

基于模糊 PA 算法的微博信息 传播分享预测研究^{*}

田占伟¹, 刘 臣², 王 磊¹, 隋 场³

(1. 东北农业大学 工程学院, 哈尔滨 150030; 2. 上海理工大学 管理学院, 上海 200001; 3. 黑龙江大学 新闻传播学院, 哈尔滨 150080)

摘要: 微博已经成为网民信息获取、分享的主要平台之一。对信息分享进行预测, 是对微博信息传播进行监管控制的基础。微博用户和信息属性中包含着用户偏好、生理特征、内容类型等数据, 基于这些数据可进行信息分享预测。分析了微博信息传播模式、分享预测理论方法, 基于 PA 算法提出了信息分享预测模型, 以新浪微博数据为例验证了预测模型。结果表明, 该模型对信息分享具有较高的预测准确率。

关键词: 微博; 信息分享; PA 算法; 模型预测

中图分类号: N99; TP393

文献标志码: A

文章编号: 1001-3695(2014)01-0051-04

doi:10.3969/j.issn.1001-3695.2014.01.011

Research on prediction of micro-blog information dissemination based on PA algorithm

TIAN Zhan-wei¹, LIU Chen², WANG Lei¹, SUI Yang³

(1. School of Engineering, Northeast Agricultural University, Harbin 150030, China; 2. School of Management, University of Shanghai for Science & Technology, Shanghai 200001, China; 3. School of Journalism & Communication, Heilongjiang University, Harbin 150080, China)

Abstract: Micro-blog has become one of the main Internet platforms for information acquiring and sharing. Prediction on information sharing is the basis for controlling and supervising information. Attributes of micro-blog users and information contain data of users' preferences, physiological characteristics, content type, etc. Based on these data can predict information sharing. This paper analyzed modes of micro-blog information dissemination, theory and methods for predicting information sharing, based on PA algorithm, it proposed information sharing prediction model. Through Sina micro-blog data verified the model, it shows that the model has high prediction accuracy of information sharing.

Key words: micro-blog; information sharing; PA algorithm; model prediction

0 引言

微博已经成为互联网用户获取信息的主要平台之一。信息产生主体为普通用户的微博信息, 具有时效性强、碎片化、病毒式传播等特征, 在社会公众事件、突发性事件的传播方面, 已经对传统媒体造成了冲击^[1]。

与传统互联网应用如 BBS 论坛、人人网等交友网站相比, 微博的重要特征是用户之间的关注关系为弱联系, 这使得微博信息传播网络的平均路径小、信息传播效率高^[2]。信息由于用户的不断分享, 在网络中产生裂变式传播, 分享成为微博中信息传播机制的基础。对信息分享进行预测, 一方面是对信息传播的广度、速度、路径和影响力进行预测的基础, 另一方面可以挖掘用户的行为模式, 为信息推荐、精准广告营销等电子商务应用提供依据。目前, 针对微博信息传播的研究, 集中在数据挖掘、电子商务应用等领域^[3, 4], 而对信息传播分享预测缺乏理论研究。

本文首先分析微博信息传播分享模式; 然后对预测微博信息分享的理论知识进行研究, 基于 PA 算法提出分享预测模型; 最后利用新浪微博数据进行实例分析, 验证模型的有效性。

1 微博信息传播分享模式

微博中信息传播的速度要远快于传统媒体和其他在线社会网络。传统传播理论将信息传播方式划分为大众传播和人际传播两种。大众传播的信息由源头向一个较大群体的单向传递, 人际传播是两个或多个个体之间的传递, 而微博中信息的传播是两种传播模式的结合^[5]。

相较于传统的媒体或在线社会网络, 微博创造了新的信息传播模式。即时通信工具如 MSN、QQ 等, 是个人一个人的单线传播, 传统媒体、博客、论坛是个人—小众传播模式, 这两种模式都是单层传播, 不能使信息形成裂变式传播, 微博恰恰弥补了这一点, 形成裂变式的个人—小众—大众的传播模式, 如图 1 所示。

