微博关注网构建与统计分析研究

刘茂福　康 乐　顾进广
（武汉科技大学 计算机科学与技术学院　武汉市 430065）

摘　要　本文通过使用微博用户数据作为社会媒体数据来源，基于微博用户之间的相互关注关系构建出微博关注网，然后利用社会网络分析和数据统计分析方法，对所构建关注网的用户节点属性特征以及网络结构特征进行了细粒度分析，并对微博这种虚拟群体的基本特征进行了总结。主要结论包括：（1）微博用户所关注的对象与用户所在地域密切相关；（2）微博用户圈中处于核心地位的往往是粉丝数目较多的用户或是职业在该用户圈中较为受关注的用户；（3）关注子网的网络密度较大，用户之间的联系十分紧密；（4）微博用户之间的互动往往发生在其所属的派系之中。

关键词　微博关注网　社会网络分析　群体特征分析

中图分类号　TP391　　**文献标识码**　A

THE CONSTRUCTION AND ANALYSIS OF MICROBLOG FOLLOWING NETWORK

LIU Mao-fu　KANG Le　GU Jin-guang
(*College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China*)

Abstract　We use Microblog data as a social media sources to construct the microblog following network based on mutual relationships between the Microblog users, then analyze the attributes of the user node and the structure of the network by the means of social network analysis and other statistical analysis technology. Finally we summarize the basic feature of virtual groups in Microblog. The main conclusions include four points: (1)Microblog user and whom he or she follows often share the same location; (2)Those who have more fans or have a distinguished job are the core of the group; (3)The density of the network is quite high and the users are closely linked to each other; (4)The interaction among the Microblog users often occurs in their own clique.

Keywords　Microblog following network　Social network analysis　Group feature analysis

0　引　言

伴随着以Web2.0为特征的新一代互联网的迅速发展，人与人的关系逐渐在网络中占据主导地位。随之而来产生了各种网络应用，如社交网络（Facebook、人人网等）、微博（Twitter、新浪微博等）、论坛贴吧、百科（Wikipedia、百度百科等），这些应用无一不强调人在其中的参与和互动，微博作为其中的佼佼者，将人与人之间的互动发挥到了极致[1]。用户通过便捷迅速的信息共享渠道，在任何时间和任何地点发布内容短小精悍的信息，根据自己的兴趣偏好和对方发布内容的类别与质量，来选择是否“关注”某个用户[2]。研究表明微博已成为人们日常生活的重要组成部分之一，在国内其影响力甚至大于社交网站[3]，但由于微博的使用门槛很低，所发微博内容本身比较随意，关注的对象也具有很大的随机性，所以微博用户这类群体的行为特征及其结构特征具有不稳定性与多样性，难以进行量化分析。

社会网络分析SNA（Social Network Analysis）是 20 世纪 70 年代以来在社会学、心理学、人类学、数学、通信科学等领域逐步发展起来的一个的研究分支。社会网络分析不仅仅是一种工具，更是一种关系论的思维方式。近些年，国内学者已经将其应用在了很多研究领域，例如分析问答社区的运行机制、知识传播和共享模式[4]；结合文本内容分析和社会网络分析进行高质量Blog社区发现 [5]；引文分析[6]探讨了引文网络的结构及其对知识的流动传播产生的影响；科研人员合著[7]研究则采用文献计量法和社会网络分析法,对某个领域论文的合著情况进行统计分析,从合著度与合著率、合著网络的中心性、凝聚子群、社群图等方面探讨领域研究热点以及作者之间的合作关系情况。

收稿日期: xxxx-xx-xx。本文承国家自然科学基金(项目号 61100133)、国家社会科学基金重大招标项目(项目号 11&ZD189)和湖北省教育厅人文社会科学基金重点项目(项目号 2011jyte126)的资助。刘茂福(1977-)，男，副教授，博士，主要研究方向：自然语言处理，社会计算，E-mail: liumaofu@wust.edu.cn；康乐（1987-），男，硕士研究生，主要研究方向：自然语言处理，社会计算；顾进广(1974-)，男，教授，博士，主要研究方向：语义网技术

