|  |  |  |  |
| --- | --- | --- | --- |
| **图书分类号：** | **（黑体五号,见图书馆首页）** | **学校代码：** | **10615（黑体五号）** |

### 在线社交网络中消息传播机制的研究

|  |  |
| --- | --- |
| **论文题目** | **在线社交网络中消息 传播机制的研究** |
| **研究生姓名** | **赵培** |
| **导师姓名** | **李平** |
| **学科专业** | **计算机科学与技术** |
| **研究方向** | **数据挖掘** |

**2016年4月**

扉页示例（英文）注：此页博士生用

### Southwest Petroleum University

### Doctoral Dissertation

在线社交网络中消息传播机制的研究

（以下各项居中列）

Candidate:

Speciality:

Supervisor:

April,2009

西南石油大学研究生学位论文知识产权声明书及

学位论文版权使用授权书

本人完全了解学校有关保护知识产权的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于西南石油大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。学校可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。同时，本人保证，毕业后结合学位论文研究课题再撰写的文章一律注明作者单位为西南石油大学。

本学位论文属于

1、保密（ ），在 年解密后适用本授权书。

2、不保密（ ）

（请在以上相应括号内打“√”）

学位论文作者签名： 指导教师签名：

年 月 日 年 月 日

西南石油大学研究生学位论文独创性声明

本人声明：所呈交的研究生学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，本论文不包含其他人已经发表或撰写过的研究成果，也不包含其他人为获得西南石油大学或其它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谢意。

学位论文作者签名：

年 月 日

摘 要

随着社交媒体逐渐深入人们的生活，人们的生活习惯、思维方式也随之改变。朋友之间现实之中的互动渐渐变少，开始倾向于线上联系老朋友，或结交新朋友，因此在线社交媒体平台的服务商收集了大量的社交行为数据。对在线社交网络的研究有利于发掘很多有价值规律，其中消息传播规律就是一个很值得我们研究的课题。

人作为节点，他所发出信息在网络上如何传播，经历怎样的路径，会有怎样的影响力，传播的深度和广度是怎样的，这些问题对于我们深刻理解在线社交媒体信息传播、掌握传播、并利用传播实现趋利避害的目的有很大的研究价值和现实意义。本文结合复杂网络理论对微博平台的信息传播从各角度进行了分析，其中重点研究了微博信息传播网络的拓扑结构、微博信息传播的相关特征、相同用户和相同消息的信息传播模型相似性等问题。并对影响微博传播的因素进行了研究。

本文从微博信息传播的拓扑结构入手，进而对微博传播的影响力、微博传播量的时间分布进行分析，发现了消息类型的不同、用户节点属性的不同对消息传播结果的影响。通过大量的数据分析，发现粉丝量特大的明星用户发布的消息往往呈现出波纹式传播模型，传播层次相对较浅;时效性较强的微博消息的传播量时间分布通常是一个长尾分布，而搞笑型、技巧型、教育型微博消息的传播模型呈现多样、爆发点随机等特征。分别对相同用户的微博记录进行抽样，发现一个微博用户的消息传播模式大体是相似的；然后对由不同人发布的相同内容消息进行抽样分析，发现不管用户的属性（粉丝数，转发量，所属分类等）有何不同，同一消息内容的传播模式总是相似的。

关键词：社交网络;微博;传播规律;信息传播树

目录

[第1章 绪论 3](#_Toc447616344)

[1.1 选题的背景 3](#_Toc447616345)

[第2章 国内外研究现状 4](#_Toc447616346)

[2.1 国内外研究现状概述 5](#_Toc447616347)

[2.1.1 对传统信息传播模型进行改进 5](#_Toc447616348)

[2.1.1.1 IC模型 5](#_Toc447616349)

[2.1.1.2 LT模型 5](#_Toc447616350)

[2.1.1.3 SIR模型 6](#_Toc447616351)

[2.1.1.4 基于经典模型的研究现状 6](#_Toc447616352)

[2.1.2 基于关键因素构建新模型 6](#_Toc447616353)

[2.2 国内外研究现状评述 7](#_Toc447616354)

[第3章 微博传播规律研究 7](#_Toc447616355)

[3.1 本文研究概述 7](#_Toc447616356)

[3.1.1 研究内容 7](#_Toc447616357)

[3.1.2数据来源 7](#_Toc447616358)

[3.1.3研究的理论基础 8](#_Toc447616359)

[3.1.4研究方法 8](#_Toc447616360)

[3.2 主要研究过程 9](#_Toc447616361)

[3.2.1 影响传播者影响力的要素研究 9](#_Toc447616362)

[3.2.1.1 数据预处理 10](#_Toc447616363)

[3.2.1.2 计算指标相关性 10](#_Toc447616364)

[3.2.1.3PCI主成分分析 11](#_Toc447616365)

[3.2.2 微博消息转发拓扑结构研究 14](#_Toc447616366)

[3.2.2.1传播模型 15](#_Toc447616367)

[3.2.2.1.1 波纹式模型 15](#_Toc447616368)

[3.2.2.1.2 蒲公英式模型 16](#_Toc447616369)

[3.2.2.1.3 菌落式模型 17](#_Toc447616370)

[3.2.2.1.4 烟花式模型 18](#_Toc447616371)

[3.2.2.1.5 蜂巢式模型 19](#_Toc447616372)

[3.2.2.1.6 双子星式模型 20](#_Toc447616373)

[3.2.2.1.7 随机引爆式模型 21](#_Toc447616374)

[3.2.2.2 实验分析 21](#_Toc447616375)

[3.2.2.2.1 明星微博的波纹式传播模式 22](#_Toc447616376)

[3.2.2.2.2 波纹式传播模型的可演变性 27](#_Toc447616377)

[3.2.2.2.3 时效性差的消息之传播模式多样性 29](#_Toc447616378)

[3.2.3 微博消息传播持久性研究 30](#_Toc447616379)

[3.2.3.1 快速引爆传播模式 30](#_Toc447616380)

[3.2.3.2 不规律引爆点传播模式 32](#_Toc447616381)

[3.2.4 影响微博传播结果的因素研究 32](#_Toc447616382)

[3.2.4.1 同一用户的微博之传播相似性研究 33](#_Toc447616383)

[3.2.4.2 相同内容的微博之传播相似性研究 35](#_Toc447616384)

[参考文献 40](#_Toc447616385)

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。

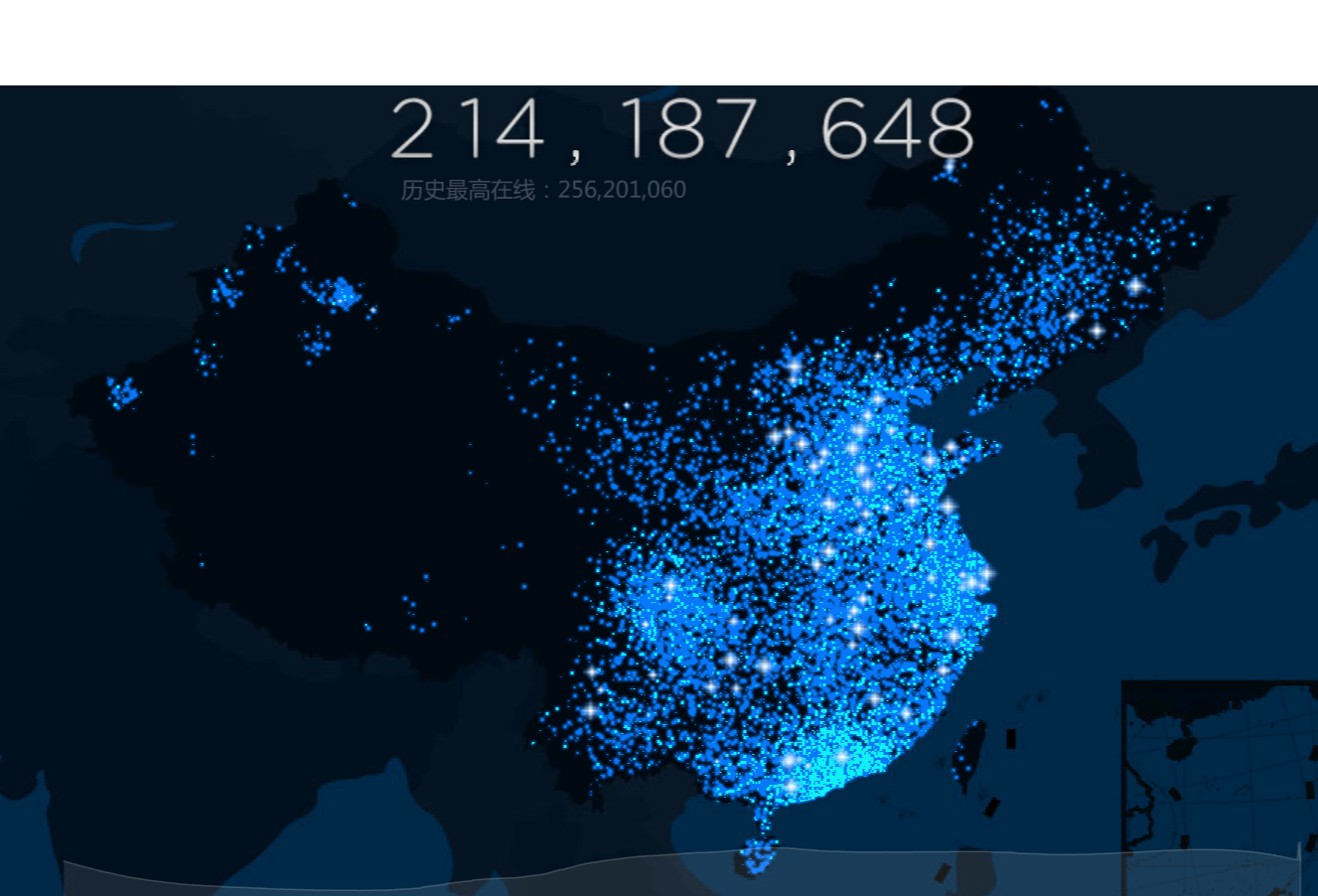


图1-1 在线社交媒体软件同时在线人数可视化

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[[1]](#endnote-2)。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”[[2]](#endnote-3)[[3]](#endnote-4)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[[4]](#endnote-5)、线性阈值（LT）模型[[5]](#endnote-6)[[6]](#endnote-7)、传染病（SIR）模型[[7]](#endnote-8)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[[8]](#endnote-9)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[[9]](#endnote-10)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[[10]](#endnote-11)。Bongwon Suh[[11]](#endnote-12)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[[12]](#endnote-13)，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[[13]](#endnote-14)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[[14]](#endnote-15)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[[15]](#endnote-16)。

# 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[[16]](#endnote-17)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

# 3.1 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。

# 3.2 主要研究过程

## 3.2.1 影响传播者影响力的要素研究

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[[17]](#endnote-18)。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 | LocalClustering 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 | 0.0198 | 0.3896 | 0.1232 | 0.0042 | 0.2536 | -0.0540 | 0.1273 | -0.0364 |
| Follows | 0.0198 | 1.000 | 0.1203 | 0.0957 | 0.0021 | -0.0049 | -0.0319 | 0.0162 | -0.0976 |
| Degree | 0.3896 | 0.1203 | 1.000 | 0.5578 | 0.0259 | 0.8839 | 0.0327 | 0.0182 | -0.0023 |
| Betweenness | 0.1232 | 0.0957 | 0.5578 | 1.000 | 0.0596 | 0.4126 | 0.0852 | 0.7385 | -0.0078 |
| Closeness | 0.0042 | 0.0021 | 0.0259 | 0.0596 | 1.000 | 0.0459 | -0.0010 | 0.0028 | 0.0522 |
| Eigenvalue | 0.2536 | -0.0049 | 0.8839 | 0.4126 | 0.0459 | 1.000 | -0.0155 | 0.0123 | 0.0358 |
| PageRank | -0.0540 | -0.0319 | 0.0327 | 0.0852 | -0.0010 | -0.0155 | 1.000 | 0.0001 | 0.1314 |
| k-shell | 0.1273 | 0.0162 | 0.0182 | 0.7385 | 0.0028 | 0.0123 | 0.0001 | 1.000 | -0.0053 |
| LocalCluster | -0.0364 | -0.0976 | -0.0023 | -0.0078 | 0.0522 | 0.0358 | 0.1314 | -0.0053 | 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。ClosenessCentrality、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 PCA 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[[18]](#endnote-19)[[19]](#endnote-20)。

