**在线社交网络中消息传播机制的研究**

目录

[第1章 绪论 1](#_Toc446969583)

[1.1 选题的背景 1](#_Toc446969584)

[第2章 国内外研究现状 3](#_Toc446969585)

[2.1 国内外研究现状概述 3](#_Toc446969586)

[2.1.1 对传统信息传播模型进行改进 3](#_Toc446969587)

[2.1.1.1 IC模型 3](#_Toc446969588)

[2.1.1.2 LT模型 4](#_Toc446969589)

[2.1.1.3 SIR模型 4](#_Toc446969590)

[2.1.1.4 基于经典模型的研究现状 4](#_Toc446969591)

[2.1.2 基于关键因素构建新模型 5](#_Toc446969592)

[2.2 国内外研究现状评述 5](#_Toc446969593)

[第3章微博传播规律研究 5](#_Toc446969594)

[3.1 本文研究概述 5](#_Toc446969595)

[3.1.1 研究内容 6](#_Toc446969596)

[3.1.2 数据来源 6](#_Toc446969597)

[3.1.3 研究的理论基础 6](#_Toc446969598)

[3.1.4 研究方法 7](#_Toc446969599)

[3.2 主要研究过程 7](#_Toc446969600)

[3.2.1 消息转发拓扑特性研究 7](#_Toc446969601)

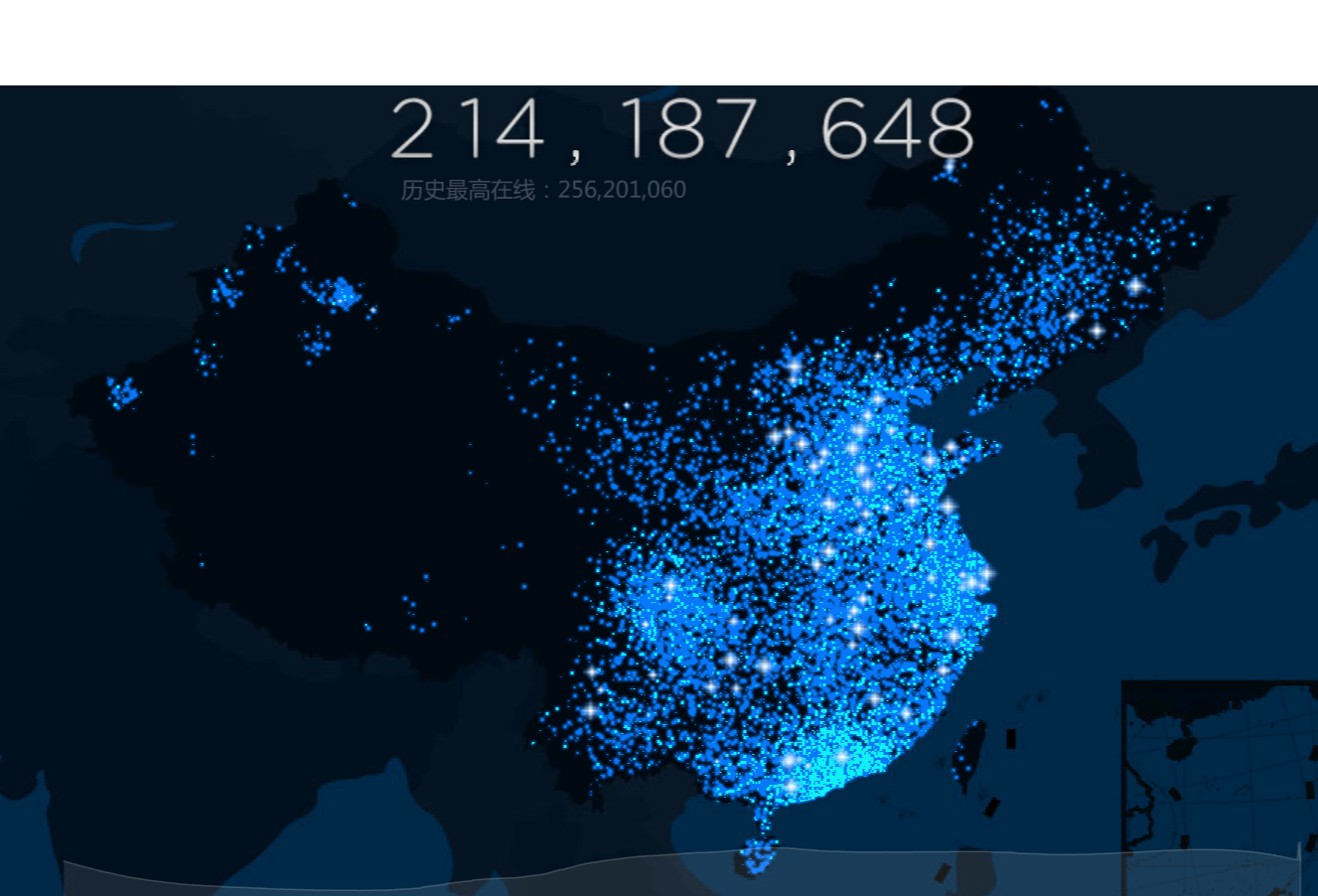
[参考文献 8](#_Toc446969602)

