doi:10.3969/j.issn.1000-436x.2012.z1.013

微博中节点影响力度量与传播路径模式研究

于洪，杨显
（重庆邮电大学 计算机科学与技术学院，重庆 400065）

摘 要：针对微博信息传播特点刻画微博传播网络并定义节点影响力来描述节点重要性，从而分析微博信息传播路径模式。首先采集单条微博的转播/评论数据并进行预处理；然后给出微博传播网络的形式化描述，从局部和全局 2 方面定义节点的影响力来刻画节点的重要性；比较实验说明新定义的影响力度量方法方法是可行的。同时，结合影响力度量给出了微博网络中信息传播路径的几种典型模式；采用软件 NodeXL 的可视化结果说明给出的信息传播路径模式具有一定的典型性。

关键词：微博；节点影响力；信息传播；路径
中图分类号：TP391 **文献标识码**：A **文章编号**：1000-436X(2012)Z1-0096-07

Studying on the node’s influence and propagation path modes in microblogging

YU Hong, YANG Xian
（Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China）

Abstract: In order to analyze microblogging information propagation path modes, the node's influence to reflect its importance was defined based on considering the characteristics of information propagation in microblogging. Firstly, the broadcast/comment data of a microblogging was collected and preprocessed. Then, the formal description of the transmission network was given, and the node’s influence was defined which reflects its significance in the local and global aspects. The results of comparative experiments show that the new definition is reasonable. Besides, some information propagation path modes were proposed, which combine the measurement of influence. Finally, the results by using visualization software-NodeXL show that the information propagation path modes are typical.

Key words: microblogging; node’s influence; information propagation; path

1 引言

作为 Web 2.0 的产物，微博在信息传播的速度和广度方面发挥着不可小觑的作用，也是其他媒体不能胜任的。一个非常重要的原因在于微博所具有的转播功能，信息通过粉丝以及粉丝的粉丝不断在传播网络中转发，从而达到信息扩散的目的。

相对于其他社交网络，如人人网，微博用户与其粉丝之间存在的更多的是一种“弱关系”，其特点是个人和社会网络异质性较强、人与人的关系并不紧密、无太多的感情维系[1,2]。表 1 是对人人网和微博区别的总结。可见，微博中信息传播路径具有自身的一些特点，了解这种单方认可的“弱关系”是如何促使信息不断向外扩散以及不同节点对信

收稿日期：2012-08-06
基金项目：国家自然科学基金资助项目(61073146,61272060)；重庆市自然科学基金资助项目(cstc2011jjA40045)
Foundation Items: The National Natural Science Foundation of China (61073146, 61272060); The Natural Science Foundation of Chongqing (cstc2011jjA40045)

表 1　　人人网与微博的区别

	人人网	微博
关注关系	双方认可（面对面）	单方认可（面对背）
圈子	是熟人小圈子	大圈子
功能	以社交为主	以媒体（信息传播、获取及分享）为主，社交为辅
信息传播速度	慢	极快
公开性	信息在小圈子内封闭	信息全面开放（内容、关系）
话题内容	就人论事，可信度高，更多局限在个人生活话题	就事论事，大众化，发布所见所闻，新闻八卦、言论表达场所
互动模式	以人为核心	以话题为核心

息传播的影响，都将有助于进一步建立起微博中信息传播路径模式。

另外一方面，对于微博中信息传播路径模式的研究非常有意义：对于个人，刻画微博信息的具体传播路径可为用户提供一个客观的依据，更好地为其服务；对于企业，使其可分析已有的传播模式是否适用于现阶段的受众人群，找出扩大微博影响力的突破点，从而优化资源的分配、利于产品的营销和推广；对于社会，社会团体和政府机构可以根据信息传播特点和规律来做好信息的发布、提高管理效率和透明度，也可据此进行一定的信息筛选和过滤，合理引导社会舆论。