这些研究主要是把社会网络分析手段应用于虚拟社区，挖掘虚拟社区的结构、核心和通信行为等。微博用户之间的相互关注与相互转发微博的关系实质上也构成了一种虚拟社会网络，然而把微博用户作为一种虚拟群体，并针对其群体的社会学特征的相关研究还比较少。利用微博“社会网络”和“开放媒体平台”的双重属性，可以采用社会网络分析（SNA，Social Network Analysis）方法来研究其网络结构特点，再结合一些辅助的统计分析方法，可以对微博用户这种虚拟社会群体的特征进行细粒度的分析研究。本文利用全自动化的数据获取与网络构建方法，基于最新的微博数据来实现对社会网络节点属性的统计分析，分析微博社会网络的结构特点，挖掘该网络群体中用户特点以及用户之间关系的实质，研究结果对于优化微博信息整合与传输、微博发展方向、维护网络内容安全、促进网络发展具有一定的理论和实践意义。

1 微博关注网构建

1.1 数据获取

利用新浪微博提供的接口可以便捷并准确地获得所需数据，新浪微博开放接口的具体使用方法可以参考《新浪微博·开放平台API文档》[8]。基于新浪微博接口的数据获取方案考虑到以下几个方面。

(1) 爬取数据类别

微博用户个人信息，包括所在地、性别、粉丝数、关注数、微博数、互粉数、是否为认证用户以及认证原因、关注用户列表、微博注册时间、微博读取控制权限；微博用户行为信息，主要为所发微博，微博类型（包括原创和转发）。

(2) 数据爬取顺序

爬取的时候是按一个一个的用户顺序进行爬取，对于每个用户获得（1）中提到的爬取数据类别。首先人工选择一个起始用户（benchmark_user）进行爬取，然后从接口获取该用户所关注的微博用户集合（L_{api}）存入到微博用户数据库(user_db)，从数据库第一个尚未获取其关注微博用户集合的用户开始爬取该用户的关注用户集，以此类推。具体的爬取模式类似树的广度遍历，如图 1 所示，爬取并存入数据库的用户顺序将是：{（起始用户），（A、B、C），（（D、E），（F、G、H），（I、J））}。

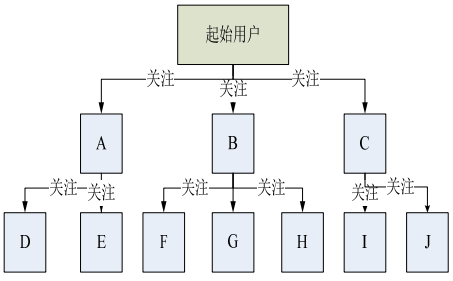

图 1 关注关系爬取模式

(3) 爬取方法

考虑到爬取的效率问题和新浪微博接口对调用频率的限制，使用多线程和缓冲队列的机制进行爬取。具体爬取过程中为每个用户设置一个爬取进度标志位 finish_flag，当某用户的所有关注用户都保存在微博用户数据库(user_db)中则称该用户爬取完毕，爬取标志位才为真。整个数据爬取的流程图如下：

