Journal of Computer Applications
计算机应用,2014,34(8):2404 - 2408,2414
ISSN 1001-9081
CODEN JYIIDU
2014-08-10
http://www.joca.cn

文章编号:1001-9081(2014)08-2404-05
doi:10.11772/j.issn.1001-9081.2014.08.2404

基于行为分析的微博信息传播效果

齐　超*,陈鸿昶,于　岩
(国家数字交换系统工程技术研究中心,郑州 450002)
(*通信作者电子邮箱 qichao352534142@163.com)

摘　要:微博的传播效果研究对于提高市场营销效率、加强舆情监控和准确发现热点具有重要作用。针对以前传播效果研究中未考虑用户个体差异的问题,提出一种基于行为分析的微博转发规模和传播深度预测方法。从微博用户自身、用户关系和微博内容3个方面提取9个相关特征,结合逻辑回归(LR)方法提出一种转发行为预测模型,并基于此模型结合信息沿用户传播特点,通过逐级对相邻用户迭代统计分析得到转发规模和传播深度预测方法。在新浪微博数据集上的实验结果表明,所提方法对转发规模和传播深度预测的正确率分别约为87.1%和81.6%,能较好地预测出信息传播效果。

关键词:微博;行为预测;转发规模;传播深度;逻辑回归
中图分类号:TP391　　**文献标志码**:A

Micro-blog information diffusion effect based on behavior analysis

QI Chao*, CHEN Hongchang, YU Yan
(*National Digital Switching System Engineering and Technology Research Center, Zhengzhou Henan* 450002, *China*)

Abstract: The research of dissemination effect of micro-blog message has an important role in improving marketing, strengthening public opinion monitoring and discovering hotspots accurately. Focused on difference between individuals which was not considered previously, this paper proposed a method of predicting scale and depth of retweeting based on behavior analysis. This paper presented a predictive model of retweet behavior with Logistic Regression (LR) algorithm and extracted nine relative features from users, relationship and content. Based on this model, this paper proposed the above predicting method which considered the character of information disseminating along users and iterative statistical analysis of adjacent users step by step. The experimental results on Sina micro-blog dataset show that the accuracy rate of scale and depth prediction approximates 87.1% and 81.6 respectively, which can predict the dissemination effect well.

Key words: micro-blog; behavior prediction; retweet scale; diffusion depth; Logistic Regression (LR)

0　引言

微博作为一种快速、便捷的信息分享与交互平台,已成为了人们生活中信息交流的重要媒介。微博接入便捷、内容极简,具有原创性、时效性、草根性、随意性和碎片性等特点。近几年微博在全球范围内掀起一股热潮。根据中国互联网络信息中心(China Internet Network Information Center, CNNIC)发布的报告显示,截止2012年12月底,我国微博用户规模为3.09亿,比2011年底增长了5 873万[1]。作为一种新兴的社交媒体,微博不仅是个人自我表达、获取信息的工具,还逐渐发展成为政府、企业、组织用于信息发布、公关营销的手段。与传统社会媒体相比,其信息传播速度、广度和效率都得到了极大的提高。

从本质上说,微博仍是一种传播媒体,其最终目的是向外界传递消息。因此,研究如何利用微博进行高效的信息传播十分必要。近年来,微博网络中的信息传播研究已逐渐成为国内外学者关注的热点。文献[2]通过分析Twitter的拓扑特征,指出微博是一种新的信息分享媒介。在微博网络中用户的转发行为是信息快速传播的重要因素。文献[3]对Twitter的转发功能作了细致分析,探讨人们如何转发,为什么转发以及转发什么的问题。文献[4]针对用户转发行为预测问题提出一种基于特征加权的预测模型。该模型提取了11个用户特征和11个文本特征,并运用信息增益方法对各个特征进行了权重分析,最后通过支持向量机(Support Vector Machine, SVM)算法训练得到预测模型。文献[5]引入了一种线性阈值模型(Linear Threshold Model, LTM)预测用户转发行为,其基本思想是节点被激活的概率随着周围激活节点个数的增加而增大。文献[6]根据文章内容提取了文章类别、客观程度、提及的人物和地名、文章来源4个特征,通过回归算法得到转发量与该4个特征的关系式,由此预测文章被分享到Twitter后会引起多少转发和点击。文献[7]指出不同的特征对微博转发的影响是有差异的,并在对用户转发数据统计分析基础上建立了一个预测微博所能得到转发总数的模型。文献[8]考虑节点度和传播机制的影响,结合复杂网络和传染病动力学理论,进而建立信息传播模型分析信息传播规律。文献[9-10]利用其他动力学模型对社交网络中的信息传播进行了分析。