收稿日期: 2013-03-08; **修回日期:** 2013-05-03 **基金项目:** 国家自然科学基金资助项目(70901023); 国家教育部人文社会科学研究青年资助项目(12YJCZH126)

作者简介: 田占伟(1983-), 男, 河北河间人, 博士, 主要研究方向为管理信息系统、信息传播、在线社会网络(tianzhanwei@gmail.com); 刘臣(1981-), 男, 讲师, 博士, 主要研究方向为信息传播、在线社会网络; 王磊(1980-), 女, 讲师, 博士, 主要研究方向为知识管理; 隋场(1985-), 女, 硕士, 主要研究方向为新闻传播。

关注与被关注在微博信息传播的核心,用户 A 关注用户 B 后,B 产生信息后随即自动推送给 A,A 通过转发、评论后,该信息进一步分享给 A 的关注者,如此反复形成裂变式传播。关注关系的增加主要通过用户浏览信息时发现感兴趣的其他用户、特意的搜索、系统好友推荐等。不同于一般的 SNS,微博中关注某个用户,并不需要经过允许,被关注用户也不必作出回应,这样有利于大量关注关系的产生,增加信息传播路径。

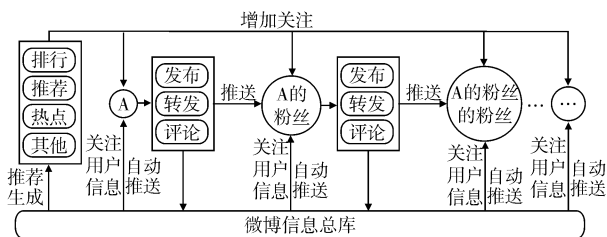


图1 微博信息裂变传播模式

用户作为节点处在微博信息传播网络中,他们之间相互关注的强关系可看做是人与人之间直接的信息交互,而大量的单向关注弱关系可看做一个用户与所有相关的其他用户之间进行一对多的信息交互。微博中信息分享的模式如图2所示。

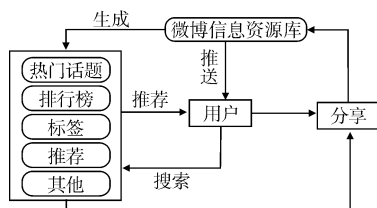


图2 微博信息分享模式

微博系统自动推送给用户其关注用户产生的信息,用户通过直接浏览、系统推荐、搜索功能可以获取非关注用户信息,信息在不同用户之间的分享形成传播网络。

2 微博信息分享预测理论

信息传播路径预测是目前的研究热点,具有重要的理论和实践意义。而信息分享预测是传播路径预测的基础。分享预测是指通过已知的网络结构、用户行为等信息,来预测网络中不同节点间产生信息分享的可能性。

不同领域的学者对信息分享、物质交换、传播路径预测进行了研究,如生物技术、社会网络分析、计算机网络等^[6,7]。其主要预测方法是利用马尔可夫链和机器学习算法,预测中用到的前提条件主要是基于节点的属性信息或者网络的拓扑结构^[8]。

基于节点属性的预测是利用不同节点属性之间的相似性、节点的行为特征进行直接的传播链路预测。这种预测结果比较准确,但节点属性需要提前获取,而节点属性的真实信息在现实网络中是很难获取的。比如微博用户许多注册信息,出于保护个人隐私的考虑,都是保密的。而在个人公开资料中,也存在着大量虚假信息。此外,用户的行为特征,需要对用户进行大量的跟踪统计调查才能获取。在获取了准确的资料后,哪些资料可用来对信息传播的链路进行预测,也需要进行判断。基于节点属性的预测掺杂了诸多主观因素和不确定性,实现起来非常困难。

