申请上海交通大学硕士学位论文

基于PageRank的社交网络用户实时影响力研究

学　　校： 上海交通大学

院　　系： 电子信息与电气工程学院

班　　级： B1003494

学　　号： 1100349162

姓　　名： 陈少钦

专　　业： 电子与通信工程

导　　师： 李建华（教授）

上海交通大学电子信息与电气工程学院

2013 年 1 月

Master. Dissertation Submitted to Shanghai Jiao Tong University

A Users' Real-time Influence Algorithm of Social Network Based on PageRank

Author: Shaoqin Chen

Specialty: Electronic and Communication Engineering

Advisor: Prof.Jianhua Li

School of Electronic Information and Electrical Engineering

Shanghai Jiao Tong University

Shanghai, P.R.China

January,2013

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：陈步钦

日期：2013 年 01月 03日

上海交通大学

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校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，允许论文被查阅和借阅。本人授权上海交通大学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。

本学位论文属于

保密□，在___年解密后适用本授权书。

不保密□。✓

（请在以上方框内打“√”）

学位论文作者签名：陈力铁　　指导教师签名：李建华

日期：2013年01月03日　　日期：2013年01月03日

基于PageRank的社交网络用户实时影响力研究

摘 要

社交网络(Social Network Service，简称 SNS)是现今互联网中的一个重要领域。SNS 是一个基于好友关系的信息发布、分享的平台。随着移动互联网的发展，越来越多的人们能随时随地在这个平台上发布、分享消息。社交网络正在扮演着一个非常重要的消息传播平台。著名的 SNS 有 Facebook、Twitter、YouTube、人人网、新浪微博等。由于社交网络是基于现实中的好友关系建立起来的网络结构，因此它既有网络结构也具有社交的特质。

新浪微博作为国内著名的社交网络之一，它拥有超过三个亿的用户，每天都有海量的信息在这个平台上发布。新浪微博网络在信息传播中扮演着非常重要的角色，甚至有超越传播媒体的趋势。如“林书豪”事件率先是在新浪微博上传播，为人们所知。因此研究新浪微博网络是具有重大的商业价值与应用价值，研究用户影响力有助于理解消息传播模型。

由于人们在社交网络上的动作也是人类的一种主观的行为，本文基于人类动力学的理论与研究方法，提出一种新的评价新浪微博中节点实时影响力的算法。本论文的主要工作有以下几个方面：

首先阐述了复杂网络的起源、基本概念与两个基本特性，即小世界效应与无标度特性。已有学者证明在线社会网络也具有小世界效应与无标度，进而说明了在线社会网络是复杂网络中的一种。论述了社会网络的发展与分析方法，其中重点阐述了“六度分割理论”与“150 定律”。之后简要介绍了在线社会网络的发展与研究现状。

其次，通过新浪微博提供的 API，我们采集了微博用户信息与微博信息。在这个数据集中，统计分析了新浪微博用户之间转发微博的行为，实验结果表明用户之间的转发行为时间间隔分布是服从幂律分布，这与人类动力学领域中的研究成果一致，即人类行为时间间隔分布是服从幂律分布，并且具有“胖尾”特性。

再次，本文基于 PageRank 算法的基本思想提出一种社交网络用户实时影响力算法,这个算法称为 Micro-blogging User Rank,简称 MURank。MURank 算法是基于用户之间转发行为的时间间隔分布规律与用户之间的网络结构，计算网络节点用户的实时影响力。通过评价微博网络中用户节点影响力，可以迅速地找到信息传播过程起关键性作用的用户节点，有助于理解微博网络中消息传播模型等。

最后，分别使用粉丝数量与传统的 PageRank 算法来评价微博网络的用户影响力，我们分析了这两个实验结果与 MURank 实验结果，分析结果表明 MURank 算法具有比另外两种算法，具有更好的实时性。

关键词：社交网络，转发行为，新浪微博，用户影响力，MURank 算法

A Users' Real-time Influence Algorithm of Social Network Based on PageRank

ABSTRACT

Social networks (Social Network Service, SNS) is an important field in the Internet today. SNS is a platform based on friend's relationship. In this platform, people can share, get and broadcast information. With the development of the mobile Internet, more and more people can share messages in this platform anywhere. SNS is playing a very important role in the message communication platform. There are many famous SNS, such as Facebook, Twitter, YouTube and so on. Since the social network is based on real friends' relationship, it has both characters of media communication and social network.

AS a famous social network in China, Sina micro-blogging has over three million users, mass information are released on this platform every day. Sina micro-blogging network plays a very important role in the dissemination of information, even beyond traditional communication media. For example, the news of "Jeremy Lin" is spread in the Sina micro-blogging at first. Therefore the researches of Sina micro-blogging network have significant commercial value

and the value of applications, it will help us to understand the model of message propagation.

Human behavior in the internet is also subjective. Based on the human dynamics theories, we advanced a new algorithm to evaluate Micro-blogging users' real-time influence. The main work of this paper is presented as following：

First, we described complex networks' originate, the basic concept and two basic characteristics. The basic characteristics are small-world effect and scale-free characteristics, which are different from random network. Some studies have shown that the online social networks also have the characteristics of small-world effect and scale-free, which means online social network is a kind of complex network. Then discussed the development of the social network analysis, and focused on the "six degrees of separation theory" and the "150 rule". The paper also made a brief introduction of the development and research status of online social networks.

Second, through Sina micro-blogging API, we got a lot of user information. In this data set, we statistical analyzed the behavior of retweet, the experimental results show that the distribution of retweet behaviors follows power-law distribution, which is consistent with research in the field of human dynamics, that the distribution of human behavior follows power-law distribution, and has

the characteristics of fat tails.

Third, based on the basic idea of PageRank algorithm, we advanced an algorithm of user real-time influence. This algorithm called Micro-blogging User Rank, namely MURank. MURank based on the distribution of users' retweet and the network structure to calculate the real-time influence of a user. Through evaluating the influence of node, we can quickly find the key nodes in the dissemination of information to understand the micro-blogging network news dissemination and contribute to a micro-blogging influence.

Finally, we used the traditional PageRank algorithm and the number of fans to evaluate the influence of user respectively, the experiments show that MURank has a better character of real-time than the others.

Keywords: social network, forwarding behavior, Sina micro-blogging, user influence, MURank algorithm

目 录

第一章 绪论

1.1 研究背景与意义

社交网络作为一个全新的互联网交友平台与信息传播平台,每天都有海量数据在这个平台上发布。社交网络是一个虚拟社会网络，它是由许多节点构成，是现实社会在网络上的体现。每个节点都代表了现实生活中的一个人或者一个组织，节点之间的好友关系也是现实社会中的社会关系。在这个虚拟社会中，人们从事着大量的社交活动，如交友、分布消息、关注好友状态与分享视频等。在社交网络的平台上，人们可以分享自己的心情、关注朋友的状态以及了解一些热门话题等。目前社交网络的形式主要有交友网络、博客、视频共享等形式。

微博服务是近年来兴起的一种信息交流、传播平台。微博用户可以通过手机、网页来发布与获得信息。与博客等传统社交网络相比，微博的最大优势在于它应用上的实时性与内容的简易性。随着移动互联网技术的快速发展，微博用户可以随时随地的发表一条消息。这条消息可以是用户的状态、心情或者只是转发了其他人的消息，这条消息可以是几个字符或者只是转发其关注者的微博消息，但是它的字符不能超过一个限值。微博，这概念起源于美国的 Twitter。Twitter 是一个著名的社交网络及微博客服务的网站。截止 2012 年五月份，Twitter 注册用户量超过 5 亿，其特点在于 Twitter 消息不能超过 160 个字符。近年来，新浪、腾讯、网易等门户网站也开发与发布了各自的微博服务，国内的微博服务得到快速的发展。其中新浪微博拥有超过 3 亿用户并依然呈现增加的趋势，是国内最大的社交网络。新浪微博用户平均每天发布微博数量超过一亿条。新浪微博已经成为一个主流媒体平台，越来越多的人使用、关注与研究微博，微博已经成为社交网络领域中的一个研究热点。

拥有庞大的用户群体以及海量的信息，微博在应用领域与研究领域都有着重大的意义。研究社交网络的主要目的在于挖掘网络用户行为以及商业、应用价值等。首先海量的微博网络用户数据信息，为人类动力学研究提供了足够的样本，使得人类行为动力学的研究成果更加的精确与方便。微博网络拥有庞大的用户，这些用户又是现实生活中的个体，微博网络用户的每个行为都是在其主观意识下支配完成。因此研究微博网络上的用户行为就是研究现实生活中人类的行为，对于理解人类行为、社会关系网络有着重大

的意义。其次，微博网络是个信息交流平台，它有着媒体的特性，每天都有海量的信息在这个平台上传播与发布。另外与传统媒体相比，微博网络有着更好的实时性，它的传播速度更快，研究微博网络使得我们更好理解消息的传播方式等。微博网络吸引着国内外的许多学者，他们希望能通过研究微博网络中用户的行为与信息传播，能够发现人类社会的规律与信息传播的规律等。

1.2 国内外研究现状

1.2.1 社交网络研究方向

目前社交网络的主要研究方向有数据挖掘研究、个性化搜索、社区挖掘、基础结构研究、网络用户行为分析和用户影响力等。目前，研究对象主要是 FaceBook 与 Twitter，并且取得了许多成果。

一、数据挖掘技术研究

根据相关报道显示，FaceBook 每天共收到 500TB 新数据[25]。根据 Twitter 网站公布的数据，目前 Twitter 每日信息发送量已达 5000 万条，即每月 15 亿条[26]。社交网络的数据量不同于互联网以往的服务。这些海量信息具有异构性和多样性，不是简单的结构化数据。针对这个问题，学者提出了解决方案，如基于社交网络服务[27,28]、隐私保护[29]以及分析社交网络的流量，设计出下一代 Internet 基础架构[30-31]等。其中也有学者利用搜索引擎来挖掘社交网络中的人际关系，如 POLYPHONET 的抽取分析算法[32]等。

二、个性化搜索

个性化搜索即是基于社交网络的搜索。与传统搜索技术不同，基于社交网络的搜索技术更在意于个性化的搜索，每个搜索用户的兴趣爱好等都不一样，所以单一的搜索结果很难满足一些用户的需求。社交网络是基于现实生活中人际关系建立起来的，社交网络的用户信息等数据蕴含了用户的习惯、爱好等个人信息。基于社交网络的搜索就是根据用户的以往喜好与好友关系，对搜索结果再次进行筛选，以达到搜索结果的个性化，为用户提供更好的搜索体验。基于社交网络的搜索或者个性化的搜索，也是搜索领域的一个新的研究热点。

三、社交网络社区挖掘

社交网络中用户因为某种爱好等构成一个社团。研究社团的结构，对于理解为网络结构等有着重要的意义。社交网络中社团挖掘主要有意识图形分割（Graph Partition）

和分级聚类（Hierarchical Clustering）两种。其中分级聚类是以 GN（Girvan-Newman）[36]算法为经典，之后很多学者也提出许多 GN 算法的改进算法[37-38]。

四、网络结构研究

与其它基础网络的拓扑结构不同，社交网络的拓扑结构是以现实世界中的社会关系为基础建立起来的。例如在人人网中，只有当用户确定另外一个用户的好友请求，他们之间才能建立关系。每一个好友关系都是社交网络结构中的边，网络用户则是网络结构中的节点。研究社交网络结构主要有度数中心性(Degree Centrality)、中介中心性（Betweenness Centrality）等[33]。另外有学者对于社交网络拓扑结构进行了静态分析，研究表明社交网络中节点度分布服从幂律分布[34]，以及社交网络具有小世界效应[35]现象。这两个重大的研究都表明了社交网络是复杂网络中的一种，为以后关于社交网络的研究提供了理论基础与方向。

五、网络用户行为的研究

社交网络是基于现实中的人际关系建立起来，用户的一切行为都是人们在网络的主观行为。研究社交网络上的用户行为，有助于人类行为的研究。这方面的研究主要集中是研究浏览网页、回复帖子、点击影片等用户行为。如 YeWu,Kurths[8]等人研究了天涯论坛上用户的回复行为，对几个帖子的回复情况上进行跟踪，发现用户对于某一帖子发表评论的时间间隔服从幂律分布。周涛[52]对 Netflix 的用户数据进行分析，按照用户的活跃度，把用户分为 20 个群体，他发现每个群体中用户点播电影的的间隔时间分布都服从幂律分布。还有一些学者对国内的 QQ 用户在群里的回复动作，表明 QQ 用户行为的时间间隔分布服从幂律分布。已有的许多研究实例表明社交网络的用户行为与网络结构都存在幂律分布现象。

六、用户影响力的研究

影响力的研究一直是各个领域的研究热点。如 Google、百度等对 Web 页面影响力的研究，学术界用 H 指数来衡量一个学者的学术成就。同样，社交网络用户影响力也吸引了许多学者与商家。目前，国内外对于社交网络的研究主要有：Jianshu Weng[3]基于用户之间的话题相似度，估算每个用户在各个话题的影响力，最后综合用户在各个话题的影响力之和，作为用户的综合影响力。Daniel M.Romero[2]等人认为用户的影响力具有消极性与积极性两个方面，消极性体现为用户对其关注者消息的抵触，积极性体现为粉丝对用户消息的转发情况。基于用户之间以往消息转发比例与 HITS 算法的基本思想，提出了 IP 算法来综合衡量 Twitter 用户的影响力。国内的新浪微博主要是用《微数据》来衡量用户的影响力，认为用户的影响力是覆盖度、传播力、活跃度三者综合的体现，此外，新浪微博还有各种排行榜，如粉丝数量、转发次数等。国外有一个叫做

Klout 的创业型企业，它专门研究了 Twitter 与 Facebook 用户影响力，使用了 Twitter 中的 35 个变量来综合衡量用户 Twitter 的影响力，后来它又将用户在 Twitter 与 Facebook 的影响力综合起来，它的最终目标是希望能有效地评估出网络用户在整个 Intelnet 的综合影响力。

1.2.2 影响力研究现状

社会网络的信息传播与影响力的研究已经存在于各个领域，如市场营销、同性、社会学、政治学等。早期的影响力研究，一方面是从无标度网络的中心亲和性，即一个固定话题传播过程的活跃用户[50]，另一方面是认为社会网络的消息传播，主要是看哪些人在传播，这些人就具有极高的影响力[51]。总结来说，他们认为节点的好友数量，在消息传播过程起着至关重要的作用。

由于社交网络的自身优势与商家的成功运营，社交网络拥有这海量的用户，它存在着巨大的商业价值与学术价值，如广告投放等。因此社交网络用户影响力的研究也是社交网络的热点之一。目前,这个研究领域主要是集中在对Twitter用户的研究。Meeyoung[1]等人研究 Twitter 上消息传播性质，分别从粉丝数量（In-degree）、转发次数（retweet）与引用次数（mention）三个指数来衡量 Twitter 用户的影响力，实验结果表明粉丝数量多的用户不一定被很多人引用（mention）以及他的消息不一定被转发很多次，因此用户的影响力与其粉丝数量的关联性很低。Daniel M. Romero[2]等人认为社交网络上的用户影响力不仅取决于其粉丝的数量，还跟粉丝的消极性相关。当一个用户看到其好友消息时，他以一个概率 P 选择转发这条消息，而这个概率 P 是与他以往转发这好友消息比例相关。最后他们提出基于 HITS 算法的 IP 算法来计算 Twitter 用户的影响力，IP 算法分别算出一个用户的积极性与消极性来综合衡量网络用户的影响力。Daniel M. Romero[2]等人的研究表明用户消息被浏览次数与用户粉丝数量之间的关联比较弱。 Jianshu Weng,Ee-Peng[3]等人认为 Twitter 用户的影响力是他在各个话题影响力的总和，他们提出基于 Topic-Sensitive PageRank 的 TwitterRank[3]算法来衡量 Twitter 用户的影响力，使用 TwitterRank 算法分别计算用户在每个话题（topic）的影响力，然后进行叠加，计算出用户在网络中全部话题(topic)的影响力之和为用户最终的影响力。Honey[9]对 Twitter 上‘@’符号进行了研究，证明 Twitter 消息在交流协作方面有一定的作用。

另外，国外研究社交网络用户影响力比较著名的企业是 Klout，它是一个美国创业型企业。Klout 公司先后分别研究了 Twitter 与 Facebook 的用户影响力，并将用户在这两个平台的影响力综合起来，Klout 指数（Klout Score）表征了用户在 Facebook 和

Twitter 的综合影响力[55]，Klout 公司致力于研究用户在多个平台上的影响力、在整个 Internet 上的影响力。Klout 在 2011 年获得 KPCB 领投的 3000 万美元巨额融资。

新浪微博是国内著名的社交网络之一，每天都有海量数据在这个平台上传播。目前，新浪微博主要是运用《微数据》来分析量化用户自身或者周围的粉丝的影响力，个人影响力是指覆盖度、传播力、活跃度三者的综合体现。国内还没有一个像 Klout 专注于研究社交网络用户的影响力、网络用户综合影响力的企业。

1.3 本文的主要工作

国外的学者主要是从用户话题相似度、用户以往转发其关注者微博消息的比例来研究节点的影响力，进而评估节点用户的全网影响力，没有从人类行为动力学方面研究节点影响力。而在微博网络上用户的行为是人们在网络上的行为，这种行为是现实世界中人在主观意识的支配下完成的，是人类行为的一部分。而已有的研究实例表明人类行为时间间隔具有一定的规律，并且微博网络结构类似于 Web 页面的网络结构。所以本文基于人类行为动力学、复杂网络的研究方法与衡量 Web 页面权威性经典算法的基本思想，研究微博网络中的用户影响力算法。另外从 Klout 例子来看，研究用户影响力具有极大的商业价值。本文的主要工作内容有：

- 通过调用新浪微博提供的 API,获取微博用户的好友列表与微博消息列表。借鉴人类行为动力学的研究方法，我们研究分析了新浪微博网络中用户之间的转发行为时间间隔分布，实验结果发现这个分布服从幂律分布。以此为基础，本文定义了新浪微博用户之间的关注度数学模型。
- 分析新浪微博网络中消息传递形式与好友关系结构，发现新浪微博网络结构类似于 Web 页面的网络结构，而 PageRank 算法是衡量 Web 页面权威性的一种经典算法。因此本文基于 PageRank[4]的基本思想，提出一种新浪微博用户实时影响力算法（Micro-blogging User Rank，简称 MURank）。
- 运用 MURank 算法评估用户节点在网络中影响力，并且分别使用 PageRank 算法、粉丝数量、用户微博消息被转发次数来衡量用户影响力。综合分析了这几种算法的实验结果，结果表明与其它算法相比，MURank 具有更好的实时性，更能反映了用户在时间段内新增粉丝数量。