综上所述,转发行为在微博信息传播中起着重要作用,而

收稿日期:2014-02-11;**修回日期**:2014-03-18。　**基金项目**:国家863计划项目(2011AA010603,2011AA010605)。
作者简介:齐超(1991-),男,江西南昌人,硕士研究生,主要研究方向:通信与信息系统;　陈鸿昶(1964-),男,河南新密人,教授,博士生导师,主要研究方向:通信与信息系统;　于岩(1989-),男,吉林通榆人,硕士研究生,主要研究方向:通信与信息系统。

转发规模和传播深度能较好反映出信息传播范围的大小。因此将转发规模和传播深度作为评价微博传播效果好坏的指标。本文从预测用户转发行为的角度出发,分析影响转发行为的特征,并从微博用户自身、用户关系和微博信息 3 个方面提取了 9 个特征,再结合逻辑回归(Logistic Regression, LR)方法提出一种转发行为的概率预测模型,在此基础上,沿着用户间连接关系逐级对各个用户的转发行为进行预测,通过迭代统计获得转发用户规模的大小,并利用用户转发概率分析出信息传播深度,该方法对转发规模预测的总体正确率约为 87.1%,对传播深度预测的正确率约为 81.6%。

1　问题描述

微博网络中,用户发表微博后,信息有一定概率被粉丝看到,如果粉丝对该微博内容感兴趣,则有可能会对该微博进行转发,如果对该内容没有兴趣,则上述行为不会发生。因此,信息沿好友关系进行传播。对于微博信息的转发路径,可用图 1 来表示。

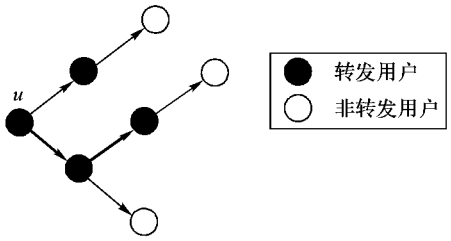

图 1　信息转发路径

为了分析微博的传播效果,需要对 3 个方面的问题进行研究:1)转发行为的预测(即用户微博发布后,其粉丝是否会对其进行转发);2)微博转发规模大小(图 1 中除发布用户外实心圈的个数);3)微博传播深度(图 1 中粗箭头的级数)。通过对这三个问题的分析,能有助于深入了解微博网络中的信息传播机制,并在市场营销、舆情管控等应用方面发挥重要作用。

微博消息的转发行为预测是机器学习中典型的二分类问题,通过对历史数据的训练,在此基础上对于用户 u 新发布的微博 w,得到其粉丝 v 对该微博的转发分类结果。在消息转发预测中,影响转发的属性特征与转发行为呈现出线性关系[7]。因此本文利用 LR 模型来完成这个分类过程,并得到每个用户转发行为发生的概率,进而沿着用户连接关系逐级对用户的转发概率计算并判断是否转发,最后通过迭代统计得到转发规模;同时利用用户转发概率的大小分析出消息传播的深度。

2　基于 LR 的转发行为预测算法

2.1　LR 算法

在转发预测问题中,用 $y_u = f(u,v,w)$ 表示粉丝 u 看见用户 v 发微博 w 后采取的行为,$y_u = 1$ 时表明转发,$y_u = 0$ 时表明不转发。则转发预测问题转化为二分类问题。逻辑回归是解决二分类问题的有效方法之一[11]。结合微博特点得到转发预测公式如式(1)所示:

$$h(u) = p(y_u = 1 \mid u,v,w) = 1/(1 + e^{-\omega C_u}) \tag{1}$$

其中:$p(y_u = 1 \mid u,v,w)$ 为已知 u(微博接收用户)、v(微博发布用户)和 w(发布微博内容)条件下用户 u 转发该微博的概率;$\boldsymbol{C}_u$ 为影响用户 u 转发行为的特征集合;$\boldsymbol{\omega}$ 为特征权值向量,表示不同特征影响转发行为发生的程度。权值的取值由极大似然函数方法得到。对于用户 u 的第 i 条历史观测数据,其状态可用式(2)表示:

$$p(y_u^i \mid u,v,w) = (h(u^i))^{y_u^i}(1 - h(u^i))^{1-y_u^i} \tag{2}$$

其中 $y_u^i \in \{0,1\}$。则对于用户 u 的 N 个独立历史数据,其似然函数如式(3):

$$L(\boldsymbol{\omega}) = p(y_u \mid u,v,w) = \prod_{i=1}^{N} p(y_u^i \mid u,v^i,w^i) = \prod_{i=1}^{N} (h(u^i))^{y_u^i}(1 - h(u^i))^{1-y_u^i} \tag{3}$$