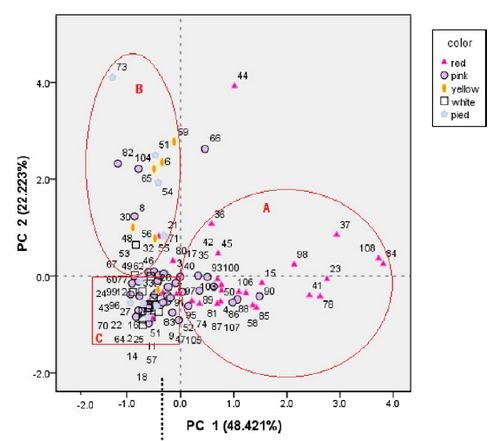


图3-x PCA主成分分析图示

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。

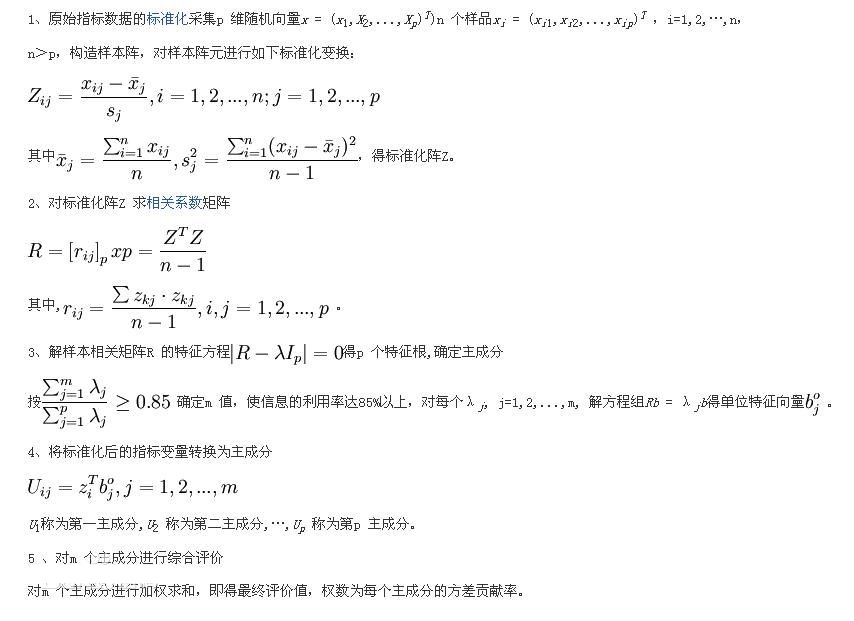


图3-xPCA主成分分析详细步骤

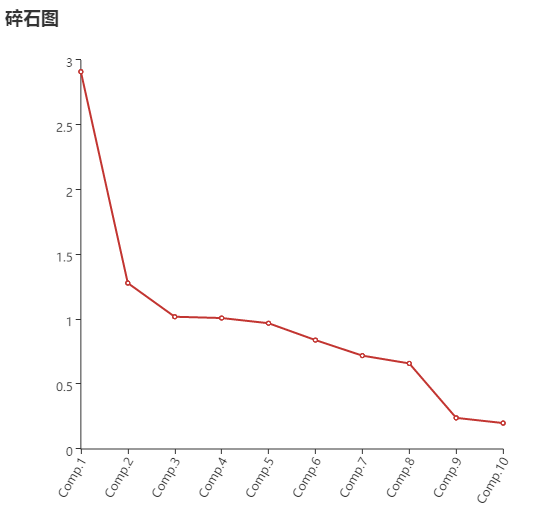


图3-x主成分分析结果的碎石图

如图所示，我们采用碎石图的形式将各指标的特征值展示出来。我们可以看出，第一二三四主成分的特征值全部大于1，而第五至第十指标的特征值都小于1，我们知道，当特征值小于1时，说明引用此主成分并不及直接引用原始变量所具有的解释力强，因此本文选用了前四个主成分。而且，从下表我们可以看到，前四个主成分累积的贡献率为65.15%，所以前四个主成分的选取是正确的选择。

表3-x主成分提取结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 初始特征值 | | | | 主成分 | | |
| 成分 | 合计 | 方差比值% | 累积比值% | 合计 | 方差比值% | 累积比值% |
| 1 | 2.91 | 29.54 | 29.54 | 2.91 | 29.54 | 29.54 |
| 2 | 1.28 | 12.99 | 42.54 | 1.28 | 12.99 | 42.54 |
| 3 | 1.02 | 10.36 | 52.89 | 1.02 | 10.36 | 52.89 |
| 4 | 1.01 | 10.25 | 63.15 | 1.01 | 10.25 | 63.15 |
| 5 | 0.97 | 9.85 | 73.00 |  |  |  |
| 6 | 0.84 | 8.53 | 81.52 |  |  |  |
| 7 | 0.72 | 7.31 | 88.83 |  |  |  |
| 8 | 0.66 | 6.7 | 95.53 |  |  |  |
| 9 | 0.24 | 2.44 | 97.97 |  |  |  |
| 10 | 0.20 | 2.03 | 100 |  |  |  |

## 3.2.2 微博消息转发拓扑结构研究

微博消息转发网络的拓扑结构是指一条消息由原创作者发布后，经由其他用户转发和多次转发后最终形成的转发网络所具有的拓扑结构，从信息传播的角度来讲可以称之为信息传播树。前人的研究表明，大多数传播树具有相似的传播模型。易成岐等人[[20]](#endnote-21)把微博信息传播树归结为八种模型，分别是：蒲公英式模型、烟花式模型、双子星式模型、菌落式模型、蜂巢式模型、波纹式模型、随机引爆模型和其它模型。这些模型的在深度、广度、节点度分布、网络结构上各有自己的特点。在对热门微博的大量数据统计下得到各种模型所占的比重，如下图所示：

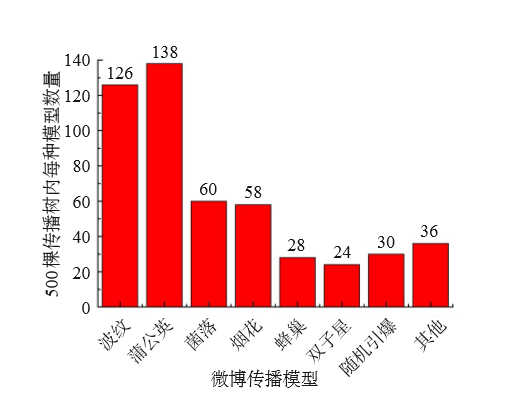


图3-1各种传播模型数量统计

### 3.2.2.1传播模型

### 3.2.2.1.1 波纹式模型

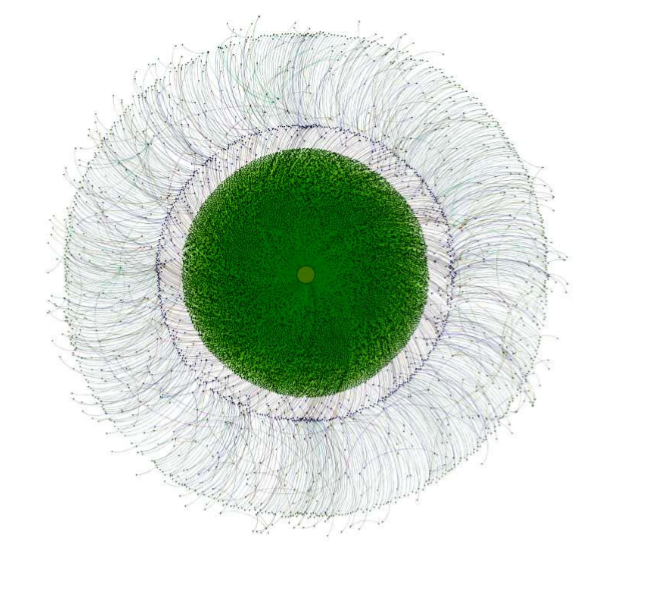


图3-1波纹式传播模型

如图3-1所示，该传播模型主要以微博信息原作者为中心，就像平静的湖面投入一粒石子激起的波纹，迅速往四周蔓延开来。这种情形大多是博主发出一条消息后，他的粉丝会同时收到该条消息，其中一部分粉丝会对这条消息进行转发，虽然转发量比较大，但二次转发及多次转发的情形较少。这种传播在所有转播模型中占有相当高的比例。

### 3.2.2.1.2 蒲公英式模型

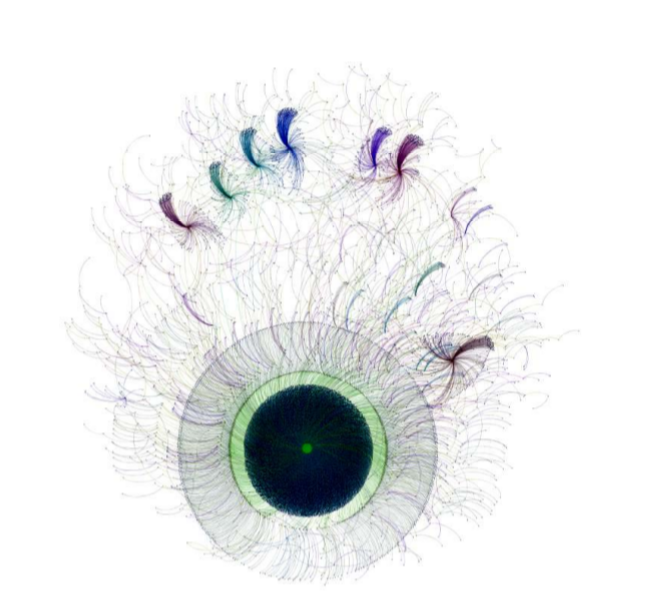


图3-2蒲公英式传播模型

如图3-2所示，这种模型的拓扑结构看起来就像是蒲公英。该模型在新浪微博官方发出的统计结果中被提及，多见于集团的官方账号。信息传播从一个官方账号发布进行传播后，由集团内多个其他官方账号转发。以其他官方账号为中心再次进行扩散的模式。被称为集团账号信息传播的蒲公英模式。官方表示从营销的角度来讲，适合此模式的企业类型：这类传播模式多见于拥有众多子品牌或业务线的企业，子品牌或业务线的目标用户既有特性又有共性。

### 3.2.2.1.3 菌落式模型

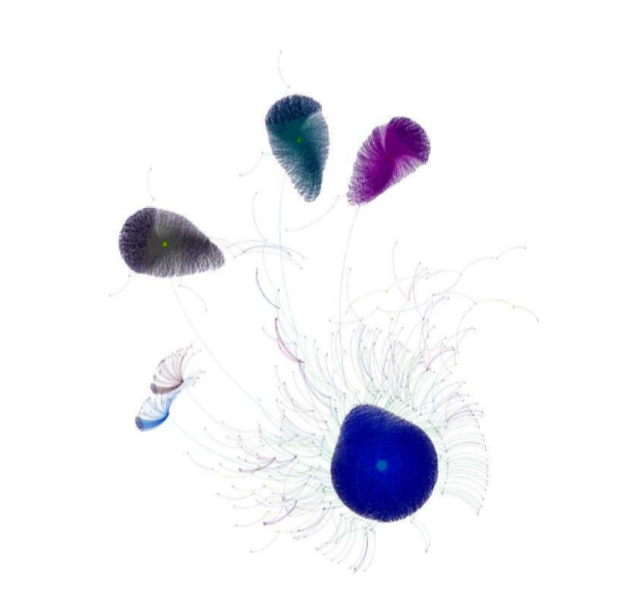


图3-3菌落式传播模型

如图3-3所示，菌落式传播模型的特点比较明显，以原创微博博主为中心传播，并在网络中形成若干个规模较大的波纹式传播模型。影响力较高节点的中心性非常强，除了几个比较大的波纹式模型外，其它转发树的规模都相当小。所以，这种模型的传播结果是由几个影响力较大的节点主要导致的，层级比较明显，从上级往下级发布消息时传播接近于这种模型。从营销的角度讲，大型企业可以试用该模型，通过企业旗下多个子品牌设计产品营销模式。