1.4 论文组织结构

本文共分为五章，第一章绪论简要介绍用户影响力算法的现状研究背景与意义、社交网络的研究方向、介绍了国内外研究社交网络用户影响力的现状；第二章首先介绍了复杂网络的相关理论，之后论述了社会网络分析法，对社交网络研究现状的综述，并且对微博网络的发展做个简要的介绍以及一些基本概念，为后续章节做理论准备；第三章与第四章是本文主要研究工作内容，第三章阐述了实验数据集的获取方法与结果，借鉴人类动力学的研究方法，研究分析了用户转发行为的时间间隔分布。以此为基础，建立微博网络用户之间关注度的数学模型。第四章基于 PageRank 算法的基本思想与用户之间关注度模型，提出一种用户实时影响力算法（MURank），并且论述了该算法的参数含义与具体实现过程；第五章是实验结果分析，运用本文中新算法评估新浪微博用户的影响力。与其它算法进行比较，研究分析新算法的有效性与实时性；第六章是结束语，对本文工作进行总结，以及提出未来工作的内容。

第二章 社交网络综述

社交网络是复杂网络中的一种，其网络结构等特性的研究分析是基于复杂网络理论。本章首先介绍了复杂网络的起源，基本概念与特性等，其次阐述了社会网络的相关理论、社交网络的起源与发展现状，最后论述了新浪微博网络中的网络结构、消息传递方式和两种用户关系，为后面章节内容提供理论支持。

2.1 复杂网络的基本理论

自然界上存在着许多的原始系统，如神经系统等。为了研究这些系统，学者往往采用网络结构来描述它们。一个典型的网络结构是由许多节点与节点之间的边构成的，其中网络中的节表示点了现实生态环境中的不同个体或者事物，网络结构中的边代表了不同个体之间的某种联系，如果系统中不同个体之间有某种联系，那么对应与网络结构中这个两个节点之间存在一条边，如果不同个体之间没有联系的话，那么这两个节点之间不存在边。这种网络结构通常可以表示成为 G=(V,E)，其中 V 是网络中的节点集合，E 是网络中节点之间的边集合。根据边是否有向，网络分为无向网络与有向网络两种。其中如果两个节点之间的边（x，y）与（y，x）相同，则成这个网络为无向网络，如公路网络等。如果两个节点之间的边不同，则这个网络称为有向网络，如神经网络，WWW 网络等。另外如果网络中每条边具有权值，则网络称为有权网络(weighted network)，否则这个网络结构称为无权网络(unweighted network)。

在描述现实的系统，学者认为可以用一些规则的结构来描述系统各个因素之间的联系，就是说网络结构中的边都遵循着某种特定的规律。这个网络结构称为规则网络[10]。到了二十世纪中期，匈牙利的 Paul Erdos 和 Alfred Renyi 提出了随机图理论[11-13]，随机图理论认为网络中两个节点之间是否存在边是个概率问题。根据随机图理论构成的网络被称为随机网络，随机网络(ER 随机网络)指任意两个节点之间是以概率 p 随机的连接。近年来，随着计算机的技术发展，计算机能够处理更加复杂的系统。学者们发现了一个现象许多现实系统既不能用规则网络来描述，也不能用随机网络来描述，这些系统有着前面两种网络所没有的统计特征。描述这些系统的网络被称为复杂网络。与其它两种网络相比，复杂网络的最大特点在于它具有小世界效应(small-world effect)与无标度特性(scale-free)。

2.1.1 小世界效应

在网络中，从一个节点到另一个节点的所有路径中长度最小的一条路径，称为这两个节点之间的最短路径。网络中所有节点对的最短路径的平均值，称为网络距离。节点的簇系数定义为它所有相邻节点之间连边的数目占最大可能连边数目的比例，网络的簇系数是网络中所有节点簇系数的算术平均值，它反映了网络节点聚类的情况。

小世界网络，这个概念最早是 Watts 和 Strogatz 在 1998 的一篇《Nature》论文中提出的，在这篇论文中他们提出了 WS 小世界模型[18]。WS 网络模型认为节点会以概率 P 放弃原先的节点而连接新的节点，构造出一个新的网络结构。当 p=0 时，相当于节点自此至终不会选择新的节点，还是规则网络；当 p=1 时，这时网络结构是一个随机网络。WS 网络模型具有大的簇系数和小的网络距离。之后，物理科学家们把网络中具有大的簇系数与小的网络距离的统计特征，则称这个现象为小世界效应。具有小世界效应的网络称为小世界网络(small-world network)。规则网络，小世界网络与随机网络之间的关系如图 2-1 所示：

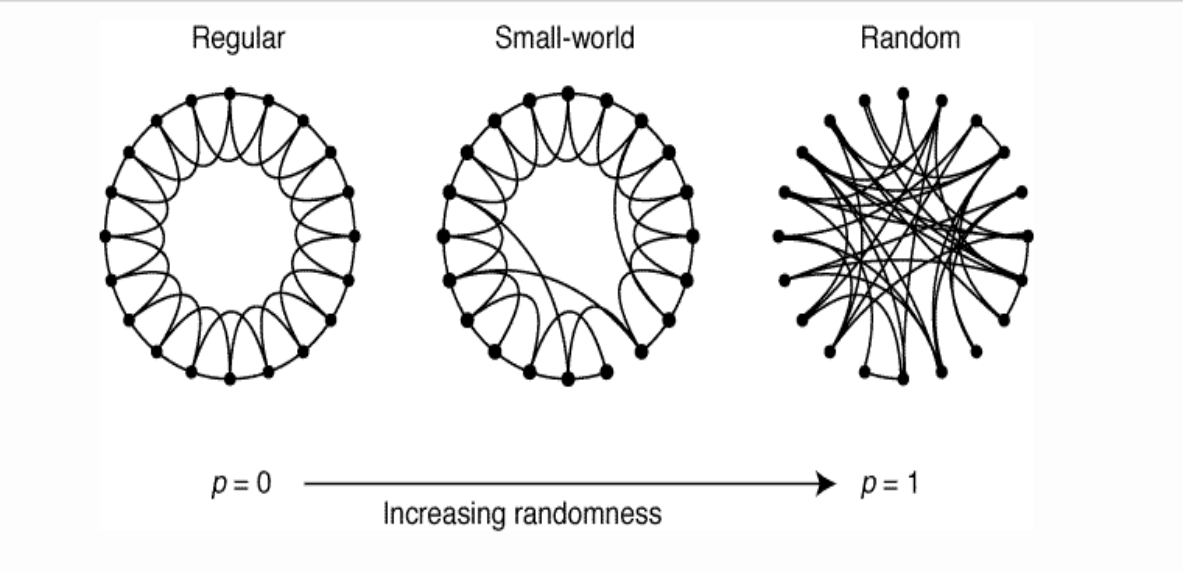

图 2-1 规则网络，小世界网络与随机网络之间的关系[18]

Fig.2-1 the Relation among Regular, Small-world and Random network[18]

2.1.2 无标度特性。

继 Watts 和 Strogatz 之后，科学家们研究发现小世界效应普遍存在于大量的真实网络，同时他们还发现很多网络的节点度服从幂律分布[19-20]。其中节点度是指与这个节点

连接的边数量。在有向网络中，节点度分为入度与出度两种，入度是指指向该节点的边的数量，出度是指节点指向其它节点的边的个数。通常使用函数 P(k)描述网络节点的度的分布情况，P(k)指网络结构中任何一个节点的度为 k 的概率。

1999 年，Barabási 与 Albert 的研究表明复杂网络具有无尺度特性[5]。复杂网络中的无标度特性是指复杂网络的度分布服从幂律分布（$P(k)\sim k^{-\alpha}$），而随机网络的节点度分布是服从泊松分布。在随机网络中有许多节点的度都是很接近的，而在复杂网络中，只有少量节点的度非常大，大部分节点的度都比较小。因此复杂网络与随机网络的最大不同点在于各自的度分布。

因此衡量一个网络是不是复杂网络，主要看它是否具有小世界效应与无标度特性。复杂网络具有规则网络、随机网络所没有的特性，而这些特性往往是现实系统网络所具有的，所以研究复杂网络具有重大的意义与应用前景。

2.2 社会网络的相关理论

社会网络是指社会行动者之间的相对稳定的关系体系，社会网络关心的是社会行动者之间的互动和联系，社会互动会影响到人们的社会行为。社会网络理论是从 20 世纪 30、40 年代开始的，德国社会学家 G.Simmel 在《群体联系的网络》（1922）中第一次提出了“网络”概念。Barnes(1954)使用“社会网络”的概念分析挪威一个渔村的社会结构，他系统化的呈现了复杂的社交网络系统下成员之间的关系模型，他是最早提出了“社会网络”的概念。

关于社会网络的形式化界定，有学者认为社会网络指由多个社会行动者(Action)和行动者之间的关系组成的[39]，这种关系可以是行动者之间的任何相互作用，比如共事、朋友与血缘等。图论是社会网络分析的基础数学理论之一，社会网络的形式化描述可分为社会关系网络图及社会关系邻接矩阵，主要用于发现和理解社会组织结构[43]。使用图来描述社会网络，其中元素：

• 点 社会网络中的点代表了社会行动者。在社会网络研究领域中，点可以代表一个社会个体与一个集体单位、组织等。例如点可以描述现实的一个人、具有共性的一个群体，可以替代一个教室、学校、甚至可以表示一个村庄、社区、国家等，当然也就可以表示虚拟网络中一个用户，或者说是一个网络账号等。

• 边 社会网络中的边表示了社会行动者之间的某种或者多种关系。如果两个点之间存在边，那么说明这两个行动者之间或多或少存在着一种或多种的社会关系。这些关

系如：1、个人之间的关系如尊重、敌意等；2、生物意义上的关系如遗传关系，亲属关系等。3、地理位置的关系，如同一个国家、同一个城市等。4、共同的兴趣爱好，如喜欢某个明星、喜欢某个话题等。

实现世界两个行动者之间是同学关系的同时，他们也可能是同事、恋人等关系，这种关系称为“多重关系”，也是社会网络分析者重点研究对象。行动者之间的联系又分为强联系与弱联系。强联系是指行动者的社会网络同质性比较强（即行动者之间从事的工作、兴趣爱好等方面都有着相似之处），行动者之间关系紧密，具有很稳定的人际关系。弱联系是指行动者的网络异质性比较强（即认识的人比较多，朋友可能来自各行各业，因此可以获得多方面的信息），行动者之间不紧密联络或是间接联络，人际关系比较不稳定。

2.2.1 社会网络结构特点

社会网络分析（Social Network Analysis，SNA）是对社会网络中行动者之间的关系进行量化研究[42]，研究其中的社会结构与社会关系等。根据研究角度的不同，社会网络分析可分为两种基本视角[40]：关系取向(relational approach)和位置取向（positional approach)。关系取向通过社会联结本身——如强度、密集性、对称性、规模大小等来说明特定的行为和过程，这个研究角度重点关注行动者之间的社会性关联关系，。这种观点认为那些强关系、集中的以及相对孤立的社会网络可以促进集体认同；位置取向则重点关注行动者之间的社会关系模式化，这些社会关系的特点在于在结构上它们处于平等地位。它研究的是两个或者以上的个体和第三方之间的关系所折射出来的社会网络结构，注重使用网络结构等效来理解人类行为。

社会网络分析中重要的理论有“六度分割理论”和“150定律”。

2.2.1.1 六度分割理论

六度分割理论[14]，该理论认为可以最多通过六个人我们就可以认识世界上任何一个人。六度分割理论起源于 20 世纪 60 年代，美国心理学家 Stanley Milgram 设计了一个实验：他随机的把一封信发给一些人，信中写着一个波士顿股票经纪人的名字，Milagram 要求每个收信人将这份信发给自己的亲朋好友，这个朋友必须是收信人认为他是最有可能接近这名股票经纪人。在信的传递过程中，每个收信人都遵守这个约定，直到这封信到达这个股票经纪人。最后，Milgram 统计这个股票经纪人收到的信件，他发现每个信件平均经过 6.2 次传递就能达到目的地。根据这个实验结果，Stanley Milgram 提出了“六度分割理论”，该理论认为世界上任何两个人之间建立某种联系，最多需要 6 个人。

六度分割理论示意图如图 2-2 所示：

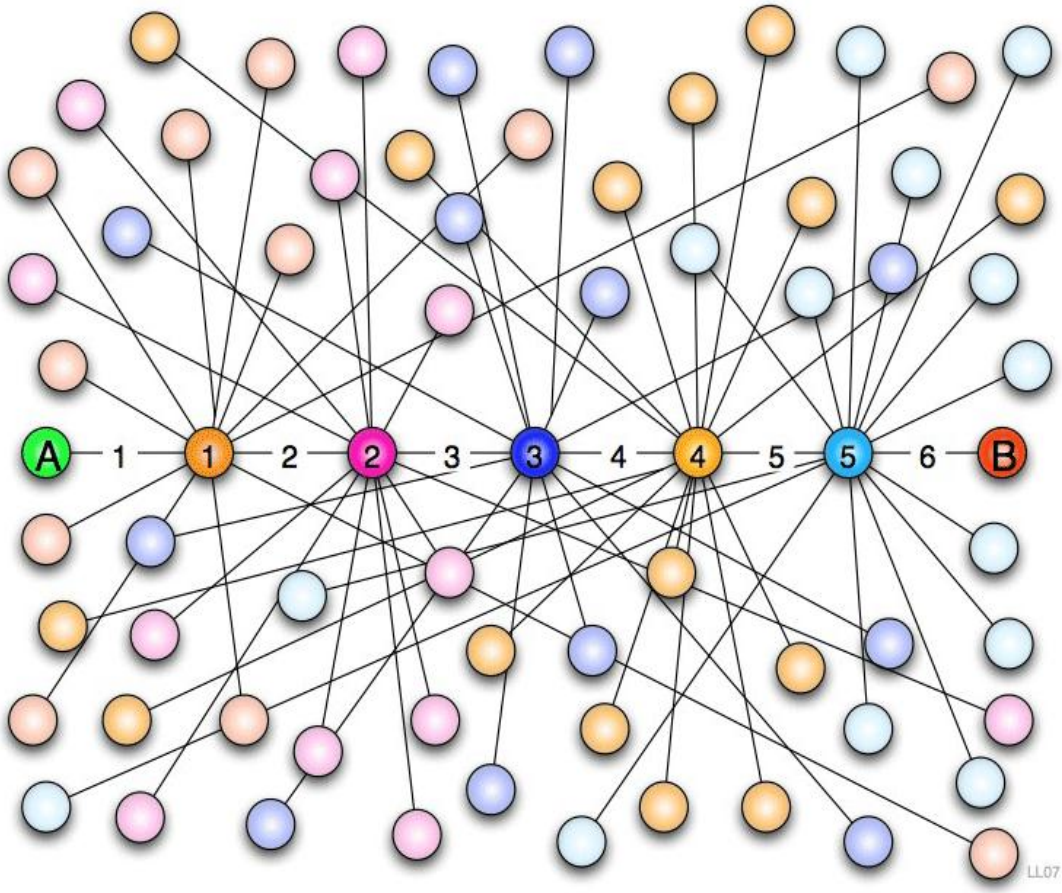

图 2-2 六度分割理论示意图[15,16]

Fig.2-2 the Six Degrees of Separation Theory[15,16]

之后，为了验证“六度分割理论”，科学们在其它领域进行了类似于Milgram的实验，如Tjaden和Wasson设计的“game of Kevin Bacon”实验和Watts用E-mail进行“小世界”实验[17]，他们的实验结果基本证明了Milgram的小世界理论。

2.2.1.2 150 定律

150定律(Rule of 150)，也称为邓巴数字，是由人类学家Robin Dunbar提出。Robin Dunbar研究了不同形态的原始社会，发现这些村落的人数量均在150人左右，根据猿猴的智力与社会网络，他提出了著名的“150定律”。该定律指出：同一个时间段内，一个人与许多人保持某种关系，但是保持着稳定人际关系或者强关系的人只有150个左右。这个理论也被Facebook内部社会学家CameronMarlow证实，他表示Facebook上的网络用户的平均好友数量在120人左右。一些Facebook用户虽然有着超过500位好友，但是他们

经常互动的好友却非常少而且相对稳定。这个定律也被应用到我们生活中来，例如我们只能存储140个人的号码在移动的SIM卡里。

2.2.2 社会网络分析方法

自上个世界二三十年代产生以来，社会网络分析法已经发展了多半个世纪，已经成为了一套完善的理论。对网络结构的分析就是对社会网络的关系结构与属性进行分析，它研究的对象就是行动者（Action）之间的关系模型。社会网络分析家Barry Wellman[41]指出，作为一种研究社会结构的基本方法，社会网络分析法应该具有如下基本原理：

- 关系纽带相互作用的强度与内容往往是有所不同的，是不对称地相互作用。
- 关系纽带直接或者间接地把网络成员关联在一起，所以需要在更加复杂、规模更大的网络结构下对它们进行分析
- 社会关系纽带结构产生了非随机网络，因而形成了交叉关联、网络集群(network clusters)以及网络界限。
- 网络群以及个体都是交叉关联在一起的。
- 由于不对称的关系纽带和复杂网络，造成了稀缺资源的分配不平等现象。
- 网络中不仅仅存在合作行为也存在竞争行为，这些行为共同点是它们以获取网络中稀缺资源为目的。

这种结构分析法的意义在于：社会科学研究是以现实社会网络结构作为研究对象。通过研究社会网络中诸多关系，它把个体之间的社会关系、社会网络的“宏观”结构与微观结构结合起来。

