

基于行为预测的微博网络信息传播建模*

吴凯, 季新生, 刘彩霞

(国家数字交换系统工程技术研究中心, 郑州 450002)

摘要: 研究微博网络中的信息传播及扩散机制在市场营销、舆情管控等方面具有重要意义。当前的传播模型大多忽视了用户间的个体差异。为解决这一问题, 提取了影响转发行为的四类特征, 利用机器学习中的逻辑回归模型分析预测个体转发行为, 并在此基础上融入用户个体差异, 建立了一种基于行为预测的信息传播模型。实验表明, 该模型能较好地模拟真实网络中的信息传播过程。

关键词: 微博; 转发; 行为预测; 信息传播建模

中图分类号: TP393 **文献标志码:** A **文章编号:** 1001-3695(2013)06-1809-04

doi:10.3969/j.issn.1001-3695.2013.06.054

Modeling information diffusion based on behavior predicting in microblog

WU Kai, JI Xin-sheng, LIU Cai-xia

(National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China)

Abstract: Studying the information diffusion model in microblog is necessary in many different fields such as viral marketing. The most current diffusion models ignored the diversity of different users. This paper used logistic regression model to predict user's retweet behavior and then built a information diffusion model. Simulation result illustrates that this model can forecast the trend of information diffusion well.

Key words: microblog; retweet; behavior prediction; information diffusion model

0 引言

微博作为一种迷你型博客, 近年来得到了爆炸式的发展。在微博中, 用户可以通过手机短信、即时通信工具、邮件或 Web 页面等方式阅读或转发好友发布的消息。简便的应用方式使微博网络中的信息传播具备一对多几何级的增长特点。与传统社会媒体相比, 其信息传播的速度、广度和效率都得到了极大的提高, 微博已经成为消息扩散和舆论传播的重要平台。分析微博网络的信息传播机制, 建立微博网络的信息传播模型, 将有助于人们深入了解信息传播的内在机理, 并在市场营销、舆情管控等应用方面发挥重要作用。

近年来, 社会网络中的信息传播研究逐渐成为国内外学者关注的热点。当前的研究集中在两个层面: 在宏观层面, 研究者从网络的角度将信息的传播过程类比为技术革新^[1]或疾病传染, 利用复杂网络传播动力学模型对社会网络上的信息传播规律进行分析。文献[2]基于 SIS(susceptible-infected-susceptible)疾病传染模型, 构建了社会网络中信息级联传播模型; 文献[3]在 SIRS(susceptible infected removed susceptible)模型的基础上, 提出了一种计算节点之间信息传播概率的算法; 文献[4]介绍了一种线性阈值模型(linear threshold model, LTM), 其基本思想是一个节点转换为活跃状态的概率会随着周围活跃节点个数的增多而增大, 如果一个节点邻居中活跃节点数量超过一定阈值, 该节点将转换为活跃状态; 文献[5, 6]利用其他动力学模型对社会网络中的信息传播进行了分析。在微

观层面, 研究者对微博用户的转发行为进行了分析。文献[7]从用户转发数据上进行统计分析, 得到转发行为的一些特征, 并建立模型预测一条给定微博的转发总数; 文献[8]提到了微博特征对转发行为的影响, 包括是否含有 url 等; 文献[9]分析了影响微博转发的特征因素, 利用机器学习中的分类方法建立了特征加权的预测模型。

微博作为一种新型的社会媒体, 单一从宏观或微观层面的研究并不能准确地描述微博网络中的信息传播。本文在宏观与微观层面之间找到结合点, 在深入分析用户转发行为的基础上, 融入用户的个体差异, 建立基于用户行为预测的微博网络信息传播模型, 并在真实数据中进行仿真验证。

1 微博消息转发行为预测

在微博网络中, 一条微博是通过用户之间的层层转发才得以在网络中形成瀑布型的扩散效果, 转发行为是信息传播的主要途径, 也是用户个体差异的重要表现。因此, 分析影响用户个体转发行为的因素, 提出用户转发行为的概率预测方法是建立基于行为预测传播模型的基础。

1.1 问题描述

考虑微博网络 $G = \langle U, E \rangle$, 其中节点 $u \in U$ 表示网络中的所有用户, 边 $(u, v) \in E$ 表示用户 u 与 v 之间的关注关系, 定义用户转发行为构成的网络是用户关系网络的一个子网, 即用户只转发来自其关注节点的消息。