第1章 绪论

# 选题的背景

随着WEB2.0和各种APP的兴起和普及，越来越多的人习惯使用网络平台进行工作、交流、学习、消遣、娱乐，社交媒体已深入到人们生活的方方面面。人与人之间是普遍联系的，以人为节点，以人与人之间关系为基础的社交网络也逐步发展起来。几乎每一个人都成为在线社交网络中的一员。

据统计，截至2014年7月，Facebook的每月活跃用户数为13亿人，WhatsApp的每月活跃用户数为5亿人，Instagram的每月活跃用户数为2亿人，Messenger的每月活跃用户数为2亿人，这些都是Facebook的重要业务，所有业务的用户数加在一起，Facebook公司的每月活跃用户总数其实已经超过了22亿人。2015年8月，Facebook创历史纪录:单日用户数突破10亿。截至2016年6月腾讯QQ最高同时在线人数为2.56亿，QQ空间月活6.53亿。微信每月活跃用户已达到5.49亿，据统计，25%的微信用户每天打开微信超过30次。55.2%的微信用户每天打开微信超过10次。



这些在线社交网站和APP的服务器通常会记录用户的资料和日常数据，随着互联网时代的到来，人们越来越发现数据的重要性，为了收集更多有价值的数据，现在的大型网站和网络服务提供商已开始尽可能多的记录所有用户的所有行为。例如中国的淘宝网不但记录了用户的姓名、性别、住址、登陆时间、登陆IP、订单记录，而且连用户打开过哪些网页、点击过哪些商品、在每个页面停留的时间、鼠标经过页面哪些位置都被服务器悉数记录。越来越多的互联网海量数据成为大数据的重要来源。其中蕴含的大量有价值的信息值得我们去挖掘，并分析利用以指导实践，使大数据实现其自身的意义，创造实实在在的产业价值。

对大数据的挖掘可以从不同角度、不同层次进行。其中，在线消息传播机制的挖掘对于很多领域（例如舆情监控、链路预测、网络控制、社团划分等）具有很大的意义。近年来，对博客的研究层出不穷，在线微型博客为消息传播的研究提供了很好的数据来源，国外的学者大多以Facebook和Twitter为研究对象，国内的学者对新浪微博的研究较多。如今是一个言论自由的时代，任何人都可以在网上发表自己的经历、感受、对事物的看法，甚至是生活中很小一件不值得一提的事情，很多人都喜欢图文并茂地发表到网络。每个人都可以浏览、评论、转发自己看到的动态消息，其他人看到被转发后的消息后自己可以选择做同样的评论、转发等行为，信息由此在在线社交网络上传播开来。虽然微博是微型博客的总称，但是在中国新浪微博出现较早、发展迅猛，造成了几乎垄断的局面。因此生活中提及微博通常指新浪微博，下文中的“微博”也代指新浪微博。

微博的传播机制是其中一个值得研究的课题。微博的传播机制，就是微博信息传播的制度与其各要素职能和联系的建立与运作，是包括传播者、传播过程、传播模式以及受众等所有构成的统一体，具有其独特性。微博的传播机制下，传播的过程具有五个要素：传播者、信息、媒介、受众、反馈，以它们为研究点可以对微博的传播过程进行系统地梳理。使用传播模式的方法可以来说明它们在传播过程中起到的作用以及相互之间的联系。

新浪微博数据的可研究性体现在以下几个方面：

1. 用户量特别大。微博月活跃用户达到2.22亿，同比增长33%，日活跃用户也达到1亿。
2. 覆盖领域比较广。从2009年兴起并风靡至今，各种社会媒体、政府机关、公众人物纷纷开通微博，微博已走入更多普通大众的生活，影响力越来越大。
3. 消息转发关系明确。在微博上每一条转发过的消息都可以追根溯源，找到所有参与过的人的行为，包括转发、回复、点赞、@等。
4. 数据可获取。新浪微博提供了官方收费API接口，使用者可以轻松获取较多数据用于研究。也可以使用网络爬虫对微博数据进行爬取，这也是比较流行的数据获取方式。
   1. 问题研究的理论和现实意义