### 3.2.2.1.4 烟花式模型

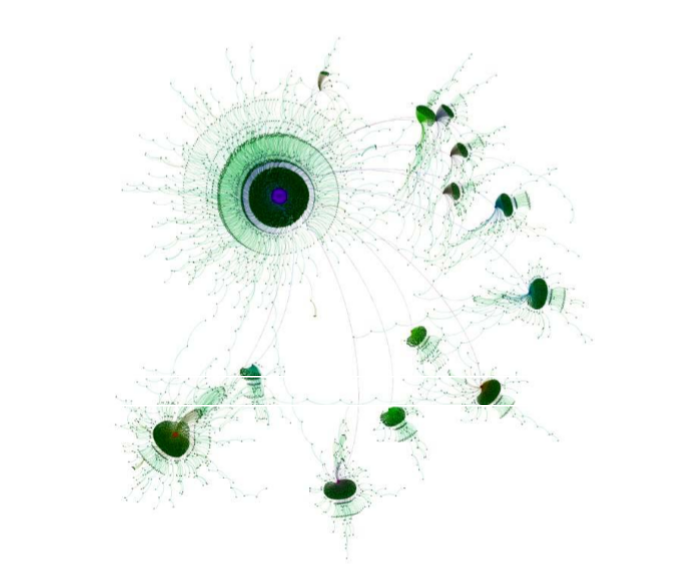


图3-4烟花式传播模型

烟花模式也是比较常见的传播类型，原创微博的博主具有十分强大的传播力，同时在二次转发和多次转发中，也有大量关键节点参与了转发，最后形成了大规模的转发网络。这种消息的影响力非常大，传播范围极广，通常是比较热门的话题，对大多数网民都有普遍吸引力。

### 3.2.2.1.5 蜂巢式模型

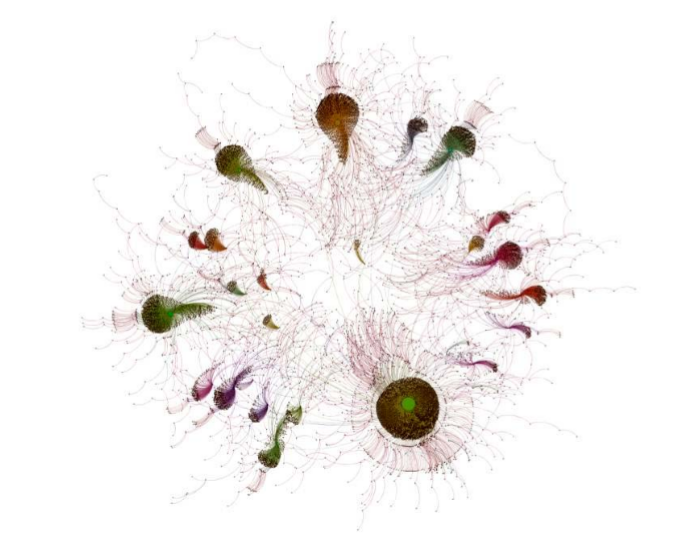


图3-5蜂巢传播模型

如图3-5所示，蜂巢式传播模型和烟花式模型在拓扑结构上存在相似性，它们最大的区别是烟花式模型的最强中心是原创微博，其它大节点虽然在传播中起了很大作用，但影响力远不及原创作者。而蜂巢式模型的传播爆发点和原创微博造成影响的规模比较相近，转发节点对扩大传播的贡献度甚至更高，所以蜂巢式模型具有持久的爆发能力和更强大的传播影响力。相当于在传播中出现的爆发点把该条微博的影响力发挥到了最大。

### 3.2.2.1.6 双子星式模型

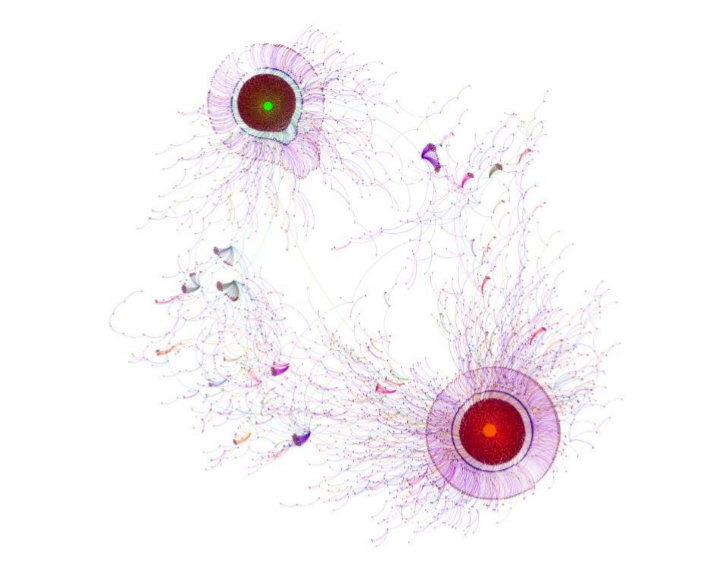


图3-6双子星式传播模型

如图3-6所示，该种传播模型中存在两个影响力非常大，而且影响力相近的节点，而且他们之间存在交互。新浪官方说明了这种传播的一个常见情形是：企业拥有官方账号，而且同时企业的领导人账号也具有较强的影响力和传播力。在互联网比较开放的今天，普通民众喜欢关注一些商界的名人，同时又会关注他们领导的企业在微博开设的账号。当这些商界大V的企业推出新产品或新政策时，企业账号会在微博发布消息，同时领导人会转发，或者刚好相反。如此以来，在最终形成的消息传播网络中，这两个账号起到了非常重要的作用，就像两颗遥相呼应的行星。

### 3.2.2.1.7 随机引爆式模型

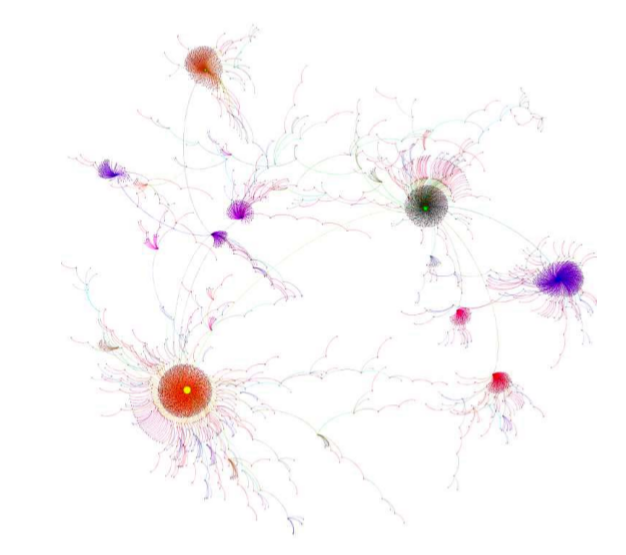


图3-7随机引爆式传播模型

随机引爆模型中有时融合了很多其它模型的特点，如图3-7所示，我们可以看到波纹式模型、蒲公英式模型、烟花式模型等。它最大的特点在于爆发点不确定，具有很强的随机性。一条消息在传播的过程中遇到了不确定的传播能力较强的节点，然后被引爆，最终造成了较大的影响力。

### 3.2.2.2 实验分析

对于消息传播树模型的划分种类有很多，上一节所述的传播模型中很多模型之间具有共性，也可以从其它角度做不同划分。经过对我们数据集里大量明星用户和热门微博的统计分析，发现了一些具有普遍性的规律。

为了了解影响微博消息转发树的相关因素，我们采集了一些转发量比较大的和一些粉丝特别多的明星用户发布的微博，并对转发链上的消息进行收集，最后集中采集了所有相关用户的个人资料，包含微博账号、年龄、性别、粉丝数、关注数、地域、微博数等信息。数据构成如下图所示：

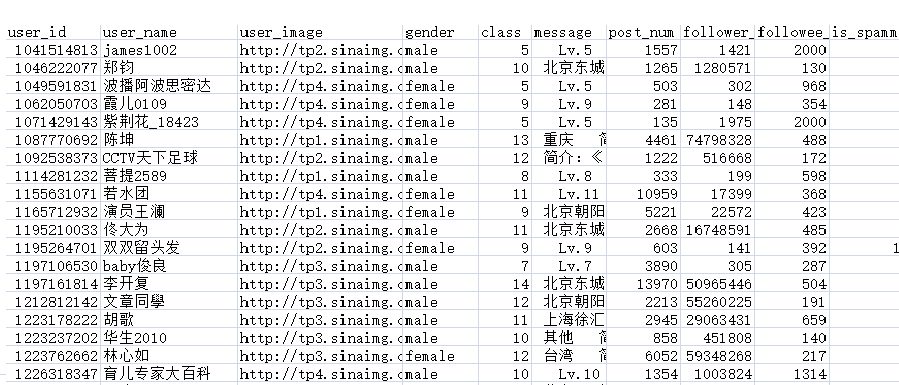


图3-8微博用户信息数据

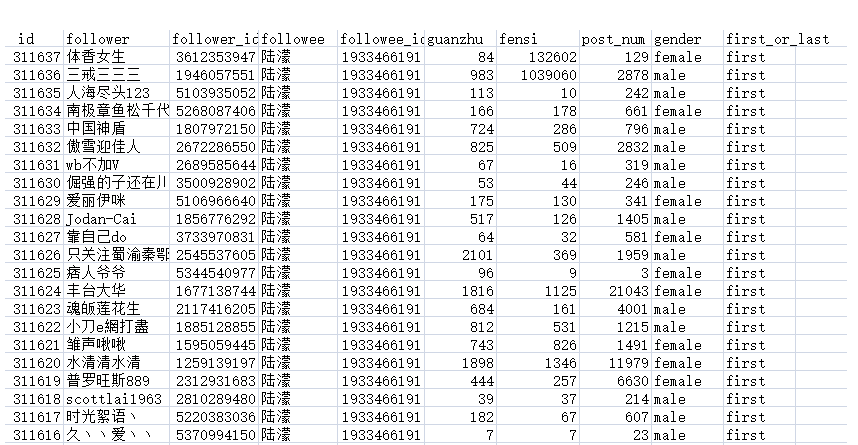


图3-9微博用户关注关系数据

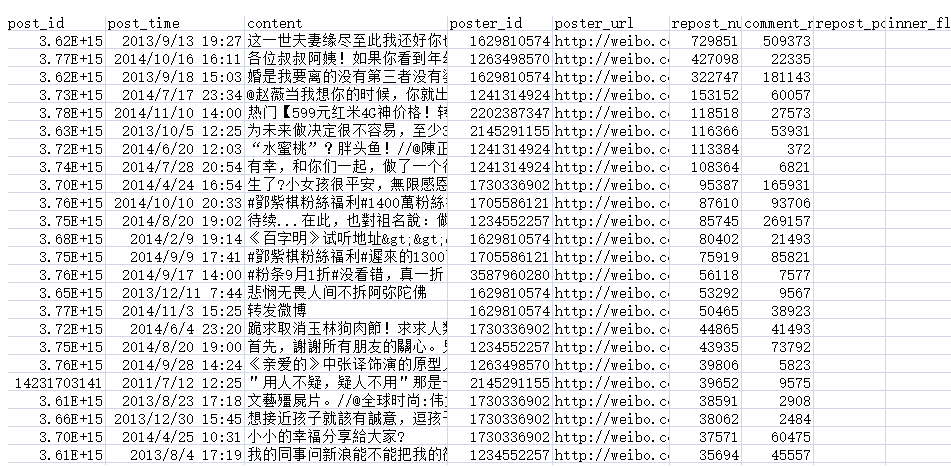


图3-10微博消息内容及详情数据

对微博源数据需要做一些数据预处理和整理，例如有一些无用的账号和无用的消息，，如果不清洗掉将会对实验的准确性造成影响。例如，在我们所采集的数据集中存在“错误的请求”之类的属性值，这是爬虫采集时遇到错误没有正确处理造成的。也有一些粉丝量特别大的用户，经过人为查看后发现是“微博小秘书”、“新手指南”之类的无意义帐号。

在微博数据之上建立消息传播网络之前，我们将消息的转发者和被转发者提取出来，以二元组的形式导出为中间数据，这在后续的很多实验中都会用到。

在本节试验中我们选取了多个转发量较大的微博消息，然后根据微博转发关系建立信息传播树，并对微博消息、微博博主、转发者构成、引爆点等属性进行统计分析。我们发现了一些规律.

