luyee	张予曦
13_macy	SecretChan 陳媽	林糊糊	张雨乔 Babijisa
A-bowlife	seina 施依娜	林小宅-	张智研
Alice 赵静	SHIWEILIANGXXX	刘钰懿-Shirleylau	张佐佐 997
ashui-AS	UVN 许微娜	卖皮草的 CC	长胖了的 AVIVA
Ciny 心霓儿	VIMAS-小维	南表妹	赵若语_Crystal
周小熊大人	周扬青	左岸潇	左娇娇 Rosemary

业报告给出明确的电商红人行业总结名单。个别与淘宝有数据合作的数据中心如 CBNDData 公开了一份女装行业 Top20 网红名单。除此之外，阿里巴巴生意参谋团队研发的一款面向淘宝系头部商家的付费大数据监控产品——数据作战室，其

2018 年双十一活动当天提供的女装行业销量 Top30 店铺名单中，有 24 位电商网红。然而这些权威名单包含的样本数据量偏少，且均是高销量网红，样本区分度很小，不能满足一个机器学习模型的建立与后续优化。经过调查与实验发现，想要从淘宝和微博上识别女装电商网红并通过爬虫技术获取名单是不可行的。首先，淘宝有完备的防爬虫体系来保护商家店铺信息数据，几乎不能大批量爬取商家信息，无法从中筛选出在微博活跃的电商网红。虽然手机淘宝端会显示“红人店铺”标签，但并非所有电商网红都开通该标签。其次，微博网页端有搜索显示限制，无法显示所有网红标签用户。且微博侧同样存在并非所有电商红人都为账户添加网红标签或开通网红认证的问题。综上所述，想通过大规模爬虫技术遍历整个淘宝与微博是不可能的。为了扩充数据集，使其样本达到一定规模且有明显区分度，本研究通过爬虫遍历了微博红 v 认证用户，结合 CBNDData、阿里数据作战室和知名网购维权、网红店铺测评博主@母神 aki 公开的其账号下讨论度较高的 110 位电商网红名单，制定了初始电商网红名单，包含 144 位电商网红。@母神 aki 拥有近 53 万粉丝，且从 2011 年 6 月 30 日就开始相关话题讨论，其经验总结，虽不能完全覆盖行业内所有电商网红，但是具有较高可信度和参考价值。综上所述，本研究用到的 108 位电商网红名单具体如表 3-1 所示。

3.3 基于 Python 爬虫的社交数据获取

本部分首先介绍了自动获取微博游客 cookie 来避开微博反爬机制的方法，接着探讨了电商网红个人基本信息及所有微博内容的自动化获取思路，实现了一整套可以批量、稳定爬取用户所有微博内容的方法。

3.3.1 自动获取游客 cookie

微博全站使用 Sina Visitor System(新浪访客系统)来进行用户和爬虫识别。该系统判断用户请求中是否携带 cookie，如果有则直接进入正常页面，否则将进入访客流程，创建一个游客 cookie 以便访问。对用户来说，使用当前浏览器登陆过微博后就会带有 cookie，之后浏览微博任何站内页面都不会再弹出该系统。但是一般的爬虫程序不携带 cookie，只有进行模拟登陆或者把已有的 cookie 放入爬虫请求中才能通过该系统。然而使用模拟登陆可能会遇到各种形式的验证码，使用已有的 cookie 也有一定的有效期。且这两种方法都需要账号，后期爬虫中如果触碰到其他反爬机制还可能被冻结账号，这些方法都不能够稳定有效地实现自动化爬取。所以本小节介绍如何跳过 Sina Visitor System 来保证后续爬虫的稳定进行。

首先打开一个微博网页，使用谷歌浏览器 Chrome 的开发者工具分析页面。可以发现，请求该微博网页时，第一次 HTTP 状态码为 302 重定向，经过一系列请求跳转后状态码变成 200。对比状态码为 302 和 200 的两次请求可以发现，状态码为 302 的请求为 set-cookie，而状态码为 200 的请求中多了 cookie，里面有三个值：YF-Page-G0、SUB 和 SUBP。这就是本研究所需要获取的游客 cookie。

接着分析 Sina Visitor System 的网页源码，从中找到游客 cookie 的产生方法。源码的整体流程是：判断用户请求中是否携带 cookie，如果有则直接进入正常页面，否则将进入访客流程。具体代码如下：

```
// 流程入口
wload(function () {
    try {
        if (!Store.CookieHelper.get('SRF')) {
            // 尝试从 cookie 获取用户身份，获取失败走创建访客流程
            tid.get(function (tid, where, confidence) {
                incarnate(tid, where, confidence);
            });
        } else {
            // 用户身份存在，尝试恢复用户身份
            restore();
        }
    } catch (e) {
        // 出错
        error_back();
    }
});
```

可以看到 incarnate() 是用来给用户赋予访客身份的。它发送了一个 get 请求，只要模仿这个 get 请求，就可以获取 cookie。找到这个函数的具体定义如下：

```
// 为用户赋予访客 cookie
var incarnate = function (tid, where, confidence) {
    var gen_conf = "";
    var from = "weibo";
```

```

var incarnate_intr = "http://passport.weibo.com/visitor/visitor?a=
incarnate&t=" + encodeURIComponent(tid) + "&w=" +
encodeURIComponent(where) + "&c=" +
encodeURIComponent(conficence) + "&gc=" +
encodeURIComponent(gen_conf) + "&cb=cross_domain&from=" +
from + "&_rand=" + Math.random();
url.l(incarnate_intr);
};

```

从上面的源码中可知，想要成功发送这个请求，需要以下几个参数：a、t、w、c、gc、cb、from、_rand。图 3-1 给出了该请求发送详情，观察可知 a、cb、from 三个参数是定值，_rand 是随机数，gc 是空值。只需要得到 t(tid)、w(where)、c(conficence)三个参数，其中 t(tid)在 mini_original.js 请求中可以找到，w(where) 在 "new_tid"为 true 的时候是 3，false 的时候是 2，c(conficence)可能有也可能没有，没有时默认为 100，实验中使用默认值。

得到全部参数后，模拟发送 incarnate 请求，得到 SUB 和 SUBP，加上固定的 YF-Page-G0，就可以得到完整的游客 cookie。将这个 cookie 填入请求头，再次发起微博网页请求即可在没有账号的情况下跳过 Sina Visitor System。

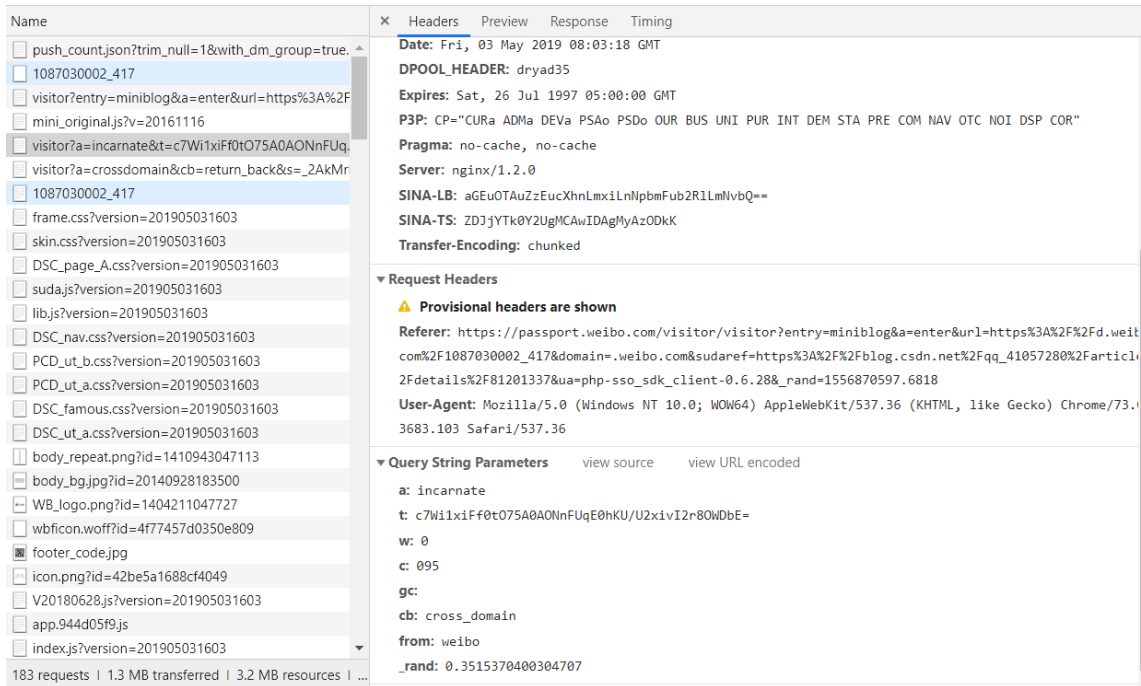


图 3-1 incarnate 请求详情

Figure 3-1 Details of incarnate request

3.3.2 用户基本信息获取

微博用户的个人主页收录了用户的个人信息及他在微博平台上发布和分享的信息，这些信息包含了用户行为，可以通过爬取电商网红的个人主页来获取其社交数据。3.2 节已经得到电商网红名单，本小节需要利用该列表进一步获取电商网红的个人主页地址和基本信息。

微博共有三个站点，分别是 <https://weibo.cn>、<https://m.weibo.com> 和 <https://weibo.com>。三个站点的页面复杂度逐渐提高，<https://weibo.com> 的内容最丰富，抓取复杂度最高，<https://weibo.cn> 的内容最少，抓取最简单。为了抓取用户个人主页地址和较多的基本信息，本节选择 <https://m.weibo.com> 站点来设计爬虫程序。



图 3-2 微博用户搜索页面

Figure 3-2 User search page on Weibo

首先观察微博用户搜索页面，例如搜索电商网红 [onlyanna](#)，如图 3-2 所示。微博搜索算法会将匹配度最高的用户放在搜索结果的第一位显示，第一位用户即是本研究需要的电商网红。查看该搜索页面的源代码，使用 Chrome 开发者工具中的 Elements 工具查看排在首位的用户表单代码，如图 3-3 所示。经分析可知，该电商网红个人主页 url 存储在 `W_texta W_fb` 标签下的 `href` 中，使用 lxml 中的 `etree` 和 `xpath` 可以获取该地址。同理，还可以找到 `person_num` 下的关注量，粉丝量与微博总量等信息。

接着，需要对 108 位网红进行遍历，获取网红主页 url 列表。遍历的过程需要所有电商网红名字的搜索页 url，解析微博用户搜索页地址发现，该地址为固定格式：`url = 'https://s.weibo.com/user?q=' + '用户名' + '&Refer=weibo_user'`，只需要

将用户名替换成不同电商网红的名字，就可以得到电商网红搜索页 url 列表。

```

<!doctype html>
<html>
<head>...</head>
<body class="wbs-user" style="background-position: center 140px; background-repeat: repeat-x;" == $0
  <div class="m-main">
    <!--搜索-->
    <div class="m-search" id="pl_feedtop_top">...</div>
    <!--/搜索-->
    <!--主导航-->
    <div class="m-main-nav s-mt28">...</div>
    <!--/主导航-->
    <!--内容-->
    <div class="m-wrap" id="pl_feed_main">
      <div class="m-con-1">
        <!--子导航-->
        <div class="m-sub-nav" id="pl_common_totalshow">...</div>
        <!--/子导航-->
        <!--高级搜索筛选-->
        <div class="m-filtertab" id="pl_user_filtertab">...</div>
        <!--/高级搜索筛选-->
        <!--无结果-->
        <div class="card-wrap" id="pl_user_feedList">
          <!--用户直达-->
          <!-- 用户card-->
          <div class="card card-user-b s-pg16 s-brt1">
            <div class="avator">
              <a href="//weibo.com/u/1736315592" target="_blank" suda-data="key=tblog_search_weibo&value=seqid:155602189286902281973|type:3|t:0|pos:1-0|q:onlyanna|ext:mpos:1,click:user_pic">...</a>
            </div>
          </div>
        </div>
      </div>
    </div>
  </body>

```

图 3-3 用户搜索页面源代码

Figure 3-3 Source code of user search page

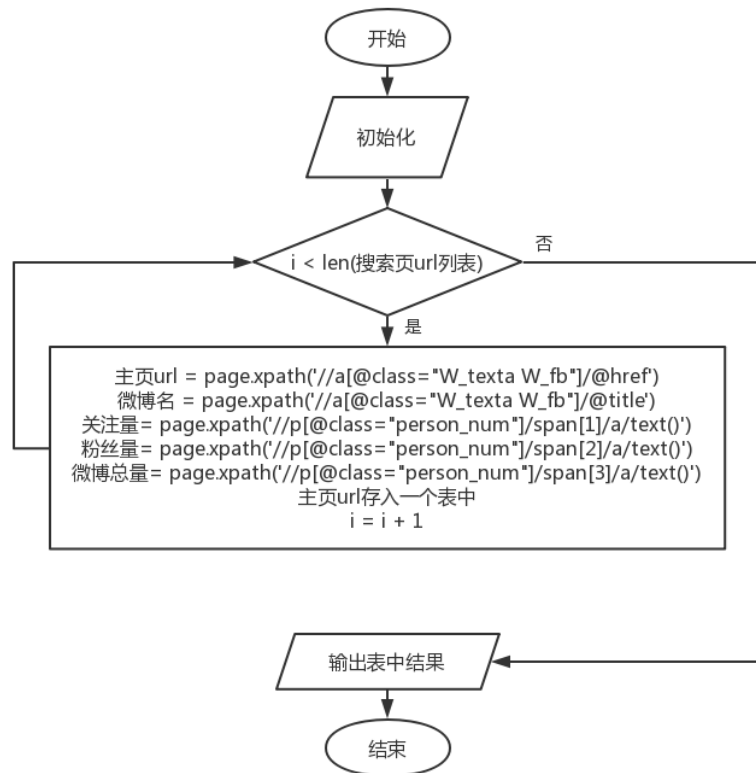


图 3-4 个人信息爬虫流程

Figure 3-4 Crawler process of personal information

电商网红微博个人主页 url 列表的爬虫程序流程如图 3-4 所示。首先初始化一个数据表用以存储最终结果,接着遍历电商网红搜索页 url 列表,解析每个搜索页,提取首位表单用户的个人主页 url、关注量,粉丝量与微博总量,并存入数据表中。遍历完毕后,输出电商网红基本信息数据。

同理,还可以使用相同的方法爬取其他用户信息。比如下面的研究中,将 108 位网红转发和点赞过的用户生成列表,用上述方法获得了这些用户的基本信息,作为补充数据。

3.3.3 用户原创微博与转发微博获取

想要获取电商网红微博个人主页中的信息,首先需要对个人主页结构进行分析。本节继续使用 <https://m.weibo.com> 站点的网页进行研究。移动版个人主页打开后如图 3-5 所示,页面顶端显示电商网红的个人信息,下半部分为电商网红发布和转发微博,不显示点赞微博。页面使用 Ajax 加载,即当鼠标拖到页面底端后会自动加载下一页,需要模拟 Ajax 请求才能通过爬虫获得所有微博数据。



图 3-5 移动版个人主页

Figure 3-5 personal homepage on mobile phone

本研究使用 Chrome 开发者工具中的 Network 工具分析该网页的 Ajax 请求。如图 3-6 所示, Ajax 请求以 `getIndex` 开头,选中一个 Ajax 请求查看它的头部信息

可知，这是一个 GET 类型的请求，请求的参数有 4 个：type、value、containerid 和 page。查看其它同类请求可以发现，type、value 和 containerid 不会发生改变，只有 page 的值发生递增。可以判定 page 参数是用来控制分页的，page=1 代表第一页，page=2 代表第二页，以此类推。