对在线消息传播机制的研究具有很重要的理论和现实意义。学术界对传播模型的研究由来已久，比较经典的的有SI、SIS和SIR等传染病模型。一般把传染病流行区域内的所有人分成几个类型：S型，易感者（Susceptible），该类人群未得病，但对疾病缺乏免疫力，一旦与感染者接触容易被感染；I型，感染者，已经得了传染病；R型，已治愈（Removal），指被隔离或因病愈而具有免疫力的人。模拟现实世界人与人之间的联系形成网络，随着时间的流逝，染病者以一定的概率向周围的邻居传播疾病，在更复杂的模型中，感染者也可以以一定的概率痊愈而改变节点性质。可将传染病模型适用于不同类型的网络，设置不同的感染概率、治愈模式，以观察网络结构、节点性质、感染机制等因素对于传播结果的影响。

学者们经过大量的实验和改进，促使了传染病动力学模型的建立。通过模拟种群或各种复杂社会因素，构建经典数学模型，定性、定量分析以发现疾病传播的普遍性规律，掌握传染病流行的原因和关键环节，从而能够在现实世界中对疾病传播做到有效预测和有力控制，发挥传播动力学的作用。

与传染病模型不尽相同，对在线社交媒体中消息传播机制的研究具有更强的现实意义，因为疾病传播模型是基于概率假设的理论模型，所使用的网络无论是随机网络，小世界网络，或是BA无标度网络都是对现实人际关系的一种近似模拟，所使用的数值都是虚拟的，而在线社交媒体记录了海量真实的消息传播数据，是人们在线活动的真实反映。类似的传播数据在现实生活中是很难完全收集到的，例如在流行疾病发生时，即使强有力的政府机构也很难完全、真实地收集到感染病持续时间内的所有有效数据。因此，在线社交媒体的消息传播理论特性和数据优势决定了其在理论研究和现实应用领域都具有重要意义。

对社交媒体消息传播机制的研究既能弥补传统传播模型的局限性，又能基于对传播规律的认知实现链路预测，对个性化推荐、广告精准投放具有重要意义，同时可根据不同节点在传播中起到的作用进行网络控制，保证网络和信息安全，这在舆情监控、谣言传播控制领域将发挥巨大作用。

第2章 国内外研究现状

# 2.1 国内外研究现状概述

国内外对在线社交媒体的研究从其兴起至今从未中断过，出现不少有价值的学术成果。最著名的“六度分离”理论是哈佛大学心理学教授通过邮件网络传递实验得到的。在2011年5月，研究人员发现在Facebook上全球两个用户之间的平均距离只有4.74，比“六度分离”缩短了很多。对信息传播模型的研究也在很多领域（例如信息科学、传播动力学、医学等）取得了显著成果，这些理论和模型也同样适用于微博消息传播。学界对微博传播规律的研究可概括为以下两种途径：改进传统模型和通过研究影响传播的关键因素构建新模型。其中国外学者多以Facebook和Twitter为研究对象，国内学者则大多关注于新浪微博数据。

## 2.1.1 对传统信息传播模型进行改进

学界对信息传播领域的研究由来已久，独立级联（IC）模型[[1]](#endnote-1)、线性阈值（LT）模型[[2]](#endnote-2)[[3]](#endnote-3)、传染病（SIR）模型[[4]](#endnote-4)是经典的三大传播模型。这些经典模型对传播领域问题具有普遍的指导意义。

### 2.1.1.1 IC模型

IC模型中的网络传播过程如下：

1. 网络中初始存在两种节点，活跃节点和非活跃节点。规定活跃节点可转变为非活跃节点。
2. 在t时刻，若节点v变成活跃状态，它将以概率的概率对其邻居节点w进行影响，其中是预设的随机值。该概率的大小决定节点对其邻居的影响力。
3. 如果节点w有多个刚被激活的邻居，则这些邻居们会以任意的次序对w产生影响。
4. 若w成功被某个邻居激活，该节点将在下一时刻成为活跃节点，并对其所有邻居产生影响，以此规则迭代，终止条件是网络中不存在可以激活其它节点的节点。

其中不存在可激活其它节点的节点是指所有被激活的节点都已丧失激活邻居节点的能力。每个节点被激活后仅有一次影响其邻居节点的机会，无论当前时刻该节点是否使它的邻居们成功激活，它在此后的时刻里成为无影响力的节点，尽管它自身也是激活节点。

### LT模型

线性阈值（LT）模型的传播机制如下：

1. 与独立级联模型相似，网络中的节点存在活跃和非活跃两种状态，节点可从非活跃状态转变为活跃状态。初始状态下，为每个节点v赋予随机阈值，该值代表节点的受影响被激活的概率，值越大，节点v越稳定而不易改变，反之表明节点越容易被影响而激活。
2. 若节点v的所有邻居中，最近处于活跃状态的节点对它的影响力之和大于该节点的阈值时，该节点将受到影响转变为激活状态。
3. 定义节点v的邻居集合为N，则对于任意w∈N，权值代表邻居节点w对节点v的影响力。因此，节点v所受其所有邻居节点(激活状态)的影响力之和可表示为：

