

硕士学位论文

**基于交互行为的微博用户影响力研究**

学位申请人：孙思禹

指导教师：饶元 副教授

类别（领域）：工程硕士（软件工程）

2017年04月

**Research on The Influence of Micro Blog Users Based on Interactive Behavior**

A thesis submitted to

Xi’an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

By

Siyu Sun

Supervisor: Assoc.Prof.Yuan Rao

(Software Engineering)

April 2017

**论文题目：基于交互行为的微博用户影响力研究**

**类别（领域）：工程硕士（软件工程）**

**学位申请人：孙思禹**

**指导教师：饶元 副教授**

摘 要

随着互联网技术的迅猛发展，社交媒体作为一种新的信息交流平台已经融入我们的生活。其中，新浪微博以其独特的性质获得了国内外学者的广泛关注，具有重大的研究价值。用户的影响力能够直接体现用户在该网络中的地位与重要性，如何在微博中选取一批用户集合，使得从他们出发进行产品的推广，能够影响到最大范围的人群，成为我们关注的焦点。影响力的度量和影响力最大化构成了微博中影响力研究的重要研究课题。本文工作概括如下：

首先，我们设计了基于cookie的微博数据爬虫，分析总结了新浪微博数据的特征，利用爬虫技术爬取的新浪微博数据，并给出了结合微博网络的基本属性和关系，构建了微博社交网络；其次，详细分析了新浪微博用户的交互行为，提出影响因子的概念，根据影响因子，建立影响力矩阵，结合经典的mdd算法提出一种影响力度量方法，综合考虑了节点的全局重要性与局部重要性，并且考虑了节点对其邻居节点的影响力贡献不同，用节点的mdd值和相邻节点之间的影响力两方面综合度量，更加符合实际情况；最后，在影响力度量的基础上，深入研究了影响力最大化问题，提出基于活跃度的改进的独立级联模型，并在基础上改进了最大化算法，取得了时间效率上的提高，并且影响范围也略有提高。通过结果的对比和分析，发现上述方法能够取得较优的结果，并且降低了时间复杂度，能在实验范围内取得相对于其他最大化算法更大的激活范围，表明了算法的有效性和合理性。

综上所述，通过真实的新浪微博数据提出并验证了本文提出的算法模型。影响力最大化算法有较高的激活范围和较低的时间复杂度，能帮助决策者进行分析，对于网络中的营销有积极作用，并且对于在微博中的影响力研究工作具有一定的指导和实践意义。

**关 键 词**：社交媒体；新浪微博；交互行为；影响力度量；影响力最大化

**论文类型**：应用研究

**Title: Research on The Influence of Micro Blog Users Based on Interactive Behavior**

**Professional Fields: Software Engineering**

**Applicant:Siyu Sun**

**Supervisor: Assoc.Prof. Yuan Rao**

ABSTRACT

With the rapid development of Internet technology, social media as a new information communication platform has been integrated into our life. Among them, Sina micro-blog with its unique nature has received wide attention of scholars at home and abroad, it has great research value. The influence of the user can directly reflect the status and importance of the users in the network. How to select a set of users in micro-blog, the promotion of products from them, can affect the maximum range of people, become the focus of our attention. The measurement of influence and the influence maximization constitute an important research topic in the study on the influence of micro-blog. The paper is summarized as follows:

First of all, we designed the micro-blog crawler based on cookie data analysis, summarizes the features of sina micro-blog data, Sina micro-blog data taken by climbing crawler technology, and gives the basic attributes and relationships with micro-blog network,established a micro-blog social network; secondly, we analysis the interactive behavior of sina micro-blog users in detail, put forward the concept of the impact factor and according to the impact factor, establish influence matrix, combined with classical MDD algorithm proposed an influence measure method, considering the importance of global and local node importance,And consider the influence of the node to its neighbor node contribution is different,with the MDD value of the node and the influence of the neighboring nodes in two aspects of comprehensive measurement, more in line with the actual situation. Finally, based on the influence on the measurement, in-depth study of the influence maximization problem, put forward independent cascade model based on improved activity, and improved the algorithm of maximizing the basis, made time efficiency improved, and the influence scope is slightly increased.By comparing and analyzing the results, the method can achieve better results and reduce the time complexity, the algorithm can achieve greater activation range relative to other maximization algorithms in the range of experiments, the results show the effectiveness and rationality.

To sum up, through the Sina micro-blog real data proposed and validate the algorithm model in this paper.The influence maximization algorithm has higher activation range and lower time complexity,it can help decision makers to analyze.It plays a positive role in the marketing of the network, and has some guidance and practical significance for the research on the influence of micro-blog

**KEY WORDS**: Social media; Sina micro-blog; Interactive behavior; Influence measurement; Influence maximization

**TYPE OF THESIS**: Application Research

# 

目 录

[1绪论 1](#_Toc22345)

[1.1课题背景和意义 1](#_Toc13057)

[1.2国内外研究现状 2](#_Toc11173)

[1.2.1影响力度量 2](#_Toc1990)

[1.2.2影响力最大化 3](#_Toc9450)

[1.3论文的主要研究内容 5](#_Toc28116)

[1.4论文组织结构 6](#_Toc21314)

[2相关理论知识与技术 7](#_Toc16099)

[2.1社交媒体与社会网络 7](#_Toc21234)

[2.1.1社交媒体 7](#_Toc20967)

[2.1.2社会网络 8](#_Toc9168)

[2.2新浪微博 8](#_Toc299)

[2.3影响力评价方法 9](#_Toc29067)

[2.3.1基于网络局部属性的最大度算法 9](#_Toc16899)

[2.3.2基于网络全局属性的距离中心算法 9](#_Toc31485)

[2.2.3基于随机游走模型的LeaderRank算法 10](#_Toc25435)

[2.2.4基于网络位置的的K-核混合分解法（MDD） 10](#_Toc30994)

[2.4影响力传播模型 13](#_Toc29033)

[2.4.1两种经典的传播动力学模型 13](#_Toc10780)

[2.4.2独立级联模型和线性阈值模型 13](#_Toc28750)

[2.5影响力最大化算法 14](#_Toc3204)

[2.5.1贪心爬山算法 14](#_Toc27760)

[2.5.2启发式算法 15](#_Toc27699)

[2.5.3其他相关算法 16](#_Toc5828)

[2.6本章小结 16](#_Toc16435)

[3微博社交网络构建与分析 18](#_Toc5231)

[3.1微博数据源分析 18](#_Toc591)

[3.1.1 微博信息数据 19](#_Toc9315)

[3.1.2 用户关系数据 19](#_Toc14526)

[3.1.3 用户信息数据 21](#_Toc4829)

[3.2微博数据获取 22](#_Toc27163)

[3.2.1微博数据存储 22](#_Toc29179)

[3.2.2基于Cookie的微博数据半自动爬虫 24](#_Toc13887)

[3.2.2关键技术点 26](#_Toc1975)

[3.3微博网络的构建 27](#_Toc25265)

[3.3.1 微博网络基本属性与关系 27](#_Toc9486)

[3.3.2 微博社交网络的构建 27](#_Toc24518)

[3.4微博网络数据可视化分析 29](#_Toc30498)

[3.4.1 整体的微博社交网络 30](#_Toc13232)

[3.4.2 局部的微博社交网络 31](#_Toc10896)

[3.5本章小结 31](#_Toc25351)

[4 微博用户影响力评价算法 33](#_Toc14930)

[4.1微博交互行为分析 33](#_Toc27279)

[4.1.1转发行为 33](#_Toc20203)

[4.1.2评论行为 34](#_Toc15439)

[4.1.3赞行为 34](#_Toc30634)

[4.1.4关注行为 34](#_Toc561)

[4.2微博用户影响因子 35](#_Toc5958)

[4.2.1微博用户的影响力 35](#_Toc15037)

[4.2.2影响因子 35](#_Toc26149)

[4.3 MIB算法模型 36](#_Toc10425)

[4.3.1算法设计 37](#_Toc16391)

[4.3.2算法步骤 37](#_Toc5294)

[4.4算法仿真与分析 38](#_Toc24564)

[4.4.1数据集描述 38](#_Toc6047)

[4.4.2实验项目 39](#_Toc1174)

[4.4.2实验环境 39](#_Toc8313)

[4.4.3实验结果 40](#_Toc2153)

[4.4.4实验结论 45](#_Toc12804)

[4.5本章小结 45](#_Toc5220)

[5 微博用户影响力最大化算法 47](#_Toc1373)

[5.1影响力度量 47](#_Toc2323)

[5.2影响力传播模型 47](#_Toc23874)

[5.2.1传统的独立级联模型 47](#_Toc22829)

[5.2.2活跃度 49](#_Toc30903)

[5.2.3基于活跃度的改进的独立级联模型 49](#_Toc10795)

[5.3影响力最大化SMC算法 50](#_Toc18376)

[5.3.1 CELF算法 51](#_Toc29306)

[5.3.2算法设计 51](#_Toc23416)

[5.3.3算法步骤 52](#_Toc18990)

[5.4算法仿真与分析 53](#_Toc4645)

[5.4.1实验数据集 53](#_Toc2665)

[5.4.2实验结果 53](#_Toc19415)

[5.4.3实验结论 55](#_Toc30924)

[5.5本章小结 55](#_Toc3682)

[6 结论与展望 57](#_Toc26544)

[6.1结论 57](#_Toc32579)

[6.2展望 58](#_Toc14593)

[致谢 59](#_Toc2453)

[参考文献 60](#_Toc24800)

CONTENTS

[1 Introduction 1](#_Toc19476)

[1.1 Background and Significance 1](#_Toc12000)

[1.2 Glance of Current Research Status 2](#_Toc17268)

[1.2.1Influence measurement 2](#_Toc3967)

[1.2.2](#_Toc3066)[Influence maximization 3](#_Toc3066)

[1.3 Main Research Contents of Thesis 5](#_Toc15078)

[1.4 Organization of Thesis 6](#_Toc11900)

[2 Theory and Related Technology 7](#_Toc19036)

[2.1](#_Toc27592)[Social Media and](#_Toc27592) [Social Networks 7](#_Toc27592)

[2.1.1 Social Media 7](#_Toc23953)

[2.1.2 Social Networks 8](#_Toc9853)

[2.2Sina micro-blog 8](#_Toc20369)

[2.3Influence evaluation method 9](#_Toc27837)

[2.3.1Maximum degree algorithm based on local attributes of network 9](#_Toc8241)

[2.3.2Distance center algorithm based on global attributes of network 9](#_Toc10328)

[2.2.3LeaderRank algorithm based on random walk model 10](#_Toc7694)

[2.2.4K- kernel hybrid decomposition method based on network position（MDD） 10](#_Toc29021)

[2.4Influence propagation model 13](#_Toc32082)

[2.4.1Two classical propagation dynamics models 13](#_Toc7736)

[2.4.2Independent cascade model and linear threshold model 13](#_Toc10720)

[2.5](#_Toc17867)[Influence maximization algorithm 14](#_Toc17867)

[2.5.1Greedy hill-climbing algorithm 14](#_Toc22418)

[2.5.2heuristic algorithm 15](#_Toc19224)

[2.5.3Other related algorithms 16](#_Toc10665)

[2.6 Chapter Summary 16](#_Toc1914)

[3Construction and analysis of micro-blog social network 18](#_Toc21167)

[3.1Micro-blog data source analysis 18](#_Toc3974)

[3.1.1 Micro-blog information data 19](#_Toc22605)

[3.1.2 User relationship data 19](#_Toc12292)

[3.1.3 User Information data 21](#_Toc17124)

[3.2Micro-blog data acquisition 22](#_Toc27558)

[3.2.1Micro-blog data storage 22](#_Toc32495)

[3.2.2Semi automatic crawler of micro-blog data based on Cookie 24](#_Toc18266)

[3.2.2Key technical points 26](#_Toc9846)

[3.3Construction of micro-blog network 27](#_Toc25209)

[3.3.1 Basic properties and relationships of micro-blog network 27](#_Toc23324)

[3.3.2 Micro-blog social networking 27](#_Toc286)

[3.4Micro-blog network data visualization analysis 29](#_Toc23946)

[3.4.1 Whole micro-blog social network 30](#_Toc8799)

[3.4.2 Local](#_Toc13637) [micro-blog social network 31](#_Toc13637)

[3.5 Chapter Summary 31](#_Toc18765)

[4 Micro-blog user influence evaluation algorithm 33](#_Toc5177)

[4.1Micro-blog interactive behavior analysis 33](#_Toc5851)

[4.1.1Forwarding behavior 33](#_Toc15219)

[4.1.2Comment behavior 34](#_Toc1057)

[4.1.3Praise behavior 34](#_Toc4838)

[4.1.4Attention behavior 34](#_Toc18327)

[4.2](#_Toc29750)[Micro-blog user impact factor 35](#_Toc29750)

[4.2.1Micro-blog user](#_Toc1334) [influence 35](#_Toc1334)

[4.2.2influence factor 35](#_Toc26831)

[4.3 MIB](#_Toc10408) [Algorithm model 36](#_Toc10408)

[4.3.1](#_Toc24030) [Algorithm Design 37](#_Toc24030)

[4.3.2](#_Toc491) [Algorithm Procedures 37](#_Toc491)

[4.4](#_Toc9443)[Algorithm simulation and analysis 38](#_Toc9443)

[4.4.1Data set description 38](#_Toc16402)

[4.4.2Experiment project 39](#_Toc3566)

[4.4.2Experimental environment 39](#_Toc779)

[4.4.3Experimental result 40](#_Toc4308)

[4.4.4Experimental Conclusion 45](#_Toc17551)

[4.5 Chapter Summary 45](#_Toc25806)

[5 Micro-blog user influence maximization](#_Toc6573) [algorithm 47](#_Toc6573)

[5.1Influence measurement 47](#_Toc8812)

[5.2Influence propagation model 47](#_Toc6540)

[5.2.1 Traditional independent cascade model 47](#_Toc10688)

[5.2.2](#_Toc23384) [Activation 49](#_Toc23384)

[5.2.3 Improved independent cascade model based on Activation 49](#_Toc5924)

[5.3Influence maximization SMC algorithm 50](#_Toc19573)

[5.3.1 CELF Algorithm 51](#_Toc8981)

[5.3.2 Algorithm Design 51](#_Toc26987)

[5.3.3 Algorithm Procedures 52](#_Toc28147)

[5.4](#_Toc30514)[Algorithm simulation and analysis 53](#_Toc30514)

[5.4.1 Experiment Data 53](#_Toc32417)

[5.4.2 Experiment Results 53](#_Toc1980)

[5.4.3](#_Toc13156) [Experimental Conclusion 55](#_Toc13156)

[5.5 Chapter Summary 55](#_Toc11250)

[6](#_Toc7788) [Conclusions and](#_Toc7788) [Suggestions 57](#_Toc7788)

[6.1 Conclusions 57](#_Toc1664)

[6.2 Suggestions 58](#_Toc21483)

[Acknowledgements 59](#_Toc20007)

[References 60](#_Toc7594)

，在印预览中看不见即可）：

# 1绪论

1.1课题背景和意义

随着互联网技术的迅猛发展，社交媒体**[1]**作为一种新的信息交流平台已经融入我们的生活。社交媒体(Social Media)指互联网上基于用户关系的内容生产与交换平台。社交媒体是人们[彼此](http://baike.baidu.com/subview/185902/11099593.htm" \t "http://baike.baidu.com/_blank)之间用来分享意见、见解、经验和观点的工具和平台，现阶段主要包括社博客、论坛、[维基](http://baike.baidu.com/view/587292.htm" \t "http://baike.baidu.com/item/_blank)、播客、视频博客、职业社交网站、企业社交网站和其他工具等等。社交媒体在互联网的沃土上蓬勃发展，爆发出令人眩目的能量，其传播的信息已成为人们浏览互联网的重要内容，不仅制造了人们社交生活中争相讨论的一个又一个热门话题，更进而吸引[传统媒体](http://baike.baidu.com/view/817079.htm" \t "http://baike.baidu.com/_blank)争相跟进。人们在虚拟的网络中中发布、共享并相互交换信息，而社交媒体正是提供这种信息交流的平台。常见的社交媒体包括微博、博客、社会网络、贴吧、百度百科等等，其中，微博以其独特的性质获得了国内外学者的广泛关注，具有重大的研究价值。

微博是一种包含了多种复杂的关系和实体关联的一种复杂网络，这种复杂网络是信息交流和传播的载体。但是，微博具有其独特的性质，它采用一种叫做关注的动作建立起整个网络。在微博中，一个用户关注另一个用户并不需要经过其批准，同时，关注的发起者被形象化地称为“粉丝”。一旦两个用户的关注关系被建立，关注者就可以看到被关注者所发布的一切消息。微博中的消息基于文本却不能超过个140字（符）的内容，甚至可以包含图片、超链接、视频等其他信息。2006年，Twitter**[2]**的横空出世标志了微博发展的新纪元。[Twitter](http://baike.baidu.com/view/843376.htm" \t "http://baike.baidu.com/item/_blank)作为一种社交功能与共享特性的社交媒体吸引着数以万计的用户，[社会化](http://baike.baidu.com/view/79745.htm" \t "http://baike.baidu.com/item/_blank)的互联网生活已经在一小批[互联网](http://baike.baidu.com/subview/6825/18005829.htm" \t "http://baike.baidu.com/item/_blank)Geek中成熟发展起来，尤其是每天泡在[Twitter](http://baike.baidu.com/view/843376.htm" \t "http://baike.baidu.com/item/_blank)上的“推客”们，每天孜孜不倦地推送各种新奇好玩的想法，网站，新闻，音乐，视频的链接，这些“推特”（Twitter）消息被一级一级过滤转发（RT，ReTweet），迅速传遍了全球。与此同时，国内的微博市场也在近年来发展的极为迅速。《2016微博用户发展报告》**[3]**指出，截止2016年9月30日，新浪微博月活跃人数已达到了2.97亿，日活跃人数达到了1.32亿。

用户的影响力能够直接体现用户在该网络中的地位与重要性。一般来说，有影响力的用户往往处于网络中的核心位置，他们发出的信息能够快速传播，并影响到其他的用户，甚至改变其他用户的行为。在微博中，由于微博的多元性和用户交互的多样性，单单从网络的拓扑结构进行度量已经不够充分，并且用户自身的一些因素也对影响力的度量产生影响，如何结合这些相关的因素来度量微博中用户的影响力是本文需要解决的一个问题。

在微博中，每一个用户都有自己的粉丝，粉丝可以收到该用户所发布的一切消息。在看到消息后，粉丝又可以传播给粉丝的粉丝，如此下去。通过这种方式，信息逐层向外传播，影响到更多的人群。在营销领域，由最初的“口碑效应”和“病毒式营销策略”**[4]**的推广方式，人们发现可以在人群中选取具有代表性的节点子集，通过它们可以引起更大的级联影响。病毒式营销是通过利用公众的积极性和人际网络，让营销信息像病毒一样传播和扩散，营销信息被快速复制传向数以万计、数以百万计的受众。它的核心在于寻找到有吸引力的病原体，也就是我们所说的造成影响力最大的种子集合。这种营销领域的战略和微博信息传播的特点不谋而合。在微博中，每个用户都可以进行信息的传播，根据用户影响力的不同，影响范围又是不一样的，影响到的用户数量也是不相同的。如何在微博中选取一批用户集合，使得从他们出发进行产品的推广，能够影响到最大范围的人群，成为我们关注的焦点。本文对这个问题会重点进行讨论。

影响力的度量和影响力最大化构成了微博中影响力研究的两大问题。对微博中影响力的研究是有着十分重要的现实意义的。一方面，微博作为一种新的销售平台，受到了越来越多商家的重视。找到网络中影响力大的一些种子节点进行推广，能够覆盖最大范围的人群，这是各商家一致关心的问题。并且研究用户的行为以及信息的传播规律将有助于网络公司对每个用户的偏好进行更为准确地把握，并将其可能感兴趣的话题信息、其他用户或者用户社群推荐给用户。另一方面，随着微博的快速发展，越来越多的人通过微博获取信息，方便快捷。还有一些需要找出影响力较大的种子节点进行抑制影响的传播，例如抑制传染病在人群中的流行，遏制谣言在社会中的扩散等。对于监管信息来源、引导微博健康有序的发展起着重要作用。下文将围绕以上提出的两个问题展开我们对微博中影响力的研究。