为了计算方便,取对数似然:

$$l(\boldsymbol{\omega}) = \ln L(\boldsymbol{\omega}) = \sum_{i=1}^{N} (y_u^i(\ln h(u^i)) + (1 - y_u^i)\ln(1 - h(u^i))) \tag{4}$$

合理回归就是选择恰当的 $\boldsymbol{\omega}$ 使得 $l(\boldsymbol{\omega})$ 最大,令 $\frac{dl(\boldsymbol{\omega})}{d\boldsymbol{\omega}} = 0$,所求得的 $\boldsymbol{\omega}$ 即为要求的权值。

2.2　特征选取

特征选取的好坏很大程度上决定了转发预测的准确性。微博网络中信息传播的根本动力来源于关注好友间的转发行为。影响转发行为发生的因素主要包含 3 个方面:发布用户、接收用户和微博消息。本文从 3 方面进行特征提取。

2.2.1　发布用户

微博发布者影响力的高低一定程度能影响转发量。如用户名为新京报发布的一条微博(如图 2)。

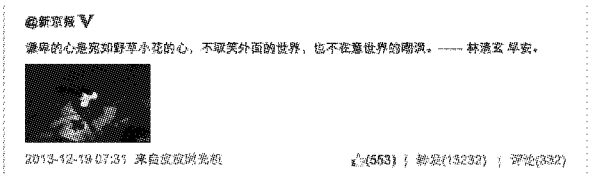

图 2　微博信息

该微博总共引起 13 232 条转发,通过“北京大学微博可视分析工具”分析了此微博的转发情况,其中发布者引起了 1 919次转发,同时对直接转发者后续引发的转发情况进行排名,排名前 3 的用户依次是姚晨、李晨和蒋锡培,他们粉丝数众多,是微博网络中较权威的用户,因此发布用户的影响力[12]是影响转发行为的一个重要因素。在这采用 PageRank 算法来评价用户的影响力[13]。该算法同时结合了用户粉丝与关注两个特征,能相对客观地反映用户影响力的大小。其计算公式如式(5):

$$PR(u_i) = d + (1 - d)\sum_{v_j \in I(u_i)} \frac{PR(v_j)}{O(v_j)} \tag{5}$$

其中:$PR(u_i)$ 为用户 u_i 的影响力;$I(u_i)$ 为用户 u_i 的粉丝集合;$O(v_j)$ 为用户 v_j 的关注数量;d 为阻尼系数,取值一般为 0.15[14],它能使得最后的结果收敛。

在用户影响力的基础上,对用户转发量进行了统计,结果如图 3 所示。

图 3 中:横轴为用户影响力由高到低的排名,纵轴为微博转发量(其值为实际转发数通过以 e 为底的对数计算后的结

果)。显然随着影响力的降低,转发量成下降趋势,表明高影响力用户的微博更容易得到转发。

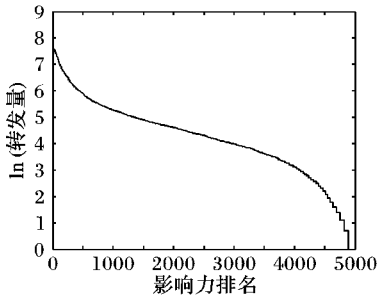

图 3 影响力与转发量的关系

2.2.2 接收用户

影响信息接收者其是否转发的因素主要有两个:习惯和兴趣。有些用户主要通过微博媒介向外表达自己的心情和观点,偏向于发布原创微博,极少去转发。而有的用户添加了许多关注者,通过转发各种微博以提高自己受关注的程度,以便吸引更多的粉丝。又假如两个用户之间相互交流较多关系亲密,则他们之间再次发生转发的概率一般会高于普通用户。再者因为个人兴趣爱好的差异,用户对于转发内容的偏好也会有很大不同。于是本文采用转发活跃度、用户亲密度与兴趣相近度对用户两方面的因素进行度量。