4. 所有在t时刻被激活的节点将在t+1时刻对它的邻居释放影响力。若节点v的阈值满足如下条件：

节点v将转化为激活状态。按以上步骤迭代。

该模型的传播终止条件为网络中所有非活跃状态的节点都不能被他们的活跃邻居的能量之和所激活，此时网络状态趋于稳定。

### SIR模型

SIR模型是在SI模型的基础上改进而来的，该模型忽略考察地区的人口出生、自然死亡、迁徙等因素影响，假设人口总量保持在一个常数，即模型中网络节点数N保持不变。其传播过程如下：

1. 网络中初始存在两种类型的节点：S型-易感者（Susceptible）， I型-感染者（Infective）。随着传播的进行可能出现第三种类型的节点：R型-已治愈（Removal），指被隔离或因病愈而具有免疫力的人。即易感染者可能转变为感染者，感染者可转化为已治愈型。
2. t时刻所有I型节点对它的所有S型邻居节点以一定概率p进行感染。
3. I型节点以一定概率q转换为R型节点，从此具有免疫力，不再容易受到感染。

迭代指定次数后终止模型。

该模型对于研究不同网络类型的传播特性有很大作用。

### 基于经典模型的研究现状

在国内外，根据微博中人与人之间网络的特点，对IC模型、LT模型、SIR模型等经典传播模型进行改进，出现了大量研究成果。张彦超[[5]](#endnote-5)等基于传染病模型利用传播动力学和网络科学理论，对在线社交网络中的消息传播从理论上构建了比较详细的模型，然后提出了基于SNS网络的独特的信息传播模型，通过对在线社交网络的模拟实验，并对I型节点、S型节点、R型节点在网络中不同的传播行为进行了分析。Kazumi Saito[[6]](#endnote-6)等人基于IC模型和LT模型，在节点和连边上引入两种延迟：Node Delay（用户行为带来的延迟）和Link Delay（消息传播带来的延迟），并对激活时间的更新机制进行修改，引入了两种节点激活方式：non-override和override，即多个父节点激活时使用最早激活时间但不更新激活时间，当多个父节点多次激活时对激活时间进行更新。结合两种原始经典模型，构建了异步IC模型和异步LT模型。进而在多种不同性质的网络中调整传播参数以研究时间延迟及更新策略的不同对消息在网络中传播带来的影响，得出了在密集网络中差异化时间延迟和更新机制对信息传播研究至关重要的结论。

## 基于关键因素构建新模型

传统的消息传播经典模型虽然在传播领域具有指导意义，但应用于新的复杂的网络环境和数据形式时存在局限性，新的传播模型不断被提出。这些新模型通常是基于对影响微博网络中信息传播的诸多要素进行研究和评估得到的。2012年，郑蕾和李生红针对微博网络提出了一种叫做IDM的消息传播模型。通过模拟消息在新浪微博用户网络中的传播流向，发现在微博传播过程中当信息具有一定敏感度时，粉丝多的用户（明星用户）的意见领袖作用呈现一种逐渐下降的趋势[[7]](#endnote-7)。Bongwon Suh[[8]](#endnote-8)等人对影响消息转发频率的因素（消息内容特征、用户节点属性等）进行了深入研究，在Generalized Linear Model（广义线性模型）之上建立了一个用于预测消息转发的新模型。通过对大量真实数据的分析，得出了如下结论：微博消息中Tag和Url链接对消息转发影响较大，对于用户在网络中的属性，用户历史所发表的微博总数量对微博转发影响较小，而粉丝数、关注数、使用时间长短对传播的影响较为突出。

Herbrich R等人提出了一个命名为Matchbox的新模型[[9]](#endnote-9)，该模型是一个基于概率理论的协同过滤模型，其中选用了用户网络中的节点特征（关注人数、粉丝数）和微博消息特征（单词数等）来预测用户对消息的转发。尽管这些特征很能反映用户的行为习惯，但在导致传播行为的影响因子中并不是最重要的。Fan等人[[10]](#endnote-10)对新浪微博的用户网络拓扑及信息传播扩散进行了深入研究，他们发现该网络的拓扑结构和小世界网络及无标度网络非常相似，网络的度分布服从幂律分布，热门新闻事件的传播路径呈现两级或星形结构。Webberley等人[[11]](#endnote-11)深入研究了信息在Twitter上产生影响的广度、传播的深度以及消息转发的延迟规律，这些分析对用户行为模式和消息转发规律具有很好的借鉴意义。Jure Leskovec等人基于博客数据对社交网络结构特性及其信息传播模型进行了深入研究，发现了博客文章热度的下降并不是一般认为的指数分布，而是非常符合幂律分布。同时，以每一个博客为节点以转载为连边构建的有向网络中，节点的入度和出度都符合幂律分布。经过使用经典的传染病模型对信息传播进行模拟，实验结果与真实博客网络具有很高的匹配度[[12]](#endnote-12)。

