doi:10.11959/j.issn.1000-436x.2016020

基于区域交互模型的 SNS 网络用户影响力评估

王楠[1]，孙钦东[1]，周亚东[2]，王汉秦[1]，隋连升[1]
（1. 西安理工大学网络计算与安全技术陕西省重点实验室，陕西 西安 710048；
2. 西安交通大学智能网络与网络安全教育部重点实验室，陕西 西安 710049）

摘 要：针对现有方法与模型未能准确体现不同距离用户之间真实交互行为的问题，提出了一种基于用户区域交互模型的用户影响力评估方法。区域交互模型利用影响力传递的不同方式，刻画不同距离之间用户的交互行为模式，能更为真实准确地反映在线社会网络用户之间的交互行为。通过计算用户对相邻用户的显性影响力与非相邻用户的隐性影响力，可有效识别在线社会网络中大影响力用户、僵尸粉用户等不同类型用户。基于新浪微博与人人网真实数据开展用户影响力评估以及相应的用户角色识别实验，结果显示，与现有方法相比，基于区域交互模型的识别方法可以准确有效地识别出在线社会网络中的大影响力用户、僵尸粉用户等各类型用户。

关键词：用户影响力评估；区域交互模型；在线社会网络；大影响力用户；僵尸粉
中图分类号：TP393 **文献标识码：**A

Study on user influence analysis via regional user interaction model in online social networks

WANG Nan[1], SUN Qin-dong[1], ZHOU Ya-dong[2], WANG Han-qin[1], SUI Lian-sheng[1]
(1. Shaanxi Key Laboratory of Network Computing and Securi y, Xi'an University of Technology, Xi'an 710048, China;
2. MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Conventional user influence researches do not accurate reflect the real interaction pattern between different users in online social networks. In order to solve this problem, a user influence evaluation method based on regional user interaction model has been proposed. The regional user interaction model can illustrate the real online social network user interaction pattern between users with different distance by the influence transfer effect. The method calculates the direct influence and the indirect influence of each user in online social networks and identifies the influential users and zombie users. Experiments are based on the real data of Sina Weibo and RenRen online social networks and the results show that compared with the existing methods the method has better accuracy and efficiency for the infl tial user and zombie user identification.

Key words: user influence evaluation, regional interaction model, online social network, influential user, zombie user

1 引言

近年来，Twitter、新浪微博、Facebook 等新兴在线社会网络（SNS, online social network services）吸引了大量网络用户关注。与传统的 E-mail、新闻站点等网络信息交换平台相比，这些新兴在线社会网络具有用户主动参与度高、信息规模巨大、信息传播速度快等特点。海量用户之间通过关注或者添加好友等行为，建立起有向或无向的连接关系，并通过信息转发或者分享等行为形成了新型的网络生态系统。用户影响力评估是在线社会网络的重要研究内容之一，其研究结果可为网络的信息传播规律、用户行为分析等研究提供理论支撑，并且可用于精准化网络营销、

收稿日期：2015-02-03；修回日期：2015-07-30
通信作者：孙钦东，sqd@xaut.edu.cn
基金项目：国家自然科学基金资助项目（No.61172124, No.61571360, No.61202392）
Foundation Item: The National Natural Science Foundation of China (No.61172124, No.61571360, No.61202392)

网络舆情管控等提供技术支持[1]。目前，用户影响力相关研究方法大多基于网络拓扑结构、用户行为[2-4]等基础特征（诸如粉丝连接数、转发行为）对用户影响力进行评估。已有方法对用户影响力评估有着重要的参考价值，但是仍然存在不足。单一拓扑结构并不能真实反映用户重要性[5]，而基于介数等复杂的拓扑结构方法同样仅考虑到网络中用户之间的连接关系，忽略了用户行为等其他在线社会网络用户特性。基于用户行为的影响力评估方法大多从相邻用户之间的交互行为为出发点，对于一定距离范围内的非直接相邻用户行为交互分析不足。此外，现有影响力分析研究中大多数方法的研究对象只针对网络的大影响力用户，而在线社会网络用户可根据用户影响力被区分为大影响力用户、普通用户、僵尸粉用户等多种类型用户。

在线社会网络中，用户之间的交互行为与真实社会类似，用户之间即使并不直接相连，由于信息在不同用户之间的多次转发也能够形成交互关系，如图 1 所示。用户影响力可由与其不同距离用户之间的交互行为体现，并且对其他用户的影响方式以及影响力大小能够体现出该用户在社会网络中的地位与角色。本文以新浪微博与人人网为研究对象，针对现有研究中所存在的问题，提出了一个在线社会网络用户区域交互模型并对网络用户影响力进行评估。通过用影响力传递的方式描述用户与其他相邻或非相邻用户之间的交互行为，反映用户在在线社会网络中真实的影响力，并以此来对网络中的用户进行类型划分。实验结果表明，区域交互模型可应用于在线社会网络中用户的影响力评估研究，并且能够对网络中不同类型角色的用户进行有效准确地识别。