因此，本文在刻画微博传播网络和定义节点影响力的基础上，寻找出网络结构中的重要节点，从而分析微博信息传播路径模式。本文首先描述了如何采集单条微博的转播/评论数据以及预处理方法。第 2 节给出了微博传播网络的形式化描述，从局部和全局 2 方面定义节点的影响力来刻画节点的重要性，实验表明：用影响力衡量节点的重要性与中心性衡量的结果具有一定程度的相似性，该方法在一定程度上是可行的。第 3 节结合影响力的定义提出了一触即发传播模式、多级传播模式、多点触发传播模式和混合传播模式几种典型传播路径模式，并通过可视化软件 NodeXL 对采集到的数据分析验证提出的微博信息传播路径模式的典型性。

2　数据采集及处理

为了能清晰地认识微博中信息向外扩散的过程，即信息传播的具体路径，本文首先随机选取了部分具有一定影响力的微博，通过 API（application programming interface）对其转播/评论列表数据进行获取；然后，对数据进行预处理；随后对预处理的结果进行可视化表示与分析。微博信息传播路径是指一条微博发布后，为达到信息扩散的目的，如何通过博主的粉丝以及粉丝的粉丝，在微博网络中所形成的具体传播路线。

2.1　数据采集

本文主要针对腾讯微博进行数据获取。使用腾讯 API，首先需要通过用户认证，因此，了解 OAuth 协议至关重要，它为用户资源的授权提供了一个安全、开放而又简易的标准。认证流程及访问资源流程可参见文献[3]和文献[4]。

通过 OAuth 认证后，就可通过 API 请求，返回 XML 或 JOSN 格式的数据。XML 是一种跨平台的强结构性标记语言，用来标记数据、定义数据类型，是一种允许用户对自己标记语言定义的源语言，非常适合 Web 传输。因此，本文采用 XML 格式来返回数据。

对微博的转播/评论数据采集过程如下。

Step1　根据选取的微博 URL 得到 ID，如：http://t.qq.com/p/t/6732096641648，ID=6 732 096 641 648。

Step2　得到 ID 后，再通过 API 获取该条微博的相关属性，包括：微博内容、转播次数、评论次数、发表人账户名、是否原创、发表时间、是否认证微博等 28 个属性。

Step3　根据步骤 Step 2 得到的转播次数和评论次数来确定获取该条微博转播/评论数据的循环次数，然后设置参数规则，通过 http://open.t.qq.com/api/t/re_list 来获取转播/评论列表，并用 txt 格式文件保存。当 API 调用次数不够时，之后程序调用结果将会返回为空。要想获取相应的数据，可对 API 的调用进行定时读取。

2.2　数据预处理

由于当粉丝对微博原文直接进行转播/评论时，粉丝可根据需要向输入框添加文本，添加前文本里

无内容，如图 1 所示。当粉丝对微博进行间接转播/评论时，文本框里会自动以“||@name：xxxxxx”格式添加进去，name 表示该条微博信息的来源，即用户通过谁转播的该条微博，xxxxxx 表示 name 对该条微博添加的文本，如图 2 所示。