1.2国内外研究现状

1.2.1影响力度量

社交影响力的定义具有明显的因果性，而人们的思想、行为等产生变化的原因则是不胜枚举且因人而异，社交影响力只是其中之一。这就给社交影响力的建模和度量带来了很大挑战，同时也是造成社交影响力模型众多的重要原因。同样，在线社交网络中的影响力也与很多因素相关，目前大部分研究工作都是针对社交网络结构及其上的交互信息和用户行为特征进行量化和分析的，因此可以把能对信息传播过程或他人行为产生影响的个体视为具有社交影响力．

传统的社会网络利用连接结构来识别具有高中心性的节点，具体的衡量指标包括度中心性、介数中心性、紧密度中心性和特征向量中心性**[5]**。通过对这些指标的计算，侧面反映了某一个节点的重要性。相对而言，该方法具有一定局限性，它的时间复杂度往往较高，不利于大规模网络的分析。

2010年Meeyoung Cha等人**[6]**在对影响力的三种度量方法：节点入度、转发行为和提及行为进行了深度的比较，使用Spearmans相关系数衡量两种算法的相关性，分别比较了所有的用户、前10%和前1%的用户在三种算法上的到结果的相关性，总结得出具有高入度值，代表用户的受欢迎程度，在转发微博和提及其他用户方面，并不一定是具有重要影响力的。此外，他们还进一步得出影响力的增加并不是自发或者偶然获得的，而是需要用户保持足够的个人参与，比如将所发微博限制在同一个主题上。

2010年《Nature》物理版上，Kitsak等人**[7]**首次提出了节点重要性依赖于其在整个网络中的位置的思想, 并且利用 K-核分解获得了节点重要性排序指标 (k-shell),该指标时间复杂度低, 适用于大型网络, 而且比度、介数更能准确识别在疾病传播中最有影响力的节点。

2011年Linyuan Lü**[8]**提出LeaderRank 算法，在已有节点外另加一个节点(ground node)，并且将它与已有的所有节点双向连接，于是得到N+1个节点的网络，这个新的网络是一个强连通的网络，再按照原始的PageRank算法计算得到原来N个节点的“重要性”排序。于是由于加入ground node后的图是一个连通图，解决的PageRank的排序的不唯一性问题。

2012年Zeng等人**[9]**考虑节点的 Ks 信息和经过Ks分解后被移除节点的信息, 提出了混合度分解方法MDD。基本思路是不仅考虑移除后剩余节点的信息，而且考虑被移除节点的度的信息，解决了在某个ks值范围内有大量相同ks值的情况。

2013年任卓明等**[10]**提出了基于最小 K-核节点邻居集合中最大Ks值的深度指标H(i), 该指标依靠最小K-核节点与网络的其他层级节点的连接关系,判断最小K-核节点的重要性。

2013年Liu 等**[11]**综合考虑目标节点自身 K-核的信息和与网络最大 K-核的距离, 提出了新的度量节点重要性的指标.。该指标解决了 Ks 指标赋予网络中大量节点相同的值导致其无法准确衡量其节点重要性的缺陷。

1.2.2影响力最大化

影响力最大化问题已被证明是一个NP问题**[12]**,也就意味着我们只能够使用启发式的算法来近似最优结果。基于节点度数或节点中心性等是社会网络分析中常见的启发式方法,在某些领域这些简单的启发式方法效果很好。但由于影响最大化问题的特殊性,这些启发式的方法在所要解决的问题上效果并不好,甚至比较差。那么如何定义一种行之有效的启发式策略,并在此基础上如何将启发式策略发挥的更好是我们当前应该关注的焦点。

2001年Richardson和Domingos**[12]**第一次将影响力最大化问题引入到社会网络领域，他们将该问题定义为在网络中找出一个小的初始种子节点集合，使得从这些节点出发传播信息可以影响到最大范围的人群。从此，越来越多的学者致力于该问题的研究，并提出了许多算法。

Kempe等**[[12]**将影响力最大化问题定义为离散优化问题，并证明其是NP难解的。与此同时，他们还提出一个贪心算法，该算法能够取得最优解63%的近似度。贪心算法的基本思想是每次选择一个节点进入种子节点集合当中，该节点使影响力增量最大。尽管该算法能够取得对最优解较高的近似，但是运行时间过长使得它不适用于大规模网络。

为了提高效率，Leskovec**[13]]**提出了一个优化的贪心算法CELF。该算法利用影响力最大化目标函数的子模特性来减少每次需要计算的节点数，在达到与贪心算法相同准确度的前提下，大大减少了运行时间。实验验证，其运行速度比贪心算法提升了700倍。尽管如此，算法在拥有数以万计节点的网络中找出大小为的种子节点集合仍需要数小时。

随后，提出了W.Chen**[14]**两种更快的贪心算法，分别是NewGreedy和MixedGreedy。NewGreedy算法是先去除原始图中对影响力传播没有作用的边，从而得到一个规模较小的图，然后在新图上进行影响力的研究。MixedGreedy算法是NewGreedy算法与算法的结合，其在第一轮使用NewGreedy算法找出种子节点，剩下轮次使用CELF算法。实验证明，MixedGreedy算法比CELF算法更高效。

2010年Y.wang**[15]]**提出CGA算法。该算法是基于社区的贪心算法，首先将网络化分为一系列社区，然后每一轮都利用动态规划的方法选择候选种子节点所在社区，在确定了社区之后就在该社区内利用贪心算法选择节点。以上方法均是贪心的或者基于贪心的，它们的共同特点是运行时间相对过长。

2010年，W.chen**[16]**证明在独立级联模型的基础上计算影响力传播是问题，由此他们提出最大影响力路径和MIA模型。他们的实验结果证明MIA模型确实可以取得较快的速度，但是由于该模型忽略了许多可能的传播路径，因此，传播过程被大大简化，与实际过程不同。当然，计算节点的度中心性、介数中心性、紧密度中心性等也都属于启发式算法的范畴。对于启发式算法而言，它们确实能够缩短运行时间，并适用于大规模网络，但是却损失了准确度。Guangzhi Zhang**[17]**认为影响力的传播与节点度数有直接关系，其运行结果和时间效率方面都很不错，但是它也基于统一独立级联模型。

2011年兰如钦**[18]**提出了CPWM算法，把节点的度数和节点指向的节点度数综合起来作为节点影响力的考虑因素,并且降低的贪心算法的复杂度。提出一个基于微博网络的消息影响力最大化算法构思，将微博用户按话题的主题划分用户，然后动态分配给每个话题的主题不同数量的初始节点数，最后通过算法找到每个话题的初始集合。解决了贪心算法时间复杂度高的问题、算法选择初始节点可能出现邻居重叠的问题。但是对于按照话题划分用户就不能够达到很好的消息影响力最大化效果。

2011年田家堂等**[19]**利用线性阈值模型的“影响力积累”特性,提出了一个该模型下影响最大化算法的框架,并在此框架基础上给出一个新的算法HPG。HPG算法相比于KK算法影响范围和时间复杂度都比较好。但是在该启发阶段并没有考虑每个节点阈值不同的情况。会出现若一个未激活节点的阈值大于已激活节点的阈值，未激活节点的潜在价值就大于已激活节点的潜在价值。显然是不太合理的。

2012年王铁彤等**[20]**利用线性阈值模型提出了一种基于节点激活阈值的启发式算法．它综合考虑了节点之间的影响力和节点的激活阈值，根据每个节点在激活过程中动态变化的阈值来计算PIN值，启发过程中，每一次都选取PIN最大的节点作为种子节点进行激活，贪心阶段中再贪心地挑选那些具有最大影响范围增量的节点作为种子节点。相比于HPG算法，该算法相对于HPG在相同的启发因子c下具有很好的激活范围以及非常低的复杂度。

2013年于淼**[21]**提出一种基于用户质量的消息影响力评估算法BiasRank。主要考虑参与消息传播的用户的质量，并在评估过程中加入去重机制、惩罚机制，综合地对消息进行评估。最后通过实验证明该方法能够有效的屏蔽虚假的广告消息和机器人用户为消息带来的不真实的消息影响力，更加准确的评估消息影响力。算法不仅考虑转发用户的质量还考虑交互图中边的质量，即用户的影响力是根据t 时间段内，用户对所有转发边的转发比例确定的。即被该转发次数高的用户分得的影响力比转发次数少的用户分得的多。用户活跃度越高其分配出的影响力也就越多。

2015年曹玖新**[22]**提出一种基于k-核的社会网络影响最大化算法，主要思想是若节点u被选为种子节点，则与u距离小于等于d(d为自定义参数)的所有节点标识为覆盖状态，被标记为覆盖状态的节点不能被选为种子节点，每轮选择核数最大或在核数相等的情况下选择度数最大且未被覆盖的节点作为种子节点。实验结果表明：在较大传播概率下CCA算法的传播范围优于其他启发式算法，且时间复杂度非常低，并随着种子节点数的增加，影响范围保持增长势头。

1.3论文的主要研究内容

我们发现针对社会网络的影响力研究已经取得了许多优秀的研究成果，但是还有一些空间值得去探讨。在影响力度量方面，虽然基于用户交互行为的方法已经存在，但是它们往往只是从行为的角度去分析问题，而忽略了交互行为与网络拓扑结构及自身的属性的关系，如何综合这些因素去分析问题值得我们进一步探索。在影响力最大化问题上，基于贪心的算法准确性较高，但是运行时间相对过长，而启发式算法虽然所需时间较短，但是准确度却往往难以令人满意。如何结合贪心法与启发式算法的优点，找到一个方法既能给出不错的准确度，又有较高的效率，也是值得我们去研究的主题。

本文的研究工作是在前人研究的基础上，针对目前影响力最大化算法中存在的一些不足提出求解影响力最大化问题的算法，以希望能够降低求解该问题的时间复杂度和提高最终被影响的节点数目。针对该研究目标，主要做以下几个方面的工作：

1.了解该课题的研究背景和意义，并详细介绍了该课题的研究现状，针对目前研究现状存在的问题展开本文的研究工作。

2.掌握相关理论知识，尤其对求解影响力最大化问题所需要的独立级联模型和线性阈值模型这两种传播模型进行详细的研究，这些相关的基础理论研究为后续的研究工作奠定了基础。

3.研究新浪微博，使用Python爬取领域下的数据，对这些真实数据进行分析。

4.研究影响力度量算法，基于节点交互行为的算法和基于K-核的MDD的算法。并详细分析了这些算法存在的不足，提出一种新型的影响力度量MIB算法。

5.研究影响力最大化算法,提出一种基于活跃度的改进的独立级联模型，并在此模型基础上研究影响力最大化SMC算法。

6..在真实网络网络上进行了实验,从得出的实验结果来验证提出算法的有效性,并对实验结果进行详细的分析。

1.4论文组织结构

本文针对微博网络，围绕微博网络中的用户影响力的度量以及消息影响力最大化算法展开研究，论文的组织结构如下：

第1章是绪论。详细的阐述了课题的背景与意义，并介绍了论文的研究内容和组织结构。

第2章是微博与社会网络相关理论。介绍了社会网络中与用户影响力评估以及用户影响力最大化相关的研究，并分析了相关研究的特点与不足。

第3章是微博社交网络构建与分析。利用数据爬虫技术爬取的特定话题下的一个月帖子作为数据集，并且由此数据集构建网络模型。最后通过可视化技术展示并分析了微博社交网络。

第4章是微博用户影响力评价算法。提出MIB算法模型，即基于用户交互行为的影响力度量模型，该算法克服以往传统算法片面性，综合考虑了节点的全局重要性与局部重要性，并且考虑了节点对其邻居节点的影响力贡献不同，用节点的MDD值和相邻节点之间的影响力两方面综合度量，更加符合实际情况，能够得到更准确的结果。

第5章是微博用户影响力最大化算法。在微博网络中引入活跃度概念，充分考虑节点的活跃程度。对独立级联模型模型进行了扩展，提出一种基于活跃度的改进的独立级联模型。并在此基础上提出影响力最大化SMC算法，并且通过实验验证了该算法合理性和有效性。

第6章是结论与展望。本章主要对之前的做的工作做了全面的总结，并指出本文中仍然存在的不足，并提出了下一步的展望。

# 2相关理论知识与技术

在开始微博中影响力的研究之前，需要先对相关理论知识与技术进行整理。本章对社交媒体、社会网络、影响力度量方法和影响力最大化等概念进行了介绍，并对影响力评价方法进行了总结。此外，还详细介绍了几种影响力传播模型以及一些代表性影响力最大化算法。

2.1社交媒体与社会网络

2.1.1社交媒体

Web在过去的十年中得到了快速的发展，同时也促使了互联网发生了巨大的变化。Web2.0技术的异军突起，使得无数互动的应用程序和社会网络站点涌现出来，社会媒体由此产生。从字面意思上看，媒体是一种传递信息的工具，如报纸、广播等，社交媒体则是一种传递信息的社会化工具。

社交媒体**[1]**种类多样，主要包括博客、论坛、微博、社会新闻、媒体共享平台等。表2-1列出了各种类型的社交媒体。虽然这些社交媒体在应用上是不同的，但是共同的特点是用户可以作为消息产生源进行传播。传统的媒体例如报纸和电视，内容和信息都是有一小部分权威的的人士作为信息源进行产生，大部分用户只是消费者，对这些信息内容只能进行阅读接收，不能参与其产生制作，只有一小部分权威人士能参与到信息的产生过程中。与传统媒体相比，社交媒体就像是一个双行道，用户既是一个产生源者又是一个接收源。在得到信息的同时，用户就可以将自己的想法表达出来，实现信息交换。，每一个用户都可以作为一个信息生产源，创建自己的内容，并通过网络实现信息共享。这种通信方式推动了实时信息的发布，有助于人们第一时间收到新的消息。

表2-1各种类型的社交媒体

|  |  |
| --- | --- |
| 论坛 | 天涯论坛、新浪论坛、百度贴吧 |
| 微博 | 新浪微博、Twitter |
| 社会新闻 | 搜狐新闻、今日头条、网易新闻 |
| 媒体共享平台 | 优酷、爱奇艺 |
| 社交网络 | LinkedIn、FaceBook |
| 博客 | Csdn、新浪博客 |

社交媒体是大批网民自发[贡献](http://baike.baidu.com/subview/1448/8031405.htm" \t "http://baike.baidu.com/_blank)，提取，创造新闻资讯，然后传播的过程。有两点需要强调，一个人数众多，一个是自发的传播，如果缺乏这两点因素的任何一点就不会构成社交媒体的范畴。社交媒体的产生依赖的是[WEB2.0](http://baike.baidu.com/view/733.htm" \t "http://baike.baidu.com/_blank)的发展，如果网络不赋予网民更多的主动权，社交媒体就失去了群众基础和技术支持，失去了根基。如果没有技术支撑那么多的互动[模式](http://baike.baidu.com/subview/37878/12538847.htm" \t "http://baike.baidu.com/_blank)，那么多互动的产品，网民的需求只能被压制无法释放。如果没有意识到网民对于互动的，表达自我的强烈[愿望](http://baike.baidu.com/subview/54473/5136452.htm" \t "http://baike.baidu.com/_blank)也不会催生那么多眼花缭乱的技术。社交媒体正是基于群众基础和技术支持才得以发展。

2.1.2社会网络

[社会网络](http://baike.baidu.com/view/1483745.htm" \t "http://baike.baidu.com/item/_blank)（social network）是一种基于“[网络](http://baike.baidu.com/view/3487.htm" \t "http://baike.baidu.com/item/_blank)”（节点之间的相互连接）而非“群体”（明确的边界和秩序）的社会组织形式，它是由许多节点和边构成的一种社会结构，其中节点表示个人或组织，而边则表示节点之间复杂的关系，社会关系包括朋友关系、同学关系、生意伙伴关系、种族信仰关系等。

社会媒体和社会网络是两个比较容易混滑的概念。社会媒体是包含博客、论坛、电子邮件、社会网络等一系列互联网服务的信息传播工具，是一种向用户传递信息、分享知识的新的方式。而社会网络则是一种服务，它为具有相同兴趣或某种联系的人们提供了交流的平台。因此，社会网络是一个相对狭义的概念，属于社会媒体的一种类型。从另一个角度来看，“社交网络（SNS）”也可以是一种媒体，因为在这个网络平台上，无数的信息被网络中的节点（人）过滤并传播着，有价值的消息会被迅速传遍全球，无价值的信息则会被人们遗忘或者只能得到小范围的传播。在各种类型的社会媒体中，人与人之间总是可以形成一种网络关系，我们习惯性的将其统称为社会网络。

社会网络一般都是大规模而且非常复杂的。这些大规模网络往往拥有一些相同的特性，而在小网络中这些特性则是不明显的，包括无尺度分布，小世界效应和强社区性等。

2.2新浪微博

新浪微博是一个由新浪网推出，提供微型博客服务类的社交网站。用户可以通过网页、WAP页面、手机客户端、手机短信、彩信发布消息或上传图片。新浪可以把微博理解为“微型博客”或者“一句话博客”。用户可以将看到的、听到的、想到的事情写成一句话，或发一张图片，通过电脑或者手机随时随地分享给朋友，一起分享、讨论；还可以关注朋友，即时看到朋友们发布的信息。用户可以像[博客](http://baike.baidu.com/view/1509.htm" \t "_blank)、聊天工具一样发布内容。用户可以把自己喜欢的内容一键转发到自己的微博（转发功能是对twitter RT功能的改良，保留原帖，避免在传播过程中被篡改）转发时还可以加上自己的评论。用户可以对自己喜欢的用户进行关注，成为这个用户的关注者（即“粉丝”）。用户可以对任何一条微博进行评论。（这是基于中国用户习惯而设置的特殊功能，之后Yahoo Meme 和Google Buzz也都有了评论功能）。用户可以用两个#号之间，插入某一话题。像这像，#某一话题XXX#，则发出的微博，可以点击这个，自动搜索微博上所有的包含有“某一话题XXX”的相关微博。可以展开讨论，实现信息的聚合。私信功能：用户可以点击私信，给新浪微博上任意的一个开放了私信端口的用户发送私信，这条私信将只被对方看到。实现私密的交流。

新浪微博的内容每条不能超过140个字符，仅两条中文短信的长度，可以三言两语，现场记录、也可以发发感慨，晒晒心情。用户可以通过互联网、客户端、手机短信彩信、WAP等多种手段，随时随地地发布信息和接受信息。用户发布一条信息，他的所有粉丝能同步看到，还可以一键转发给自己的粉丝，实现裂变传播。用户可以通过搜索找到其他微博用户在几秒前发布的信息，比传统搜索引擎的搜索结果更有时效性，更鲜活。“分享到新浪微博”的按钮被添加到了百度百科词条的下面，用户可以直接分享词条到新浪微博。

新浪微博是一种包含了多种复杂的关系和实体关联的一种社交网络，这种复杂网络是信息交流和传播的载体，以微博为代表的社交网络已经成为当前最重要的媒体之一**[22]**。由于新浪微博的用户关系丰富，目前是最受欢迎的微博类社交媒体，因此，我们以新浪微博作为我们研究的对象。

2.3影响力评价方法

复杂网络中节点影响力评价方式有很多。从网络拓扑结构入手是研究这一问题常用的方法之一。刘建国、任卓明等**[23]**人从网络的局部属性、全局属性、网络的位置以及随机游走等四个角度出发, 介绍了基于网络结构的节点重要性排序的不同指标。

2.3.1基于网络局部属性的最大度算法

基于网络的局部属性度量方法主要考虑节点自身信息和其邻居信息**[24]**，计算简单，时间复杂度低。节点的出度和入度可以衡量社交网络中与用户影响力相关的指标（如好友数、推荐数、跟帖数等），在一定程度上可表示节点的影响力大小，而它们的方向可以表示用户影响力或者信息的传播方向．度中心度则可以用来衡量节点对其邻居的平均影响力。

最大度算法的一个基本假设是：越是拥有更多和其他节点链接的节点，越是可能让更多人看到自己发布的信息。因此，一个朴素的想法是，社交网络中度数越大的节点，其影响力也就越大。定义如下：

 　　　　　　　　　　　　(2-1)