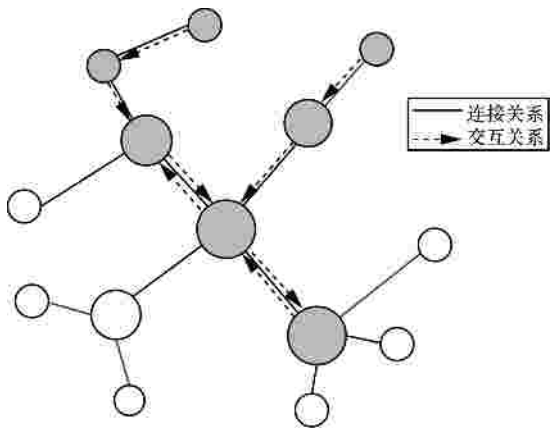

图 1　在线社会网络用户交互

2　相关研究

目前，在线社会网络用户影响力研究大多针对大影响力用户的识别，现有方法可分为基于拓扑结构与基于用户行为特征等。早期研究大多将简单的拓扑结构属性直接作为评估与识别网络中重要节点的依据，如 Leavitt 等[6]直接将用户粉丝规模的大小作为判断用户影响力大小的依据。Kitsak 等[7]根据计算用户的介数（betweenness）、中心度（centralities）等特征值来对用户影响力进行评估，进而筛选网络中大影响力的用户。Brown 等[8]通过 *K*-shell 分解的方法识别 Twitter 中的大影响力用户，该方法通过节点在网络中所处的位置对节点影响力进行评估，认为越靠近中心位置的节点其影响力越大。由于根据介数与中心度评估用户影响力的方法难以应用到大规模复杂的网络中，Chen 等[9]在中心度等用户结构特征的基础上结合时间开销因素，对网络中节点进行影响力评估。

基于用户行为特征的方法是在线社会网络大影响力用户识别常用的一类方法。此类方法大多基于用户转发、评论等行为，再采取相应的评估手段对用户影响力进行评估。Huang 等[10]将用户行为与 PageRank 算法相结合对微博社会网络中的用户影响力进行评估，研究结果发现网络中活跃用户的影响力更大，并且此现象与粉丝规模的大小并无严格的相关关系。Tang 等[11]研究了用户转发行为、交互信息内容以及相应时间等属性与用户影响力之间的关系，并在此基础上提出了一个在线社会网络用户影响力评估架构。

此外，针对大影响力用户识别还有一些其他类型方法。Uysal 等[12]根据用户转发微博的习惯，提出了一种用户微博的排序方法，并以转发微博的可能性作为用户影响力评估的标准。Sun 等[13]根据在线社会网络话题传播过程中用户行为的差别，将用户分为不同角色，并利用相应方法对分类过的用户进行影响力分析。

对于僵尸粉识别，早期研究主要根据一些指标并通过简单的规则进行僵尸粉的识别[14]，这些方法虽然简单易于实现但是准确率偏低，难以应用到实际的僵尸粉识别工作中。目前，比较有效的僵尸粉识别方法大多通过特征选择，选出与用户身份存在密切关联的特征集，然后通过机器学习的方法对僵尸粉进行识别，如 Chu 等[15]研究分析了多个正常用户、僵尸粉

用户等类型用户的特征，并提出了一个基于熵、用户属性以及文本处理的僵尸粉、正常用户分类系统。Bhat 等[16]根据群组特性来对网络中僵尸粉进行研究，通过分析群组交互性、用户连接、用户是否为核心节点等多个属性，对网络用户类型进行划分。

3 数据集

实验过程所使用的数据通过爬虫程序采用广度优先的策略从新浪微博以及人人网获得，并且为保障用户隐私所有数据均进行了匿名化处理。在采集新浪微博数据时，利用新浪提供的 API 获取相关数据，采集人人网数据则使用基于页面内容解析方式的网络爬虫进行爬取。最终得到的微博数据如表 1 所示，采集得到的微博用户数据中分为用户信息以及用户的微博信息，其中，用户信息包括用户 UID、昵称、微博数、粉丝数、关注数以及注册日期等。微博信息则包括了发布时间、转发量以及转发列表信息等。

表 1 新浪微博数据集

数据种类	数量	时间信息
用户节点	427 629	2013-1
微博	15 087 703	2013-1-1 至 2013-1-31

对于人人网，由于其有向图性质以及受限于隐私保护策略，在爬取数据时选取的实验室内部成员为根节点，筛选可以浏览到新鲜事分享的用户对其信息进行存储。最终得到的人人网数据如表 2 所示。其中，用户信息包括用户 UID、好友数、学校信息、用户基本信息等。新鲜事信息包括参与信息分享过程的用户链以及信息 ID、分享数等。

表 2 人人网数据集

数据种类	数量	时间信息
用户节点	17 216	2013-6
新鲜事信息	597 831	2013-6-1 至 2013-6-30

4 区域用户交互模型

磁场、引力场等物理学的场模型理论描述了物理场中节点之间的相互作用关系，以及物体之间的能量传递效应。在线社会网络用户之间的交互行为与场模型中节点间的交互作用相类似，具有相近的特征。作者在前期研究中发现，用户之间的交互行为与影响力相关，用户影响力由于与其相邻和非相邻用户的信息转发行为具有与场模型类似的传递效应[17]。本文在考虑用户交互行为与影响力传递关系的基础上，提出了用户区域交互模型，用户区域交互行为模式与影响力传递机制如图 2 所示。

交互行为与影响力传递过程可描述如下。

有社会网络 $G(E,V)$，其中，V 表示社会网络的节点集合，E 为边集合，表示节点之间有无连接关系，其值的大小表示节点之间的距离。$V=\{V_1,V_2,V_3\}$，其中，V_2 是 V_1 的粉丝节点，V_3 是 V_2 的粉丝节点。如果 V_2 转发了 V_1 的信息，由于信息内容或者用户真实身份等因素，V_1 所发布的信息有一定的概率被 V_2 的粉丝再次转发。转发过程使 V_1 的影响力沿着转发链传递下去，同时节点由于信息被转发其影响力得到了增加，此过程与能量反馈相类似。根据参与转发用户之间的距离，本文将影响力划分为 2 种不同的形式：显性影响力与隐性影响力。显性影响力表示距离为 1（E=1）的情况下，用户转发所传递的影响力，即由于粉丝用户转发所产生的影响力。隐性影响力表示距离大于 1（E>1）的情况下，用户转发所产生的影响力传递效应，即由于非直接相连接用户转发所产生的影响力。

