

基于微博的六度空间理论研究^{*}

刘宏杰¹, 陆浩², 张楠³, 郑晓龙²

(1. 陕西师范大学, 西安 710062; 2. 中国科学院自动化研究所, 北京 100190; 3. 中国科学院自动化学会, 北京 100190)

摘要: 通过研究微博用户群体领域范围中社交关系网络, 提出小社会理论猜想证明方法。以物联网为例选做系列对比实验, 选取随机基点验证六度空间理论在网络微社会中的正确性, 并结合蒙特卡罗法模拟小样本事件推至全网, 证明网络社会中的六度理论。该方法主要涉及领域范围内随机选取、最远不相关性的节点联系选择、最近相关节点的选取以及对比生成图的重合验证。

关键词: 六度空间; 微博客; 社会关系网络; 蒙特卡罗法模拟

中图分类号: TP309 **文献标志码:** A **文章编号:** 1001-3695(2012)08-2826-04

doi: 10.3969/j.issn.1001-3695.2012.08.006

Theory research based on microblog of six degrees space

LIU Hong-jie¹, LU Hao², ZHANG Nan³, ZHENG Xiao-long²

(1. Shaanxi Normal University, Xi'an 710062, China; 2. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; 3. Automation Congress, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper proposed a method to prove the theoretical conjecture of little society by studying the social networks among the microblog users, and did some comparative experiments of Internet of things as a case study. It selected random nodes combined with the simulation of small sample events by Monte Carlo method to validate the correctness of six degree theory in social networks. The proposed method mainly involved the random selection within the areas, the selection of the farthest unrelated nodes, the selection of the closest related nodes and the validation of generated figures of contrasting.

Key words: six degrees space; microblog; social network; Monte Carlo method simulation

0 引言

随着网络技术的发展, 微博^[1]以新型社会化媒体建立起一个微型社会关系网络, 作为一种崭新交流方式, 在人们生活中占据越来越大的比重, 微博交流已经成为日常生活中不可或缺的一部分。据2011年7月19日中国互联网络信息中心(CNNIC)在北京发布的《第28次中国互联网络发展状况统计报告》^[2], 截至2011年6月底, 中国网民规模达到4.85亿, 较2010年底增加2770万人, 增幅仅为6.1%, 网民规模增长减缓。2011年上半年, 我国微博用户数量从6311万快速增长到1.95亿, 半年增幅高达208.9%。随着微博客数几何增长, 微博客之间交流的日益简便, 微博系统俨然已经形成了一个微型小社会。

1967年Travers等人^[3]通过连锁信件实验提出了六度空间理论, 又称做六度分隔(six degrees of separation)理论。2001年Dodds等人^[4]通过电子邮件进行了一个现代版实验来验证该理论, 有超过六万个来自166个不同国家的志愿者参加了本次实验。由于社会关系网络的复杂性以及传播过程中的各种不可抗因素, 最终只有384(不到2%)封邮件成功地传达, 其

中完整的邮件链平均只有四人, 之后很多学者相继对该理论做了大量相关实验, 并取得了一定的研究成果^[5]。

由于社会关系中存在诸多制约条件, 在前人研究过程中仍存在许多不完善之处, 如连锁信件实验的实验者数量偏少, 电子邮件在实验传播过程中丢失; 培根与爱多士的小世界实验^[3]存在范围偏小等诸多制约条件。因此, 至今在社会化媒体中仍未取得结论性证明。

本文将通过对微博系统的监控以及对相应微博数据的采集分析, 建立一个微型小社会缩影, 以此来说明六度空间理论在微博小世界中的真实性^[6]。文中选取特定网络环境作为论证的先行条件, 在网络环境中社会复杂性干扰因素将被极大削弱, 网络连通性保证实验者数量恒定, 数据严密采集亦使社交传递的丢失率降到最低, 大范围的数据采集以及小范围的精确定位将实验从小范围延伸至全网。

实验成功绕过传统实验中存在的社会性制约条件:

a) 社会环境强弱连接关系——权值问题, 相对于现实世界中关系亲疏所造成的人为性传递误差, 即传递偏好。使用数据采集建立非否认性连接关系从根本上避免了因社会复杂性所带来的倾向性传导误差。

b) 传递成本和个人意愿代价——阻尼问题, 与传统依赖

收稿日期: 2012-01-03; **修回日期:** 2012-02-20 **基金项目:** 中国科学院规划与决策科技支持系统建设资助项目(2F11N06); 国家自然科学基金资助项目(71103180, 91124001)