2.3.2基于网络全局属性的距离中心算法

基于网络全局属性的度量方法要考虑网络的全局信息，这些指标一般准确性相对较好，但时间复杂度高。主要有紧密中心度和介数中心度等方法。紧密中心度可用来度量当前节点对其他节点的间接影响力，或者信息从该节点传播到其他节点的距离，也可间接度量该用户的社会关系强度。该值越大，表示当前用户和其他用户之间的距离越短，该用户影响其他用户的速度越快。介数中心度衡量节点在网络结构中所处位置的重要性。该度量值越大，表示网络中信息流动时经过该节点的信息量越大，即该节点在信息传播过程中的影响力越大。

距离中心算法**[25]**的一个基础假设是：社交网络中的节点，如果可以更短的路径到达其他节点，那它应该有更好的机会来影响其他节点。考察网络中某个节点到其他节点所需的平均路径的长度，也是衡量影响力的一个方法。定义如下：

 (2-2)

2.2.3基于随机游走模型的LeaderRank算法

基于随机游走模型的度量方法主要是基于网页之间的链接关系的网页排序技术，链接关系提供了相互之间的关联和支持，从而可以判断重要程度。主要的度量方法有PageRank、LeaderRank等算法。

Google创始人Larry Page发明的PageRank算法**[26]**是一个著名的排序算法，Google 的成功很大程度上决定于网页这个排序算法。它的基本思想是：从网页A指向网页B的链接被看作是页面A对页面B的支持投票，但是网页的“重要性”不能简单地看投票数(即链接数目)，而是要看是否有重要的网页链接，重要的网页链接你那么你就变得重要。有了这个想法，很容易建立网页重要性指标的数学模型，转化为一个列随机矩阵的1特征值对应的特征向量问题，特征向量的分量就是网页的“重要性”指标。

但是原始的PageRank模型仍然存在几个问题，其中一个问题是当网络有r个不连通的子网的时候，该网络的网页排序方式至少有r种，于是出现了排序不唯一的问题。PageRank的办法是作素性修正，构造一个所有元素都为1/n的列随机矩阵，与原来的矩阵按加权（一般取0.15和0.85）叠加变成所谓的素矩阵 (primitive matrix),使得网络连通（解决了排序不唯一），并且仍然保持为列随机矩阵（保证了1特征值的存在）。这显然解决了排序不唯一的问题，但是新的矩阵全连接了，但是原来的结构也遭到很大的变形。

[吕琳媛](http://bbs.sciencenet.cn/home.php?mod=space&uid=329471" \t "__blank)等**[8]**针对这个问题提出LeaderRank，在已有节点外另加一个节点(ground node)，并且将它与已有的所有节点双向连接，于是得到N+1个节点的网络，这个新的网络是一个强连通的网络，再按照原始的PageRank算法计算得到原来N个节点的“重要性”排序。于是由于加入ground node后的图是一个连通图，排序的唯一性就解决了。结果表明LeaderRank比PageRank算法排序更精准，并且对网络噪音有更好的容忍性。

2.2.4基于网络位置的的K-核混合分解法（MDD）

在2010年《Nature》物理版上，Kitsak等人**[7]**首次提出了节点重要性依赖于其在整个网络中的位置的思想, 并且利用 K-核分解获得了节点重要性排序指标 (k-shell),该指标时间复杂度低, 适用于大型网络, 而且比度、介数更能准确识别在疾病传播中最有影响力的节点。近几年不少学者受到这种思想的启发, 对 K-核进行了扩展和改进, 使其应用范围更广, 准确性更好，并且通过实验表明在影响力传播方面，核数比度数和介数等节点属性具有更稳定的传播力，并且提出了基于覆盖的最大核算法**[27]**和最大度算法**[28]**。

通过SIR模型和SIS模型的建模分析指出，对于单个传播源情形，Hubs节点或者高介数的节点不一定是最有影响力的节点，而通过K-shell分解分析确定的网络核心节点（即K-shell值大的节点）才是最有影响力的节点。

K-shell是图论里的一个经典的概念**[29]**，网络的外壳和边缘的K-shell为1，然后往内像剥洋葱一样进入网络的核心（K-shell值大的区域）。之所以产生Hubs节点或者高介数的节点不一定最具有影响力的原因在于，它们如果在整个网络的边缘，那么它在传播中的作用就很微弱了。而某些度数虽然比较小却位于网络核心的节点将对传播过程产生重大的影响。以CNI的网络为例说明，一个传播过程中感染的规模并不一定与传播起源节点的度k有关，即使从度数相近的不同Hub节点开始，感染结果也可能会非常不同。相反，从相同的Ks层节点开始，即使有不同的度K，传播规模却很相近，说明初始传播节点的Ks值可以准确地预测感染的规模。同时Ks值对于不同的传播率显得相当健壮。这也说明节点所处网络的位置比一个节点的局部属性度K更能决定它的影响力。同样地，高介数的节点和Hub节点一样，也远不如Ks值更能准确地预测感染的规模。

当然文章也指出，当初始存在多个传播源的时候，传播的规模很大程度依赖于初始传播源之间的距离。尽管高Ks点是最好的单一传播者，但是在多个传播源情况，度大的Hub节点往往比Ks大的节点具有更高传播效率。这是因为传播存在交叉感染现象（感染者传染给感染者），Ks大的节点往往在网络的核心，它们聚集在一起，而度大的Hub节点可以分散在网络的不同区域。因此，如果考虑多个传播源有更大的传播规模的话，应该选择处于不同shell（层）、不直接相连的较大的 k的节点作为初始传播源。这篇文章对于我们认识传播网络中到底哪些节点最有影响力，评价网络中节点的重要性，以及从事网络控制、信息传播、疫苗具有重要意义。

Ks的一个缺陷是赋予了大量节点相同的值，导致在一些较为密集的网络中无法衡量其节点的重要性。

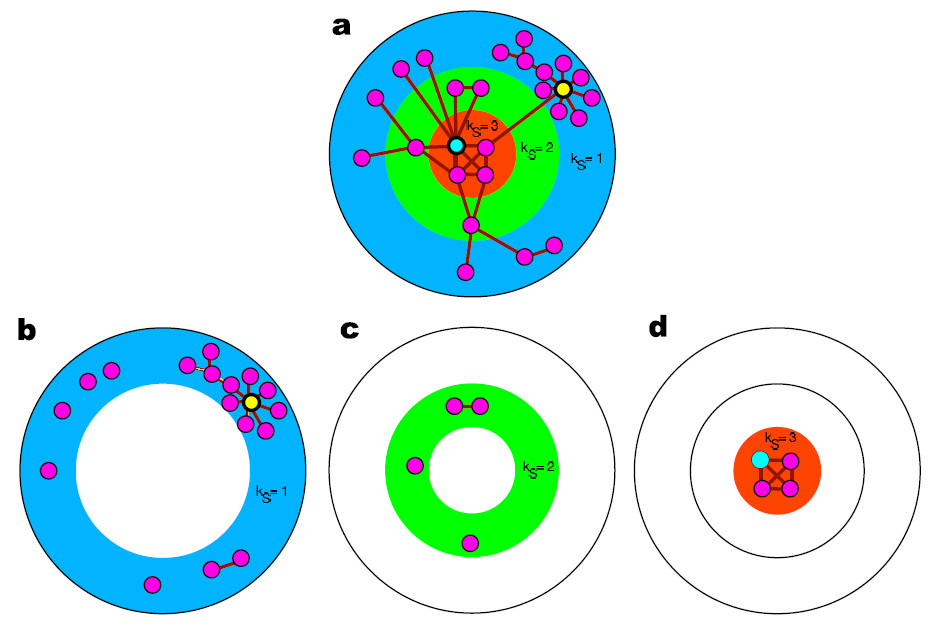


图2-1 Ks计算示意图

针对这个问题，An Zeng**[9]**在[《Physics Letters A](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=journaluri%3A%28fa5516ef2180f2c2%29%20%E3%80%8APhysics%20Letters%20A%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited" \t "_blank" \o "《Physics Letters A》)》上发表的一篇文章中提出了 MDD（Mixed degree decomposition method），这是一种混合K-核分解法。基本思路是不仅考虑移除后剩余节点的信息，而且考虑被移除节点的度的信息，解决了在某个ks值范围内有大量相同ks值的情况。下面是MDD的度量方式：

k(m) =k(r) + λ ∗ k(e) (2-3)

λ是一个可调节的参数，值在0-1之间。详细的分解步骤如下：

1. 初始化，在网络中没有节点移除时,每个节点的k(m)等于k(r)。

2. 移除最小的k(m)值的节点（标记为M），并把他们标记为M-shell.

3. 通过公式 k(m) =k(r) +λ ∗ k(e) ，更新所有剩余节点的k(m)。然后，移除所有节点小于等于M并标记为M-shell。重复地执行直到所有节点的k(m)值大于M。

4.重复2 and 3，直到网络中的节点的值都被分配shell值。

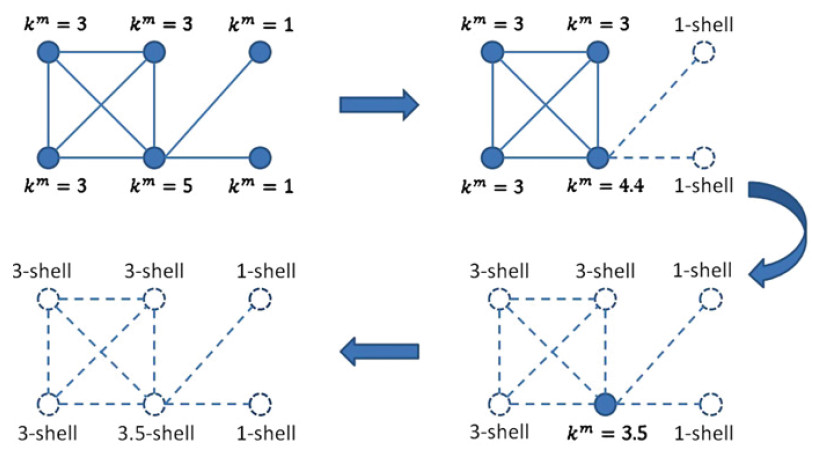


图2-2 MDD的分解示例

图中虚线表示当前步骤需要移除的节点。MDD的参数λ设置为0*.*7. 在每步中，k(m)值被更新。

2.4影响力传播模型

### 2.4.1两种经典的传播动力学模型

复杂网络上传播研究的对象极广, 比如通信网络中的病毒传播、社会网络中的信息传播、经济网络中的危机扩散等. 在评价各种节点重要性挖掘方法时广泛采用的是传染病模型, 主要包括SIS 模型和SIR模型**[30]**。

经典传播模型有SIR模型和SIS模型，很多研究学者在经典的传播模型的基础上，提出了多种传播模型，如基于社交网络的传播模型SEIR，谣言传播模型 HKASI 等。

在经典传染病模型中，个体状态主要分为如下几类：易染状态 S（Susceptible），感染状态 I （Infected）和移除状态 R （Removed 或 Recovered）。初始时，设定网络有一个或几个个体处于感染状态，其他个体为易染状态，处于感染状态的个体以一定概率感染处于易染状态的邻居个体。

SIR传播模型中，网络中的个体有 3 种状态，即 S ，I ，R 三种状态，与 SI 模型一样，S状态的个体接触 I 状态的个体后，以概率β变成 I 状态，不同的是，I 状态的个体以概率γ变成 R 状态，处于 R 状态不能被感染或继续传染。记 t 时刻， s(t)，i(t) 和 r(t) 为网络中 3 种状态的个体数所占比例，则 SIR 模型方程如下所示：

 (2-4)

SIR 传播模型有两种传播模式：单点接触和全点接触。前者认为一个时间步内，状态为 I 的个体随机选择一个邻居节点中状态为 S的个体，以概率β将其感染为 I 状态，后者是选择其邻居节点中所有S状态的个体，这些个体都以概率β变为I状态。

SIS 传播模型中，网络中的个体有 2 种状态，即 S， I 两种状态，SIR 模型中， I 状态的个体以概率γ变成 R状态，而 SIS 模型中，I 状态的个体以概率γ变成 S 状态。记 t 时刻，s(t)和i(t)为网络中 2 种状态的个体数所占比例，则 SIS模型方程如下所示：

 (2-5)

### 2.4.2独立级联模型和线性阈值模型

独立级联模型（Independent Cascade Model）**[31]**，是基于相互粒子系统（Interacting Particle System）设计的一个信息扩散的模型，这是一个概率模型。在其传播机制中关键点是把一个节点对另一个节点的激活用概率p表示，激活失败的概率是1-p，概率越大表示激活成功的可能性越大。此外该模型做了这样的假设：即当节点试图激活其邻接节点时，一旦没有激活成功，那么它将永远失去再次激活节点的机会。

独立级联模型的影响力传播具有次模特性**[32]**。次模函数的定义如下：

当任意一个函数满足以下条件是被称为次模函数：

* 函数f将有限集合S的子集映射到一个非负实数集U
* 函数具有非严格单调递增的属性，即添加一个元素到集合里面，不

致f减小：

* 定义N为一个有限集合，
* 
* 

第三条也称为自然收益递减属性：添加一个元素到集合所得到的边际收益至少会和添加相同的元素到集合是的父集所获得的边际收益一样多。当前社会网络影响最大化算法大都基于独立级联模型的次模函数性质进行研究，这是因为对于独立级联模型的任意一个实例，影响传播函数都是次模函数，其影响传播具存次模特性，而线性阈值模型却没有良好的次模特性。

线性阈值模型**[33]**是所有基于节点特异性阈值模型的核心。模型的扩散过程如下：

给定初始传播集合，所有节点的特异性阈值和节点之间的影响力，在第t步扩散时，将t-1时刻满足激活条件的节点加入到t时刻的集合中去。重复这一过程，知道不再有新的节点加入进来。

针对线性阈值模型的扩散过程进行研究，此模型具有一个很好的“影响积累”特性。在线性阈值模型中，当一个激活节点尝试去激活它的未激活邻居而没有成功时，节点对节点的影响力被“积累”下来，而不是被抛弃。这种影响力的积累对后面节点的其它邻居对的激活是有贡献的，直到节点被激活或传播过程结束。

2.5影响力最大化算法

### 2.5.1贪心爬山算法

Kempe等**[12]**将影响力最大化问题定义为离散优化问题，并证明其是NP难解的。与此同时，他们还提出一个贪心算法，该算法能够取得最优解63%的近似度。贪心算法的基本思想是现将所有节点标记为未激活，然后在这些节点中每次选取一个能激活节点数最多的节点，标记该节点为激活，并加入到最终的集合S中去，然后在剩下的未激活节点中在进行上一步操作，知道选出所要求的k个节点为止。将蒙特卡洛模拟次数设置为1000，就可以得到相当准确的近似值。虽然原始贪心算法可以给出的近似度，但是它的运行时间过长，无法应用在大规模网络中。

在影响力最大化问题中，我们定义σ(S)为影响力传播函数，它表示信息从初始活跃节点集合S开始扩散，最终能够影响的节点数。可以看到，该函数首先是非负的，因为能够影响的节点个数不可能小于零。其次，该函数满足单调性。当加入任意一个新的节点v到集合S中时，σ(S)值不会减小。最后，Kempe证明，无论是在独立级联模型还是线性阈值模型中，影响力传播函数都具有子模特性。σ函数满足次模函数的三个特性，唯一一点区别在于次模函数的值是精确的，而σ值是未知的。为解决这个问题，我们使用蒙特卡洛模拟来对值进行近似估计。通过模拟信息传播过程，我们可以得到相当准确的近似值。正因为如此，借鉴于贪心爬山法，Kempe等人给出了原始贪心算法如下所示，该算法的时间复杂度为O(knRm)。

将社会网络图表示为G=(V,E)，种子节点数k，影响力最大化的种子节点S，蒙特卡洛模拟次数R。

|  |
| --- |
| Input:G=(V,E),k |
| Output:S |
| Initialize:S={},R=1000 |
| 1 for i=0 to k do  2 for each u∈V\S  3 for j=0 to R  4 calculate σ(S∪u) -σ(S)  5 end for  6 calculate avg(σ(S∪u) -σ(S))  7 end for  8 select u which make max(σ(S∪u) -σ(s))  9  10 end for  12 output S |

### 2.5.2启发式算法

基于节点度数或节点中心性等**[34]**是社会网络分析中常见的启发式方法。在社会网络和其他网络中，以度数选择个最大度数节点的启发式节点选择策略，是长期以来的一个标准方法，因为度数越大说明这个节点和其他节点之间的关联就越多，且节点的度数很容易得知，在社会科学中被称为“度中心性”。

社会网络图表示为G=(V,E)，种子节点数k，影响力最大化的种子节点S。

|  |
| --- |
| Input:G=(V,E),k |
| Output:S |
| Initialize:S={} |
| 1 for each u∈V  2 calculate  3 end for  4 for i=1 to k  5 select u which make  6  7 end for  8 end for  9 output S |

在某些领域这些简单的启发式方法效果很好。但是由于每次是静态选取最大度数的节点，没有考虑影响的扩散过程，并不代表这些节点能达到最多的影响范围。因此这些启发式的方法在本文所要解决的问题上效果并不好,甚至比较差**[35]**。

### 2.5.3其他相关算法

基于用户的交互行为算法**[36]**主要是根据节点之间的交互关系，结合贪心算法，将节点的度量应用于最大化问题的计算中去。周东浩等**[37]**提出一种基于信息偏好的2阶段启发式影响结点挖掘策略L\_GAUP：第1阶段，基于网络中各结点对于信息主题的偏好程度，得到易感染结点网络；第2阶段，在易感染网络中，基于贪心策略进行影响结点的挖掘核心思想：

基于社区的影响力最大化算法主要是利用现有的社区发现算法**[38]**,将网络分成若干个社区,然后在每个独立的社区上应用已有的影响力最大化算法求解,其中OASNET算法和CGA算法是目前两种应用社区发现算法来求解影响力最大化问题的算法。

OASNET算法**[39]**是等作者利用网络的社区性质把影响力最大化问题看成一种最佳的资源动态分配问题而提出的一种算法。作者利用社区发现算法把网络划分为各个独立的社区,然后利用动态规划的方法把初始的节点最佳的分配到各个社区,最后将各个社区中被影响的节点累加得到最终被影响的节点数目。

CGA算法**[15]**的总体思想和OASNET算法类似,也是根据社区发现算法将网络分成各个独立的社区,然后把问题转化为将初始激活节点个数最佳的分配到各个社区,最后将各个社区影响力值求和得到最佳的影响力值。唯一不同的是CGA算法使用贪心算法作为各个社区寻找初始的影响力节点的算法,而OASNET算法是使用度数最大的算法。由于贪心算法的时间复杂度比度数最大算法的时间复杂度高很多,因此CGA算法的时间复杂度也比OASNET算法高很多。CGA算法由于将网络划分成各个小的网络,因此它相比贪心算法时间复杂度有很大数量级的降低。但是CGA算法使用贪心算法作为社区的寻找算法,时间复杂度依然比较高,而且由于网络分成各个独立社区之后会导致边的减少而影响效果。

2.6本章小结

通过以上对当前国内外研究现状的总结，我们发现针对社会网络的影响力研究已经取得了许多优秀的研究成果，但是还有一些空间值得去探讨。在影响力度量方面，MDD算法已经取得非常好的效果，但是如何应用于社交网络特别是微博网络中，值得我们进一步去探索。在影响力最大化问题上，基于贪心的算法准确性较高，但是运行时间相对过长，而启发式算法虽然所需时间较短，但是准确度却往往难以令人满意。如何结合贪心法与启发式算法的优点，找到一个方法既能给出不错的准确度，又有较高的效率，也是值得我们去研究的主题。

# 3微博社交网络构建与分析

前一章介绍了社交网络分析以及微博中的影响力的一些基本理论，本章开始本文的主要内容。本文的核心工作是在微博网络中研究影响力最大化问题，即找出一个用户集合，从这个集合出发，按照我们的传播模型，可以传播到到最多的用户。因此，微博数据是本文研究工作的数据基础。目前，影响力最大化在一些数据集相对容易获取 的社会网络中的研究比较多，但是在微博网络中的研究较少，当前还没有比较规范的、成熟的微博数据集可供研究。所以，构建合理的微博社交网络是本文在进行影响力最大化研究的前提。由于微博数据的数量比较巨大，所以采取手动统计的方法不太现实。本章的主要研究内容是基于网络爬虫的微博社交网络的构建及对数据的的可视化分析。