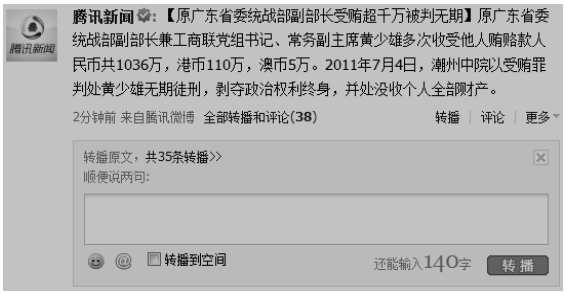


图 1 对微博进行直接转播

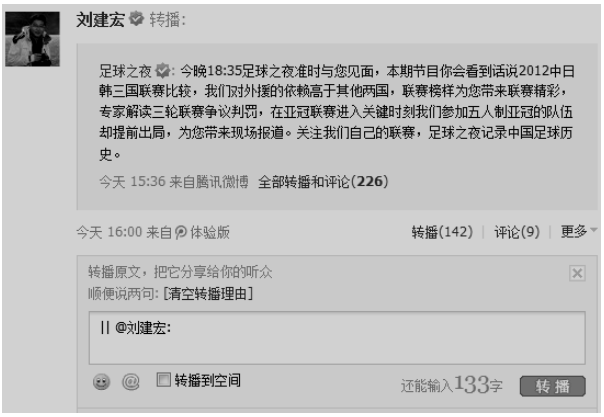


图 2 对微博进行间接转播

因此，在对通过 API 获得一条微博的转播/评论列表进行预处理时，XML 用标签<origtext></origtext>保存用户的转播/评论内容：当内容中不含有“||@name xxxxxx”时，则判断为直接转播/评论，信息来源者为标签<source> </source>的子标签<name></name>中内容；否则为间接转播/评论，信息来源者为标签<origtext> </origtext>中第一个“||@”与“：”之间的 name；信息到达者为标签<info></info>中的<name> </name>。如果用户进行间接转播时，删除了系统自动为其添加的转播/评论来源信息时，则本文视为对微博原文进行直接转播/评论。

数据预处理后的格式为“信息来源者 信息到达者”，并用 txt 文件格式保存。

3 节点影响力度量

在信息传播的每一个网络中，由于用户自身属性及其所在网络中位置的不同，使得它在信息传播过程中的作用也大相径庭。因此，如果能找到个人或组织在微博信息传播网络中具有什么样的影响力或者说居于什么样的中心地位，都具有一定的实用价值，这对网络舆情和营销具有一定指导作用：当需要对信息进行传播时，合理地对节点进行布局，有利于信息的扩散，延长信息传播时间，如活动策划、产品推广等；当需要对信息进行遏制时，通过对关键节点的监管和控制，从而破坏网络传播的结构达到控制的目的[5,6]。

为了描述方便，本节首先给出了微博信息传播网络的形式化描述，随后在分析微博信息传播自身特点的基础上给出了节点影响力的度量方法，最后通过实例分析给出了度数中心性、中间中心性、接近中心性和本文提出的影响力 4 种指标对节点重要度的刻画结果，比较结果说明本文定义的影响力是有效的。

3.1 微博信息传播网络

首先用有向图 $G=\{V,E\}$ 表示微博信息传播的实际网络：V 是微博节点集合，即用户节点的集合，节点可发布微博信息，也可对其感兴趣的信息进行转播/评论；E 是连接微博网络中节点与节点之间的边所组成的集合。

V 也可以表示为微博信息原创节点与其他传播节点的并，即 $V=\{v_0\}\bigcup V'$，$V'=\{V_1,\cdots,V_i,\cdots,V_n\}$。其中，$v_0$ 表示微博信息原创节点，V_1 表示信息由原创节点 v_0 发出直接到达节点组成的集合（路径距离为 1），即一级传播节点集合；V_2 表示信息由原创节点 v_0 发出并且经过 V_1 中某些节点所到达的节点组成的集合（路径距离为 2），即二级传播节点集合；其他 V_i 含义以此类推。显然，对于 $\forall V_i,V_j$，$V_i\bigcap V_j=\varnothing$。

设 $v_k\in V_k$，即 $v_1\in V_1,v_2\in V_2,\cdots$，则称 $v_0\to v_1$ 为一级传播路径；$v_0\to v_1\to v_2$ 为二级传播路径，其他以此类推。

3.2 节点重要性

评估网络中节点重要性的方法大都以图论为基础，目前主要的研究方法[7,8]有以下 3 种。

1) 社会网络分析法。该方法将节点的重要程度等同于和其他节点的连接而使其具有显著性，常用指标包括度数中心性(degree centrality)、中间中心性(betweenness centrality)、接近中心性(closeness

centrality)。度数中心性指节点在网络中与其他节点有直接联系，那么该节点就处于中心地位；中间中心性指一个节点位于多个交往网络的路径上，那么该节点具有控制其他节点间交互的能力，则认为该节点居于重要的位置；接近中心性考察一个节点不受其他节点控制的程度，节点离其他节点越近，则节点在网络中越不依赖其他节点。