基于网络拓扑结构的预测,主要是利用网络中路径、演化过程、节点的结构属性等来进行信息分享预测。这种方法的优

点是,网络的拓扑结构特征可以非常容易获取,而且是客观存在的因素,如果预测的方法选择正确,预测结果也具有较高的准确性,具有相似拓扑结构的网络之间可以相互借鉴预测指标选取、预测算法;这种方法的缺点是,没有考虑节点的属性,网络的复杂性不仅体现为网络结构复杂,也体现在节点的复杂行为方面。

无论是基于节点属性,还是基于网络结构特征的预测,都是尽可能真实地描述现有数据,在此基础上实现预测。每种方法适用的网络不同,如果网络的结构特征非常清晰,基于网络结构的预测方法不仅能够降低计算的复杂度,而且能够取得较好的预测结果。如果网络的结构复杂,含有加权边、有向边,利用网络结构进行预测,计算复杂度会呈几何级增加,准确性也会下降很多。

微博中信息的分享更多的是用户主观情绪作用的结果,因此可基于节点属性进行预测。从微博中用户的个人信息、传播行为、信息相关属性等方面总结节点属性特征,可对信息分享进行预测。这里的难点是这些属性测量时的量化、属性对信息分享预测的量化。

3 微博信息分享预测模型

微博中信息涉及的转发用户、非转发用户数量众多,数据量较大时使用决策树、关联规则等传统算法进行预测,速度较慢。本文基于 PA 算法,构建信息分享预测模型。该算法适宜处理大规模数据的二分类、多级分类、序列预测等问题,速度较快,但需要较大的数据量进行训练,对存储空间要求较高,而且预测的结果容易被预测指标干扰。

3.1 PA 算法

不同于传统的基于已知整体数据集的预测,通过微博的时间线(timeline)可以获得微博中发布的实时信息。在信息发布出来后,即需要进行信息分享与否的预测,处理这种实时的预测问题需要在线学习算法。本文对 PA(passive-aggressive)算法^[9]进行改进,利用用户和信息的属性对分享进行预测。

对信息是否进行分享是个二分类问题。PA 算法的基本思路是:微博信息产生是一个连续的时间序列,在每一个序列,对新出现的信息进行用户是否分享预判,将用户标志为 0(不分享)或 1(分享),预判结束后,在该时间序列结束时,用户是否进行了分享的结果将出现,算法将会出现瞬时的损失(instantaneous loss)来反映算法判断错误的程度,并利用该时间序列内产生的新的属性和是否分享的数据,来对已有判别规则进行更新,利用新规则对下一个时间序列进行判断。

算法的形式化表示如下:假定 x_t 为时序 t 内的实例,表示为向量空间 R^n 内的一个向量, $y_t \in \{0, 1\}$ 为判断类别,则 (x_t, y_t) 可作为一个样本。算法利用一个函数进行样本的分类识别:

$$y_{t+1} = \arg \max_y (w^t \times \varphi(x_t, y_t)) \quad (1)$$

该函数基于权重向量 $w \in R^n$,算法的主要任务即是通过循环学习来更新权重。PA 算法提供了三种策略来更新权重。权重可表示为

$$w_{t+1} = \arg \min_w \frac{1}{2} \|w - w_t\|^2 + C\xi$$

$$\text{s.t. } \ell(w; \varphi(x_t, y_t)) \leq \xi \quad \xi \geq 0 \quad (2)$$

其中: $\ell(w; \varphi(x_t, y_t))$ 为瞬时损失函数,定义为

$$\ell(w; \varphi(x_t, y_t)) = \begin{cases} 0 & \gamma(w; \varphi(x_t, y_t)) > 1 \\ 1 - \gamma(w; \varphi(x_t, y_t)) & \text{其他} \end{cases} \quad (3)$$

$$\begin{aligned} \gamma(w; \varphi(x_t, y_t)) &= w^t \times \varphi(x_t, y_t) - w^t \times \varphi(x_t, y') \\ y' &= \arg \max_{z \neq y} (w^t \times \varphi(x_t, z)) \end{aligned} \quad (4)$$