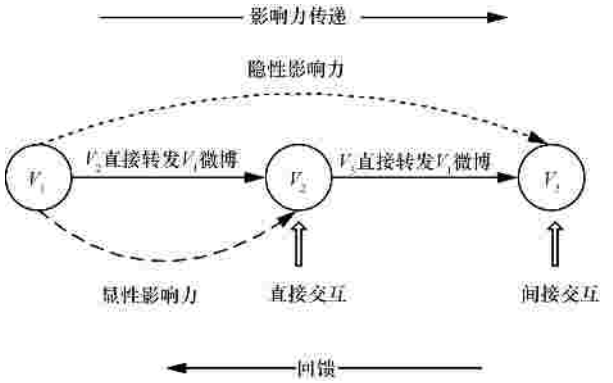

图 2 用户区域交互行为与影响力传递

图 2 中由用户交互产生的影响力传递效应可被推广到一般模型。假设有社会网络图 $G(E,V)$，V 表示节点集合 $V=\{V_1,V_2,\cdots,V_3\}$，E 表示边集合 $E=\{E_1,E_2,\cdots,E_m\}$。则可以得到如下定义。

定义 1 t 时刻节点（即用户）之间的连接关系（距离）矩阵为

$$\boldsymbol{D}^t_{\text{connect}}=\left(d^t_{ij}\right)_{n\times n}=\begin{pmatrix} d^t_{11} & d^t_{12} & \cdots & d^t_{1n} \\ d^t_{21} & d^t_{22} & \cdots & d^t_{2n} \\ \vdots & \vdots & & \vdots \\ d^t_{n1} & d^t_{n2} & \cdots & d^t_{nn} \end{pmatrix} \tag{1}$$

其中，d_{ij}^t 的取值为边集合 E 中的值。

定义 2　t 时刻用户 V_i 转发 V_j 信息的转发关系矩阵为

$$\boldsymbol{C}_r^t = \left(c_{ij}^t\right)_{n\times n} = \begin{pmatrix} c_{11}^t & c_{12}^t & \cdots & c_{1n}^t \\ c_{21}^t & c_{22}^t & \cdots & c_{2n}^t \\ \vdots & \vdots & & \vdots \\ c_{n1}^t & c_{n2}^t & \cdots & c_{nn}^t \end{pmatrix} \tag{2}$$

其中，$c_{ij}^t \geq 0$，表示 t 时刻用户 V_i 转发 V_j 信息的行为关系，值为 0 表示用户 V_i 与 V_j 之间没有转发关系，值大于 0 时表示存在转发关系。在此假设前提为一个用户可以转发同条信息多次。根据转发关系矩阵可得出以下 2 个结论。

结论 1　t 时刻用户 V_k 转发总量为 $\sum_{j=1}^{n} c_{kj}^t$，即为 C_r^t 中第 k 行的总和。

结论 2　t 时刻用户 V_k 信息被转发总量为 $\sum_{i=1}^{n} c_{ik}^t$，即为 C_r^t 中第 k 列的总和。

定义 3　所有节点（即用户）在 t 时刻的发帖数目向量为

$$\boldsymbol{V}_p^t = (v_i^t)_{1\times n} = (v_1^t, v_2^t, \cdots, v_n^t) \tag{3}$$

定义 4　t 时刻用户 V_k 活跃粉丝的规模为 $B(V_k,t)$，其值为参与转发的粉丝数，即 $B(V_k,t) = \sum_{i=1}^{n} b_{ik}^t$，其中，$b_{ik}^t = \begin{cases} 0, c_{ik}^t = 0 \\ 1, c_{ik}^t > 0 \end{cases}$，$c_{ik}^t$ 为定义 2 中转发关系矩阵 $\boldsymbol{C}_r^t$ 中第 k 列的值。

定义 5　t 时刻用户 V_i 若转发了 V_k 的信息，且他们之间的距离为 r，则称 V_k 为 V_i 的 r 距父节点，记为 $V_{k,r}^{fa}$；相对应地，称 V_i 为 V_k 的 r 距孩子节点，记为 $V_{k,r}^{ch}$。若与 V_k 距离为 r 的孩子节点有多个，则 $V_{k,r}^{ch} = \left\{V_{k,r}^1, V_{k,r}^2, \cdots, V_{k,r}^m\right\}$，其中，$V_i \in V_{k,r}^{ch}$。

用户的信息越多地被转发表示该用户的吸引度越大，其影响力传递效应越强，基于上述定义，本文建立的用户区域交互模型中在 t 时刻用户 V_k 的吸引度为 e_k^t，可由以下公式计算

$$e_k^t = B(V_k,t)\frac{1}{n}\sum_{i=1}^{n}\frac{c_{ik}^t}{\sum_{j=1}^{m} c_{ij}^t} \tag{4}$$

其中，c_{ik}^t 为粉丝 V_i 转发 V_k 的信息数量，$\sum_{j=1}^{n} c_{ij}^t$ 为 V_i 总的转发数量。用户吸引度为用户吸引粉丝，并使其信息被转发的能力。用户吸引度与粉丝转发其信息占粉丝转发平均比例成正比，粉丝转发其信息的比例越高，表示该用户对其粉丝的吸引越大。此外用户活跃度与其粉丝规模 $B(V_k,t)$ 成正比，活跃粉丝越多表示该用户的信息具有被更广泛传播的可能性。