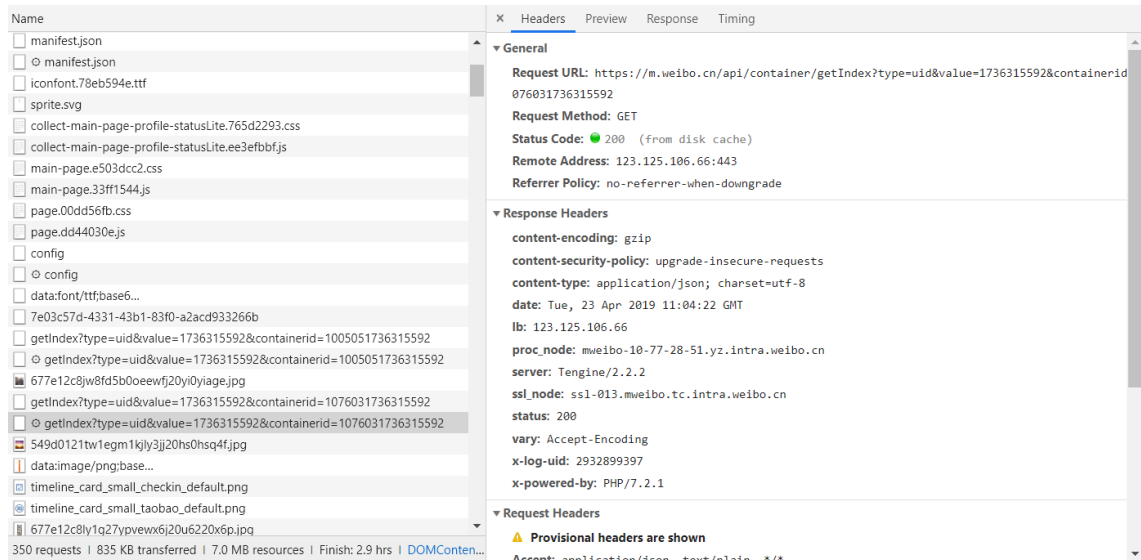


图 3-6 Ajax 请求头部信息

Figure 3-6 header information of Ajax requests

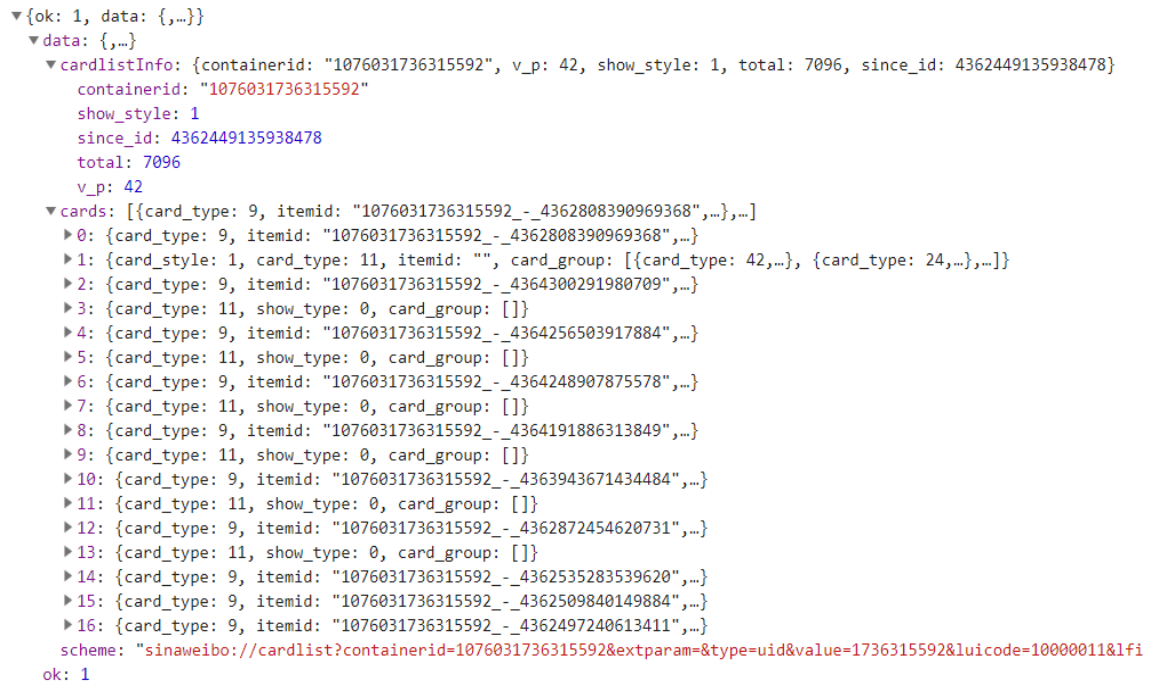


图 3-7 Ajax 请求响应信息

Figure 3-7 Response for Ajax requests

接着分析 Ajax 请求的响应，如图 3-7 所示。响应数据是 JSON 格式，最关键的两部分信息放置在 `cardlistInfo` 和 `cards` 内。`cardlistInfo` 中的 `total` 下的数据代表微博总量，`cards` 内包含 17 个元素，每个元素是一条微博数据。

展开单条微博数据，如图 3-8 所示。可以发现这个元素中的 `mblog` 字段中包含着微博信息，比如 `attitudes_count` (点赞数)、`comments_count` (评论数)、`reposts_count` (转发数)、`created_at` (发布时间)、`text` (微博正文) 等。若该微博为点赞或转发微博，则原博主信息存储在 `retweeted_status` 字段中。

```

▼4: {card_type: 9, itemid: "1076031736315592_-4364256503917884",...}
  card_type: 9
  itemid: "1076031736315592_-4364256503917884"
  ▼mblog: {created_at: "3小时前", id: "4364256503917884", idstr: "4364256503917884", mid: "4364256503917884",...}
    attitudes_count: 175
    bid: "HqXBwgrdG"
    can_edit: false
    comments_count: 0
    content_auth: 0
    created_at: "3小时前"
    edit_at: "Tue Apr 23 15:09:09 +0800 2019"
    ▶edit_config: {edited: true,...}
    edit_count: 1
    extern_safe: 0
    favorited: false
    hide_flag: 0
    id: "4364256503917884"
    idstr: "4364256503917884"
    isLongText: false
    is_paid: false
    mblog_vip_type: 0
    mblogtype: 0
    mid: "4364256503917884"
    more_info_type: 0
    ▶number_display_strategy: {apply_scenario_flag: 3, display_text_min_number: 100000, display_text: "100万+"}
    pending_approval_count: 0
    pic_types: ""
    raw_text: "孕妈妈给大家参考呀 面料的银丝帮我拍出来了哈哈哈哈哈 "
    reposts_count: 10
    ▶retweeted_status: {created_at: "4小时前", id: "4364251328285579", idstr: "4364251328285579", mid: "4364251328285579",...}
    reward_exhibition_type: 2
    reward_scheme: "sinaweibo://reward?bid=1000293251&enter_id=1000293251&enter_type=1&oid=4364256503917884&seller=17363155"
    show_additional_indication: 0
    show_attitude_bar: 0
    source: "iPhone客户端"
    text: "孕妈妈给大家参考呀 面料的银丝帮我拍出来了哈哈哈哈哈"
    ▶user: {id: 1736315592, screen_name: "onlyanna",...}
    version: 1
    ▶visible: {type: 0, list_id: 0}
    weibo_position: 3
    scheme: "https://m.weibo.cn/status/HqXBwgrdG?mblogid=HqXBwgrdG&luiicode=10000011&fid=1076031736315592"
    show_type: 0

```

图 3-8 单条微博的内容

Figure 3-8 Content of a microblog

根据 Ajax 请求分析可知，只要做一个循环，使用 `page` 参数改变接口，用总微博量估计请求次数，不断地发起请求，每次都可以获得包含固定条数微博内容的响应，循环完成即可获得所有原创与转发微博内容。

该爬虫程序流程如图 3-9 所示，首先进行一些初始化和定义，接着使用生成 Ajax 请求地址函数和估算请求次数函数得到循环所需的地址和循环次数，进入循环后，通过获得请求结果函数得到每次请求的响应结果，再通过提取数据函数对响应结果进行解析，得到本研究想要的的数据，最后保存数据。上述流程中使用到的 4

个函数将在下面分别进行详细介绍。

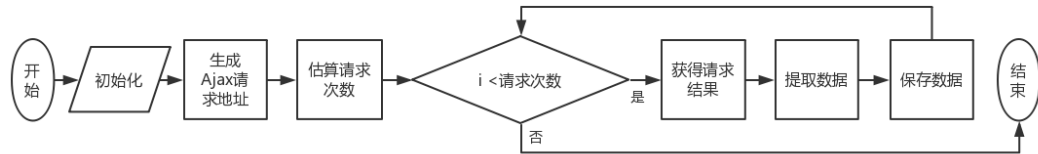


图 3-9 原创与转发微博爬虫程序设计流程

Figure 3-9 The process of crawler programming for posts and retweets

(1) 生成 Ajax 请求地址:

首先构造参数字典, 按照上一节中的分析, 在请求时 type、value 和 containerid 是固定参数, page 定为 1。type 的值固定为'uid', value 的值为用户 id, 该值可以从个人主页地址中直接解析出来, containerid 为'107603'+value'。接着调用 Python 的 urlencode()函数将参数转化为请求 url。

(2) 估算请求次数

用 requests 请求上一个函数生成的请求地址, 加入初始化中的 headers 参数, 判断响应的状态码是否为 200, 如果是则调用 json()函数将响应内容解析, 并从响应内容中提取出 cardlistInfo 中 total 字段下的微博总量数据, 并统计一次请求的微博条数, 对这两个数据取余加 1, 即可得到请求次数。

(3) 获得请求结果

与函数(1)类似, 使用 type、value 和 containerid 这个三个固定参数和 page=i, 构造本次请求 url, 接着与函数(2)类似, 将请求对应的响应内容解析为 JSON 返回。

(4) 提取数据

遍历 cards 元素, 使用 get()函数在 mblog 下的对应字段下取出想要的信息, 并生成字典一个存储这些信息。

最后, 只需要遍历 108 个网红, 完成 108 次循环爬虫流程就可以得到所有网红的原创微博与转发微博数据。

3.3.4 用户点赞微博获取

上一小节的分析提到 https://m.weibo.com 站点仅能获取用户的原创与转发微博, 缺失了用户的点赞微博数据。用户点赞微博数据仅能从 https://weibo.com 站点中获取, https://weibo.com/用户 id/like?page=1#feedtop 是用户点赞过的微博集合页面, 只需要改变网页 url 中用户 id, 就可以得到所有网红的点赞微博页面。

用户点赞微博页面在使用了 Ajax 加载的同时还设有点击“下一页”按钮进行翻页。使用浏览器的开发者工具对用户点赞微博页面进行分析可知，每页有两次 Ajax 动态加载，对比两次动态加载的请求 url，发现只有 `_rnd` 和 `pagebar` 这两个参数取值不同。多次测验后发现 `_rnd` 不起作用，`pagebar` 取 0 和 1 分别代表两次动态加载。接着点击其他页数，发现请求 url 中的 `page` 参数对应页数。综上所述，通过改变 `pagebar` 参数，可以获取每页 Ajax 动态加载的内容；通过改变 `page` 参数可以进行翻页。所需数据通过上一节中的方法，寻找对应 Xpath 即可提取。

电商网红点赞微博数据爬虫程序流程如图 3-10 所示，首先进行初始化，定义所需变量，接着构建第一页的第二个 Ajax 请求，从请求结果中找到 `action-data` 标签中 `countPage` 的值，该值即为总页数。接着遍历每一页，对每一页生成对应的网页请求地址和两个 Ajax 请求地址，得到 3 个请求的结果，从中提取出所需数据，最后保存数据。流程使用到的函数与上一节中介绍的大致相同，不再赘述。

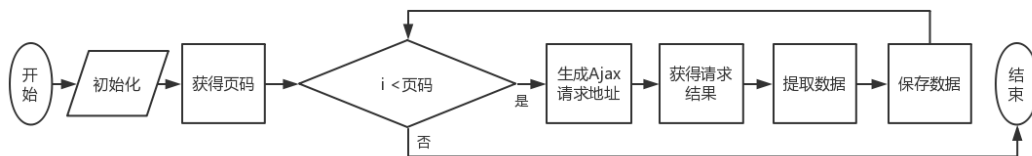


图 3-10 点赞微博爬虫程序设计流程

Figure 3-10 The process of crawler programming for likes

3.4 基于 Web Scraper 的销量数据获取

电商网红的社交行为可以从微博数据中提取，那么同理电商网红的商业数据可以从淘宝平台中获得。由于淘宝反爬严重且淘宝店铺页面中有很多动态数据，使用 Python 自己编写程序进行淘宝爬虫将会遇到很多困难。本研究使用 Chrome 浏览器的爬虫插件 Web Scraper 可以轻松解决这些问题。

在使用插件爬虫之前，与电商网红微博信息爬虫思路一样，首先需要获取电商网红的淘宝店铺 url 列表。电商网红的淘宝店铺名与微博账户名不相同，无法通过搜索的方法精准匹配，所以本研究采用人工查找的方法获得电商网红的淘宝店铺商品详情页 url 列表。

淘宝不向商家以外的人直接显示店铺任何商业数据，必须通过别的方法收集得到。本研究爬取电商网红店铺中所有商品销量数据，通过对所有商品销量求和的方式得到电商网红的总销量作为其商业数据。需要注意的是，淘宝不显示单个商品的累计销量数据，只显示该商品 30 天内的销量数据，所以本研究最终得到的是电商

网红一个月的总销量。

确定商业数据的计算方法后，开始使用 Web Scraper 插件进行爬虫设计。Web Scraper 操作简单，首先新建一个 Sitemap，即爬虫，输入 108 个电商网红的店铺商品详情页 url，接着如图 3-11 所示，建立一个 selector 后点击 select 激活选框，只需要在页面上点击想要获取的元素，selector 会自动获取该元素的 Xpath，给想要的字段取名保存即可。

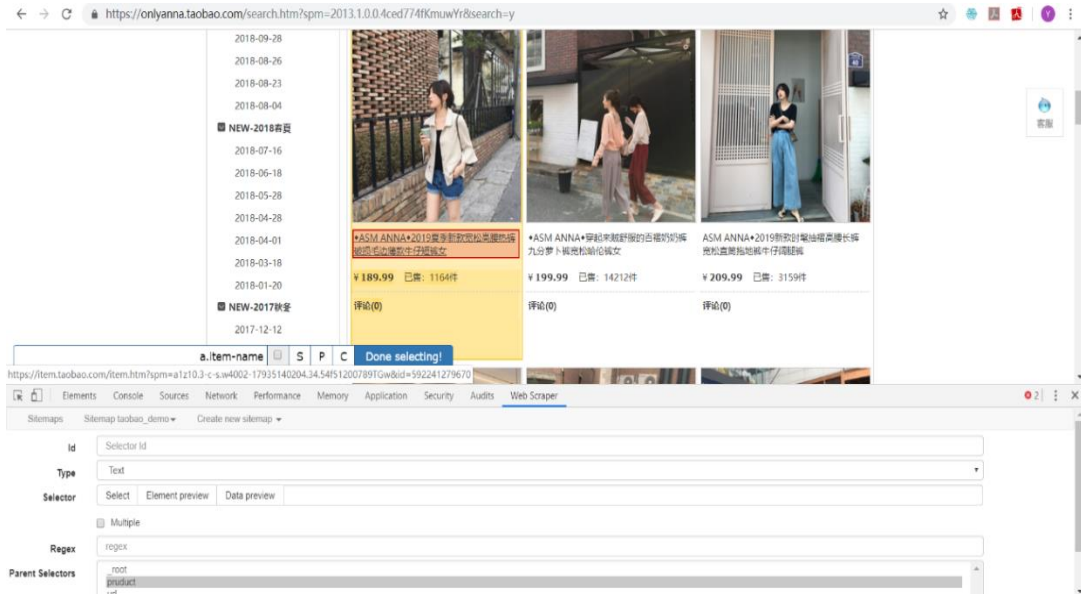


图 3-11 Web Scraper 使用示范

Figure 3-11 The use of Web Scraper

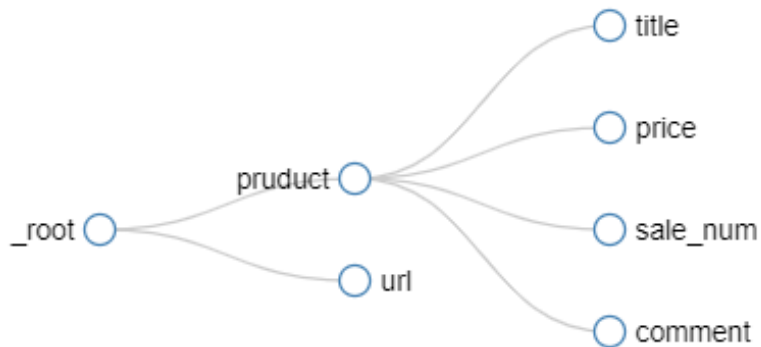


图 3-12 淘宝爬虫 workflow

Figure 3-12 The workflow of Taobao crawler

本研究设计的爬虫 workflow 如图 3-12 所示, 打开一个电商网红店铺主页后, 首先记录该页的 url, 并选取该页面所有商品表单, 接着遍历每个商品表单, 提取商品名称、商品价格、商品销量和评论量等信息。另外, 只需要在填写电商网红店铺商品详情页 url 的时候, 只需要令 url 中的 page=[1-‘总页数’], 该爬虫就会实现自动翻页功能。