3.1微博数据源分析

新浪微博作为近几年最为活跃的社交网络平台，不仅其内容几乎涵盖了所有的方面，而且在新浪微博中用户存在各种各样的关系和交互行为。由这些关系形成的微博社交网络存在着巨大的潜在价值，对于这些关系本身的分析和在此基础上进行我们影响力传播模型的构建就非常有意义。由于缺乏相对合理的数据集，所以对这些关系的获取是我们需要做的工作。和这些关系相关的数据集是进行微博社交网络影响力最大化研究的前提，而建立微博社交网络的关键点在于微博节点的关系以及自身属性的处理与挖掘。微博中用户的关系是人们互相交流的主要途径，也是用户表达自己、体现自己的社交关系的渠道之一。很多用户都比较重视自己的社交关系，在每个用户的主页中，有很多用户的信息进行展示。比较活跃的用户会把自己吃穿住行展示上去，以吸引更多的用户关注，并且在用户主页中，还包含用户产生的历史行为信息。另外，对于其他的用户，在进行访问以及一些行为也会留下有价值的信息，比如回复和转发等都会作为媒介传递重要的信息。用户可以通过其他人发表的微博来发现和获取信息，并且通过转发时更多的人获取到信息，并且可以找到特定的对象进行信息传递，微博中信息的获取和传递依赖于这些关系。对于这些类型的网络数据，往往是非常重要的，通过网络爬虫技术检索并获取，进行数据预处理后，可以作为构建微博社交网络的基础数据源。

微博数据具有较强的多样性和即时性，由于范围比较广，获取存在一定的难度，本文将数据源定位到新浪微博中的一个特定的话题下。在这个特定的话题中，我们需要取得的数据往往包含如下几个方面。首先，用户发表的微博信息数据，用户的微博数据往往是反应用户动态的最直接的体现。其次，是用户的关系数据。包括了用户的一些关系，如评论、转发等等，这是体现用户关系，展现出用户在社交网络中的影响范围和地位。用户的关系数据是构成网络的较为重要的数据，并且是进行影响力最大化传播模拟的基础。再次，用户信息数据也作为一个重要的方面。用户的信息使用户个人在网络中产生行为的直接体现，通过观察用户的信息，能给出对用户相对直观的判定，也会带来有价值的信息。在这些数据中，包含了大量的微博数据的基本信息。综上，微博数据源包括以上这三种，每种数据的特点如下：

3.1.1 微博信息数据

微博信息是用户所发表的内容，信息的发布是自主进行的，并且是实时的，可以随时随地进行发表微博。微博内容的特点是内容短小精悍。[微博](https://www.baidu.com/s?wd=%E5%BE%AE%E5%8D%9A&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1YkuhmYmHnYmHw-PWbznhn40ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPHckP1m1nHbd" \t "https://zhidao.baidu.com/question/_blank)的内容限定为140字左右，内容简短，不需长篇大论，门槛较低，可以插入图片，并且可以加“@”来通知特定的用户来阅读。微博内容的模块基本都相似，主要包括了昵称、发布时间、微博内容，转发数量、评论数量和赞数等。这里以微博昵称为“四月没有钱”的这位用户发表的一条微博为例，图3-1是一条微博内容的信息。可以观察到微博信息数据包含的上述这些信息，明确的可以看到微博的具体内容以及发布的时间等一些有价值的信息。我们将这些有价值的信息进行爬取，作为微博信息数据，

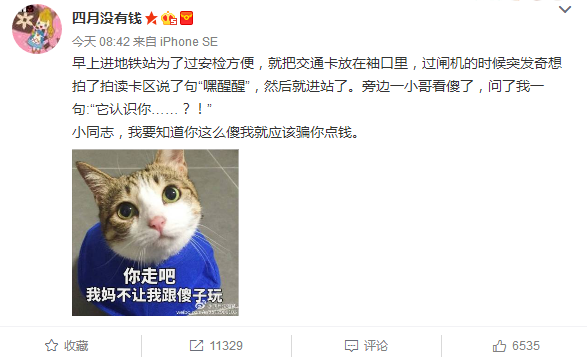


图 3-1 微博信息数据

3.1.2 用户关系数据

用户关系数据是用户在微博网络中的关系形成的数据集合。这种关系有多种，比如关注关系、评论关系和转发关系等。用户之间的关注关系是单向的，是用户的自主意愿。如果一个用户对另一个用户感兴趣，就可以通过关注这个用户，即成为这个用户的粉丝。关注关系是有向无权的，表明的一个用户受到的在乎程度，直接体现出这个用户的价值。而评论和转发关系是针对用户发表的某条微博而言，其他用户可以根据自己的意见，对这个用户的这条微博进行转发或评论。这种关系也是针对此用户的，也能从间接反映出该用户的重要程度。下面以微博昵称为“四月没有钱”这位用户的关系来说明。由图3-2我们可以看到用户的关注关系中直接可以看到其粉丝和关注人，并且可以知道这些用户的一些基本的个人信息。我们将此用户的粉丝和关注人的这样的关系爬取，作为用户间的关注关系。



图3-2 用户关注关系

图3-3是用户的转发和评论关系，是单向有权的。与关注关系不同的是，转发和评论关系不像关注关系是无权的，即只能关注或者不关注，转发和评论关系是可以进行多次评论和转发的，并且可以通过评论和转发的内容，直观的进行一些有价值的交互。通过对用户转发和评论关系图的观察，我们可以获得此用户的评论和转发的用户，以及评论和转发的内容等有价值的信息，并且可以得到转发的次数和赞数以及评论的赞数。将这种关系和在这种关系下产生的内容进行获取，分别作为用户之间的转发关系和评论关系。



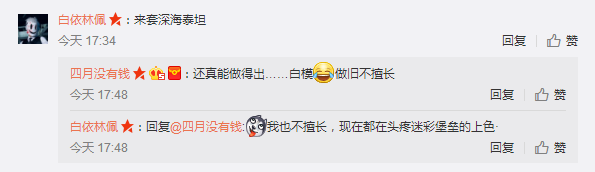


图3-3 用户转发和评论关系

3.1.3 用户信息数据

用户信息数据是用户个人信息的最直接体现，也是我们获取用户价值的比较重要的来源。通过用户的信息，我们可以直观的了解到在用户在网络中的一些基本信息，能给我们提供比较重要的数据。图3-4是“四月没有钱”这位用户的个人信息，通过观察，用户的关注数、粉丝数、微博数等这些信息，是用户在网络中的行为总体的反映，这些信息体现出用户在全局的重要性，可以获取到用户这些基本信息，作为用户的全局属性度量的方面。



图3-4 用户信息数据

3.2微博数据获取

在上小节中通过对新浪微博网络中的数据分析，确定了微博数据源。通过对这些数据源的详细分析，明确了需要的数据。在此基础上，本小节开始研究微博网络数据的获取。

3.2.1微博数据存储

微博数据的获取利用网络爬虫技术实现的，在确定微博网络的数据源之后，本文使用编写了用于获取新浪微博数据的网络爬虫程序。

微博数据的存储需要数据库，而数据库表的设计是我们进行数据存储的关键。根据数据库设计和命名规范，我们的微博数据库表的设计如下表，包括了用户信息表、微博信息表、关注关系表、评论关系表、转发关系表：

表3-1 微博信息表

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 备注 |
| mid | varchar | 微博id |
| attitudesCount | int | 赞数量 |
| commentsCount | int | 评论数量 |
| createdTime | varchar | 发布时间 |
| repostsCount | int | 转发数量 |
| weibotext | longtext | 微博内容 |
| uid | varchar | 用户id |
| state | varchar | 状态 |

表3-2 用户信息表

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 备注 |
| id | varchar | 微博id |
| attnum | int | 关注数 |
| fansNum | int | 粉丝数 |
| mblogNum | int | 微博数 |
| name | varchar | 昵称 |
| zanNum | int | 赞数 |
| comment | int | 评论数 |
| state | varchar | 状态 |
| repost | int | 转发数 |

表3-3 关注关系表

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 备注 |
| id | varchar | 关注id |
| fromId | varchar | 关注用户id |
| toId | varchar | 被关注用户id |

表3-4 转发关系表

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 备注 |
| id | varchar | 转发id |
| content | int | 转发内容 |
| mid | int | 微博id |
| uid | varchar | 用户id |
| muid | int | 转发人id |
| state | varchar | 状态 |

表3-5 评论关系表

|  |  |  |
| --- | --- | --- |
| 字段 | 数据类型 | 备注 |
| id | varchar | 评论id |
| content | int | 评论内容 |
| mid | int | 微博id |
| uid | varchar | 用户id |
| muid | int | 评论人id |
| state | varchar | 状态 |

3.2.2基于Cookie的微博数据半自动爬虫

我们使用python语言开发了一个半自动的微博数据爬取程序。Python语言，是一种面向对象、解释型计算机程序设计语言，语法简洁而清晰，具有丰富和强大的类库。它常被昵称为胶水语言，它能够很轻松的把用其他语言制作的各种模块（C/C++、Java等）轻松地联结在一起。在获取cookie方面，我们选取Fiddler为调试工具。Fiddler是一个[http协议](http://baike.baidu.com/view/70545.htm" \t "http://baike.baidu.com/item/_blank)调试代理工具，它能够记录并检查所有你的电脑和互联网之间的http通讯，设置断点，查看所有经过浏览器进行的HTTP请求的数据（例如[cookie](http://baike.baidu.com/subview/835/5062332.htm" \t "http://baike.baidu.com/item/_blank)，html，js，css等）。 数据存储我们使用Mysql数据库进行存储。表3-6列出了使用的相关软件工具。

表 3-6 微博爬虫相关软件

|  |  |
| --- | --- |
| 名称 | 软件 |
| 调试工具 | Fiddler |
| 语言 | Python |
| 数据库 | MySQL5.5 |

Cookie是在 HTTP协议下，服务器或脚本可以维护客户工作站上信息的一种方式。Cookie可以包含有关用户的信息。当我们登录网站时，浏览器将检查你是否就该唯一的服务器的预先定义的首选项（cookie）。如果有的话，浏览器将此 cookie 随你对网页的请求一起发送给服务器。我们可以将这个Cookie记录的信息保存下来，利用它来进行登录。因此，我们可以先手动进行新浪微博用户登录，登录成功后，我们使用Fiddler工具获取我们本次登录的Cookie，取得Cookies后，再使用Python利用获取到的Cookie进行请求登录，就相当于模拟了一次用户的登录，跳过了验证码等的验证信息。

根据本文研究的内容，主要爬取从指定版块出发，按广度优先策略，逐层深入，爬取规定时间内的所有微博的信息，通过这些微博信息，爬去相应的用户信息，与此同时，再针对微博中评论与转发中的信息，爬取所属用户的信息，并且抽取评论和转发关系。技术上是手动获取Cookie，利用Python模拟登陆，获取信息，抓取到相应的Html后，用正则表达式解析提取相关的信息。对于新浪微博，因为存在中反爬虫机制，对于一个账号的多次请求有次数限制，为此我们进行了一些处理，完成了该爬虫的编写。新浪微博爬取按照以下的几个步骤：

(1)爬取微博数据源页面

由于微博内容种类丰富，数量过于庞大，因此我们选定特定的话题“#西安身边事#”，如图3-5。将这个话题下的微博进行获取。在这个主题下讨论了大量的西安相关的新闻，讨论量和用户比较多。我们选取2016年7月1号到2016年8月1号这一个月的数据作为我们的微博数据源。首先登陆账号，利用Fiddler工具获取到此用户的Cookie。我们使用Python的requests库中的get方法模拟HTTP请求，将获取到的Cookie、需要访问的url以及一些相关的数据作为参数，方法get方法中模拟浏览器进行请求。微博信息的地址为“http://m.weibo.cn/p/index”,我们需要手动登录，用Fiddler工具获取到登录的Cookie和用户的uid，设置超时时间timeout=3。获取到返回的信息后，需要解析返回的页面，我们利用正则表达式获取微博内容的相关信息，由于微博信息的格式是相同的，所以我们正则表达式只需根据一个页面的信息格式进行提取就足以。通过分析HTML，针对微博内容的正则表达式如下：

r'<span class="time">(.+?)</span>.+?class="weibo-text".+?/a>(.+?)</div>.+?<span node-type="zfCount">(.+?)</span>.+?<span node-type="cmtCount">(.+?)</span>.+?<i node-type="zanCount">(.+?)</i>.+?"mid":"(.+?)"}'

我们提取了发布之间、微博内容、赞数、评论数、转发数和微博id这几个维度的信息作为微博信息内容。最后再进行判断数据是否在我们的规定的时间段，如果合理，则将数据存入到数据库中。



图3-5 “西安身边事”主题

(2)爬取微博用户页面

在获取到微博之后，接下来便是取得这条微博的用户的信息。获取信息分为两个步骤，一是根据上一步，使用正则表达式提取到用户的昵称和用户的id，二是通过这个用户id，访问用户的主页，以便获取用户信息数据。个人主页的访问格式为“http://m.weibo.cn/u/”+用户的id，我们通过这个地址就可以得到用户的个人信息数据。对于获得的用户信息数据，通过与现有的数据进行对比，将目前不存在数据存入数据库，已经存在的则不用处理。

(3)抽取微博用户关系

对于这条微博存在评论、转发和关注的关系，需要分别进行抽取。先通过正则表达式提取到评论的链接，然后通过一次request请求进行访问，解析返回的页面，提取评论的用户以及评论的内容，将这种关系以评论人+用户+评论内容这种格式进行抽取。其次在用同样的方法抽取转发关系。对于关注关系，我们先获取用户的粉丝列表与关注列表，对于粉丝列表，采取粉丝+用户进行关系的抽取；对于关注列表，采取用户+关注进行关系的抽取，最后存入数据库。

(4)爬取新的微博数据

将用户的信息和关系抽取完成后，我们进行再获取下一条微博的信息，重复步骤2到步骤4，可以实现多次循环的爬取。通过这个方式，微博网络数据库中的节点和关系不断增多。

爬取系统分为数据抓取层和存储层，如图3-6。数据抓取层，主要利用 cookies模拟登陆从新浪微博抓取需要的数据，并用用正则表达式执行过滤操作；异常处理层，则对在数据获取过程中发生网络错误、数据解析错误、编码错误等情况进行处理，保证系统可以无监督不间断运行；存储层，则是将抓取到的数据组织后存储到数据库中。

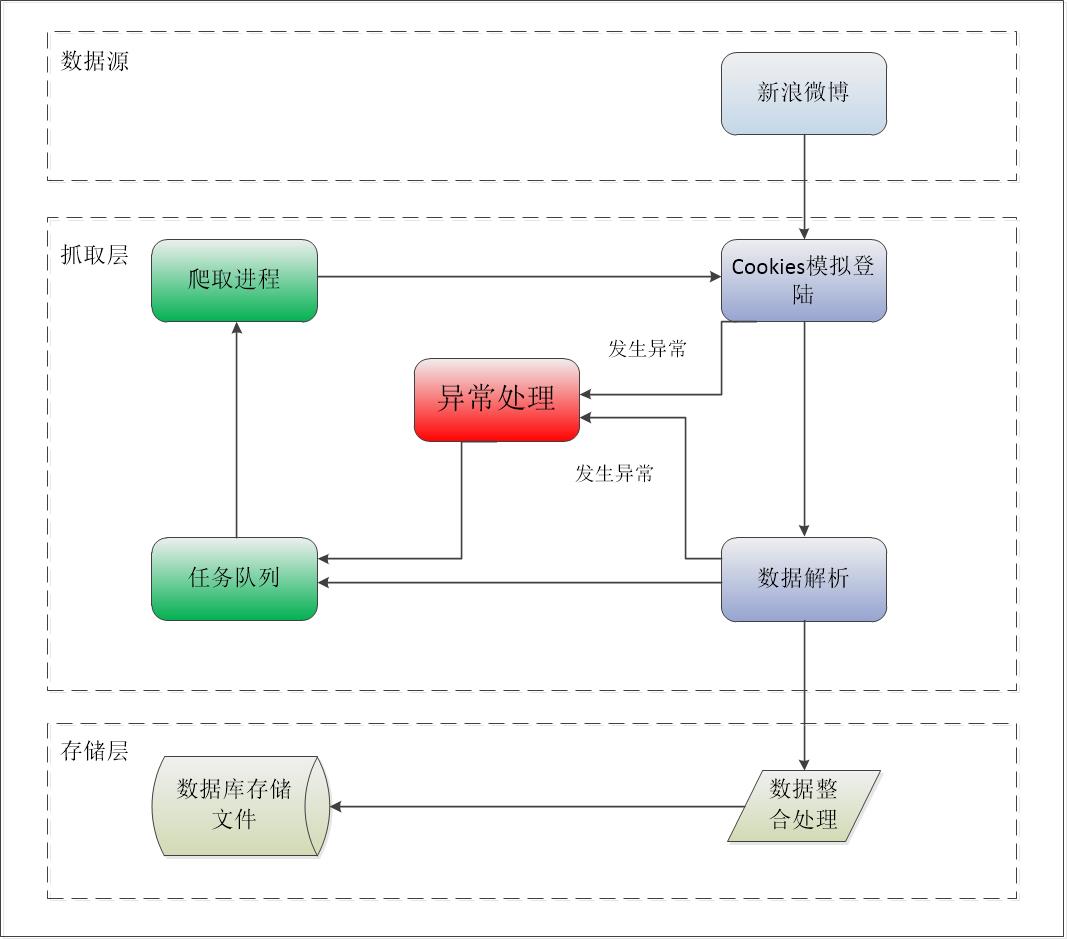


图 3-6 数据主抓取系统框架图

3.2.2关键技术点

在数据抓取层，有几点需要注意。首先，新浪微博对每个账号的请求频率限制为1000 次/小时，超过1000次/每小时的请求将返回403，服务器拒绝请求。我们采取多账号的方式绕过此限制，此次系统中，共使用20个帐号，4个进程同时抓取的策略，以绕过新浪微博对请求频率的限制。同时，在实际的系统测试中，我们发现，除超限外，过快的请求频率，依然会被新浪微博封锁掉，以防止这种情况的发生，我们在系统中加入了异常处理层，对系统执行过程中出现的异常进行针对性处理，并记录异常发生的时间和发生异常的用户，同时将该用户加入二次抓取队列，以等待再次尝试抓取，二次抓取失败则放弃该用户。

3.3微博网络的构建

微博网络的构建，实际上就是利用爬虫爬取的这些节点和关系，进行建立的过程。微博数据的获取从种子数据开始，通过爬取微博信息、用户信息，再抽取用户的关系形成网络链接关系。不断重复该过程，微博节点和关系链接会越来越多，意味着网络不断的建立。微博社交网络的构建分为三个步骤：第一步，确定数据源和获取用户的数据；第第二步，定义和挖掘微博社交网络中的节点之间的边；第三步，丰富节点和边相关的信息。其中第二步是构建社交网络的核心，因为节点和边是微博社交网路的核心组成部分，这部分是下面所有工作的基础。第三步通过丰富节点和边的信息，使微博社交网络的数据结构更加准确与合理，为下一步的工作提高准确度。

3.3.1 微博网络基本属性与关系

1)微博网络节点

微博社交网络中的节点就是用户本身，即社交网络活动中的参与者。每个用户拥有自己的昵称和id，都是唯一且不可重复。在新浪微博的社交网络结构中，节点表示已经注册的用户账号。节点可以是普通的个人，也可以是明星或者某一机构。节点是研究微博社交网络的基本单位。

我们需要从页面上识别出用户的昵称并且爬取节点的信息。本文中利用正则表达式解析网页中用户的昵称，并且根据网页链接，得到用户的id，用户的id是不能直接在网页中显示的，需要我们获取。用户节点的信息包括了id、昵称、关注数、粉丝数、微博数这几种重要的信息，还包括了所在地、性别、生日等这些基本信息。

由于所在地性别等这些基本信息与用户在网络中的关系不大，所以我们只提取重要相关的方面作为我们构造节点的属性。

2)微博网络关系的定义和抽取

微博网络中的关系可以是两个用户之间的关注关系，也可以是转发关系和评论关系。在新浪微博中，当用户A关注了用户B之后，变成为B的粉丝，用户A可以看到B的发布的动态。他们之间的关系是单向的，即用户A关注的用户B，用户B不一定关注了用户A，也就是说，把新浪微博社交网络看成是一个图的话，这个图就是一个有向无权图。我们在也页面中获取用户的粉丝和关注列表，采取抽取微博的关注关系，抽取的规则定义如下：

