基于转发传播过程的微博转发量预测

赵惠东,刘 刚,石 川,吴 斌
(北京邮电大学智能通信软件与多媒体北京市重点实验室,北京 100876)

摘 要: 微博已经成为日常生活中最流行的信息分享工具.转发是微博中信息传播的核心方法,所以转发量预测不仅是一个有趣的研究问题,也有较大的实际意义.然而,当前大部分研究只是把问题视为分类或回归问题,没有考虑转发的传播过程.本文中,我们提出一个符合转发传播过程的转发量预测模型.本文认为转发信息来自两方面:直接粉丝和间接粉丝,而粉丝带来的转发量由转发意愿和影响力决定.我们用历史行为和内容相关性来估算一名直接粉丝的转发意愿,并用他/她的影响力来估算通过他/她的间接粉丝的转发量.新浪微博上的实验表明我们的模型比其他已有的方法效果好.

关键词: 转发量预测; 转发意愿; 转发影响力
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)12-2989-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.12.025

Retweet Number Prediction Based on Retweet Propagation Process

ZHAO Hui-dong, LIU Gang, SHI Chuan, WU Bin
(*Beijing Key Lab of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijng* 100876, *China*)

Abstract: Micro-blog has become the most popular information sharing tool in our daily life. The retweet behavior is a main method of information propagation in micro-blog. So the retweet number prediction is an interesting research topic and has much practical significance. However, most of current researches only regard this problem as aclassification or regression problem, and they did not consider the retweet propagation process. Considering the retweet propagation process, we propose a retweet number prediction model BCI. In our model, we think retweet messages are from two parts, direct followers and indirect followers. The retweet number of followers is decided by their retweet intention and influence. We use behavior and content information to estimate retweet intention for a direct follower and use the influence to estimate the indirect followers' retweet number. Experimental results on Sina Weibo dataset show that our retweet number prediction model has much better performance than other well-established methods.

Key words: retweet number prediction; retweet intention; the influence on retweeting

1 引言

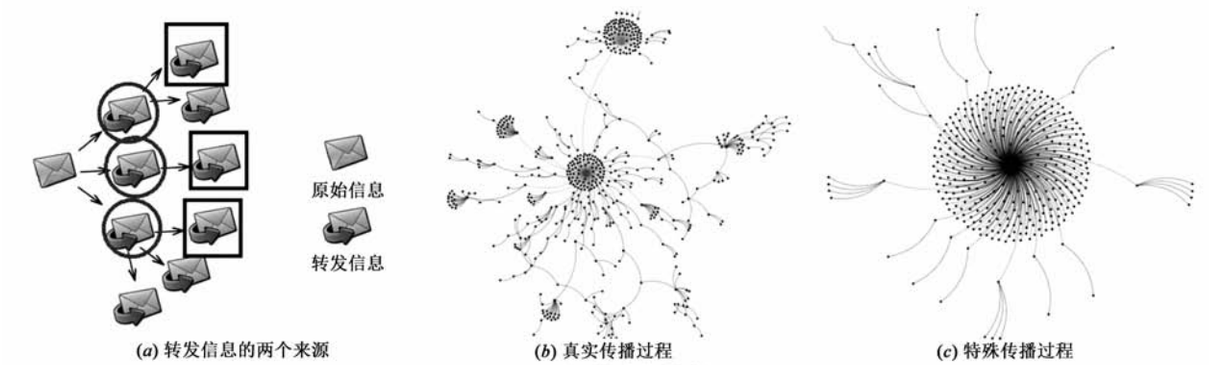
微博为人们提供了一个通过互联网和智能手机等设备就能够随时随地和朋友或陌生人分享、传播、获取信息的平台.这些年来微博服务越来越流行.例如,美国著名微博 Twitter 在 2012 年 3 月就已经拥有一千四百万活跃用户.而作为中国最有代表性的微博服务,新浪微博在 2013 年 3 月时已经拥有超过五千万的注册用户.

微博服务已经成为信息传播的重要媒体之一.在微博网络中,信息主要通过转发行为来传播.当用户发布一条微博,微博就会被推送给他/她的粉丝.当粉丝看到这条微博,他们将决定是否转发这条微博.如果转发,这条微博就会再次推送给该粉丝的粉丝.通过这种方式,信息在微博网络中传播.转发量是指一条微博被转发了多少次.它是转发行为的重要衡量指标.转发量预测在真实世界中具有重大意义.例如,我们可以在开始时就阻止谣言的传播.