# 2.2 国内外研究现状评述

综上所述，国内外近年来对于在线社交网络中信息传播的研究较多，因为进入互联网时代后社交网络和人们的生活息息相关，甚至改变了人们的思维、行为方式，对社交网络信息传播的研究有助于人们更好的了解人与人之间的普遍联系，挖掘普遍存在的传播规律。目前，国内学者对在线社交网络中信息传播的研究，较多停留在宏观层面的分析，其中一部分研究从传播学角度入手，对在线社交网络中信息传播模式作定性分析；另一部分研究往往利用已有传播模型对社交网络的信息传播进行模拟。在预测用户对消息的转发时，往往采用了多种影响微博互动的特征，然而在对用户关系属性进行分析时，有些使用的算法对真实社交网络的适用性经不起推敲或效率较低。很多针对在线社交网络中信息传播影响力和传播路径的分析及预测不具有普遍性，得到的大多是特定网络（Twitter或Facebook等）上对信息传播基于统计的规律，例如有的研究[[13]](#endnote-13)表明邮件网络的传播树往往具有传播广度教小、深度较大的特性，然而并没有针对具体信息预测其在传播过程中可能经过的网络路径。

第3章 微博传播规律研究

# 本文研究概述

前人的研究表明，不同性质的网络在消息传播中呈现的特性也不尽相同，甚至差别很大。例如微博网络、博客网络与Email网络中信息传播的规律各不相同，对在线社交网络的信息传播的研究往往因为数据和领域的不同而必须区别化对待。由于一个人的精力有限，笔者仅仅基于新浪微博数据对新浪微博中信息传播机制进行研究，希望挖掘出有价值的传播规律，使人们能更清晰地认识传播、深入了解传播、实现传播最大化或控制谣言传播。然而这也不是一项小的工程，因为新浪微博的用户群体和消息内容实在超出传统数据分析处理的能力，本人将使用大数据技术完成该项研究。

## 3.1.1 研究内容

本文将从以下几个角度对新浪微博信息传播规律进行研究：

* 1. 消息转发拓扑特性研究。微博消息的转发树有很多种常见形态，例如：蒲公英式、双子星式、烟花式等。微博消息转发拓扑结构与微博用户节点的特征、消息的内容特征、粉丝群的属性等因素之间是否存在联系，存在何种关联，这对于我们从宏观角度了解传播有重要意义。
  2. 研究影响微博消息传播广度、深度的要素。消息传播的广度和深度是衡量传播影响力的重要指标，消息的传播模式不同导致消息传播的广度及深度千差万别。
  3. 对用户基于消息转发聚类。对于同一个博主的粉丝，有的人喜欢转发这样的内容，有的人则偏重于转发那样的消息。其中某些人的爱好、行为习惯或其他特质是一样的，可以对这些人进行聚类建模，了解不同人群对于消息传播所做的贡献或影响力，对于预测传播行为具有一定意义。
  4. 预测用户对于微博消息的转发。转发行为是微博消息传播研究的关键，本文将根据用户节点属性、微博消息内容等其它可能影响用户转发的要素来建立模型，预测用户对于消息的转发。

## 3.1.2数据来源

新浪微博拥有丰富的海量数据，这对于数据挖掘、传播分析来说是个大型的宝库。微博一直以来都是学界、商届都很青睐的数据来源，它具有用户量大、覆盖面广、数据量大等特点。为了使研究结果具有普遍性和可比较性，本文选用了国内在线社交网络研究领域领域里最常用的新浪微博数据。对于微博数据的获取通常有两种方式：官方API接口查询和网络爬虫爬取。这两种方式各有优缺点：

* 1. API方式。微博为应用开发提供了开放API接口，提供JAVA、PHP、PYTHON等多种编程语言支持，开发者根据相关文档调用接口可完成数据动态获取。这种数据获取方式比较方便快捷，但是为了避免数据集被恶意大量复制，新浪微博对API调用次数做了限制：微博开放接口限制单个IP、单个用户、单个应用在1小时内请求不能超过一定的次数。这对于我们短时间内获取大量数据带来了很大障碍。
  2. 爬虫获取方式。爬虫其实就是一个机器人，开发爬虫的目的是为了让它代替人来浏览网页并收集数据。爬虫对于开发人员的技术要求较高，因为它不但要模拟登陆、识别验证码，而且要有较好的爬取策略以应对反扒程序的封锁，有时可能导致所在局域网都无法正常访问源站的局面。然而道到一尺，魔高一丈，爬取和反爬的斗争从未停息过，爬虫开发人员一直坚信只要人能够访问到的内容爬虫就能获取到。利用爬虫收集数据是目前比较流行的做法。

本文所使用的数据是经过网络爬虫获取的，包含了本研究所需要的各方面数据，其中有微博消息数据、转发记录、用户信息、关注关系等，经过数据清洗、整理后形成格式化数据，在此基础上进行实验。