表 2 为上述 3 种指标的计算方法；此外，还有特征向量、网络直径、累计提名等[9]指标。表 2 中：$b_{jk}(i)=g_{jk}(i)/g_{jk}$，即 i 处于点 j 和 k 间的最短路径上的概率 $b_{jk}(i)$ 为点 j 和 k 之间存在的经过点 i 的最短路径数目 $g_{jk}(i)$ 与点 j 和 k 之间存在的最短路径数目 g_{jk} 之比；d_{ij} 表示节点 i 与 j 之间距离。

表 2　　中心性计算

度数中心性	中间中心性	接近中心性
$C_{Di}=i$ 的度数	$C_{Bi}=\sum_{j}^{n}\sum_{k}^{n}b_{jk}(i), j\neq k\neq i$	$C_{Ci}=\left(\sum_{j=1}^{n}d_{ij}\right)^{-1}$

2) 系统科学分析法。它将删除网络中某个节点所造成的破坏程度等同于节点的重要性，该方法没有完全体现网络拓扑结构的差异，因此对节点重要性评估不是很准确[8]。

3) 信息搜索领域分析方法。代表性算法是 Larry page 和 Sergey Brin 提出的 pagerank 算法以及和 Kleinberg 提出的 HITS 算法。

与此同时，能否从其他角度对节点的重要性进行度量来反映其综合性能，比如充分结合微博信息传播自身的一些特点，将是本文要思考的问题。比如在这个有向网络中：

① 和节点 i 直接相连的节点越多，也就是说该节点的出度（*outdegree*）越大，显然该节点号召力越强；

② 和节点 i 间接相连通的节点越多，那么在含有节点 i 的传播路径上所承载的信息量越大；传播路径的距离越远，那么节点 i 的传播力度显然也越深。

因此，结合上述 2 个特点，从局部和全局两方面出发，本文用节点 i 的影响力来刻画节点的重要性。

定义 1　节点 i 的影响力　节点 i 的影响力定义为节点 i 的号召力与信息承载平均距离的乘积。即：$Influence(i)=outdegree(i)\cdot\sum_{j=0}^{n}d_{ij}\Big/count(i)$。

其中，$Influence(i)$ 表示节点 i 的影响力，j 表示节点 i 能与其连通的节点，d_{ij} 表示节点 i 与节点 j 之间的距离；$count(i)$ 表示从节点 i 出发与其连通的其他所有节点的个数。$outdegree(i)$ 反应的就是节点 i 的号召力，主要是从局部来考虑节点的影响力；$\sum_{j=0}^{n}d_{ij}\Big/count(i)$ 表示的就是信息承载平均距离，主要是从全局来考察节点的影响力。

3.3　比较实验结果与分析

为了说明 3.2 节中用影响力刻画节点重要程度的合理性，在此用一个简单的示例进行说明。

图 3 是一条微博信息的传播网络，节点 1 是原创节点。以节点 57 为例观察，其出度为 5，信息承载平均距离=（1×5+2×8+3×2）/ 15=1.8，那么节点 57 影响力为 9，其他节点以此类推。