### 3.2.2.2.1 明星微博的波纹式传播模式

以下统计来自转发量大于100的200条明星微博，其中明星微博内容来自随机抽样。我们对每条微博的转发量及转发层次进行了统计，在此忽略了评论数，因为经验证每一层次的转发量和评论量成正相关，而且评论仅仅在被评论点形成实质性传播，除一级传播的评论（博主的直接粉丝对于微博内容的评论）以外其它传播层次的评论量通常极少，对我们研究整体传播来讲意义不大。统计结果如图所示：

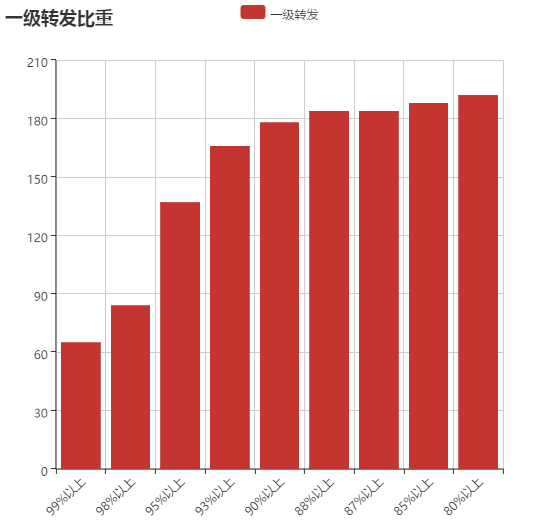


图3-x一级转发比重各比值所占数量

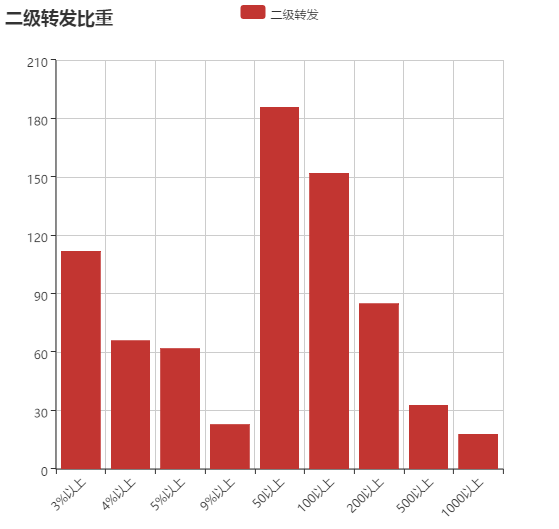


图3-x二级以上转发各层次所占数量

我们发现明星用户（自然人，不含机构或组织的官方账号）的绝大多数日常微博都呈波纹式传播，博主本人在传播中起了主要作用，也就是说大规模的传播结果是靠博主本人的个人魅力造成的。这一信息传播树模型也和普通用户的日常微博传播结果相似，差别是明星用户的粉丝量基数比较大而已。

下列图中信息传播树的拓扑结构用环形显示，节点大小代表微博消息在当前结点被转发的次数。



图3-x 明星微博原文示例-方中信微博原文1

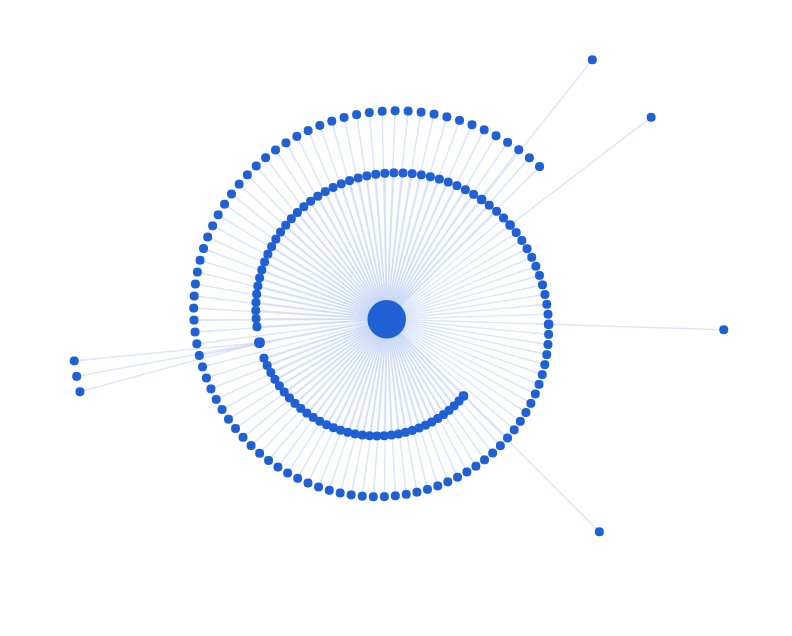


图3-x明星日常微博-方中信微博统计1



图3-x 明星微博原文示例-方中信微博原文2

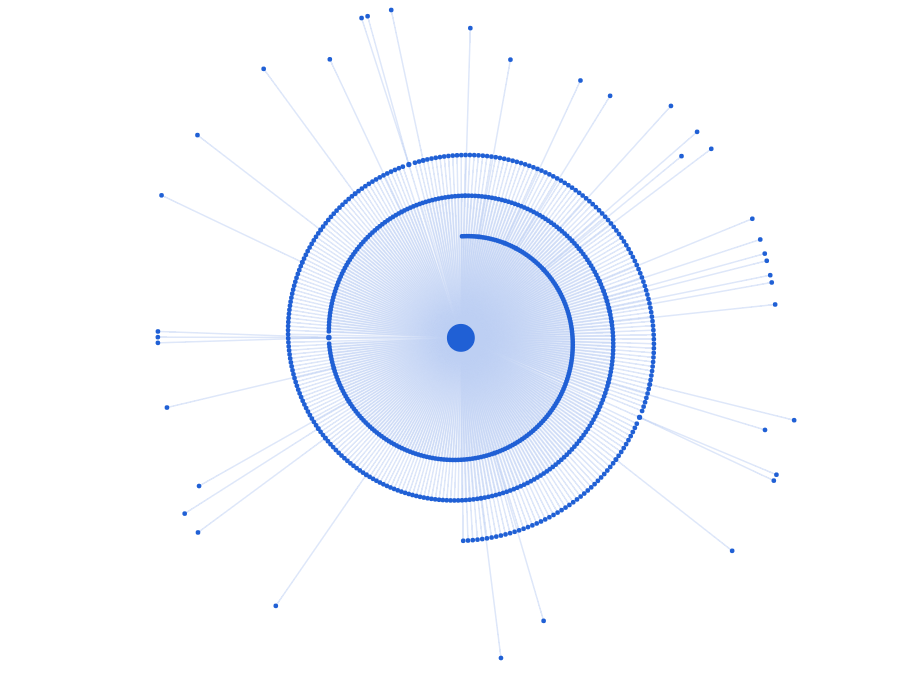


图3-x明星日常微博-方中信微博统计2

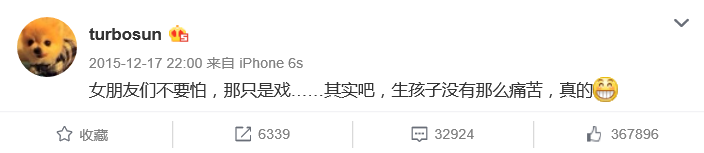


图3-x 明星微博原文示例-孙俪微博原文1

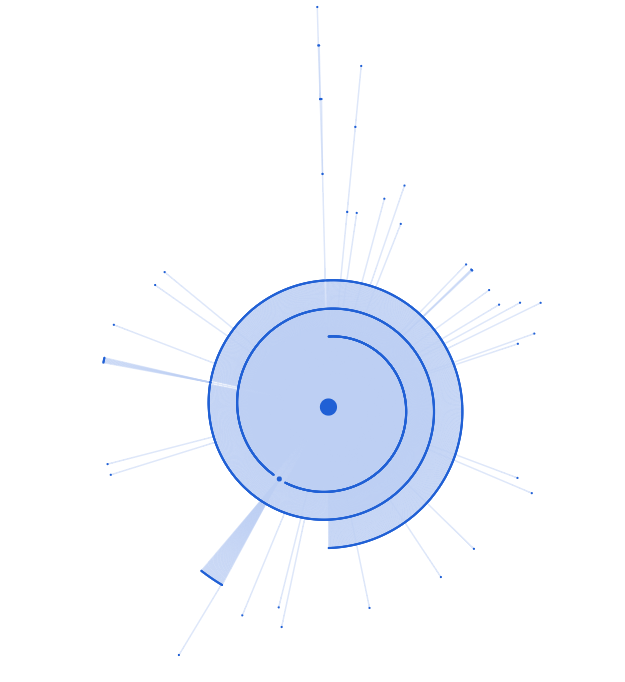


图3-x明星日常微博-孙俪微博统计1

在该条消息中虽然出现了传播层次较深的几条传播链，但基本是单线传播，没有在线上被广泛传播从而形成爆发点，没有造成较大的影响力。在手动查看这几个链条上相关节点的微博，发现这些用户之间关系密切，大多都是双向关注，而且互动比较频繁，导致了彼此之间经常互相转发彼此微博的习惯。

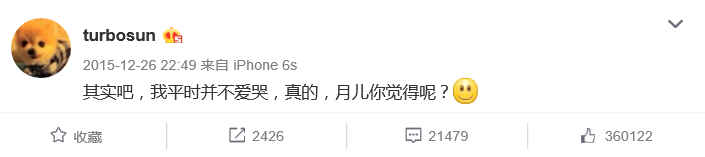


图3-x 明星微博原文示例-孙俪微博原文2

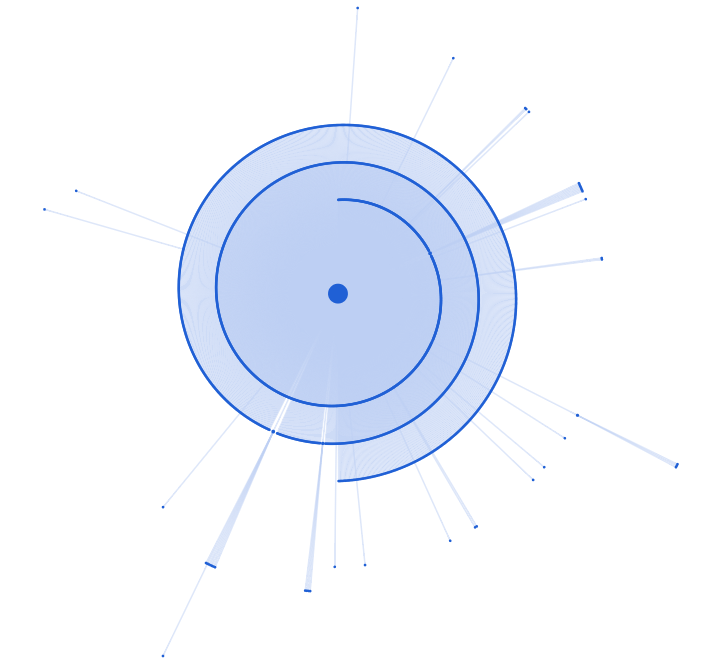


图3-x明星日常微博-孙俪微博统计2

综上所述，明星微博虽然具有巨大的转发量和影响范围，但内容大多数是日常生活中的平淡小事，其传播往往呈现出比较单调的波纹式传播，由自身节点大往外范围扩张。在这种情形下，微博所起到的作用主要是博主和粉丝的消息传递和互动，或者是维持一个以博主为中心的在线社交团体。这种传播具有广度大、深度小的特点。