社会网络分析主要有两个流派，整体网模型和自我中心网模型。

为了分析网络结构以及研究社会群体中角色的关系结构，整体网络结构模型采用图论为分析工具，使用社会网络结构关系图表示社会小群体中错综复杂的人际关系，如随机网络、图论、PERT图、决策树等。后来，它为了能够研究群体数量更为庞大、网络关系更为复杂的情况以及能够全面研究一定范围内的人际关系，它引入了社会计量学、数学中的矩阵方法，能够定量的、全面的描述整个网络关系结构。目前，整体网络分析重点在于研究网络结构中个体之间间接或者直接的关系方式和网络结构时间的特性。其中主要定义有：簇（Clusters）、桥（Bridges）、紧密性、中心性、中距性等；侧重研究不同社会地位的网络成员的明星（Stars）、联络人（Liaisons）、孤立者（Isolates）、结合体（Coalitions）、小集团（Cliques）等。

自我中心网络（Ego-centric Networks，EN）分析，主要通过研究个体的人际网络

是如何影响到个体行为，挖掘人际关系网络和社会团体网络形成的原因，学术渊源主要来自于英语人类学院的社区研究。目前，自我中心网络分析集中在新经济社会学的研究上，并且逐渐延伸到社会研究的各个领域，如社会阶级、社区、人口流动、社会变迁等。在此使用的主要概念有：网络的密度，网络的结构大小以及网络结构的多元性，人际关系中的强弱联系情况。在这个研究领域的著名学者有：提出嵌入性概念的Granoveter，提出市场网络观的White等。从数据采集与数据处理方面来看，自我中心网络分析主要有以下几种方法：互动方法、角色关系和情感方法以及社会交换法。

目前，这两种网络分析方法都得到了广泛应用，社会网络分析已经广泛的应用于经济学、人类学以及心理学等领域。

2.2.3 社会网络分析的指标

汪小帆等人[44]详细介绍了复杂网络的相关理论，认为社会网络分析的指标可以从节点的度、密度、中心性、凝聚子群、平均路径长度、聚类系数六个方面展开。

- 节点的度是节点所有属性中最简单的属性，一个节点的度定义为网络其它节点与该节点之间存在边的总条数，即该节点与网络中其它节点存在多少个联系。度在不同网络有着不同的值。在无向网路中，度等于该节点与网络中其它节点连接的总边数。在有向网络中，一个节点的度分为出度和入度两种，其中节点的出度是指从该节点指向网络中其它节点的边的总数目，节点的入度是指从网络中其它节点指向该节点的边的总数目。一个节点的度大就意味着这个节点在社会网络中与许多节点有联系。
- 中心性是一个人重要的个人结构位置指标，是社会网络分析中一个非常重要的指标，体现一个人的职务地位重要性或特权性，和个人社会声望等常用这一指标。弗里曼总结了三种关于中心性的度量指标，即亲近中心性（Closeness Centrality）、程度中心性（Degree Centrality）以及中介中心性（Betweenness Centrality）[45-46]。程度中心性，用来衡量节点在网络结构中的中心位置，有向图的程度中心性又非为外向程度中心性和内向程度中心性。亲近中心性，是以距离为概念来计算一个节点在网络结构上的中心程度，与其他节点越亲近，其中心性越高。中介中心性是指节点作为媒介的能力，中介中心性高的节点掌握了信息流以及商业机会，从而获得中介利益。
- 密度描述一个图中各个点之间关联的紧密程度。一个图的密度定义为图中实际拥有的连线数与最多可能拥有的线数之比，一个拥有n个点的无向图最多可能拥有

n(n-1)/2条边。所以密度的测度取值范围为[0,1]，显然一个完全图的密度为1。

- 凝聚子群,试图发现一个社会网络中存在多少个凝聚子群（小团体），也就是所谓的社团，研究这些社团内部成员的关系特点，此外还能分析社团之间的各种影响和联系。因此，凝聚子群分析也被形象的称为小社团分析。社团分析或挖掘是社会网络分析领域的一个研究热点。
- 聚类系数，网络有相应的社团分布，社团内的节点之间有密集的联系，而社团之间的联系则相对较少。即一个网络社团中各个节点与周围节点相连的边数除以理论总的可能边数，所得的比值是这个节点的聚类系数，一个社团的聚类系数就是其内部各个节点的聚类系数平均值。因此聚类系数的取值范围也是[0,1]，仅当社团内部为完全图时其值等于1，即网络中任意两个节点都直接相连，而许多大规模的实际网络都具有明显的聚类效应。
- 平均路径长度 定义为网络上任意两个节点之间距离的平均值，用于分析网络的规模，一般对于复杂的综合性网络来说，如果有特殊需要可以对此进行深入研究，由于其计算量很高一般不需要对这个指标做详细探讨。

对现实网络系统中的各种关系进行量化分析，社会网络分析量化了社会网络结构与社会行动者之间的关系，为分析社会网络提供了可能性。社会网络分析主要是挖掘错综复杂的社会网络中隐含的人性行为特征、社会本质等。社会网络分析法已经在多个领域得到了应用，如发现和理解社会结构[47]、研究组织通信行为[48]、学术合作网络研究[49]。

2.2.4 社会网络与复杂网络的区别

不管是社交网络还是复杂网络，它们使用节点代表现实系统中的行动者，使用边表示现实系统中行动者者之间的关系。社会网络分析方法也是研究复杂网络的一个重要方法。复杂网络与社会网络分析不同在于复杂网络侧重从数学理论层面去解释和处理问题，而社会网络分析则通过数学统计等多方式研究社会现象，揭示人类社会行为的特征。已经有学者[34-35]研究证明了社会网络具有小世界效应与无标度特性，证明了社会网络是复杂网络的一种，它是复杂网络研究领域的一种特殊的网络。

2.3 社交网络

2.3.1 社交网络起源与发展

社交网络，这个词是由 J. A. Barnes 在 1954 年提出的[5]。社交网络即是在线社会网络，是人们通过爱好、朋友等形式在 Internet 上建立的社会关系网络结构。社交网络起源于网络社交，最早的网络社交形式通过 E-mail，人们通过 E-mail 进行通信与联系，从而建立起网络结构。之后 BBS 的出现，它实现了消息从点到面的传递，不是点到点的交流。后来即时通信工具，如 MSN、QQ 等提高了消息的及时性。最后随着网络社交的发展与人们交流的需求，网络用户在网络上越来越呈现出作为人的形象与思想，这时候社交网络出现了。

近年来，社交网络得到迅速的发展，已经成为互联网行业的一个热门的网络服务应用，它渗透到人们生活的方方面面，为人们提供了便利的信息交流平台。成功的社交网络平台有，基于好友关系的 Facebook、基于视频分享的 YouTube 等。国内的社交网络平台有人人网，开心网、朋友网等。这些社交网络平台让一群有着相同兴趣等的人聚集在一起，不管这些人身在何地。来自世界各地的社交网络用户，在网络这个大平台上交流兴趣、建立好友关系等。社交网络平台缩短了人们之间的距离，也减少了交友的成本。由于有着显著优点，社交网络服务已经遍布了互联网。2007 年的 compete.com 的报告显示[23]，世界上社交网络服务的访问量占互联网流量的 40%以上。一份来自 CR-Nielsen 的统计报告[24]显示，全球有超过三分之二的网民使用过社交网络服务。

由此看出，社交网络聚集着大量的网络用户。享受社交网络平台服务的同时，社交网络用户也在为互联网提供了海量的数据，这些数据里包含着庞大的人类行为数据与社会消息数据等。通过挖掘这些数据中我们所关系的信息，运用社会网络分析的办法对于这些数据进行处理与分析，我们可以挖掘出人类行为特征、人类社会规律与社会网络结构等。因此分析社交网络有着重要的现实意义和应用价值。

2.3.2 微博综述

微博，也称为微博客、微型博客。博客来源于英文的“Blog”，是“Web Bog”的缩写。在 2006 年，Evan Williams 创建的新兴公司 Obvious 推出了 Twitter 服务，开创了“Micro Blog”的形式，与传统的博客不同，Twitter 的消息被限制在 140 字符之内，人们可在 140 字符之间表达自己的最近动态、对某个事件的看法等。Twitter 还允许用

户使用手机等移动终端发布消息。由于 Twitter 的内容简明性与使用的移动性强，Twitter 已经成为一个受广大用户欢迎的社交网络及微博客服务的网站。随着 Twitter 在国外的成功运营，国内的各大门户网站也争前恐后的推出各自的微博服务，如新浪微博、网易微博和腾讯微博等。现在我们几乎每天都与微博打交道，如往往在电视上能看到，以及周围的亲戚朋友都在使用微博这一网络服务等。

在众多的国内微博服务中，新浪微博是国内最为著名的微博。我们可以在一些大型晚会上看到新浪微博的身影，如春节晚会等。新浪微博在消息传播扮演着非常重要的角色。如在“林书豪”事件里，新闻借助新浪微博传播的速度不亚于传统媒体，甚至有超越传统媒体的趋势。由于新浪微博是社交网络中的一种，社交网络是在线的社会网络，因此我们也可以对新浪微博进行形式化界定。我们使用图论理论来构造出新浪微博的网络结构，以及结合新浪微博的特点，定义了新浪微博的网络结构的基本要素，为后面章节的研究内容提供基本的概念。新浪微博的基本要素有：

- 节点　即网络中的参与者，是指社交网络上活动的参与者，即在社交网络中与其他人相关联的具体个人、组织等。在新浪微博网络结构中，一个节点代表了新浪微博的一个注册账号，这个账号可以是一个具体的个人、也可以是明星，如姚晨、组织，如南方报纸等。研究新浪微博的网络结构，我们是以节点为基本单位。
- 边　表明了网络结构中两个节点之间的一种或者多种的联系,代表了两个用户节点之间存在或多或少的关系。这些关系有好友关系、消息转发关系、相互关注关系等。在其它社交服务网络中，用户之间的关系是双向的，即一个用户 A 加为用户 B 为好友，那么用户 B 默认的加用户 A 为好友，如人人网。与这些社交网络服务不同，新浪微博网络中节点 A 关注了节点 B，并不代表节点 B 就关注了节点 A。因此新浪微博网络是一个有向网络，节点之间的关注是有方向的。
- 社区　社区是某部分节点的集合，社区的形成不是人为的作用，它是有节点自身的固有属性造成的，固有属性如学校、公司或者共同兴趣等。例如来自同一个学校的成员，就自然而然构成了一个该学校的社区。在同一个社区中，社区成员交流更加频繁、好友关系更加密集。与社区外的网络节点相比，该社区网络中的密度会明显的比较大，有着明显的划分。
- 关注关系 如果一个用户 A 关注了另一个用户 B，那么用户 A 是用户 B 的粉丝，他们之间存在着关注关系，新浪微博网络中，关注关系是单向，即用户 A 关注了用户 B，不代表用户 B 就关注了用户 A。这点与其它社交网络不同，例如在人人网中，如果用户 A 确定了用户 B 的好友请求，则他们之间的关注是相互的，

双向。

与其它社交网络不同，新浪微博网络中消息的传递方式是单向的。为了后面章节的关注度模型与影响力算法的描述，我们有必要阐述下新浪微博网络中消息传递的形式，如图 2-3：

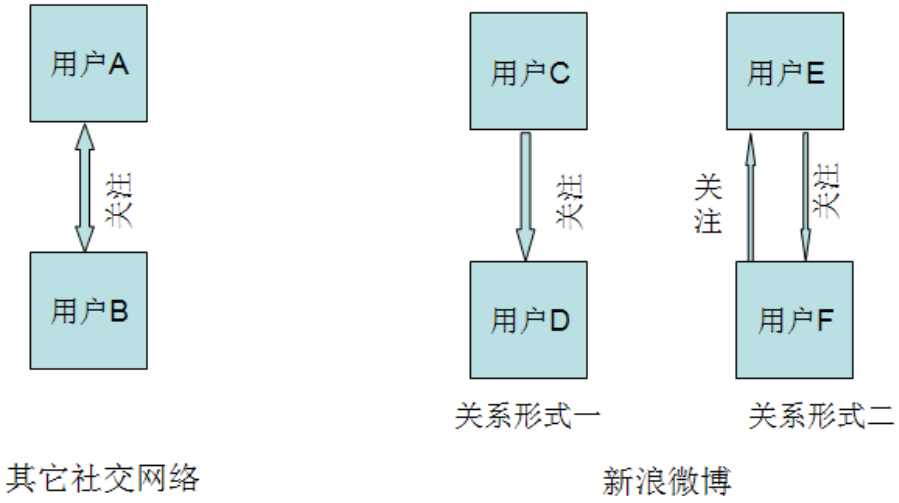

图 2-3 社交网络的消息传递方式

Fig.2-3 Message propagation in Social Network

在其它社交网络中，如人人网，如果用户 A 确定了用户 B 的好友请求，加用户 B 为好友，那么任何一方发布消息，这个消息都会出现在另外一方的个人消息中心页面上。他们之间的消息是双向传递；在新浪微博服务中，用户之间的关系分为粉丝关系与关注关系。用户 D 关注了用户 C，则称用户 C 为用户 D 的关注者,用户 D 是用户 C 的粉丝。当用户 C 发布公共微博消息之后，该消息会自动出现在用户 D 的个人信息中心页面，但是用户 D 发布的任何微博消息，都不会自动出现在用户 C 的个人信息中心页面上。与其它社交网络服务的消息传播方式相比，新浪微博网络最大的不同点在于是单向传递。只有用户之间相互关注，他们之间的消息传递才是双向传递。例如图 2-2 中的新浪微博形式二中的用户 E 与用户 F，由于他们相互关注另一方，任何一方都能即时看到另一方的微博消息。

2.4 本章小结

社交网络是一种在线社会网络，而社会网络又是复杂网络中的一种。本章首先介绍了复杂网络的起源以及相关理论，重点阐述了复杂网络中的小世界效应与无标度特性。其次论述了社会网络的基本理论,包括社会网络理论中的“六度分割理论”与“150 定律”，简要阐述了社交网络的历史发展。最后做了关于新浪微博网络的描述，其中涉及到新浪微博网络的形式化界定与新浪微博中消息的传递形式，为后面的章节提供了基本概念、理论。

第三章 消息传播时间间隔分布

为了分析社交网络中用户行为时间间隔分布的规律，我们需要从新浪微博网络的海量信息中获取所需的信息和数据，并对这些原始信息进行处理。另外新浪微博用户是现实生活中有着自主意识的个人、或者一个组织，用户在微博网络上的一切行为都是受其主观意识支配的。因此研究新浪微博用户的转发行为规律，可以把人类行为动力学的理论作为理论依据。本章首先简单介绍了实验数据集获取的方法与过程。然后，简要阐述了人类行为动力学的起源与研究现状，其中重点阐述了 Barabási 等人研究人类行为时间间隔分布的方法，以及这个分布的物理意义。借鉴人类行为动力学的研究方法，本章着重统计分析了新浪微博用户的转发行为时间间隔分布。实验结果表明新浪微博用户的转发行为时间间隔分布服从幂律分布，即具有无标度特性，也为社交网络是复杂网络提供了一个实证分析。最后，分析这个时间间隔分布的物理意义以及对这个分布进行了数学拟合，我们建立了新浪微博用户转发行为的数学模型，为下面章节的用户影响力算法提供了重要参数。

3.1 数据集的获取

数据挖掘（Data Mining）是从海量的、模糊的、杂乱的、随机的实际应用数据中，提取隐含在其中的、潜在有用的数据信息或者规律等的过程[50]。数据挖掘的工作内容有分类分析、关联分析、聚类分析、异常分析等。

目前获得微博数据的方法主要有两种，一种是通过网络爬虫程序去读取 Web 页面的微博消息。另外一种是通过微博网络官方的 API 获取用户微博数据。通过网络爬虫的方式是指通过程序模拟用户登陆页面的操作，直接访问 Web 页面，获得 HTML 文本文件。将 HTML 文本读到内存，然后通过正则表达式来进行信息抽取，获得指定的数据。爬虫的基本原理如下：从一个指定的 URL 出发，访问该 URL 指向的页面，读取文本数据以及该页面所包含的 URL 集。抽取所关心的信息数据，再根据广度优先搜索或者深度优先搜索的方式访问下一个 URL。最后爬虫根据某种标准停止运行。如图 3-1 所示：

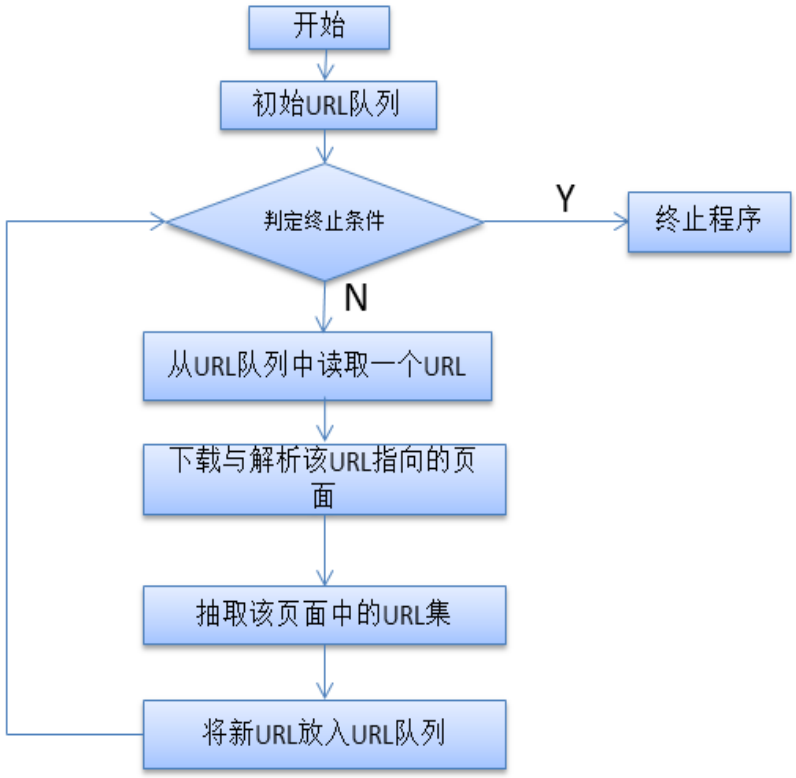

图 3-1 网络爬虫流程图

Fig.3.1 Flow diagram of Web crawler

通过 API 方式获取主要是程序调用官方的 API，会返回 API 相应的官方信息数据.然后根据需求,我们再次抽取信息来获取数据。与 Web 爬虫的方式相比，调用 API 方式的优点在于程序实现简单、返回值内容比较少，处理起来方便与快速。其劣势在于 API 接口调用次数受到户门网站的限制，以及调用 API 的返回值是官方指定的数据，这些数据不一定是开发者所需要的全部信息数据。比较两种获取信息数据方法之后，本文采用调用 API 方式来获取新浪微博网络好友信息与微博信息，主要是基于以下理由：