“当前用户A”+ “关注人B”或“粉丝A”+ “当前用户B”

其中，前一个用户关注了后面的用户，即用户A是用户B的粉丝。

3.3.2 微博社交网络的构建

微博网络是一个有向网络，这是由于微博中存在单向的关注关系。在微博网络中，节点代表微博用户，而边代表了关注关系。我们定义来表示微博网络，V表示节点的集合，E表示有向边的集合。此外，每一个用户都有自己的个人信息，比如发布的微博数、转发的微博数、评论的微博数等等。与此同时，用户之间包含了很多交互行为，比如转发、评论等。对于一条特定的有向边e，表示一个用户关注了另一个用户。

微博社交网络的构建过程如图3-7：



图3-7 微博社交网络的构建

本文获取2016年7月份在“西安身边事”一个月的发布的4472篇微博作为种子数据，用于微博网络数据的获取。部分数据如图3-8所示；

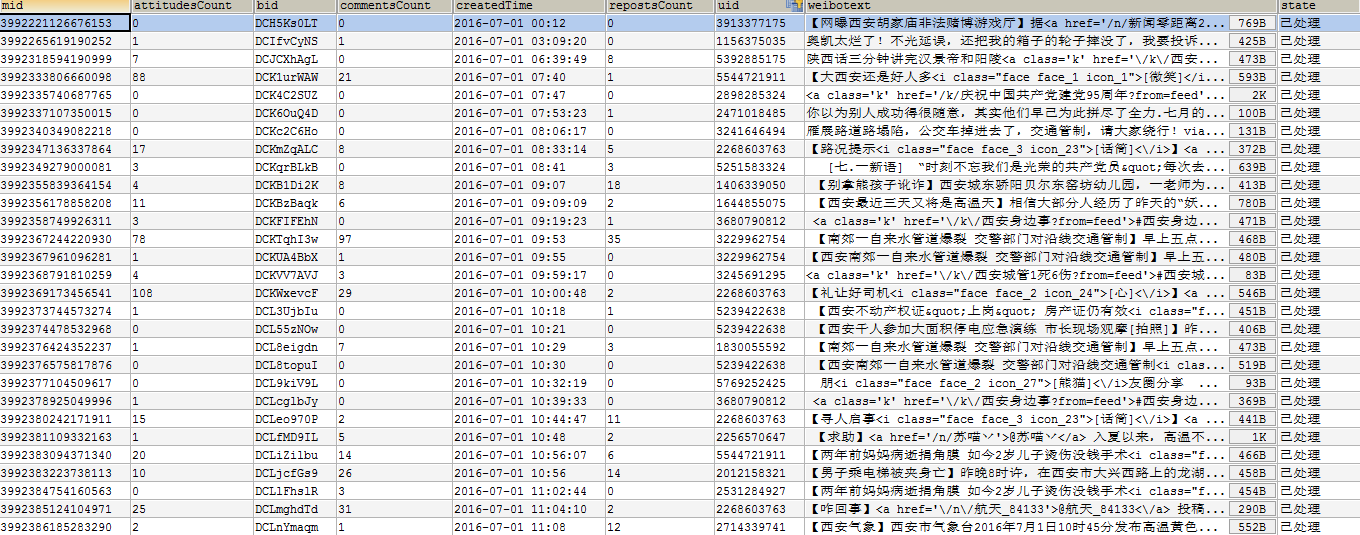
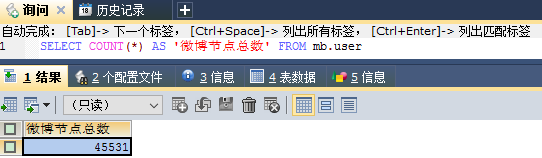


图3-8 部分微博数据

利用上述微博数据，并且使用微博社交网络的构建过程，本文构建的微博社交网络的用户节点和关系的数量如图3-9所示：



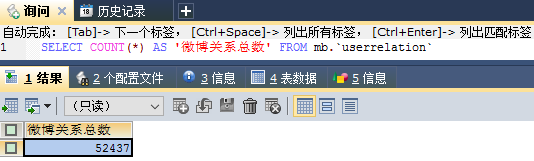


图3-9 用户节点和关系总数

3.4微博网络数据可视化分析

上小节通过爬虫技术获取到了微博数据，并由微博网络中的节点和关系建立了微博社交网络，为后续的影响力及最大化算法打下了基础。在前面微博网络关系分析和网络构建的基础上，本小节通过数据可视化将微博社交网络的构建结果用可视化技术展示，从更加直观的可视化结果中获取更多微博社交网络相关的信息，探索微博社交网络形成的主要因素和微博社交网络的特征，将这些相关的信息运用到下一步的微博中影响力度量和影响力最大化问题中去。

3.4.1 整体的微博社交网络

由于是通过一个月的微博数据爬取的用户节点信息，在这一个月中用户不是所有的微博用户，关系也不可能将所有的关系全部体现出来，因此构建的网络也不可能是完全联通的。本文通过Gephi对构建的微博社交网络做数据可视化，将网络中关系抽取，存为.csv文件，导入Gephi。希望从宏观的角度得到并分析微博社交网络的特点。可视化结果如图3-10所示：

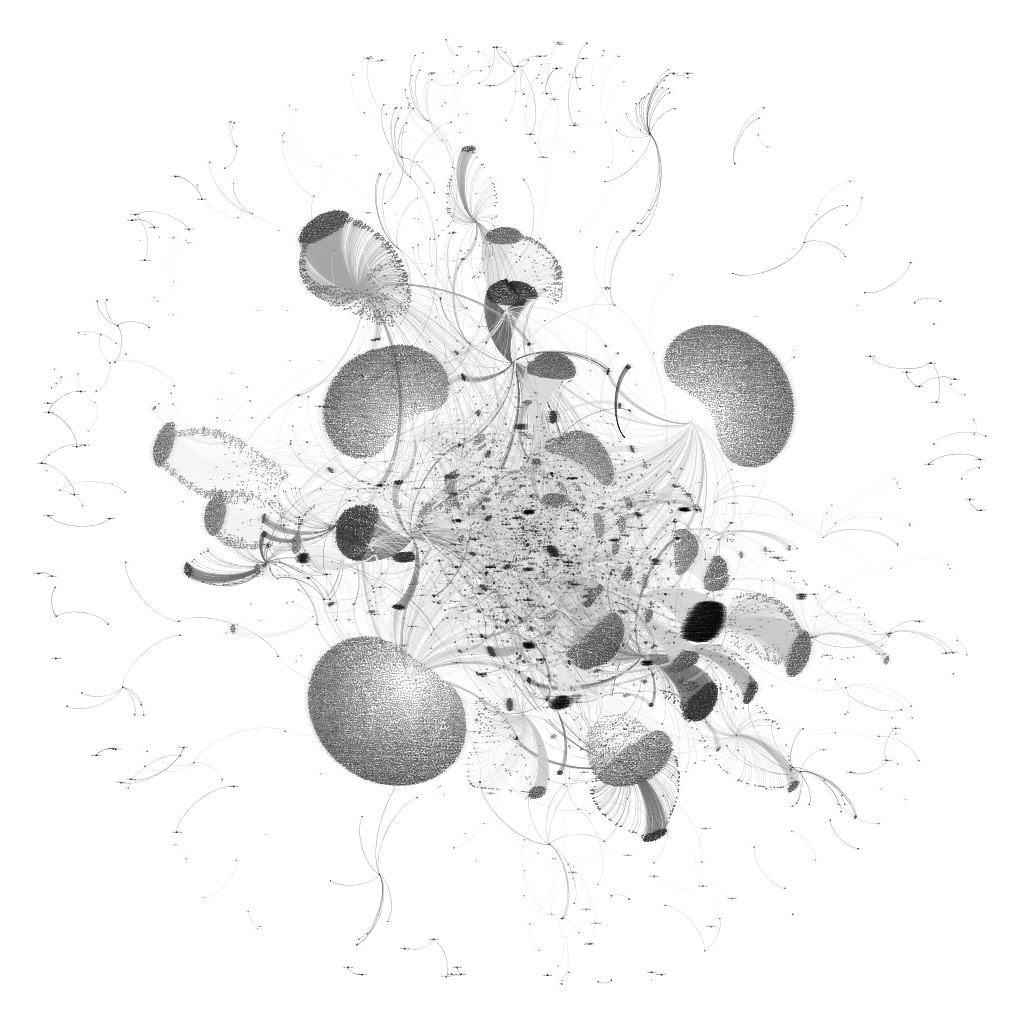


图3-10 微博社交网络数据可视化结果

图中的黑色节点代表微博网络中的用户，黑点之间的连线代表微博网络之间的关注关系。这张图是本文建立的微博社交网络的整体图，包含的用户节点和关系较多。从网络整体的角度来分析，我们发现微博社交网络中存在一定的社区化现象，而且在同一社区中的用户之间的关系形成的结构近似“星型结构”。贺涛在研究微博舆论传播中发现网络群体内节点间信息的传递是一种类似于“星型结构”的辐射性扩散模式，并且“意见领袖”的影响力也与这种结构有一定的关系。所以，这种社区化中的结构也与用户的影响力有一定的关系。

3.4.2 局部的微博社交网络

接下来，考虑局部的用户信息，我们中某个用户出发，建立微博社交网络，希望在相对小规模的情况下，探索微博社交网络之间的特点。下面以用户id为5251583324的用户节点为起始节点，建立规模较小的局部网络。

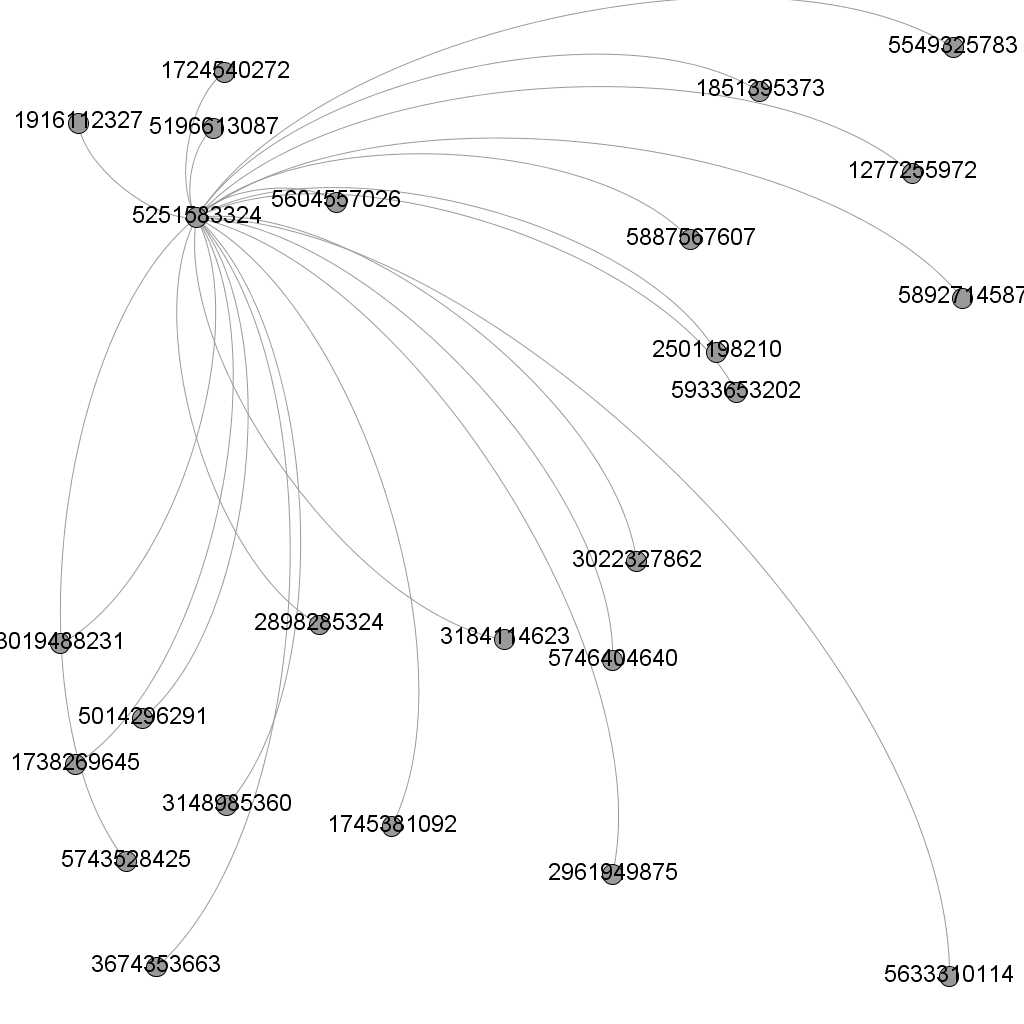


图3-11 微博局部网络结构图

分析上图可以发现，在以用户为中心的局部网络中，也存在着“星型结构”，这种机构的主要关联对象是用户的邻居节点，邻居节点提供给用户的影响是最为直接的，局部的影响往往也是节点影响力的一个重要的因素。

3.5本章小结

本章首先研究了微博网络上的数据，确定了以新浪微博特定主题下的一个时间段的数据作为构建微博社交网络的数据源。接着分析了微博网络中用户节点和关系的定义以及抽取方法，再次基础上研究了微博社交网络的构建方法。利用爬虫技术爬取的微博用户的信息以及用户关系，为构建微博社交网络提供了数据支持。最后对构建的微博社交网络进行了数据可视化分析，从宏观的角度得到微博社交网络中网络的影响力与社区的结构存在着一定关系，从具体的某一用户构建的网络分析得到这些影响与邻居节点对其的影响存在着密不可分的关系。

# 4 微博用户影响力评价算法

本文给出了用户影响力的定义：微博用户的影响力是指用户通过传播信息或者与他人的交互行为而影响他人的思想或者激发他人产生更多交互行为的能力。针对当前微博用户影响力评价研究中存在的问题以及本文对微博用户影响力的定义，本章在对微博用户间交互行为分析和影响力的关键因素分析的基础上，构建一种基于用户间交互行为的微博影响力评价模型。

复杂网络研究学者提出了多种复杂网络中重要节点发现算法，但是没有考虑到节点对邻居节点的影响是不同的，例如在人际关系网络中，人与人之间连接紧密程度是存在差异的，基于此，本章引入影响因子概念衡量节点之间的重要度贡献，基于MDD分解算法，提出本文的微博用户影响力评价算法。

4.1微博交互行为分析

用户之间的交互行为主要有关注、转发、评论、提及，除此之外，在新浪微博中，用户间的交互行为还有发私信、收藏微博、“赞”等。其中，关注、转发、评论这几种行为基本涵盖了微博用户间的所有交互行为特征，本节将对这几种主要行为进行分析。

### 4.1.1转发行为

用户转发消息，可以在所转发的消息前加入文本也可直接转发，用户将消息转发后消息会被发送到用户所有的粉丝页面中，并且被转发的用户的界面会收到特定的提示。

转发关系是用户间就微博信息进行的主要交互关系，推动着信息在微博网络中的不断传播。如果微博用户发布的一条信息被粉丝转发，该条信息就会被再次传播给下一层的粉丝用户，如此下去，该微博信息就会被越来越多的人看到。转发关系是用户间的一种信息交流与反馈的关系。目前对微博用户的转发行为的研究主要集中对转发行为的预测方面，通过预测用户的微博信息能够引起的转发规模，来预测微博信息的传播效果。用户之间通过转发行为传播信息的过程，也是用户间的影响力传播的过程。用户转发他人的微博信息会显示在自己的微博页面上，并被推送给自己的粉丝，转发者成为信息的再次传播者，可以让更多的人看到该微博信息和信息的原发布者，扩散了信息原发布者的影响力。 转发关系可以体现转发者在微博中的活跃性、参与性，微博信息对转发者产生的影响效果以及被转发者的信息在微博网络中的传播广度和被转发者的影响力。

### 4.1.2评论行为

用户对消息进行评论，评论消息会被发送到所评论消息的用户页面。关注关系可以视作是一种信息传播和接收的关。评论关系也是用户之间对于微博信息进行的另外一种主要行为，也可以体现出评论者在微博中的参与度。但是，与转发行为不同的是，用户在评论信息的时候，该信息不会被传播到下一级的用户，且评论的内容仅能在信息发布者的粉丝圈中传播（当然是也会被其他能看到该信息，但不是信息发布者的粉丝的用户）。

用户可以通过评论行为吸引信息发布者的目光，增加用户之间的联系强度，吸引其他粉丝用户参与对微博信息的讨论，也可以提高微博信息的受关注度，从而反应出用户的影响力。

### 4.1.3赞行为

用户在阅读某一条微博时候，如果产生的共鸣或者对这条微博的观点持有赞同态度，就会产生点赞的行为。用户点赞之后，也可以在此点击取消赞，赞数代表的用户发表的微博被大众的认可程度。

在用户对用户之间有赞和没有赞行为对其产生的影响力是有所不同的。赞代表的是积极的、认同的，是高度的同意。所以赞这种行为也是统计用户影响力的一种不可或缺的因素

### 4.1.4关注行为

当用户在浏览微博时，对某个人发表的微博感兴趣时，就会产生关注关系，就是成为某个人的粉丝。在以后这个人发的微博就可以非常方便的看到，并且可以作为粉丝进行转发和评论等一系列的操作。

与博客、Facebook、人人网等在线社交网络不同的是，微博中用户间的关系是一

种单向的、不对称的好友关系。一般来说，微博信息的流向是通过关注关系网络从微博用户到其粉丝，微博用户通过关注行为来获取信息，通过被关注行为传播信息。 用户间关系的不对称造成用户间影响力的不对称，关注关系能够体现出被关注者对关注者的影响。

用户之间的关注关系是构建网络模型的主要的关系。通过关注某一个用户，粉丝可以进行评论、转发和赞等相关交互行为的操作。

表4-1本文中使用的符号

|  |  |
| --- | --- |
| 符号 | 描述 |
| V | 微博网络中的节点集合 |
| E | 微博网络中的边的集合 |
| v | 网络中任意一节点 |
| e | 网络中任意一边 |
| r(u,v) | 用户u转发v的数量 |
| c(u,v) | 用户u评论v的数量 |
| r(v) | 用户v的评论数量 |
| c(v) | 用户v的转发数量 |
| A(v) | 用户v的活跃度 |
| Inf(u,v) | 用户v受到u的交互影响力 |
| IS(u,v) | 用户v受到u影响的影响因子 |
| Inf(v) | 用户v的影响力 |
| InfLv(v) | 用户v受到u影响的影响度 |

4.2微博用户影响因子

### 4.2.1微博用户的影响力

社交网络中的影响力只有通过人们之间的交互活动才能够体现出来，比如用户Ａ在网上的发帖吸引了用户Ｂ，使得后者成为Ａ的粉丝，即Ａ对Ｂ产生了影响力。由于社交影响力的研究工作涉及众多学科和领域，对影响力的的定义也大不相同。

目前大部分研究工作都是针对社交网络结构及其上的交互信息和用户行为特征进行量化和分析的，因此可以把能对信息传播产生影响和对他人行为产生影响视为用户的影响力。在本文所研究的工作中，在针对新浪微博的研究中发现，用户的交互行为是影响力的一个重要的影响因素，用户A发表的微博被B所看到，B产生了共鸣或者不赞同的态度，就会产生一系列的行为，比如说赞、评论、转发和关注等等。这样A便对B产生了影响。影响力不仅仅是由这些行为决定的，而且和产生影响力的用户自身属性以及该用户在网络中的重要性综合判定的。在研究了社会网络中节点的重要性，我们根据K-核理论及其改进的MDD算法，用来作为衡量节点在网络中位置的重要性的一个基本的条件，并且结合网络中的交互行为，定义出节点的影响因子，由这几个方面来综合判定微博网络中用户的影响力。