## 3.1.3研究的理论基础

本文是基于社交网络分析和统计分析的方法所进行的研究。

社交网络分析是基于信息学、数学、社会学、管理学、心理学等多学科的融合理论和方法，为理解人类各种社交关系的形成、行为特点分析以及信息传播的规律提供的一种可计算的分析方法。从社会网络的角度出发，人在社会环境中的相互作用可以表达为基于关系的一种模式或规则，而基于这种关系的有规律模式反映了社会结构，这种结构的量化分析是社会网络分析的出发点。

统计分析是运用统计方法及与分析对象有关的知识，从定量与定性的结合上进行的研究活动。在相关科学理论的指导下，利用统计调查并整理所掌握的大量资料及相关资料，运用统计的方法，对客观存在的社会经济现象及与之相关联的现象，进行分析研究，探求其发展变化的规律，为社会经济活动的实际决策提供依据。

统计学是通过搜集、整理、分析、描述数据等手段，以实现推断所测对象的本质，甚至和有效预测对象未来等目的的一门综合性科学。其中用到了大量的数学和其它学科的专业知识，其使用范围几乎覆盖了自然科学和社会科学的各个领域，也是很多研究领域的基础。

## 3.1.4研究方法

为了从海量数据中发掘具有理论和现实意义的规律或其它有价值信息，本文采用了多种研究方法。具体如下:

* 1. 文献研究方法。通过对国内外文献进行梳理，对复杂网络理论、信息传播理论、传播心理学理论等有了较深刻的认知。
  2. 网络分析方法。运用了复杂网络理论中的网络拓扑结构分析法，选取多个指标，对微博信息传播复杂网络的拓扑结构进行了分析。
  3. 应用机器学习方法建模和仿真模拟分析。结合微博自身特征，对网络演化和信息传播过程，分别进行建模，并利用软件对其进行仿真。
  4. 可视分析方法。通过运用可视化技术，将数据从不同维度展示出来，以发掘其中蕴含的规律，为深入研究提供突破口。

试验中还用到其它普遍的研究方法，此处不再详尽列举。

# 主要研究过程

## 3.2.1 影响消息传播的要素研究

微博是现实世界的缩影，在微博消息传播中，不同的人扮演不同的角色，有些节点是整个传播过程的核心，有些节点是传播的桥梁。因此，用户的传播能力与微博传播的结果紧密相联，信息传播能力高的用户在增强传播影响力方面有着相当重要的作用，这些人对微博话题和舆论的形成起了引导作用。为了比较深入地对微博网络中的信息传播规律进行研究，我们有必要对在传播过程中所表现的信息传播能力高的用户进行挖掘。同时这对于公正、客观地评价节点的价值，为舆情监控及广告营销等应用提供理论依据和方法支持。微博网络建立的基础是用户之间的相互关注关系。信息传播是在在线社交网络的基础之上进行的，其传播效果与用户间关系的强弱有密切的联系。在虚拟的社交网络上一个人的影响力大小不仅与所在网络的结构有关，而且依赖于网络结构随时间的动态演化。所以我们在研究在线社交网络中用户的信息传播能力时，应该综合在信息传播中用户节点的基本属性特征及其所在社交网络结构中特征等要素。下文将提取微博用户的基本属性指标以及所在微博网络结构的一些指标，以此来建立模型对用户的传播影响力进行评估，这对于我们了解传播规律具有重要意义。

进过初步筛选，我们选用了如下可能对传播影响较大的指标:

* 1. 粉丝数。一般情况下，用户的粉丝数是用户微博转发量的保障因此很多用户千方百计地拉粉丝，以增加自己的被关注量，提升自己的影响力。从传播学的角度来讲，拥有大量活跃粉丝的用户在信息传播中发挥的作用往往较大，不但他们自己原创的微博能够很快得到扩散，而且还能利用自己的粉丝优势通过转发别人的微博消息使其得到广泛关注。由此可见，粉丝数是社交网络信息传播中对传播影响较大的一个显著因素。
  2. 关注数。从理论上讲，用户的关注数与其微博的转发量之间也存在某种关联。孔子曰：来而不往非礼也。如果一个微博用户没有几个关系好的好友，那么即使粉丝再多也难以构成一个良性的健康生态环境，在社交网络中不是一个健康的节点，对消息传播的稳定性有一定影响。
  3. 微博总数。和粉丝数相似，微博用户所发的微博数量是可能影响消息传播另一个重要因素。例如某用户经常转发微博，那么他对于邻居新消息做出转发行为的概率在理论上稍大一些。但这只是一种基于理论的猜测，具体情形我们将在接下来的试验中予以分析。