定义 1 消息转发预测问题。在给定微博 r 、用户网络 $G =$

收稿日期: 2012-11-07; 修回日期: 2012-12-17 基金项目: 国家“863”计划资助项目(2011AA7116011, 2011AA010604)

作者简介: 吴凯(1988-), 男, 河北武安人, 硕士研究生, 主要研究方向为社会网络分析(wukai10@gmail.com); 季新生(1968-), 男, 教授, 主要研究方向为移动通信技术; 刘彩霞(1974-), 女, 副教授, 主要研究方向为无线通信安全、社会网络分析。

$\langle U, E \rangle$ 、历史转发消息集合 M 的情况下,对用户 u 转发微博 r 的概率进行预测。

对消息转发的预测问题是机器学习中的典型的二分类问题,在给定用户网络的情况下,通过对历史消息集合 M 进行训练,可以得到用户 u 对微博 r 的转发行为分类结果。在消息转发预测问题中,影响转发的属性特征与转发行为呈现出线性关系^[9]。逻辑回归(logistic regression, LR)模型可以在误差较小的情况下很好地完成线性分类的任务,并且得到概率型的预测结果,将 LR 模型运用到微博转发预测中,得到预测公式如下:

$$p(y_u = 1|x) = \frac{1}{1 + \exp(-\omega(1 + F_u(r, G)))} \quad (1)$$

其中: $F(r, G)$ 为影响用户 u 转发行为的特征集合; y_u 表示用户 u 的转发行为,取值为0或1, $y_u = 1$ 表明用户 u 对消息 r 进行转发; ω 为权值向量。权值的获取利用极大似然函数的方法,对用户 u 利用 N 条历史微博数据形成训练集 $\{r, y_u^i\}_i^N$,用户 u 形成 N 个观测值 $\{y_u^1, y_u^2, \dots, y_u^N\}$, i 为转发的条数。因此, N 个观测值的似然函数为

$$l(\omega) = \prod_{i=1}^N (1 - p(y_u = 1|x))^{(1-y_u^i)} (p(y_u = 1|x))^{y_u^i} \quad (2)$$

当 $\frac{dl(\omega)}{d\omega} = 0$ 时求得的 ω 值即为要求的权值。

1.2 特征提取

特征的选取将在很大程度上影响转发预测的准确性。通过对微博应用特点的分析,影响转发行为的因素主要来自于微博内容、用户属性和网络关系。本文从以下四个方面对用户特征进行了提取。

1) 微博内容与用户兴趣

微博原文内容与用户兴趣的相似程度可以看做源自消息内部对消息传播起到驱动作用的动力。在微博中,热门主题体现了网络中的全局兴趣倾向,往往被大多数用户提及,而比较小众的内容只在特定兴趣的用户中讨论。也就是说,全局及个体兴趣与微博内容的相似程度将对微博消息的传播起到积极的贡献作用。

为了对兴趣进行提取,并计算兴趣之间的相似关系,利用了 Jaccard 相似度,即 $\frac{|A \cap B|}{|A \cup B|}$,其中 A 表示微博内容的词语集合, B 表示从用户历史数据中提取的高频词语集合。Jaccard 相似度越大,说明两者之间兴趣相关性越高。

2) 微博用户社会关系

在微博网络中,信息传递是通过用户节点之间的关注关系发生的,关注者—被关注者的关系可以看做是一种拟社会关系,类似于传统媒体,这种关系中关注者与被关注者之间的相互了解程度是不对等的,关注者更信赖自己所关注的人,并且更倾向于转发被关注者所发布的消息。因此,信息发布者的影响力、发布者与接收者之间的亲密程度将对用户的转发行为产生影响。

3) 微博文本与用户属性

微博文本与用户属性特征在统计意义上对用户的转发行为产生影响,相关研究指出,微博中是否含有 hashtag、url、@ 符号等,都对用户的转发行为产生影响。包含这些符号表明微博内容具有更多的延伸信息,这些延伸信息将影响用户对微博的接受程度。

