改进的 SIR 模型在微博信息传播中的应用

Application of Improved SIR Model on Information Diffusion in Microblog

杨 曦 刘艳华
Yang Xi Liu Yanhua
(厦门大学信息科学与技术学院，福建 厦门 361005)
(School of Information Science and Technology, Xiamen University, Fujian Xiamen 361005)

摘 要：研究微博中信息的传播规律对舆论预测与管控、市场营销等方面具有重要意义。当前的信息传播模型大多忽视了不同信息间、不同用户间的个体差异。为解决这一问题，本文选取了影响用户浏览和转发信息行为的特征，采用集成逻辑回归与 SVM 的二分类算法预测个体行为。预测多个用户对于同一信息的浏览与转发行为构成了本文中的信息传播模型。结果表明，该模型能较好地预测现实微博网络中的信息传播过程。

关键词：微博；浏览；转发；信息传播模型 DOI:10.13838/j.cnki.kjgc.2015.02.002
中图分类号：TP391 **文献标识码**：A **文章编号**：1671-4792(2015)02-0012-05

Abstract: Studying diffusion rules of information in microblog is significant for public opinion predicting and controlling, marketing and etc. The most current information diffusion models ignore the diversity of different information and different users. In order to solve the problem, this paper, by choosing the characteristics of users' browsing and retweeting behavior, uses binary-class classification algorithm based on logistic and SVM to predict individual behavior. In this paper, the information diffusion model is to predict users' browsing and retweeting behavior for given information. Simulation result shows that this model can better predict the process of information diffusion in real microblog network.

Keywords: Microblog; Browse; Retweet; Information Diffusion Model

0 引言

微博，微型博客(MicroBlog)的简称，是一种通过发布、关注、转发机制分享、获取、传播信息的实时广播式社交媒体平台。微博中的信息传播具备一传多的几何级增长特点；与传统媒体渠道相比，信息的传播速度、广度和效率都得到了极大提高，现今微博已经成为消息扩散和舆论传播最重要的平台之一。因此，通过构建信息传播的数学模型来定量分析信息在微博网络中的传播规律对舆论预测与管控、市场营销等方面具有重要的理论价值与现实意义。

近年来，随着在线社交网络的飞速发展，相关信息传播研究已逐渐成为国内外学者关注的热点。现有的在线社交网络信息传播模型可划分为三大类[1]：第一类，基于传播过程的模型，描述了用户对信息的接受状态与状态变化。文献[2]基于 SIS 模型，构建了博客网络中信息级联传播模型；文献[3]建立了基于 SIR 模型的在线社交网络信息传播模型；文献[4]将微博的信息交流过程分为信息发布、信息接收、信息加工、信息传播四个阶段，并提出了竞争窗口模型。第二类，基于用户影响力的模型，通过节点和节点间

的影响力预测信息的传播方向和传播概率。文献[5]提出了一种基于对节点影响力的评估预测信息传播趋势的模型;文献[6]提出了一个通过用户个体的多种特性评估用户影响力的多阈值信息传播模型。第三类,基于转发因素的模型,将微博特征与其被转发次数进行统计分析,建立模型预测一条给定微博的转发总数。文献[7]分析了是否包含 URL 等微博特征对转发行为的影响;文献[8]分析了 Twitter 的转发行为如何受到用户博文和时间因素的影响。

第一类模型未考虑不同用户间的个体差异,第二类未考虑不同微博间的特征差异,第三类缺少信息的整体传播过程。本文在三者之间找到结合点,将不同用户间的个体差异和不同微博间的特征差异融入信息的整体传播过程中,建立了改进的 SIR 信息传播模型,并通过真实的数据进行仿真分析。

1 微博网络信息传播规则

以新浪微博为例,一个用户发布信息后,该信息出现在此用户的每个"粉丝"(跟随者)的微博主页面,每个粉丝以一定的概率浏览该信息,然后根据自己对该信息的感兴趣程度,选择是否分享转发该信息。

2 经典 SIR 模型在微博网络信息传播中的应用

给定微博网络中一个子网络,子网络中的用户分为三类:未感染用户 S、感染用户 I 和免疫用户 R。给定一条信息:感染用户表示此用户传播,即发布或转发了该信息;免疫用户表示此用户已经浏览过该信息,但是选择不转发;未感染用户表示此用户没有浏览过该信息,但有可能之后会浏览此信息并选择是否转发,即由概率转变为感染节点或免疫节点。定义以下传播规则[3]:

(1) 感染用户只能将信息传播给其粉丝。

(2) 未感染用户一定的浏览概率 p_b 在浏览感染用户传播的信息后,依据其对信息所持态度,以转发概率 p_r 转化为感染用户,或者以 $1-P_r$ 的免疫概率转化为免疫用户。

(3) 浏览概率 p_b 和转发概率 p_r 为常数。

3 改进 SIR 模型在微博网络信息传播中的应用

如前所述,经典 SIR 模型应用于微博网络信息传播中时,每个用户的浏览概率和转发概率都一样,为常数。但现实情况是,由于个人微博使用时长和个人兴趣偏好等因素的不同,不同用户对同一信息的浏览概率和转发概率极有可能是不同的;由于微博内容与个人兴趣的契合度,同一用户对于不同信息的转发概率也极有可能是不同的。所以在模型中将浏览概率和免疫概率设为常数是不合理的,因此本文针对这一问题提出了在微博网络信息传播中应用的改进的 SIR 模型。该模型建立在微博子网络 $G=<U,E>$ 上,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,用户只转发来自其关注用户的消息。

3.1 个人消息浏览和转发预测模型

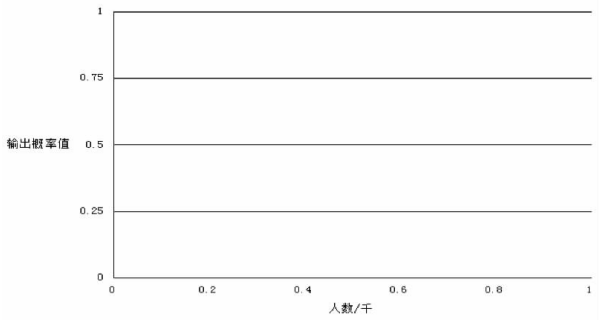
对消息是否浏览和是否转发的预测问题是典型的二分类问题;在消息浏览、转发预测问题中,影响浏览、转发的属性特征与转发行为呈现出线性关系[9];在给定用户网络、历史转发消息集合的情况下,可以应用适合线性情况的分类算法对网络中任一用户浏览和转发任一微博的概率和结果进行预测,常用的算法有逻辑回归算法和 SVM 算法等。但传统的逻辑回归方法一般以 0.5 为分界点,对于处在 0.5 附近的两个区间的样本点的误判风险较大;而传统的 SVM 算法需要占用较大的内存空间。所以本文采用集成逻辑回归与 SVM 的二分类算法[10],它综合了 SVM 和逻辑回归这两种算法的优点,减弱了这两种算法的缺点对计算过程和结果的影响。因为除了特征集合不同外,浏览行为预测算法和转发行为预测算法的步骤完全相同,故仅详述浏览行为的预测算法步骤。具体应用规则如下:

(1) 基于训练集,计算权值向量 α,即:

$$p_{b_u}(y_u=1\mid x)=\frac{1}{1+\exp(-\alpha(1+E_u(r,G)))} \quad (1)$$

式(1)中，$E_u(r,G)$为影响用户 u 浏览行为的特征集合；y_u表示用户 u 的浏览行为，取值为 0 或 1，为 1 表示用户 u 对消息 r 已进行浏览，反之亦然；采用极大似然法估计权值向量 α 。

(2) 将可能的逻辑回归输出概率值划分为四个等大小的连续的区间 I_1,I_2,I_3,I_4，如图一所示。



图一 概率区间划分示意图

(3) 计算训练集的逻辑回归输出概率值，并依据输出概率值在已划分的 I_1,I_2,I_3,I_4 这四个概率区间的对应情况将训练集分为四个子集 G_1,G_2,G_3,G_4，使用 SVM 算法计算子集 G_2 和 G_3 的输出概率值。

(4) 分别计算逻辑回归和 SVM 算法在子集 G_2 和 G_3 上的分类正确率，记为 f_i^L,f_i^S (i=2,3) 。

(5) 计算测试集的逻辑回归输出概率值，输出概率值为 $\hat{y}_i^L$ (i=1,2,3,4) ，并依据输出概率值在已划分的 I_1,I_2,I_3,I_4 这四个概率区间的对应情况将测试集分为四个子集 $\hat{G}_1,\hat{G}_2,\hat{G}_3,\hat{G}_4$。