有很多关于微博网络中信息传播和转发行为的研究[1~3].其中大部分研究将此问题看作微博是否被转发的二分类问题.通过提取适当的特征和选择合适的分类器,这些方法都会得到一个转发预测模型.也有一些人认为这个问题是回归问题,但取得的结果一般.然而,

收稿日期:2015-02-03;修回日期:2015-08-04;责任编辑:梅志强
基金项目:国家重点基础研究发展计划(No.2013CB329606);国家自然科学基金(No.61375058,No.71231002);北京市高等教育青年英才项目

所有这些方法都忽略了对于转发行为来说很重要的转发传播过程. 通过分析传播过程,我们认为微博的转发主要有两个部分:来自用户直接粉丝的转发(图1(*a*)中的圆圈)和来自用户间接粉丝的转发(图1(*a*)中的方块). 粉丝的影响力对于预测来自间接粉丝的转发量很重要. 图1(*b*)展示了一个真实的转发传播过程. 其中来自用户间接粉丝的转发量可能会很大. 如果我们忽略了转发过程,就可能只关注用户自己的影响力,只能处理特殊的转发传播过程,例如图1(*c*)那样. 这样会大大简化问题的难度并导致错误的预测.



(*a*) 转发信息的两个来源 (*b*) 真实传播过程 (*c*) 特殊传播过程

图1 转发传播过程

在本文中,我们研究了转发量预测问题,并提出了合理利用多种信息的转发量预测模型. 使用的信息包括历史行为、内容相关性、粉丝的影响力三种. 我们利用行为和内容相关性来估计用户的直接粉丝是否愿意转发,即转发意愿,然后用直接粉丝的转发影响力来估计来自用户间接粉丝的转发量,合理整合转发意愿和转发影响力,通过模拟转发传播过程我们能预测转发量. 我们搜集了新浪微博数据作为数据集并得到了良好的实验结果.

2 相关工作

随着微博服务的高速发展,出现了越来越多与微博相关的研究热点. 最基本的研究点是网络结构和用户特征. 文献[4]从各种方面比较两种不同微博平台(新浪微博和Twitter)上的用户行为. 文献[5]研究了与政治相关的微博,并发现这些微博中的情感会影响它的转发量.

转发是微博的核心功能之一,保证了微博网络中的信息传播. 所以许多研究者把注意力放在转发行为上. 文献[6]用主成分分析的方法分析了那些高转发量的微博的特征,自动提取出那些可能被大规模转发的微博. 文献[7]提出用条件随机场来对转发模型进行建模. 他们研究了划分社会关系图的方法,构建出用来转发预测的网络关系. 文献[8]通过预测社会影响来回答"谁分享什么"的问题,提出了一种混合因子非负矩阵分解方法.

现在也有许多基于中文微博的转发量预测研究. 文献[9]提出一个动态预测用户转发模式的方法,发现了许多以前传统方法没有捕捉到的外生特性,这些外生特性也很重要. 文献[10]提出一个基于SVM(Support Vector Machine)算法的预测模型. 在新浪微博上取得了很好的效果,但他们提取了太过复杂的特征. 文献[11]将问题分成了2步. 他们先将微博基于潜在的转发量分成几类,再对每类分别做回归. 新浪微博上的实验得到了比传统的没有提取复杂特征的一阶段模型更好的效果.

3 转发量预测模型

3.1 问题定义

对于所有用户,我们有一个四元组 $U=(F,T,N_r,\boldsymbol{M}_r)$. F 表示用户的粉丝集合,其中 F_i 表示第 i 名粉丝. T 表示该用户已经发过的微博的集合,其中 T_j 表示第 j 条微博. 为了方便,在没有声明时下文的 i 表示粉丝,j 表示微博. N_r 表示一条微博的真实转发量,包括所有直接粉丝和间接粉丝带来的转发量,其中 $N_r(j)$ 是指第 j 条微博的转发量. 矩阵 $\boldsymbol{M}_r$ 定义为粉丝和微博的转发关系矩阵. $\boldsymbol{M}_r$ 中的每行表示一个粉丝,每列表示一条微博. $\boldsymbol{M}_r$ 的规模是 $|F|*|T|$. $\boldsymbol{M}_r$ 中的值如下:

$$\boldsymbol{M}_r[i,j]=\begin{cases}1, & F_i \text{ 转发了 } T_j\\ 0, & \text{否则}\end{cases} \tag{1}$$