作者简介: 刘宏杰(1988-), 男, 陕西潼关人, 硕士研究生, 主要研究方向为无线传感器网络、普适计算(tiangao yundannin@163.com); 陆浩, 男, 河北人, 工程师, 主要研究方向为复杂网络、社会计算; 张楠, 女, 河北人, 助理研究员, 主要研究方向为数据挖掘、文本处理; 郑晓龙, 男, 安徽人, 助理研究员, 主要研究方向为复杂网络、社会计算。

差异性个体传递和个体意愿变化期望而言,进行非主观依赖性实验,以连接的确性模拟传递过程以确保外在阻尼最小化。

微博采集系统随机采集小样本数据,通过数学模型递推至全网,扩大数据范围,为数据分析提供广域性数据连通,突破传统小圈子的局限性,分析数据不再局限于某一行业或者某一领域,突破旧观念中依靠人力传递所带来的各种分析误差,真正意义上实现了网状人际关系。

数据分析系统精密分析采集数据,确保数据随机性并以此寻找各个随机独立节点之间所隐藏的关系,建立准确清晰的可视化图像,把握节点关联关系。

借助蒙特卡罗法模拟论证小范围事件,并递推至全网,完成理论证明。

1 特定社会环境构建

随机选取特定网络社会环境,实验通过随机范围选取方式来搭建虚拟社会环境,以此保证数据代表广泛性。本次实验以物联网为话题,检索微博系统中所有存在与话题相关的节点以此构建博主集,博主集模拟社会环境自然人,通过搜索博主集内任意博主之间关系的方式来论证小世界理论。

1.1 媒体选择

针对网络环境中存在的众多微博传媒,综合其关注度、更新频率、访问方式(是否开放API)、PR排行以及影响力选取采集媒体。

图1为排名靠前的三个网站数据指数分析(来自百度指数)。

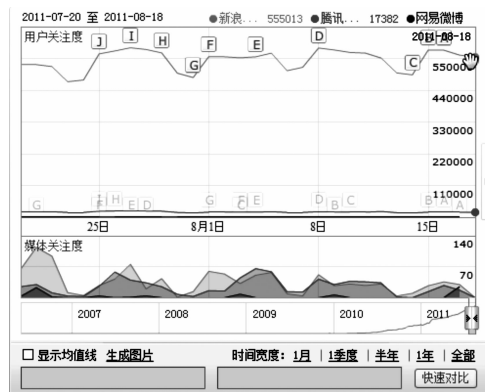


图1 网站数据指数

综合考虑各项参数,本文最终选取新浪微博为采集网站,以确保数据充分准确及涵盖的范围。

1.2 话题建立

微博话题是以当前社会生活中存在的种种事件为讨论对象,在微博中建立一个针对事件或者能够代表该事件的名称,即为一个话题,也可将其理解为微博搜索中的一个关键字。为实现话题建立公正性,本文特意选取了近十个话题,通过数量检测,对热门讨论话题以及冷门话题进行综合考虑,选取介于两者之间的中值话题作为本次话题选取的对象(中值话题,即话题热度排名居中的话题)。根据数据采集过程中重复数据概率计算,排名第一话题“铁道部”其二层节点冗余最高达到47%,而排名最低的话题“物联网”则只有5%的冗余。综合考虑所需数值及其冗余度,选择中值话题,可以避免热门话题中过多无关联数据所造成的数据冗余,亦可避免生僻话题相关

数据过少难以分析的窘境。综上,本文最终将微博话题定为“物联网”。详细见表1。

表1 微博站点热度

话题	热度	时间段
铁道部	292 715	7.26 ~ 8.1
变形金刚	43 132	7.26 ~ 8.1
香格里拉	25 908	7.26 ~ 8.1
实话实说	21 462	7.26 ~ 8.1
郝海东	8 327	7.26 ~ 8.1
基金会	10 437	7.26 ~ 8.1
爱因斯坦	8 783	7.26 ~ 8.1
物联网	17 080	7.26 ~ 8.1