由于用户影响力分为显性影响力与隐性影响力，所以在模型中 t 时刻用户 V_k 总的影响力为所有传递效应产生的显性与隐性影响力之和，表达式为

$$I_k^t = I(V_k,t) = I_d(V_k,t) + I_r(V_k,t) \tag{5}$$

其中，$I_d(V_k,t)$、$I_r(V_k,t)$ 分别表示 t 时刻用户 V_k 总的显性与隐性影响力。

由图 2 可以看出，用户的显性影响力为相邻用户间的影响关系，其物理意义可由某一时刻邻接用户转发引起的用户影响度变化率表示，其表达式如下

$$G_{V_k}(t) = \frac{\mathrm{d}e_k^t}{\mathrm{d}t}\sum_{i=1}^{n}(c_{ik}^t d_{ik}^t) \tag{6}$$

其中，$\frac{\mathrm{d}e_k^t}{\mathrm{d}t}$ 表示用户 V_k 的吸引度在时刻 t 的变化速度；d_{ik}^t 表示在 t 时刻用户 V_i 与 V_k 之间的显性关系，其取值为 $d_{ik}^t = \begin{cases} 1, d_{ik}^t = 1 \\ 0, d_{ik}^t > 1 \end{cases}$；$c_{ik}^t$ 为定义 2 中转发关系矩阵 $\boldsymbol{C}_r^t$ 中第 k 列的值。

由于用户影响度变化率 G_{V_k} 表达式为导数形式，需要将其离散处理，采用向前差分格式，最终的表达式为

$$G_{V_k} = \frac{\mathrm{d}e_k^t}{\mathrm{d}t}\sum_{i=1}^{n}\left(c_{ik}^t d_{ik}^t\right) \approx \frac{e_k^{t+\Delta t} - e_k^t}{\Delta t}\sum_{i=1}^{n}\left(c_{ik}^t d_{ik}^t\right) \tag{7}$$

那么，t 时刻用户 V_k 的显性影响力 $I_d(V_k,t)$ 则可由累计的相邻用户影响度变化率表示

$$I_d\left(V_k,t\right) = \sum_{0}^{t} G_{V_k} \tag{8}$$

对于时间尺度的间隔 $?t$，本文实验取 $?t$=1，t=0,1,2,…,T（T 是考虑到的最大时间，时间单位为天）。在初始时刻用户之间没有信息传递，其显性影响力为零，因此在初始时刻（即零时刻）规定 $I_d(V_k,0)$=0。

对于用户的隐性影响力，采取遍历连接图中所有父节点的方式，计算每个父节点与其孩子节点的

传递效应总和衡量该父节点的隐性影响力。所以 t 时刻用户 V_k 的隐性影响力 $I_r(V_k,t)$，本文主要考虑 V_k 与其 r 距孩子节点（$r>1$）的影响关系。假设此时用户 V_k 有 m 个孩子节点，则 $I_r(V_k,t)$表达式为

$$I_r(V_k,t)=\sum_{r=2}^{\infty}\left(\sum I_d\left(V_{k,r}^{fa},t\right)p^r\right)=\sum_{r=2}^{\infty}\left(\left(\sum_{j=1}^{m}I_d\left(V_{k,r}^{j},t\right)\right)p^r\right) \tag{9}$$

其中，p 为转发概率（其值是通过抽样得到的分布概率），r 为用户间信息转发的路径距离。

实际中，由于影响力作用的距离 r 不可能是无穷远。若已知 t 时刻用户之间的连接距离矩阵中的最大值 $d_{\max}^t$，则相应地修正隐性影响力 $I_r(V_k,t)$表达式为

$$I_r(V_k,t)=\sum_{r=2}^{d_{\max}^t}\left(\sum I_d\left(V_{k,r}^{fa},t\right)k^r\right)=\sum_{r=2}^{d_{\max}^t}\left(\left(\sum_{j=1}^{m_{k,r}^t}I_d\left(V_{k,r}^{j},t\right)\right)k^r\right) \tag{10}$$

其中，$m_{k,r}^t$ 表示 t 时刻用户 V_k 的 r 距离孩子节点数。

综上所述，t 时刻用户 V_k 总的影响力为

$$\begin{aligned}I_k^t&=I(V_k,t)=I_d(V_k,t)+I_r(V_k,t)\\&=\sum_0^t\left(\frac{e_k^{t+\Delta t}-e_k^t}{\Delta t}\sum_{i=1}^{n}\left(c_{ik}^t d_{ik}^t\right)\right)+\sum_{r=2}^{d_{\max}^t}\left(\left(\sum_{j=1}^{m_{k,r}^t}I_d\left(V_{k,r}^{j},t\right)\right)k^r\right)\\&=\sum_0^t\left(\left(e_k^{t+1}-e_k^t\right)\sum_{i=1}^{n}\left(c_{ik}^t d_{ik}^t\right)\right)+\sum_{r=2}^{d_{\max}^t}\left(\left(\sum_{j=1}^{m_{k,r}^t}I_d\left(V_{k,r}^{j},t\right)\right)k^r\right)\end{aligned} \tag{11}$$

用户任意 t 时刻影响力可基于上述过程计算得到。由于实验过程中所需要处理的用户数据都是十万级别以上的，此时得到的转发关系矩阵、连接关系（距离）矩阵等是稀疏且相当庞大的，而大数据的存取也制约了模型的求解。为了解决上面的问题，本文采用图论中树形结构的方式来表达用户间的连接关系，使数据的存取和模型的求解得到极大的简化。

5 用户交互行为实证分析

在线社会网络中，相邻用户交互行为可以通过直观的数据进行分析，而不相邻用户之间的交互行为则难以直接被观测到。本节通过分析表明非直接相邻用户之间是否存在交互行为且具有一定规模能够为区域交互行为模型提供支撑。