式中: y' 为分类识别的最佳类别。

对于式(4),可以利用拉格朗日算法求出最优解:

$$w = w_t + a^* \varphi(x_t, y_t) - \varphi(x_t, y') \quad (5)$$

$$a^* = \min(C, \frac{\ell_t}{\|\varphi(x_t, y_t) - \varphi(x_t, y')\|^2}) \quad (6)$$

式中: ℓ_t 为 t 时序的损失函数值,依赖于损失函数。PA 算法提供了三种计算 w 的变形,具体流程如下所示:

INPUT: aggressiveness parameter $C > 0$

INITIALIZE: $w_1 = (0, \dots, 0)$

for $t = 1, 2, \dots$

receive instance: $x_t \in \mathbb{R}^n$

predict: $y_t = \text{sign}(w_t \cdot x_t)$

receive correct label: $y_t \in \{-1, +1\}$

suffer loss: $\ell_t = \max\{0, 1 - y_t(w_t \cdot x_t)\}$

update:

set:

$$\tau_t = \frac{\ell_t}{\|x_t\|^2} \quad (\text{PA})$$

$$\tau_t = \min\left\{C \cdot \frac{\ell_t}{\|x_t\|^2}\right\} \quad (\text{PA-I})$$

$$\tau_t = \frac{\ell_t}{\|x_t\|^2 + \frac{1}{2C}} \quad (\text{PA-II})$$

update: $w_{t+1} = w_t + \tau_t y_t x_t$

3.2 基于PA算法的分享预测模型

研究表明,超过50%的微博信息转发出现在微博发布后的1 h内^[10],微博是否被分享受时间因素影响。微博用户每天登录时段存在三个高峰期,显然在这三个时段内发布的信息被转发的概率大。这样仅仅建立一个基于PA算法的全局预测模型会产生较大的误差。因此,考虑到时间段因素,本文对模型进行如下构建:

a) 利用PA算法建立全局模型,利用整体数据对全局规则进行更新。

b) 针对一天之中的每一个小时,同样利用PA算法构建24个局部模型,利用该小时内的数据进行训练,更新局部规则。

c) 进行预测时,修改原有算法的判别函数,使得判别结果由全局和局部规则共同决定:

$$y = \arg \max_y (w^t \times \varphi(x_t, y_t) + \alpha(w^\zeta \times \varphi(x_\zeta, y_\zeta))) \quad (7)$$

式中: w^t 为全局权重; w^ζ 为特定小时 ζ 时间段的局部权重; α 为权重系数,可以定义为两部分训练样本的比例,即 $\alpha = \zeta_n / N$; ζ_n 为时间段的样本数量; N 为所有样本量。这样 ζ 时间段内的样本量越大,该时间段内的局部判别函数起到的作用也越大。

4 实例分析

4.1 数据获取

以新浪微博数据为例来验证上述预测方法的有效性。对于训练数据集,本文采集1 h内的数据进行研究。这里随机选取了微博中的部分原创性微博进行研究,以匿名方式登录新浪

微博,通过“大家正在说”可以查看新浪微博实时更新的消息,这其中大量的转发性消息(以是否包含“@”为判断依据),去除这部分消息后,获得原创性消息的ID和信息发布者ID。由于新浪微博限制,只能获取1 000个用户粉丝数据,因此首先对发布者资料进行分析,如果其粉丝数量超过1 000,舍弃该消息;否则,将消息ID、消息发布者ID、发布者粉丝ID保存。1 h后利用API接口通过消息ID读取消息评论,获得全部转发用户ID,利用用户ID可获取信息发布者和受众的基本特征指标数据。对于用户和信息类型数据目前没有自动的识别方法,只能通过人工识别的方式进行。最终数据包含微博信息5 142条,涉及到的转发用户1.4万,非转发用户70余万,每个信息发布者发布的微博信息大约只有2%的粉丝会进行转发。