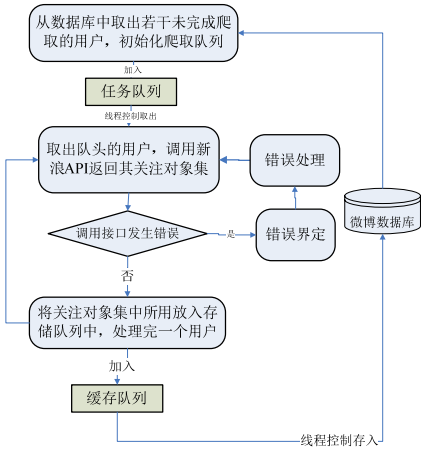

图 2 基于新浪微博接口的数据获取流程图

爬取过程的核心算法如下：

L_{DB}：已爬取的最终将存储在数据库中的微博用户列表；

finish_flag：标识一个用户的关注关系是否齐全，所谓“齐全”是指其关注的所有用户对象都储存到了数据库中；

current_user：从队列中取出并正在为其调用新浪微博接口取得其关注对象集合的用户；

```
算法 1：数据爬取
输入：基准用户 benchmark_user
输出：LDB
{
    LDB=empty; // 开始LDB为空
    benchmark_user.finish_flag=false;
    current_user= benchmark_user;
    //如果该用户关注关系不全，并且数据库中用户数
    量还未达到理想个数，则循环爬取
    while(current_user.finish_flag==false&&
    LDB.size<WANTED_SIZE){
        //调用微博接口，返回其关注对象集合
        Lapi= current_user.callWeiboApi();
        current_user.setFinishFlag(true);
        Lapi.setFinishFlag(false);
        Lapi.setBeenFollowedBy(current_user);
        LDB.add(Lapi); //添加到数据库用户列表中
        //LDB中取出下一个尚未爬取完毕的用户
        current_user= LDB.getNext();
    }
    user_db.save(LDB); //存储到微博用户数据库
}
```

(4) 数据预处理

在利用微博用户数据库中的用户数据进行分析之前，有必要首先过滤出没有研究价值的用户记录。我们利用这些用户基本信息：微博注册时间、微博数、微博读取控制权限、关注数，将以下四类没有研究价值的用户记录剔除。

① 微博账号注册时间太短的用户（注册时间至今小于半年）；

② 发微博频率很小的用户(注册至今发表微博数量平均每星期不足一篇）；

③ 对微博数据读取有限制的用户(将个人微博读取权限设置为他人不可见）；

④ 不关注任何人的用户（关注对象个数为 0）。

接下来根据关注网的形式化定义，利用预处理之后的用户数据对关注网进行可视化。

1.2 关注网可视化

微博关注网 MFN（Microblog following network）是一个二元组，记为 MFN=(W,R)，其中 W 是微博用户（Weibo User）的非空有限集合，R 是有序集 W×W 的一个子集，其元素是关注网的关注关系（Following Relation）

从微博用户数据库（user_db）中随机选择若干用户作为样本集（S_w）的可视化算法如下:

```
算法 2：关注网络可视化
输入：Sw 微博用户样本集
输出:关注网络图
{
    generateNodes(Sw); //为用户生成节点
    for(Wi in Sw)      //为节点生成边
        for(Wj in Sw && Wi≠Wj){
            if(Wi follows Wj)
                generateAEdge(Wi ,Wj);
        }
}
```

对于一个 50 人样本的关注网进行可视化之后，如图 3 所示。

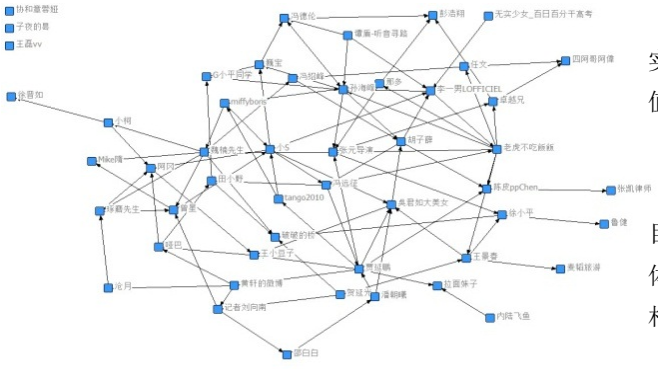

图 3 关注网络可视化示意图

关注网的数据获取和可视化完成后，下面就可以基于这些数据进行分析了。

2 关注网数据分析

在所构建的微博关注网的基础上，对网络中的个体及网络结构特性进行分析。首先从关注网络的节点属性特征出发，对关注网中微博用户的特征进行统计分析，从宏观层面得出该群体的统计特征规律。然后针对关注网节点之间的连接关系，利用社会网络分析方法分析其网络结构特点以及节点之间的关联性，最后综合节点特征与社会网络结构特征，研究微博用户自身特点对该虚拟社区特征的影响。