- 经过观察新浪微博开发的 API 返回值中信息数据有本文所需要的信息数据。即本文需求的信息数据均可从调用 API 获得。
- 新浪站点对新浪微博网络做了防爬虫机制，即每个 IP 地址在一个小时内访问微博页面的次数受到限制。
- 分别使用了两种方式进行获取消息，我们观察到一个小时内，这两种方式获得数据量差不多。而通过 API 方式获取的信息数据比较“干净”，在本地处理起来

更加快捷。

所以最后我们采用调用 API 的方式来获取实验数据集。

基于广度优先搜索原则，本文通过调用 API 方式、多节点获取新浪微博用户的好友列表信息与微博消息列表，将获得的信息集中保存在本地服务器机器上。具体流程如下：从一个特定微博用户出发，获取其粉丝列表，然后将其粉丝列表作为下一次的搜索对象，搜索其粉丝的粉丝列表，如此地一层层获取用户关系，直到数据集满足预期要求。

我们选取 MySQL 作为后台数据库。根据实验要求后续章节的需求，对 API 返回信息数据做了进一步的处理，只存储了用户的微博信息列表与粉丝关系列表。在新浪微博服务中，每个注册账号都有全网络唯一的 ID,每条微博信息都有一个唯一的 ID。因此我们设计的 MySQL 数据库结构图，如图 3-2：

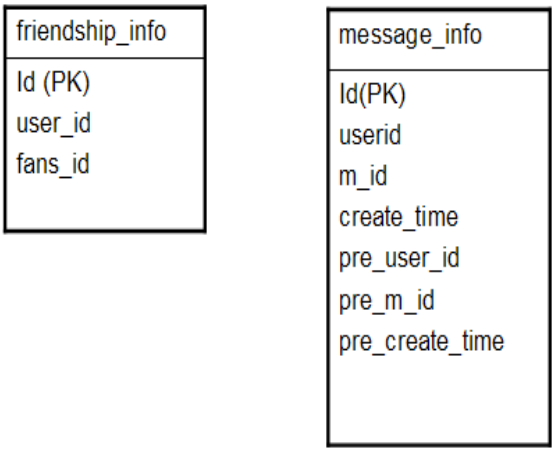

图 3-2 MySQL 数据库表结构

Fig.3-2 table structure of MySQL database

其中，表 friendship_info 存储用户的好友关系，表中 userid 表示微博用户的 ID，fans_id 表示用户粉丝的用户 ID。表 message_info 中存储了表 friendship_info 用户的微博消息，userid 是每条微博消息的发表者 ID，m_id 是每条微博消息的 ID，create_time 是每个 m_id 的创建时间，如果某条微博消息是转发了其他用户的微博消息，那么我们还要记录原始微博消息的 user_id、m_id 和创建时间 create_time。表中 pre_user_id、pre_user_id 和 pre_create_time 存储了原创消息的用户 ID、消息 ID 与创建时间。

截止到 2012 年 9 月份，新浪微博 API 有两个版本。其中第一版的 API 调用限制比较宽松，调用次数为 1000 次/小时，因此我们采用第一版本的 API 来获取我们需求的数据

信息。采用广度优先原则，我们从姚晨微博节点开始获取粉丝列表，并将用户的粉丝数据压入表 friendship_info 中，之后把表 friendship_info 中的 fans_id 作为下一个获取信息的起始节点。表 message_info 存放了表 friendship_info 所有 user_id 与 fans_id 的微博信息列表。本文在获取用户好友信息与用户微博信息的流程图如图 3-3、图 3-4 所示：

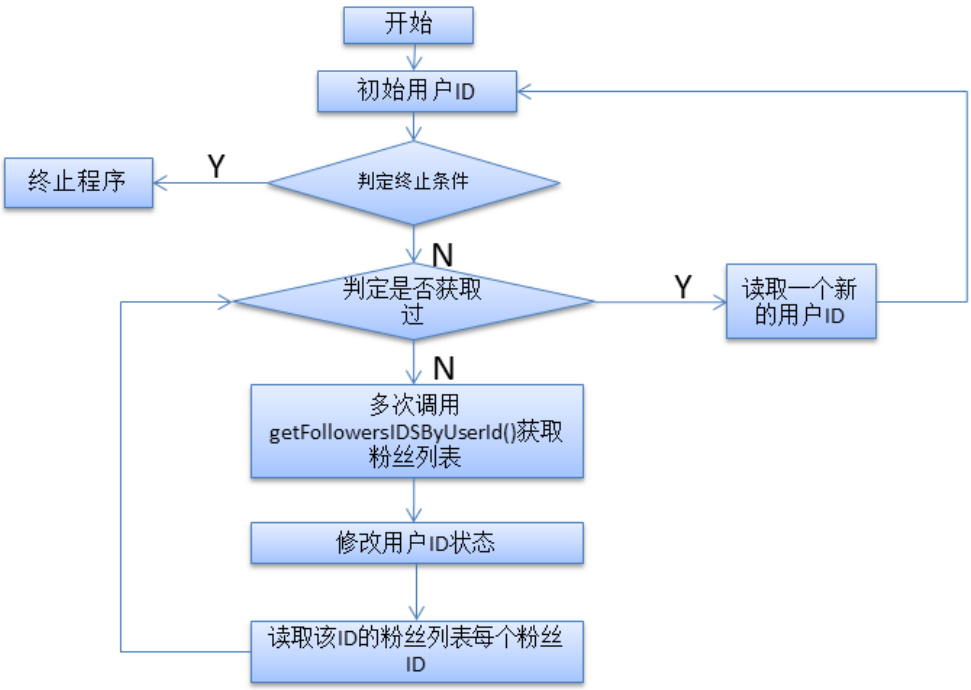

图 3-3 获取用户好友关系流程图

Fig.3-3 Flow diagram of getting friends relationship

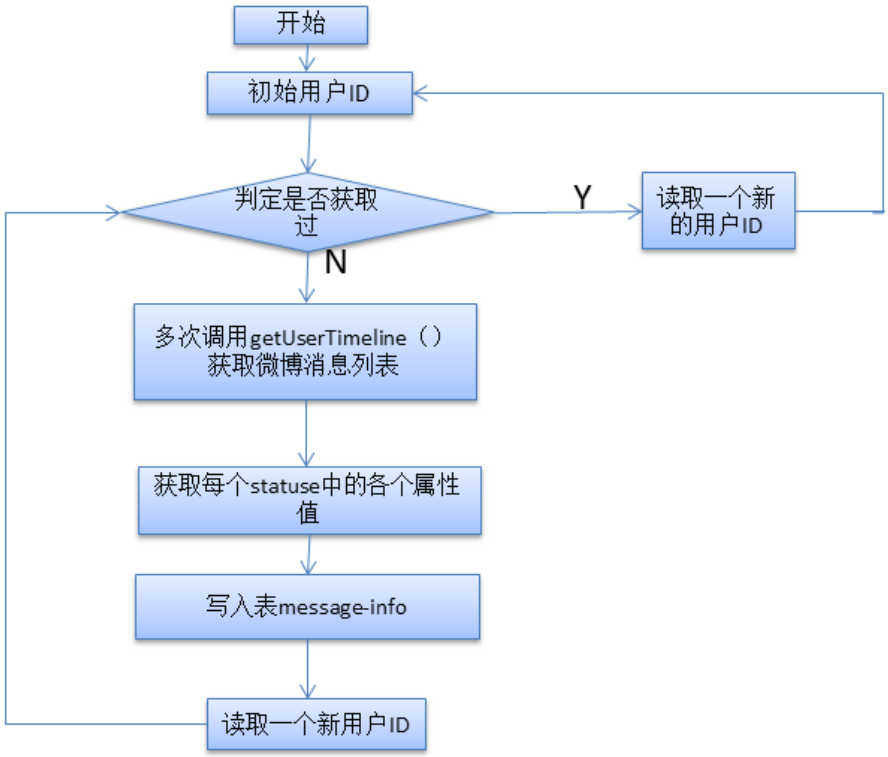

图 3-4 获取用户微博消息流程图

Fig.3-4 Flow diagram of getting users' messages list

根据图 3-3、3-4 中的流程图，我们从 2012 年一月份开始抓取新浪微博数据，到 2012 年八月份共获取了近六千万条用户微博消息。由于调用 API 的返回值中含有一些的重复数据，我们进一步对 friendship_info、message_info 两张表中数据进行处理与统计分析，得出最终采集到的数据集情况如表 3-1：

表 3-1 获取新浪微博数据集

Table.3-1 The information of data set

微博用户数量(ID 数)	170, 612
好友关系数	13, 797, 528
用户微博消息数量	59, 697, 234

3.2 人类行为动力学

人是构成现实社会的基本单元,人类行为无时无刻的影响着社会的发展与进步,研究人类行为有着重大的社会意义，吸引着各个领域的学者们。研究人类行为的学科称为人类行为动力学。人类行为动力学，这个词最早在 1890 年被 AlfredEspinas 提出的。人类动力学主要研究：人类和其他生物的各种个体和群体行为的统计特征，关注各种行为的时间、空间和强度统计特性以及不同/相同行为之间的相关性。

这一百多年来，对于人类行为的研究一直吸引着社会学、心理学等领域的学者。然而人类行为自身具有复杂性和多样性等特点，研究人类行为是科研界的一个难题与挑战。早期的金融、社会学方面的研究中，往往把人的行为假设成可以用泊松分布的稳态随机过程来描述。这假设导致了一个推论：人的行为的时间统计特征是较为均匀的，并且不可能存在两个时间间隔很大的相继行为[52]。

但是 A.L.Barabási[5-7]通过对于 E-mail 以及爱因斯坦、达尔文等名人邮件的发送与回复的时间间隔的统计分析，研究表明人类行为的时间间隔分布服从幂律分布，而不是之前研究中假设的泊松分布。就此研究成果，Barabási 于 2005 年在《Nature》发表一篇论文，该论文中统计分析了达尔文的 7591 份信与爱因斯坦的 14500 封信的回复情况，统计情况如下图 3-5：

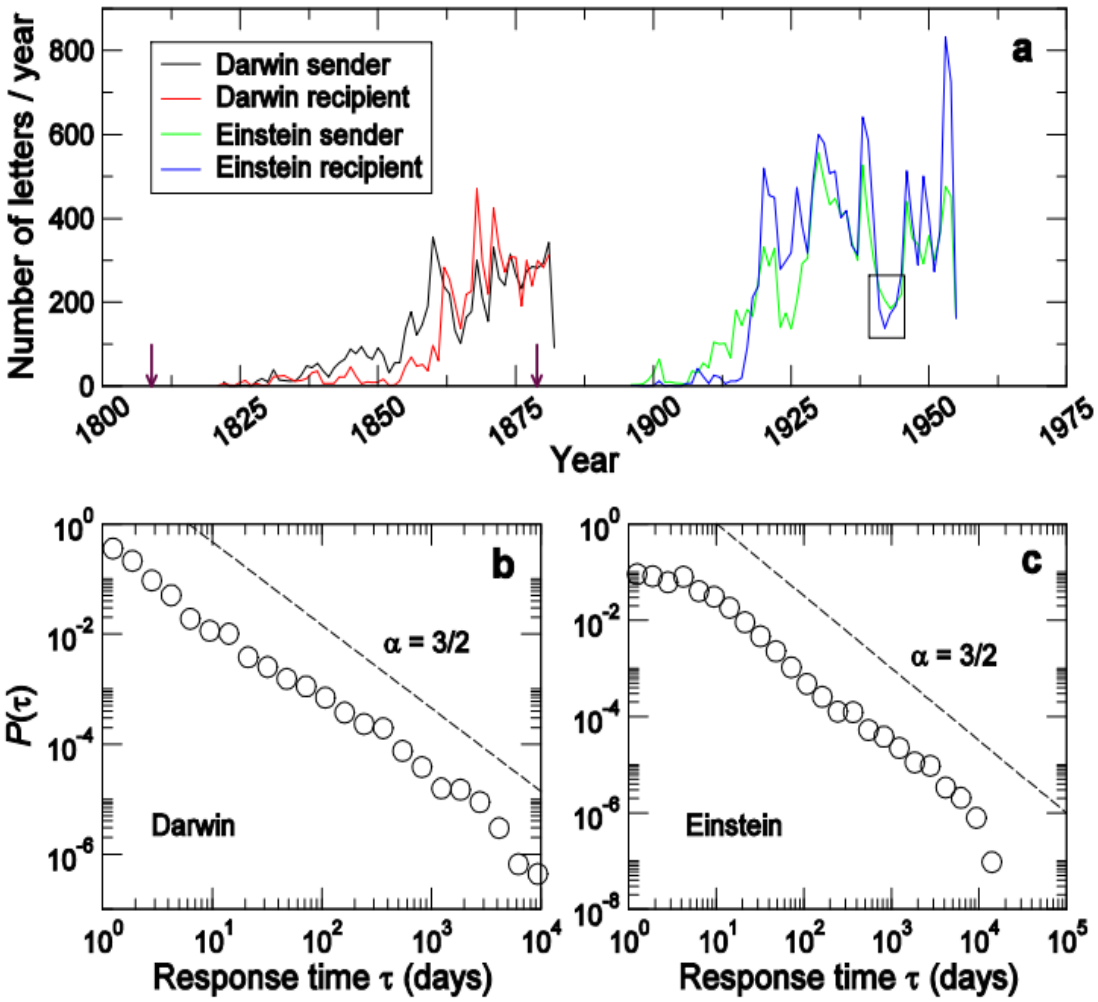

图 3-5 达尔文、爱迪生的通信模式[6]

Fig.3-5 The correspondence patterns of Darwin and Einstein[6]

图 3-5 中，a 图分别统计了达尔文与爱因斯坦信件数量的时间分布，图 b 与 c 是这两位学者回复信件的时间间隔分布，横坐标是时间间隔，指从他们收到一封信到他们回复这封信之间的时间长度，纵坐标是比例，即在时间间隔（τ）回复的信件数量，占总回复信件数量的比例。Barabási 等人拟合了两个分布，惊奇的发现虽然达尔文与爱因斯坦生活在不同年代、不同地点、生活方式也不同，但是他们的回信行为时间间隔分布都是服从幂律分布（$P(\tau) \sim \tau^{-\alpha}$）。这个幂律分布表明了达尔文等人收到一封信件，这封信会近似以 $P(\tau)$ 概率在第 τ 天被回复。之后 Barabási 等人又在不同领域做了很多研究，如电子邮件等，研究发现人类的许多行为时间间隔分布服从幂律分布，但是幂律指数（α）不相同。幂律分布表明了人类的行为在前期具有爆发特性，之后会呈现严重的“胖尾”现象。

Barabási 等人认为人类行为的这种幂律分布现象的原因在于当人们同时面临多个任务的时候，人们会根据任务优先级来处理这些任务而不是根据任务的时间顺序。就此，Barabási 等人开创了“人类动力学”新的研究方向[5-7]。

由于 Barabási 的研究结论与传统的泊松分布假设之间存在反差，学者们好奇是不是这种幂律分布现象也普遍存在于其它人类行为中。另外，互联网技术的发展与各种社交网络的掘起，给研究人员带来海量的用户数据。这些数据为学者们研究人类行为提供足够的样本。已有的研究实例也证明了许多人类行为的时间间隔分布是符合幂律分布而不是泊松分布，如电影点播[53]，网络音乐欣赏[54]，手机通讯，论坛发表评论等。下面是香港城市大学祝建华教授 2009 年的一个演讲稿 Global Regularity and Individual Variability in Dynamic Behaviors of Human Communication，这个演讲稿中例举出不同领域中，学者们已经发现的人类行为幂律分布(Power law)现象.如图 3-6：

More Empirical Tests of Human Dynamics across Diverse Behavioral Domains

Author	Year	Subject	Data Source	Exponents (γ)
Mobile Phone Usage				
Gonzalez et al	2008	frequency of visiting different locations	100K anonymous mobile phone users	-1
Candia et al	2008	Interevent time distribution for mobile phone calling activity.	mobile phone calling record in US	-0.9
Hong et al	2009	interevent time distribution between two consecutive SMS	6 person's calling record	-2.1 to -1.5
		interevent time distribution between consecutive conversations	4 individuals' calling record	-1.65 to -1.25
Video-on-Demand				
Crane & Sornette	2008	waiting time between two consecutive viewing behaviors	5 mil time-series of activities on Youtube collected over 8 months	-1.4
Webpage Browsing				
Huberman et al	1998	clicks to each website	Web users at an university for 3 weeks	-1.5
Dezso et al	2006	time interval between consecutive HTML requests by the same visitor	log files of the largest Hungarian news and entertainment portal	-1.2
		visitation pattern of news documents		-0.3
Geczy et al	2008	use of web services in intranet	intranet web log data of an organization	Evident long tail
Goncalves & Ramasco	2008	Distribution of times between consecutive clicks by the same user to the same URL	Logs of the Web server of Emory University.	-1
		Distribution of times between consecutive clicks by the same user to the Emory domain		-1.25
Social Network Sites				
Grabowski & Kosinski	2008	Number of days since the time of an individual was added (invited to the network) to the date of last logging)	a large social network of an Internet community (Grono)	-0.6
Online Games				
Grabowski & Kruszewska	2007	N of individuals spending time T playing game	30K individuals in a virtual world of MMORPG Games	-1
		N of individuals whose activity lasted T days		-1