注:以上数据来源于 weibo.com

1.3 博主抽取方式

以内容检测为主,设计实现采集程序搜索话题“物联网”,以“物联网”为关键词进行搜索获得所有内容相关博主信息或名称相关博主信息,建立节点,形成节点集,构造初级自然环境圈。本次采集获得与上述话题相关的1 249个博主,其中不重复数据为728,重复率为41%。

1.4 博主粉丝抽取

以选定节点为中心,分别对本文选择博主“中国国际物联网大会”和“移动物联网”进行深度遍历采集五层粉丝博主,建立粉丝集合,保存粉丝博主连通关系,构建次级自然环境圈。其中“移动物联网”采集获得粉丝博主数418 217,其中不重复数据为395 116,重复率为5.5%。

使用博主及其粉丝共同搭建的网络环境,模拟现实社会环境中人与人之间关系,形成社交网络。

1.5 数据处理

对数据进行去重处理,博主和粉丝之间相互均有大量相关关系,所以在采集过程中会有一些重复数据,因此结束采集后,应对数据库进行数据去重操作,去掉所有重复节点。去重处理后一级博主节点为728个,重复率为41%。图2为搭建特定社会环境的过程。

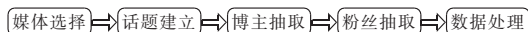


图2 特定社会环境构建

2 数据分析网络图构建

2.1 数集建立

以博主集作为起始节点,通过对其名下粉丝进行深度遍历,涵盖整个范围内节点,构成节点集,为分析提供原始数据。

2.2 数据分析

寻找博主集中任意两个节点的相关关系,以部分连通跳数构建正态分布,根据蒙特卡罗法模拟计算小样本事件^[4]最终获得特定社会环境内各节点的连通性。

2.2.1 任意节点选取

多次随机选择不同的节点进行样本分析,本例中,博主集存在728个非重复性博主(2.3节中获得),因分析过程基本一致。本文只给出一次分析的具体过程以及多次分析的数据结果。

以“中国国际物联网大会”和“移动物联网”作为主要分析

2.2.2 数据分析处理

a) 以 1 号节点为中心遍历其粉丝节点, 遍历深度为五层, 每层获得的子节点与 2 号节点进行匹配, 若是相同则将该连通保存为一条路径; 若匹配不成功, 则继续寻找。对一号节点循环结束时, 选择最小跳数作为其最短路径保存转向 b)。若未获得任意一条通路则将最小跳数设置为 6 转向 b)。

```
String[] path = new ArrayList[] ;
int pathNum = 6;
for (int i = 0; i < list1. size() ; i ++ ) {
    if (i. getName() == 2. getName() && pathNum > ( i. getLevel() +
1)) {
        pathNum = i. getLevel() + 1 ;
        int num = pathNum;
        path[ num ] = i. getName() ;
        num--;
        if (num > = 1 ) {
            path[ num ] = i. getParentName() ;
            num--;
        }
    }
}
void loop(6,0,pathNum) ;
```

```
String[] path = new ArrayList[];
void loop( int level1 ,int level2 ,int pathNum) {
    Level1 = level1--;level2 = level2 + 1;
    for( int i = 0; i < list1. size() && list1( i ). getLevel( ) < level1 ; i ++ )
    {
        for( int j = 0; j < list2. size() && list2( j ). getLevel( ) < = level2 ;
        ) {
            if( i. getName( ) == 2. getName( ) && pathNum > ( i. getLevel( )
            ) ) {
                pathNum = i. getLevel( ) + 1;
                int num = pathNum;
                path[ num ] = i. getName( ) ;
                num--;
                if( num > = 1 ) {
                    path[ num ] = i. getParentName( ) ;
                    num--;
                }
            }
        }
    }
}
```

```

if( level2 >= 5 ) {
    break;
}
void loop( level1 , level2 , pathNum );
}

```

又如,寻找“中国国际物联网大会”与“移动物联网”之间的关联,通过 a) 没有直接通路到达,故设置最小跳数为 6,对“移动物联网”的粉丝节点进行遍历,查找到三条不同路径,如图 3 所示。