表 3-2 电商网红微博基本信息字段介绍

Table 3-2 The introduction of e-commerce celebrities' Weibo information fields

数据字段	含义
userid	用户 ID, 用户的唯一识别号
user_name	用户微博名
follows	用户关注量, 用户所关注的其他账号数量
fans	用户粉丝量, 关注该用户的账号数量
vip	用户账户的 Vip 等级
location	用户地域, 省份-城市
creation_time	用户账号创立时间, 精确到日
posts	用户总微博量, 包括用户原创、转发和点赞的所有微博

表 3-3 用户社交数据字段介绍

Table 3-3 The introduction of user social data fields

数据字段	含义
userid	用户 ID, 用于将每条数据与用户对应
time	发布微博的时间
text	微博内容
comment	评论量, 用户某条微博下的评论数量
repost	转发量, 用户某条微博被转发的数量
like	点赞量, 用户某条微博被点赞的数量
re_user	原微博作者名, 若非转发或点赞微博, 则该字段为空
re_comment	原微博评论量, 若非转发或点赞微博, 则该字段为空
re_repost	原微博转发量, 若非转发或点赞微博, 则该字段为空
re_like	原微博点赞量, 若非转发或点赞微博, 则该字段为空

3.5 采集成果与数据字段介绍

使用前文介绍的爬虫方法，最终获取了电商网红社交数据和销量数据，同时，本研究还爬取了被电商网红转发和点赞过的用户的微博基本信息作为补充数据，以便接下来的研究使用。各数据表的具体字段介绍如表 3-2、表 3-3、表 3-4 和表 3-5 所示。

表 3-4 电商网红淘宝字段介绍

Table 3-4 The introduction of e-commerce celebrities' Taobao fields

数据字段	含义
userid	用户 ID，用于将每条数据与用户对应
sales	店铺一个月的总销量

表 3-5 被转发点赞用户账号信息字段介绍

Table 3-5 The introduction of alters' information fields

数据字段	含义
alter_name	被转发和点赞的用户名
alter_fans	被转发和点赞的用户粉丝量
alter_follows	被转发和点赞的用户关注量
alter_posts	被转发和点赞的用户总微博量

3.6 数据预处理

本章所用数据集为爬虫获取，数据格式相对杂乱，需要对数据进行预处理后才能研究。本章的数据预处理包括两个方面：一是数据整理，二是字段处理。数据整理部分，主要是清洗掉一些爬虫中断后重新爬虫产生的重复值。需要进行字段处理的有 comment、repost、like、time 和 text。comment、repost 和 like 三个字段中有些数据格式不为数字，例如某条微博下的评论数是 0，那么这条数据的 comment 字段会显示成“评论”两个字，另外两个字段同理。所以本研究将这三个字段下所有显示为文字的部分都填充为数字 0。为了后续研究的方便，本研究还需将 time 字段下的时间表示统一为 python 常用时间格式。由于本研究的最小单位精确到天，所以将 time 字段下所有数据统一修改为 YYYY-MM-DD 格式。为了研究网红的微博内容，需要将网红发表的内容从字段中提取出来。text 字段不仅包括了网红的发文内容，还包括了前几次转发中他人的发表内容，微博用“//”标志分隔每个转发

用户的发文内容。所以本研究将 text 字段内容中第一次出现“//”后的内容都除去，只保留了电商网红的发文内容。

通过这种方法最终得到 108 电商位网红共 449489 条微博数据。表 3-6 是数据的统计信息。电商网红的平均微博数据量为 4034 条，网红与网红之间有一定差异。数据量最大的电商网红有 13758 条数据，只有少量网红的数量低于千条。

表 3-6 数据统计信息

Table 3-6 The statistics of data

	数据量
平均值	4162
标准差	2791
最小值	353
第一四分位	2173
中位数	4034
第三四分位	5509
最大值	13758

3.7 本章小结

本章主要分别使用 Python 和 Web Scraper 设计微博和淘宝的爬虫程序，采集了电商网红的社交数据与销量数据，构成电商网红数据集。主要工作包括：(1)确定了网红大致规模，综合各类数据报告与微博博主经验总结，尽可能多地列举符合研究条件的电商网红，获得电商网红列表；(2)在无需账号的情况下跳过 Sina Visitor System，从根源上解决了账号被封的问题；(3)基于 Python，通过解析微博用户搜索页面获取电商网红的基本信息；(4)基于 Python，通过解析微博个人主页 Ajax 请求的响应结果，获取电商网红在微信平台从 2010 年 1 月 1 日至 2018 年 4 月 30 日的原创、转发和点赞微博；(5)使用(3)中相同的策略，获取了被电商网红转发和点赞过的 46470 个微博用户的基本信息，作为后续研究的补充数据；(6)使用 Web Scraper 插件设计淘宝爬虫程序，避开淘宝的反爬措施，获取电商网红的销量数据。

4 电商网红营销行为测量分析

电商网红社交行为特征的发现，需要以行为分析为基础。本章首先从微博内容中提取电商网红的三类营销行为，即电商网红的广告、促销和口碑营销行为，接着对这三种营销行为的进行深入分析，探索行为特征构建的方法，并初步探究营销行为带来的商业价值。

4.1 问题描述

第一章中提到电商网红近年来引起市场的高度关注的一个重要原因是电商网红的长期生命力。他们不断出现在各类消费领域，且在粉丝——交易转化率上有明显的优势。电商网红所打造的人格化品牌对消费者的吸引力已经远远超过普通品牌，他们所引申出的社交商业模式也在不断冲击、改革着传统在线商业模式。那么为了适应社交商业化，深入了解电商网红在社交网站上的行为规律是非常必要的。

想要了解电商网红的社交行为并探究其商业模式，需要全面而且深入地研究电商网红行为数据和其社交网络数据。电商网红在社交平台上的行为比普通用户更加丰富，这意味着电商网红行为有更多值得关注和深入研究的地方。例如电商网红是如何在社交平台上展开营销活动的？电商网红早期的营销行为与近期的营销行为相比有什么变化呢？他们如何与粉丝实现高效沟通与互动？如何调动粉丝、发挥粉丝的自媒体价值？如何给消费者提供更加丰富的购买动力？然而之前的研究者没有对大量电商网红进行过系统研究，大部分研究停留在理论观察阶段，只以几个网红为样本，基于对网红的采访和肉眼观察结果论述电商网红的各种营销行为，没有落实到真实数据上。为了深入了解电商网红，必须使用数据科学的方法进行定量分析。所以本章基于第三章采集的电商网红社交数据集、销量数据集和补充数据集，从多个尺度分别分析了电商网红的广告、促销和口碑营销三种营销行为规律，探索行为特征构建的方法，并初步探究营销行为带来的商业价值。

4.2 电商网红广告行为分析

电商网红主要的广告行为是在店铺发布新产品前开展的宣传行为，业界将这种行为称为“上新行为”。本节主要对预处理过的数据集进行更进一步的分析，挖掘电商网红上新行为模式，从多个尺度对其行为变化进行深入研究从而了解电商网红的商业模式。

4.2.1 上新微博内容分析

第一章中提到电商网红使用微博进行低成本、高效率的个人品牌宣传。而品牌宣传中,店铺上新前的预告宣传是最重要的一个环节。商品上新售卖前的宣传效果越好,本次上新的流水就会越多。如上一节中所说,本研究的电商网红的店铺均在淘宝平台,消费者在淘宝端关注电商网红店铺后,可以通过微淘消息收到电铺的上新通知与新品预览。但是淘宝虽然是一个拥有社交功能的电商平台,但用户已经习惯性将其当成购物工具而非社交工具,其社交积累与其他专业社交平台相比相差很大。这导致微淘社交氛围相对冷淡,店家与消费者无法进行深入有效的沟通和互动,在微淘上进行的上新宣传效果较差。除此之外,消费者可以通过手机淘宝进入店铺主页,选择“新品”栏目同样可以看到即将上架的商品预览。该途径需要消费者自发地查看,且中间需要的交互次数较多,用户体验较差,这并不是卖家能够控制的宣传途径。综合来说,淘宝平台社交能力很弱,导致单一地在淘宝平台上宣传不能达到良好的宣传效果,电商网红必须选择社交能力更强的微博来进行商品上新宣传。那么电商网红是如何在微博上进行上新宣传的呢?

表 4-1 各类型微博占比统计信息

Table 4-1 The statistics of different types of posts

	原创微博	转发微博	点赞微博
平均值	63.91%	15.83%	20.26%
标准差	16.04%	9.67%	15.51%
最小值	20.49%	2.63%	0.00%
第一四分位	55.76%	8.08%	9.55%
中位数	64.14%	14.01%	15.76%
第三四分位	75.97%	22.21%	30.10%
最大值	99.73%	51.38%	69.86%

电商网红在微淘上发布的消息全部与商品相关,但他们在微博上发布的消息种类却相对较多。微博上的消息分为三种:原创微博、转发微博与点赞微博。表 4-1 展示的是电商网红三种微博占比的统计情况,可以看到,电商网红原创微博平均占 63.91%,其次是平均占比为 20.26%点赞微博,最后是转发微博,平均占比为 15.83%,原创微博成为电商网红的主要发文形式。原创微博中,按内容大致可以分为:日常分享、上新微博和抽奖微博,其中有些微博可能同时拥有多个属性。三种属性的原创微博中,与商品直接相关的是上新微博。表 4-2 展示了 3 条随机抽样的上新微博当做示例,可以发现网红使用活泼明快的词汇和句式描述新商品特点,配

以高视觉效果的图片，有时还会发起抽奖活动来吸引粉丝和其他微博用户的注意。

表 4-2 上新微博示例

Table 4-2 Examples for new product previews

电商网红	上新预告原文
ALU_U	留言送出一整套 look! 这次新品里我超愛的一个牛仔背心很酷 很街頭俏皮的短裙也可以混搭的很有型穿上那一刻覺得自己...又帥了
张大奕 eve	只有用心才会让人记住绝对不做流星🌟🌟🌟。 19 号 10 点上新 帮你们过年,年会战袍出个主意盖楼啦
onlyanna	上新先剧透一款我个人很偏好喜欢的绑带衬衫, 这款属于看着不是多惊艳但是上身很有气质的一款单品, 搭配上一期的阔腿长裤和牛仔裤都很有感觉, 袖子是有小小廓形设计的哦版型真的特别特别推荐! 喜欢这款的宝贝赞里我们选 3 个

4.2.2 上新微博比例分析

电商网红在微淘上发布的消息全部与商品相关，但他们在微博上发布的消息种类却相对较多。人工阅读随机抽样的 10 位网红近 3 个月的所有微博后发现，电商网红的原创微博内容基本由上新预告、福利抽奖结果公示、日常生活分享等三部分构成。那么，上新预告占比的多少是否对网红有所影响呢？接下来，本节研究不同网红上新微博比例的差异。

上新微博比例定义为上新微博占网红所有原创微博的比例，想要得到上新微博比例，首先要从数据集中筛选出所有与上新相关的微博。这里采用以下处理：筛选 text 字段中，包含“上新”或“新品”的微博数据。表 4-3 给出上新微博比例的基本统计结果。电商网红上新微博比例平均为 7.72%，最高达 36.04%，网红间有一定差距。分别计算上新微博比例与网红粉丝量、网红淘宝销量的皮尔森相关系数来探究上新微博比例与粉丝量和销量的相关性。可以发现，上新微博比例与网红粉丝数的相关系数 $r=-0.149$ ，呈弱相关；上新微博比例与网红淘宝销量的相关系数 $r=0.414$ ，呈中等正相关。这个结果说明，上新微博比例几乎不会对网红粉丝量造成影响，但是对网红销量有所贡献。上新微博比例提高，网红的销量也会增长。这个结果也充分说明了网红在微博发布上新预告能够获得较好的宣传效果。大量的上

新预告虽然无法带来更多的粉丝流量，但能在现有的粉丝群体中催化出更高的转化率，实现更高的商业变现。

表 4-3 上新微博比例统计信息

Table 4-3 The statistics of preview percentage

上新微博比例	
平均值	7.72%
标准差	5.42%
最小值	0.31%
第一四分位	4.57%
中位数	6.65%
第三四分位	10.49%
最大值	36.04%

4.2.3 上新闻隔分析

接着，本章研究不同电商网红上新闻隔的差异。想要得到上新闻隔，必须找出网红每次上新的起始时间。本研究将网红每次上新时段中第一次发上新预告的时间定义为上新起点，使用图 4-1 所示的算法得到所有上新起点和上新闻隔。首先提取出上新微博的发布时间，形成一个时间戳序列。再初始化一个空表用来存储最终结果，空表中有两个字段，一个用来存储每次上新的起点，另一个用来存储该上新起点与下一次上新之间的上新闻隔。同时，将上新起点初始化为第一条上新微博的时间戳。接着，循环输入时间序列。每次循环中先计算一次相邻时间间隔，即本条上新微博与下一条上新微博的时间差。如果相邻时间间隔小于等于 7 天，可以认为下一条上新微博与本条上新微博同属于一个上新周期内，此时继续循环。如果相邻时间间隔大于 7 天且小于等于 90 天，可以认为已经进入到下一个上新周期中，此时计算上新闻隔，上新闻隔为下一条上新微博与上新起点的时间差。计算完成后，将上新起点改为下一条上新微博的发布时间再继续进入循环，接下来会识别新的上新周期，计算下一个上新闻隔。如果相邻时间间隔大于等 90 天，可以认为电商网红因故停止上新。需要注意的是，如果不设置停止上新门槛，会导致一些网红上新闻隔数据标准差过大，不符合实际情况。经多次实验后发现，停止上新门槛定为 90 天最合理，符合实际情况。此时，不需要计算上新闻隔，只需将上新起点改为下一条上新微博的发布时间，直接循环进入新的上新周期。最后，输出网红的上新闻隔数据。下面本研究将对上新闻隔数据进行两个尺度上的分析。

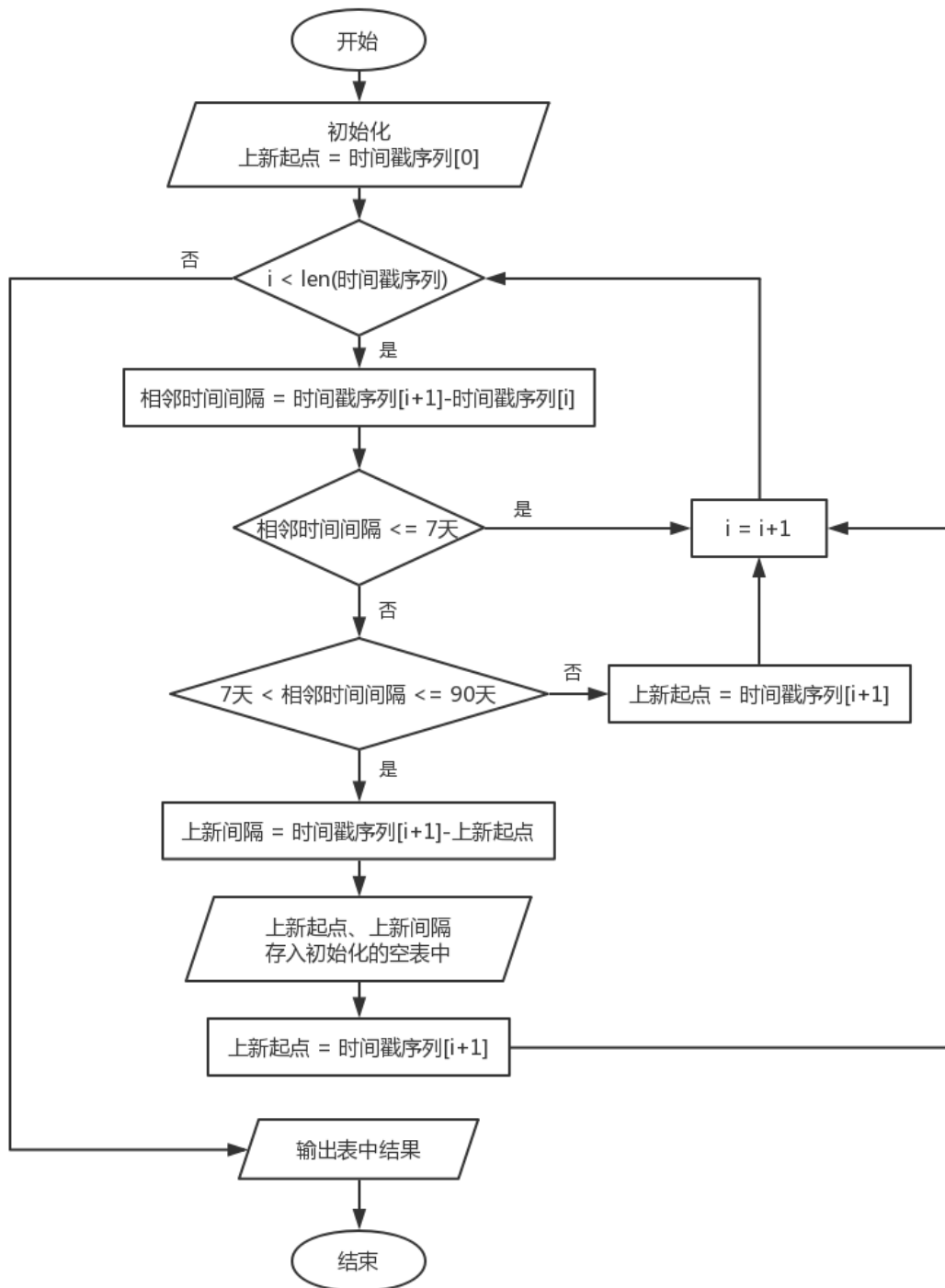


图 4-1 上新间隔计算流程

Figure 4-1 The process of interval calculation

表 4-4 上新闻隔统计信息

Table 4-4 The statistics of interval

上新闻隔（单位：天）	
平均值	28.31
标准差	5.68
最小值	19.21
第一四分位	24.86
中位数	27.45
第三四分位	30.99
最大值	54.90