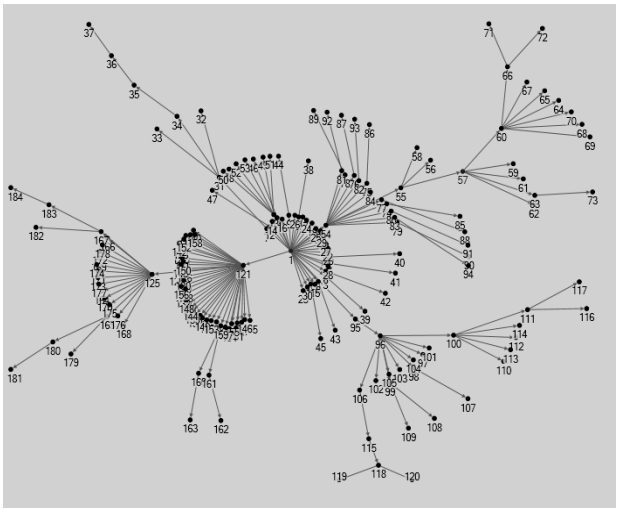


图 3　微博信息传播网络例子

以图 3 这个微博信息传播网络为例，分别用度数中心性、中间中心性、接近中心性和本文定义的影响力度量方法测量网络中各点的重要性，表 3 给出了 4 种度量方法排名前十的统计结果。其中，表 3 中的元素记录的就是对应测量方法相应排名的节点标号，例如度数中心性方法中，排名为 1 的节点就是标号为 121 的节点。

由此可看出，排名前十的影响力节点分别与度数中心性、中间中心性、接近中心性排名前十的节点具有相同的节点个数为 11、8、6。为了说明方法的合理性，在更多的微博信息传播网络上进行了统计，实验结果说明用影响力衡量节点的重要性与 3 种中心性衡量的结果具有一定程度的相似性，因此该方法在一定程度上是可行的。

表 3　4 种节点重要性度量方法比较结果

排名	度数中心性	中间中心性	接近中心性	影响力
1	121	1	1	1
2	1	121	121	121
3	125	54	54	54
4	54	96	95	125
5	96	95	7	96
6	60	125	13	57
7	13	55	2~6,8~12	60
8	57,100	57	14~30	55
9	31,55	60	125	13,31
10	66,99,111,118,159,165,167	100	159	159

4　微博信息传播路径模式

有了对微博信息传播过程中节点的影响力的分析后，就可以建立起微博信息传播路径模式。为了能对微博信息传播路径有一个直观的认识，本文将预处理后的数据用软件 NodeXL 进行了可视化表示。NodeXL 是一个免费、开源、功能强大且易于使用的交互式网络可视化和分析工具，它以 Microsoft Excel(Excel 2007/2010)模板的形式，利用 Excel 作为数据展示和分析平台；NodeXL 可定制图像外观、无损缩放、移动图像、动态过滤顶点和边，提供多种布局方式，查找群和相关边，支持多种数据格式输入和输出[10]。

通过导入预处理后的数据至 NodeXL 中，观察发现一条微博信息的传播路径具有一定的规律。本文将其称为微博信息传播路径元模式：一触即发传播模式、多级传播模式或多点触发传播模式。一般说来，一条典型的微博信息传播路径要么是元模式之一，要么是某几种元模式的组合形式。下面给出元模式的定义。

定义 2　一触即发传播模式　如果$V=\{v_0\}\cup V'$，其中，$V'\approx V_1$，n 基本上等于 1，称之为一触即发传播模式。

该传播模式由微博原创节点 v_0 带动一级传播节点 V_1，即 v_0 的粉丝 V_1 进行信息的转播/评论，V_1 中各节点关系平等。信息由影响力大的原创节点 v_0 向外扩散，传播路径呈发散状，且每条分支路径长度较短，用户基本都停留在一级传播节点的位置，即 $V'\approx V_1$。一触即发式的传播模式由于信息原创节点具有一定影响力带动其粉丝进行快速转发，而转播/评论粉丝普遍是草根平民，传播路径呈现出发散状，信息停留在一级传播节点位置就基本终止。