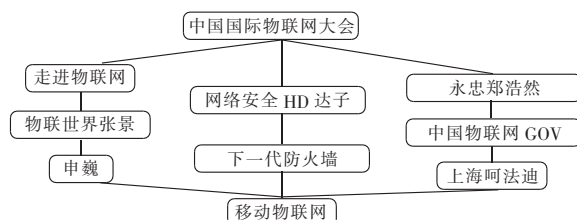


图 3 路径

“中国国际物联网大会,网络安全 HD 达子,下一代防火墙,移动物联网”,最小跳数为4。

此处需要说明的是:因为网络问题和网站本身的问题,粉丝采集过程可能存在个别遗漏粉丝,虽不能确保所得路径一定是最短路径,但亦相差不远。

2.2.3 最小跳数集建立

重复选取多个不同节点进行数据分析,观察不同节点最小跳数分布情况,提取所有涉及最短路径节点构建最小跳数集。

2.3 网络图构建

实验通过 Java 图示工具 Prefuse^[7] 生成实验图, 如图 4 所示。

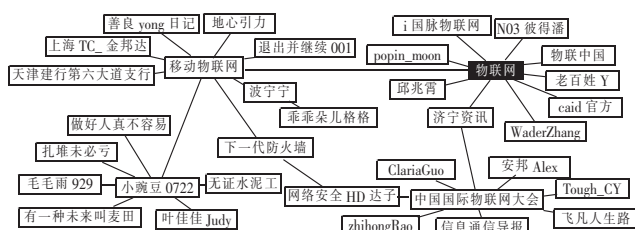


图 4 实验结构

注:图4中以“中国国际物联网大会”和“移动物联网”为中心,构建其粉丝结构图,生成网状关系图(为了方便显示,图中删除了许多不相关节点)

实验路径标注如图 5 所示。

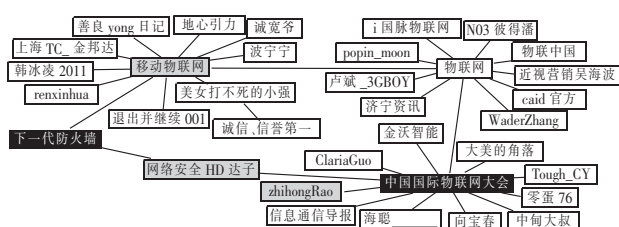


图 5 实验路径标注图

注:图 5 中将 1 号节点与 2 号节点之间路径进行路线标志

“移动物联网”和“中国国际物联网大会”均是“物联网”的一级博主,1号节点和2号节点通过其粉丝建立联系,为方便查看,图5中仅取了部分粉丝节点的粉丝进行标注。

通过该连接实验,找出其连通路经:“移动物联网,下一代防火墙,网络安全 HD 达子,中国国际物联网大会”。

3 结论

3.1 结论分析

根据最小跳数集建立概率分布图,如图6所示。根据蒙特卡罗法模拟计算小样本事件可靠性^[8]分析,由多次小样本事件建立柱状图可推出其在全网范围内的可靠性。

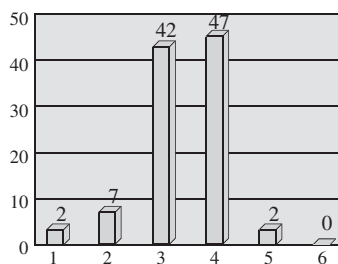


图6 关联跳数分布

蒙特卡罗的基本思想是:当所求解的问题是某事件出现的概率时,模拟法首先建立一个概率模型,通过用数字进行的假象实验得到抽样值,经统计处理,以其结果转换为问题的解。针对本例而言,随机抽样抽取100个一级博主,观察分析跳数分布,最大为6度,最小为直接相连节点为1度,其中直接相连节点占1%,跳数为2的节点占2%,跳数为3的节点占41%,跳数为4的节点占45%,跳数为5的节点为1%,不存在跳数为6的节点。此处需要说明的是根据文中算法,查找为最短跳数,并非不存在跳数为6的连通路经,但因关注的是最小跳数所以跳数为6的节点均已被覆盖。