### 4.2.2影响因子

K-核分解算法对网络进行一层层剥落，最后剩下网络中最核心的节点集，考虑的只是节点自身在网络中所处的位置，并没有考虑节点所处的局部环境，而且ks指标往往赋予大量节点相同的值，导致没法区分处于同一核的节点的重要程度。K-Shell 分解算法得到的结果层次性较差，但是能够快速有效的发现网络中核心节点集，但是并不能进一步区分节点重要性差异。MDD根据引入了分解节点的度的思想，解决了K-核分解存在的大量节点ks值相同的问题，但是只是解决大量高ks值的节点区分问题，对于度为1的核数还是没有有效的方法。Liu等提出根据某一节点到最大K-核节点集合的最短距离之和来区分大量ks值相同的节点。任卓明等提出一种度量指标，根据最小K-核节点和网络中其他层节点的连接关系来区分最小K-核节点的重要性。上述方式只是考虑解决ks值相同的问题，并没有考虑到邻居节点的差异性，不同的节点对其产生的影响力是不同的。

基于这个出发点，本文引入影响因子概念，充分考虑节点的局部环境，即节点与其邻居节点的连接关系。

微博用户网络是一个有向无权网络，边代表用户之间的关注关系。在微博社交网络中，如果用户v和用户w同时关注用户u，也就是说v和w是u的粉丝，由于v和w的节点自身的属性和与u的交互各不相同，因而他们对u的影响也不相同。自身的属性由这个用户的粉丝数、发表微博数和关注数来衡量，与u的交互影响用转发、评论和关注信息来度量，用户之间的影响因子是由这两方面的影响共同决定。因此，引入影响因子来表示节点对其邻居节点的贡献度是不同的，节点v受到节点u的影响因子是由u的自身属性和u和v之间的的交互行为信息来综合度量。

节点u的自身属性的定义：

 (4-1)

I(u)表示用户u的自身属性，mbnums表示用户u发表的微博数，attnums表示用户的关注数，fansnums表示用户u的粉丝数，zannums表示用户u的收到的赞数。α、β、γ、δ是参数。

节点u对v的交互影响：

 (4-2)

Inf(u,v)表示用户v受到用户u的交互影响，c(u,v)表示用户u评论用户v的数量，r(u,v)表示用户u转发用户v的数量，c(v)表示用户v的评论数量，r(v)表示用户v的转发数量,表示u对v的关注关系。取值为0或1，表示存在关注关系或者不存在关注关系。

节点u对v的影响因子：

 (4-3)

IS(u,v)表示用户u对 v的影响因子,I(u)是用户u的自身属性，Inf(u,v)是u对v的交互影响。λ和σ是参数。

4.3 MIB算法模型

根据MDD算法，本文提出了一种基于交互行为的微博影响力度量算法MIB(Mixed degree decomposition Interaction Behavior)，该算法不仅考虑考虑了用户在网络中的位置，并且考虑了邻居节点对其的影响力，即局部属性，从全局重要性和局部重要性综合度量用户的影响力。

### 4.3.1算法设计

上一步给出了u和v之间的影响因子，在本节中我们根据MDD算法得到每个节点的核数值，再结合用户之间的影响因子给出MDD-IB算法的模型。MDD-IB算法根据节点的ks值，并且考虑了邻居节点的影响力贡献，参照网络拓扑结构的邻接矩阵表示形式，构造影响力矩阵I。

 (4-4)

其中I表示影响力矩阵。m(n)表示经过MDD分解后的数值，表示节点i对节点j的影响因子。矩阵中，,表示节点u 对节点 v 的影响力，节点u的m(u)越大，而且对其邻居节点v的影响因子越大，那么它对节点v的影响力就越大。

根据影响力矩阵I，考虑自身ks值和邻居节点的影响力贡献，给出用户的影响力的定义：

 (4-5)

用Inf(v)表示节点的影响力，节点v 的影响力由两部分乘积构成，一个因素是在网络基于网络中节点的位置，即网络全局的重要性，用此节点自身的MDD值来度量，另一个因素是节点的局部重要性，即其相邻节点对节点v的影响力之和，该模型综合考虑了局部重要性以及全局重要性，使得在网络中能较为准确的发现重要的节点。

### 4.3.2算法步骤

本小节，本文给出具体的算法步骤和算法的伪代码实现。

具体步骤：

1、由公式3.1计算出每个节点自身的影响力。

2、遍历用户集合，取得所有和当前用户产生交互行为的用户。

3、定义一个影响力计数器inf；初始值为零。

4、循环遍历与当前用户产生交互行为的用户，分别根据公式3.2计算他们之间的交互影响力，根据第1步计算出自身影响力和公式3.2计算他们之间的影响因子，计算出来的值加入inf。

5、循环步骤4，遍历完成后，得到的inf值即为当前用户的交互影响力。

6、循环步骤2，得到每个用户的交互影响力。

7、根据K-Shell的改进MDD算法，得到每个节点的MDD值。

8、将用户的交互影响力和用户的MDD值合并。首先遍历用户集合，判断用户的交互影响力是否为空，若不为空，根据公式3.5更新结果;若为空，用户最终的影响力即为自身影响力和MDD分解后的影响力乘积，

9、将最终的影响力进行排序遍历输出，即为每个节点的重要度排名。

算法的伪代码如下：

社会网络图表示为G=(V,E)，节点的重要度结果rMap，节点的自身属性I(u),节点之间的交互影响力Inf(u,v),节点之间的影响因子IS(u,v)。

表4-2 MIB算法伪代码

|  |
| --- |
| Input:G=(V,E),uList |
| Output:rMap |
| Initialize:rMap={} |
| 1 //MDD  2 for each u∈V  3 Initial k(u) =k(r)  4 end for  5 while(size(V)!=0)  6 minSet = getMinSets();  7 update k(u) cby k(u) =k(r) + λ ∗ k(e)  8 mdd(u) = k(u)  9 remove minSets;  10 end while  11 // reaction  12 for each u∈V  13 calculate I(u)  14 for each (u,v)∈E  15 calculate Inf(u,v)  16 calculate IS(u,v)  17 end for  18 end for  19 //merge  20 for each u∈V  21 Double result=0;  22 for each v∈neib(u)  23 result += IS(u,v)  24 end for  25 rMap.put(u,result);  26 end for |

4.4算法仿真与分析

### 4.4.1数据集描述

本文重点研究的是微博中用户间的交互行为。交互主要包括发布微博、转发微博、评论微博等。实验选取新浪微博作为数据来源，通过模拟cookies登陆得到所需的数据，将收集到的数据按照以下几个方面进行存储统计：

（1）用户信息表：用户ID、昵称、粉丝数、关注数、微博数；

（2）微博信息表：微博ID、微博内容、发布时间、转发数、评论数、赞数；

（3）用户转发关系表：转发用户ID、被转发用户ID、转发内容；

（4）用户评论关系表：评论用户ID、被评论用户ID、评论内容。

（4）用户评论关系表：关注用户ID、被关注用户ID。

本人爬取了2016年新浪微博7月1日到8月2日#西安身边事#主题下的一个月的微博相关的数据。总共4472条微博，爬取了相关的评论数、转发数、赞数和内容。共有45531个用户、20383条转发关系和59404条评论关系。统计数据如下表3-3所示：

表4-3数据统计资料

|  |  |
| --- | --- |
| 数据项 | 数据量 |
| 用户 | 45531 |
| 微博 | 4472 |
| 转发关系 | 20383 |
| 评论关系 | 59404 |
| 关注关系 | 52437 |

### 4.4.2实验项目

本节为了突出算法的优劣性，分别从以下的几个方面进行了实验：

1）本节算法与其他算法在微博数据下的节点评价效果

3）本节算法与其他算法之间的相关性对比

4）本节算法与其他算法之间的影响范围对比

### 4.4.2实验环境

我们采取的实验环境如下表所示。

表 4-4 实验软硬件环境

|  |  |
| --- | --- |
| 计算机 | 联想G470(lenovo) |
| 操作系统 | 64位Windows 7专业版 |
| 处理器及内存 | Intel(R) Core(TM) i5 CPU 760 @ 2.80GHz 2.93 GHz 12GB RAM |
| 数据库软件 | MySQL |
| 编程语言 | Java |
| JDK | JDK1.7.0\_25 |

### 4.4.3实验结果

为了评估算法的有效性，我们选择MDD算法来做比较。

1)算法结果

通过MIB算法和MDD算法的计算，我们将得到的前100名用户分别通过直观的方式展示出来，如下图所示：

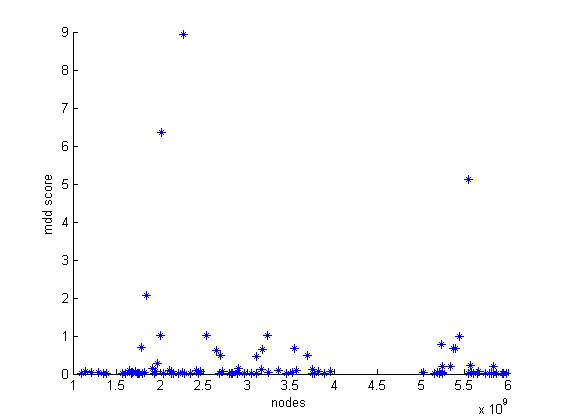


图4-1 MDD算法的前100节点对应的排名值

图中横坐标表示用户的id，纵坐标表示有MDD算法的得到的各个用户的排名值。通过上图可以发现，MDD算法能很好的区分出不同核数的节点的信息，对于同一核数下的节点区分不是很明显，是因为MDD是根据移除的节点的信息和剩下的节点的信息进行判定，但是在微博中，很多的用户交互行为没有考虑到的。图中的存在些相同排名的节点。

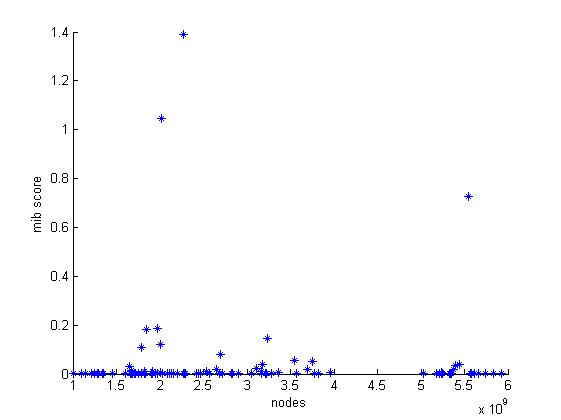


图4-2 MIB算法的前100节点对应的排名值

图中横坐标表示用户的id，纵坐标表示有MIB算法的得到的各个用户的排名值。与MDD算法进行对比，可以发现在高排名的节点区分上结果近似，但是在一些核数相同的节点的区分上明显的效果好于MDD算法。通过上图可以发现，在MDD算法中排名不太容易区分的节点，基本上可以很好的区分。

为了更加具体了说明情况，下表分别给出了MDD算法和MIB算法排出的网络中最具有影响力的前二十个用户。

表4-5 MDD选出的排名前20用户

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | id | Name | Value | Repost | Comment | Fans |
| 1 | 2268603763 | 西安直播 | 8.9354 | 6015 | 17319 | 285625 |
| 2 | 2012158321 | 漂西安 | 6.3538 | 1794 | 11812 | 322265 |
| 3 | 5544721911 | 西安城事儿 | 5.1119 | 2478 | 9659 | 277432 |
| 4 | 1849527897 | 关注三秦 | 2.0645 | 771 | 2878 | 169992 |
| 5 | 2002587837 | 西安校园 | 1.0173 | 569 | 1117 | 232083 |
| 6 | 2531284927 | 瞭望西安 | 1.0032 | 519 | 1507 | 13825 |
| 7 | 3229962754 | 陕视新闻 | 1.0016 | 546 | 1410 | 273935 |
| 8 | 5446004161 | 网说陕西 | 0.992 | 710 | 920 | 75664 |
| 9 | 5239422638 | 西安最前沿 | 0.7847 | 230 | 1113 | 12370 |
| 10 | 1787544013 | 西安同城会 | 0.7104 | 70 | 2141 | 304213 |
| 11 | 5379286126 | 西安新鲜事 | 0.6698 | 473 | 727 | 52266 |
| 12 | 3550995170 | 掌中西安 | 0.6662 | 542 | 714 | 165926 |
| 13 | 5392885175 | 线上西安 | 0.6595 | 766 | 267 | 105768 |
| 14 | 3174967910 | 陕西美食攻略 | 0.6384 | 811 | 157 | 117363 |
| 15 | 2643743024 | 陕西新鲜事儿 | 0.6229 | 172 | 867 | 62113 |
| 16 | 3690085175 | 陕西城事儿 | 0.4752 | 60 | 791 | 72270 |
| 17 | 2697087061 | 西安吃喝玩乐排行榜 | 0.4745 | 333 | 579 | 336789 |
| 18 | 3108778417 | IN古城西安 | 0.4616 | 147 | 653 | 74087 |
| 19 | 1975995305 | 西安晚报 | 0.2786 | 81 | 384 | 1317290 |
| 20 | 5566882921 | 西安月老牵线 | 0.2106 | 75 | 670 | 5691 |

表4-5 MIB算法选出的排名前20用户

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | id | Name | Value | Repost | Comment | Fans |
| 1 | 2268603763 | 西安直播 | 1.390756588 | 6015 | 17319 | 285625 |
| 2 | 2012158321 | 漂西安 | 1.044358857 | 1794 | 11812 | 322265 |
| 3 | 5544721911 | 西安城事儿 | 0.725969546 | 2478 | 9659 | 277432 |
| 4 | 1975995305 | 西安晚报 | 0.186813865 | 81 | 384 | 1317290 |
| 5 | 1849527897 | 关注三秦 | 0.181796773 | 771 | 2878 | 169992 |
| 6 | 3229962754 | 陕视新闻 | 0.146634741 | 546 | 1410 | 273935 |
| 7 | 2002587837 | 西安校园 | 0.12225006 | 569 | 1117 | 232083 |
| 8 | 1787544013 | 西安同城会 | 0.109288007 | 70 | 2141 | 304213 |
| 9 | 2697087061 | 西安吃喝玩乐排行榜 | 0.081414947 | 333 | 579 | 336789 |
| 10 | 3550995170 | 掌中西安 | 0.056818666 | 542 | 714 | 165926 |
| 11 | 3757167087 | 西安发布 | 0.051361312 | 261 | 60 | 900874 |
| 12 | 5446004161 | 网说陕西 | 0.03938855 | 710 | 920 | 75664 |
| 13 | 3174967910 | 陕西美食攻略 | 0.038090008 | 811 | 157 | 117363 |
| 14 | 5392885175 | 线上西安 | 0.035369974 | 766 | 267 | 105768 |
| 15 | 1644855075 | 华商网 | 0.031435456 | 34 | 91 | 744404 |
| 16 | 3108778417 | IN古城西安 | 0.022101223 | 147 | 653 | 74087 |
| 17 | 2643743024 | 陕西新鲜事儿 | 0.020170499 | 172 | 867 | 62113 |
| 18 | 5379286126 | 西安新鲜事 | 0.017755929 | 473 | 727 | 52266 |
| 19 | 3690085175 | 陕西城事儿 | 0.017373027 | 60 | 791 | 72270 |
| 20 | 1676457292 | 中国电信客服 | 0.013615645 | 0 | 0 | 13610838 |

图4-4和图4-5分别给出了MDD和MIB排出的网络中最具有影响力的二十个用户，在这二十个用户中，有十七个用户是重叠的。通过MIB算法，用户西安发布、华商网和中国电信客服成为涌进前十的新用户，但是这三个用户在MDD算法下的排名分别是第26、第32和第62。虽然这些用户的排名都得到了提升，但是原因是不同的。对于用户西安发布来说，尽管发表的博客数相对较少，但是其微博被别人转发的次数非常多，也就是说，作为用户交互行为的一种，转发行为的密集使得用户排进了前二十。对于华商网和中国电信客服而言，尽管用户交互行为也不密集，但是却粉丝的数量众多，而且用户中国电信客服被排名第二和排名第六的用户的用户关注。粉丝的质量高，用户的影响力也会得以提升。此外，在表中，用户关注三秦的排名比用户西安晚报要高，这是因为用户关注三秦与其他用户做了更多的交互。总之，实验结果告诉我们，MDD算法依赖于网络局部的位置，即是用户的粉丝的数量和质量，而我们的算法不仅依赖于粉丝的数量和质量，还充分考虑了用户之间的交互行为。

2)算法之间的相关性

相关性分析是指研究两个变量之间的相关紧密程度，衡量这两个变量之间是否存在某种关联。社交网络用户影响力研究领域中，学者们通常采用统计学中相关系数指标来衡量算法之间的相关性。

统计学中的斯皮尔曼等级相关系数（Spearman’s correlation coefficient）是用来衡量两个变量之间的相关性。它对两个变量的分布没有要求与不管样本容量的大小，都可以使用斯皮尔曼等级相关来研究这两个变量之间的相关性。它对两个变量的分布没有要求与不管样本容量的大小，都可以使用斯皮尔曼等级相关来研究这两个变量之间的相关性。对于样本容量为n的样本，按照升序或者降序原则把n个原始数据和转化为等级数据和，则原始变量 X和Y之间的相关系数为：

 (4-6)

式子中, 分别是原始数据在数据集 X，Y 中的平均排名位置。而在实际应用中，可以采用如下公式来近似计算斯皮尔曼等级相关系数：

 (4-7)

其中，斯皮尔曼等级相关系数的取值范围[-1,1],当越接近 1，则表明两种排序算法之间相关性越大，当越接近-1，则表明两种排序算法之间相关性越小，

甚至相反的。在本文中提到的相关性系数均指按照降序原则，对两个变量的初始值进行排序，然后计算这两个变量的斯皮尔曼等级相关系数。

研究社交网络用户影响力的领域中，我们采用以下两种算法与新算法做个关联性分析。

MDD：基于K-Shell算法，根据节点所处的位置，结合节点的度移除节点，再根据剩下的节点和移除节点的度的信息更显剩下的节点的影响力值。节点的位置是衡量用户影响力的一个非常直观的因素，所以在研究复杂网络问题时，学者们把它作为一个重要衡量节点影响力的算法，如 Zeng 等人，他们研究表明MDD划分算法更好的解决的K-Shell无法区分大量节点影响力相同的情况。

MBUserRank：基于不同的用户交互，即不同的用户对用户影响力，一个用户的影响力被不等划分，并传递给其关注者，对用户交互行为的充分考虑使得我们的方法更符合微博的特性，计算得到的影响力值也更准确。与PageRank类似，阻尼系数被引入，这是因为微博行为也会存在各种随机偶然性，比如转发微博的偶然性等。该算法基于PageRank，但是不同之处在于在该算法中，任意一个用户对其关注者的影响力贡献是不均等的，这依赖于用户之间的交互程度，也就是不同的用户对用户影响力。MBUserRank算法认为不仅依赖于粉丝的数量，并且依赖于粉丝的质量。

表4-6 算法之间的相关性

|  |  |  |
| --- | --- | --- |
|  | MIB和MDD | MIB和MBUserRank |
| 斯皮尔曼等级相关系数 | 0.62 | 0.34 |

由表可看出 MIB算法的运行结果与 MBUserRank的运行结果之间关联性不强，即MIB算法的运行结果与用户粉丝数量没有必然的联系，这与国外研究 Twitter 用户影响力的结论一致。MIB与MDD之间的相关系数为0.62，表明 MURank 算法虽然是考虑到用户的交互行为，但是其运行结果与 MDD结果之间的关联性还是较强

3)验证

在排序出前二十的用户后，下一步就是要验证这二十个用户到底是不是最具有影响力的。我们选择了其他的一些算法作比较，包括距离中心算法、K-Shell算法、MBUserRank算法和MDD算法。

将种子集合的大小标注为k。对于每个算法，每次挑选出排名最靠前的k个节点（对于ks算法是随机选出核数最高的k个节点），从它们出发传播信息，并获得最终激活的节点数。图4-5展示了不同方法选出的种子集合最终激活的用户数量，这里主要统计了从到的情况。将节点u对节点v的交互影响作为激活概率，反应了用户之间的交互行为，交互影响的取值是在0到1之间。此外，激活概率不是我们重点考虑的，主要的是模拟信息的传播。我们采取独立级联模型进行实验的验证。