微博用户的网络结构指标和该用户的基础属性指标有着同等重要的属性，主要体现在节点的中心性上。网络节点的中心性在社交网络结构分析方面有着重要的地位，并且节点的中心性指标值能够反映节点在网络中的重要性程度。节点用户的信息传播能力与节点在所属网络中所具有的重要性的联系甚至比节点自身的属性（关注数、粉丝数、微博数等）更紧密。Rumi Ghosh和Kristina Lerman通过对在线社交网络Digg进行研究，建立了对网络中用户节点影响力大小进行排序的模型，发现基于节点中心性的度量是对预测用户影响力的最优可行方案之一[[14]](#endnote-14)。下文将选用下表所示的网络指标来度量用户的信息传播能力。

表3-1 度量用户网络中信息传播能力的网络指标

|  |  |
| --- | --- |
| 指标 | 描述 |
| Degree Centrality | Degree Centrality是网络中节点关联关系的基本度量。反映了网络节点在信息传播的过程中受欢迎的程度和活跃度。 |
| Betweenness Centrality | Betweenness Centrality测量了节点在网络中位置的重要性。反映了网络节点在信息传播的过程中对其他节点间发生交互信息行为的控制能力。 |
| Closeness Centrality | Closeness Centrality测量了节点在网络中的独立性。反映了网络节点在信息传播的过程中能够不受其他节点控制的程度。 |
| Eigenvector Centrality | Eigenvector Centrality测量了节点在网络中的贡献度。反映了网络节点在信息传播的过程中对目标节点所具贡献的大小。 |
| PageRank | PageRank Value可以用来评价网络中节点的重要性程度。该指标对网络节点的邻居节点的数量和质量进行了综合考虑。 |
| K-Core Value | K-core分解值对网络按节点重要性程度进行了层次的划分。离网络中心越近的节点的K-core分解值相对越大，对应节点的重要性较高。 |
| Local Clustering Coefficient | Local Clustering Coefficient测量了节点的聚集程度。节点的该指标越大，说明网络节点在信息传播的过程中和周围邻居节点的互动越频繁。 |

### 3.2.1.1 数据预处理

根据数据集提取出来的各项指标的值差别太大，不利于指标之间的比较和加权，为了去除数据的单位限制，将其转化为无量纲的纯数值，我们需要用到数据标准化技术（normalization）。该方法是把数据按照一定比例映射到另一个特定的区间内，实现了数据缩放功能，是很多基于数据的分析工作的前提。

数据标准化的方法有很多，常用的有Min-max 标准化、z-score 标准化、Decimal scaling小数定标标准化。在此我们选用z-score标准化方法，因为z-score对于指标的最大值和最小值提前未知的情况能够较好处理。处理过程如公式所示：

其中代表处理后得到的值，为指标的样本值，代表样本均值，s是样本方差。

### 3.2.1.2 计算指标相关性

相关系数是变量之间线性相关程度的衡量指标，它所表示的是一种非确定性的关系，并且假定两个变量都是随机变量，而且之间是线性关系。除此之外，所研究的样本数据中不应该出现非常大或非常小的极端值，否则将会对相关系数的值产生很大影响，这也是在之前我们进行数据标准化的原因之一。我们采用公式求解相关系数：

其中为相关系数，为X，Y的协方差，D(X)，D(Y)分别代表X，Y的方差。

在此给出协方差的计算公式：

通过两两对比我们所收集指标之间的相关系数以了解它们之间的相关性。为了方便比较，在此我们建立了相关系数矩阵。所得结果如下表所示：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fans | Follows | Degree | Betweenness | Closeness | Eigenvalue | PageRank | k-shell | LocalCluster |
| Fans | 1.000 |  |  |  |  |  |  |  |  |
| Follows |  | 1.000 |  |  |  |  |  |  |  |
| Degree |  |  | 1.000 |  |  |  |  |  |  |
| Betweenness |  |  |  | 1.000 |  |  |  |  |  |
| Closeness |  |  |  |  | 1.000 |  |  |  |  |
| Eigenvalue |  |  |  |  |  | 1.000 |  |  |  |
| PageRank |  |  |  |  |  |  | 1.000 |  |  |
| k-shell |  |  |  |  |  |  |  | 1.000 |  |
| LocalCluster |  |  |  |  |  |  |  |  | 1.000 |

表3-2 各项指标间相关系数矩阵

从图3-2可以看出节点度中心性、Eigenvector Centrality、k-shell值的相关系数比较大，说明节点的度越大，它在网络中对其它节点的贡献也相应较大，也就越靠近网络的中心位置。Betweenness Centrality和K-Shell值的相关性也比较大，说明了如果某节点在网络中网络中控制其它节点进行信息交互的能力越强，它在网络中就越靠近中心位置。同时还有一个有趣的现象，我们发现尽管Degree Centrality、Betweenness Centrality均和K-Shell值的相关性较大，然而Degree Centrality和Betweenness Centrality的相关性却并不是很大，这说明了度大的节点控制其它节点交互信息的能力并不一定大，这两个指标之间没有必然联系。Closeness Centrality、PagaRank Value、Local Clustering Coefficient与其它指标之间都不具有很大的相关性。