图 3-6 人类行为的幂律分布研究

Fig.3-6 the Power Law about Human Behavior

图 3-6，表明了已经研究的一些人类行为大体是服从幂律分布，但是幂律指数不一定是相同的，如其中浏览新闻行为的幂律指数是 0.3，而点击网页行为的幂律指数是

1.5。这个现象也与 Barabási 的结论相符，即人类行为的时间间隔分布是服从幂律分布，但是不同行为的幂律指数可以不相同。

3.3 用户行为研究分析与模型

研究社交网络的用户行为一直是社交网络研究领域的一个热点。社交网络拥有海量的用户数据，可以为研究结果带来更高的准确度，所以一直吸引着人类动力学、复杂网络领域等的学者去挖掘用户行为背后的规律与商业价值，这对于理解人类行为有着非常重大的意义。学者们已经研究的网络用户行为有电影点击、浏览网页、QQ 聊天与发送邮件等。新浪微博作为一个著名的社交网络平台，同样它拥有着海量的用户与用户信息。所以研究分析新浪微博用户的行为，可以帮助我们更好的理解人类行为、理解信息传播以及为其它研究提供理论依据等。本小节主要是借鉴人类动力学的研究方法，分析了新浪微博上用户转发行为时间间隔分布，建立了用户之间关注度的数学模型。

3.3.1 微博用户转发行为分析

社交网络中，消息的传播主要是通过对其评论与转发两种方式。当用户对某条消息的评论时候，该条消息并不会被下一跳用户节点看到，即消息的传播范围仅限于当前用户的粉丝圈子里，浏览行为主要体现了用户之间的交流。但是如果一条消息被用户转发之后，它的内容就会被下一跳的用户节点获悉，如此下去，这条消息可能被网络上所有的用户看到。我们知道一个用户拥有多个粉丝，通过转发方式的消息传播速度呈现几何指数增长,可以看出用户转发行为是消息传播的一个重要途径，因此研究用户的转发行为对于消息传播有着重大的意义。

如 3.2 的论述中，人类在网络上的一些行为也是服从幂律分布。微博网络用户对某个用户的微博消息进行转发，这个行为也是人类在网络上的行为，即它是人类的一种主观行为。所以我们借鉴 Barabási 等人的研究方法，研究新浪微博网络中用户的转发行为。由于本文是研究一个用户对其某个特定关注者的微博消息转发情况，所以在研究分析过程中以网络中每个关注关系作为研究基本单位，以每个关注关系建立的时间长度作为时间间隔，研究分析粉丝转发的微博在时间间隔上的分布。我们定义了如下公式 3-1：

$$W_{(\upsilon,u)}(\tau)=\frac{N_{(\upsilon,u)}(\tau)}{N_{(\upsilon,u)}} \qquad (3-1)$$

式中τ ：用户υ关注了用户u多长时间，即τ是个时间间隔；$N_{(\upsilon,u)}(\tau)$：用户υ关

注了用户 u 的第 τ 周/天内，用户 υ 转发用户 u 微博消息的次数； $N_{(\upsilon,u)}$：用户 υ 转发用户 u 微博消息的总次数。

如果粉丝对于用户的关注程度基本不变，那么粉丝会在不同转发其微博数量基本不变，则 $W_{(\upsilon,u)}(\tau)$ 应该是个常量或者一个围绕着一个值小幅度波动。根据公式 3-1，本文统计分析了实验数据集中每一个关注关系的转发情况，实验结果表明新浪微博中用户转发行为时间间隔服从幂律分布，如图 3-7：

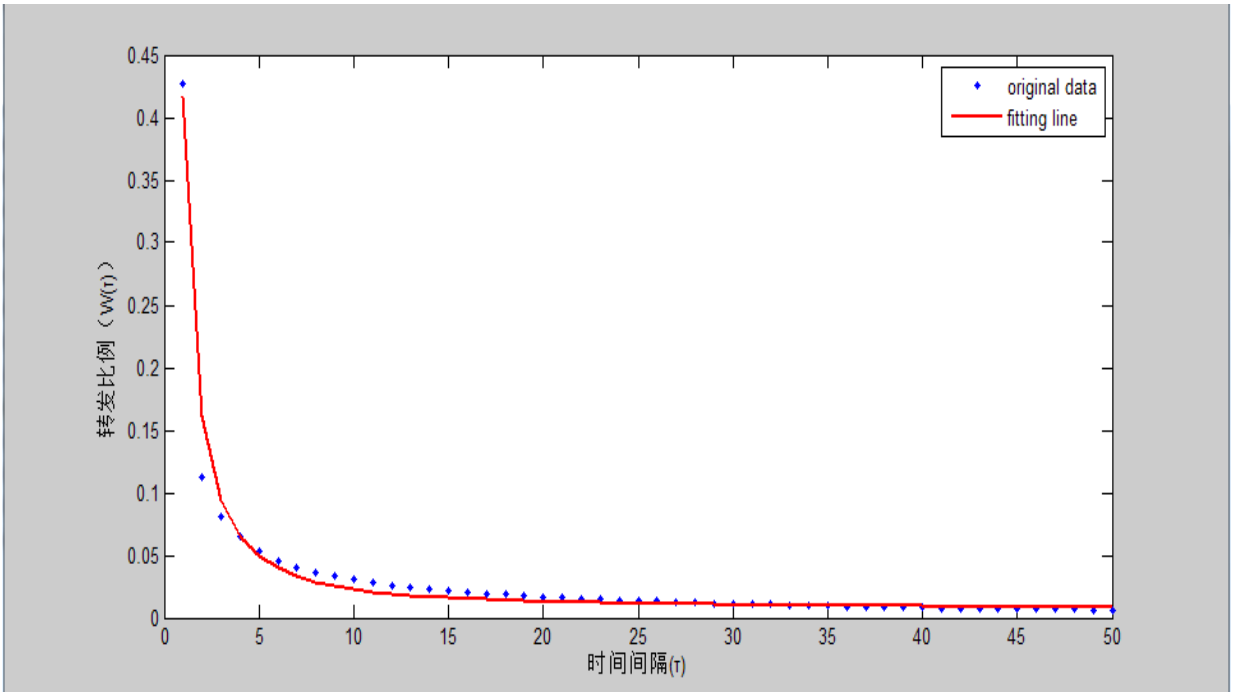

图 3-7 用户转发情况的时间间隔分布

Fig.3-7 Distribution of time intervals about retweet

图 3-7 中，横坐标是每个好友关系建立之后的时间间隔，单位为周；纵坐标是单位时间段内转发微博次数与总转发次数的比值（$W_{(\upsilon,u)}(\tau)$），

图 3-7 表明，微博用户转发某个特定关注者的微博消息，前期转发的数量占了总数量的很大比例，之后转发数量占总数量很小的比例，换而言之，转发行为主要爆发在用户建立关系的前期，之后显示非常严重的“胖尾”现象。即用户在关注某个用户的前一段时间，他会常常转发该关注者的微博消息，之后这种转发行为会逐渐衰减或者保持一个稳定的状态。经过拟合，我们发现这个时间间隔分布服从幂律分布（$P(\tau) \sim \tau^{-\alpha}$）。

3.3.2 用户之间的关注模型

在新浪微博网络中，一个用户可以同时关注多个微博用户。由于与关注者的关系、兴趣爱好等原因，微博用户不可能对其所有关注者的即时消息都一视同仁对待，即一个用户不可能看到其关注者的微博消息就转发，也不可能对其微博消息都不转发。所以用户对其不同关注者微博消息的转发概率是因人而异的。例如在社交网络中，用户更喜欢与亲昵朋友互动交流，对亲昵朋友的微博消息更加关注或者对于新关注者的消息更加关注等等。而只有用户对于其关注者的微博消息进行转发，微博消息才能沿着网络结构传播下去。因此某个用户对其特定关注者的关注程度，体现在用户对关注者微博消息的转发行为，我们在研究用户之间关注度模型是以用户之间转发行为的规律作为理论基础。

本文定义用户对其关注者微博消息的关注程度称为关注度。本文主要研究是微博消息传播，因此一个用户对其关注者的某时刻关注度，体现在这个时刻该用户是否转发该关注者的微博消息的概率。借鉴社会网络的形式化界定方法，我们定义关注度为用户节点对其关注者节点消息的转发概率。关注度越大，用户转发关注者的微博消息概率越高，则表明该用户经常关注这个关注者的 消息或者与这个关注者有着亲密关系、共同的兴趣爱好、共同话题等。

用户对某个用户微博消息的关注，只有在他们建立关注关系之后才会存在的。因此我们把用户之间建立关注关系的初时刻，作为研究用户之间关注程度的起点。在 3.3.1 小节中，根据$W_{(v,u)}(\tau)$的数学表示式，$W_{(v,u)}(\tau)$表示了建立关注关系之后的第τ天，用户就已经转发了总转发微博数量的比例。换而言之，$W_{(v,u)}(\tau)$也表明了建立关注关系之后的第τ天，如果关注者发布了一条微博消息，那么用户会近似以$W_{(v,u)}(\tau)$的概率转发该条微博。因此$W_{(v,u)}(\tau)$就是我们定义的用户之间关注度的数学模型。

为了得到精确的数学表达式，我们对$W_{(v,u)}(\tau)$进行数学拟合，图 3-7 中的红线即为数学拟合之后的曲线，实验结果发现拟合的效果非常好，相关性达到 0.987，其数学表达式为$W_{(v,u)}(\tau)=a*\tau^{-\alpha}+b$，其中 a=0.4097, $\alpha=1.5\pm0.086$, b=0.007626，τ为时间间隔，即用户建立关注关系的第τ周/天。根据上述分析，$W_{(v,u)}(\tau)$的物理意义为：用户 v 关注了用户u的第τ天，用户 v 会以$W_{(v,u)}(\tau)$的概率转发用户u的微博消息。从$W_{(v,u)}(\tau)$的拟

合效果来看，用户对一个特定关注者的关注程度不会一直不变的，会显现一个快速衰减的状态，最后是一个很弱的关注程度。这个现象也是符合社会网络的“150 定律”理论，即同一时间内，人们会认识很多人，但是保持强联系的人数一般不超过 150。

3.4 本章小结

本章首先介绍了社交网络中获取信息的方法，通过比较两种获取数据的方式，本文最后决定采用 API 方式来采集所需的实验数据集；其次综述了人类行为动力学的起源与社交网络中人类行为的幂律分布现象，其中重点阐述了 Barabási 的研究方法。文章阐述了转发行为对于消息传播的意义，借鉴 Barabási 的研究方法，统计分析了微博用户的转发行为时间间隔分布。我们发现新浪微博用户转发行为的时间间隔分布是服从幂律分布，即用户转发行为具有前期的爆发性与后期的严重“胖尾”现象；最后我们定义了用户之间关注度，分析关注度与转发行为时间间隔分布之间的关系，定义了关注度的数学模型，这个研究结果对于文章后面的影响力算法研究有着重大意义，本章中的关注度也是新算法的重要一部分。

第四章 用户影响力算法（MURank）

第三章研究表明新浪微博网络中好友关系结构类似于 Web 页面的链接结构。我们基于 PageRank 算法的基本思想与用户之间的关注度模型，提出了一种评价新浪微博网络中节点影响力的算法 MURank（Micro_blogging User Rank），该算法通过反复迭代计算出每个用户的 MURank 值，从而找出最具影响力的用户。由于 MURank 算法需要计算的用户数量很大，本文采用了多线程机制来实现 MURank 算法的计算，提高运算速度。最后本章详细阐述了 MURank 算法的运行流程与收敛条件。

4.1 社交网络影响力的概念

社会网络的信息传播与影响力的研究已经存在于各个领域，如市场营销、同性、社会学、政治学等。在传播学理论中认为人际传播网络存在一种人，这种人经常为他人提供建议或者见意等，他的想法、信息传播给其他人或者影响到其他人的行为、思想等。通过这个人人之间相传的传播方式，信息传播的成本远远小于其它方式的传播，并且传播速度也比其它方式快速。传播学中称这种人对信息的效应为影响力。

SNS 类的社交网络是基于现实社会人际关系网络组成，社交网络用户是现实世界中的个体，所以社交网络用户的影响力类似于现实世界的个人影响力。现实世界中的个人影响力不仅仅与其社会地位、教育背景等因素有关，其中包含了许多主观因素，比在线社交网络更为复杂。而在线社交网络中，我们可以提取用户的相关属性值来定量衡量一个用户在网络中影响力，对用户潜在的传播学价值与商业价值进行评估。

在社交网络中，如人人网、新浪微博等，一个用户对另外一个用户的影响力只能通过他们之间的关注关系，以及其他用户对其发布的信息内容的感兴趣程度等。一个用户对信息的处理动作主要有转发、浏览、评价与原创等动作，其中对信息传播起到作用，只有用户的转发行为。当用户的微博消息被转发之后，这条微博消息会沿着网络结构传播下去，被更多的网络用户看到。我们提出用户影响力算法的目的在于快速地找出影响力高的用户，便于我们理解信息传播模型等，而用户的转发行为是信息传播过程中的一个重要动作，因此我们把用户之间的转发行为作为用户影响力算法的一个重要因素。

前面章节的讨论，表明了新浪微博的网络结构类似于 Web 页面的网络结构。评估 Web 页面的权威性或者影响力的算法有 PageRank 算法、HITS 算法等。其中 PageRank 算法是

著名 Google 的创始人提出的，之后它吸引了许多学者对它的研究，学者们提出了许多基于 PageRank 的改进算法。衡量用户在网络中的影响力就是评估用户在网络中的权威性，每个用户对应于一个 Web 页面，而 PageRank 算法是评估 Web 页面在网络中的权威性，并且它是非常经典的评估 Web 页面权威性的算法。因此，我们基于 PageRank 算法的基本思想，提出适用于评估社交网络用户影响力的算法。

4.2 PageRank 算法的基本思想

PageRank 算法[4]是由拉里·佩奇和谢尔盖·布林于 1998 年提出并发表，该算法是个非常经典的 Web 页面排名算法。凭借着 PageRank 算法的基本思想与成功的商业经营，Google 已经成为全球非常优秀的互联网企业。之后，许多学者提出了许多 PageRank 改进算法。PageRank 算法的思想主要是基于网络结构的分析，Web 页面之间的链接关系构成了一个网络生态系统，每个页面都有自己的职能、地位等。

PageRank 算法的基本思想：它认为当某个页面里有个链接（URL），那么该页面就对这个 URL 指向页面的内容认可或者肯定，那么这个 URL 页面内容的权威性会有所增加。例如网页 A 上有网页 B 的链接，那么网页 A 认为网页 B 的内容是有价值。由此推出，如果有许多网页中有网页 B 的链接（URL），那么表明网页 B 的内容被许多个网页认可，表明了页面 B 的内容具有很高的价值，PageRank 算法认为网页 B 内容具有权威性或者极大价值。PageRank 算法还有一个优点在于，它不仅仅统计网页被链接的数量，还衡量了链接该网页的 URL 本身权值。如果一个网页被权威很高的网页指向，比被一般网页指向，它得到更高的权值。这点类似于社交网络分析中等级权威的观念。在 PageRank 算法中，表示每个网页权威性的值称为 PR 值，每个网页的 PR 值不仅仅取决于被链接网页的数量，还受到指向该网页的 URL 的质量和重要程度影响。PageRank 算法认为每个网页的 PR 值都被均匀的分配到它指向的网页，如此迭代下来，网络中每个页面的 PR 值达到稳定、收敛状态。PR 值的分配过程如下图 4-1：

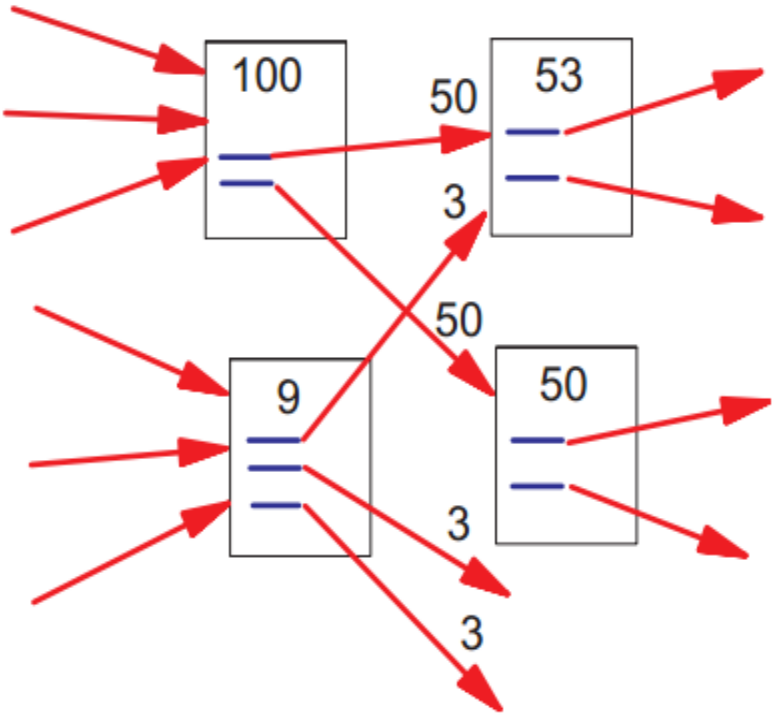

图 4-1 PR 值分配示意图[4]

Fig.4-1 Simplified PageRank Calculation[4]

PR 值为 100 的页面中含有两个链接，于是它将自身的 PR 值平均分配给这两个链出页面，对应的，这两个链出页面都得到了 50 的 PR 值，链出的页面得到的 PR 值与其自身的 PR 值进行累加后继续分配给这个页面的链出页面，如此 PR 值一层一层的传递下去，使得全网络总的页面都得到一个稳定的 PR 值。最后根据页面各自的 PR 值，我们可以很快找出 PR 值高的页面。上述 PR 值分配过程的数学表达式，如公式 4-1：

$$PR(v) = c \sum_{u \in U(v)} PR(u) / N(u) \qquad (4-1)$$

式中 PR(v)是网页 v 的 PR 值；u 是指向网页 v 的网页，U(v)是链接到网页 v 的网页集合；$N(u)$：网页 u 里的所有链接数。

根据公式 4-1，网络中 Web 页面都经过不断的迭代，直到每个 Web 页面的 PR 值都达到一个收敛、稳定的状态，那么就停止迭代，最后的 Web 页面得到的 PR 值即为页面的最终影响力值。其示意图如图 4-2：