5.1 用户关系的确定

研究不同距离用户之间的交互行为，需要确定转发链中各个用户之间的连接关系。由于各 SNS 站点都设置了隐私保护机制，因此信息传播链中用户之间是否存在关注关系需要进行判断分析。新浪微博的共同关注功能显示了 2 个用户之间是否关注了同一个用户，本文通过共同关注判断 2 个用户之间是否存在关注关系。由于请求限制以及转发链中用户数量规模，难以准确判断全部用户的连接关系。本文根据抽样推断的方法，从获取到的转发链中随机选择了一部分用户，并判断他们的关系，基于此结果来估计转发链中各个用户之间的关注关系。具体方法如下。

1) 随机选取 N 条转发链。

2) 统计距离为 d 且存在关注关系的用户数目，并计算其占整个转发链的比例 P_d。

3) 为了减少抽样统计的分布与总体分布的误差，采取多次抽样取平均的方法，即重复步骤 1）和步骤 2），完成 m 次抽样统计得到一系列的距离为 d 且存在关注关系的用户比例 $P_d^i\left(i=1,2,\cdots,m\right)$。

4) 最终的总体分布表示为

$$\overline{P}_d=\frac{1}{m}\sum_{i=1}^{m}P_d^i \tag{12}$$

例如，分析长度为 3 的转发链中用户之间的距离关系，从数据集中选取相应长度的转发链，并判断不同位置的用户之间是否存在关注关系。

人人网提供了与微博类似的共同好友查看功能。在判断转发链中用户之间的链接关系时，共同好友可作为判断依据之一。由于人人网部分用户设置了非好友的访问权限，因此针对有向图中用户关系采取以下机制进行判别。

1) 根据有向图节点之间连接关系以及相应的用户转发行为方式，在转发链中相邻的 2 个用户为互为好友的用户。

2) 对于转发链中非直接相邻的用户，若能访问用户详细信息，则进一步判断 2 个用户是否拥有共同好友。

3) 若不能访问用户详细信息，用户通常会填写

学校信息以及籍贯等，通过个人信息相似性对用户之间是否存在好友关系进行判断。

5.2　测量结果

根据转发链中用户距离的分析过程，将基于转发顺序的用户序列，转化成基于距离排列的用户序列，并且对不同距离用户转发进行统计分析。数据集中信息被不同距离用户转发比例如图 3 和图 4 所示。从图中可知，当用户之间距离大于 1 时，用户之间的交互行为是存在并且活跃的。对于无向图网络，虽然用户之间若非直接好友关系并不能直接访问，但是非相邻用户之间同样存在一定规模的交互行为。

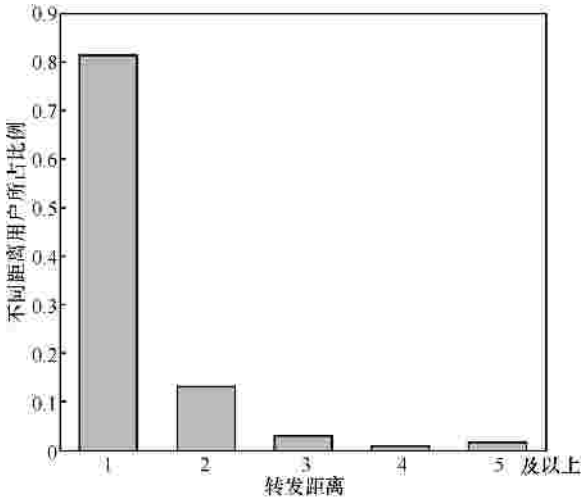

图 3　微博转发链中不同距离用户比例

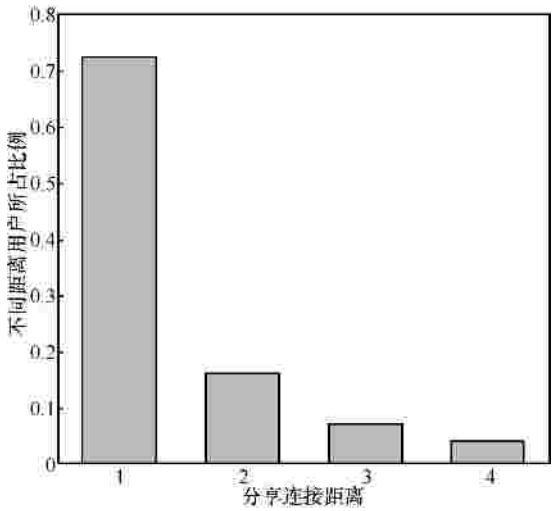

图 4　人人网分享链中不同距离用户比例

由于人人网数据规模及其隐私限制，为了减小判断误差带来的影响，在分析时定义用户的距离最长为 4。根据上述结果，在线社会网络用户之间的交互关系存在于不相邻的用户之间并具有一定数量规模，能够对基于不同邻接距离用户交互行为的区域交互行为模型提供理论依据。

6　实验结果与分析

6.1　影响力评估与用户划分结果分析

为了验证区域交互模型在用户影响力评估以及基于影响力分析的用户角色划分研究的有效性，本文基于用户显性、隐性影响力对网络中的大影响力用户、普通用户以及僵尸粉用户进行识别研究。

图 5 为基于微博数据得到的部分大影响力用户、普通用户以及僵尸粉用户的显性、隐性影响力分布。大影响力用户的隐性影响力与显性影响力都具有较大的数值规模，此结果表明大影响力用户的微博信息不仅能够被大量的粉丝转发，还能够由传递效应传播到距离较远的用户。普通用户的信息传播能力较弱，因此其 2 类影响力分布取值区域较小。僵尸粉用户的影响力分布显示出极为不平均的结果，这是由于僵尸粉用户的信息很难被正常用户转发，其影响力分布也与普通用户有明显差别。