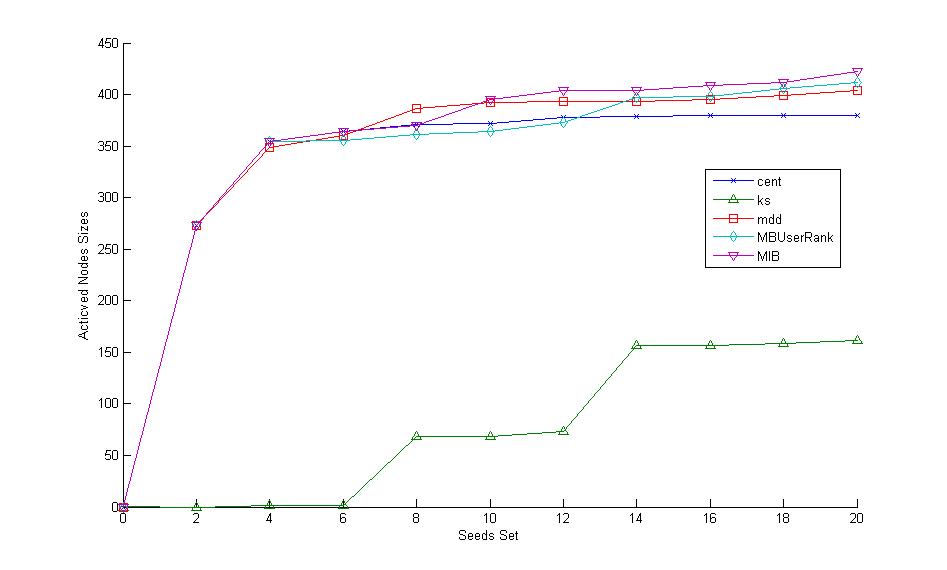


图4-5 不同方法激活用户的对比

这里主要统计了k从0到20的情况，可以看到，由从MIB选出的种子节点集合每次都会比其他方法激活更多的用户，这足以证明我们的影响力计算方法确实能够给出相对较优的结果。

### 4.4.4实验结论

通过以上实验，我们通过观察和总结，得到以下结论，（1）一般来说，用户的粉丝数量在一定程度上影响了该用户的影响力，但并不是唯一的因素；（2）—个用户的影响力与该用户和其他用户的交互行为密切相关；（3）—些用户在粉丝数量和交互行为上并没有优势，但是只要他们能获得影响力大的用户的关注，他们依然可以成为具有影响力的用户。综上实验结果表明，改进的MDD算法的比传统的算法要明显。并且改进的算法的效率比传统的算法效率要高。综合以上几个方面，本章提出的算法是比较具有实用性的。

4.5本章小结

本章先介绍了微博中主要的几种交互行为，并且给出了微博网络构建的模型。MDD算法是较为理想的粗粒度的度量节点重要性的一种方法，但是由于划分结果层次性较差，得到的结果并不能进一步区分节点的重要性差异。因此本文引入“影响因子”概念，构造节点之间的影响力矩阵来表明邻居节点之间的影响力的贡献度。最后提出MIB算法模型，即基于用户交互行为的影响力度量模型，该算法克服以往传统算法片面性，综合考虑了节点的全局重要性与局部重要性，并且考虑了节点对其邻居节点的影响力贡献不同，用节点的MDD值和相邻节点之间的影响力两方面综合度量，更加符合实际情况，能够得到更准确的结果。

# 5 微博用户影响力最大化算法

尽管针对影响力最大化问题已经有许多算法被提出，但是它们不是太耗时，就是缺乏准确度，或者是前提条件太多。本章对该问题进行了进一步的探讨。首先，对算法做了修改，得到新的影响力度量算法。在此基础上引入启发，提出了新的影响力最大化算法。最后通过实验对该算法进行了分析与验证。

5.1影响力度量

上一章重点讨论了微博中影响力的度量方式。首先引入用户对用户交互影响力用来度量一个用户对另一个用户的影响程度，并分别从直接和间接的角度出发，讨论了体现用户交互程度的几个个因素。由于交互是两个人的行为，因此我们从关注者与被关注者的角度分析双向的交互行为，并给出了用户对用户影响力计算公式。其中，转发、评论是用户直接交互行为相关的，而赞和关注是用户和用户之间的间接的交互行为。接着，基于MDD算法，结合用户对用户之间的交互关系，提出了用户之间的的影响因子，最后根据MDD和用户间的影响因子，构造用户影响力矩阵，提出了计算微博中用户影响力的算法。该算法与MDD算法的本质区别在于充分考虑到用户之间的影响不同，这取决于用户自身的属性和用户之间的交互。

在本小节中，各个节点的影响力值用的是上一章计算出来的数值。采取了将用户行为和网络位置结合其来的方法，给出一个较为合理的影响力度量方式，在接下来的最大化问题的研究中，各个节点的影响力值就以这个为前提。

5.2影响力传播模型

在影响力传播模型问题上，微博网络中节点传播不同于在其他社会网络中的节点传播，传统的独立级联模型并不适合微博网络的信息的传播，所以本小节给出一种基于独立级联模型的一种扩展的传播模模型。

### 5.2.1传统的独立级联模型

关于影响最大化问题的模型选择，目前研究最广泛的、最基本的两个传播模型是：线性阈值模型([Linear Threshold Model](http://www.baidu.com/link?url=t9ImdMd4MZiehbaFqVSLXe-NYhC5FNx3huNVUEQadjX1UMQQTVOTPRt256W3YQkbVRRLfMII4SpqEfQN4gt4A3QK-A20ISAj6tEsXER2eBH5df9s0dAuW3Wsw-3IxgT6" \t "https://www.baidu.com/_blank))和独立级联模型(Independent Cascade Model)，所有的研究工作都是基于这两个模型或者其拓展模型。

独立级联模型，是基于相互粒子系统（Interacting Particle System）设计的一个信息扩散的模型，这是一个概率模型。在其传播机制中关键点是把一个节点对另一个节点的激活用概率p表示，激活失败的概率是1-p，概率越大表示激活成功的可能性越大。此外该模型做了这样的假设：即当节点试图激活其邻接节点时，一旦没有激活成功，那么它将永远失去再次激活节点的机会。

给定初始传播节点集合，以及所有节点之间相互激活成功的概率。当传播至第t步时，利用在t-1步中被激活的节点，根据激活概率试图去激活它们的邻居节点，并将在这一步中被激活的节点加入到激活节点的集合中。重复这一过程，直至不再有新的节点被激活。下图是独立级联模型的激活过程。这个激活概率是一个系统变量，与其它尝试激活节点而未成功的节点无关，这也是该模型命名的来历。



图5-1 IC模型传播过程

图中白色的节点表示未激活的节点，黑色的节点表示激活的节点。以A为初始激活节点，在IC模型传播中，先是激活到C和D，接着再由C激活到E，而D并没有激活F，最后再由节点E激活节点G。节点的激活概率是随机的，是概率事件。并且一旦某个节点没有激活另一个节点，那么它就永远失去了激活这个节点的机会。

传统IC模型中认为处于激活状态的节点就会自发地去尝试激活其邻居节点，即一旦节点v处于激活状态，节点v就以100％的概率去对其邻居节点执行尝试激活行为，而邻居节点是否被激活则由节点间的激活概率决定，这个激活概率是是一个系统变量，通过多次蒙特卡洛模拟得到。显然，在微博网络中激活概率并不是随机的，与节点之间的影响以及节点的活跃度有必然的联系。而且，并不是出于激活状态的节点必然会激活其他节点，也存在着激活状态的节点对其他节点没有产生任何影响的情况。

### 5.2.2活跃度

在微博中，存在着许多“僵尸粉”，即是有部分用户的存在是为了“刷粉”，并且在T时间段内没有任何网络行为的用户，这些用户即我们通常所说的“僵尸用户”。在微博平台中，僵尸粉可以通过对信息发布者及所接收信息本身进行反馈来提升用户影响力评判指标，进而间接影响真实用户对信息发布者的判断，其中，粉丝数量这一指标对用户的影响是最直接的。僵尸粉的大量存在，使人们倾向于在粉丝数量与影响力之间建立近似等价关系。这种普遍的问题造成了僵尸粉的泛滥，真实的影响力被隐藏，取而代之的是简单的数字比较。此外，从众心理使得用户可以凭借虚假粉丝数量获取大量真实用户，从而间接提高用户的真实影响力。可见，僵尸粉的存在对信息传播的影响是巨大了，我们需要时识别出这些僵尸粉，争取减少对真实的信息传播产生的干扰。

在独立级联模型模拟信息传播时，每一条边都会对应一个0到1之间的数值，该值表示信息有多大可能从一个节点扩散到另一个节点，称为激活概率。由于在微博中的激活概率并不是像在IC模型中随机的，而是根据节点的一些行为因素来进行决定的。所以独立级联模型在微博网络的传播中并不适用，我们需要考虑节点之间的影响，并且要考虑到节点的活跃程度，将这两个因素综合起来作为模型的激活概率才是比较合理的。

基于这个出发点，本文引入活跃度概念，充分考虑节点的活跃程度。活跃度用来度量发布者在此网络中的活跃程度，通过该用户在网络中的发微博数和参与的回复数以及参与的转发数来衡量。

用户u的在网络中的参与程度为；

 (5-1)

AC(u)表示用户u的参与程度，表blog(u)示节点u发布的微博数，comment(u)表示用户u参与的回复数，repost(u)表示用户u参与度转发数。

用户u的活跃度为：

 (5-2)

A(u)表示用户的活跃度，max(AC)表示最大活跃值的用户。

### 5.2.3基于活跃度的改进的独立级联模型

在研究影响力最大化问题时，我们的方法基于独立级联模型（Independent Cascade Model），但是在激活概率的设置上进行了调整。众所周知，统一独立级联模型将图中的每条边都赋予一个相等的概率值，而在加权独立级联模型中，每条边对应的激活概率等于起始节点出度的倒数。无论是统一独立级联模型还是加权独立级联模型，都不适合直接用于微博网络，这是因为微博网络中每条边的激活概率与这条边连接的两个用户之间的交互行为息息相关。用户对用户影响度表示一个用户对另一个用户的影响程度，它基于用户交互行为。

IC模型是对实际社交网络的一种抽象，模型中节点的主要属性是节点度与节点间的激活概率，节点影响力的定义也都是基于这两个属性。本节对IC模型进行了扩展，引入节点之间的影响度属性，使节点影响力的定义更科学。

节点u对v影响度：

 (5-3)

InfLv(u,v)表示u对v 的影响度，Inf(u,v)表示u对v的影响因子，A(u)为用户u的活跃度。

从用户交互行为计算得到的用户对用户影响力代表了过去一段时间内信息从一个用户到另一个用户的传播情况，在乘以用户的活跃度，构成了用户之间的影响度。我们将此作为我们模型的激活概率。不同的用户对用户影响力体现了不同的信息传播可能性。此外，它的值位于0到1之间，符合激活概率的数学要求。考虑到我们的目标是对影响力最大化问题进行研究，而不是激活概率的精确计算，因此将用户对用户的影响度贡献近似为激活概率是合理的。下图是改进的IC模型图：



图5-2改进的IC模型图

图中白色的节点表示未激活的节点，黑色的节点表示激活的节点。两个节点之间的数字表示激活概率，是由两个节点的影响度决定，即节点u对节点v的影响度等于u的活跃度乘以u对v的影响因子。

5.3影响力最大化SMC算法

有关影响力最大化问题的算法可以归纳为两种类型，一种是基于贪心的算法，它的特点是结果准确性较高，但是运行时间过长；另一种是启发式算法，它在运行时间上具有巨大的优势，但是准确性较低。为了结合这两种算法的优点，本节提出了一个新的算法SMC（Sina Microblog CELF），该算法可以取得近似最优解，而且运行时间相对较短。

### 5.3.1 CELF算法

CELF算法是为了提升原始贪心算法的效率而提出的，它利用目标函数的次模特性来实现优化，使得算法的时间复杂度得到提升。次模特性中的最重要的一条为自然收益递减属性：添加一个元素到集合所得到的边际收益至少会和添加相同的元素到集合是的父集所获得的边际收益一样多。即将一个用户加入到一个集合中产生的影响大于等于将这个用户加入到他的父集合的影响。

CELF算法利用次模特性，在每一步选择初始种子节点时，大量节点的增量影响不需要被重新计算，这是因为它们的增量影响在之前步骤中的值已经小于其它节点在当前步骤中的值。如在第i步时得到的增量影响为420、400、380、320、280、200、70，在第i+1步时计算得到有一个节点的增量影响为320，则在第i+1步中计算出来增量影响为280、200、70的这些节点在第步就不需要被重新计算了，因为根据次模函数的性质，这些节点的增量影响肯定不会大于节点的增量影响。因此它缩短了Greedy算法的时间。

### 5.3.2算法设计

上一节介绍了MIB算法，该算法在计算出用户影响力后，会对用户进行排序。一个用户的影响力代表了该用户在网络中的权威性以及影响其他用户的可能性，相比排名较低的用户，排名靠前的用户更具有话语权，能够影响到范围更大的人群。在影响力最大化问题中，我们的目的是找到大小为的种子节点集合，从该集合出发传递信息能够激活最多的节点。因此，在选择种子节点时，不需要考虑普通节点，因为这些节点对于信息传播的意义不大。相反，要考虑具有一定影响力，也就是排名靠前的用户，因为这些用户往往能够使信息传播的更远。在选择大小为的种子节点集合时，首先根据活跃度去除活跃度值是零的一些用户，接着通过MIB算法对全网络用户进行影响力排序，然后选出排名前ck的用户组成候选种子节点集合，c是启发因子，取值在[0,1]。再从剩下的集合中利用CELF算法挑选出剩下的k-ck个种子节点。当c==0时候，退化成了CELF算法；当c==1时，即是启发式的MIB算法。我们使用贪心方式从剩下未激活的集合挑选种子节点，但是候选种子数却从全网络用户减少到了只有k-ck个用户，这是与贪心算法的区别。

我们选取CELF算法从剩下的未激活节点中找寻剩余的节点，是因为在独立线性模型具有次模特性。当前社会网络影响最大化算法大都基于独立级联模型的次模函数性质进行研究，这是因为对于独立级联模型的任意一个实例，影响传播函数都是次模函数，其影响传播具存次模特性。CELF算法比Greedy算法时间复杂度提高了700多倍，可以实现对原始贪心算法高达700倍的加速。也正因为如此，可以用它代替原始贪心算法完成挑选个种子节点的工作。改进后的算法由于从第二轮开始减少了每一轮需要计算的节点数，该算法的运行时间得以大大降低。能够在较短时间内取得近似最优解。

与原始CELF算法相比，SMC算法考虑了影响旳整个传播过程，每一步都需要计算所有未激活节点的边际影响。刚开始时，所有的节点均处于未激活状态，计算每个未激活节点的边际影响均要遍历整张网络图来进行影响的传播，非常耗时。SMG算法由于找出ck个节点后激活了大量的节点，接着是从未激活中节点中寻找k-ck节点，但是由于己经经历了启发阶段，己经有大量的节点被激活，此时未激活的节点比原始数据集会少很多，相应的会比算法少遍历很多遍网络图，因此的时间复杂度大大降低。

### 5.3.3算法步骤

在影响力最大化问题上，基于贪心的算法准确性较高，但是运行时间相对过长，而启发式算法虽然所需时间较短，但是准确度却往往难以令人满意。

基于这个出发点，在下一步工作中，以前面的影响力度量算法为基础，提出一种基于两阶段的算法。由于贪心爬山算法的时间复杂度主要集中在前部分的节点选择，所以准备将影响力最大化算法的分为三个阶段：

第一阶段是根据上述公式计算用户的活跃度，去除活跃度为零的用户。

第二阶段是根据第三章计算出的用户的影响力，在所有节点中启发式的选出ck个影响力最大的节点作为激活节点，进行传播。c是启发因子，取值在[0,1]

第三阶段再在剩下未激活节点中选出的中用CELF算法求出剩下的k-ck个节点。

算法的伪代码如下：

社会网络图表示为G=(V,E)，种子节点数k，影响力最大化的种子节点S。

表5-1 SMC算法的伪代码

|  |
| --- |
| Input:G=(V,E),k,c |
| Output:S |
| Initialize:S={} |
| 1 for each u∈V  2 calc A(u)  3 if(A(u)==0)  4 Remove(u)  5 end if  6 end for  7 for i=1 to ck  8 select u which make max(inf(u))  9  10 end for  11 for i=1 to k-ck  12 select u which make  13  14 update state  15 end for  16 output S |

5.4算法仿真与分析

### 5.4.1实验数据集

在上一章中，我们使用爬取的新浪微博的数据来验证MIB算法。本节依旧采用该数据做实验。为了进一步保证准确度，将粉丝数为0的用户从原始数据集中删除。众所周知，社会网络中节点的度会遵从幂律分布，因此，为了进一步验证子数据集的有效性，给出用户粉丝人数分布分布，分别如图所示。由图得知，它们服从幂律分布。因此将此数据集做实验数据集是合理。

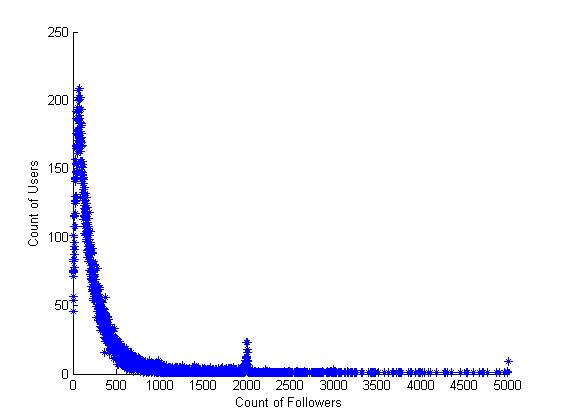


图5-3 用户粉丝和用户数量的关系

### 5.4.2实验结果

为了验证算法的有效性，在此选择了著名的Greedy算法、MaxDegree算法和MBCELF算法来作比较。

Greedy算法：影响力最大化问题的原始的贪心算法。本思想是现将所有节点标记为未激活，然后在这些节点中每次选取一个能激活节点数最多的节点，标记该节点为激活，并加入到最终的集合S中去，然后在剩下的未激活节点中在进行上一步操作，知道选出所要求的k个节点为止。将蒙特卡洛模拟次数设置为1000，就可以得到相当准确的近似值。

MBCELF算法：中科大李翔提出的一种影响力最大化算法。根据用户交互行为的双向性，从转发力度、评论强度、提及密度和关键字相似度四个因素对用户交互程度进行评估。然后，考虑到一个用户对其关注者的影响力贡献值是变化的，且依赖于双方的交互行为，提出一种新的影响力度量算法。接着选取排名前2k的用户,组成候选种子集合，再从该集合中挑选出最终的个种子节点。我们使用原始贪心算法从候选集中挑选种子节点，但是候选种子数却从全网络用户减少到了只有个用户，这是与贪心算法的本质区别。

MIB算法:结合MDD和用户的交互关系的一种算法。构造节点之间的影响力矩阵来表明邻居节点之间的影响力的贡献度。提出MIB算法模型，即基于用户交互行为的影响力度量模型，该算法克服以往传统算法片面性，综合考虑了节点的全局重要性与局部重要性，并且考虑了节点对其邻居节点的影响力贡献不同，用节点的MDD值和相邻节点之间的影响力两方面综合度量。