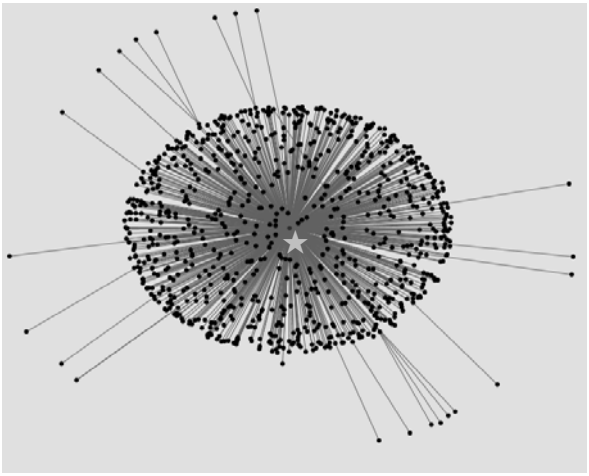


图 4　一触即发传播模式例子

图 4 是“中国新闻周刊”发出的一条关于“核恐怖”的微博。显然，其传播路径模式是典型的一触即发式。其中，五角星节点表示微博原创节点。

定义 3　多级传播模式　如果$V=\{v_0\}\cup V'$，$V'=\{V_1,\cdots,V_k,\cdots,V_n\}$，其中，$n>1$，存在节点 $v'\in V'$，且当 $\left|Influence(v')-Influence(v_0)\right|>\tau$ 时，称之为多级传播模式，其中，τ 为阈值。

该传播模式主要由原创节点 v_0 向外扩散，由于 $n(n>1)$ 级传播节点集合 V' 中某些节点自身具有一定程度的影响力，会带动自身粉丝再次对信息进行转播/评论，从而扩展了信息传播的广度和深度。

例如图 5 所示，该条微博由“公安部打四黑除四害”发出的一条关于“爱心接力-寻人”的微博，其一级传播、二级传播非常明显，三级传播已初步形成，一方面源于粉丝传播的质量，另一方来自微博信息内容的吸引度。由图 6 可见，该传播网络的 4 级传播已初步形成，其中，传播得最远的节点已经达到 11 级，用灰色标注的节点对本次传播深度具有重要作用。

定义 4　多点触发传播模式　如果 $V=\{v_0\}\cup V'$，$V'=\{V_1,\cdots,V_k,\cdots,V_n\}$，其中，$n>1$，存在节点 $v'\in V'$，且当 $\left|Influence(v')-Influence(v_0)\right|\leqslant\tau$ 时，称之为多点触发传播模式。