转发活跃度 A_r 一段时间 t 内用户转发微博占所有发布微博的比率,计算公式如式(6):

$$A_r = \left(\sum_{i\in t} r_i\right)/\left(\sum_{i\in t} p_i\right) \tag{6}$$

其中:r_i 为用户第 i 天转发的微博数量,p_i 为用户第 i 天发布的微博数量。

用户亲密度 C 一段时间内用户之间相互交流频度,用户交互行为一般有转发、评论与提及 3 种。则 C 可由式(6)得到:

$$C = \frac{1}{6}(r_{uv} + c_{uv} + m_{uv} + r_{vu} + c_{vu} + m_{vu}) \tag{7}$$

其中:r_{uv} 为用户 u 转发 v 的微博次数,c_{uv} 为评论次数,m_{uv} 为提及次数。

兴趣相近度 R_s 新微博消息与用户历史兴趣空间的相似程度。为了得到该值,需要对兴趣空间进行构造,步骤如下:

1)兴趣收集。收集用户一段时间内发布的 n 条微博信息,构成语句级的兴趣空间 $A_s = \{m_1, m_2, \cdots, m_i, \cdots, m_n\}$,其中 m_i 为第 i 条信息。

2)词语提取。利用中科院计算技术研究所研发的汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System, ICTCLAS)[15] 法对 A_s 进行词语提取,得到词语级的兴趣空间 A_w。

3)停用词剔除。A_w 中常常包含了许多使用广泛但实际意义不大的词,如“你”“的”等,还有一些像“啊”“唉”之类的语气助词。这些词统称为停用词,这些词对于兴趣空间的构建不起作用,因此需要对其进行去除。通过与 CSDN(China Software Developer Network)提供的停用词列表进行对比来剔除。剔除后的兴趣空间记为 A。

4)相近度计算。通过上述步骤得到用户兴趣空间 A 和新微博的特征空间 B。Jaccard 系数常被用来计算符号度量或布尔值度量的个体间的相似度[16]。本文也采用该系数计算 R_s,如式(8)所示:

$$R_s = (A \cap B)/(A \cup B) \tag{8}$$

2.2.3 微博信息

微博的文本特征是影响转发行为的一个重要因素,其重要性和用户特征等同[4]。本文选取如表 1 的 5 个特征作为文本特征。

表 1 微博文本特征

特征序号	特征名称
1	该微博中提及他人的次数
2	该微博是否为回复
3	该微博中 hashtag 的数量
4	该微博以前是否被转发
5	该微博中 URL 的数量

其中特征 2 和 4 采用 0 和 1 表示,其他特征可以由信息中直接获取。再结合前文中的发布用户影响力、转发活跃度、用户亲密度和兴趣相似度 4 个特征,一共 9 个特征用于式(4)的权值计算,然后得到用户转发行为发生概率的预测模型。

3 微博传播效果预测模型

通过上述模型可以得到用户面对一条微博信息时转发的概率,从而反映出个体用户的行为规律。但微博作为一个信息扩散的重要平台,研究信息的传播效果更有价值。因此本章在预测个人转发行为的基础上利用用户间的关系网络结构对信息传播规模和深度的预测进行建模分析。

根据微博网络中用户间相互关注的关系可以构造一个用户关系网,用 $G(V,E)$ 来表示,节点 $u \in V$ 为网络中用户节点的集合,$(u,v) \in E$ 表示用户 u 和 v 存在关注关系。定义网络中每个节点可能有两个状态:活跃状态 I 和移除状态 R。活跃状态表明用户发生转发行为,移除状态表示用户没有发生转发行为。图 4 为节点在网络中的状态。

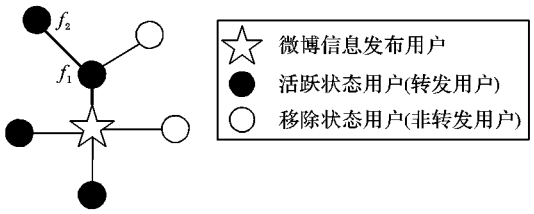

图 4 节点状态

图 4 中:黑粗线为传播距离最远的一条路径。为了方便描述,将与目标节点距离为 1 的 f_1 节点称为一级粉丝,距离为 n 的 f_n 称为 n 级粉丝。假设只有当 m 级节点处于活跃状态后,信息才能由该用户继续向 $m+1$ 级节点传播。同时各节点之间的转发行为的发生是相互独立的。图 3 中该用户的微博得到了 3 个一级粉丝和 1 个二级粉丝的转发,因此转发规模为 4,距离目标用户最远的转发用户是 f_2,因此传播深度为 2。下面建立基于转发行为的微博传播效果预测模型。