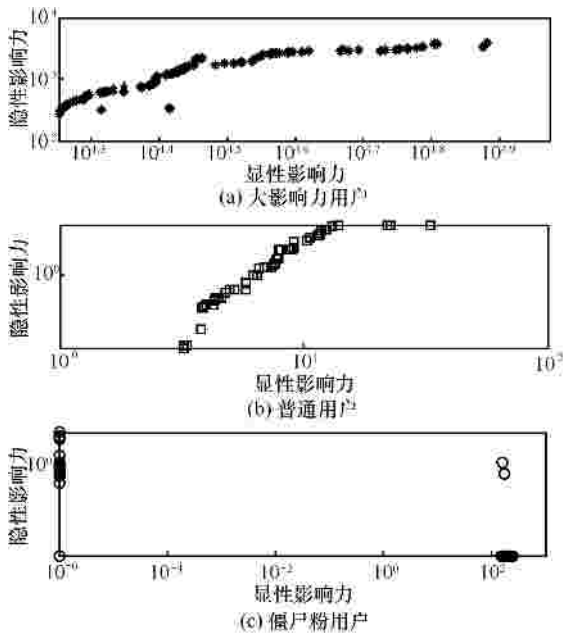

图 5　新浪微博大影响力用户、普通用户以及僵尸粉用户显性、隐性影响力分布

由于人人网的有向图性质，用户之间的好友关系建立需要用户审核确认，因此本文研究内容不包括人人网中僵尸粉的识别。图 6 为人人网中大影响力用户与普通用户的显性、隐性影响力分布。由于本文实验所使用的人人网数据中不包括明星账号、

机构账号或者媒体账号等用户，并且人人网用户整体的活跃性与新浪微博相比相对较低，因此用户影响力的计算值相对较小。

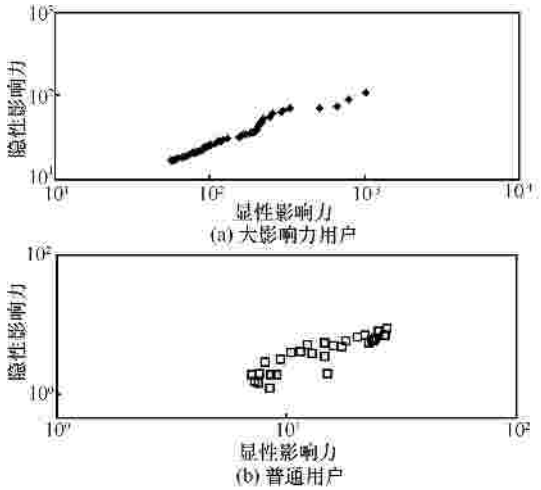

图 6 人人网大影响力用户、普通用户显性、隐性影响力分布

从图 6 所示的各类型用户的显性与隐性影响力分布可以看出，大影响力用户的信息能够被大量的粉丝转发，并且还可被大量非连接用户转发，因此其区域交互性十分明显。普通用户与相邻、非相邻用户之间也存在交互行为但规模相对较小。僵尸粉用户的信息难以被大规模转发。即使目前存在通过僵尸粉团等模拟正常用户的僵尸粉，其不同类型的影响力分布与正常用户相比仍有明显区别。

6.2 有效性分析

6.2.1 大影响力用户

为了分析区域交互模型在识别大影响力用户时的有效性，本文与基于粉丝数、PageRank[18]以及信息级联模型[19]的大影响力用户识别方法进行对比分析。图 7 和图 8 为新浪微博与人人网中影响力排序前 50 用户的粉丝粘性对比结果。

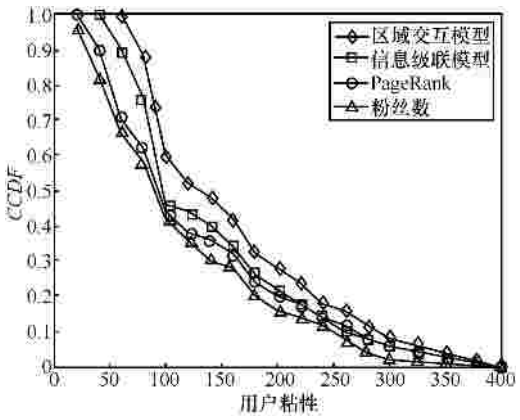

图 7 微博用户粉丝粘性对比

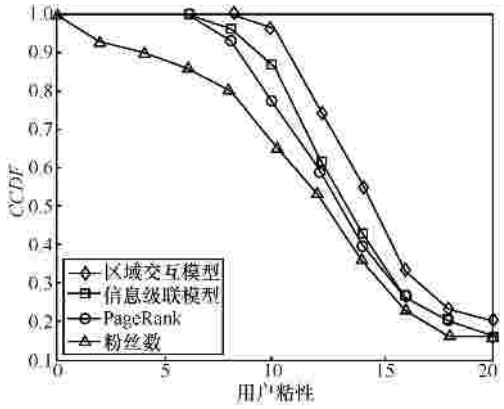

图 8 人人网用户粉丝粘性对比

本文用粉丝用户转发次数 2 次上的用户数来表示用户粘性，用以反映用户与其粉丝之间的交互频度与依赖关系。根据图 7 和图 8 结果，无论有向图网络(新浪微博)或无向图网络(人人网)，基于区域交互模型识别得到的大影响力用户要优于其他方法。虽然通过不同方法识别得到的大影响力用户具有重合部分，但是基于区域交互模型识别得到大影响力用户在整体上具有较大粉丝粘性。