4) 受激活次数

在社会网络中,当邻居节点频繁对目标节点传递信息时,目标节点对该信息的接受程度将相应地发生变化^[4]。因此,人们认为在微博网络中,当用户的朋友中转发微博的人数增多时,用户转发微博的倾向性将发生变化。

综上所述,共提取了影响微博转发的 17 个数值化特征,如表 1 所示。将特征分为四类:A 类表征了兴趣相似程度;B 类表征了社会关系影响;C 类表征文本特征与用户属性影响;D 类为用户受激活次数的影响。

表 1 影响消息转发特征列表

特征类别	特征序号	特征名称
A	1	微博 r 与微博热点话题的 Jaccard 值
	2	微博 r 与用户 u 的 Jaccard 值
	3	微博 r 与用户关注节点 v 的 Jaccard 值
	4	发布者的粉丝数
B	5	发布者被转发微博数
	6	发布者是否为实名认证
	7	用户之间共有的关注者数目
	8	用户之间共同转发数目
	9	用户之间的历史转发数
	10	用户之间的历史评论数
C	11	是否包含 url
	12	是否包含 hashtag
	13	是否包含 @ 符号
	14	用户的关注数
	15	用户的粉丝数
	16	用户所发微博数
D	17	用户关注节点的已有转发次数

特征 6、11、12 和 13 采用 0 和 1 二元表示,0 表示否,1 表示是,其余特征可以从历史数据集中直接提取。这些特征与转发行为呈现出线性关系,可以利用式(1)和(2)训练得到特征权重,从而建立用户转发行为的概率预测公式为

$$Pr_u = p(y_u = 1|x) = \frac{1}{1 + \exp(-\omega(1 + F_u(r, G)))} \quad (3)$$

2 RPT 网络传播模型

通过分析影响用户转发行为的四类特征,利用式(3)可以计算出微博网络中个体用户的转发概率,这反映出当一条微博呈现在用户面前时个体的行为规律。然而,微博网络作为一个快速而庞大的信息传播网络,众多个体的转发行为形成了信息的瀑布级联效应,分析多用户的信息行为,研究信息在微博网络中传播的速度、广度和深度将更为重要。为解决这一问题,本文将用户个体的转发行为分析与信息传播过程结合建立基于转发行为预测的阈值传播模型(retweet prediction based threshold model, RPTM)。

2.1 相关定义

定义微博网络中的信息传播底图为有向图 $G = \langle U, E \rangle$,其中 U 和 E 分别为图的节点集合和边集合,每个节点代表一个微博用户,每条有向边代表信息由边出发的节点传递到边指向的节点。定义网络中节点的两种状态为未激活状态 S 和激活状态 I 。定义节点 u 的邻居节点集合为

$$N(u) = \{v | (u, v) \in E\}$$

为了便于模型的描述,引入 $A_T(u)$ 为时间步 t 内节点 u 的邻居节点中处于状态 I 的节点集合:

$$A_t(u) = \{v | v \in N(u), u \text{ 的状态为 } I\}$$

2.2 RPT 模型的建立

转发是微博网络中信息传播的重要方式。在 RPT 模型中,对于一条微博,定义已经转发的用户处于已激活状态,未转

发的用户处于未激活状态。节点在网络中的状态如图1所示,灰色节点为待激活节点,黑色节点表示已激活节点。节点在网络中受到多个邻居节点的激活行为,不同邻居节点的激活行为之间是相互独立的。

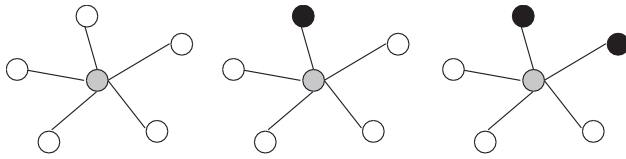


图1 节点状态示意图

为了确定节点被激活的规则,为每个用户 u 设置阈值函数 $f(u)$ 与门限值 θ_u ,其中 $f(u) \in [0,1]$, $\theta_u \in [0,1]$ 。当 $f(u) \geq \theta_u$ 时,用户 u 被激活。由第1章的分析可知, Pr_u 体现了用户对一条微博的转发倾向。因此,在考虑用户间一对一转发消息的情景时,阈值函数可表示为

$$f(u) = Pr_u = \frac{1}{1 + \exp(-\omega(1 + F_u(r, G)))} \quad (4)$$

在网络环境中,邻居中有多个节点试图对用户 u 进行激活。当邻居节点中每新增一个激活状态节点时,则要对阈值函数 $f(u) = Pr_u$ 中的B类特征、D类特征进行更新。由此可以刻画出一条微博在目标网络中的传播过程。