预测时从微博发布用户开始沿粉丝路径逐级进行转发行为预测,并判断节点是否会转变为活跃状态,依次迭代直到不再有新的活跃节点产生,统计网络中活跃节点的个数得到转发规模,并分析出距离目标用户最远的活跃节点,从而得到传

播深度。具体步骤如下:

1) 准备阶段。通过统计分析用户的历史数据,得到每个节点的转发行为预测模型参数 $\boldsymbol{\omega}$。

2) 初始阶段。对于用户 u,得到其粉丝集合记为 F_u,用 F_u^i 表示集合中的第 i 个用户。

3) 预测阶段。遍历集合 F_u 中的粉丝 F_u^i,计算其对于发布微博 w 的转发概率 $h(F_u^i)$,同时设定一个 θ 作为阈值门限,且 $h(F_u^i),\theta \in [0,1]$,当 $h(F_u^i) \geqslant \theta$ 时,粉丝 F_u^i 发生转发行为成为活跃状态;否则为移除状态。将增加的活跃节点加入集合 $A(N)$ 和 $A_t(N)$。

4) 迭代阶段。针对 $A_t(N)$ 中的用户依次重复粉丝获取和转发行为预测过程,将新增加的活跃节点作为新的 $A_t(N)$,并将其加入 $A(N)$。

5) 结束阶段。重复预测迭代过程,直到 $A_t(N)$ 为空停止,得到最终的 $A(N)$。

在迭代过程中,如果遇到对 $m(m > 1)$ 级粉丝进行转发预测,则需要注意它可能同时连接多个 $m-1$ 级粉丝。假设处于 m 级的节点 i(记为 f_m^i)同时与 n 个 $m-1$ 级节点有边相连,且该 n 个节点都发生转发行为,如图 5 所示。

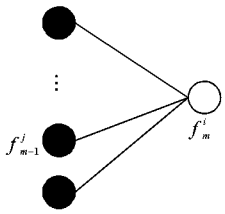

图 5　多连接转发概率

则可以得到节点 i 的转发概率计算公式,如式(9):

$$h(f_m^i) = \max\{p(y_i = 1 \mid i,j)h(f_{m-1}^j)\}; j \in \{1,2,\cdots,n\} \tag{9}$$

在此基础上判断其与 θ 值的大小关系,从而确定该节点是否会转发。同时在此过程中能得到各级节点的转发概率,将满足式(10)条件的最大级数作为传播深度 R_D:

$$\exists i, h(f_m^i) \geqslant \theta, \forall j, h(f_{m+1}^j) < \theta, R_D = \max(m) \tag{10}$$

最后计算集合中 $A(N)$ 的节点个数,得到转发规模 $R_S = card(A(N))$。

该模型的伪代码实现如下。

```
Input: G(V,E), post user u and weibo w。
Output: R_S and R_D。
Initialization: Set A(N) and A_t(N) Null, R_S = R_D = 0
Procedures:
1)   Put u in A_t(N)                          // 将用户 u 加入初始集合
2)   While A_t(N) is not Null
3)     For each user j in A_t(N)
4)       Get j's follower set F_j             // 获取 j 的粉丝集合
5)       For each user k in F_j
6)         Calculate h(k)                     // 计算 k 的转发概率
7)         If h(k) > θ
8)           Put k in A(N)                    //将用户加入转发集合
9)     End
10)  End While
11)  Update A_t(N)                            //更新初始集合
12)  If A_t(N) is Null
13)    Get R_D according to formula           //获得传播深度
14)    R_S = card(A(N))                       // 计算转发规模
15)  Else
16)    back to step 2)                        //对初始集合用户进行迭代
17)  End
```

4　实验分析

4.1　实验数据集

本文的实验数据来源于新浪微博,利用其公开的应用程序编程接口(Application Programming Interface,API)实现数据爬取[17],从某一用户出发,逐层爬取用户信息与微博信息。本文爬取了从 2013-03-01—2013-03-30 的部分用户微博信息作为研究数据。获取的信息统计如表 2 所示。

表 2　微博数据统计

统计信息	数据值
统计时间	2013-03-01—2013-03-30
统计人数	5 000
微博信息总数	220 526
原创微博数	73 259
人均发微博数	44.1

将数据按时间分为两部分:3 月 1 日至 3 月 15 日作为训练数据集(包含 119 145 条微博数据),用于建立用户的预测模型;之后的作为测试数据集(包含 101 381 条数据,其中转发微博 63 724 条,非转发微博 37 657 条),对模型的性能作分析。