### 3.2.1.3 PCI主成分分析

度量用户节点属性及网络节点属性的指标有很多，从直观上很难判断哪些指标比较重要，哪些是对研究无关紧要的，而且有些指标间有着千丝万缕的联系，这样给我们的实验增加了复杂度，也会影响算法的效率。为了有效地从众多指标中提取主要成分，我们采用了主成分分析（Principal Component Analysis，PCA）方法。

主成分分析（Principal Component Analysis，PCA）， 是一种统计方法。将一组可能存在相关性的变量，通过正交变换之后，转换成一组新的线性不相关的变量，得到的转换之后的这组变量叫主成分。当使用统计分析方法对一些具有复杂影响因素的课题进行研究时，如果使用的变量个数太多，课题研究的复杂度也相应较高，但变量太少又不能很好反映现实事物之间的联系。如何在不影响分析效果的前提下精简变量是研究人员比较关注的问题。此外，在大多情况下，虽然我们采用了很多个变量，但有些变量之间却存在某种关联，造成他们在对课题产生影响的成分里有部分重叠，但如果盲目去掉其中某些部分，又会造成信息的缺失。主成分分析为这个问题提供了很好的解决方案。PCA所做的是是对所有提出的变量进行过滤，删除其中重复出现的部分，精简后形成新的变量，其中这些变量之间是独立关系，但变量总体基本上保留了原变量对于课题所能反映的信息。总的来讲就是设法将原有变量重新组合为新的无关的几个综合变量，然后选取较少的几个变量来尽可能多地反映原有信息的一种统计分析方法。

PCA的一般处理过程如下：

1、特征中心化。每一维的数据减去这个维度的均值。在此每一“维”指的是一个属性（或特征，通常用向量来表示）。经过变换后，每一维度上的均值都变成了0。原始所有维度都减去其列均值后得到新的矩阵，在此记作B。

2、计算新的到的矩阵B的协方差矩阵C。

3、计算协方差矩阵C的特征值和特征向量。

4、选用较大的特征值对应的特征向量，最后得到新的数据集。

参考文献

1. Goldenberg J, Libai B. Using Complex Systems Analysis to Advance Marketing Theory Development: Modeling Heterogeneity Effects on New Product Growth through Stochastic Cellular Automata[J]. Acad.marketingSci.rev, 2001. [↑](#endnote-ref-1)
2. Granovetter M. Threshold models of collective behavior[J]. American Journal of Sociology, 1978, 83(6):1420-1443. [↑](#endnote-ref-2)
3. Schelling T. Micromotives and Macrobehaviour[J]. Norton, 1978. [↑](#endnote-ref-3)
4. Kermack W O, Mckendrick A G. Contributions to the mathematical theory of epidemics—III. Further studies of the problem of endemicity[J]. Royal Society of London Proceedings, 1927, 115(772):700-721. [↑](#endnote-ref-4)
5. 张彦超, 刘云, 张海峰,等. 基于在线社交网络的信息传播模型[J]. 物理学报, 2011, 60(5):60-66. [↑](#endnote-ref-5)
6. Saito K, Kimura M, Ohara K, et al. Generative Models of Information Diffusion with Asynchronous Timedelay[C]//ACML. 2010: 193-208. [↑](#endnote-ref-6)
7. 郑蕾,李生红. 基于微博网络的信息传播模型[J]. 通信技术, 2012, 45(2): 39-41. [↑](#endnote-ref-7)
8. Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network[C]//Social computing (socialcom), 2010 ieee second international conference on. IEEE, 2010: 177-184. [↑](#endnote-ref-8)
9. Zaman T R, Herbrich R, Van Gael J, et al. Predicting information spreading in twitter[C]//Workshop on computational social science and the wisdom of crowds, nips. Citeseer, 2010, 104(45): 17599-601. [↑](#endnote-ref-9)
10. Fan P, Li P, Jiang Z, et al. Measurement and analysis of topology and information propagation on Sina-Microblog[C]//Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on. IEEE, 2011: 396-401. [↑](#endnote-ref-10)
11. Webberley W, Allen S, Whitaker R. Retweeting: A study of message-forwarding in twitter[C]//Mobile and Online Social Networks (MOSN), 2011 Workshop on. IEEE, 2011: 13-18. [↑](#endnote-ref-11)
12. Mcglohon M, Leskovec J, Faloutsos C, et al. Information Propagation and Network Evolution on the Web[J]. Port & Waterway Engineering, 2010, 13(22):11375-11389. [↑](#endnote-ref-12)
13. Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using Internet chain-letter data[J]. Proceedings of the National Academy of Sciences, 2008, 105(12): 4633-4638. [↑](#endnote-ref-13)
14. Ghosh R, Lerman K. Predicting Influential Users in Online Social Networks[C]// In: Sna-kdd: Kdd Workshop on Social Network Analysis. 2010. [↑](#endnote-ref-14)