### 3.2.2.2.2 波纹式传播模型的可演变性

然而有些明星日常微博本应该是常见的波纹式传播模型，却因为其它重要用户的加入从而改变了既有的传播形态，形成具有其它模型特点的传播局面。例如如下微博：



图3-x 明星微博原文示例-方中信微博原文3

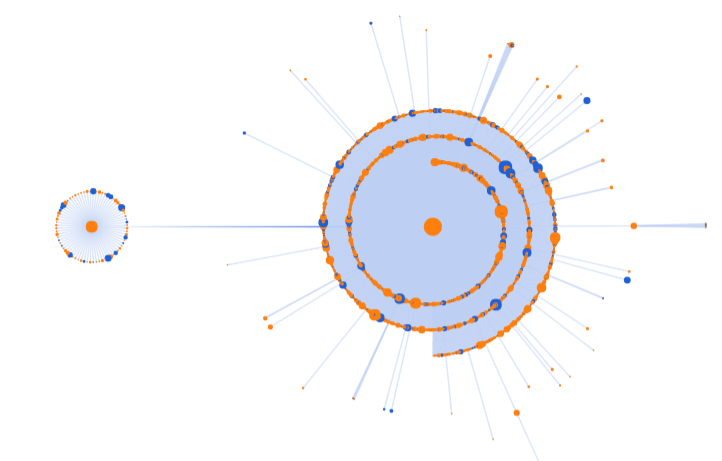


图3-x其它因素导致传播形态转变

这条微博和图3-x – 3-x类似，都是演员发布关于剧情的自言自语式的微博，但由于账号“芈月传官微”的转发而造成了具有类似于“双子星式”的传播结果。账号“芈月传官微”在一定程度上起了推波助澜的作用，同时也提高了该账号的影响力，利用演员的微博做到了宣传作用。

该现象说明不同传播模型之间是可以相互转化的，不同的消息传播树模型是各方面因素综合起来形成的结果，只有合理利用各种有利因素才能实现消息传播影响的最大化。

### 3.2.2.2.3 时效性差的消息之传播模式多样性

由于微博的消息传播树模型太多，而且即使了解了每一种传播的情形对我们的研究目的没有太大帮助。为了简化传播模型分析，我们将传播分为波纹式（中心化）、演化式（半中心化）、多极式（去中心化）三种模型。通过对消息传播的广度及深度进行统计，我们发现中心化的波纹式传播的二级以上（含二级）转发数往往小于5%；人能主观觉察到的具有半中心化的传播模型的二级以上转发数在10%左右，这种情形通常是原博主的具有影响力的好友参与了转发，造成了另一个转发中心，但该中心的影响力远远比中心节点小；当二级以上转发量达到20%时，人能明显感觉到多个中心的存在，而且不同中心都具有较强的影响力。

经过实验分析，我们发现非名人类热门微博往往呈现出多样化、多极化的传播模式。我们统计了二级以上转发量大于12%的微博博主的身份、粉丝量、日常微博，发现样本里76%以上的博主不是娱乐明星及公众人物。即这些微博不像名人一样靠个人魅力赢得关注，而是靠不断推出新的、有意思、有价值的内容而受到关注。他们的粉丝量与娱乐明星相比较少，处于一个中等数量级。

示例1：关于裙子颜色的微博



图3-x关于“裙子颜色”微博原文

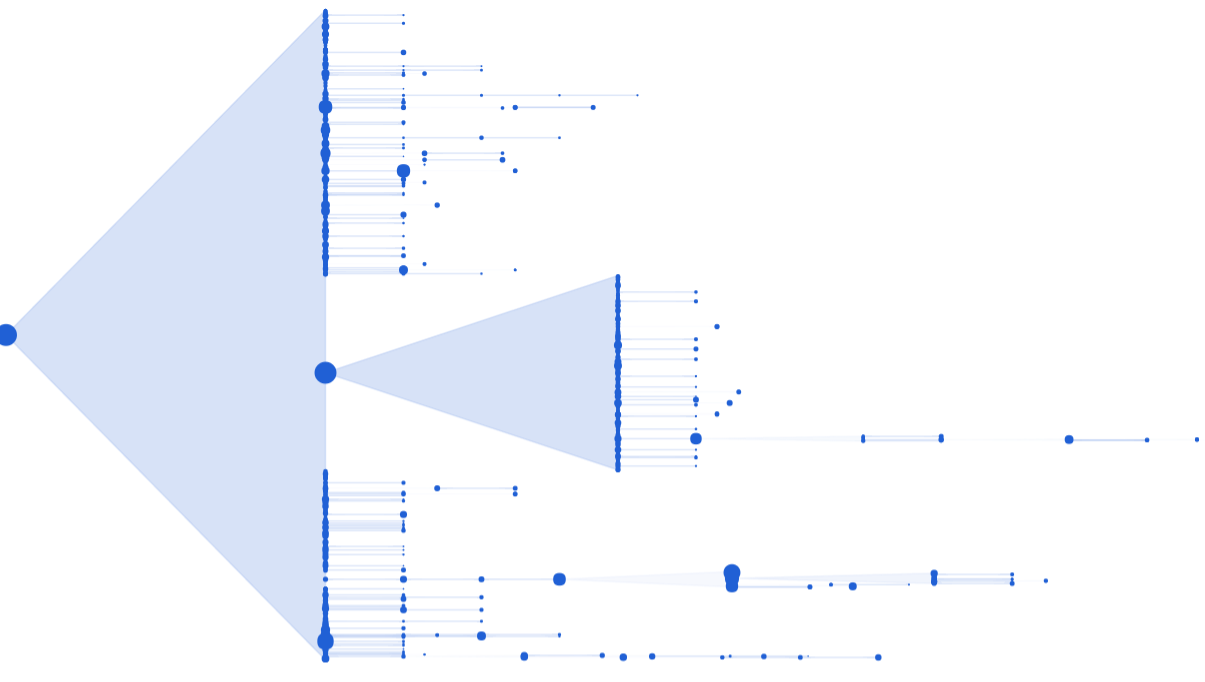


图3-x关于“裙子颜色”微博转发拓扑

示例2：关于“重要的事情说三遍”



图3-x关于“说三遍”的微博原文



图3-x关于“说三遍”的微博转发拓扑

显而易见，内容质量较高的微博，再加上适量的粉丝数，也能达到规模较大的传播效果。这给我们的启发是:在没有明星微博基数大的先天优势时，就要想办法来提高微博内容的质量来吸引眼球。换成很通俗的一句话来讲：如果不能靠脸吃饭，那么就一定要有才华了。

## 3.2.3 微博消息传播持久性研究

### 3.2.3.1 快速引爆传播模式

微博消息的类型涵盖社会生活的方方面面，不同类型的消息具有的时效性不同。由于消息的性质不同，所传播的时间也有不同的特点。按照人们的主观认识，一般事件的传播应该经历开始、上升、巅峰、下降、结束几个过程，在各个阶段的传播量大致服从正态分布，往往只有一个较大波峰。我们对数据进行了大量统计，发现大部分微博在传播时间上往往呈现出急速上升、缓慢下降、渐渐停止的发展特征，有一个长长的尾巴，如下图所示：

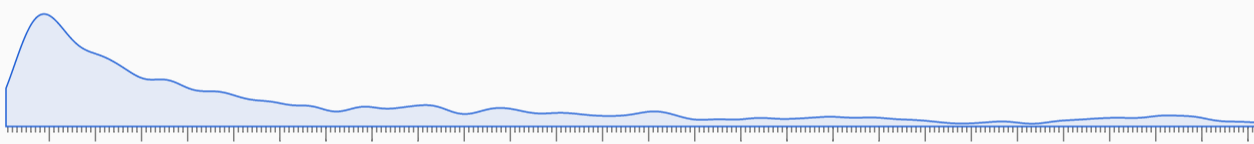


图3-x快速引爆式传播量时间分布示例1



图3-x快速引爆式传播量时间分布示例2

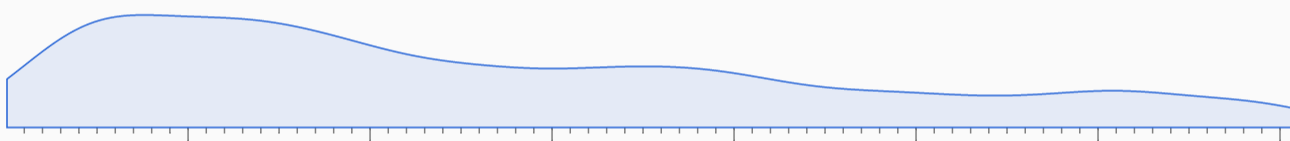


图3-x快速引爆式传播量时间分布示例3

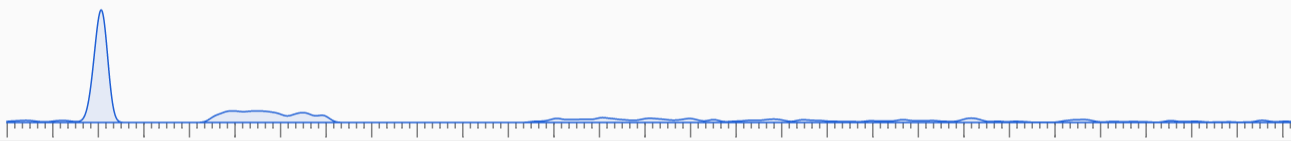


图3-x快速引爆式传播量时间分布示例4

在试验中，我们希望统计在微博消息传播的过程中，符合以上几种情形的传播所占的比例。如何根据传播过程的时间分布数据从数学层面来衡量一个消息的传播趋势，或者说如何定义一种上升速度快、下降过程长、大的波峰唯一且出现较早的传播趋势，这是一个值得考虑的问题。针对这个问题，我们采取了如下指标来衡量:

1. 传播到达最大波峰所用的时间占整个传播过程的比例，即传播到达顶峰所需要的时间长短。这一指标对于衡量信息传播量上升的迅速程度有用。
2. 传播到达最大波峰所用的时间值。这是一个绝对的时间，因为通常消息一定的时效性，在这个期间消息往往作用最大。我们选用天数来做时间单位。
3. 绝对波峰的个数。从单纯数学意义上来讲，波峰只要满足前后值都小于该点即可，但如图3-x所示，我们从直观上看并不认同那些小的波峰是我们想要的波峰。
4. 相邻点所在直线斜率分布。我们知道基于常用平面坐标系，两点间直线斜率与0的大小反映了直线的趋势，大于0则上升，小于0则下降。如果在整个微博消息传播的过程中，斜率小于0的相邻点对占了绝大部分，那么就满足传播量持续下降的特征，也就是说在绝大多数时间里，传播量呈下降趋势。
5. 大的波峰传播量所占传播总量的比例。快速爆发的事件往往在很短时间内获得了大量的传播量，因为消息的传播范围是有限的，就像疾病传播一样，在一段时间内被感染的人数大规模上升，那么在总人数一定的情况下，剩下可以被感染的人就相对少了很多。因此在微博消息引爆点的周围如果聚集的传播量总体较大，这种传播就符合传播量上升迅速的特点，与我们目标统计的类型相同。