4.2　转发行为预测分析

为了评价转发行为预测模型的效果,用图 6 的混淆矩阵形式表示预测结果。

	预测转发	预测非转发
实际转发	a	b
实际非转发	c	d

图 6　混淆矩阵预测结果

图 6 中:a、b、c 和 d 分别表示各种情况下的百分比,同时选用总体命中率 T_r 指标来评价整体预测的效果。该指标的计算公式如式(11)所示:

$$T_r = (a+d)/(a+b+c+d) \tag{11}$$

采用 T_r 作为阈值 θ 选择的一个评判,图 7 为不同 θ 下该指标的变化趋势。

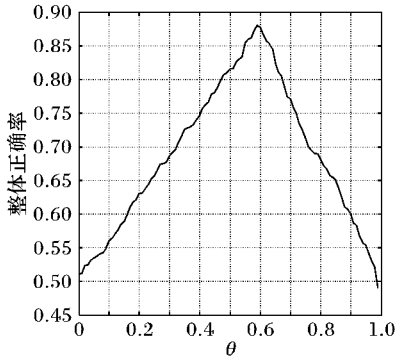

图 7　T_r 与 θ 关系

图 7 表明,在 $\theta = 0.58$ 时,该指标的值最大,说明此时转发预测整体准确率较高,约为 88.2%,因此在预测模型中将阈值 θ 选为 0.58,并得到预测结果如图 8 所示。

	预测转发	预测非转发
实际转发	87.1%	12.9%
实际非转发	10.9%	89.1%

图 8 实际预测结果

4.3 传播效果预测分析

在转发行为预测的基础上,对传播效果进行了仿真实验。传播效果包含对转发规模和传播深度两个方面的评判,因该两个数值数量级可能相差较大,因此采取不同的评价标准。

转发规模符合冥律分布,因此根据数量级来划分数量规模,数量规模[18]定义如下。

假设正整数 a、b、m,满足 $a < b$ 且 $10^a < m < 10^b$,m 的数量规模为以 m 为中点,左右各扩展数量级所在区域的一半,即:

$$S_m = \left[m - \frac{10^b - 10^a}{2}, m + \frac{10^b - 10^a}{2} \right] \tag{12}$$

故当转发规模真实值为 N_f,预测值为 N_p 时,如若满足式(13)判为预测正确:

$$|N_p - N_f| < (10^{\lg\lceil N_f \rceil} - 10^{\lg\lfloor N_f \rfloor})/2 \tag{13}$$

其中:$\lceil \cdot \rceil$为向上取整,$\lfloor \cdot \rfloor$为向下取整。

然后对 1 000 个用户的 13 297 条原创微博进行了转发规模预测,并在上述评价标准下计算了预测准确率,结果如图 9 所示。

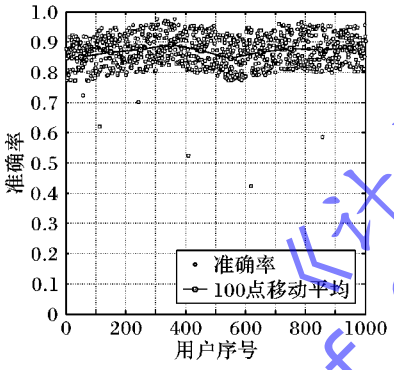

图 9 转发规模预测准确率

图 9 表明该预测模型对于不同用户的预测准确率基本都能达到 80% 以上,总体准确率约为 87.1%,能较好地预测转发规模。

传播深度的级数一般较少,数值较小,因此采用偏差率 γ 来衡量,假设实际传播深度为 D_f,预测传播深度为 D_p,则 γ 由式(14)得到:

$$\gamma = |D_p - D_f| / \max\{D_p, D_f\} \tag{14}$$

对上述数据进行实验得到传播深度预测的偏差率如表 5 所示。

表 5 传播深度预测结果

偏差率范围	所占比值/%
$\gamma = 0$	81.6
$0 < \gamma \leqslant 0.4$	16.7
$0.4 < \gamma < 1$	1.4
$\gamma = 1$	0.3