多次抽取样本进行构建图分析,通过对多次图像的重合对比,去除杂乱的异常节点发现其概率分布完全符合图8的概率分布。

由上文分析可知,通过蒙特卡罗法模拟计算小样本事件可推知在以该话题为前提的网络关系中,各个一级博主之间均存在着小于6度连通路经。而各个一级博主名下存在的粉丝亦可通过上述途径加以证明。

至此,同一话题下的博主及粉丝关系得证。结合上述可知,基于微博的六度空间理论在以话题为前提的小范围环境中成立。

3.2 结论扩展

另外,从数据角度而言,笔者发现绝大部分节点之间的关系都停留在3~4跳之间,而相对于话题而言,一个博主节点或者粉丝节点往往对应很多话题,以“物联中国”为例,其名下拥有1829个微博,每个微博涉及话题以2个计算,即为3658个相关话题,所以一个博主可以跨越的领域并不仅仅局限于采集话题处。在本文中以“物联网”为话题的一级博主拥有728个,若仅以每个节点拥有2000个相关话题计算,则一层节点可涉及的话题数即为 $728 \times 2000 = 1\,456\,000$ 。通过不同话题,

可以建立的相关联系的节点可达到 $1456000 \times 728 = 1\,059\,968\,000$ 个。据新浪新闻中心2011年6月29日统计,新浪微博客占中国国内微博用户的57%,约1.4亿,根据数据对比,一级博主已经足够涵盖新浪微博中所有的微博客。当然,不排除前面计算中存在的大量重复节点,以前文计算的重复率41%而言,不重复的一级博主约有 $1059968000 \times 41\% = 423\,987\,200$,亦足以覆盖整个新浪微博用户。当然,重复数据会随着考察用户的增多而增多,但就以六度空间呈几何数的增长而言,其网络涵盖的范围也远远大于重复数据的增长。

因此,从上述数学涵盖范围角度考虑,亦可判定微博站点中的节点在六度范围内存在联系。

4 结束语

本文提出了基于微博的六度空间理论研究,一方面对前人算法中不可避免的社会复杂性现实约束条件进行了削弱,使用非否定性连接作为本次实验的前提条件,很大程度上避免了因人为因素造成的传递误差。另一方面在该体系下使用局部范围内的随机小样本事件作为本次实验的考察范围,针对小世界理论^[9]范围内的局限性,提出通过蒙特卡罗法模拟计算进行范围推广,并联合数据统计量进行二次验证,最终将结论推至整个微博系统乃至全社会。

本文下一步工作将考虑将蝴蝶效应^[10]引入微博系统,观察微小事件在全网范围内所引起的量变质变交替,最终从消息在网络中传播的角度证明六度空间理论。

参考文献:

- [1] 潘跃华. Web2.0时代的网络新闻传播[J]. 新闻爱好者:理论版, 2009(4):37-38.
- [2] 中国互联网信息中心. 第28次中国互联网络发展状况统计报告[R/OL]. (2011-07-19). http://www.cnnic.net.cn/dtygg/dtgg/201107/t20110719_22132.html.
- [3] TRAVERS J, MILGRAM S. An experimental study of the small world problem[J]. *Sociometry*, 1969, 32(4):425-443.
- [4] DODDS P S, MUHAMAD R, WATTS D J. An experimental study of search in global social networks[J]. *Science*, 2003, 301(5634):827-829.
- [5] The Erdos Number Project. Information about Grossman's theory[EB/OL]. <http://www.oakland.edu/enp/>.
- [6] NEWMAN M E J, WATTS D J. Scaling and percolation in the small-world network model[J]. *Physical Review E*, 1999, 60(6):7332-7342.
- [7] The prefuse visualization toolkit[EB/OL]. (2012-01-13). <http://www.prefuse.org/>.
- [8] 吴社群,何建国,蒲洁. 蒙特卡罗法模拟计算小样本事件可靠性[C]//2004年全国机械可靠性学术交流会论文集. 2004.
- [9] 李军利,赵红领,范明. 邮件社区划分和小世界网络[J]. 计算机应用, 2008, 28(6):146-149.
- [10] CAHYADI C, CHAN L W, COLOMBO P, et al. The butterfly effect: a physical phenomenon of hypromellose matrices during dissolution and factors effecting its occurrence[J]. *International Journal of Pharmaceutics*, 2011, 406(1-2):31-40.