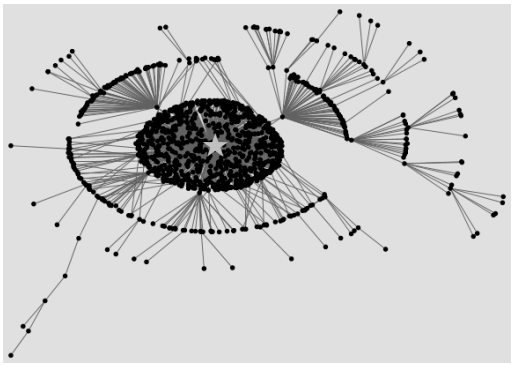


图 5　多级传播模式例子 1

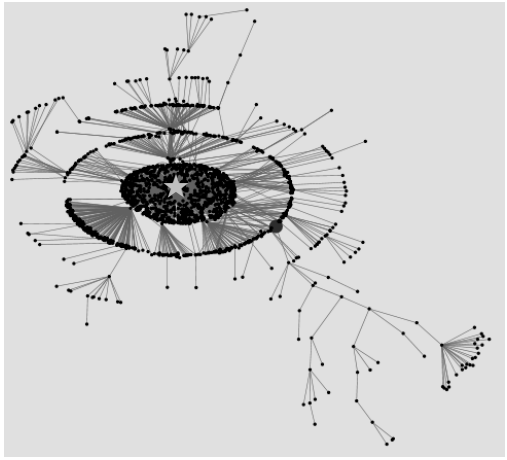


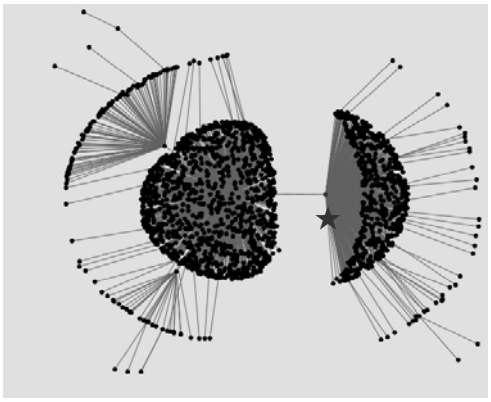
图 6　多级传播模式例子 2

该传播模式由原创节点 v_0 发出信息后，由于节点自身的影响力会带动其粉丝 V' 进行一定数量的转播/评论，但是在 V' 中，部分节点同样具有较大的影响力，甚至比它的影响力还大、粉丝数量还多。因此，在 $n(n>1)$ 级传播后，传播效果又会进一步扩大，乃至达到更强的传播效果。从而形成了多个影响力较强的节点相互呼应的传播局面，扩大了信息的覆盖面。

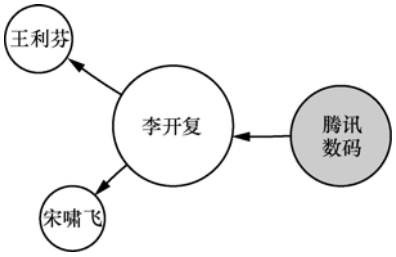
例如图 7(a)所示，由“腾讯数码”发出的一条关于“苹果 CEO 库克现身北京苹果专卖店”的微博，当信息到达一级传播节点时，其自身粉丝带来一定数量的转播/评论，V_1 中“李开复”对该信息进行转播进一步扩大了信息的传播力度，而“李开复”的粉丝 V_2 中由于“王利芬”和“宋啸飞”的转播，使得信息继续向外传播。图 7(b)简单地描述了上述信息传播网络中的关键人物（节点）转播关系。

当然，大量微博传播路径模式更多的是混合模式，即可能包含以上 3 种模式中的多种模式。

如图 8(a)所示，它是一条由“佛山市公安局”发

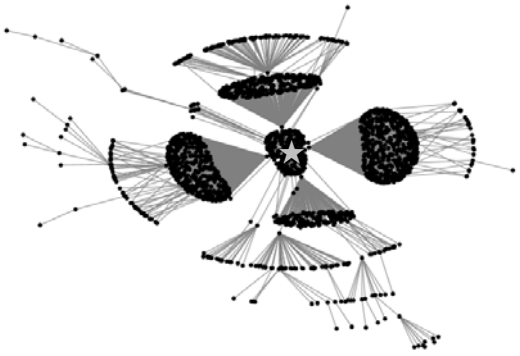


(a)传播关系

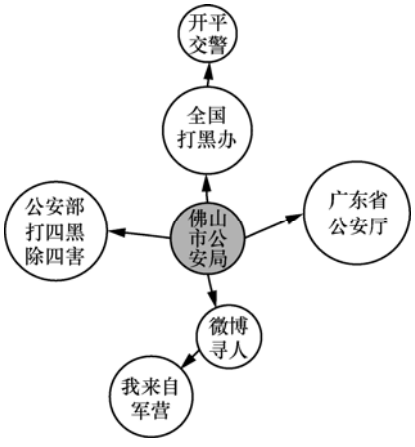


(b)传播网络关键节点

图 7　多点触发传播模式例子



(a)传播关系



(b)传播网络关键节点

图 8　混合传播模式例子

出的一条关于“寻人”的微博。微博发出后，在一级传播路径上形成多点触发模式，图 8(b)更直观地描述了在该模式上主要由“全国打黑办”、“广东省公安厅”、“公安部打四黑除四害”、“微博寻人”4 个影响力较大的用户构成，然后通过这 4 个用户又按照多级触发传播模式促使三级传播形成，如“我来自军营”这个用户又引起了下一级传播。