4.2 预测指标选取

微博信息分享过程涉及用户和信息主体。用户又可以分为信息发布者和信息受众。结合本文分析和其他学者的研究成果^[11, 12],对两个主体可以进行预测指标选取,结果如表1所示。

表1 用户和信息属性中选取的预测指标

主体类型	指标名称	备注说明
用户	1. 粉丝数量	这几个指标与用户的活跃性有关。文献[11]中的相关研究验证了部分指标的显著性
	2. 关注用户数量	
	3. 微博数量	
	4. 用户类型	
	5. 是否VIP认证	
	6. 性别	
信息	1. 内容类型	文献[12]的研究发现,这些指标与信息是否被分享具有相关性
	2. 长度	
	3. 标签数量	
	4. URL数量	
	5. 时间段	

需要指出,这些指标只是初步预选结果,在预测的过程中,会对指标进行交叉组合,对显著性作进一步检验,确定预测最佳指标。

4.3 基本预测

在进行预测之前,对数据进行预处理,利用min-max方式(新数据=(原始数据-极小值)/(极大值-极小值)),使得原始数据映射成[0,1]之间的新值。利用信息指标和转发用户指标进行转发的训练学习,利用信息指标和非转发用户指标进行非转发训练。相较于非常大的信息不被转发概率,本文更关心信息被转发识别,因此利用判断为被转发的准确率对模型进行评估。利用随机识别、多数识别(将所有数据均判断为转发)、标准PA模型、改进PA模型四种方式对样本数据进行判断,得到的预测识别准确率如表2所示。

表2 四种分类方法得到的预测识别准确率结果 /%

结果	模型			
	随机	多数	标准PA	改进PA
准确率	2.93	2.37	35.5	47.2

由表2可知,利用随机和多数识别方法,判断的准确率非常低。利用标准PA算法模型后,可以提高准确率到35.5%,而经过改进后的PA模型,又可以提高将近12%。

4.4 不同种类指标预测

考虑了时间因素后,预测的准确率会有所提升,这促使本

文进一步思考,所选择的测试指标哪些是对时间段敏感的。通过分别利用用户和信息指标进行预测,得到的预测识别准确率结果如表 3 所示。

表 3 用户和信息指标分别进行预测的识别准确率结果 /%

指标	模型	
	标准 PA	改进 PA
用户相关	37.6	38.2
信息相关	20.3	34.6

由表 3 可知,改进的 PA 模型对信息相关的指标进行的预测准确率提升要比用户相关指标高很多。因此,可知信息相关指标对时间段相对敏感。这其实可以按照常理来解释,不同的用户在一天之内的任何时间段都可能发布信息,而某类内容的信息可能只出现在某时间段内,如早饭相关信息更多地会出现在早晨。

4.5 不同指标的显著性分析

目前预测用到的指标是表 1 中所有的指标,通过观察改进 PA 模型全局预测部分分配给各个指标的权重如表 4 所示,可以发现这些指标对预测的贡献率不同。

表 4 PA 模型中各指标的权重

指标	权重
粉丝数量	0.43
关注用户数量	0.22
微博数量	0.38
用户类型	0.14
是否 VIP 认证	0.27
性别	0.31
内容类型	0.19
长度	-0.17
标签数量	-0.06
URL 数量	-0.09

由表 4 可知,各指标的权重存在较大差别,因此可以对指标进行精简,进而提高模型的效率。分三个阶段分别去掉权重绝对值小于 0.1、0.2、0.3 的指标,利用改进的 PA 模型进行预测准确率如表 5 所示。