(6) 子集 $\hat{G}_1,\hat{G}_4$ 直接采用逻辑回归输出概率值。在子集 $\hat{G}_2,\hat{G}_3$ 上，如果 $f_i^L \geq f_i^S$ 时，选择逻辑回归输出概率值 $\hat{y}_i^L$；反之，则再使用 SVM 算法计算并选择其结果作为输出概率值 $\hat{y}_i^S$。

(7) 若输出概率值大于 0.5，则判断用户已浏览过该信息；若输出概率值小于 0.5，则判断用户未浏览过该信息；若输出概率值等于 0.5，则进行等概率的随机判断分类。

3.2 特征选取

通过对微博用户行为特点的分析，影响浏览行为的因素主要有：用户活跃度、用户接收给定信息的总次数和给定时间内用户接收的所有信息总数[11]。

(1) 用户活跃度

因为微博自身的字数限制，阅读不同微博所用时长可视为相同，以便简化模型。用户活跃度越高，则相同时间内浏览微博数量就越多，可能浏览到某条信息的概率越大；反之亦然。但用户活跃度无法直接统计，故通过用户微博数进行表征。

(2) 用户接收给定信息的总次数

某一用户关注的人中可能有多人对同一信息进行了发布或转发，这会影响到该用户浏览到此信息的概率。

(3) 给定时间内用户接收的所有信息总数

使用微博的时间有限，接收到的信息也越多，错过了某一信息的概率越大。一般而言，关注的用户越多，相同时间内接收到的信息也越多。

综上所述，选取了 3 个影响浏览行为的数值化特征：用户微博数、关注用户对给定信息的已有转发次数、用户关注数。

关于影响微博转发行为的因素有不少研究[12]，可分为发布者的社会影响力 A、微博文本属性 B、用户个人属性 C、微博内容与用户兴趣 D 这四大类。具体的数值化特征如表一所示。

3.3 改进的 SIR 信息传播模型

通过分析和选取影响用户浏览、转发行为的特征，利用 3.1 所述的步骤可以分别算出个体用户的浏览和转发概率。多个用户浏览后的转发行为形成了信息的整体传播。本文将信息的整体传播过程与

表一　影响微博用户转发行为特征列表

特征类别	特征序号	特征名称
A	1	发布者的粉丝数
	2	发布者是否为实名认证
	3	发布者平均每条微博被转发的次数
B	4	是否包含图片
	5	是否包含标签
	6	是否包含 URL
C	7	用户转发微博总数
	8	用户粉丝数
D	9	微博 r 与微博热点话题的 Jaccard 值
	10	微博 r 与用户 u 的 Jaccard 值

个体用户的浏览转发行为分析相结合，建立了改进的 SIR 信息传播模型。

在改进的 SIR 信息传播模型中，对于一条微博，定义未浏览的用户为未感染用户 S、浏览后转发的用户为感染用户 I、已浏览但未转发的用户为免疫用户 R。建立未感染用户集合 S(0)，S(1)，…，S(n)；感染用户集合 I(0)，I(1)，…，I(n)；免疫用户集合 R(0)，R(1)…，R(n)。具体过程如下：

(1) 数据预处理。选定子网络，获取目标网络的历史数据，子网络中的所有用户置为未感染用户 S，并加入集合 S(0) 中。

(2) 在初始时刻即时间步 1 时，微博进入目标网络，网络入口用户置为感染用户 I。并将该用户加入集合 I(1) 中。

(3) 在时间步 t，以 3.1 中所述的集成逻辑回归与 SVM 的算法计算预测同时属于集合 I(t－1) 里的用户的粉丝和集合 S(t－1) 里的用户的浏览概率 Pb_u 与浏览行为。若未浏览，则将用户加入集合 S(t)；若已浏览，计算预测用户的转发概率 Pb_r 与转发行为。若未转发，则将用户加入集合 R(t)；若已转发，则将用户加入集合 I(t)。