经过指标的制定，我们统计了所有样本数据里符合以上传播时间分布的微博消息数量所占的比重，发现该比值是78%。虽然可能由于数据采集不均匀造成该种转播模型在总体中所占比值的不精确，但在一般意义上反映了大部分类型的微博消息像新闻一样都具有一定的时效性，他们从发布时间起，在一定的时间段内迅速传播，然后被人们慢慢淡忘。经过对这些消息的内容、作者、类别进行分析，我们发现这些消息大多是日常的琐事，不具有很大意义、情感比较平淡的明星微博，例如很多人喜欢把生活中微不足道的小事，小到吃一碗面、买一件衣服，更不要说自拍党去什么地方旅游、体验了什么新奇的东西。这些事情本身不具有影响力，但由于博主的粉丝对博主有较深的感情依赖，所以会对其微博进行转发，形成频繁的互动。此外，一些新闻媒体或其他用户报道的新闻时事也大多具有这样的传播特征。

### 3.2.3.2 不规律引爆点传播模式

经过统计，在样本中有19.7%的微博消息呈现出引爆点随机分布的不规则传播模式。这种分布的特点是传播量时起时落，有若干个随机的时间段是爆发期，整个传播过程是不规则的，有可能期间遇到影响力大的节点而使传播迅速加剧，此种情况下与迅速引爆式传播的最大区别在于这个点并不一定而且大多数情况下并不在消息传播的初期。

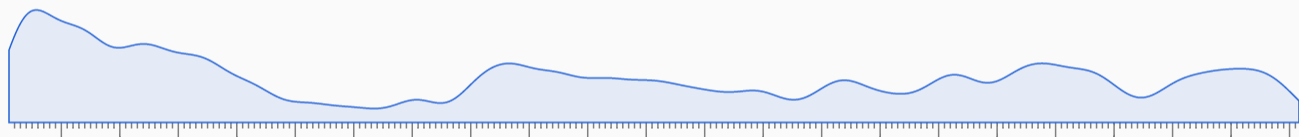


图3-x随机引爆式传播量时间分布示例1

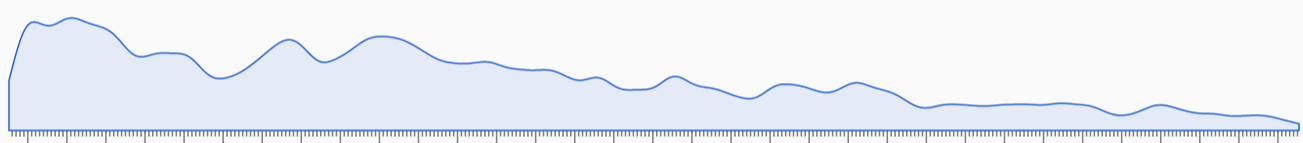


图3-x随机引爆式传播量时间分布示例2

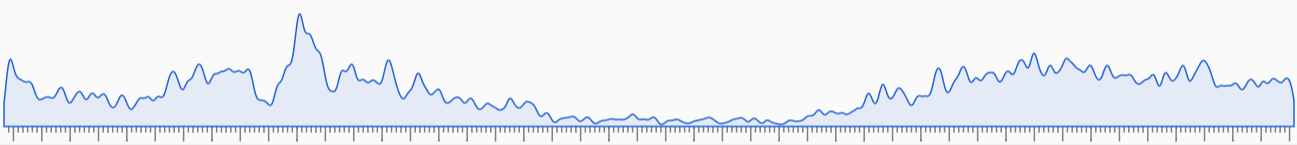


图3-x随机引爆式传播量时间分布示例3



图3-x随机引爆式传播量时间分布示例4

经过对这些消息的内容、作者、类别进行分析，我们发现这些消息大多是时效无关型。通常是笑话、教育、趣事、流行梗之类的微博消息。这样的微博在传播过程中就像一条大河在大地上流淌，遇到平坦的地势时水流相对缓慢，遇到瀑布、峡谷时水流的速度就被激发出来，但在整个河流的流动中，何处平坦，何处有瀑布是无法断定的。同理，在复杂的微博环境中，这些消息的传播会随着转发者的影响力强弱而起伏不定，有可能一直很平缓，有可能有很多个爆发阶段，也有可能在很长一段时间过来被引爆。

虽然这类消息的时效性差，但不会影响其整个过程造成的影响力。而且，这类微博传播的时间通常很久，在试验中我们发现有的微博在2个月之后被人转发，并重新被大量传播开来。

## 3.2.4 影响微博传播结果的因素研究

不同的消息传播树模型，以及消息传播量的时间分布与博主、消息内容、粉丝之间有何相关性。为了研究这个问题，我们做了两个实验。其中一个试验中我们选取n个特定用户，针对每个用户分析其所发微博消息最终的传播结果，以分析同一用户所发微博的传播特征是否相似；另一个实验我们挑选相同内容的微博查看经不同用户发布后形成的传播结果，以分析同一消息内容的微博传播特征是否具有相似性。

### 3.2.4.1 同一用户的微博之传播相似性研究

在试验中，我们选取了20个粉丝量较大的活跃用户，对其微博消息的传播模式进行研究。我们首先对如下指标做了统计：一层转发量与总转发量之比、二层及以上转发量与总转发量之比、传播中心个数、传播深度、传播者男女比例。经过统计我们发现相同博主的微博传播模式大体是相似的。

例如如下明星微博的传播模型属于典型的波纹式演化型，在传播的过程中原博主的影响最大，但总有一两个节点起到了较大的辅助作用，因而脱离了单调的中心性特别强的波纹式模型。