表 5 去掉不同指标后的预测准确率结果

结果	指标			
	全部	去掉 <0.1	去掉 <0.2	去掉 <0.3
准确率/%	47.2	45.5	39.4	24.8

由表 5 可知,去掉权重小于 0.2 的指标后,准确率相比利用全部指标减小了 7.8%,而去掉小于 0.3 的指标后,准确率减少了 22.4%,因此利用去掉小于 0.2 的指标进行预测是比较合理的,这样最终剩余 5 个指标,准确率达到 39.4%。由于用户分享信息的随意性和复杂性,利用改进 PA 模型的最

终准确率也不是很高,但相较于随机不到 3% 预测准确率,已有了较大提高。

5 结束语

在基于节点属性的信息预测中,从信息传播的两个主体用户和信息属性中选择了 11 个指标,利用改进的 PA 算法模型进行用户信息分享预测,通过分析模型赋予指标的权重,比较了不同指标对预测的重要程度。为了提高预测程序的运行效率,最终选择了 5 个指标,预测的准确率达到 39.4%,相较于随机预测不到 3% 的准确率有了较大提高。

参考文献:

- [1] 邢立双. 突发事件中微博信息的传播与管理[J]. 网络传播, 2012, 25(3): 28-31.
- [2] 樊鹏翼,王晖,姜志宏. 微博网络测量研究[J]. 计算机研究与发展, 2012, 49(4): 691-699.
- [3] EFRON M. Information search and retrieval in microblogs[J]. Journal of the American Society for Information Science and Technology, 2011, 62(6): 996-1008.
- [4] 刘鲁, 刘志明. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1): 1-4.
- [5] WALTHER J B, CARR C T, CHOI S S W. Interaction of interpersonal, peer, and media influence sources online[M]//PAPACHARISSI Z. A Networked Self: Identity, Community, and Culture on Social Network Sites. [S. l.]: Routledge, 2010: 246-267.
- [6] YU Hai-yuan, BRAUN P, YIDMM M A. High-quality binary protein interaction map of the yeast interactome network[J]. Science, 2008, 322(5898): 104-110.
- [7] 叶东海, 蒋国平, 宋玉蓉. 多局域世界复杂网络中的病毒传播研究[J]. 计算机工程, 2010, 36(23): 130-132.
- [8] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661.
- [9] CRAMMER K, DEKEL O, KESHET J, et al. Online passive-aggressive algorithms[J]. Journal of Machine Learning Research, 2006, 7: 551-585.
- [10] KWAK H, LEE C, PARK H. What is Twitter, a social network or a news media? [C]//Proc of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 591-600.
- [11] STERN D H, HERBRICH R, GRAEPEL T. Matchbox: large scale online Bayesian recommendations[C]//Proc of the 18th International Conference on World Wide Web. New York: ACM Press, 2009: 111-120.
- [12] SUH B, HONG Li-chan, PIROLI P. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network[C]//Proc of the 2nd IEEE International Conference on Social Computing. Washington DC: IEEE Computer Society, 2010: 177-184.

(上接第 50 页)

- [5] 钟秋波, 高超, 方宝富. 一种复杂环境下的仿人机器人路径规划算法[J]. 华中科技大学学报: 自然科学版, 2011, 39(S2): 192-203.
- [6] 赵娟平, 高宪文, 符秀辉, 等. 移动机器人路径规划的改进蚁群优化算法[J]. 控制理论与应用, 2011, 28(4): 458-461.
- [7] 程国建, 颜宇甲, 强新建, 等. 基于多 agent 的生态复杂适应系统建模和仿真[J]. 西安石油大学学报: 自然科学版, 2011, 26(2): 99-

103.

- [8] 宁爱兵, 马良, 熊小华. 基于复杂适应系统的蚂蚁群体智能研究[J]. 微计算机信息, 2008, 24(1): 265-267.
- [9] 李建会, 张江. 数字创世纪——人工生命的新科学[M]. 北京: 科学出版社, 2006: 118-128.
- [10] 孟志刚. 蚁群觅食仿真和动画的研究[D]. 长沙: 中南大学, 2011.