模型的具体过程可以描述如下:

a)数据预处理。获取目标网络历史数据,利用逻辑回归模型为每个节点建立转发行为的预测公式 Pr_u 。所有节点置为未激活状态 S ,并放入集合 N ,为 N 中的每个节点 i 设置阈值 $\theta_i = 0.5$,并建立集合 $A_0(1), A_0(2), \dots, A_0(n)$ 。

b)初始时刻,微博进入目标网络,网络入口节点置为激活状态 I ,并将其放入相应的 $A_0(i)$ 中。

c)在时间步 t ,根据集合 $A_{t-1}(1), A_{t-1}(2), \dots, A_{t-1}(n)$ 中的更新情况,更新特征值,计算相应节点的阈值函数 $f(u) = Pr_u$,当 $f(u) \geq \theta_u$ 时,节点 u 被激活,将 u 从 N 中移除。在此步骤被激活的节点并不参与激活行为。

d)将在上一步中被激活的节点并入相应集合 $A_{t-1}(i)$,从而形成 $A_t(i)$ 。

e) $t = t + 1$,重复步骤c)d),直到网络中没有新的节点被激活。

RPT模型的实现算法如下:

输入: Weibo network $G(U, E)$, retweet R , user set U 。

输出: retweet user set P 。

```

1 all users are in state S
2 generate user set  $A_0(1), A_0(2), \dots, A_0(n)$ 
2 generate a random number  $\theta_u$  for each user
3 for each user  $u$  in  $U$  at time  $t$ 
4   if  $A_t(u)$  have new element
5     calculate  $f(u)$ 
6     if  $f(u) \geq \theta_u$ 
7       take  $u$ ' state to  $I$ , push  $u$  to set  $P$ 
8     else
9        $t = t + 1$ , back to step 3
10    end if
11  else
12     $t = t + 1$ , back to step 3
13  end if
14 end

```

3 实验与仿真

为了验证本文提出的微博转发预测模型以及网络传播模型,需要获取两类数据:用户发布的微博数据和用户之间的关

注关系数据。微博数据用来提取用户的兴趣、微博项特征、用户属性等特征以及分析这些特征对转发行为的影响;关系数据可以用来考察用户之间的亲和力对消息转发的影响以及为RPT网络传播模型建立仿真环境。本文利用新浪微博提供的API接口爬取微博数据来进行传播模型的仿真与验证,数据集中包含了23万条微博以及1万条用户信息。微博网络用户度分布情况如图2所示。从图中可以看出,微博网络的度分布呈现幂率特性,这表明微博网络是一个与尺度无关的网络。

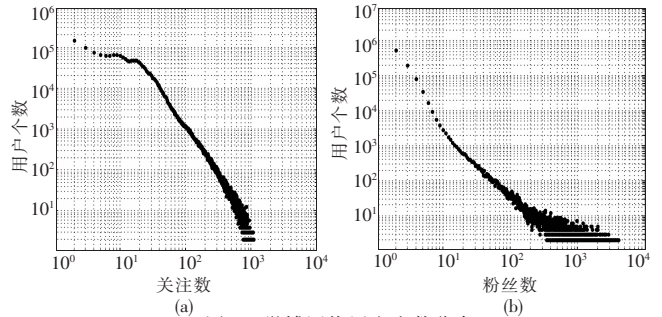


图2 微博网络用户度数分布

本文在数据集中选择一名用户作为根节点,提取了具有完整关注关系的2000名活跃用户形成了传播子网以及这些用户发布的微博52548条。传播子网数据集应满足如下条件:

- a)包含每个用户的所有活跃粉丝。
- b)包含每个用户一段时间内所发布的所有微博。

将数据集分为两部分,每部分数据包含26274条微博数据和2000条用户信息,分别用做训练集和测试集。训练集用来建立每个用户的转发行为预测模型,测试集用来对用户的转发行为进行预测,预测结果如表2所示。