首先观察网红的上新闻隔总体水平，本章对每个网红的上新闻隔数据求均值，所有电商网红的上新闻隔水平结果如表 4-4 所示。电商网红的上新闻隔平均值为 28 天，标准差远小于均值，说明大部分网红按月上新，网红与网红间上新闻隔差距很小。平均上新闻隔与粉丝的皮尔森相关系数 $r=-0.075$ ，平均上新闻隔与网红销量的皮尔森相关系数 $r=0.080$ ，均为弱相关。上述结果表明，网红与网红的上新闻隔基本是相同的，约为一个月。上新闻隔与粉丝量和销量关系都不大，这可能是由电商环境，供需关系和国内网红供应链能力共同决定的。

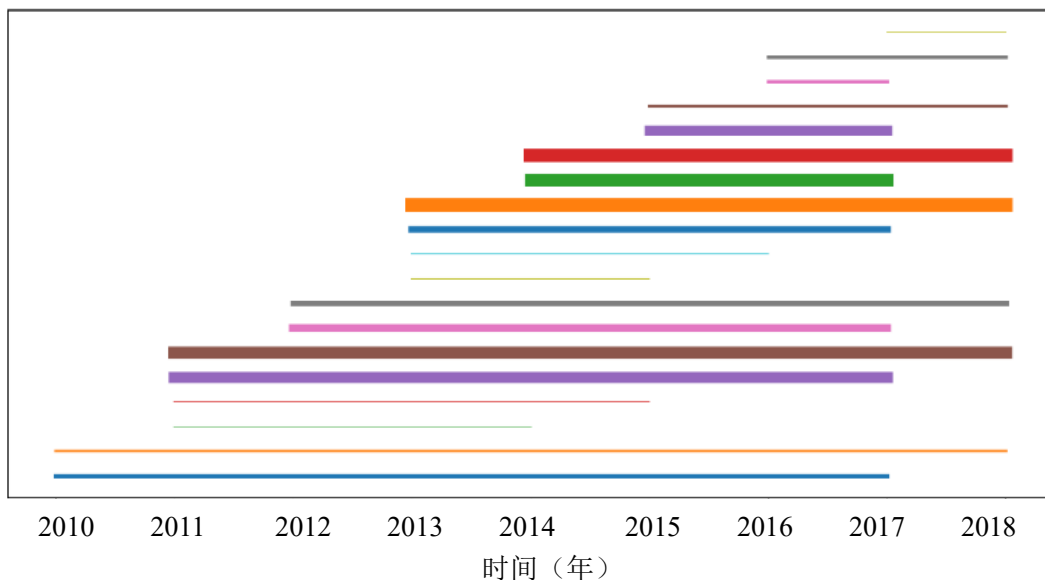


图 4-2 上新资历分布

Figure 4-2 The distribution of business qualification

最后，本研究从时间尺度观察不同资历的网红其上新间隔变化趋势。由于网红的上新间隔与粉丝量和销量关系不大，故本节将资历定义为电商网红第一次上新的年份，年份越早则资历越深。图 4-2 表示电商网红上新资历分布，其中每条线段的左右端点分别代表上一节中估算的电商网红上新的起点与终点，线段的粗细代表同一时间段的人数。从图中可以看出，电商网红开始进行“淘宝卖货——微博运营”商业模式的高峰期为 2011 年、2013 年与 2014 年。

本研究选取 2011 年与 2014 年开始在微博进行上新宣传的两组网红，画出她们的上新间隔随时间的变化趋势图，如图 4-3 所示，其中(a)表示 2011 年开始上新的网红，(b)表示 2014 年开始上新的网红。可以发现，不同上新资历的网红，其上新间隔变化趋势相同，均为围绕一个稳定值上下波动。

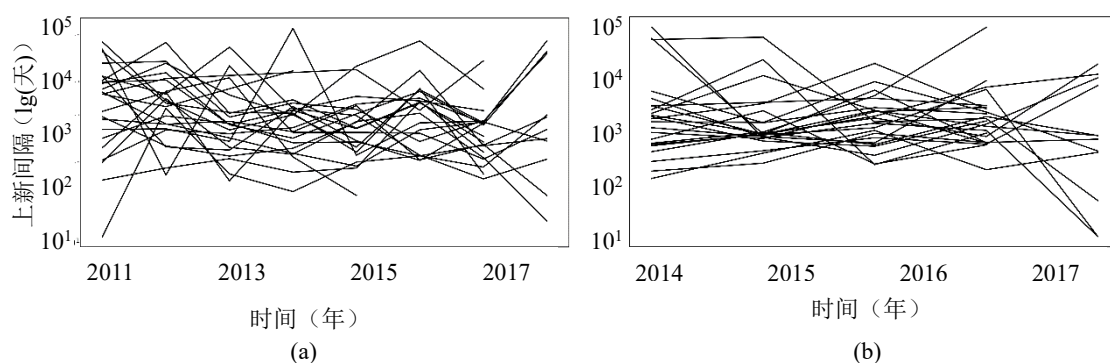


图 4-3 上新间隔变化趋势

Figure 4-3 The trend of interval

4.3 电商网红促销行为分析

本部分主要挖掘电商网红促销行为特点，从不同的抽奖手段分析其带来的效果，从而了解电商网红的商业模式。

4.3.1 抽奖微博内容分析

上一节提到，电商网红通过微博发布上新预告，吸引更多粉丝关注新品并产生购买欲，然而仅仅发布上新预告是不能达到足够的吸引力的。那么，如何才能让一则上新预告不仅带来更大的阅读量，而且能够吸引消费者认真看完。促销是一种有效手段，电商网红经常通过抽奖的方式来进行促销。通过对抽样微博的阅读，本研究发现电商网红在上新的过程中，使用抽奖的方式来吸引用户阅读、参与互动和进

行二次扩散。表 4-5 展示了抽奖微博的示例，可以发现电商网红在介绍新品的时候同时开展抽奖活动，抽奖的形式为：发起粉丝参与转发、留言、点赞三项中的一项或多项。抽奖奖品为店铺内即将售卖的新品、现金、店铺红包和其他对消费者有吸引力的实体物品。这样的抽奖微博做到了一举两得，成本低，却促进了粉丝与网红的互动，极大化了宣传效果。

表 4-5 抽奖示例

Table 4-5 Examples for lotteries

电商网红	抽奖原文
狼宝-LangBoom	12.12 我来啦！这个月很多高级的东东我的女王殿下们。👉转抽 3 个 BB 送，喜欢哪款大声告诉我，都是高级货抽中了好划算啊
呛口小辣椒	来剧透了，都在等我吗？九图一下 26 号上新的一些单品喜欢哪一款快来留言点赞哦从点赞留言中抽 3 位宝宝送一套
onlyanna	YESWOMEN 双 11 今晚凌晨 12 点准时开拍（30 款新品 8.5 折包邮）关注@yeswomen 小宜 转发评论此微博@ 三位好友抽取等奖（i6s 一台）等奖美图手机一部等奖 200 元现金卷 10 张[好棒]

4.3.2 抽奖偏好分析

微博抽奖主要分为三种参与形式：转发抽奖、评论抽奖和点赞抽奖，有时电商网红会选择超过一种形式的组合方式让用户帮助扩散上新预告微博。那么，不同的抽奖形式是否会影响网红的粉丝量或者销量呢？本节研究不同抽奖形式的比例与网红粉丝量和销量的相关性。

首先筛选出 3 种抽奖微博。本研究将微博内容中包含“抽”字和“转”字的微博筛选为转发抽奖微博；将包含“抽”字和“评”字的微博筛选为评论抽奖微博；将包含“抽”字和“赞”字的微博筛选为点赞抽奖微博；分别计算 3 种抽奖微博占原创微博的比例，基本统计结果如表 4-6 所示。转发抽奖的占比最高，均值为 3.47%，最高达 16.65%。包含评论抽奖与点赞抽奖的微博数量较少，且网红与网红之间差异很大，有的网红几乎不评论或点赞抽奖，有的网红进行大量评论和点赞抽奖。

接着，表 4-7 给出不同抽奖方式与电商网红粉丝及淘宝销量的相关性。只有转发抽奖与淘宝销量产生了中等正相关，其他配对均没有太大相关性。说明转发抽奖

可以给电商网红带来了更多的曝光率，从而增加了消费者的数量，提升了销量。而本研究发现有 10%的网红长期或间断关闭评论功能，且其中很多是销量非常高的电商网红。这可能是评论抽奖比例低，网红间差距较大，且与淘宝销量相关性较弱的主要原因。用户点赞过的微博会在用户首页显示，但不一定会在用户信息流中显示，微博复杂的推荐机制也不会把普通用户点赞过的微博优先显示在信息流顶端，这导致抽奖给网红带来的二次流量效果很差，大部分网红为了更大的流量，不使用点赞抽奖形式。

表 4-6 抽奖统计信息

Table 4-6 The statistics of reward

	转发抽奖	评论抽奖	点赞抽奖
平均值	3.47%	1.71%	1.37%
标准差	3.07%	2.38%	1.55%
最小值	0.00%	0.00%	0.00%
第一四分位	1.47%	0.18%	0.19%
中位数	2.59%	0.81%	0.78%
第三四分位	4.78%	2.26%	2.26%
最大值	16.65%	12.54%	6.03%

表 4-7 抽奖方式相关性分析

Table 3-12 The correlation analysis of lottery

	转发抽奖	评论抽奖	点赞抽奖
粉丝量	-0.055	0.022	0.042
淘宝销量	0.390	-0.016	-0.008

4.4 电商网红口碑营销行为分析

本部分主要挖掘电商点赞转发行为特点，从多个尺度分析电商网红的点赞转发目的，从而了解电商网红的商业模式。

4.4.1 转发点赞内容分析

在上文中提到，电商网红除了发布原创微博，还有点赞微博与转发微博。普通用户按照自己的兴趣和需求选择转发或点赞他们所阅读的微博，网红的转发与点

赞行为是否也是由兴趣趋势的呢？本节同样随机抽样了 10 个网红近 3 个月的转发和点赞微博来进行阅读，表 4-8 给出了被转发点赞的原微博内容示例。可以发现这些原博内容都与电商网红的商品相关。这些由非电商网红本人发出的、与电商网红商品相关的微博，叫做“买家秀”。消费者在微博原文或评论中@电商网红账号，电商网红都会收到@提醒，网红会在收到的买家秀微博中选择一部分高质量的进行转发或者点赞。这样，电商网红的粉丝能在个人的微博信息流中刷到被网红转发和点赞的微博，达到通过买家秀来实现商品二次宣传的效果。可以猜测，电商网红的转发与点赞行为与普通用户不同，他们的转发与点赞行为很可能是为了扩散买家秀微博，这是经典的口碑营销策略。

表 4-8 被转发点赞的原微博内容示例

Table 4-8 Examples for repost and like

电商网红	原作者	原微博内容
Lin 张林超	AmazingQuella	一眼就爱上的连衣裙一定要买啊 这件真的太特别了 满满的高级感@Lin 张林超
ZY 喜哥	魚寶鳗鳗	得瑟一下火箭般的预售 细节满满的精致小黑裙果然没有失望❤️@ZY 喜哥 @ZY 大暖
delicious 大金	阿 Yun0618	去年在大金买了棒球服，条纹斜挎包之前一直看中但是没有买到，还有离家出走包真的很实用,打底裤最喜欢就是不反光也不透！来一次大合集，下次要把吊牌一起存着！

4.4.2 转发点赞行为倾向分析

为了进一步分析电商网红转发点赞行为模式，需要识别出不同转发点赞行为的倾向。上一节提到，买家秀微博中会出现@网红账号的字段。于是，本研究筛选出网红转发点赞的微博原文中包含“@网红微博名”字段的内容，发现筛选出的数据只占有所有转发点赞数据的 5%，与肉眼观察结果相差较大。这说明有大量的买家秀选择在评论中@网红账号，导致了大量数据的缺失。所以，本研究采用分析原微博作者身份的方法，分析电商网红转发点赞的倾向，估计网红的转发点赞行为中属

于传播买家秀性质的行为的占比。

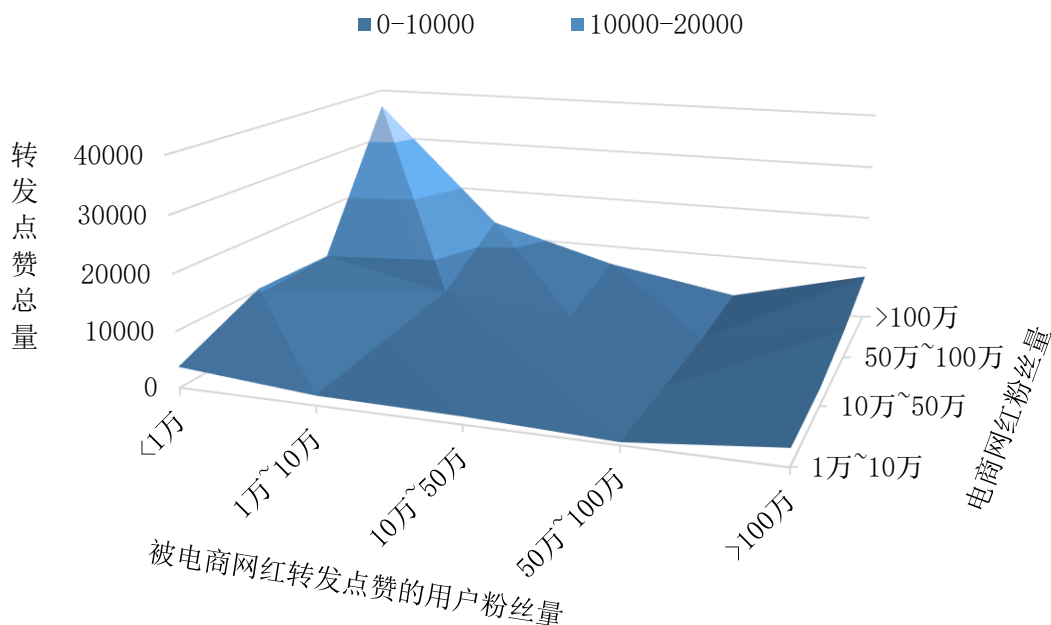


图 4-4 转发点赞不同粉丝量用户的趋势

Figure 4-4 The trend of repost and like

微博用户的身份划分与其影响力有关，粉丝量是用户在微博影响力的重要指标之一，所以本研究使用用户的粉丝量来区分用户身份。表 4-9 展示的是被 108 位网红转发和点赞过的 46470 位原作者的粉丝量统计情况，大部分原作者粉丝量很小，小部分原作者粉丝量甚至过亿。按照原作者粉丝量的分布情况，参考表 4-10