2.1 微博用户个人特征统计分析

利用微博用户数据库中用户的属性数据，针对如下几方面对用户特征进行统计分析：

(1) 用户关注与被关注情况，用户发微博的频率；

(2) 关注对象在地域上的分布；

(3) 关注对象在职业上的分布；

随机从微博用户数据库中取出 N 个微博用户，利用用户节点的属性数据对上述特征进行统计的具体算法如下：

```
算法 3：数据统计分析
输入：随机选取N个微博用户的集合Lsu;
输出：所输入节点的统计信息。
{
    for(user : Lsu){
        //获取其关注对象的地域统计信息
        update(user,locationInfo);
        //获取其关注对象的职业分布信息
        update(user,jobInfo);
        //统计关注、微博数量
        friendsCount+= user.getFriendsCount();
        followsCount+= user.getFollowerCount();
        weiboCount+= user.getWeiboCount();
    }
}
```

使用算法 3，计算出来的数据可用来分析用户关注与被关注的情况，用户发微博的频率，用户关注对象所在地域的人数，用户关注对象所处行业的人数。

2.2 微博用户个人特征统计分析

(1) 社会网络密度

社会网络密度 SND（Social Network Density）是指网络中实际包含的关系总数与其理论上所能包含的最大关系数的比值，对于关注网这种有向图，计算公式如（1）所示。

$$SND=m/(n(n-1)) \quad (1)$$

其中n指社会网中节点个数，m为网络中实际包含关系数目。网络的密度越大，该网络联系得越紧密，在微博关注网中体现为其中的微博用户之间相互关注程度高，彼此之间更容易相互影响[9]。

(2) 网络中心性

社会网络中“中心性”研究的是节点所代表的用户在网络中居于怎样的中心地位，常用的几类中心度包括：度数中心度、中间中心度、接近中心度等。

节点 A 的度数中心度就是与点 A 直接相连的其它点的个

数，在关注网这样的有向图中，每个点的度数可分为点入度（被关注）和点出度（关注别人），关注网络中入度直接反映该用户在网络中直接受关注的程度，所以本文使用度数中心度的入度来衡量一个用户是否处于关注网的中心。与之相似的概念是用户的粉丝数，代表了该用户在微博上有多少人关注，通过对比用户在其所处局部网络上的入度与其微博平台（可以说是整体网）上的粉丝数，可以很清晰的看出该局部网络的中心用户特点。

(3) 社会网络中的派系

社会网中“凝聚子群”反映的是集合中的节点之间具有相对较强、直接、紧密、经常的或者积极的关系。派系是最基本的凝聚子群，其定义为：其成员之间的关系都是互惠的（在关注网中是指相互关注），并且不能向其中加入任何一个成员，否则将改变这种性质。利用图论语言给出派系的定义为：在一个图中，派系指至少包含三个点的最大完备子图。

在我们所构建的关注网络中体现为一个数目为 N（N>=3）的最小完备子图，该子图的网络密度为 1，也就是说子图中用户节点之间均相互关注。

(4) 网络关联性

社会网络的关联性是指一个群体中的成员之间的关系把该群体联系在一起的程度。这里使用关注关系的可达性（Reachability）来测量网络的关联性，即两个点之间的途径越多，关联性越大，由此给出关联度的测量公式（2）。

$$C=1-[V/(N(N-1)/2)] \quad (2)$$

其中 V 是该网络中不可达的点对数目，N 是网络的规模。使用公式（2）对关注网的关联性进行测量。

距离频次数据（Frequencies of Geodesic Distances）用来统计节点之间距离长度出现的频次；网络距离的描述性数据（Descriptive Statistics）统计网络中的节点之间的平均距离[9]。