表2 转发行为预测结果 /%

实际	预测	
	转发	非转发
转发	71.3	28.7
非转发	13.1	86.9

利用LR模型进行训练其结果显示,一些特征在转发行为上起到的作用更为明显,如用户与微博兴趣相似度、用户粉丝数、与信息发布者的关系亲密度等。而一些特征的作用相对较弱,如是否含有url、url数量等。

为验证RPT网络传播模型,在数据集中选取了两条微博用于仿真。微博1与微博2在传播子网中分别被转发了69次和45次,将这两条微博输入RPT传播模型,利用算法1得到了预测转发次数随时间的演化趋势,考虑到信息接收阈值函数 $f(u)$ 表示了用户接收信息的概率值,因此将阈值 θ_u 设置为0.5。与真实数据的对比如图3所示。为了得到模型在传播速率上的预测能力,将真实数据与预测数据在相同的时间轴上展开,其中(a)为微博1的转发情况,(b)为微博2的转发情况。

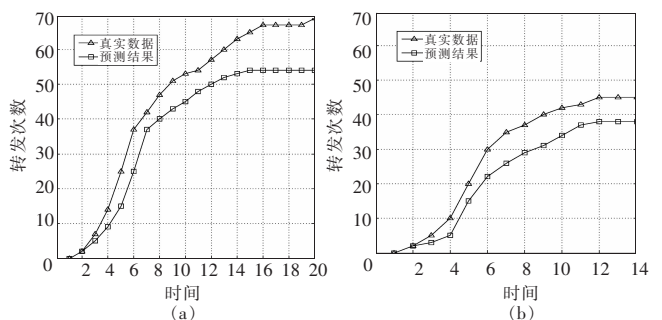


图3 转发次数随时间变化趋势

图3展现了RPT模型在个体层面的转发行为和网络层面的传播速率的预测结果。在最终的转发次数上模型的预测结果与真实数据存在误差,这主要是由于选取的训练集有限,其中并未包含一些用户的转发微博,导致在计算用户转发概率 Pr_u 时出现偏差。

为验证RPT模型在消息传播范围上的预测能力,定义信息扩散密度 $\rho = \frac{R}{N}$, R 为信息传播结束时网络中获得信息的节点个数, N 为网络中总的节点个数。微博1与微博2的扩散密度预测情况如表3所示。

表3 扩散密度预测结果

微博	扩散密度/%		预测准确率/%
	真实数据	预测结果	
1	62	53	85.5
2	46	41	89.1

微博网络中,转发行为是信息传播的关键。在进行RPT传播模型仿真实验时,计算了传播子网中2000个用户的信息接受阈值 $f(u)$,也就是用户对一条微博的转发概率。为了解网络中用户转发行为背后的规律特征,对计算得到的概率值进行了统计,结果如图4所示。

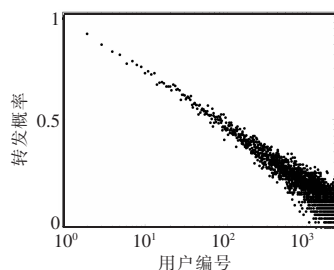


图4 用户转发概率分布

从图4可以看出,网络中用户的转发概率分布呈现显著的幂律特性,网络中对于一条微博,只有少数用户倾向于转发,而大多数用户则不会转发,这种特性说明在网络中存在一定数量的影响力用户对信息传播起着决定性的作用。

(上接第1808页)空间监测系统具有重要的理论意义和实践意义。本文从地下空间的基本特点出发,提出了地下空间节点部署算法的约束指标,建立了基于网格模型的面向地下空间的节点部署算法,并从网格结构、能量消耗及约束目标性能等方面对算法进行了分析和仿真。结果表明,该算法能有效地解决近地表地下空间三维节点部署问题。下一步工作将在本文提出的部署算法的基础上进一步讨论认知无线电技术和链路质量的探测算法,以提升算法的自适应性能。

参考文献:

- [1] YICK J, MUKHERJEE B, GHOSAL D. Wireless sensor network survey [J]. *Computer Networks*, 2008, 52(12): 2292-2330.
- [2] 韩文霆, 吴普特, 郝晓庆. 农业环境信息无线传感器网络监测技术研究进展[J]. *农业工程学报*, 2011, 27(S2): 336-340.
- [3] AKYILDIZ I F, STUNTEBEEK E P. Wireless underground sensor networks: research challenge [J]. *Ad hoc Networks*, 2006, 4(6): 669-686.
- [4] YU Hua-ping, GUO Mei. An efficient oil and gas pipeline monitoring systems based on wireless sensor networks [C]//Proc of International Conference on Information Security and Intelligent Control. 2012: 178-181.