信息转发是在线社会网络中最具特色的功能，信息的转发规模能够体现用户影响力。若排序序列中越靠前的用户其信息传播具有越大的覆盖人数，相应的影响力评估方法具有更好的效果。因此，本文对影响力排名靠前的用户信息转发规模进行统计分析，进一步验证区域交互模型的有效性，结果如图 9 和图 10 所示。根据图中结果可知，采用区域交互模型的方法识别得到用户在信息覆盖人数上要高于其他方法。从上述分析可以看出，基于区域交互模型的用户影响力评估方法识别得到的大影响力用户具有较高的活跃性，并且能够吸引大量其他用户关注与转发其信息，该模型能够有效体现出在线社会网络中用户的真实影响力。

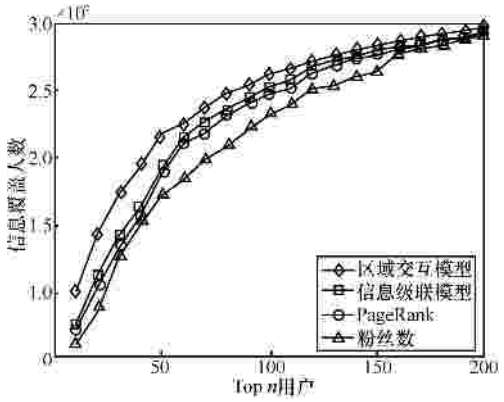

图 9 微博用户信息传播规模对比

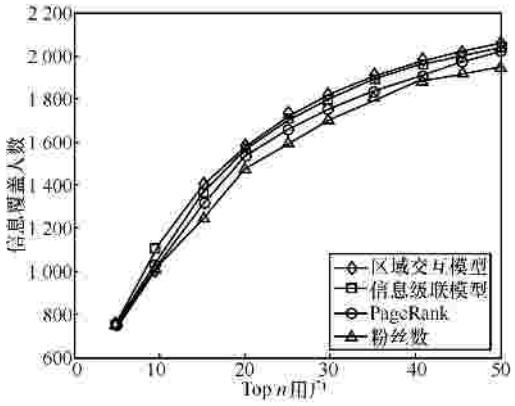

图 10　人人网用户信息传播规模对比

6.2.2　僵尸粉用户

僵尸粉用户是在线社会网络中对正常用户造成较差用户体验度的一类虚假用户，简称为僵尸粉，部分僵尸粉由机器人程序自动注册生成，以广告为目的发布大量垃圾信息。还有一些用户在注册后并没有任何使用站点服务的行为，也属于僵尸粉。为了验证在识别僵尸粉时模型的有效性，本文采用人工方式从微博中的僵尸粉进行标注，具体判断原则为：1）判断用户发布微博内容中是否明显为广告信息，如果比例超过 90%则判定其为僵尸粉用户；2）判断用户微博内容的差异性，若用户微博内容中含有大量短链接或者图片等，判断文字内容与短链接内容是否相关；3）若微博内容为纯文字信息，判断信息内容是否具有意义，是否含有生僻、乱码等字符。最终通过人工标注方式得到 3 000 个僵尸粉样本。

目前识别僵尸粉的方法大多是基于用户的特征指标，然后利用机器学习的方法来分类识别。对于基于用户特征的识别方法，单一指标虽然可以被用来识别网络中的僵尸粉用户，但是准确率偏低，实际应用效果较差。表 3 所示内容为根据单个特征进行僵尸粉识别时的准确率（由于数据集的差异，本文中的结果与文献[15]稍有差异）。因此，为了保证识别效果，此类方法必须要有足够多的特征指标。

表 3　　单一指标僵尸粉识别准确率

特征	准确率/%
账号名誉度	70
身份认证	20.7
微博数	61
URL	70.9
评论数	52
注册日期	62
发布设备	58.3

在使用较多用户属性进行僵尸粉识别时，基于机器学习方法，如 SVM 识别方法能够达到 90%的正确率，但是如果相关算法使用的特征较少时准确率则相对较低。选取 2 个属性作为特征向量并使用 SVM 做 2 类分类，进行僵尸粉识别，结果如表 4 所示。从表 5 中的结果可以看出少量特征并不能保证识别结果能够有很高的准确性，因为特征对于识别方法的权重也是有区别的。

表 4　基于区域交互模型的僵尸粉识别结果

类别	指标	平均结果/%
僵尸粉	准确率	92.3
	召回率	93.1
正常用户	准确率	91.6
	召回率	92.1

表 5　　基于少量特征的僵尸粉识别准确率

特征	准确率/%
微博数和评论数	75.2%
身份认证和 URL	79.1%
微博数和身份认证	83%

基于区域交互模型采取的僵尸粉识别方法为：1）获取用户显性、隐性影响力值；2）显性、隐性影响力值阈值设定；3）根据用户相应影响力值对其身份进行标定。为了设定合理的影响力阈值取值，本文根据人工筛选得到的数据集以 300 个僵尸粉与 300 个普通用户一组，将用户分成 10 组作为训练与测试数据集，并采用循环估计的方法选取平均准确率最高时相应显性、隐性影响力数值作为僵尸粉识别过程的阈值。最终僵尸粉判定条件为选取显性影响力大于 100 且隐性影响力小于 5，隐性影响力大于 150 且显性影响力小于 10 以及选取显性、隐性影响力同时小于 1 为僵尸粉用户，其他则认为是正常用户。僵尸粉识别实验结果如表 4 所示。

根据对识别错误的用户进行分析发现，误判的主要原因在于某些正常用户其活跃性非常低，在实验周期中发微博的行为十分稀疏，虽然在用户影响力上与僵尸粉用户极为相似，但通过人工筛检并不能被归为僵尸粉用户。此外把僵尸粉误认为普通用户的原因为其信息在本文实验数据中截止时间的原因并不完整，因此计算出结果未能满足僵尸粉筛选条件。在分析得到的僵尸粉后，