表 4-9 原作者统计信息

Table 4-9 The statistics of original authors

	粉丝量 (单位: 人)
平均值	421804
标准差	3299073
最小值	0
第一四分位	302
中位数	999
第三四分位	8735
最大值	210023472

表 4-10 电商网红统计信息

Table 4-10 The statistics of e-commerce celebrities

	粉丝量 (单位: 人)
平均值	1606223
标准差	1614619
最小值	11725
第一四分位	51829
中位数	1004896
第三四分位	2107624
最大值	6738735

电商网红的粉丝量分布情况，本研究将粉丝量划分为五个等级：粉丝量大于等于 100 万的为明星和大网红用户；粉丝量大于等于 50 万且小于 100 万的为小网红用户；粉丝量大于等于 10 万且小于等于 50 万的为潜力网红用户；粉丝量大于等于 1 万且小于等于 10 万的为潜力普通用户；粉丝量小于 1 万的为普通用户。本研究将电商网红粉丝量与被电商网红转发点赞的用户的粉丝量均按上述标准划分，那么电商网红对五种身份用户的转发点赞倾向，可以代表其行为目的。那么电商网红更喜欢转发点赞粉丝量为多少的用户呢？实验结果如图 4-4 所示，其中横向的 x 轴代表被电商网红转发点赞的用户的粉丝量分桶，纵向的 y 轴代表电商网红自身的粉丝量分桶，竖直的 z 轴代表转发点赞的总量，可以明显地看出粉丝量越高的电商网红越喜欢转发点赞普通用户的微博，且大部分电商网红都倾向于转发点赞粉丝量比自身小的用户。

表 4-11 买家秀占比统计信息

Table 4-11 The statistics of shows percent

买家秀（单位：个）	
平均值	76.36%
标准差	17.41%
最小值	28.57%
第一四分位	63.62%
中位数	80.21%
第三四分位	90.60%
最大值	98.21%

本研究进一步统计了每个电商网红每年分别转发点赞不同等级用户的数量，得到电商网红转发点赞倾向随时间变化的趋势，如图 4-5 所示。其中，(a)、(b)、(c)、(d)和(e)子图分别代表粉丝量从高到低五个等级的用户每年被电商网红点赞和转发的数量。图中横坐标为年份，纵坐标为人数，每条折线代表一个电商网红的倾向趋势，折线颜色代表电商网红的销量，销量越高颜色越红，反之则越蓝。图中可以得到以下结论：(1)从 5 张子图折线分布的高度可以看出，电商网红最喜欢转发和点赞普通用户，其次是潜力普通用户、明星和大网红用户及潜力网红用户，最不喜欢转发和点赞的是小网红用户。(2)从时间尺度上看，从 2016 年起电商网红转发和点赞普通用户及潜力普通用户的次数有明显上升趋势，而对其他级别的用户没有明显的偏好变化。(3)不同销量的电商网红转发和点赞偏好不同，具体到每个子图来说如下：(a)图显示，任何销量的网红转发点赞明星和大网红用户的倾向差不多。(b)

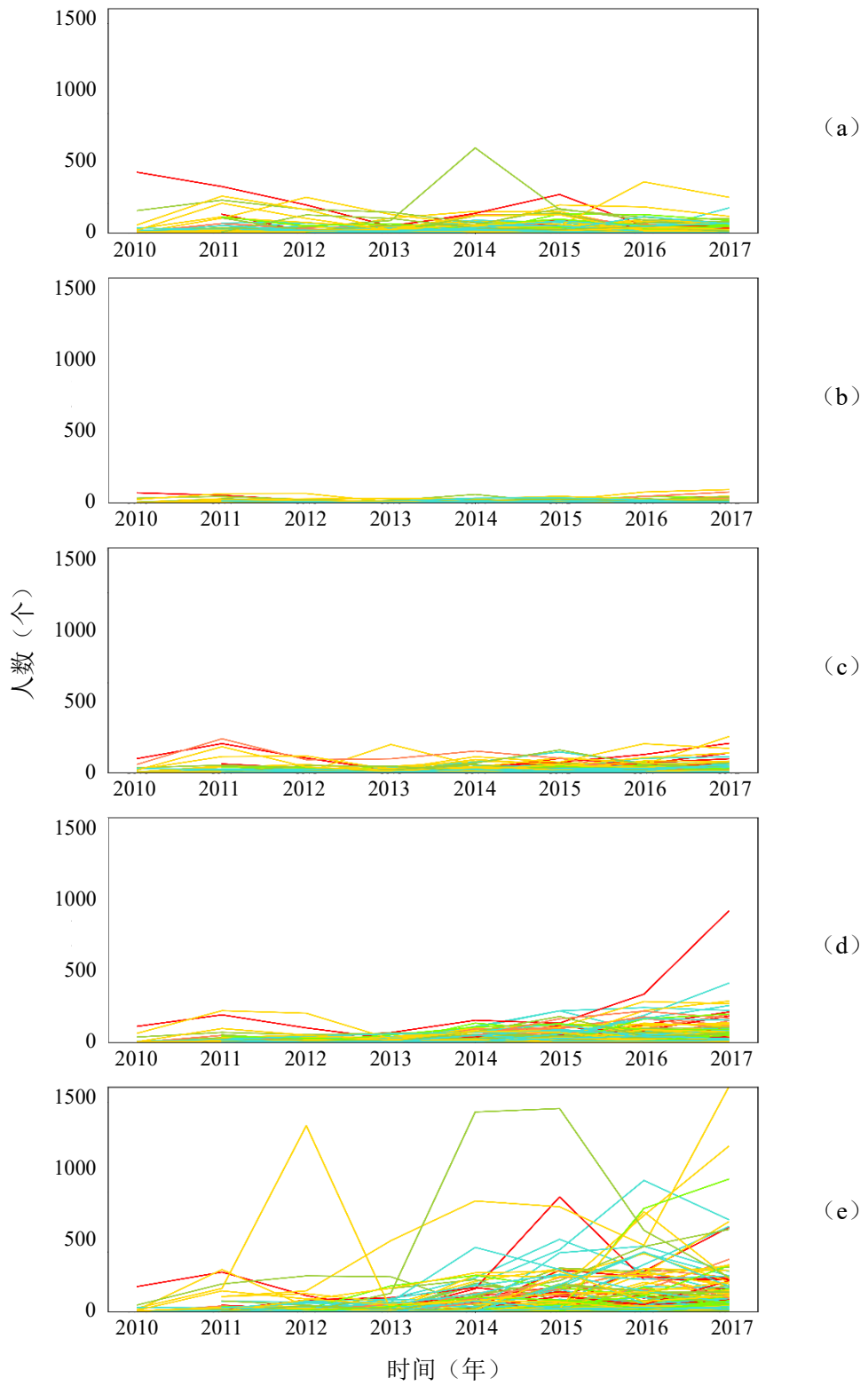


图 4-5 转发点赞不同粉丝量用户的趋势

Figure 4-5 The trend of repost and like

图显示,所有网红都不喜欢转发和点赞小网红的微博;(c)图显示销量越高的网红,越喜欢转发和点赞潜力网红用户;(d)图显示高销量、较高销量和一些销量垫底的网红越来越喜欢转发和点赞潜力普通用户的微博;(e)图显示无论销量高低,普通用户都很受电商网红欢迎。

根据上述结果,可以揣测电商网红转发点赞行为背后的意图。首先,电商网红会转发粉丝量很高的用户,这部分用户可能是电商网红关注的对象,所以电商网红会分享这些用户的微博到自己的主页上,这与普通人转发明星微博的动机相同。其次,电商网红不喜欢转发和点赞粉丝量比自身略低一个档次用户的微博,这部分用户极大可能是竞争对手,电商网红会控制自己给竞争对手带去的流量。其三,电商网红每年都会转发和点赞许多普通用户,这部分用户是给电商网红发买家秀的主力军。其四,一些高销量网红和销量很低的网红开始增加对潜力普通用户的转发点赞。潜力普通用户的粉丝量比普通人多,他们提供的买家秀可能更有吸引力,同时他们自身有更好的自媒体扩散能力。与潜力普通用户进行转发点赞互动,能够以最低成本获取最大的收益。潜力普通用户帮助高销量网红进一步稳固行业龙头地位,也能帮助低销量网红快速成长。电商网红的转发和点赞行为有明显的倾向,证明了转发与点赞是电商网红重要的营销行为之一。

最后根据上述规律,将普通用户、潜力普通用户和潜力网红用户作为主要给电商网红提供买家秀的群体,据此估算出电商网红推广买家秀行为的占比。表 4-11 给出该结果的统计情况,推广买家秀的平均占比为 76.36%。

4.5 本章小结

本章主要从电商网红微博内容出发,提取了电商网红的三个重要的营销行为,并分别对它们进行了深入分析,得出电商网红的营销行为规律。主要工作包括:(1)从原创微博内容中提取网红的上新行为,发现所有网红的上新行为间隔均为一个月,该间隔不随时间的变化而变化。(2)从原创微博内容中提取电商网红的抽奖行为数据,发现电商网红的抽奖行为往往伴随着上新行为,电商网红更偏向转发形式的抽奖且转发抽奖与淘宝销量存在中度相关性。(3)从非原创微博中提取网红的转发点赞行为,挖掘网红转发点赞的倾向,发现网红和普通人一样,也会分享自己关注的博主和内容;但同时,他们更愿意大量转发点赞低粉丝量用户的微博,这是因为低粉丝量用户为网红提供的买家秀可以为网红创造更大的商业价值;最后估算出电商网红转发点赞买家秀的比例为 76.36%。这些对网红营销行为分析揭示了网红的商业模式,为营销行为特征和其他行为特征的构建和提取提供了基础。

5 电商网红销量评估模型

本章首先基于行为分析，构建了电商网红营销行为特征和日常行为特征，基于个人信息数据提取了电商网红个人基本信息特征。接着分析了不同模式的电商网红特征的总体分布情况并给出了每个特征与销量的相关性。最后基于分类算法，使用所有特征建立了电商网红销量水平评估模型，分析了不同种类的特征对电商网红销量水平的影响，并给出最影响销量的 10 个特征量。

5.1 问题描述

第一章中提到在线商业模式以不可阻挡之势向社交商业模式转变，电商网红、网红经纪公司和各行各业都将成为社交商业趋势中的受益者。对电商网红来说，她们在社交平台上的所有表现都可能与商业利润相关，那么哪些社交行为对电商网红的商业水平提升最大？网红经纪公司又该怎样评估网红的商业价值？从哪些方面培养电商网红？

上述几个问题，归根到底，是如何设计一个基于社交数据的、可以有效评估网红销量水平关系的模型。本研究基于第 4 章的行为分析，构建电商网红的营销行为特征和日常行为特征，再结合电商网红个人基本信息特征，使用这些特征来构建电商网红的销量水平模型。通过加入不同种类的特征，观察该模型的性能变化得到不同类型特征对销量的影响。通过算法得到销量模型中重要性最高的特征，给电商网红的自我优化与网红经济公司的培养环节提供建议。

5.2 电商网红有效社交网络定义

想要深入并准确地捕捉电商网红社交特征量，必须先确定电商网红的有效社交网络。第一章提到网红对自身粉丝的自媒体的发动和管理能力更重要。第四章的研究显示，对于电商网红来说，粉丝的自媒体价值主要体现在“买家秀”上。粉丝购买电商网红产品后，会将商品以照片评语及形式发布在微博上，通过@电商网红引起电商网红的关注，电商网红会转发或点赞有“种草”价值——能够吸引其他粉丝产生购买欲望的高质量买家秀，形成一种大流量带动小流量，小流量为大流量提供低成本或无成本广告的互利互惠的运营模式。这种运营模式普遍存在于网红营销中，不难发现，转发与点赞成为网红的核心社交手段。但是，微博存在大量水军为原创微博刷评论量、转发量和点赞量的情况，这是明星运营热度

的一种常用手段，网红中也不乏有团队买水军的情况。水军识别不是本研究的重点，但水军的客观存在势必影响本研究对网红社交能力的判断。另一方面，本研究发现很多高销量电商网红（如“onlyanna”、“MALInv”等）长期或间歇性关闭了自己账号的评论功能，导致其原创微博评论量大部分都是“0”，与未关闭评论功能的网红产生巨大差距。综上所述，可以猜想电商网红与粉丝的有效社交手段并非基于其原创微博，而是依赖于网红转发和点赞其他用户微博的形式。这样的社交形式是由网红自身主导的，不仅排除了机器人“水军”刷热度的可能性，而且充分代表了电商网红的主观社交意向。所以，本章将重点研究那些被网红转发和点赞过的用户。首先定义网红的有效社交网络如下：网红的有效社交网络是以网红为中心，被网红转发过或点赞过的用户为节点，被转发和点赞的次数为边权重的图。其中，被网红转发或点赞过的用户称作有效网络用户，被网红转发或点赞过的微博称作有效微博，被转发和点赞的总次数叫做互动量。

5.3 营销行为特征量的构建

本部分基于第4章提取出的营销行为数据，根据不同营销行为的特点对电商网红的营销行为进行细分和量化，构建营销行为特征量。

5.3.1 上新行为特征量

上新行为是电商网红社交营销中的关键环节，根据第四章中的研究结果可知，电商网红的上新周期约为1个月，所以本节将时间窗口设置为1个月，从电商网红上新微博数据中提取了4种上新行为特征量，具体定义如表5-1所示。

表 5-1 上新行为特征

Table 5-1 Sale behavior features

等级	划分条件
平均上新微博量	电商网红平均每月发上新微博的数量
上新微博平均被转发量	电商网红上新微博被转发的平均量
上新微博平均被评论量	电商网红上新微博被评论的平均量
上新微博平均被点赞量	电商网红上新微博被点赞的平均量

5.3.2 抽奖行为特征量

根据4.3.2中的分析可知，电商网红的抽奖行为倾向与其销量有相关性，所

以本节将抽奖行为的不同倾向提取出来，作为抽奖行为特征量，具体定义如表 5-2 所示。

表 5-2 抽奖行为特征

Table 5-2 Reward behavior features

等级	划分条件
平均转发抽奖微博量	电商网红平均每月转发抽奖微博的数量
平均评论抽奖微博量	电商网红平均每月评论抽奖微博的数量
平均点赞抽奖微博量	电商网红平均每月点赞抽奖微博的数量

5.3.3 转发点赞行为特征量

根据 4.4 中的研究成果可知，口碑营销行为存在与电商网红的转发和点赞行为中，换句话说，即存在于电商网红的有效社交网络中。所以，本小节基于电商网红的有效社交网络，提取转发点赞行为特征量。

(1) 互动量 (Interaction)：一个有效网络用户与某电商网红的互动量定义为其在该电商网红有效社交网络中的连边权重数。大部分普通用户与电商网红的互动量均为 1，少部分头部用户与电商网红的互动量超过 4。那么网红与哪种类型的用户互动更多能提高她的销量呢？为了解决这个问题，本小节根据互动量分布图将互动量细分成 5 个等级来代表电商网红的转发点赞倾向，如表 5-3 所示。

表 5-3 互动量等级划分

Table 5-3 The level of interaction

等级	划分条件
Interaction_lv1	某有效网络中互动量为 1 的用户个数
Interaction_lv2	某有效网络中互动量为 2 的用户个数
Interaction_lv3	某有效网络中互动量为 3 的用户个数
Interaction_lv4	某有效网络中互动量为 4 的用户个数
Interaction_lv5	某有效网络中互动量大于 4 的用户个数

(2) 跨网络度 (Overlap)：一个有效网络用户的跨网络度定义为与其互动过的电商网红数。大部分用户的跨网络度是 1，与多个电商网红互动过的用户里有少部分用户的跨网络度超过了 50。本研究将跨网络度为 1 的用户称为专一型用户，将跨网络度超过 1 的用户称为圈用户。电脑上网红与哪种类型的用户互动能够带来更多的商业利益？跨网络度为多少的用户能提供最大的帮助？为了解决

这个问题,本小节根据跨网络度分布图将跨网络落度细分成 5 个等级来代表电商网红的转发点赞倾向,如表 5-4 所示。

表 5-4 跨网络度等级划分

Table 5-4 The level of overlap

等级	划分条件
Overlap_lv1	某有效网络中跨网络度为 1 的用户个数
Overlap_lv2	某有效网络中跨网络度大于 1, 小于等于 10 的用户个数
Overlap_lv3	某有效网络中跨网络度大于 10, 小于等于 30 的用户个数
Overlap_lv4	某有效网络中跨网络度大于 30, 小于等于 50 的用户个数
Overlap_lv5	某有效网络中跨网络度大于 50 的用户个数