3 实验结果与分析

3.1 微博用户个人统计特征

从关注网中随机抽取 500 个用户节点，根据算法 3 对这 500 个用户进行统计计算，得出如下实验结果：

(1) 地域分布

通过查看所统计出的用户的关注对象的地域分布，发现用户的关注对象与用户自身所处地域联系十分紧密，用户与其关注对象位于同一地域的概率是 43.7%。图 4 是某个用户（所在地为北京朝阳区）关注对象的地域分布图。

(2) 关注与发微博

每个微博用户平均关注 150 名用户，被 180 名用户所关注，相互关注的对象占 68%；发微博频率很频繁，平均每人每天发微博 4 条，也就是说每名用户平均一天可以收到 600 条微博更新，由此可见微博用户每天所获取的信息量比较大。

(3) 职业分布

微博个人用户数据中并未直接指明用户的职业，在数据获取部分所得到的实际上是微博用户的认证原因，通过对认证原因的人工比对，发现用户所关注对象通常也是和自身职业相关的领域。例如明星所关注的微博用户大致有这几种职业：媒体人（媒体编辑、报社总监、导演等）、其他艺人、行业精英、作家、媒体官方微博等。图 5 是某个用户（知名艺人）关注对象的职业分布分布图。

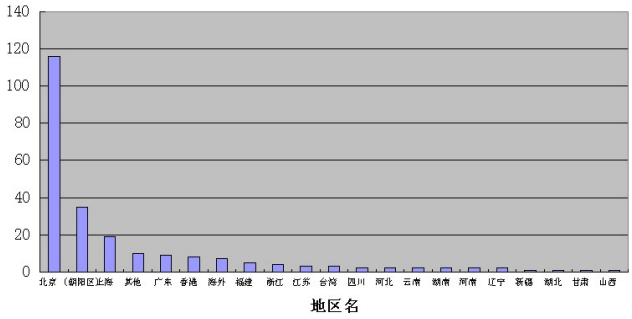

图 4 某个用户关注对象地域分布示意图

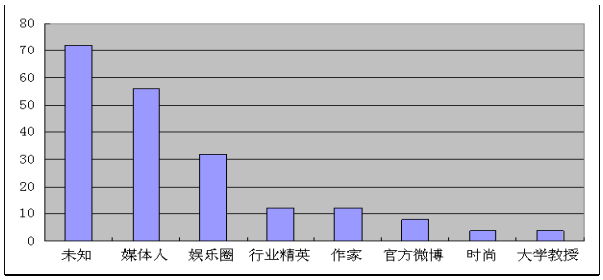

图 5 某个用户关注对象职业分布示意图

3.2 关注网络结构特征

从所构建的关注网中取出一个样本局域网络进行社会网络分析。为了使局域网络中用户节点的关系具有实际意义，局域网络的构造规则为：人工选取一个基准节点（实验中使用微博人气排行第一的用户），选取与该基准节点相连接的 200 个邻接节点作为局域网络中的所有节点。

(1) 关注网络密度

使用公式（1）计算得出所选取的局域网络的密度为 ρ =0.0443，200 个微博用户之间有 1765 条关系，也就是说平均每个微博用户大约与其他用户之间有 9 条关系，局域网络群体之间关系非常紧密。

(2) 中心性分析

我们这里主要是研究对于局域网络，通过上文中提到的网络中心性的定义，计算出节点的入度，然后与“微博用户自身粉丝数目”进行对比分析。图 6 是样本节点中分别根据用户粉丝数目以及该用户节点在本网络中的度数中心度入度排序之后的对比图。

图中红色部分是左边和右边共有的微博用户，一共有四位。从图中可以看出微博用户拥有的粉丝数目很大程度上决定了该用户在小规模局域网络中的受关注程度，几乎一半的拥有最多粉丝的用户在局域网络中排名靠前。通过分析蓝色部分的用户，我们有理由认为：这些处于局域网核心，而粉丝并非最多的用户，可以很好地反映该局域网的核心倾向。例如对于一