表 5 表明对于不同用户的微博该预测方法的正确率约为 81.6%,传播深度预测十分不理想的微博约占 1.7%。该方法总体能较好地反映出微博传播深度的趋势。

5 结语

微博的传播效果研究对于市场营销、舆情监控和热点发现等方面具有重要作用。而转发规模和传播深度是衡量微博传播效果的重要指标。本文通过分析影响转发行为的因素,从用户自身、用户关系和微博内容 3 个角度提取了 9 个特征,并结合 LR 算法建立了转发行为预测模型。通过对新浪微博数据的实验表明该方法对转发行为预测的准确率约为 88.2%。在此基础上,提出了一种基于转发行为的转发规模和传播深度预测算法,并给出了转发规模和传播深度的评价方法。实验结果表明,该方法对转发规模预测的总体准确率约为 87.1%,传播深度预测的正确率约为 81.6%,能较客观反映出微博发布后的后续传播效果。后续将在本文基础上对微博信息传播规律和方式以及传播路径的形成作进一步的研究。

参考文献:

[1] The 31st report of China Internet development statistics [R]. Beijing: China Internet Network Information Center, 2013. (第 31 次中国互联网络发展状况统计报告[R]. 北京: 中国互联网络信息中心, 2013.)

[2] KWAK H, LEE C, PARK H, *et al*. What is twitter, a social network or a news media? [C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 591 – 600.

[3] BOYD D, GOLDER S, LOTAN G. Tweet, tweet, retweet: conversational aspects of retweeting on twitter [C]// HICSS 2010: Proceedings of the 43rd Hawaii International Conference on System Sciences. Piscataway: IEEE Press, 2010: 1 – 10.

[4] ZHANG Y, LU R, YANG Q. Predicting retweeting in microblogs [J]. Journal of Chinese Information Processing, 2012, 26(4): 109 – 114. (张旸, 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] NARAYANAM R, NARAHARI Y. A shapley value-based approach to discover influential nodes in social networks [J]. IEEE Transactions on Automation Science and Engineering, 2011, 8(1): 130 – 147.

[6] BANDARI R, ASUR S, HUBERMAN B A. The pulse of news in social media: forecasting popularity [C]// ICWSM 2012: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media. Menlo Park: AAAI Press, 2012: 26 – 33.

[7] SUH B, HONG L, PIROLLI P, *et al*. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network [C]// Proceedings of the 2010 IEEE Second International Conference on Social Computing. Piscataway: IEEE Press, 2010: 177 – 184.

[8] ZHANG Y, LIU Y, ZHANG H, *et al*. The research of information dissemination model on online social network [J]. Acta Physica Sinica, 2011, 60(5): 60 – 66. (张彦超, 刘云, 张海峰, 等. 基于在线社交网络的信息传播模型[J]. 物理学报, 2011, 60(5): 60 – 66.)

[9] LAHIRI M, CEBRIAN M. The genetic algorithm as a general diffusion model for social networks [C]// Proceedings of the 24th AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2010: 494 – 499.

[10] ZHAO L, YUAN R, GUAN X, *et al*. Bursty propagation model for incidental events in blog networks [J]. Journal of Software, 2009, 20(5): 1384 – 1392. (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.)

(下转第 2414 页)

参考文献：

[1] MERATNIA N, ROLF A. Spatiotemporal compression techniques for moving point objects [C]// EDBT 2004: Proceedings of the 9th International Conference on Extending Database Technology. Berlin Springer, 2004: 765 – 782.

[2] CAO H, WOLFSON O, TRAJCEVSKI G. Spatio-temporal data reduction with deterministic error bounds [J]. The VLDB Journal—the International Journal on Very Large Data Bases, 2006, 15(3): 211 – 228.

[3] DOUGLAS D H, PEUCKER T K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature [J]. Cartographica: the International Journal for Geographic Information and Geovisualization, 1973, 10(2): 112 – 122.

[4] BELLMAN R. On the approximation of curves by line segments using dynamic programming [J]. Communications of the ACM, 1961, 4(6): 284.

[5] POTAMIAS M, PATROUMPAS K, SELLIS T. Sampling trajectory streams with spatiotemporal criteria [C]// SSDBM'06: Proceedings of the 18th International Conference on Scientific and Statistical Database Management. Piscataway: IEEE Press, 2006: 275 – 284.

[6] GUTING R H, SCHNEIDER M. Moving objects databases [M]. Amsterdam: Elsevier, 2005: 25 – 30.

[7] YUAN J, LUO J, YUE X. Principle and application of satellite navigation [M]. Beijing: China Astronautic Publishing House, 2003: 191 – 218. (袁建平, 罗建军, 岳晓奎. 卫星导航原理与应用[M]. 北京: 中国宇航出版社, 2003: 191 – 218.)

[8] KAPLAN E D, HEGARTY C J. Understanding GPS: principles and applications [M]. Boston: Artech House, 2006: 379 – 390.