(3) 用户粉丝量 (Fans of alters): 定义为电商网红的有效社交网络中的用户的粉丝量。用户的粉丝量代表用户自身的影响力。少部分用户粉丝量达到千万,而大部分用户粉丝量都在 10000 以下。那么,网红与大粉丝量用户社交能带来商业利益吗? 哪种粉丝规模的用户能带来利益最大化? 为了解决这个问题,本小节将采用与 4.4.2 中相同的划分规则,将用户粉丝量分成 5 个等级来代表电商网红的转发点赞倾向,如表 5-5 所示。

表 5-5 用户粉丝量等级划分

Table 5-5 Fans level of alters

等级	划分条件
Fans_lv1	某有效网络中用户粉丝量小于 10000 的用户个数
Fans_lv2	某有效网络中用户粉丝量大于等于 10000, 小于 100000 的用户个数
Fans_lv3	某有效网络中用户粉丝量大于等于 100000, 小于 500000 的用户个数
Fans_lv4	某有效网络中用户粉丝量大于等于 500000, 小于 1000000 的用户个数
Fans_lv5	某有效网络中用户粉丝量大于等于 1000000 的用户个数

(4) 用户关注量 (Follows of alters): 定义为电商网红的有效社交网络中的用户关注其他账号的数量。用户的关注量代表其兴趣面的广度,也是用户在社交平台上的活跃程度的表现之一。同样的,哪种关注规模的用户能为网红带来最多的商业利益? 为了解决这个问题,本小节将用户关注量按照分布情况分成 4 个等级

来代表电商网红的转发点赞倾向，如表 5-6 所示。

表 5-6 用户关注量等级划分

Table 5-6 Fans level of alters

等级	划分条件
Follows_lv1	某有效网络中用户关注量小于 100 的用户个数
Follows_lv2	某有效网络中用户关注量大于等于 100，小于 500 的用户个数
Follows_lv3	某有效网络中用户关注量大于等于 500，小于 1000 的用户个数
Follows_lv4	某有效网络中用户关注量大于等于 1000 的用户个数

(5) 用户微博量 (Posts of alters): 定义为电商网红的有效社交网络中的用户微博总量，包括其转发和点赞微博。用户的微博量是用户在社交平台上的活跃程度的重要表现。越活跃的用户越能给电商网红带来商业价值吗？为了解决这个问题，本小节同样将用户微博量按照分布情况分成 4 个等级来代表电商网红的转发点赞倾向，如表 5-7 所示。

表 5-7 用户微博量等级划分

Table 5-7 Posts level of alters

等级	划分条件
Posts_lv1	某有效网络中用户微博量小于 100 的用户个数
Posts_lv2	某有效网络中用户微博量大于等于 100，小于 1000 的用户个数
Posts_lv3	某有效网络中用户微博量大于等于 1000，小于 5000 的用户个数
Posts_lv4	某有效网络中用户微博量大于等于 5000 的用户个数

(6) 有效微博平均被转发量 (Average repost of valuable posts): 定义为某电商网红的有效社交网络中，有效微博下转发量的平均值。

(7) 有效微博平均被评论量 (Average comment of valuable posts): 定义为某电商网红的有效社交网络中，有效微博下评论数的平均值。

(8) 有效微博平均被点赞量 (Average like of valuable posts): 定义为某电商网红的有效社交网络中，有效微博下点赞量的平均值。

(9) 平均转发微博量(Average Repost): 定义为电商网红平均每月转发的微博数量。平均转发微博量越大，则电商网红越重视转发形式，转发中很可能包含大量的买家秀。

(10) 平均点赞微博量(Average like): 定义为电商网红平均每月点赞的微博数量。平均点赞微博量越大，则电商网红越重视点赞这项功能，点赞中很可能包含

大量的买家秀。

其中, (6)、(7) 和(8)三个特征量都反映了有效微博的传播能力, 可以进一步理解为一个有效网络中所有用户能够给电商网红带来的宣传能力; (9)和(10)则反映了电商网红转发点赞的倾向。

5.4 日常行为特征量的构建

第4章中提到, 电商网红的社交行为分为营销行为与日常行为。除了营销行为为特征外, 日常行为也可能影响电商网红的销量。所以, 本研究按照构建营销行为特征的思路, 从电商网红的日常行为数据中构建了以下几个特征量, 如表 5-8 所示:

表 5-8 日常微博行为特征量

Table 5-8 Non-commercial features of weibo

等级	划分条件
平均日常微博量	电商网红每月原创微博中, 非上新和抽奖微博的平均数量
日常微博平均被转发量	电商网红每月原创微博中, 非上新和抽奖微博的平均数量被转发的平均量
日常微博平均被评论量	电商网红每月原创微博中, 非上新和抽奖微博的平均数量被评论的平均量
日常微博平均被点赞量	电商网红每月原创微博中, 非上新和抽奖微博的平均数量点赞的平均量

除了行为特征, 电商网红的个人基本信息也可能影响电商网红的销量。本研究爬取的电商网红数据集中包含电商网红以下基本信息: 粉丝量、关注量、vip 等级和所在地。由于电商网红的 vip 等级差异与地域差异较小, 故本研究只选取粉丝量和关注量用于社交特征量与销量的建模。

5.5 社交特征量分析

5.3 节与 5.4 节共构建和提取了 35 个电商网红的营销行为特征量、4 个日常行为特征量和两个个人基本信息特征量, 一共 41 个特征量。接下来本节将分析不同模式的电商网红在这些特征上的表现情况, 以及特征与销量的相关性。

5.5.1 基于聚类的电商网红模式分析

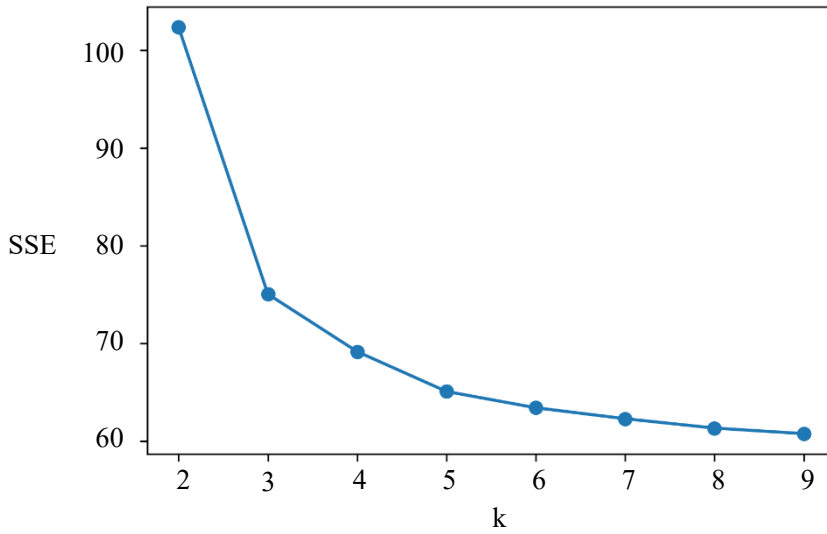


图 5-1 k 值对 SSE 的影响

Figure 5-1 The influence of k on SSE

建立社交特征与销量的模型前，本研究使用聚类分析对现有电商网红特征进行探索性的分析。聚类分析是数据挖掘的主要任务之一，它能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步地分析。本研究使用 k 均值聚类算法（k-means）对 108 位电商网红进行聚类分析，探索聚类结果的潜在含义以及不同簇的电商网红的特征分布情况。

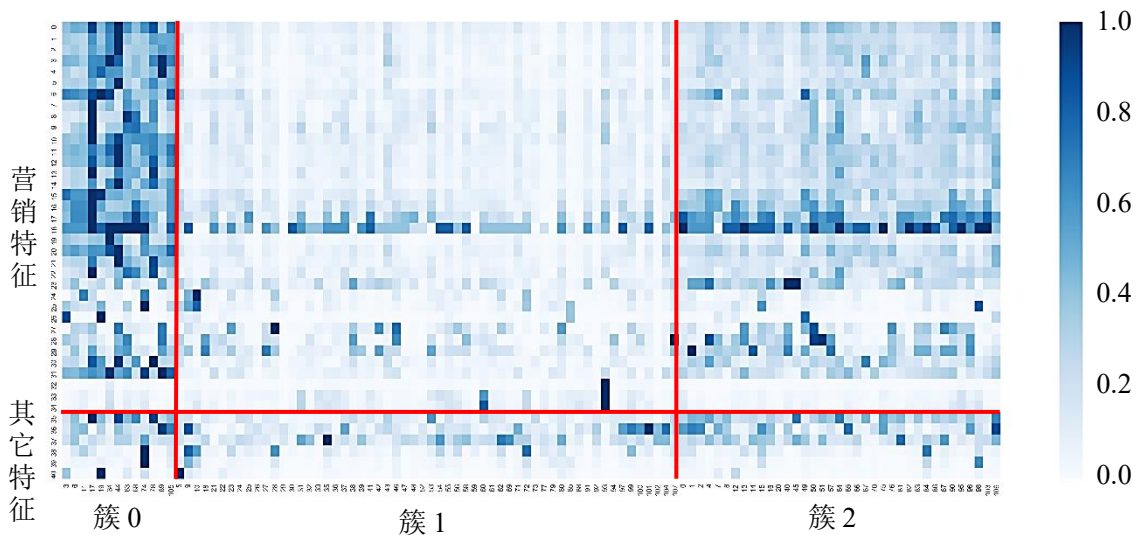


图 5-2 电商网红特征热力图

Figure 5-2 The feature heatmap of all e-commerce celebrities

k-means 聚类需要给定簇数目 k ，通常使用手肘法进行最佳 k 值的选取，图 5-1 表示 k 的取值对 k-means 算法的误差平方和的影响。可以看到“手肘”出现在 $k=3$ 处，故最佳的 k 值为 3。

设定 k 为 3，k-means 将 108 个网红聚为 3 个簇，聚类结果如图 5-2 所示。图 5-2 是不同簇的电商网红的特征热力图，横坐标小字为 108 位网红的编号 id，按照聚类的结果排序，从左到到右依次是簇 0、簇 1 和簇 2。纵坐标为特征编号，按照特征类型排序，从上到下依次是营销特征和其他特征。作图时，对所有特征数值作归一化，图中每一个小方块的深浅代表特征的数值大小水平，颜色越深，该特征数值越大，反之则数值越小。从图中不难看出，3 个簇的主要区别是特征大小整体水平不同，簇 0 的电商网红各特征水平较高，簇 2 次之，簇 1 最低。其中特征水平区分度最大的特征块集中在 0~22 号特征上，这 23 个特征全部属于营销行为中的转发点赞行为特征，从 0 到 22 依次是互动量特征 5 个，用户粉丝量特征 5 个，用户关注量特征 4 个，跨网络度特征 5 个，用户微博量特征 4 个。这说明电商网红的转发点赞行为确实存在很明显的倾向，簇 0 的网红转发点赞行为非常频繁，簇 1 的网红转发点赞则最不活跃。簇 2 的网红尤其注重转发点赞跨网络度高的用户微博。图 5-3 表示按簇排序的 108 位网红的第一次上新年份的分布图，其中横坐标是网红编号 id，不同的颜色代表不同的簇，从左到右依次是簇 0、簇 1 和簇 2，纵坐标代表第一次上新年份。网红的第一次上新年份可以认为是网红开始电商活动的年份。从图中可以看到，簇 0 的用户大部分在 2011 年前就展开了电商活动，簇 2 的用户大部分在 2014 年后才开始电商活动，簇 1 的用户或早或晚开始电商经营，但总体都在 2011 年后才开始电商活动。结合图 5-2 的结果，可以将电商网红分为以下三个模式：

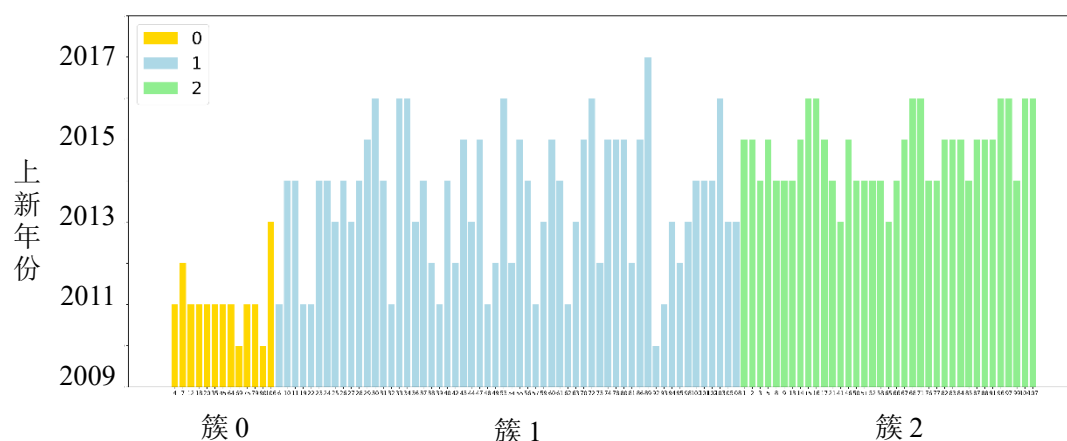


图 5-3 首次上新年份分布图

Figure 5-3 The distribution for the year of the first new product display

(1) 模式 0: 聚类后被分在簇 0 的电商网红。这类网红很早就开展电商活动，且在各个社交环节中都很活跃，尤其专注于口碑营销。

(2) 模式 1: 聚类后被分在簇 2 的电商网红。这类网红开展电商活动的时间较晚，但在各个社交环节中较为活跃，尤其专注于转发点赞活跃在多个网红社交圈的用户。

(3) 模式 2: 聚类后被分在簇 1 的电商网红。这类网红开展电商活动的时间或早或晚，在所有特征上的表现都不活跃。

下面本研究将探索上述聚类结果与电商网红销量的关系，图 5-4 给出了销量排名前 18 位的电商网红特征热力图，其中横坐标的电商网红编号按照该网红所

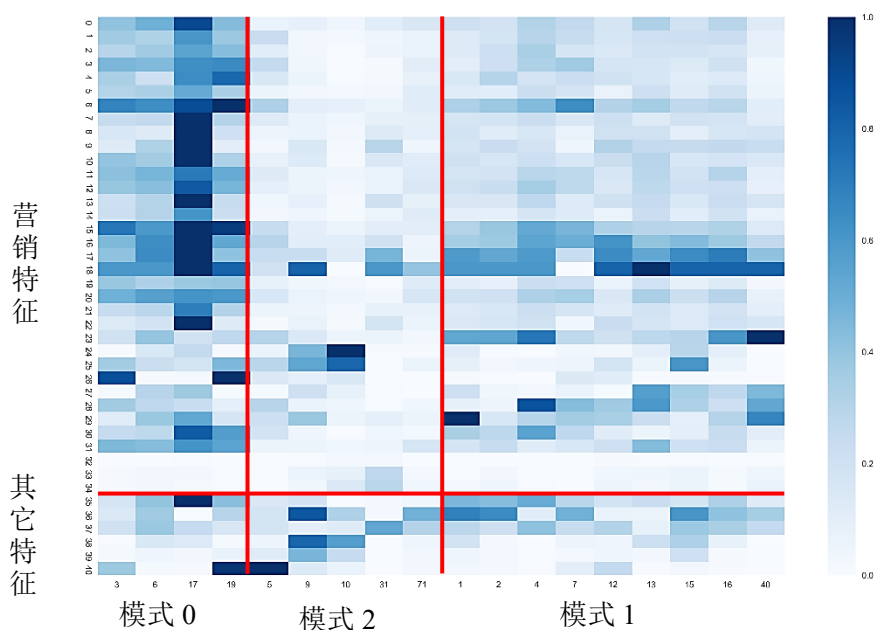


图 5-4 高销量电商网红特征热力图

Figure 5-4 The feature heatmap of high sales e-commerce celebrities

属的模式排序，从左到右依次是：模式 0、模式 2 和模式 1。可以看出销量与聚类出的模式之间没有明显相关性，3 个模式的电商网红都有可能成为高销量电商网红。但模式 0 与模式 1 的电商网红占比更大，说明社交行为活跃的网红更有可能成为高销量的电商网红。

5.5.1 社交特征量与销量的相关性分析

上一节通过聚类和特征热力图的方式给出了电商网红在所有特征上的数值分布情况，本节将继续细化特征分析，将每一个特征与销量直接作相关性分析，探索哪些特征在数值上与销量相关性最大。