4 结束语

本文在分析微博用户转发行为的基础上,建立了一种基于行为预测的网络传播模型。实验仿真结果表明,RPT模型可以较好地模拟真实网络中的信息扩散速率与扩散范围。此外,通过统计发现,网络中用户转发概率分布呈现出幂律特性,这说明在网络中有少部分用户热衷于转发,从而对信息的扩散起到关键性作用。本文将为进一步进行的微博网络舆论管控研究打下基础。

参考文献:

- [1] 王晶, 朱珂, 汪斌强. 基于信息数据分析的微博研究综述[J]. *计算机应用*, 2012, 32(7): 2027-2029.
- [2] LESKOVEC J, McGLOHON M, FALOUTSOS C, et al. Cascading behavior in large blog graphs patterns and a model [C]//Proc of the Society of Applied and Industrial Mathematics: Data Mining. 2007: 551-556.
- [3] GRUHL D, GUHA R, LIBEN-NOVELL D, et al. Information diffusion through blogspace [C]//Proc of the 13th International Conference on World Wide Web. New York: ACM Press, 2004: 491-501.
- [4] NARAYANAM R, NARAHARI Y. A shapley value-based approach to discover influential nodes in social networks [J]. *IEEE Trans on Automation Science and Engineering*, 2011, 20(1): 130-147.
- [5] LAHIRI M, CEBRIN M. The genetic algorithm as a general diffusion model for social networks [C]//Proc of the 24th AAAI Conference on Artificial Intelligence. 2010: 494-499.
- [6] 赵丽, 袁睿翕, 管晓宏, 等. 博客网络中具有突发性的话题传播模型[J]. *软件学报*, 2009, 20(5): 1384-1392.
- [7] SUH B, HONG Li-chan, PIROLI P, et al. Want to be Retweeted? Large scale analytics on factors impacting retweet in twitter network [C]//Proc of IEEE International Conference on Social Computing. 2010: 177-184.
- [8] CHA M, HADDADI H, BENEVENUTO F, et al. Measuring user influence in twitter: the million follower fallacy [C]//Proc of the 4th International AAAI Conference on Weblogs and Social Media. 2010: 10-17.
- [9] 张旻, 路荣, 杨青. 微博客中转发行为的预测研究[J]. *中文信息学报*, 2012, 26(4): 109-114.
- [5] VURAN M C, KYILDIZ I F. Channel model and analysis for wireless underground sensor networks in soil medium [J]. *Physical Communication*, 2010, 3(4): 245-254.
- [6] 李莉, 刘元安. 无线地下传感器网络节点放置算法[J]. *吉林大学学报: 信息科学版*, 2007, 25(6): 15-20.
- [7] 张启忠. 应用于土壤环境监测的传感器若干理论与技术研究 [D]. 杭州: 浙江大学, 2009: 29-48.
- [8] 王珂. 矿井无线传感器网络节点部署关键技术的研究 [D]. 北京: 中国矿业大学, 2011: 36-48.
- [9] 余华平, 郭春学, 郭梅. 移动 sinks 无线传感器网络的能量效率分析[J]. *计算机工程*, 2009, 35(14): 127-129.
- [10] 余华平, 郭梅, 胡杰. 移动 sinks 无线移动传感器网络时延性能分析[J]. *计算机应用研究*, 2010, 27(1): 317-320.
- [11] HOSSAIN A, CHAKRABARTI B. Sensing models and its impact on network coverage in wireless sensor network [C]//Proc of the 3rd IEEE Region 10 and International Conference on Industrial and Information Systems. 2008: 1-5.
- [12] 刘明, 曹建农, 郑源, 等. 无线传感器网络多重覆盖问题分析[J]. *软件学报*, 2007, 18(1): 127-136.
- [13] 王永华, 杨健, 程良伦, 等. 无线认知传感器网络的研究[J]. *计算机科学*, 2011, 38(7): 47-51.