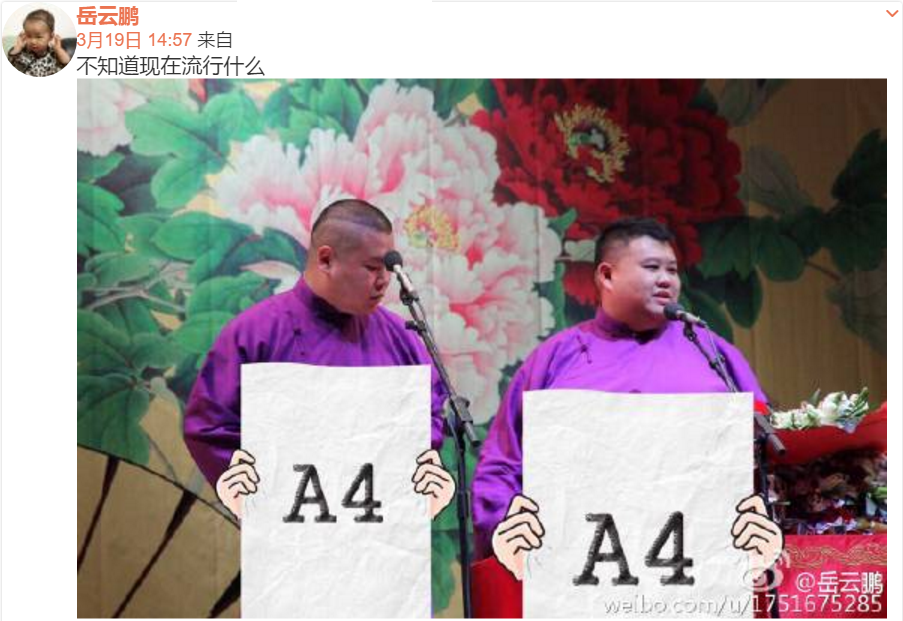


图3-x 同用户传播实验-微博原文1

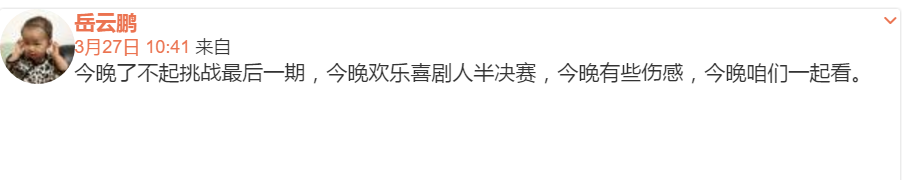


图3-x 同用户传播实验-微博原文2

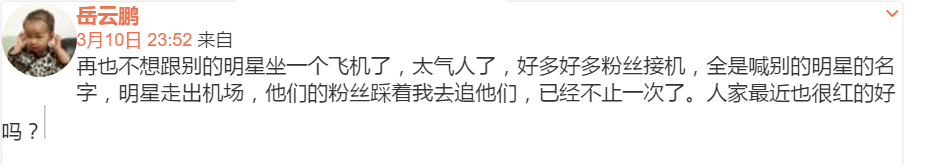


图3-x 同用户传播实验-微博原文3

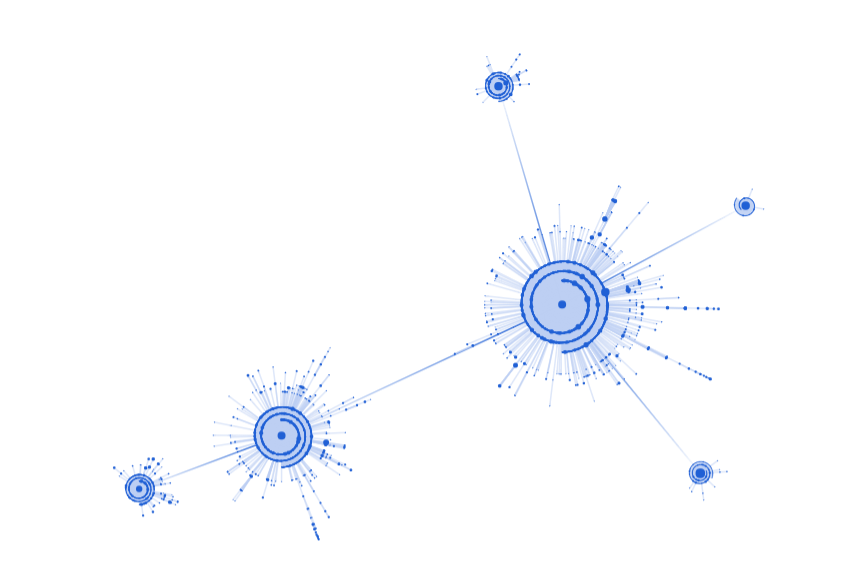


图3-x 同用户传播实验-传播结构1

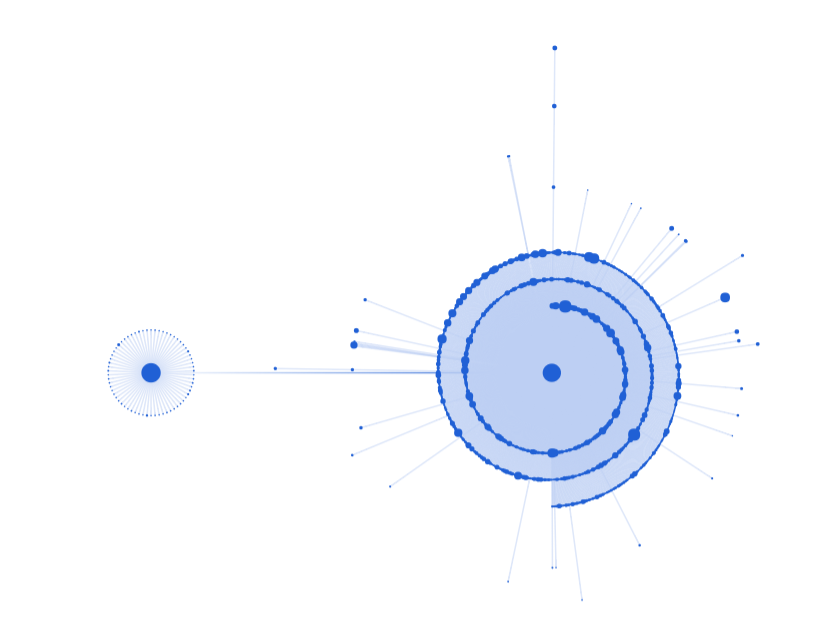


图3-x 同用户传播实验-传播结构2



图3-x 同用户传播实验-传播结构3

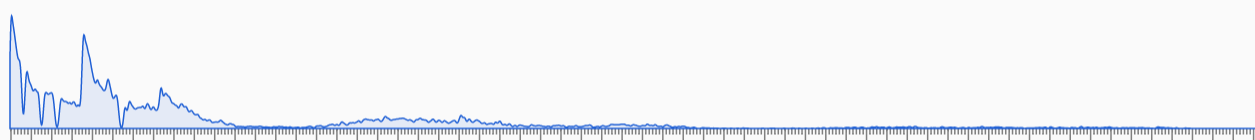


图3-x 同用户传播实验-传播量时间分布1

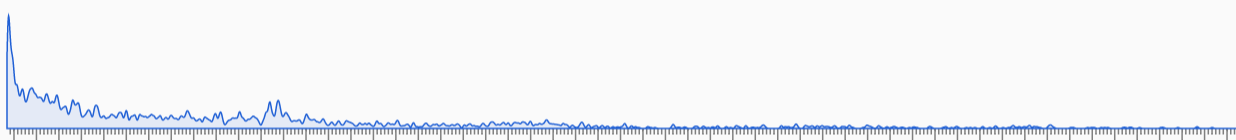


图3-x 同用户传播实验-传播量时间分布2

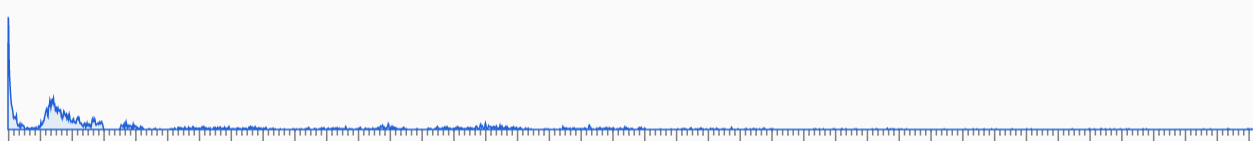


图3-x 同用户传播实验-传播量时间分布3

以上情形在我们的实验中占到的比重是86.3%，由此可见，一个人或者一个微博账号的信息传播模型具有很大相似性。这可能是由于人的社会属性、人的生活习惯、账号的功能属性以及博主的粉丝构成等原因导致的结果。

### 3.2.4.2 相同内容的微博之传播相似性研究

与上一小节的实验类似，我们采用了相同的指标来衡量某一消息传播的模型特征，然后与其它相同消息作比较。由于转发量大的微博才有研究的价值，对于同样内容的微博被不同大V用户转发的情形相对较少，在此仅抽样5条微博，它们均被3个以上大V用户以原作者的身份转发过。统计结果如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 传播男女比 | 一层转发比 | 二层转发比 | 二层以上比值 | 中心点个数 | 主要层数 |
| 样本1 | 0.0052 | 0.0003 | 0.003 | 0.003 | 0 | 0.6668 |
| 样本2 | 0.0034 | 0.0006 | 0.002 | 0.002 | 0 | 0.6668 |
| 样本3 | 0.0028 | 0.0011 | 0.005 | 0.0005 | 0.6534 | 2 |
| 样本4 | 0.0069 | 0.0008 | 0.003 | 0.0004 | 0 | 0 |
| 样本5 | 0.0013 | 0.0002 | 0.0001 | 0.0002 | 0 | 0.6668 |

表中对应的值是该样本下所有转发情况在该指标上的方差，例如:样本1由三位用户转发，对应三条消息的传播男女比分别为0.16、0.33、0.20，则各值为三个比值的方差0.0052。