表 5-9 上新行为特征相关性分析

Table 5-9 The correlation analysis of sales behavior features

上新行为	相关系数
平均上新微博量	0.268
上新微博平均被转发量	0.301
上新微博平均被评论量	0.092
上新微博平均被点赞量	0.187

表 5-10 抽奖行为特征相关性分析

Table 5-10 The correlation analysis of reward behavior features

抽奖行为	相关系数
平均转发抽奖微博量	0.416
平均评论抽奖微博量	0.045
平均点赞抽奖微博量	0.012

表 5-11 口碑营销行为特征相关性分析

Table 5-11 The correlation analysis of word of mouth marketing behavior features

社交特征量，相关系数									
互动量		跨网络度		用户粉丝量		用户关注量		用户微博量	
Interaction_lv1	0.198	Overlap_lv1	-0.125	Fans_lv1	0.116	Follows_lv1	0.174	Posts_lv1	0.152
Interaction_lv2	0.036	Overlap_lv2	0.276	Fans_lv2	0.346	Follows_lv2	0.172	Posts_lv2	0.188
Interaction_lv3	0.020	Overlap_lv3	0.013	Fans_lv3	0.358	Follows_lv3	0.294	Posts_lv3	0.136
Interaction_lv4	0.322	Overlap_lv4	-0.025	Fans_lv4	0.172	Follows_lv4	0.185	Posts_lv4	0.159
Interaction_lv5	0.304	Overlap_lv5	-0.085	Fans_lv5	0.187				
有效微博平均被转发量	0.121	有效微博平均被评论量	0.308	有效微博平均被点赞量	0.184	平均转发微博量	0.442	平均点赞微博量	0.294

本节首先分析营销为特征量与电商网红淘宝销量之间的相关性。如表 5-9、和表 5-11 所示，上新行为特征中，上新微博平均被转发量相关系数达 0.301，其他特征量相关性较弱，相关性最低的特征量是上新微博平均被评论量。新微博平均被转发量相关性强，说明粉丝转发上新微博的行为是其购买欲的象征，网红上新微博被转发得越多，商品可能越受到粉丝的欢迎，销量也就越高。

其次如表 5-10 所示，抽奖行为特征量中，平均转发抽奖微博量达到中度相关，另外两种倾向的特征量则相关性则很弱。由于这组特征量是 4.4.2 中研究对象的均值版本，所以与 4.4.2 中研究结果不谋而合。每个月转发抽奖微博的数量越多，销量越高，这可能是由于转发抽奖能激发用户的购买兴趣。

接着，分析转发点赞行为特征量，如表 5-11 所示。互动量中 Interaction_lv4 和 Interaction_lv5 与电商网红销量相关性最高。这说明互动量越高的用户越能给网红带来商业价值。跨网络度中 Overlap_lv2 的相关性最高，但仍没有达到中度相关。用户粉丝量中 Fans_lv2 和 Fans_lv3 与电商网红销量相关性最高，接近中等相关。如 4.4.2 中的研究结果类似，高粉丝量的用户不一定为网红带来更高的商业利益，相比之下中等粉丝量与低粉丝量的一部分用户可以带来更有商业意义的有效互动。用户关注量中，最高的是 Follows_lv3。关注数较多的用户，可能是使用微博更加活跃的人群，这部分人也可能影响电商网红的销量。用户微博量与电商网红相关性较弱。有效微博平均被转发量、被评论量和被点赞量中，被评论量相关性较高。经常被评论的用户，可能给电商网红形成更好地二次宣传效果，从而影响了电商网红的销量。平均转发微博量的相关性达到中度相关，高于平均点在微博量的相关性，这可能代表转发是比点赞更好的口碑营销方式。

表 5-12 日常行为特征及个人基本信息特征相关性分析

Table 5-12 The correlation analysis of daily behavior features and information features

上新行为	相关系数
关注量	0.118
粉丝量	0.507
平均日常微博量	0.212
日常微博平均被转发量	0.228
日常微博平均被评论量	-0.026
日常微博平均被点赞量	0.304

最后，分析日常行为特征量和个人信息特征量与电商网红淘宝销量之间的相关性，如表 5-12 所示。其中粉丝量与销量的相关性是所有特征中最高的，其次是平

均日常微博量与日常微博平均被点赞量。从相关性分析中可以看出日常行为特征量和个人信息特征量也会影响电商网红的销量。

本节提出的 41 个特征量,其分别与淘宝销量的相关性都处于中度相关或以下,没有强相关量,这说明单独一个或少数几个特征都无法决定电商网红的销量。下一节中本研究使用机器学习模型来分析特征量组合后的效果。

5.6 电商网红淘宝销量预测模型

单独一个或几个社交特征量不足以决定电商网红的销量,所以本节将使用预测的方法,使用 41 个社交特征构建电商网红淘宝销量预测模型。首先,本节介绍模型的框架结构以及具体的操作过程;其次,对比不同机器学习算法的性能优良,选出最优的机器学习算法;最后,计算每个特征参数的重要性,找到影响电商网红淘宝销量的最重要因素。

5.6.1 模型构建

正如 5.1 节中所说,理解电商网红的微博社交行为,使用其行为特征和个人信息特征预测电商网红的淘宝销量,找到影响销量的关键特征,有助于帮助电商网红提升自己社交营销中的不足。例如,对于销量低的网红,她可能需要通过运营帮她提升粉丝量,全面增加与粉丝的互动。对于销量较高的网红,则需要细化到社交的各个行为来进行优化。通过 5.5 的特征分析本研究又发现,不同模式的网红特征数值分布不同,但网红的模式并不对应网红的销量高低;另一方面,从特征与销量的相关性分析中可知,没有强相关的特征量,无法单独使用一个或几个特征量建立销量模型。所以,需要使用其他机器学习方法来挖掘特征与特征之间的隐含关系,建立所有社交特征与销量水平之间的模型。本研究选择使用分类的方法来构建这个模型。

首先对电商网红销量高低进行分桶定义。电商网红销量的分布大致呈现 3 个阶层,划分条件如表 5-13 所示。因此,电商网红按照销量被分成“高销量网红”、“中销量网红”和“低销量网红”,通过电商网红社交数据预测其淘宝销量就变成了一个三分类问题。为了训练分类器,本节用三元变量“0”、“1”和“2”分别代表“高销量网红”、“中销量网红”和“低销量网红”。

最终本文将电商网红自注册微博账号以来的所有微博特征送入分类器进行淘宝销量的预测。数据包含 31 名“低销量网红”,59 名“中销量网红”,18 名“高销量网红”,三类数据的比例大约为 1.7:3.3:1。模型训练集和测试集的比例设置为 6:4,

即全部数据的 60%用来训练模型，剩下的 40%用来测试模型的准确性。

表 5-13 电商网红分类

Table 5-13 The classification of e-commerce celebrities

类型	划分条件
高销量网红	销量大于 (均值 + 1/2 标准差)
中销量网红	销量大于 (均值 - 1/2 标准差) 且小于 (均值 + 1/2 标准差)
低销量网红	销量大于等于 (均值 - 1/2 标准差)

5.6.2 模型性能评估

本节分别采用了逻辑回归、随机森林、KNN 和三种不同的分类器并且对比了不同分类器的性能。在本节预测电商网红淘宝销量的三分类问题中，本研究使用第二章介绍的混淆矩阵来计算精确率和召回率两种指标来进行性能评价。由于三个类别都有各自的精确率与召回率，所以本节将整个模型的精确率和召回率分别用它们的均值来表示，具体如下：

$$precision_{average} = (precision_{high} + precision_{medium} + precision_{low})/3 \quad (5-1)$$

$$recall_{average} = (recall_{high} + recall_{medium} + recall_{low})/3 \quad (5-2)$$

本文的实验通过 Scikit-Learn 中的分类器进行，分为以下五个对比实验。

(1) 全时段非营销行为特征模型：使用日常行为特征和个人基本信息特征训练分类模型，探究非营销行为特征对电商网红淘宝销量水平预测的影响。

(2) 全时段营销行为特征模型：使用营销行为特征训练分类模型，探究营销行为特征对电商网红淘宝销量水平预测的影响。

(3) 全时段全特征模型：分类模型中将输入所有类型特征量，探究所有特征量加入后预测性能的变化情况。

(4) 近三年全特征模型：分类模型中输入 2014.4.30-2017.4.30 这三年数据对应的所有特征量。与全时段特征模型与近一年特征模型相比，近三年全特征模型评估网红最近三年的综合表现对预测的效果。

(5) 近一年全特征模型：分类模型中将输入使用 2016.4.30-2017.4.30 这一年数据所挖掘的所有特征量。与全时段特征模型相比，近一年全特征模型评估网红最近一年的综合表现对预测的效果。

表 5-14 罗列了四个对比实验的模型性能，表中分别给出了每个模型在每个实验下预测每个类别的精确率、召回率和均值，可以得到如下结论：(1) 当模型中只用营销行为特征量和只用非营销行为特征量时，效果最不好的模型是逻辑回归，但

表 5-14 模型预测性能

Table 5-14 The prediction performance of different models

		全时段 非营销行为特征 模型	全时段 营销行为特征 模型	全时段 全特征模型	近三年 全特征模型	近一年 全特征模型
		精确率, 召回率				
LR	高销量用户	0.40,0.40	0.43,0.50	0.75,0.60	0.75,0.60	0.75,0.60
	中销量用户	0.73,0.58	0.80,0.67	0.71,0.63	0.76,0.68	0.74,0.74
	低销量用户	0.46,0.67	0.45,0.56	0.50,0.67	0.50,0.67	0.60,0.67
	均值	0.53,0.55	0.56,0.57	0.65,0.63	0.67,0.65	0.70,0.67
RF	高销量用户	0.50,0.60	0.50,0.50	0.67,0.80	0.80,0.80	0.78,0.74
	中销量用户	0.75,0.63	0.71,0.63	0.83,0.79	0.84,0.84	0.89,0.84
	低销量用户	0.55,0.67	0.58,0.70	0.78,0.78	0.78,0.78	0.82,0.90
	均值	0.60,0.63	0.60,0.61	0.76,0.79	0.81,0.81	0.83,0.83
KNN	高销量用户	0.60,0.50	0.50,0.57	0.67,0.67	0.67,0.67	0.83,0.83
	中销量用户	0.75,0.67	0.71,0.67	0.74,0.67	0.82,0.78	0.79,0.83
	低销量用户	0.50,0.67	0.60,0.60	0.75,0.67	0.70,0.78	0.75,0.67
	均值	0.62,0.61	0.60,0.61	0.72,0.70	0.73,0.74	0.79,0.78

仍旧高于 0.5，其余两个模型效果差不多，准确率与召回率都超过了 0.6。综合来说，只用一种特征量的分类效果较差，没有孰强孰弱之分；(2)全特征量模型的性能比只用非营销行为特征量和只用营销行为特征量模型好很多，全时段数据的精确率均值与召回率均值最高可达 0.76 与 0.79。说明营销行为特征量、日常行为特征量和个人基本信息特征量在销量水平分类模型中相互补充了信息量，所以利用所有特征量的模型效果最好。(3)几乎所有情况下，随机森林的表现都比其他分类器更好。尤其是特征维度增加的情况下，随机森林有很大优势。(4)全时段、近三年和近一年的实验对比中，使用近一年的全特征效果最好，精确率均值与召回率均值最高可达 0.83。说明网红的销量表现有时效性，使用较新的数据比使用累积数据达到更好地预测效果。

表 5-15 特征量重要性排名 Top10

Table 5-15 Top 10 most important features

排名	特征	重要性
1	Interaction_lv5	6.43%
2	平均转发微博量	6.29%
3	粉丝量	5.87%
4	日常微博平均被点赞量	5.81%
5	Interaction_lv4	4.65%
6	Overlap_lv3	4.39%
7	Fans_lv2	4.21%
8	平均上新微博量	3.97%
9	有效微博平均被评论量	3.72%
10	平均转发抽奖微博量	3.51%

随机森林给出了特征量的重要性排名 Top10，如表 5-15 所示。这 10 个特征量重要性之和约占总 41 个特征量的 50%，特征量之间重要性相差不大。Top10 中有 8 个营销行为特征，1 个日常行为特征量，1 个人基本信息特征量。且排名第一的是营销行为特征，证明了营销行为特征的重要性。

首先对入围排名的营销行为特征进行深入分析。与电商网红销量水平最相关的特征量是 Interaction_lv5，即互动量大于 4 次的用户数。另外重要性排名第五的 Interaction_lv4 也是同类特征，这说明高互动量的用户最影响网红销量水平。平均转发微博量排名第二，说明口碑营销行为中，基于转发的方式更能影响电商网红的销量水平。Overlap_lv3 和 Fans_lv2 分别排名第六第七，说明跨网络度在(10,30]区间内和粉丝量在(10000,100000]区间的用户有较大影响力。经筛查验证发现跨网络

度高于 30 和粉丝量高于十的用户大部分是微博大 V 和以内容运营为主的媒体账号。这些用户的粉丝基数更大, 粉丝面广, 在网红圈中有较大跨网络度是正常的。网红转发和点赞这些用户的微博, 对销量贡献很小。相比之下, 中等跨网络度和粉丝规模处在“潜力普通用户”更可能与电商网红产生互利互惠的商业社交行为, 所以他们对销量预测模型的贡献也较大。平均上新微博量排第八, 这说明多发上新预告微博, 或者上新更多的商品都可能影响网红的销量水平。有效微博平均被评论量排名第九, 这说明用户所发的与网红产品相关的种草微博中, 评论量最能代表这条“买家秀”的价值。平均转发抽奖微博量排名第十, 说明转发抽奖确实是能够带动用户传播网红微博, 提高用户购买欲望的有效社交互动形式。

最后分析非营销行为特征。排名第四的是日常微博平均被点赞量, 说明粉丝给网红日常微博的点赞行为与他们是否购买网红的产品有很大的关系。结合第四章中的研究结果, 有一部分网红关闭了评论功能, 导致这个特征不够“公平”, 可能导致了日常微博平均被评论量的重要性很低。另一方面, 在转发、评论和点赞三种行为中, 点赞的交互更简单。没有任何弹窗和页面跳转, 粉丝只需要点一下点赞按钮就可以完成点赞行为。简单的交互使得点赞成为粉丝日常支持电商网红的方式, 这很有可能代表着黏性较高的粉丝的一种行为习惯, 所以对日常微博互动来说, 点赞是最好的互动方式。粉丝量仅排名第三, 正如其他文献中推测的那样, 粉丝量固然影响网红的商业利润, 但并不一定是决定性因素。一方面, 网红的粉丝中存在“僵尸号”这种作弊成分; 另一方面, 网红使用社交平台与粉丝的互动能力和其利用特定社交行为对粉丝的调动能力是电商网红得以成功的更重要的因素。

5.7 本章小结

本章主要基于行为数据构建了电商网红的在微博这个社交平台上的营销行为特征和日常行为特征, 基于用户基本信息数据提取了电商网红基本信息特征。分析了这些特征的总体分布, 探究了这些特征与电商网红淘宝销量水平的相关性, 并介绍了利用这些特征预测电商网红淘宝销量水平的方法和模型性能, 最后给出了特征的重要性排名。具体来说: (1)提取了电商网红的营销行为特征, 并观察了各特征量与淘宝销量的相关性, 发现上新微博平均被转发量、平均转发抽奖微博量、Fans_lv2、Interaction_lv4、平均转发微博量和有效微博平均被评论量相关性较高。(2)构建了电商网红的日常行为特征, 提取了电商网红的个人基本信息特征, 并分析了它们与淘宝销量的关系, 发现粉丝量与日常微博平均被点赞量与淘宝销量相关性最高。(3)通过聚类分析与相关性分析证实了使用一个或少量特征建立销量水平评估模型不可行, 然后利用机器学习分类算法和网红近一年, 近三年, 和全部历

史数据对网红淘宝销量进行预测。实验结果显示，当模型中只有营销行为特征或只有日常行为特征和个人基本信息特征时，性能不是非常好，但高于 50%。全时段实验中同时使用所有特征量的全特征模型的性能达到最优。说明营销行为特征、日常行为特征与个人基本信息特征在网红淘宝销量预测模型中互相补充，使得模型效果 0.76。另外，几乎所有情况下，随机森林比其他分类器的综合表现更好。且使用近期的社交数据会比使用长年累积数据达到更好地模型性能，说明网红的社交数据具有较强的时效性。(4)通过随机森林算法给出模型中重要性排名前十的特征。此研究有助于电商网红衡量自己现有的能力，参照特征重要性排名可以有针对性地指出电商网红提升自己社交运营中的不足。