在本文的实验数据中，阈值$\tau \in [20,50]$时，对多级传播模式和多点触发模式的区分较为合理。从本节分析可见，利用本文定义的节点影响力来刻画微博信息传播路径的模式具有一定的典型性。

5 结束语

微博中信息传播路径模式的研究对于资源的优化分配、产品的营销和推广、信息的发布管理、信息的筛选和过滤等意义重大。因此本文在分析微博信息传播特点的基础上定义了节点影响力来刻画节点的重要性，从而分析微博信息传播路径模式，给出了一触即发传播模式、多级传播模式、多点触发传播模式和混合传播模式几种典型传播路径模式的形式化描述，并通过实例分析验证影响力定义的合理性，使用可视化软件 NodeXL 对采集的微博数据的分析结果也验证了提出的微博信息传播路径模式的典型性。如何从传播网络中自动发现这些典型传播路径模式将是下一步工作的重点。

参考文献：

[1] EYTAN B. Facebook research report: the importance of social network of weak ties[EB/OL].http://tech.sina.com.cn/i/2012-01-18/ 13286651169.shtml, 2012.

[2] 张玮.透析人人网：大学生使用 SNS 的传播学意义分析[D]. 成都：西南交通大学, 2009.
ZHANG W. Analysis the Communication Significance of College Students Using SNS by Renren Network[D]. Chengdu: Southwest Jiaotong University, 2009.

[3] Development documents of Tencent open microblogging platform [EB/OL].http://wiki.open.t.qq.com/index.php/API%E6%96%87%E6%A1%A3, 2012.

[4] 廉捷, 周欣, 曹伟. 新浪微博数据挖掘方案[J].清华大学学报(自然科学版), 2011, 51(10):1300-1305.
LIAN J, ZHOU X, CAO W. SINA microblog data retrieval[J]. Journal of Tsinghua University(Science and Technology), 2011, 51(10): 1300-1305.

[5] 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, 34(10):1956-1965.
TIAN J T, WANG Y T, FENG X J. A new hybrid algorithm for influence maximization in social networks[J]. Chinese Journal of Computers, 2011, 34(10): 1956-1965.

[6] HAN Y N, LI D Y, WANG T. Identifying different community members in complex networks based on topology potential[J]. Frontiers of Computer Science in China, 2011, 5(1): 87-99.

[7] 赫南, 李德毅, 淦文燕. 复杂网络中重要性节点发掘综述[J]. 计算机科学, 2007, 34(12): 1-5.
HE N, LI D Y, GAN W Y. Mining vital nodes in complex networks[J]. Computer Science, 2007, 34(12): 1-5.

[8] 张翼, 刘玉华, 许凯华. 一种基于互信息的复杂网络节点重要性评估方法[J]. 计算机科学, 2011, 38(6):88-89.
ZHANG Y, LIU Y H, XU K H. Evaluation method for node importance based on mutual information in complex networks[J]. Computer Science, 2011, 38(6): 88-89.

[9] NARDELLI E, PROIETTI G, WIDMAYER P. Finding the most vital node of a shortest path[J]. Theoretical Computer Science, 2001, 296(1): 167-177.

[10] HANSEN D, SHNEIDERMAN B A, SMITH M. Analyzing social media networks with NodeXL: insights from a connected world[EB/OL]. http://deca.cuc.edu.cn/Community/cfs-filesystemfile.as hx/__key/ Community Server.Components.PostAttachments/00.00.01.17.38/Analyzing-Social-Media-Networks-with-NodeXL.pdf, 2012.

作者简介：

于洪（1972-），女，重庆人，博士，重庆邮电大学副教授、硕士生导师，主要研究方向为数据挖掘、粗糙集理论和 Web 智能等。

杨显（1987-），男，重庆人，重庆邮电大学硕士生，主要研究方向为数据挖掘。