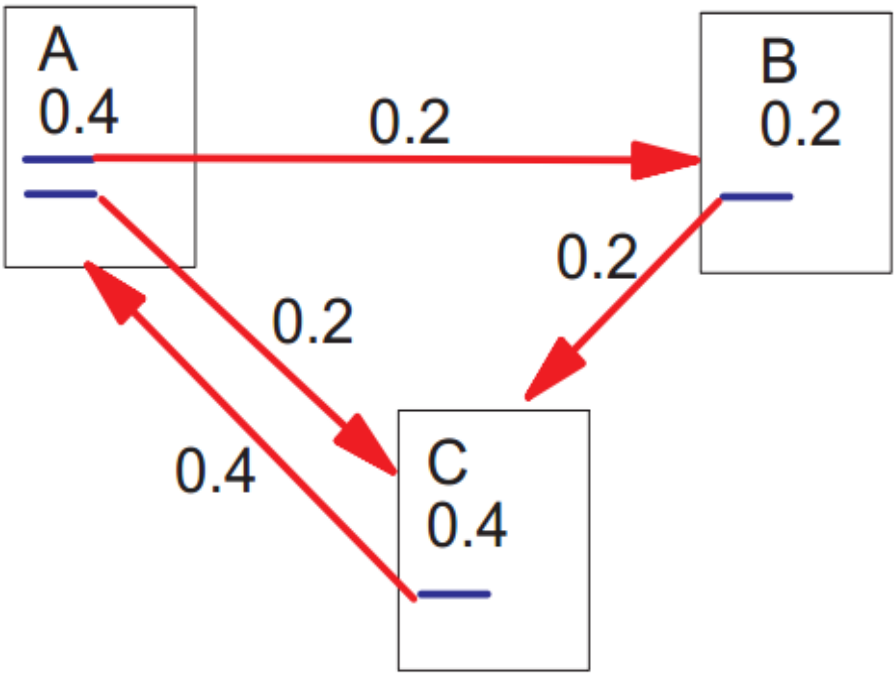

图 4-2 PageRank 算法的简单计算示意图[4]
Fig.4-2 Simplified PageRank Calulation[4]

在图 4-2 中，网页 A、B 与 C 经过不断的迭代，其 PR 值分别是 0.4、0.2 与 0.4。之后，即使再多次迭代，但是他们的 PR 值不再发生变化，因此他们最终的 PR 值分别是 0.4、0.2 与 0.4。

拉里·佩奇等人发现在 web 页面结构中会存在一个现象：有些页面自成一个环，这些页面没有指向外部的链接，只有被环外的页面指向。那么根据公式 4-1，这些环里网页的权值随着不断迭代而渐渐的在环中消耗掉，最后这个环中每个网页的 PageRank 值都将全为 0。显然网页的 PR 值为 0 是符合实际情况的，拉里·佩奇等人称这种现象为 Rank Sink 现象,如图 4-3 所示：

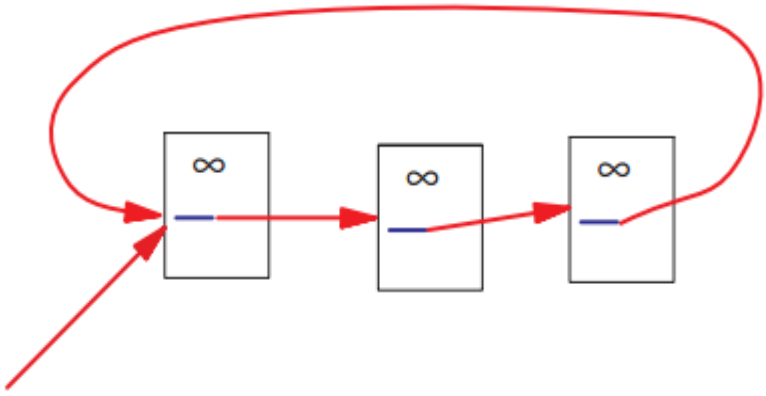

图 4-3 Rank Sink 现象[4]

Fig.4-3 Loop Which Acts as a Rank Sink[4]

为了解决 Rank Sink 现象，拉里·佩奇等人认为每个页面都拥有一个初始 PR 值，这个初始 PR 值是页面会被随机访问的概率。所以拉里·佩奇等人将公式（4-1）修改成如下公式 4-2：

$$PR(v)=(1-d)+d\sum_{u\in U(v)}PR(u)/N(u) \qquad (4-2)$$

其中 d：阻尼系数，表示网络用户在浏览某个 Web 页面之后以 d 概率继续浏览页面中的某个链出的 Web 页面。拉里·佩奇等人经过多次实验发现 d 取 0.85 的时候，实验效果最好。加入阻尼系数既可以解决了 Rank Sink 现象，也可以保证了 PageRank 算法运算收敛。传统的 PageRank 算法只考虑了网页的链接结构，而没有考虑网页内容的权威性，这样子容易造成主题偏移问题。例如一个全新网页没有内容，但是有个 PR 值很高的网页指向它，那么这个全新网页的 PR 值也就变得很高，显然这个全新网页虽然没有内容但是拥有很高的 PR 值，这种情况是显而易见的不合理。针对 PageRank 算法的这个问题，许多学者提出了基于 PageRank 算法的改进算法，比如加入时间因素、话题因素等。

在互联网网页链接结构中，存在着链入关系与链出关系两种。链入关系指的是一个网页被其它网页指向的关系，链出关系指的是一个页面里有其它网页的 URL。在微博网络中，每个用户都有其关注列表与粉丝列表，其中关注列表是指该用户关注了哪些人，在关注列表中有着这些关注者的主页 URL，粉丝列表是指该用户被多少人关注，即这些人都拥有他的主页链接。在社交网络结构中，这些粉丝关系称为出度，关注关系称为入度。考虑新浪微博用户影响力算法的时候，我们提出基于 PageRank 算法思想的理论依据如下：

- 每个新浪微博用户都拥有关注列表与粉丝列表，其中关注列表类似于 Web 网页

中的链入列表，粉丝列表类似于 Web 网页的链出列表，因此新浪微博网络结构类似于 Web 页面的网络结构。

- 评价一个用户在网络中的影响力，本质上就是评估该用户在网络中的排名。而 PageRank 是个非常经典的 Web 页面排名算法，并且基于 PageRank 算法的 Web 搜索算法已经取得了非常大的成功。

因此，我们在计算每个新浪微博用户影响力的时候，就是计算该用户的类似 PR 值，来量化每个新浪微博用户的权威性。通过对新浪微博用户的权威性排名，我们可以较快的找到权威新浪微博用户，来达到信息预测等。

4.3 基于 PageRank 的 MURank 算法

在新浪微博网络中，微博用户的基本动作主要有浏览信息、发布信息与评论其他人的微博消息，其中发布信息有两种形式：1、自己原创微博消息，消息内容包括某时刻的心情、路上见到的事物等；2、看到自己关注者的微博消息，认为这个消息有价值、有趣等，对这个消息进行转发。在 2.3.2 小节中，我们讨论了新浪微博网络中微博消息传递形式，发现新浪微博信息是单向传递。

如果一个用户的消息只是被粉丝浏览，那么该条消息的传播范围仅仅限于该用户的粉丝圈子里，而一个用户的粉丝数量是有限的。但是如果这条微博消息被所有粉丝转发之后，粉丝的所有粉丝也转发了这条微博，如此一层层的转发下去，根据“六度分割理论”理论，这条微博消息最多被转发六层，那么微博网络上绝大部分的人都可以看到这条微博。因此用户的转发微博行为，对于微博消息传播有着至关重要的意义。

研究用户影响力就是研究用户对其周围用户的影响程度。如果一个用户对周围用户影响程度越大，那么他的影响力越大，即很多人认可或者采纳了他的观点与意见等。在传播学研究领域中，这种用户被称为意见领袖。意见领袖[56]（opinion leaders）是拉扎斯菲尔德等最早在《人民的选择》中提出的概念，在《个人影响》一书中做了进一步阐释。指在人际传播网络中，经常为他人提供意见或建议等并影响到他人的行为、观点等的人物。在新浪微博网络中，如果用户的微博消息被粉丝转发，那么粉丝认为这条微博是有价值，间接的认可了该用户的见识等。因此如果某用户的微博消息被许多人转发，那么该用户就是一个意见领袖，进而这个用户具有极高的权威性。而现实生活中并不是所有粉丝都以一个恒定的概率转发用户的微博消息。我们可以得出结论：衡量用户权威性的一个重要因素是其粉丝是否转发他的微博消息。所以这个转发概率是用户影响力算法中的一部分。

在第三章中，我们讨论与研究了用户之间的转发行为时间间隔分布，发现该分布是服从幂律分布，定义了关注度的概念和分析了这个分布的物理意义。以这个分布为基础，我们定义用户之间关注度的数学模型，即这个分布的数学拟合表达式。在本章中，基于 PageRank 算法的思想与用户之间的关注度数学模型，我们分析了微博网络节点结构，提出了一种实时性的用户影响力算法，名为 MURank（Micro_blogging User Rank，MURank）算法。

在 PageRank 算法的基本思想中网页的 PR 值是均匀的分配给链出的网页上，这个思造成旧网页由于存在时间比较久，那么它会被链接的次数会明显高于新网页，而旧网页的信息往往是过时的、没有价值的，这种现象导致了旧网页的 PR 值高但是内容陈旧，而新网页的内容用过，但是 PR 值却很低。本文分析了用户转发行为的时间间隔分布，表明用户之间的关注程度不是一成不变的。因此把 PageRank 算法模型运用到新浪微博网络用户影响力模型中的同时，我们将用户之间的关注度引入影响力算法的模型中，认为如果一个粉丝高度关注某个用户，那么该粉丝分配给该用户相对较高的 MURank 值，反之，如果一个粉丝对某用户关注程度比较弱，那么该粉丝分配给该用户相对较少的 MURank 值。这样子，同一时刻粉丝将分配不同的 MURank 值给不同的关注者，也就体现了粉丝对不同关注者的关注程度是不一样的。这个分配机制也是符合现实生活中一个常见现象：一个人不可能对其朋友一视同仁的，总会存在对亲昵朋友相对比较好，对泛泛之交的朋友相对比较疏远。基于上述的分析与 PageRank 算法基本思想，我们提出了 MURank 算法。MURank 算法的数学表达式如公式 4-3：

$$MUR(v,t)=(1-\gamma_v(t))+\gamma_v(t)\sum_{u:(v,u)\in E}A_{(u,\upsilon)}(t)MUR(u,t) \qquad (4-3)$$

MURank 算法模型中的参数描述如下：

(1)：$\gamma_v(t)$

$\gamma_v(t)$ 为 t 时刻，用户 v 的微博消息被粉丝转发的平均概率。本文定义了这个消息转发的概率 $\gamma_v(t)$ 的数学表示式，如公式 4-4：

$$\gamma_v(t)=\frac{\sum_{u:(u,v)\in E}\omega_{(u,v)}(t)}{N_v} \qquad (4-4)$$

式中 $W_{(u,v)}(t)$：t 时刻，(u,v) 关系的关注度。N_v：t 时刻，用户 v 的粉丝数。E：用户 v

的粉丝集合。例如，在某个时刻，有许多新粉丝加一个用户为关注者，那么该用户的微博消息被新粉丝们转发的概率就会比较大，反应在公式 4-4 中，就是$\gamma_v(t)$就会相对比较大。$\gamma_v(t)$类似于 PageRank 中的阻尼系数(d)，表明了用户对其所有粉丝的影响程度。

(2):影响力分配比例 $A_{(u,\upsilon)}(t)$

$A_{(u,\upsilon)}(t)$是用户u分配给其关注者υ的 MURank 值的比例，是根据用户u对用户υ的关注度占用户u对所有关注者的关注度总和的比例。$A_{(u,\upsilon)}(t)$的数学表示式如公式 4-5 所示：

$$A_{(u,\upsilon)}(t)=\frac{\omega_{(u,\upsilon)}(t)}{\sum_{k:(u,k)\in E}\omega_{(u,k)}(t)} \qquad (4-5)$$

其中 $\omega_{(u,v)}(t)$：t 时刻，用户u对用户υ的关注度。$\omega_{(u,\upsilon)}(t)=a*(t-t_{(u,\upsilon)})^{-\alpha}+b$式中$t_{(u,\upsilon)}$：用户$u$关注用户$\upsilon$的初时刻。

便于理解 MURank 算法中用户 MURank 值的传递过程，本文做了一个简单的 MURank 值的分配比例示意图，如图 4-4 所示：

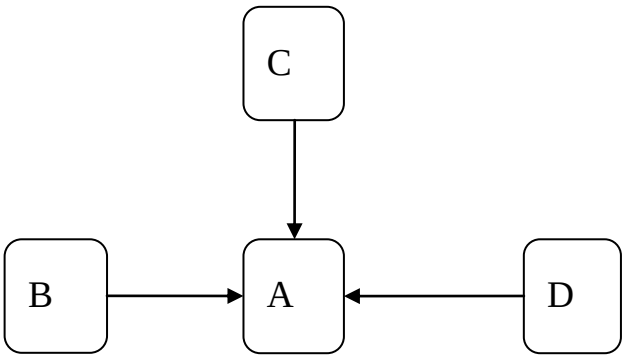

图 4-4 t 时刻，用户的 MURank 值分配示意图
Fig.4-4 time t, MURank Calculation

如图 4-4 所示，用户 A 总共关注了用户 B、C 和 D 用户。其中用户 A 分别在$t_{(A,B)}$、$t_{(A,C)}$与$t_{(A,D)}$时刻关注了用户 B、C 和 D。那么在 t 时刻，用户 A 对于这三个用户的关注度分别为$\omega_{(A,B)}(t)$、$\omega_{(A,C)}(t)$和$\omega_{(A,D)}(t)$。根据公式 4-5 可知道，t 时刻，用户 A 分配给用户 B 的 MURank 值比例是：$\frac{\omega_{(A,B)}(t)}{\omega_{(A,B)}(t)+\omega_{(A,C)}(t)+\omega_{(A,D)}(t)}$。同理，用户 C、D 得到用户 A 的 MURank 值比例分别为$\frac{\omega_{(A,C)}(t)}{\omega_{(A,B)}(t)+\omega_{(A,C)}(t)+\omega_{(A,D)}(t)}$、$\frac{\omega_{(A,D)}(t)}{\omega_{(A,B)}(t)+\omega_{(A,C)}(t)+\omega_{(A,D)}(t)}$。从上述分析，可以看出用户 A 分配给用户 B、C 和 D 的 MURank 值是相对值，而不是绝对值，这与 PageRank 算法中的 PageRank 值分配情况相似，也是保证 MURank 算法收敛性的一个重要步骤。由于$\omega_{(A,B)}(t)$、$\omega_{(A,C)}(t)$和$\omega_{(A,D)}(t)$的值都不相同，则导致了用户 B、C 与 D 将从用户 A 得到不同的 MURank 值，从而区分了用户 A 对不同用户的关注程度不同。MURank 算法的用户 MURank 值分配机制有别于传统 PageRank 算法，这点可以有效的减少类似于 PageRank 算法中的主题偏移现象。

4.4 MURank 算法的计算流程：

根据 PageRank 的权值分配原理与公式 4-3，我们可以知道 MURank 的计算过程是一个不断的迭代计算过程，直到网络中用户的 MURank 值达到一个稳定值。因此，计算 t 时刻用户 MURank 值的基本步骤如下：

1. 根据用户关注度的数学模型，计算出每个关注关系在 t 时刻的关注度。计算 t 时刻，一个用户的微博消息被其所有粉丝转发的概率。
2. 每个用户的 MURank 的初始值均设置为 1；
3. 根据公式 4-3，进行计算每个用户的新的 MURank 值
4. 将步骤 2 中的新的 MURank 值作为下次迭代的用户 MURank 初始值。
5. 如此重复步骤 2、3 的计算过程，直到两次迭代之间的每个用户 MURank 值之差的和的绝对值小于某个极小数（ε），我们就认为用户的 MURank 值达到一个稳定状态，停止 MURank 运算。
6. 退出计算流程。最后根据网络中每个用户的 MURank 值，进行从小到大的排名。

因此，MURank 算法的程序流程如下：

$$
\begin{aligned}
&for \quad each \quad (v,u):\\
&\qquad \omega_{(u,v)}(t) = a*(t-t_{(u,v)})^{-\alpha}+b\\
&\qquad A_{(u.,v)}(t) = \frac{\omega_{(u,v)}(t)}{\sum_{k:(u,k)\in E}\omega_{(u,k)}(t)}\\
&\gamma(t) = \frac{\sum_{(u,v)\in E}\omega_{(u,v)}(t)}{N_E}\\
&MUR_0(t) \leftarrow E_{n\times 1}\\
&i = 0\\
&loop:\\
&\qquad MUR_{(i+1)}(t) \leftarrow (1-\gamma(t)) + \gamma(t)*A(t)*MUR_i(t)\\
&\qquad d \leftarrow \| MUR_{(i+1)}(t) - MUR_i(t) \|_1\\
&\qquad i \leftarrow i+1\\
&while(d > \varepsilon)\\
&end
\end{aligned}
$$

本文在实验过程中ε取值为10^{-7}。

我们的实验环境如下所示：

1. 硬件平台：1 台 DELL 服务器，2.4GHz Intel E56020,32GB，1TB 硬盘，2 台 DELL PC 1.8GHz Intel，2G 内存，160G 硬盘；
2. 软件平台：Windows 7 操作系统，Eclipse 编辑器，MySQL 5.5 数据库。

运算过程中，本文采用多线程机制实现 MURank 算法的运行。具体实现如下：

1. 建立表 MURank_info，表字段有 Userid，rank0,rank1 …，将每个用户的 rank0 设置为 1，rank1，rank2 …设置为-1；
2. 每台 PC 运行三个线程，根据 Userid 的范围，平均分为六份，分别交给这六个线程计算每个用户的 MURank 值，更新用户的 MURank 值。
3. 当一个线程提前完成任务，则读取数据库中还没完成 MURank 值计算的用户 ID，读取其中 Userid 号最大值，进行计算该 Userid 号的 MURank 值。
4. 当本次迭代中每个用户都计算完成，进入下一次迭代。例如第 i 次迭代中，如

果数据库里没有用户的 ranki 是-1,则表明所有用户的本次 MURank 值计算完成，进入下一次迭代。

5. 如此不断的迭代计算，直到用户的相邻迭代 MURank 值的差的和的绝对值小于事前设定的极小数，停止 MURank 算法的运行

4.5 本章小结

本章首先阐述了传播学中关于影响力的概念,描述了社交网络中用户影响力的定义，然后阐述了 PageRank 算法的基本思想、SinkRank 现象与主题偏移的问题。基于 PageRank 算法的基本思想，本章提出了适用于新浪微博网络用户影响力算法（MURank），并将第三章中关注度数学模型引入 MURank 算法模型中，使得算法具有时间因素，并且阐述了 MURank 算法中的参数含义与数学表达式。最后详细介绍了 MURank 算法中 MURank 值的分配流程与 MURank 算法的具体程序实现流程。