6 结论

本章主要对本文的主要贡献进行了详细的总结，然后对未来的工作内容以及待研究问题进行了展望。

6.1 本文工作总结

本部分主要对本文工作成果进行总结归纳，关键性研究成果分为三个部分：一部分是测量并获取了电商网红数据集，而是基于电商网红社交网络数据，提取出电商网红的营销行为和日常行为数据，揭示了电商网红在社交平台上的营销行为规律。三是基于行为分析，构建了营销行为特征量和日常行为特征量，结合网红的个人基本信息特征量。基于分类的方法，使用这些特征建立了电商网红淘宝销量水平的评估模型。下面对这三部分成果进行重点阐述。

6.1.1 电商网红社交与销量数据的获取

本部分主要总结了本文测量与获取电商网红数据集方面的工作成果，具体的贡献包括以下几个方面：

(1)综合各类数据报告与微博博主经验总结，尽可能多地列举符合研究条件的电商网红，获得电商网红列表；

(2)在无需账号的情况下跳过 Sina Visitor System，从根源上解决了爬虫时使用的微博账号会被冻结的问题；

(3) 基于 Python，通过解析微博用户搜索页面获取电商网红的基本信息；通过解析微博个人主页 Ajax 请求的响应结果，获取电商网红在微博平台从 2010 年 1 月 1 日至 2018 年 4 月 30 日的所有社交数据；

(4) 使用(2)中相同的策略，获取了被电商网红转发和点赞过的 46470 位微博用户的基本信息，作为后续研究的补充数据；

(5) 使用 Web Scraper 插件设计淘宝爬虫程序，避开淘宝的反爬措施，获取电商网红的销量数据。

以上研究为电商网红基于实际网络测量数据的定量分析提供了基础。

6.1.2 电商网红营销行为测量与分析

本部分主要总结了本文从电商网红微博内容出发，提取了电商网红的三个重要的营销行为，并分别对它们进行了深入分析，挖掘电商网红的营销行为规律方面的工作成果，具体的贡献包括以下几个方面：

(1) 从原创微博内容中提取网红的广告行为——上新行为，发现所有网红的上新行为周期围绕着 28 天上下小幅度波动。

(2) 从原创微博内容中提取电商网红的促销行为——抽奖行为，发现电商网红的抽奖行为往往伴随着上新行为，电商网红更偏向转发形式的抽奖，且转发抽奖与淘宝销量呈中度相关。

(3) 从非原创微博中提取网红的口碑营销行为——转发点赞行为，挖掘网红转发点赞的倾向，发现网红和普通人一样，也会分享自己关注的博主和内容；但同时，他们更愿意大量转发点赞低粉丝量用户的微博，这是因为低粉丝量用户为网红提供的买家秀可以为网红创造更大的商业价值；最后估算出电商网红转发点赞买家秀的比例为 76.36%。

以上研究成果在社交商业化快速推进的今天，为各行各业的人揭示了网红的商业模式，这是社交商业化向其他领域推进的基础。只有了解电商网红在社交平台上的营销行为，才能在各行各业中，成功复制甚至优化电商网红的商业模式。

6.1.3 电商网红销量评估模型

本部分主要总结本文基于电商网红行为分析，构建了电商网红的营销行为特征与日常行为特征，从个人信息数据中提取了电商网红个人基本信息特征。分析了这些特征的总体分布情况，探究每个特征与电商网红淘宝销量的相关性，并介绍了利用这些特征预测电商网红淘宝销量水平的方法和模型性能，最后给出了特征的重要性排名等方面的工作成果。具体的贡献包括以下几个方面：

(1) 构建了 35 个电商网红营销行为特征，并分别对各个特征与淘宝进行相关性分析，发现上新微博平均被转发量、平均转发抽奖微博量、Fans_lv2、Interaction_lv4、平均转发微博量和有效微博平均被评论量是相关性较高的特征。

(2) 构建了电商网红的日常行为特征，提取了电商网红的个人基本信息特征，并分析了它们与淘宝销量的关系，发现粉丝量与日常微博平均被点赞量与淘宝销量相关性最高。

(3) 通过聚类分析与相关性分析证实了使用一个或少量特征建立销量水平评估模型不可行，然后利用机器学习分类算法和网红近一年，近三年，和全部历史数据

对网红淘宝销量进行预测。实验结果显示,使用所有特征量的全特征模型的性能比单独使用营销行为特征或者其他两类特征要好。说明营销行为特征、日常行为特征与个人基本信息特征在网红淘宝销量预测模型中互相补充。另外,几乎所有情况下,随机森林比其他分类器的综合表现更好。且使用近期的社交数据会比使用长年累积数据达到更好地模型性能,最高性能可达 0.83。说明网红的社交数据具有较强的时效性。

(4)通过随机森林算法给出模型中重要性排名前十的特征。

以上研究有助于网红经纪公司衡量网红的销量水平,或者帮助想成为电商网红的人衡量自己现有的能力,特征重要性排名可为电商网红社交优化方向和电商网红培养方向提供参考。这一研究具有理论意义和实用价值。

6.2 未来工作展望

目前本研究只能使用三个等级来评价电商网红的销量,然而准确预测出电商网红销量的具体数值是更具应用意义的。所以本文未来的第一个工作是训练出一个预测电商网红销量的高性能回归模型。另一方面,随着社交商业化的一步步推进,各行各业都需要利用社交数据来优化商业模式。电商网红的商业模式相对简单,具有良好的参考价值,但无法直接复制到每个商业场景中。如何以电商网红商业模式为基础,融合其他商业场景,构建更复杂的业务模型必将成为未来的探索热点。因此,本文未来的第二个工作重点将会放在如何把网红销量预测模型投射到社交商业模式中,增加社交行为对消费者体验、消费者黏性的影响分析,使用消费者特征丰富社交商业模式。

参考文献

- [1] 陈艺文, 潘瑾。女装类电商“网红”的粉丝购买意愿影响研究[J]。中国市场, 2018(4): 169-171。
- [2] Mayfield A. What is Social Media? [M]. Spannerworks, 2008.
- [3] Dave Evans. Social Media Marketing [M]. Sybex, 2010.
- [4] Kozinets Robert V, de Valck Kristine, Wojnicki Andrea C, Wilner Sarah J. S. Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities[J]. JOURNAL OF MARKETING, 2010, 74(05): 71-89.
- [5] Kerr G, Mortimer K, Dickinson S, et al. Buy, Boycott or Blog: Exploring Online Consumer Power to Share, Discuss and Distribute Controversial Advertising Message[J]. European Journal of Marketing, 2012, 46(3/4): 387-405.
- [6] Lance P, Golan G J. From Subservient Chickens to Brawny Men[J]. Journal of Interactive Advertising, 2006, 6(2): 4-33.
- [7] Trusov M, Bucklin R.E, Pauwels K. Effects of Word-Of-Mouth versus Traditional Marketing: Findings from an Internet Social Networking Site[J]. Journal of Marketing, 2009, 73(5): 90-102.
- [8] Batra R, Keller K L. Integrating Marketing Communications: New Findings, New Lessons and New Ideas[J]. Journal of Marketing, 2016; jm.15.0419.
- [9] Kumar A, Bezawada R, Rishika R, et al. From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior[J]. Journal of Marketing, 2016, 80(1): 7-25.
- [10] Gopinath S, Thomas J.S, Krishnamurthi L. Investigating the Relationship Between the Content of Online Word of Mouth, Advertising, and Brand Performance[J]. Marketing Science, 2014, 33(2): 241-258.
- [11] Naylor R, W, Lambel R, C, P, West P, M. Beyond The“Like”Button: The Impact of Mere Virtual Presence on Brand Evaluations and Purchase Intentions in Social Media Settings[J]. Journal of Marketing, 2012, 76(6): 105-120.
- [12] Schulze C, Schöler L, Skiera B. Not all Fun and Games: Viral Marketing for Utilitarian Products[J]. Journal of Marketing, 2014, 78(1): 1-19.
- [13] Goldenberg J, Oestreicher-Singer G, Reichman S. The Quest for Content: How User-Generated Links Can Facilitate Online Exploration[J]. Journal of Marketing Research, 2012, 49(4): 452-468.
- [14] 宋丹丹。社会化媒体营销与品牌资产关系研究[D]。南京师范大学, 2017。
- [15] 李东进, 刘建新, 马明龙等。微信红包, 消费者抢还是不抢——基于参与动机与心理抗拒中介模型的解释[J]。营销科学学报, 2016, 12(1): 18-37。
- [16] 杜文龙, 徐光辉, 冯现永。高校微信用户利用行为实证分析与研究——以西安航空学院为例[J]。图书馆学研究, 2015(3): 67-70。
- [17] 张芝明, 赵晓林, 范国庆。中学生对微信的“使用与满足”研究[J]。中小学信息技术教育, 2015(3): 92-94。
- [18] 何渔阳。基于使用与满足理论的大学生微信使用研究[D]。河北大学, 2014。
- [19] 黄楚筠, 彭琪琳。高校微信公众平台使用动机与传播效果研究——以中南大学微信平台为例的实证分析[J]。东南传播, 2014(8): 122-124。
- [20] 王玲宁。采纳、接触和依赖: 大学生微信使用行为及其影响因素研究[J]。新闻大学, 2014

- (6): 62-70。
- [21] 契约。2014 年微信公众号用户行为习惯研究报告 [EB/OL]。 http://www.time-weekly.com/html/20150210/28656_1.html。 截取自 2015-2-20。
- [22] 易观。中国网红经济下的女性社会化电商发展专题研究报告 2016, 2016-11。
- [23] Kang Y R , Park C . Acceptance factors of Social shopping[C]// International Conference on Advanced Communication Technology. IEEE, 2009.
- [24] Benyoucef A A R M . A model for understanding social commerce[J]. Journal of Information Systems Applied Research, 2011, 4(2):6373.
- [25] 宗乾进。国外社会化电子商务研究综述*[J]。情报杂志, 2013 (10)。
- [26] 戴世富, 韩晓丹。基于传播诱因的社会化营销传播研究[J]。华南理工大学学报(社会科学版), 2015, 17 (2)。
- [27] 韩永丽。国内社交媒体营销现状及发展趋势研究[D]。河南大学, 2014。
- [28] 赵佳英。基于扎根理论的社会化电子商务商业模式的跨案例研究[D]。南京大学, 2013。
- [29] 朱小栋, 陈洁。我国社会化电子商务研究综述[J]。现代情报, 2016, 36 (1): 172-177。
- [30] Barabási, Albert-László. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005.
- [31] Oliveira J G, Barabási A L. Human dynamics: Darwin and Einstein correspondence patterns[J]. Nature, 2005, 437(7063):1251.
- [32] 邓竹君, 张宁, 李季明。截止时间对人类动力学模型的影响[J]。上海系统科出版社。2008, 29-34。
- [33] Analysis S, Section. Analyzing, Modeling, and Simulation for Human Dynamics in Social Network[J]. Abstract & Applied Analysis, 2012, 2012(6684):552-582.
- [34] Freeman L C . A Set of Measures of Centrality Based on Betweenness[J]. Sociometry, 1977, 40(1):35-41.
- [35] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. 1999.
- [36] Caldarelli G , Capocci A , Rios P D L , et al. Scale-Free Networks from Varying Vertex Intrinsic Fitness[J]. Physical Review Letters, 2003, 89(25):258702.
- [37] Kudlicki A . Bayesian modeling of protein interaction networks[J]. 2004.
- [38] Muchnik L, Pei S, Parra L C, et al. Origins of power-law degree distribution in the heterogeneity of human activity in social networks[J]. Scientific Reports, 2013, 3(19):1783.
- [39] Malmgren R D, Hofman J M, Amaral L A N, et al. Characterizing individual communication patterns[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2009.
- [40] Vazquez A , Rácz, Balázs, Lukács, András, et al. Impact of Non-Poissonian Activity Patterns on Spreading Processes[J]. Physical Review Letters, 2007, 98(15):158702.
- [41] Ni S, Weng W. Impact of travel patterns on epidemic dynamics in heterogeneous spatial metapopulation networks[J]. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), 2009, 79(1):16111-0.
- [42] 韩筱璞, 周涛, 汪秉宏。基于自适应调节的人类动力学模型[J]。复杂系统与复杂性科学, 2007, 4 (4): 1-5。
- [43] Yan Q , Yi L , Wu L . Human dynamic model co-driven by interest and social identity in the MicroBlog community[J]. Physica A, 2012, 391(4):1540-1545.
- [44] 樊鹏翼, 王晖, 姜志宏等。微博网络测量研究[J]。计算机研究与发展, 2012, 49 (4)。

- [45] H.Kwak,C.Lee,H.Park, and S.Moon. What is Twitter, a social network or a news media?[C]. Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA. 2010, 591-600.
- [46] Cha M , Haddadi H , Benevenuto F , et al. Measuring user influence in Twitter : the million follower fallacy[J]. 2010.
- [47] Maia M , Almeida J , Almeida, Virgílio. [ACM Press the 1st workshop - Glasgow, Scotland (2008.04.01-2008.04.01)] Proceedings of the 1st workshop on Social network systems - SocialNets '08 - Identifying user behavior in online social networks[J]. BMC Oral Health, 2008:1-6.
- [48] Krishnamurthy B , Gill P , Arlitt M . A few chirps about Twitter[C]// Proc Workshop on Online Social Networks. 2008.
- [49] Petrovic S , Osborne M , Lavrenko V . RT to Win! Predicting Message Propagation in Twitter[J]. Dentistry Today, 2011, 19(11).
- [50] Wang Y , Liu H , Lin H , et al. SEA: A system for event analysis on chinese tweets[C]// Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.
- [51] Xiao Y , Wang B , Liu Y , et al. Analyzing, Modeling, and Simulation for Human Dynamics in Social Network[J]. Abstract & Applied Analysis, 2012, 2012(6684):552-582.
- [52] Java A , Song X , Finin T , et al. Why we Twitter: Understanding microblogging usage and communities[M]. 2007.
- [53] 郭慧玲。“天使小屋” 淘宝网店营销策略研究[D]。2013。
- [54] 周志华。机器学习[M]。清华大学出版社，2016-1-1： 63-65
- [55] 周志华。机器学习[M]。清华大学出版社，2016-1-1： 182
- [56] Anaconda Distribution. Anaconda[EB/OL].<https://docs.anaconda.com/anaconda/>. 截取自 2019-3-21。
- [57] Documentation of scikit-learn 0.20.3. scikit-learn. <https://scikitlearn.org/stable/documentation>. 截取自 2019-3-21。

作者简历及攻读硕士学位期间取得的研究成果


盛烨，女，1994年5月生。2012年9月至2016年7月就读于北京交通大学电子信息工程学院通信工程专业，取得工学学士学位。2016年9月至2019年6月就读于北京交通大学通信与信息系统专业，研究方向是信息网络，取得工学硕士学位。攻读硕士学位期间，主要从事电商网红社交行为商业影响的分析与预测方面的研究工作。

二、参与科研项目

- [1] 社会化信息网络的内容分发需求感知与整形，国家自然科学基金面上项目，No.61572071
- [2] 需求感知的分布式缓存网络的研究，国家自然科学基金，No. 61271199
- [3] 基于用户观看时间测量和模型的网络视频系统优化，国家自然科学基金，No. 61301082

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名： 签字日期：2019 年 5 月 31 日

学位论文数据集

表 1.1: 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
社交商业化; 电商网红; 行为分析; 预测; 机器学习	公开			国家自然科学基金 No.61572071, 61271199 , 61301082
学位授予单位名称*		学位授予单位代 码*	学位类别*	学位级别*
北京交通大学		10004	工学	硕士
论文题名*		并列题名		论文语种*
电商网红社交行为商业影响的分析 与预测				中文
作者姓名*	盛烨		学号*	16120116
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直 门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
通信与信息系统		信息网络	3	2019
论文提交日期*	2019.04.24			
导师姓名*	郭宇春		职称*	教授
评阅人	答辩委员会主席*		答辩委员会成员	
电子版论文提交格式 文本 () 图像 () 视频 () 音频 () 多媒体 () 其他 () 推荐格式: application/msword; application/pdf				
电子版论文出版 (发布) 者		电子版论文出版 (发布) 地		权限声明
论文总页数*	71 页			
共 33 项, 其中带*为必填数据, 为 21 项。				