下图是样本1的微博原文和信息传播树模型图示：

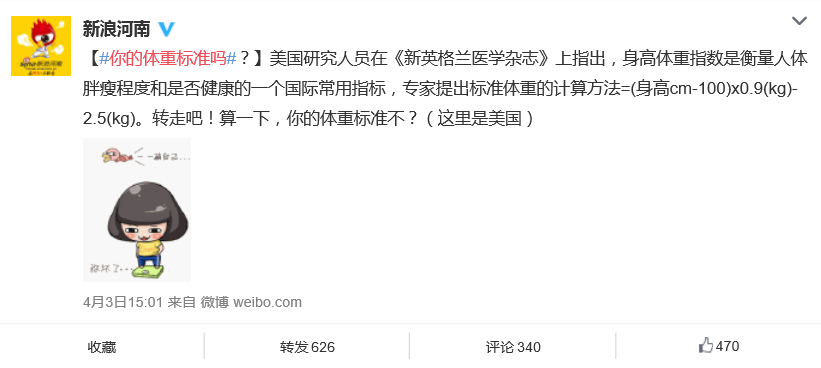


图3-x关于“体重”微博原文1

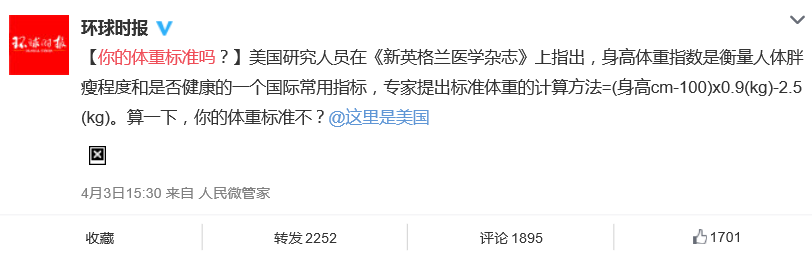


图3-x关于“体重”微博原文2

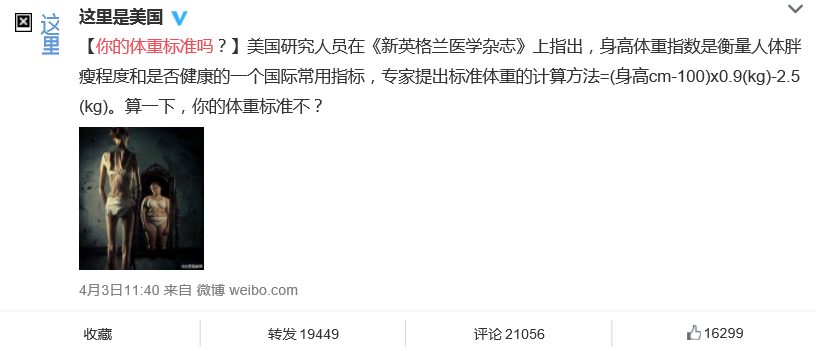


图3-x关于“体重”微博原文3

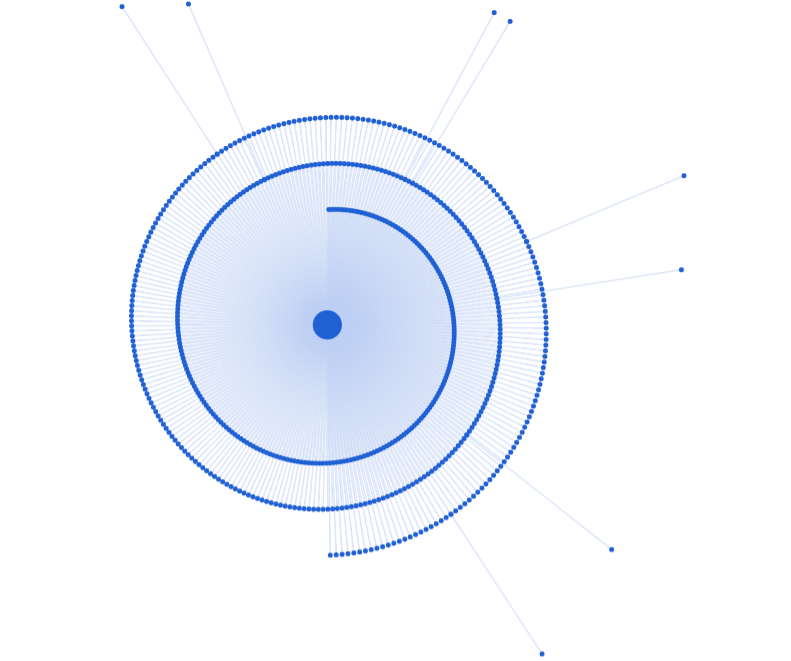


图3-x关于“体重”微博传播模型1

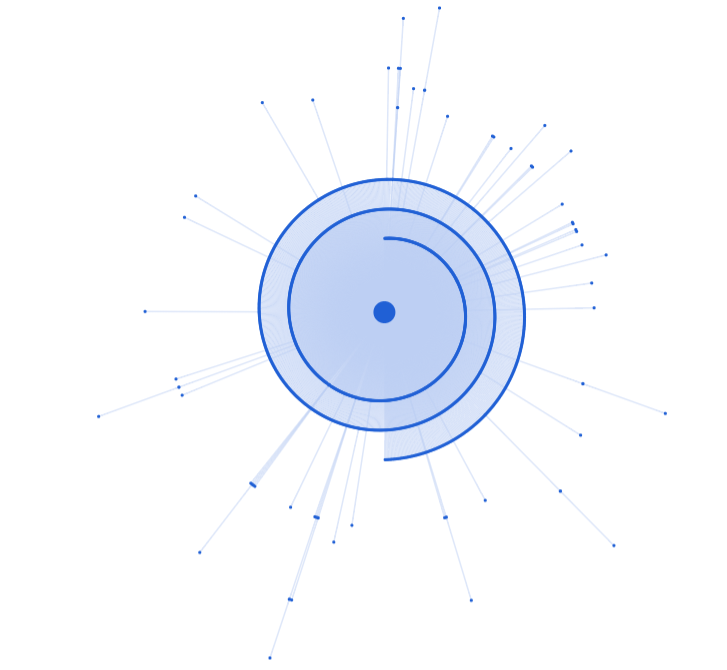


图3-x关于“体重”微博传播模型2

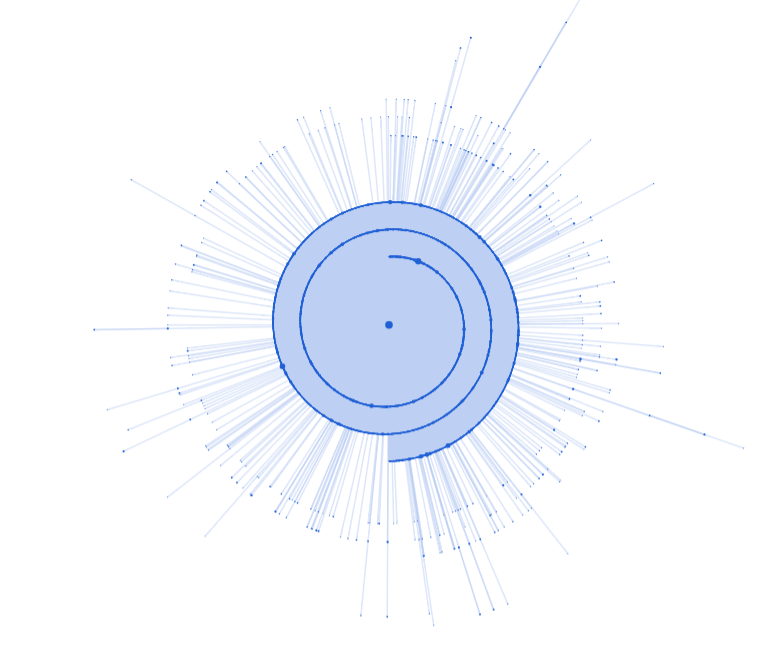


图3-x关于“体重”微博传播模型3

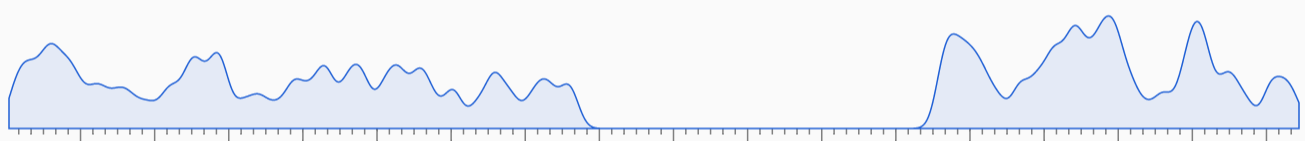


图3-x“体重”微博传播量时间分布示例1

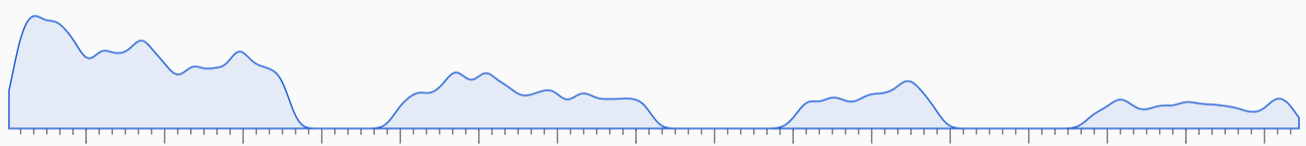


图3-x“体重”微博传播量时间分布示例2

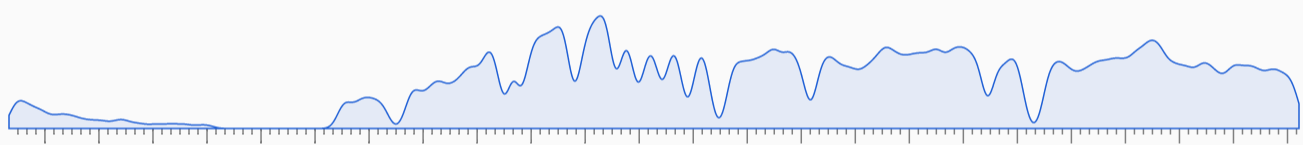


图3-x“体重”微博传播量时间分布示例3

如图所示，经过大量统计，我们发现同样内容的微博经过不同博主转发后，尽管总转发量上差别明显，而且不在一个层次上，但总体的传播模式是高度相似的。由此可见，微博消息传播的拓扑结构、影响深度、传播时间分布等因素主要是由微博消息内容决定的。在消息内容一定的条件下，消息传播虽然在广度上差别可能较大，但总体的传播模式是固定的。

第4章 微博粉丝聚类与转发预测研究

# 4.1 研究内容

在微博中存在大量的超级用户，就是我们常说的大V用户，他们拥有非常多的粉丝，其中粉丝多的可高达7800万以上，这个数量比很多大的发达国家的人口总量还要大，在微博信息传播中具有重大的影响力。对微博粉丝的研究对于微博内容的受众分析具有重要意义。

## 4.1.1 粉丝聚类

和现实生活中一样，微博中存在着各种各样的用户，他们的职业、地域、年龄、受教育程度、兴趣点、经历存在各种各样的差别，导致在微博中的行为习惯有所不同。然而物以类聚，人以群分，我们总能找到合适的方法来把用户归入一些类中，在同一个类中的所有用户在一些方面具有较高相似性。在微博中，对全局用户的聚类对于从宏观研究传播具有一定意义，但全局的微博形式错综复杂，对于我们研究特定消息的传播来讲，对局部用户聚类具有更大的适应性。在本实验中我们将对同一博主的粉丝进行聚类，目的是为了预测粉丝对于消息内容的转发。

尽管影响力大的用户的粉丝很多，但对于该博主所发微博参与互动的人则是很少一部分（相对而言，由于粉丝基数大，该值也响应很大），对于博主发布的内容，粉丝们通常会根据自己的爱好作选择性转发。所以，博主所发布的历史消息中总会有不同的人群转发他们喜欢的那部分。在下面的试验中，我们将对博主的粉丝按照转发内容相似性进行聚类，我们选择了常用的。

### 4.1.1.1 K-MEANS算法

K-means算法[[21]](#endnote-22)[[22]](#endnote-23)是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

k个初始类聚类中心点的选取对聚类结果具有较大的影响，因为在该算法第一步中是随机的选取任意k个对象作为初始聚类的中心，初始地代表一个簇。该算法在每次迭代中对数据集中剩余的每个对象，根据其与各个簇中心的距离将每个对象重新赋给最近的簇。当考察完所有数据对象后，一次迭代运算完成，新的聚类中心被计算出来。如果在一次迭代前后，J的值没有发生变化，说明算法已经收敛。

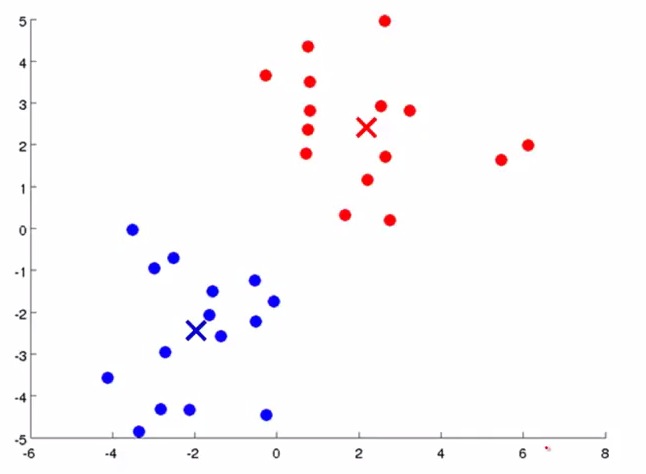


图4-1 k-means算法简单聚类效果示例

k-means 算法的处理流程：

1. 从 n个数据对象任意选择 k 个对象作为初始聚类中心；
2. 根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分；
3. 重新计算每个（有变化）聚类的均值（中心对象）；
4. 循环步骤 2 和 3，直到每个聚类不再发生变化为止；

k-means 算法接受输入量 k ；然后将n个数据对象划分为 k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”（引力中心）来进行计算的。

### 4.1.1.1 聚类实验

周所周至，聚类的依据是点与点之间的距离，比较近的点群适合被聚为一类。在本实验中，我们用任意两个人同时转发相同微博的次数作为衡量两个人之间距离的依据，如果两个人经常转发相同的微博，那么主观上可以判断他们之间距离较近，适合被聚到一个类中。

在进行实验的过程中，我们发现一个棘手的问题: 我们所研究所采用的样本博主大多是粉丝量巨大的用户，如果按照传统做法来衡量任意两个粉丝用户之间的距离，那么不管是计算的过程带来的时间复杂度还是建立索引带来的空间复杂度都是相当大的。例如对于一个有1000万粉丝的用户，我们计算任意两用户之间距离时将至少计算10,000,000 \* 10,000,000次，因此当我们进行大量实验时，所花费的时间将是一个不小的数字。对于解决这个问题，我们想到了如下的解决方案：逐条分析博主所发布的微博，在单次微博中对于所有参与转发的用户，使他们两两之间的次数值分别加1，并建立索引，进而在后续的微博分析中通过对索引的访问来累记用户之间共同转发的次数。然而单机的处理能力毕竟是有限的，我们把这个事情交给Hadoop集群[[23]](#endnote-24)[[24]](#endnote-25)[[25]](#endnote-26)来处理，实现了很好的效率，这不是本文研究的重点，在此不做深入讨论。

对于聚类个数的选取，我们分别使用了k=3和k=4，由于实验量较大而并没有有更多k的聚类尝试。

此实验中的聚类是为了接下来进行的转发预测，因此在此阶段没有一个既定的标准来评测聚类效果的好与坏。而且转发预测的准确性也作为评价聚类效果，因为影响转发预测结果的因素可能来自于其它因素。因此，本节不再对聚类算法的效率进行讨论。

## 4.1.1 转发预测

对于微博转发的预测可以分为宏观上对于消息传播的总体趋势和形态进行预测，以及从微观上对用户是否转发某消息进行预测。本节主要从微观上来预测用户的消息转发，微观转发预测的研究对于消息的精准推荐具有现实意义。如果用户对于收到的推荐消息都比较喜欢，那么至少对于微博官方提升用户体验具有帮助。

基于上一节我们对微博博主粉丝所做的聚类，来完成对用户转发的预测。主要分为以下几个步骤：

* 1. 首先我们将博主大部分历史微博信息作为训练集；
  2. 基于训练集对所有参与转发的用户聚类；
  3. 选取除训练集之外的较新的微博作为测试集；
  4. 在测试集中选取参与转发时间较早的1/3用户作为新的训练集，剩下2/3用户作为测试集；
  5. 根据新训练集中的用户，预测该用户所在社团（类）会对该条信息进行转发；
  6. 评测预测结果

我们采用了F1-Measure来衡量模型的效率。F-Measure又称为F-Score，是IR（信息检索）领域的常用的一个评价标准，计算公式为：

其中其中F1-Measure、准确率、召回率是参数，P是准确率(Precision)，R是召回率(Recall)。[[26]](#endnote-27)[[27]](#endnote-28)

经过对10个样本的大V用户进行实验，对每个博主预测其中三条微博的粉丝转发情况，对30个样本的统计结果如下：

其中平均准确率为，平均召回率为，平均F1值为。这个实验结果表示预测的准确率普遍较高，其中有部分原因在于这个是基于博主粉丝的局部预测，加上我们预测的对象是我们在历史转发数据的基础上聚类所得的集合，针对性较强。此实验的缺点为不是以用户的属性特征来对用户进行行为进行预测，对预测对象的实验要求较高，应用场景的局限性较大。但对于研究微博信息传播有一定的价值，从一个角度说明了在掌握了用户的行为历史信息的情况下，用户的传播行为具有可预测性。

结论

虽然国内外对于在线社交媒体的研究较多

参考文献

1. Kermack W O, McKendrick A G. A contribution to the mathematical theory of epidemics[C]//Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences. The Royal Society, 1927, 115(772): 700-721. [↑](#endnote-ref-2)
2. Watts D J. Six degrees: The science of a connected age[M]. WW Norton & Company, 2004. [↑](#endnote-ref-3)
3. Milgram S. The small world problem[J]. Psychology today, 1967, 2(1): 60-67. [↑](#endnote-ref-4)
4. 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-5)
5. Granovetter M. Threshold models of collective behavior[J]. American Journal of Sociology, 1978, 83(6):1420-1443. [↑](#endnote-ref-6)
6. Schelling T. Micromotives and Macrobehaviour[J]. Norton, 1978. [↑](#endnote-ref-7)
7. 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-8)
8. 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-9)
9. Saito K, Kimura M, Ohara K, et al. Generative Models of Information Diffusion with Asynchronous Timedelay[C]//ACML. 2010: 193-208. [↑](#endnote-ref-10)
10. 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-11)
11. 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-12)
12. 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-13)
13. 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-14)
14. 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-15)
15. 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-16)
16. 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-17)
17. Ghosh R, Lerman K. Predicting Influential Users in Online Social Networks[C]// In: Sna-kdd: Kdd Workshop on Social Network Analysis. 2010. [↑](#endnote-ref-18)
18. Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52. [↑](#endnote-ref-19)
19. Abdi H, Williams L J. Principal component analysis[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459. [↑](#endnote-ref-20)
20. 易成岐, 鲍媛媛, 薛一波,等. 新浪微博的大规模信息传播规律研究[J]. 计算机科学与探索, 2013, 7(6):551-561. [↑](#endnote-ref-21)
21. Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C]// Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297. [↑](#endnote-ref-22)
22. Forgy E W. Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classification[J]. International Journal of Environmental Studies, 1992, 42(42):41-52. [↑](#endnote-ref-23)
23. Ghemawat S, Gobioff H, Leung S T. The Google file system[C]//ACM SIGOPS operating systems review. ACM, 2003, 37(5): 29-43. [↑](#endnote-ref-24)
24. Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data[J]. ACM Transactions on Computer Systems (TOCS), 2008, 26(2): 4. [↑](#endnote-ref-25)
25. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113. [↑](#endnote-ref-26)
26. C. J. Van Rijsbergen. Information Retrieval[J]. Information Processing & Management, 1979, 2(2-3):320–322. [↑](#endnote-ref-27)
27. Perruchet P, Peereman R. The exploitation of distributional information in syllable processing[J]. Journal of Neurolinguistics, 2004, 17(s 2–3):97-119. [↑](#endnote-ref-28)