我们选取c的值为0.5，下面是各个算法的实验对比图：

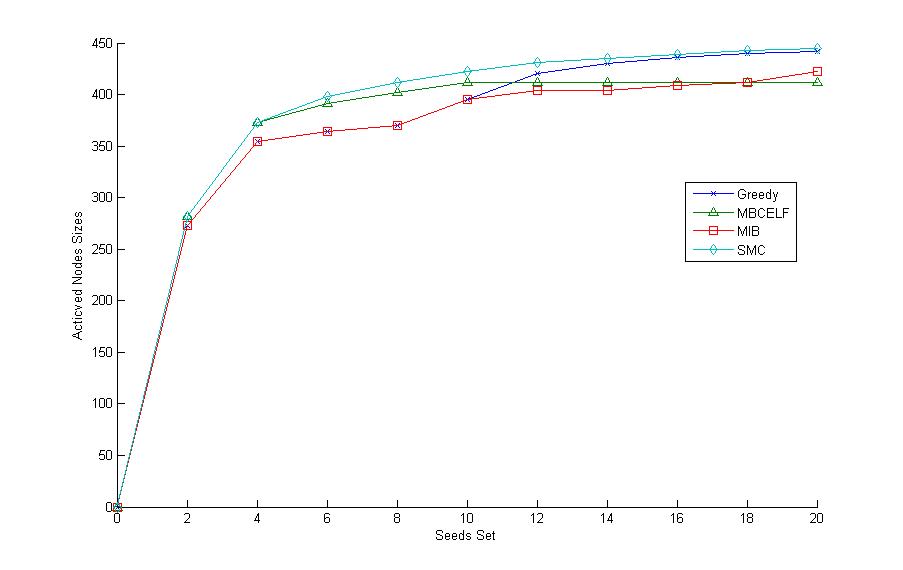


图5- 不同算法的影响范围

由上图我们可以观察到，由SMC算法计算的得到的种子节点集合能取得激活范围比较大的结果。我们选取种子节点结合从1到20，分别计算了Greedy算法、MBCELF算法和MIB算法的激活范围。当种子节点为2个的时候，这些算法的激活的节点一样；当种子节点选取为4个的时候，我们的算法SMC和MBCELF算法激活范围已经超过了MIB算法，这是因为MIB算法只是影响力的度量，但是在影响力最大化问题上，影响力最大的k个节点并不一定导致从这几个节点出发影响的范围会最大，而MBCELF也是由贪心爬山算法的改进，因此它的范围大于MIB算法。随着种子节点的增大，我们的算法SMC的激活范围慢慢超过了MBCELF算法，我们分析这是因为MBCELF算法是选取前2k个用户作为候选集合，在从这前2k个用户中选取k个，但是引起影响力的影响范围最大的k个用户不一定包含在这前2k的用户只内，因此到了后面逐渐把种子节点加入，会明显的看到影响范围逐渐平稳，增大的趋势不明显，可能是由于有影响范围较小的节点包含在这2k节点内。MIB算法由于是取影响力最大的前k用户，因此它的激活范围是低于其他三种算法的。Greedy由于是一步步将造成边际影响值最大的节点加入，所以在后面的节点选择上范围优于了MBCELF。我们的算法SMC先是选取k/2种子节点，在贪心的选取剩下的种子节点，在后续的节点影响范围上明显优于MBCELF，并且在前段的影响范围也明显高于贪心爬山算法。由此可以表明，我们提出的SMC算法是有效的，能给出相对较为理想的效果。

### 5.4.3实验结论

通过以上实验，我们通过观察和总结，得到以下结论，（1）一般来说，用户的影响力排名会对影响力最大化有影响，但不是绝对的影响；（2）排名前k的用户组成的集合并不一定会激活最多范围的节点。（3）SMC算法结合了原始贪心算法与启发式算法的优点，能够在较短时间取得近似最优解。综上实验结果表明，SMC算法的比传统的影响力最大化算法激活范围相对较大。并且改进的算法的效率比传统的算法效率要高。综合以上几个方面，本章提出的算法是合理的。

总的来说，影响力排名前k的用户并不一定会导致影响力最大化。排名靠前只能表示一个用户自身的影响力比较大，这是从个人的角度来说的。然而，影响力排名最大的个用户合在一起并不一定能使信息传播的最远。就好像在现实生活中，由几个最优秀的人组成的小组往往不是最优秀的。相反，只要组内各成员能够各自发挥优点，互相弥补缺点，即便是几个普通人组成的小组也会变得非常强大。另一种合理的解释与社会网络的社区特性有关。如果种子节点能够分布在个不同的社区里，那么每一个种子就能够在所在社区内发挥自己的力量，其所构成的种子节点集合就会变得更具有影响力，就能够影响到更大范围的人群。

5.5本章小结

本章对影响力最大化问题进行了深入探讨。先介绍了微博中的传播模型，并且根据微博中存在的“僵尸用户”的问题，引入了活跃度的概念。接着，提出了基于活跃度的改进的独立级联模型。在此基础上，提出了影响力最大化SMC算法（Sina Microblog CELF）。算法分为三个步骤，先是去除活跃度为0的用户，然后根据上章计算出的用户的影响力，在所有节点中启发式的选出ck个影响力最大的节点作为激活节点，进行传播。c是启发因子，取值在[0,1]，最后在剩下未激活节点中选出的中用CELF算法求出剩下的k-ck个节点该算法结合了综合最大化问题中贪心算法和启发式的优点。最后通过实验，验证了能在实验范围内取得相对于其他最大化算法更大的激活范围，并且时间复杂度也相对较小，该算法是有效并且合理的。

# 6 结论与展望

6.1结论

随着网络的普及和Web2.0技术的迅猛发展，不同种类的在线社交网站如雨后春夢般地涌现出来。以微博为代表的社交网络已经成为当前最重要的媒体之一，随着上网移动终端（如智能手机，平板电脑等）的普遍，几乎每个移动终端上都会有这些社交网站的应用，这样更是助推了这种在线社交网络的多样性和及时性。这些在线社会网络的成员数目庞大，它们的出现对传统社会网络中的影响最大化算法，包括传播模型均提出了巨大的挑战。由最初营销领域的“口碑效应”和“病毒式营销策略”的推广方式，人们发现可以在人群中选取具有代表性的节点子集，由它们可以引起更大的级联影响，逐步产生出了影响力最大化问题。找到影响力最大的节点和影响力传播最大的集合，研究用户的行为以及信息的传播规律将有助于网络公司对每个用户的偏好进行更为准确地把握，并将其可能感兴趣的话题信息、其他用户或者用户社群推荐给用户。

本文的主要工作和结论如下：

1)微博数据爬取和网络建模

首先是新浪微博数据的爬取。根据本文研究的内容，主要爬取话题#西安身边事#下的微博，爬取规定时间内的所有微博的信息，通过这些微博信息，爬去相应的粉丝与评论转发信息，再针对评论与转发中的信息，爬取所属用户的信息。爬虫采用cookies模拟登陆，半自动爬虫，抓取相应的Html后，用正则表达式解析提取相关的信息。对于新浪微博，因为存在中反爬虫机制，对于一个账号的多次请求有次数限制，为此我们我们采取多账号的方式绕过此限制。同时，为了防止过快的请求频率，我们在系统中加入了异常处理层。完成了该爬虫的编写。然后是微博网络的建模，在微博网络中，节点代表微博用户，而边代表了关注关系。我们定义来表示微博网络，V表示节点的集合，E表示有向边的集合。

2)微博网络影响力评价MIB算法

本文引入“影响因子”概念，构造节点之间的影响力矩阵来表明邻居节点之间的影响力的贡献度。最后提出MIB算法模型，即基于用户交互行为的影响力度量模型，该算法克服以往传统算法片面性，综合考虑了节点的全局重要性与局部重要性，并且考虑了节点对其邻居节点的影响力贡献不同，用节点的MDD值和相邻节点之间的影响力两方面综合度量，更加符合实际情况，能够得到更准确的结果。

3)基于活跃度的改进的独立级联模型

在独立级联模型模拟信息传播时，每一条边都会对应一个0到1之间的数值，该值表示信息有多大可能从一个节点扩散到另一个节点，称为激活概率。由于在微博中的激活概率并不是像在IC模型中随机的，而是根据节点的一些行为因素来进行决定的。所以独立级联模型在微博网络的传播中并不适用，我们需要考虑节点之间的影响，并且要考虑到节点的活跃程度，将这两个因素综合起来作为模型的激活概率才是比较合理的。基于这个出发点，本文引入活跃度概念，充分考虑节点的活跃程度。活跃度用来度量发布者在此网络中的活跃程度，通过该用户在网络中的发微博数和参与的回复数以及参与的转发数来衡量。对IC模型进行了扩展，引入节点之间的影响度属性，使节点影响力的定义更科学。

4)微博网络影响力最大化SMC算法

有关影响力最大化问题的算法可以归纳为两种类型，一种是基于贪心的算法，它的特点是结果准确性较高，但是运行时间过长；另一种是启发式算法，它在运行时间上具有巨大的优势，但是准确性较低。为了结合这两种算法的优点，本节提出了一个新的算法SMC（Sina Microblog CELF）。算法分为三个步骤，先是去除活跃度为0的用户，然后根据上章计算出的用户的影响力，在所有节点中启发式的选出ck个影响力最大的节点作为激活节点，进行传播。c是启发因子，取值在[0,1]，最后在剩下未激活节点中选出的中用CELF算法求出剩下的k-ck个节点该算法结合了综合最大化问题中贪心算法和启发式的优点。最后通过实验，验证了能在实验范围内取得相对于其他最大化算法更大的激活范围，并且时间复杂度也相对较小，该算法是有效并且合理的。

6.2展望

本文针对微博社交网络中的用户影响力的评价及影响力最大化问题进行了研究，通过与传统算法的对比，表明了算法的合理性和有效性。但是在本文中仍存在若干的不足之处丞待改进。

首先，在影响力传播模型方面，都采用户交互行为，来对用户激活概率进行简单估计。然而，根据真实的信息传播记录对其进行估计可能更准确。作为领域内的一个研究热点，这可能需要未来重点关注。

然后，在影响力最大化算法上，SMC结合了原始贪心算法和启发式算法的优点，算法可以在较短时间内取得近似最优解。这确实是一个巨大的进步。但是与纯粹的启发式算法相比还有不少差距。接下来还要继续寻找新的算法，使得其在能够保持准确度的同时，进一步缩减运行时间。

最后，当前的模型都基于一段时间内微博网络的静态数据。但实际上，微博网络是动态变化的，与此同时影响力也是动态的。影响力的时序动态分析也是未来我们要研究的主题。

# 致谢

时光如梭，弹指间，三年的研究生生活即将进入尾声。回顾自己在这短暂的三年研究生生活中的点点滴滴，感慨万千，有收获，有教训；有欢乐，有泪水；在工作、学习和生活中得到了父母、老师和同学的指导和帮助，在这里，我要向太多太多关心我、爱我的人表达我的感激之情。

首先我要感谢我的导师饶元教授，因为您教给我的不仅仅是书本上的知识，更重要的是工作上的经验以及做人的道理。您让我懂得如何才能成为一个真真正正对社会有用的人。在这三年里，我最骄傲的一件事，就是有一个对待学生像孩子一样的导师。您的慈祥以及对待学生的态度，一直以来都记在我们师兄弟的脑海里，终生难忘。当然，在三年的生活中，饶老师经常请高教或者海归博士教授等，来我们实验室交流，让我们了解到世界前沿的科技技术以及理论。感谢每次在自己最迷茫的时候，饶老师总是能够给予非常多的建议，让我人生目标更加准确。在此，感谢您三年的培养，感谢您在我未来道路上的建议。总之，我发自肺腑地道一句：您辛苦了！

感谢实验室的兄弟姐妹们。感谢宋明爽、毛伟宁、冯妮等师兄对我的关心与照顾。最初对社交网络的数据挖掘研究没有概念，十分迷茫，宋明爽师兄耐心的为我解惑，用各种形象的比喻让我更好的理解问题。感谢社会计算小组的范刘兵、郑杰群和闫宝彬，在研究生期间一起做过项目，互帮互助，也一起从这个实验室毕业。三年建立了深厚的感情，希望毕业后能常联系，祝大家工作顺利，前程似锦！

感谢软件学院的陈老师、宋老师、吴老师、金老师、朱老师、侯老师，是你们教会了我很多专业知识，教会我从大局看问题。让我在提高理论的同时，提高了自己的修养。在您们那学的每一堂课都是我人生一笔巨大的财富。

感谢我的父母，你们无私的爱让我一直以来都倍感温馨，你们是我前进路上的坚强后盾，感谢你们一路以来对我的信任和支持。感谢我的爸爸妈妈，感谢你们对我的理解，感谢你们对我的包容，感谢你们在我陷入低谷时的启发，给了我继续奋斗的动力和勇气。

# 参考文献

1. 社交媒体.http://baike.baidu.com/item/社交媒体
2. Twitter.http://baike.baidu.com/item/Twitter
3. 2016微博用户发展报告.http://data.weibo.com/report/reportDetail?id=346
4. [E Serrano](http://xueshu.baidu.com/s?wd=author:(Emilio Serrano) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/_blank),[CA Iglesias](http://xueshu.baidu.com/s?wd=author:(Carlos%20A.%20Iglesias)%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/_blank).Validating viral marketing strategies in Twitter via agent-based social simulation[J].[《Expert Systems with Applications》](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=journaluri:(3e47ec45def3c282)%20%E3%80%8AExpert%20Systems%20with%20Applications%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=publish&sort=sc_cited" \t "http://xueshu.baidu.com/_blank" \o "《Expert Systems with Applications》), 2014, 50:140-150
5. [吴信东](http://epub.cnki.net/kns/popup/knetsearchNew.aspx?sdb=CJFQ&sfield=%e4%bd%9c%e8%80%85&skey=%e5%90%b4%e4%bf%a1%e4%b8%9c&scode=07062932;15418699;07064301;30721218;07064263;" \t "http://epub.cnki.net/kns/brief/knet), [李毅](http://epub.cnki.net/kns/popup/knetsearchNew.aspx?sdb=CJFQ&sfield=%e4%bd%9c%e8%80%85&skey=%e6%9d%8e%e6%af%85&scode=07062932;15418699;07064301;30721218;07064263;" \t "http://epub.cnki.net/kns/brief/knet),[李磊](http://epub.cnki.net/kns/popup/knetsearchNew.aspx?sdb=CJFQ&sfield=%e4%bd%9c%e8%80%85&skey=%e6%9d%8e%e7%a3%8a&scode=07062932;15418699;07064301;30721218;07064263;" \t "http://epub.cnki.net/kns/brief/knet).[在线社交网络影响力分析](http://epub.cnki.net/kns/detail/detail.aspx?QueryID=0&CurRec=1&recid=&FileName=JSJX201404002&DbName=CJFD2014&DbCode=CJFQ&pr=" \t "http://epub.cnki.net/kns/brief/_blank)[J].计算机学报.2014.4
6. Meeyoung Cha.Measuring User Influence in Twitter: The Million Follower Fallacy. [International Conference on Weblogs & Social Media](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=confuri:(6f338407a7d8fadd)%20International%20Conference%20on%20Weblogs%20&%20Social%20Media&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=publish&sort=sc_cited" \t "http://xueshu.baidu.com/_blank" \o "International Conference on Weblogs & Social Media), 2010, 14
7. Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E,Makse H A.Identifying influential spreaders in complex networks .2010. Nat. Phys. 6 888
8. Lu LY, Zhang Y C, Yeung C H, Zhou T 2011. PLo S One 6 e21202
9. Zeng A, Zhang C J.Ranking spreaders by decomposing complex networks.2013.Phys. Lett. A 377 1031
10. Liu J G, Ren Z M, Guo Q .Ranking the spreading influence in complex networks.2013 Physica A 392 4154
11. Ren Z M, Liu J G, Shao F, Guo Q 2013 Acta Phys. Sin. 62 108902
12. Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence in a social network//Proceedings of the 9th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 137-146
13. A. Goyal, W. Lu, and L. V. S. Lakshmanan .CELF++: Optimizing the greedy algorithm for influence maximization in social networks. in WWW (Companion Volume) 2011
14. Wei Chen,Yajun Wang and Siyu Yang.Efficient influence maximization in social networks. KDD’09. Paris, France.ACM.2010:P199-207
15. Yu Wang et al. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.2010:P1039-1048
16. W. Chen, Y. Yuan, and L. Zhang,Scalable influence maximization in social networks under the linear threshold model.in ICDM 2010：88-97
17. Guangzhi ZHANG.Discovering High-quality Users from Sina Weibo Based on Trust Transfer Model.Journal of Computational Information Systems 9: 16 (2013) 6467-6478
18. 兰如钦.社会网络上的影响力最大化算法研究.北京交通大学[D].2011:30-32
19. 田家堂,王轶彤,冯小军.一种新型的社会网络影响最大化算法[J].计算机学报2011.10
20. 陈浩,王轶彤.基于阈值的社交网络影响力最大化算法[J].计算机研究与发展2012,49(10)
21. 于淼.微博消息影响力评估及最大化算法研究[D].哈尔滨工业大学2013.1
22. 曹玖新.新浪微博网信息传播分析与预测[J].计算机学报.vol.37No.4Apr.2014
23. 刘建国,任卓明等.复杂网络中节点重要性排序的研究进展[J].物理学报.Acta Phys. Sin. Vol. 62, No. 17 (2013) 178901
24. Hajian B,white T. Modelling influfence in a social network:Metrics and evaluation //Proceddings of the 3rd IEEE Inter-national Conderence On Social Computing .Boston.USA.2011:497-500
25. Guille A,Hacid H,Favre C,Zighed D A.Information diffusion in online social networks: A survey .SIGMODRecord ,2013,42(2):17
26. [S Brin](http://xueshu.baidu.com/s?wd=author%3A%28Sergey%20Brin%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank),[L Page](http://xueshu.baidu.com/s?wd=author%3A%28Lawrence%20Page%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson" \t "_blank) [The anatomy of a large-scale hypertextual web search engine](http://xueshu.baidu.com/s?wd=paperuri%3A%28832c753f0a461dded105104e1c7ec653%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.sciencedirect.com%2Fscience%2Farticle%2Fpii%2FS1389128612003611&ie=utf-8" \t "_blank)[J] Computer Networks,2012,56(18):3825-3833
27. 曹玖新,董丹,徐顺.一种基于k-核的社会网络影响最大化算法[J].计算机学报2015,38(2)
28. 顾亦然,王兵.一种基于K－Shell的复杂网络重要节点发现算法.计算机技术与发展[J]Vol.25 No.9.Sep2015
29. Bryan K, Leise T 2006 SIAM Rev. 48 569
30. C Vargas-De-León.[On the global stability of SIS , SIR and SIRS epidemic models with standard incidence](http://xueshu.baidu.com/s?wd=paperuri:(766e3d2ce311649ce9c30be078810569)&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http://www.sciencedirect.com/science/article/pii/S0960077911001743&ie=utf-8&sc_us=15291331288955727172" \t "http://xueshu.baidu.com/_blank).[Chaos Solitons & Fractals](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=journaluri:(6c8019d0aa42a608)%20%E3%80%8AChaos%20Solitons%20&%20Fractals%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=publish&sort=sc_cited" \t "http://xueshu.baidu.com/_blank" \o "《Chaos Solitons & Fractals》).2011.44(12):1106-1110
31. [Z Lu](http://xueshu.baidu.com/s?wd=author:(Zaixin Lu) University of Texas at Dallas&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/_blank)，[L Fan](http://xueshu.baidu.com/s?wd=author:(Lidan%20Fan)%20University%20of%20Texas%20at%20Dallas&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/_blank).[Efficient influence spread estimation for influence maximization under the linear threshold model](http://xueshu.baidu.com/s?wd=paperuri:(717a1c4900b319d24133f72bfa051711)&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http://link.springer.com/article/10.1186/s40649-014-0002-3&ie=utf-8&sc_us=4628010924937953883" \t "http://xueshu.baidu.com/_blank).[Computational Social Networks](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=journaluri:(6e64c412460b7396)%20%E3%80%8AComputational%20Social%20Networks%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=publish&sort=sc_cited" \t "http://xueshu.baidu.com/_blank" \o "《Computational Social Networks》). 2014. 1(1):2
32. 樊兴华，赵静，方滨兴，李欲晓．影响力扩散概率模型及其用于意见领袖发现研究[J]．计算机学报，2013,36(2):360-367
33. Jimeng Sun and Jie Tang.A survey of models and algorithms for social influence analysis.Social Network Data Analytics.2011: P177-214.
34. 王洪伟,任豪.社交网络的有效信息量研究:基于节点强度的分析[J] 北京工业大学学报2015,41(5)
35. 王兵.复杂网络的节点重要性度量算法研究[D].南京邮电大学.2015.03
36. 李翔.微博中影响力的研究[D].中国科技大学.2014.04
37. 周东浩,郭景峰,吕加国，基于信息偏好的影响最大化算法研究[J] 计算机研究与发展2014.4
38. 赵之滢.于海.基于网络社团结构的节点传播影响力分析[J].计算机学报
39. Cao, Tianyu，Wu, Xindong Wang, Song，Hu, Xiaohua OASNET: an optimal allocation approach to influence maximization in modular social networks [J] Acm Symposium on Applied Computing, 2010:1088-1094

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．交回学校授予的学位证书；

2．学校可在相关媒体上对作者本人的行为进行通报；

3．本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。

4．本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）： 日期： 年 月 日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．学校可在相关媒体上对本人的失察行为进行通报；

2．本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。

3．本人接受学校按照有关规定做出的任何处理。

指导教师（签名）： 日期： 年 月 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）： 日期： 年 月 日

指导教师（签名）： 日期： 年 月 日

(本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用)