第五章 实验结果与分析

根据 4.2 节所讨论的 MURank 算法模型与 4.3 节的算法实现流程，我们运行 MURank 算法评估新浪微博用户实时影响力，观察不同时刻 MURank 算法的收敛性。然后分别根据原始的 PageRank 算法、粉丝数量来评估用户的影响力，采用统计学中的相关系数来衡量这三种算法两两之间的相关性。实验结果表明，MURank 算法与另外两种算法之间相关性比较小，并且 MURank 算法更加能体现一个用户在一个时间段内的新增粉丝数量以及影响力变化情况。与根据原始 PageRank 的用户影响力算法比较，MURank 具有更好的实时性。

5.1 MURank 算法的收敛性

本文取数据集中关系初时刻的最大值（t1）为研究时刻。根据公式（4-5）与 4.3 节中的算法计算流程，计算微博用户在 t1 时刻的影响力值（MURank 值），计算并跟踪计算过程中的收敛情况，如图 5-1：

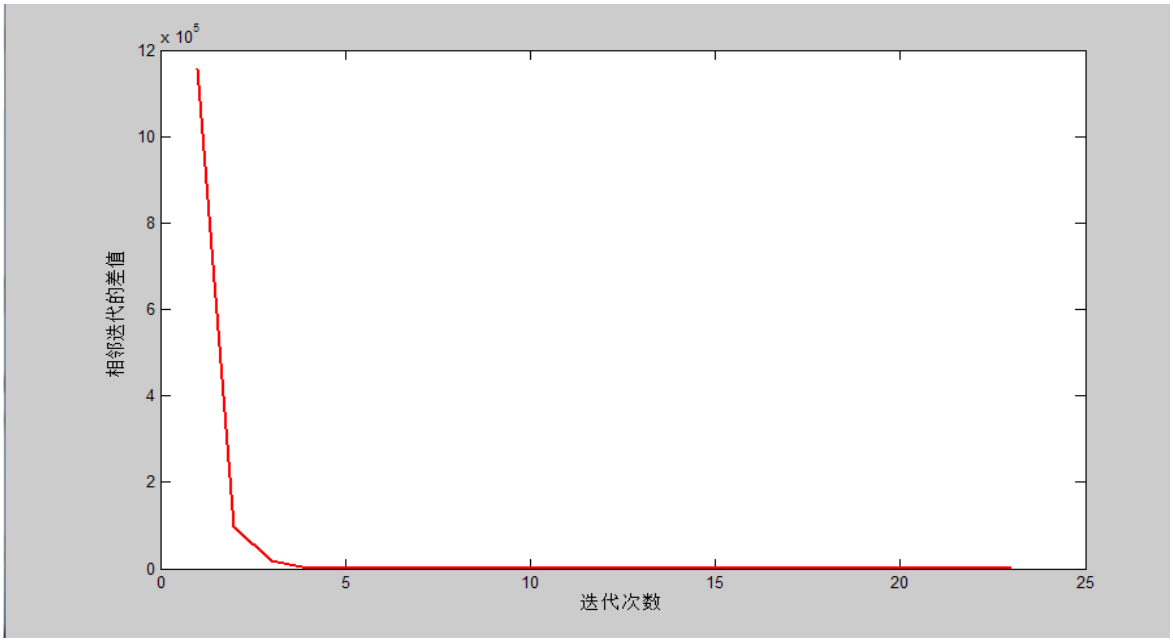

图 5-1 t1 时刻，MURank 算法收敛性

Fig.5-1 Time t1,Convergence of MURank Computation

横坐标为迭代次数，纵坐标为节点的相邻迭代的差的和。

在计算过程中，我们采用了两台计算机同时进行计算用户的 MURank 值，每台机器运行了三个线程，将用户集近似平均分配给这六个线程。图 5-1 中可以看出，MURank 算法在第五次迭代趋于收敛状态，并且在实验过程中发现从第二十次迭代开始用户排名情况不再发生变化。因此，MURank 算法在第二十四次迭代的运行结果是个稳定、收敛状态，每个用户都得到一个稳定的 MURank 值。这也表明了 MURank 具有良好的收敛性。

5.2 与其它算法的相关性

5.2.1 与 PR、FR 算法

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，如 Daniel[2]等人把用户被浏览次数作为用户影响力的一个客观指标，然后将它与其它算法进行了相关性的研究，这些算法有根据粉丝数量、根据 PageRank 算法以及 Danial 的算法（IP 算法）等，实验表明用户微博消息被浏览次数与 IP 算法相关性比较大（0.95），与粉丝数量相关性比较低（0.59）。Jian[3]等人的研究也表明用户影响力与其粉丝数量之间的相关性比较弱。本章采用斯皮尔曼等级[9]相关系数来衡量各个算法之间的相关性，并且研究时间内段用户微博消息被转发次数与各自算法排名结果之间的相关性。

统计学中的斯皮尔曼等级相关系数（Spearman’s correlation coefficient）[9]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。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 X_i, Y_i 转化为等级数据 x_i, y_i，则原始变量 X，Y 之间的相关系数 ρ 为：

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y - \bar{y})^2}} \tag{5-1}$$

式子中 $\bar{x}, \bar{y}$ 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，公式 5-2：

$$\rho = \frac{6\sum_i d_i^2}{n^2(n-1)} \qquad (5-2)$$

其中 $d_i = x_i - y_i$。斯皮尔曼等级相关系数 ρ 的取值范围[-1,1],当 ρ 越接近 1，则表明两种排序算法之间相关性越大，当 ρ 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 ρ 。

研究社交网络用户影响力的领域中，学者会采用以下两种算法与新算法做个关联性分析，这两个算法是：

- FR 根据用户在 t 时刻的粉丝数量，来衡量微博用户在该时刻的影响力。粉丝数量是衡量用户影响力的一个非常直观的参数，所以在研究 Twitter 等用户影响力问题，学者们把它作为一个重要衡量用户影响力的算法，如 Daniel[2]等人。他们研究表明用户影响力与用户粉丝数量之间没有必然联系。
- PR 根据原始 PageRank 的基本思想，来衡量微博用户在某个时刻的影响力，其中阻尼系数 d 取 0.85。PageRank 算法是一个非常经典的 Web 页面排名算法，这个算法与 MURank 算法最大不同在于，PR 算法认为用户将自身的 PR 值平均的分配给其关注者，不区分对不同关注者的关注程度。本文实验结果表明，应用在新浪微博网络结构中，该算法仍然具有收敛性。

本文对于实验数据集中的用户分别用 MURank、PR 与 FR 衡量 t1 时刻的用户影响力，并对这三个算法的运行结果做了两两之间的相关性对比，如表 5-1：

表 5-1 算法之间的相关性

Table 5-1 The corelation between algorithms

	PR vs FR	MURank vs FR	MURank vs PR
斯皮尔曼等级相关系数	0.74	0.43	0.302

从表 5-1，可看出 MURank 算法的运行结果与 FR 的运行结果之间关联性不强，即 MURank 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。MURank 与 PR 之间的相关系数只有 0.32，表明 MURank 算法虽然是基于 PageRank 的基本思想，但是其运行结果与 PR 结果之间的关联性不强。

5.2.2 与转发次数

用户影响力体现了用户节点能给其它节点提供有用的建议、观点等，具体表现形式

有浏览、转发与加为关注者。其中浏览动作只是表示了粉丝看到该用户的消息，而没有进一步的动作，无法判断粉丝是否认同这条消息等。而转发动作表明粉丝认可这条消息以及希望粉丝喜欢他的粉丝也能看到这条消息，这对于信息传播来说是一个重要的动作。还有一个动作是加某个用户为关注者，当用户在其它地方看到用户 A 的消息时候，他认为与用户 A 兴趣一致或者他对于用户 A 的状态等很感兴趣，希望接着还能看到用户 A 的消息时候，那么用户就会加用户 A 为关注者。例如，在新浪微博中，有许多人对于明星感兴趣，他们就会加明星为关注者，喜欢能实时得到明星的动态，还有基于共同兴趣爱好的，如微博用户“冷笑话“等的粉丝数量都会达到十几万，甚至更多。

我们把一个时间段内用户微博消息被转发的总次数,作为该用户影响力的一种指标。本文定义 t2 时刻为 t1 时刻之前一个月的时间，在表 5-2 中，我们列出 MURank 算法在 t2 时刻的前十名用户，和这十个用户在 PR、FR 算法中的排名情况，以及在 t2 时刻之后的一周内他们的微博消息被转发的次数。如表 5-2：

表 5-2 t2 时刻，MURank 算法的前十名用户在各个算法中的情况

Table 5-2 time t2, information of the top 10 users in MURank

用户 ID	被转发次数	转发次数排名	MURank	PR	FR
1713926427	1763	1	1	3	3
1644395354	774	3	2	2	1
1618051664	820	2	3	6	8
1660209951	570	4	4	12	9
1252373132	467	5	5	4	5
1764222885	325	6	6	9	4
1097201945	296	7	7	7	14
1567852087	268	8	8	10	7
1660209251	164	10	9	5	6
1615743184	233	9	10	11	11

通过观察，我们发现 MURank 算法的前十名用户，他们在 t2 时刻之后的一周内被转发次数的大小排序基本与他们 MURank 名次相符。但是我们发现用户 ID 为 1644395354 被转发的次数低于用户 1618051664，但是他的 MURank 排名更靠前一位，经过分析我们发现用户 1644395354 的粉丝数量是 68101 个，而用户 1618051664 的粉丝数量为 35938，比用户 1644395354 少了三万多个粉丝。MURank 算法在分配 MURank 值的过程中，用户从每个粉丝那里得到 MURank 值，粉丝数量相差悬殊，那么用户从粉丝得到的 MURank 值之和也是会有差别。并且用户 1644395354 的出度远远高于用户 1618051664 的出度，那么他

发一条微博消息，就会被更多人看到，所以他的影响力就相对的比较大也是合理的。另外，用户 1660209251 只被转发了 164 次，而位居第九位，我们统计分析前面八个用户都关注了他。我们知道如果一个权威者认为一个人重要，那么这个人就显得更加权威或者重要。因此用户 1660209251 被前八个 MURank 高的人关注，那么他的 MURank 值就会很高。所以虽然他的消息只被转发了 164 次，但是他的 MURank 排名比较靠前。其它两种算法的排名情况与被转发次数之间没有存在一个非常强的关联，如用户 1660209951 的 PR 排名在第十二名，而他的微博被转发了 570 次。在 FR 算法中，也存在类似的情况，如用户 1097201945。

我们采用斯皮尔曼等级相关系数来衡量表 5-2 中 MURank、PR、FR 与转发次数排名之间的相关性，实验结果显示转发次数与 MURank 之间的相关系数是 0.96，与 PR 之间是 0.925，与 FR 之间是 0.823。实验结果表明 MURank 算法与其前十名用户的被转发总次数之间的相关性比其它算法都要强。

5.3 与 PR 的对比分析

在 PageRank 算法的思想中，用户在某一时刻将自身的权值平均分配给该用户的关注者。而 MURank 算法认为，用户分配给其新关注者比旧关注者相对更多的 MURank 值，因此在 MURank 算法中，一个用户拥有更高 MURank 值不仅仅与粉丝数量、质量有关，还应该与其粉丝们加入他为关注者的时间长短有关系。

在第三章中，我们分析了用户转发行为的时间间隔分布情况，实验结果表明了一个现象：一个用户关注其关注者微博消息主要是发生在前期，之后会保持一个相对稳定的关注程度。因此，在前期的时间段内，用户分配给其关注者相对较多的权值，而之后会逐渐的减少或者不变。另外，我们在 5.2 节中，讨论了当一个用户被新粉丝关注，那么说明了这个新粉丝希望长期得到这个用户的消息等，因此新粉丝表现了对用户的极大关注，所以用户对于这个粉丝产生了很大的影响。所以一个时间段内新增粉丝数量也能反映了用户在这个时间段内被多少粉丝认可，或者说影响了多少人，这就体现了一个用户在一个时间段的影响力。

本文选取了数据集中关系初时刻的最大值（t1），分别计算 t1 时刻的 MURank 结果与 PR 结果，并且选取出 MURank 与 PR 结果不同的前 6 个用户，并分析这些用户在(t1-4,t1)时间内的粉丝变化情况，如表 5-3 所示：

表 5-3 MURank 结果与 PR 结果对比

Table 5-3 The compare between MURank and PR

编号	Userid	UserName	MURank	PR	(t1-4,t1) 内新粉丝数量
1	1097201945	治愈系心理学	6	7	8
2	1618051664	头条新闻	7	6	4
3	1780417033	实用小百科	14	17	4
4	1657421782	生活小智慧	17	14	1
5	1793285524	王力宏	18	19	1
6	1644572034	精彩语录	19	18	0

根据编号,我们将表 5-3 中的六个用户分为三组来分析,发现编号组合(1,2),(3,4),(5,6)中的 MURank 值与 PR 值正好相反，例如表 2 中（1,2）组合，用户“治愈系心理学”在 PR 排名第 7 名，而在 MURank 中排名第 6 名，根据事后采集在(t1-4,t1)时间段内有 8 个新粉丝加入。用户“头条新闻”在 PR 排名第 6 名，而在 MURank 中排名第 7 名，在(t1-4,t1)时间段内有 4 个新粉丝加入。用户“治愈系心理学”的 MURank 结果比用户“头条新闻”更靠前，其原因在于用户“治愈系心理学”新加入的粉丝数多于用户“头条新闻”，这是由于 MURank 算法认为粉丝会分配给新关注者更多的 MURank 值，而 PageRank 算法只是平均分配给所有的关注者。我们分析了另外两组情况，发现其原因均与(1,2)类似。

在 5.2 节中，我们讨论了转发与关注这两个动作，其中关注体现出用户对另一个用户的关注程度，高于转发其微博消息所体现的关注程度。因为转发一条微博，只是认为这条微博有价值，而关注了这个用户，表明他希望看到该用户的更多的微博消息，是个持久关注的行为。如果一段时间内某个用户增加了许多新粉丝，说明这段时间内，这些粉丝都想关注该用户的微博消息，所以该用户在这个时间段内影响力比较大。表 5-3 的分析结果表明与 PR 算法相比，MURank 算法的运行结果对于新粉丝数量变化更敏感，它更能反映了一个用户在一段时间内有多少新粉丝关注了该用户。

5.4 不同时刻的 MURank 结果分析

MURank 是一个具有实时性的用户影响力算法，因此我们有必要观察其时效性。所以我们另取一个与 t1 时刻之前一个月的时刻 t2。计算数据集中用户在 t2 时刻的 MURank 值，然后根据用户的 MURank 值对用户进行排名,得到每个用户在 t2 时刻的影响力排名情况。实验结果显示，t2 时刻的 MURank 算法仍然具有良好的收敛性，并且与 t1 时刻相比，t2 时刻有 1,379,579 个用户的排名发生变化，占总用户数量的 90.2%。在 MURank

算法计算过程中，t2 时刻用户 MURank 值随着迭代的变化情况如下图 5-3：

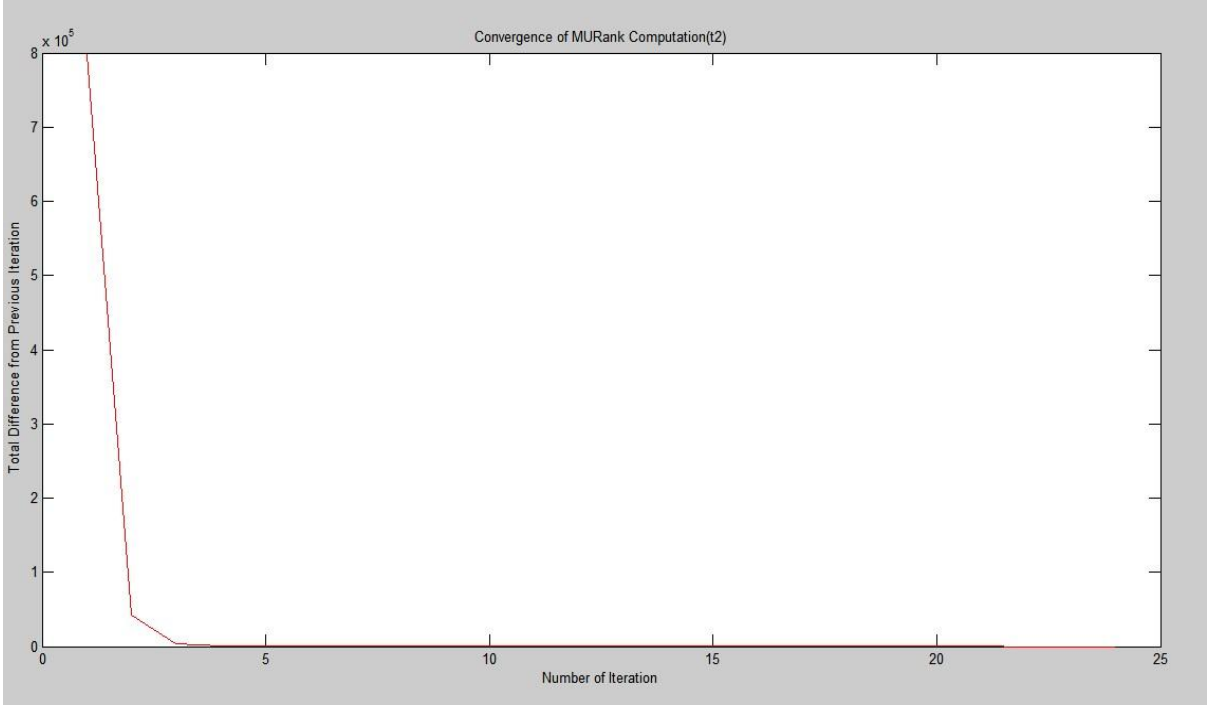

图 5-3 t2 时刻，MURank 算法收敛性

Fig.5-3 time t2, Convergence of MURank Computation

图 5-3 中，横坐标是指迭代次数，纵坐标为节点的相邻迭代的差的和（绝对值）。可以看出 t2 时刻 MURank 算法的收敛性与 t1 时刻的收敛性情况类似，也再次证明了 MURank 具有良好的收敛性。