个艺人明星节点，位于该基准用户局域网核心的主要是编辑、编剧、作家和公司总裁经理等此类职业的人。

按粉丝数目进行排列（前十名）：	按度数中心度的入度进行排列（前十名）：
1.小S(台湾知名主持人)：19611993 2.杨幂(演员，代表作《宫》《仙剑奇侠传三》等)：17195830 3.冯绍峰(演员冯绍峰)：8499758 4.李小璐Super璐(演员李小璐)：7158415 5.黑人建州(台湾艺人“黑人”陈建州)：6174621 6.徐小平(真格基金创始人、新东方联合创始人)：4790528 7.房祖名(香港演员)：4433998 8.韩寒(作家,赛车手韩寒)：3712404 9.电影工厂(暂无（电影工厂）)：3245595 10.南派三叔(《超好看》主编，南派小说堂 11.会创始人，《盗墓笔记》等书作者。)：2982623	韩寒(作家,赛车手韩寒)：72——3712404 小S(台湾知名主持人)：46——19611993 徐小平(真格基金创始人、新东方联合创始人)：39——4790528 斜江明私人围脖(网易副总编辑)：38——726537 袁莉wsj(《华尔街日报》中文网主编)：38——323818 杨幂(演员，代表作《宫》《仙剑奇侠传三》等)：37——17195830 鹦鹉史航(知名编剧 策划人)：37——236633 王利芬(优米网的创始人)：35——1288020 彭浩翔(作家 编剧 制片人 导演)：34——1766663 王长田(王长田，光线传媒有限公司总裁)：33——2145382

红色：左边和右边共有的微博用户
蓝色：右边所独有的微博用户

图 6　粉丝数目与关注网度数中心度的对比

(3) 凝聚子群分析

根据上文中提到的派系的定义，计算出局域网络中派系成员个数为 33 个，接着研究“微博用户转发其他用户微博的情况”与“微博用户所属派系”之间的关系，通过分析样本局域网中微博用户最近转发微博的情况，绘制出在这 33 个派系上微博转发的情况，如图 7 所示：

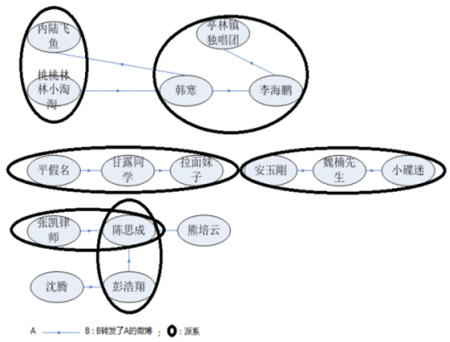

图 7　微博转发示意图

通过局域网络中派系与转发情况可以看出：同一个派系之内的微博相互转发比较普遍，转发微博这种行为发生在派系中的比例为 70%左右，说明微博信息在派系之间传递更为快速，处于同一个派系中的用户交流更为频繁。

(4) 网络关联性分析

使用公式（2）计算得出该局域网的关联度为 0.9927。说明除了个别点以外，大部分的用户之间是可达的。

Overall graph clustering coefficient: 0.359

Frequencies of Geodesic Distances

	Frequency	Proportio
1	8268.000	0.117
2	43447.000	0.614
3	17574.000	0.248
4	1428.000	0.020
5	32.000	0.000

Descriptive Statistics

		1
1	Mean	2.173
2	Std Dev	0.648
3	Sum	153756.000
4	Variance	0.420
5	SSQ	363870.000
6	MCSSQ	29718.174
7	Euc Norm	603.216
8	Minimum	1.000
9	Maximum	5.000
10	N of Obs	70749.000

图 8　网络关联性指数

从图 8 中距离频次数据（Frequencies of Geodesic Distances）中可以看出，距离是 1 的情况出现了 8268 次，距离是 2 的情况出现 42447 次，距离是 3 的情况出现 17574 次，也就是说距离在 1 到 3 之间的比例为 0.98。这说明绝大数人之间的距离在 3 以内。