(4) t=t+1 重复步骤 (3) 直到所有用户的状态不再变化。

4　仿真与分析

本文使用了数据堂提供的新浪微博相关数据，数据堂是国家相关部门支持建设的科研数据资源共享的公益性服务平台。由于新浪微博的限制，最多只能获取到每个用户的 200 个关注人信息。本文在数据集中提取了具有关注关系的 1000 名活跃用户的个人信息以及这些用户一段时间内所发布的所有微博共 27142 条。

为了分析验证个人浏览转发行为预测模型的准确度，将数据集分为两部分，每部分数据包含 13571 条微博数据和 1000 条用户信息，分别用做训练集和测试集，结果如表二所示。

表二　个人浏览转发行为预测与实际结果对比

实际	预测	
	转发	非转发
转发	70.6%	29.4%
非转发	15.0%	85.0%

从表二可以看出，模型预测的准确度仍有相当提升空间，影响准确度的因素可能有特征的选取、训练集的有限性以及基于训练集的模型对未发生过的情况无法良好预测等。

为了分析验证本文改进的 SIR 模型在整体传播范围的准确性，仿真了 5 条传播范围相对较广的微博，预测与实际结果对比如表三所示。

表三　整体传播范围预测与实际结果对比 (单位：个人)

微博	预测	实际
1	42	49
2	37	30
3	36	45
4	40	49
5	33	23

由表三可以看出，影响整体模型的预测准确度的因素主要是个人浏览转发行为预测模型的准确度。

5　结束语

本文通过对现在研究不足的分析，建立了一种基于微博网络的、考虑了用户间不同与微博间差异

的改进型 SIR 信息传播模型。实验仿真结果表明，该模型可以较好地预测实际网络中用户的浏览转发行为与信息整体的扩散范围。由于特征的选取、训练集的有限性以及基于训练集的模型对未发生过的情况无法良好预测等等，存在着许多会影响模型预测准确性的因素，因此该模型还有很大的改善空间。本文有助于更深刻地信息传播模型研究，如何在动态网络中对信息传播行为进行建模将是今后的研究方向。

参考文献

[1]陈慧娟，郑啸，陈欣.微博网络信息传播研究综述[J].计算机应用研究，2014，31（02）：333-338.

[2]LESKOVECJ，McGLOHONM，FALOUTSOSC，et al.Cascadingbehavior inlargebloggraphspatternsandamodel[C].Procof theSociety of AppliedandIndustrial Mathematics：DataMining，2007.

[3]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（05）：60-66.

[4]WUMing，GUOJun，XIEJian-jun. Social media communicationmodel researchbasedonSina-weibo[C].Procof the6thInternational Conference on Intelligent Systems and Knowledge Engineering.Berlin：Springer-Verlag，2011.

[5]YANGJ，LESKOVECJ.Modelinginformationdiffusioninimplicit networks[C].Procof the10thIEEEInternationalConferenceonDataMining.WashingtonDC：IEEEComputerSociety，2010.

[6]Li Chao，LUOJun，HUANGJ，et al.Multi-layer networkfor influencepropagationover microblog [C].Procof PacificAsiaConferenceonIntelligenceandSecurityInformatics.Berlin：Springer-Verlag，2012.

[7]CHAM，HADDADI H，BENEVENUTOF，et al.Measuringuser influence intwitter：the million follower fallacy [C].Proc of the 4th International AAAI ConferenceonWeblogsandSocial Media，2010.

[8]YANGZi，GUOJing-yi，CAI Ke-ke，et al.Understandingretweeting behaviorsinsocial networks[C].Procof the19th ACM International Conferenceon Inform ationand Knowledge Management. NewYork：ACMPress，2010.

[9]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（04）：109-114.

[10]谢玲，刘琼荪.集成 Logistic 与 SVM 的二分类算法[J].计算机工程与应用，2011，47（29）：149-151.

[11]赵文兵.Web2.0 环境下在线社交网络信息传播仿真研究[D].南京：南京大学，2013.

[12]吴凯，季新生，刘彩霞.基于行为预测的微博网络信息传播建模 [J]. 计算机应用研究，2013，30（06）：1809-1812.

作者简介

杨曦（1990—），女，侗族，贵州凯里人，硕士在读，主要研究方向：社交网络中的信息传播；

刘艳华（1965—），男，汉族，山东诸城人，博士，助理教授，主要研究方向：移动通信、图像处理、社交网络中的信息传播。