我们直观的知道，如果有一段时间内，有许多新粉丝加用户 A 为关注者，那么表明用户 A 的微博消息在这段时间内被高度关注，进而表明用户 A 在这段时间内具有比较高的影响力，或者用户的影响力有所上升。因此一段时间内新粉丝数量与质量，是衡量用户的影响力情况的一个重要因素。

基于上述理论，本文获取了 MURank 结果的前二十用户中 t2、t1 两时刻 MURank 值不同的用户，统计分析这些用户的新增粉丝数量与质量情况，研究用户 MURank 值变化的原因。这些用户信息如下表 5-4：

表 5-4 不同时刻，用户户 MURank 值变化

Table5-4 The change of User's MURank at different moments

编号	UserID	MURank(t2)	MURank(t1)	(t2,t1)内的新增粉丝数	新粉丝中最高MURank(t1)
1	1660209951	4	5	13	10
2	1252373132	5	4	39	17
3	1642591402	13	16	26	54801
4	1671526850	14	13	30	54
5	1657421782	16	17	10	31990
6	1780417033	17	14	34	36

本文分别从各个用户在(t2,t1)时间段内新增加粉丝的数量、质量（新粉丝在 t1 时刻的影响力排名）来分析用户影响力变化的原因。表 5-4 中 t2 时刻，第一个用户、第二个用户的 MURank 排名分别是第四名、第五名，并且两者在（t2，t1）时间段内新增粉丝 MURank 最高值基本相同。但是新增粉丝数量不同，第一个用户的新增粉丝数量为 13，而第二个用户为 39，表明（t2，t1）内，与第一个用户相比，有更多的微博用户关注了第二个用户，因此第二用户的影响力有所上升，体现在 MURank 算法中，第二个用户的 t1 时刻 MURank 排名上升了一位。因此在新增粉丝质量差不多的提前是，新增粉丝数量对于用户 MURank 算法的排名变化有着重要的作用；我们接下来分析了第三个用户、第四个用户的变化情况，发现这两个用户的新增粉丝数量差不多，但是第四个用户的新增粉丝质量远远高于第三个用户的。因此第四个用户在 t1 时刻的影响力排名上升了一位。所以新增粉丝质量也是影响用户影响力算法的一个因素；最后，我们分析了第五个用户与第六个用户的情况，发现第五个用户的情况与第三个用户情况类似，第六个用户与第二用户情况类似。

综上分析结果，我们发现在一个时间段内，一个用户影响力的变化情况，不仅与新增粉丝数量有变，还与这些新增粉丝质量有关。这一结论与现实相符：一、如果有影响力高的用户关注了某个用户 A，那么用户 A 影响力也会水涨船高。二、如果有许多用户关注了用户 A，那么表明在这个时间段内，用户 A 的意见或者消息为这些新粉丝所认可与肯定。

5.5 本章小结

本章在实验数据集中运行了MURank算法来评价用户的t1和t2时刻影响力。同时我们也分别根据传统的PageRank算法、粉丝数量来衡量新浪微博用户的影响力。我们研究分析了这三种算法的实验结果，实验结果表明根据MURank算法评价的用户影响力，与用户

的粉丝数量之间不存在强关联，这个结论与学者们研究Twitter用户的结论是一致的。同时与根据传统的PageRank算法评估用户影响力结果相比，我们发现运用MURank算法评价用户影响力的实验结果更能反映一个用户在一段时间内有多少新粉丝关注了他。最后，我们研究分析了不同时刻的用户MURank值变化情况，发现MURank算法具有良好的实效性。本章的实验结果表明了，运行MURank算法来评估用户的某时刻影响力，其用户影响力(MURank值)不仅仅与在该时间段内的新增粉丝数量相关，还与这些新增粉丝的质量相关(即与这些粉丝的影响力相关)。

第六章 总结与展望

6.1 论文工作总结

本文在总结基于互联网的社会网络分析研究现状、人类动力学方面的理论与研究方法，分析社交网络用户转发行为的时间间隔分布，以此为基础，建立了新浪微博网络中用户转发行为的数学模型。将经典的搜索算法 PageRank 算法的基本思想应用到社交网络的用户影响力研究中，进而提出了社交网络用户的实时影响力算法(MURank)。通过新浪微博的 API 等来获取实验数据集，对新浪微博用户进行实验分析，证明了算法的有效性。主要研究成果如下：

- 借鉴现有的网络内容挖掘技术，利用新浪微博官方的 API 来获取新浪微博用户的信息（好友信息与微博消息）。由于新浪微博是一个在线社会网络，它具有复杂网络的特性，如小世界特性、无标度特性等。因此我们运用了复杂网络的研究方法，分析了新浪微博的用户转发行为。实验结果表明了新浪微博用户之间的转发行为时间间隔分布服从幂律分布，这与近年来人类动力学中的一个理论相符，即许多人类行为时间间隔分布是服从幂律分布且具有“胖尾”现象，而不是泊松分布。
- 根据观察到的转发行为时间间隔分布，本文定义了用户之间的关注度，并对它进行了数学建模，从而量化了用户之间在特定时刻的关注程度。基于用户的关注度与传统 PageRank 的基本思想，本文提出了评估新浪微博用户实时影响力算法（MURank)。实验表明了该算法具有良好的实时性，能比较好的反映了一个用户在某时刻的网络影响力。通过分析用户的影响力，我们可以比较迅速的找出网络中最具有影响力的节点，对于这些节点进行分析，有助于信息传播、控制的研究。

本文中也有存在一些不足之处，主要有以下几个方面：第一，数据量的问题。由于受到 API 调用的限制，我们只获取了新浪微博中的一小部分数据，因此对于该数据集是否比较全面的代表整个新浪微博网络还需要进一步的研究与分析。第二、在建立用户之间关注度模型上，我们只是考虑了时间间隔这个因素。而实现生活中，用户之间的兴趣爱好、地理位置等因素都会影响到了用户之间的关注程度。

6.2 研究展望

本文针对微博网络结构分析、用户影响力进行了探索性研究，实验结果表明了新算法具有良好的实时性与有效性。在新算法(MURank)中，用户的关注度是个非常重要的参数。而关注度的模型中只考虑了时间间隔这个因素，没有涉及到用户之间的微博消息相似度等因素。因此我们进一步的研究重点：第一：研究用户之间微博消息的相似度等，分析出关注度更多的因素，建立更为复杂、精确的关注度模型。最终目标是希望能只分析用户的粉丝情况，就能得到一个用户的全网影响力。第二：新浪微博网络中的社区挖掘，社区挖掘一直是国内外研究社交网络的一个热点。研究社区结构，有助于研究微博的信息推荐机制，也有助于个性化搜索的研究。第三：研究微博网络中其它的行为，如用户发表微博、关注的行为等。研究这些行为，对于人类行为动力学的研究有着重大的意义。

微博网络已经成为一个强大的信息交流平台，它已经成为一个新的传媒介质。研究微博网络上的用户关系、用户行为，有助于我们理解与研究人类社会。另外，微博网络具有媒体性质，研究它有着应用价值与商业价值，如广告平台等。

参 考 文 献

[1] Meeyoung Cha,Hamed Haddadi,Fabricio Benevenuto ets.Measuring User Influence in Twitter: The Million Follower Fallacy[C]. International AAAI Conference on Weblogs and Social Media(ICWSM),May 2010.

[2] Daniel M.Romero, Wojciech Galuba,Sitaram Asur,Bernardo A.Huberman. Influence and Passivity in Social Media [C]. ECML PKDD,2011.

[3] Jianshu Weng,Ee-Peng Lim,Jing Jiang ets.TwitterRank:Finding Topic-sensitive Influential Twitterers[C].WSDM'10, Feb 2010.

[4] Page L,Brin S MOtwani Ret al.The pagerank citation ranking:Bringing order to the web[R].Stanford Digital Libraries,1999.

[5] Barabási A L. The origin of bursts and heavy tails in human dynamics[J]. Nature,2005,435:207-211.

[6] João Gama Oliveira, Albert-László Barabási. Human dynamics: Darwin and Einstein correspondence patterns[J].Nature,2005,437:1251-1251.

[7] A.Vazquez ,J.G Oliveira, K.-I. Goh, I. Kondor, A-L Barabási. Modeling bursts and heavy tails in human dynamics. Physical Review E,73(036127),2006.

[8] YeWu,Jurgen Kurths.Human comment dynamics in on-line social systems[C].physicaA,July 2010.

[9] Honey.C,Herring,S.C. Beyond Microblogging:Conversation and Collaboration via Twitter[C].System Sciences,2009. HICSS ' 09.42nd Hawaii International Conference on.Pages 1-10.Big Island, HI.

[10] Reka Zsuzanna Albert : Statistical mechenics of complex networks [D]. Department of Physics Notre Dame 2001.

[11] ErdSs P.，Renyi A. On the evolution of random graphs, Pub.Math.Inst.Hung . Acad. Sci.，1960，5：17—6.

[12] ErdSs P.，Renyi A. On random graphs，Publicationes Mathematicae，1959，6：290—297.

[13] ErdSs P.，Renyi A. On the strength of connectedness of arandom graph，1961，2：261—267.

[14] Stanley Milgram，The Small-World Problem[J].Psychology Today.1967.2:60.

[15] Six degrees of Separation. Retrieved Dec 14,2010,from http://zh.wikipedia.org/zh-cn/File:Six_degrees_of_separation.png.

[16] 康书龙.基于用户行为及关系的社交网络节点影响力[D].北京：北京邮电大学。2011:1-59.

[17] Dodds P S,Muhanad R,Watts DJ.An experimental study of search in global social network[J].Science,2003,301:827-829.

[18] Watts D J,Strogatz S H.Collective dynamics of ‘small-world’network[J].Nature,1998,393:440-442.

[19] Faloutsos M,Faloutsos P,Faloutsos C.Computer Communications Review,1999,29:251.

[20] Ebel H,Mielsch L I,Borbholdt S.Phys.Rev E,2002,66:035103.

[21] Jeong H et al. Nature,2001,411:41.

[22] 百度百科 http://baike.baidu.com/view/1405540.htm.

[23] Notable social networking websites. Searcher, 2007: 36-37.

[24] Watts D. J., Peretti J., Frumin M. Viral marketing for the real world[J]. Harvard Business Review. 2007, 85(5): 22-30.

[25] Facebook 每天数据处理量超 500TB, 08,23 2012, http://net.chinabyte.com/241/ 12412241.shtml.

[26] 数据显示twitter日信息发送量达5000万 23,02,2010, http://news.cnblogs.com/n/57308.

[27] Wilm.P. van der Aalst, Song M. Mining social networks: uncovering interaction patterns in business processes [J]. Business Process Management. 2004(3080): 244-260.

[28] Bacciqalupo C., Plaza E. Mining music social networks for automating social music services [A]. Workshop Notes of the ECML/PKDD2007 Workshop on Web Mining, 2007: 123-134.

[29] Gross R., Acquisti A. Information revelation and privacy in online social networks [A]. Proceedings of the2005 ACM workshop on Privacy in the electronic society, 2005:71-80.

[30] Benevenuto F., Rodrigues T., Cha M. Y. et al. Characterizing user behavior in online social networks [A]. Proceedings of the 9th ACM SIGCOMM conference

on Internet measurement conference, 2009: 49-62.

[31] Watts D. J., Peretti J., Frumin M. Viral marketing for the real world [J]. Harvard Business Review. 2007, 85(5): 22-30.

[32] Bergamaschi, S., Castano,S. and Vincini, M. Semantic Integration of Semistructured and Structured Data Sources. SIGMOD Record,28(1).54-59.

[33] S. Wasserman and K. Faust. Social Network Analysis:Methods and Applications.Cambridge University Press, Cambridge, UK, 1994.

[34] Michael M. A brief history of generative models for power law and lognormal distributions [J]. Internet Mathematics, 2004, 1(2): 226-251.

[35] Li L., Alderson D., Doyle J. C., et al. Towards a theory of scale-free graphs: definition, properties, and implications [J]. Internet Mathematics, 2005, 2(4): 431-523.

[36] Girvan M., Newman M. E. J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 99(12): 7821-7826.

[37] Traag V. A., Bruggeman J. Community detection in networks with positive and negative links [J]. Physical Review E, 2009, 80(3): 23-29.

[38] Gregory S. A fast algorithm to find overlapping communities in networks [J]. Machine Learning and Knowledge Discovery in Databases, 2008(5211), 408-423.

[39] Jennifer J. X., Chen H. C. Crime Net Explorer: A Framework for Criminal Network Knowledge Discovery [J], ACM Transactions on Information Systems, 2005, 23(2): 201-226.

[40] Emirbayer M., Goodwin J. Network Analysis, Culture, and the Problem of Agency. American Journal of Sociology, 1994, 99(6): 1411-1454.

[41] Wellman B. Computer Networks as Social Networks [J]. Science Magazine. 2001, 293(5537): 2031-2034.

[42] 朱庆华, 李亮. 社会网络分析法及其在情报学中的应用[J]. 情报理论与实践, 2008(2): 179-183.

[43] Jennifer J. X., Chen H. C. Crime Net Explorer: A Framework for Criminal Network Knowledge Discovery [J], ACM Transactions on Information Systems, 2005, 23(2): 201-226.

[44] 汪小帆，李翔，陈关荣. 复杂网络的理论及应用[M]，北京：清华大学出版社，2006.

[45] Freeman L.C. Centrality in social networks. Conceptual Clarification [J]. Social Networks, 1979, 1(3):215-239.

[46] Freeman L.C. A set of measures of centrality based on betweenness [J]. Sociometry, 1977, 40(1): 35-41.

[47] Berkowitz S. D. An Introduction to Structural Analysis: The Network Approach to Social Research [M]. Butter worth, Toronto, 1982.

[48] Breiger R. L. The analysis of social networks [A]. Handbook of Data Analysis. Sage Publications, London, UK, 2004: 505 - 526.

[49] Liu X. M., Bollen J., Nelson M. L., et al. Co-authorship networks in the digital library research community [J]. Information Processing & Management: 2005(41):1462-1480.

[50] Wu, F.,Huberman,B.A.,Adamic,L.,Tyler.J.:Information Flow in Social Groups.Physica A 337,327-335(2004).

[51] Domingos,P.,Richardson,M.: Mining the network value of customers.In: SIGKDD(2001).

[52] 韩筱璞，汪秉宏，周涛.人类行为动力学研究[J]. 复杂系统与复杂性科学.2010,1.

[53] DezsoZ.AlmaasE.LukacsA，Fifteen minutes of fame:the dynamies of information aeeess on the Web[J].Physieal，2006(06).

[54] H.B.Hu,and D.Y.Han,Physica A 387,(2008)5916.

[55] 36氪，关注互联网创业，Twitter影响力指数Klout开始支持Facebook, 2010, accessible in June 2011。http://www.36kr.com/klout-supports-facebook/.

[56] http://baike.baidu.com/view/368550.htm

攻读硕士学位期间已发表或录用的论文

[1] 陈少钦，范磊，李建华. MURank: 一种社交网络用户实时影响力算法，信息安全与通信保密（已录用）.

攻读硕士学位期间参与的科研项目

致 谢

光阴荏苒，转眼二年半的研究生生活即将结束。回顾这研究生两年半以来的点点滴滴，在学习、工作和生活中，得到了父母、老师和同学一路上的指导与帮助，使自己受益良多。

首先，我要感谢我的导师李建华教授、范磊副教授。他们严谨的科研态度、朴实的生活作风等格深深地感染和激励着我。在硕士研究生期间，我得到了两位老师给予的无数帮助以及耐心指导。他们幽默风趣的言语、平易近人的态度以及豁达的胸襟将影响着了我为人处世的作风。

再次，感谢学院以及学校为我提供这么好的学习环境，一流的教学质量和先进的科研条件，帮助我很好的完成学业。感谢实验室的全体老师和同学，对本人在学术、科研上的关心和支持。与大家相处的这段时间，不仅使我感受到科研团队的力量，也使我感受到生活的愉悦！

最后，我要感谢我的父母，感谢他们一路上对我的支持和鼓励，使我能够坚持下来，最终完成硕士的学业。

在此祝愿以上所有人学业事业双丰收。

上海交通大学硕士学位论文答辩决议书

1100349162

姓　名	陈少钦	学号	1100349162	所在学科	电子与通信工程
指导教师	李建华	答辩日期	2013-01-04	答辩地点	电信群楼1号楼二楼会议室
论文题目	基于PageRank的社交网络用户实时影响力研究				

投票表决结果：5/5/5（同意票数/实到委员数/应到委员数）　答辩结论：☑通过　☐未通过

评语和决议：

用户影响力算法在社交网络领域是一个研究热点，它具有较高的学术意义与商业价值。论文基于经典搜索算法 PageRank 与人类动力学研究方法，提出一种社交网络用户实时影响力算法，选题具有理论意义和重要应用背景。

论文通过新浪微博网络的 API 获取好友关系与用户微博信息作为分析数据，验证了新浪微博用户转发行为时间间隔分布服从幂律分布，以此为基础定义用户之间的关注度模型。基于 PageRank 算法的基本思想，提出适用于社交网络的用户实时影响力算法(MURank)。实验结果表明了 MURank 算法的有效性。

论文选题先进，内容较充实，表述条理清晰，并有较好的新见解。论文表明，作者具有扎实的理论基础、专业知识与良好的自身见解，以及较强的科研工作能力。论文已达到硕士学位论文水平。

陈少钦同学在答辩过程中，能够清楚地报告自己所从事的工作，回答问题正确。经答辩委员会认真讨论。投票表决，一致同意通过其硕士学位论文答辩，并建议校学位评定委员会授予其专业硕士学位。

2013 年 1 月 4 日

答辩委员会成员签名	职务	姓名	职称	单 位	签名
	主席	薛质	教授	上海交通大学	薛质
	委员	戎蒙恬	教授	上海交通大学	戎蒙恬
	委员	唐俊华	副教授	上海交通大学	唐俊华
	委员	范磊	副教授	上海交通大学	范磊
	委员	李生红	教授	上海交通大学	李生红
	秘书	姜红	中级	上海交通大学	姜红