通过观察网络距离的描述性数据（Descriptive Statistics），发现该网络的平均距离为 2.173，标准差为 0.648，方差为 0.420，最小距离为 1，最大距离为 5，也就是说在这个局域网络中，任何两个用户最多只需要 3 个中间人就可以建立联系了。该局域网络也从侧面印证了六度分离理论，说明该样本局域网是一个小世界网路。

4　结　论

本文主要介绍了如何使用新浪微博提供的开放接口爬取微博用户数据及用户之间的相互关注数据，然后根据用户之间的关注关系建立关注网络，通过对该关注网络的用户节点属性统计分析以及网络结构分析，得出了以下微博用户群体特征：

（1）用户与其关注对象的地域关联性很高；

（2）微博用户的关注对象往往和自己的职业领域有关，可以通过对“节点的度数中心度的入度”和“微博用户自身粉丝数目”之间的关系的分析来判断一个局域网的核心趋向；

（3）微博关注网络的密度大，群体之间的联系十分紧密；

（4）微博用户之间的互动（相互转发微博）往往发生在其所在派系之间，处于同一个派系中的用户交流更为频繁。

下一步的工作将会基于用户个人特征与网络结构特征，针对网络中的某一特定用户，向其推荐该用户潜在可能关注的用户对象。

参　考　文　献

[1] Kaplan, Andreas M.; Michael Haenlein (2010) "Users of the world, unite! The challenges and opportunities of Social Media". Business Horizons 53(1): 59–68.

[2] 兴亮，微博的传播机制及未来发展思考[J]. 新闻与写作，2010，(3)：43—46

[3] 李军，陈震，黄霁崴. 微博影响力评价研究. 信息网络安全，2012, (3).

[4] 张兴刚. 中文问答社区信息传播机制研究. 华东师范大学，2010.

[5] 罗方. 基于社会网络分析的Blog社区发现. 安徽工业大学，2011.

[6] 徐媛媛，朱庆华. 社会网络分析法在引文分析中的实证研究. 情报理论与实践，2008, 31(2)

[7] 李凌云，王海军，王佳. 虚拟实验研究领域作者合著的社会网络分析. 实验室研究与探索，2011, 30(11)

[8] 新浪微博，开放平台API文档. http://open.weibo.com，2012

[9] 刘军，整体网分析讲义：UCINET软件实用指南[M].上海：格致出版社，2009

第三次审稿修改说明

感谢贵刊的耐心细致地审稿，针对您提出的审稿意见，在第三版的基础上，我们对本文做出了如下修改。

1 修改全文组织，注意加强各节内容之间的贯穿与联系。

对全文中为同一事物的名词的名称进行统一，增加衔接各节内容的过渡性语句，下文中使用上文中已定义或者说明的概念、算法、公式；清晰阐明之前的数据获取、数据分析方法是为了后面实验的进行必须的数据准备与理论基础。

2 修改全文表述，注意算法形式化表述的严谨性。

对整篇文章的文字表述进行优化，修改了多处表达不恰当的地方，例如：

摘要：

“然后利用社会网络分析和其它数据统计分析方法”**改为**“然后利用社会网络分析和数据统计分析方法”

引言：

“用户通过便捷迅速的信息共享渠道(各种连接网络的平台),在任何时间和任何地点发布内容短小精悍的信息”**改为**“用户通过便捷迅速的信息共享渠道，在任何时间和任何地点发布内容短小精悍的信息”

“基于最新的微博数据实现对社会网络节点属性的统计分析，动态研究微博社会网络的结构特点，试图挖掘该网络群体中用户特点以及用户之间关系的实质”**改为**“基于最新的微博数据来实现对社会网络节点属性的统计分析，分析微博社会网络的结构特点，挖掘该网络群体中用户特点以及用户之间关系的实质，”

等等

算法：

修改了算法中逻辑错误的地方，增加了一些辅助步骤，简化了注释说明文字。