[9] LIU G, IWAI M, SEZAKI K. An online method for trajectory simplification under uncertainty of GPS [J]. Information and Media Technologies, 2013, 8(3): 665 – 674.

[10] LONG C, WONG R C W, JAGADISH H V. Direction-preserving trajectory simplification [J]. Proceedings of the VLDB Endowment, 2013, 6(10): 949 – 960.

[11] TRAJCEVSKI G, CAO H, SCHEUERMANNY P, *et al.* On-line data reduction and the quality of history in moving objects databases [C]// Proceedings of the 5th ACM International Workshop on Data Engineering for Wireless and Mobile Access. New York: ACM Press, 2006: 19 – 26.

[12] PARK H, LEE Y J, CHAE J, *et al.* Online approach for spatio-temporal trajectory data reduction for portable devices [J]. Journal of Computer Science and Technology, 2013, 28(4): 597 – 604.

[13] FAN C. The study of geometric algorithms: color-spanning set, Voronoi diagram and Fréchet distance [D]. Changsha: Central South University, 2011. (范成林. 基于颜色支撑点集 Voronoi 图和 Fréchet 距离的几何算法研究[D]. 长沙: 中南大学, 2011.)

[14] JENSEN C S, LAHRMANN H, PAKALNIS S, *et al.* The INFATI data [EB/OL]. [2014- 01- 11]. http://arxiv.org/abs/cs/0410001.

[15] HÖNLE N, GROSSMANN M, REIMANN S, *et al.* Usability analysis of compression algorithms for position data streams [C]// Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2010: 240 – 249.

[16] LANGE R, DÜRR F, ROTHERMEL K. Online trajectory data reduction using connection-preserving dead reckoning [C]// Proceedings of the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services. Brussels: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 2008: 52.

[17] SOROUSH E, WU K, PEI J. Fast and quality-guaranteed data streaming in resource-constrained sensor networks [C]// Proceedings of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing. New York: ACM Press, 2008: 391 – 400.

[18] LI Y. The spatial index method based on minimum bounding circle and minimum encircling sector [D]. Harbin: Harbin University of Science and Technology, 2009. (李杨. 基于最小边界圆和最小包围扇形的空间索引方法[D]. 哈尔滨: 哈尔滨理工大学, 2009.)

(上接第 2408 页)

[11] SHI C, ZHANG M. Analysis of logistic regression models [J]. Computer Aided Engineering, 2005, 14(3): 74 – 78. (施朝健, 张明铭. Logistic 回归模型分析[J]. 计算机辅助工程, 2005, 14(3): 74 – 78.)

[12] CHA M, HADDADI H, BENEVENUTO F, *et al.* Measuring user influence in twitter: the million follower fallacy [C]// ICWSM 2012: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media. Menlo Park: AAAI Press, 2010: 10 – 17.

[13] ZHANG Y, ZHANG H, ZHANG W. Quick ranking algorithm for network user based on power law distribution [J]. Journal of Chinese Information Processing, 2012, 26(4): 122 – 128. (张玥, 张宏莉, 张伟哲. 基于幂律分布的网络用户快速排序算法[J]. 中文信息学报, 2012, 26(4): 122 – 128.)

[14] RICHARDSON M, DOMINGOS P. Combining link and content information in Web search [M]// Web dynamics: adapting to change in content, size, topology and use. Berlin: Springer, 2004: 179 – 193.

[15] LIU Q, ZHANG H, BAI S. An open resource platform for Chinese NLP [J]. Applied Linguistics, 2002(4): 50 – 56. (刘群, 张浩, 白硕. 自然语言处理开放资源平台[J]. 语言文字应用, 2002(4): 50 – 56.)

[16] LIN X, WANG W. Set and string similarity queries: a survey [J]. Chinese Journal of Computers, 2011, 34(10): 1853 – 1862. (林学民, 王炜. 集合和字符串的相似度查询[J]. 计算机学报, 2011, 34(10): 1853 – 1862.)

[17] LIAN J, ZHOU X, CAO W, *et al.* SINA microblog data retrieval [J]. Journal of Tsinghua University: Science and Technology, 2011, 51(10): 1300 – 1305. (廉捷, 周欣, 曹伟, 等. 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.)

[18] LI Y, YU H, LIU L. Predict algorithm of micro-blog retweet scale based on SVM [J]. Application Research of Computers, 2013, 30(9): 2594 – 2597. (李英乐, 于洪涛, 刘力雄. 基于 SVM 的微博转发规模预测方法[J]. 计算机应用研究, 2013, 30(9): 2594 – 2597.)