发现存在僵尸粉团的现象存在，部分账号其信息内容与行为跟普通用户相比并无较大差异，该账号微博由其他僵尸粉进行转发但账号之间并不存在关注关系，这些模拟正常用户行为的僵尸粉也被基于区域交互模型的识别方法检测获得。综合上述几部分实验，结果表明区域交互模型能够较为真实地反映用户之间的交互行为，基于行为不同模式的差异可对在线社会网络中不同类型用户进行识别。

7 结束语

本文基于新浪微博与人人网数据发现用户之间的转发、分享等交互行为在一定邻接距离范围内是广泛存在的，并不仅限于相邻用户。根据不同距离的交互行为提出了一个区域交互模型。该模型基于不同邻接距离用户之间的交互行为，对在线社会网络中用户的影响力进行判断分析。用户对相邻节点的显性影响力以及非相邻节点的隐性影响力可应用于在线社会网络用户类型划分，能够从用户行为、用户影响范畴等方面更真实地体现出用户在网络中所处的地位。实验结果表明，不论是对于大影响力用户识别，还是僵尸粉识别，本文的方法在准确度等方面具有一定的有效性。

区域交互模型是针对在线社会网络用户之间交互行为的抽象，本文开展包括的用户影响力研究以及相应的用户角色划分研究仅是基于该模型展开的部分基础研究。在下一步研究工作中，将开展在本文工作基础上的算法复杂度优化研究，并对算法有效性进行更为细致的分析研究。

参考文献：

[1] KANNA A F, YACINE A, AJITH A. Models of influence in online social networks[J]. International Journal of Intelligent Systems, 2013, 29(2): 161-183.

[2] LIM S H, KIM S W, PARK SUN J. Determining content power users in a blog network: an approach and its applications[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans Archive, 2011, 41(5): 853-862.

[3] LI X，CHENG S Y, CHEN W L. Novel user influence measurement based on user interaction in microblog[C]//The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and MiningNiagara Falls. Canada, c2013: 615-619.

[4] WU X M，WANG J M. Micro-blog in China: identify influential users and automatically classify posts on Sina micro-blog[J]. Journal of Ambient Intelligence and Humanized Computing, 2014, 5(1): 51-63.

[5] CHA M Y, HADDADI H, BENEVENUTO F. Measuring user influence in Twitter: the million follower fallacy[C]//The Fourth International AAAI Conference on Weblogs and Social Media. Washington, DC, USA, c2010: 10-18.

[6] LEAVITT A, BURCHARD E, FISHER D, et al. The Influentials: New Approaches for Analyzing Influence on Twitter[R]. Web Ecology Project, 2009.

[7] KITSAK M, GALLOS L K, HAVLIN S. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11): 888-893.

[8] BROWN P, FENG J L. Measuring user influence on Twitter using modified *K*-shell decomposition[C]//The 2011 ICWSM Workshop on the Social Mobile Web. Barcelona, Spain, c2011: 18-23.

[9] CHEN D B, LV L Y, SHANG M S. Identifying influential nodes in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1777-1787.

[10] HUANG Y L，LI L. Analysis of user influence in social network based on behavior and relationship[C]//The 2nd International Conference on Measurement, Information and Control. Harbin, China, c2013: 682-686.

[11] TANG X N, YANG C C. Ranking user influence in healthcare social media[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(4): 565-582.

[12] UYSAL I, CRFOFT W B. User oriented tweet ranking: a filtering approach to microblogs[C]//The 20th ACM International Conference on Information and knowledge Management. Glasgow, Scotland, c2011: 2261-2264.

[13] SUN B M, VINCENT T Y. Identifying influential users by their postings in social networks[C]//The23rd ACM Conference on Hypertext and Social Media Workshop on Modeling Social Media. Milwaukee, USA, c2012: 1-8.

[14] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks[C]//The 26th Annual Computer Security Applications Conference. New York, NY, USA: ACM, c2010: 1-9.

[15] CHU Z, GIANVECCHIO S, WANG H N. Detecting automation of Twitter accounts: are you a human, bot, or cyborg[J]. IEEE Transactions on Dependable and Secure Computing, 2012, 9(6): 811-824.

[16] BHAT S Y, ISLAMIA J M, DELHI N. Community-based features for identifying spammers in online social networks[C]//The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara Falls, Canada, c2013: 100-107.

[17] SUN Q D, WANG N, ZHOU Y D, et al. Modeling for user Interaction by influence transfer effect in online social networks[C]//The 39th IEEE Conference on Local Computer Networks. Edmonton, Canada, c2014: 486-489.

[18] LIANG H, LU G, XU N S. Analyzing user influence of microblog[C]//2012 IEEE fifth International Conference on Advanced

Computational Intelligence (ICACI). Nanjing, China, c2012: 15-22.

[19] BAKSHY E, HOFMAN J M, MASON W A, et al. Everyone's an influencer: quantifying influence on Twitter[C]//The 4th ACM International Conference on Web Search & Data Mining. HongKong, China, c2011: 65-74.

作者简介：

王楠（1983-），男，河南安阳人，西安理工大学博士生，主要研究方向为在线社会网络、数据挖掘等。

孙钦东（1975-），男，山东莒南人，博士，西安理工大学教授，主要研究方向为网络安全、在线社会网络、物联网等。

周亚东（1982-），男，陕西汉中人，博士，西安交通大学讲师，主要研究方向为在线社会网络、Web 挖掘等。

王汉秦（1987-），男，陕西西安人，西安理工大学硕士生，主要研究方向为在线社会网络。

隋连升（1972-），男，陕西韩城人，博士，西安理工大学副教授，主要研究方向为计算机图形学、数字图像处理以及计算机视觉等。