同时,对于每一个粉丝 F_i,我们有一个二元组 $F_i=(E_i,R_i)$. F_i 发布的微博组成集合 E_i. 对于 E_i 中的每条微博,我们有其内容. R_i 表示所有被 F_i 转发的微博,注意这些微博可能不是来自于用户 U. 对于 R_i 中的第 k 条微博,我们有它的转发时间 $t_{R_i}(k)$ 和真实转发量 $N_{R_i}(k)$.

前文提到过,转发量包括两部分. 因为直接粉丝有更紧密的关系和更丰富的信息,模型预测直接粉丝的转发意愿. 同时,因为间接粉丝的信息太多而不好获

取,我们利用直接粉丝的转发影响力来估算来自间接粉丝的转发量. 最后,我们从这两方面来估算转发量. 转发量预测模型的核心函数如下.

$$N_p(j) = \sum_{i \in F} f(i,j) \times (1 + P_i) \tag{2}$$

$N_p(j)$表示对微博 T_j 的预测转发量,其中包括来自 F 中所有粉丝的转发量及通过 F_i 的间接粉丝的转发量. $f(i,j)$表示粉丝 F_i 对微博 T_j 的转发意愿. 它是属于0、1之间的概率. P_i 表示粉丝 F_i 的转发影响力,用于估算间接粉丝的转发量. $f(i,j)$和$(1+P_i)$的乘积是来自直接粉丝 F_i 和其间接粉丝的预测量总和. 所以对于一条微博 T_j,它的转发量就是所有乘积的总和.

所有主要符号定义见表1.

表1　所用的主要符号

符号	含义	符号	含义
F	用户的粉丝集合	T	用户的微博集合
N_r	微博的真实转发量	$\boldsymbol{M}_r$	粉丝和微博的转发关系矩阵
E_i	F_i 发布的微博集合	R_i	被 F_i 转发的微博集合
t_{R_i}	R_i 中微博的转发时间	N_{R_i}	R_i 中微博的真实转发量
N_p	微博的预测转发量	$f(i,j)$	粉丝 F_i 对微博 T_j 的转发意愿
P_i	粉丝 F_i 的转发影响力	$\boldsymbol{M}_p$	粉丝对微博的转发意愿矩阵
$\boldsymbol{M}_c$	粉丝的微博和用户的微博的内容相关性		

3.2　转发意愿的计算

本文从两方面信息估算$f(i,j)$的值,粉丝对该用户微博的过去转发行为和该粉丝发过的微博与要预测的微博的内容相关性. 不同的粉丝会表现出不同的行为习惯. 一些人喜欢转发而另一些不喜欢. 过去转发行为代表一个粉丝的转发习惯,是否喜欢转发. 内容相关性则表示一条微博是否和该粉丝的日常兴趣点相关. 一般来说,用户只对自己关心的领域的微博感兴趣. 如果一条微博属于该用户的兴趣领域,被转发的可能性更大.

3.2.1　过去转发行为

我们通过矩阵分解模型[12]来对过去转发行为建模. 矩阵分解的目的是补全矩阵中的空缺. 对于一条新微博,我们能预测所有粉丝对其的转发可能性.

该模型的核心观点是将转发关系矩阵 $\boldsymbol{M}_r$ 分解成两个更小的矩阵. 首先,我们基于数据集中粉丝和微博的关系构建转发关系矩阵 $\boldsymbol{M}_r$. 然后构建隐特征矩阵 $\boldsymbol{p}$ 和 $\boldsymbol{q}$,分别对应粉丝和微博. 矩阵 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 的规模为$|F| \times K$和$|T| \times K$. $\boldsymbol{p}_i$ 表示粉丝 F_i 的隐特征,$\boldsymbol{q}_j$ 表示微博 T_j 的隐特征. K 表示隐特征的个数. 转发矩阵 $\boldsymbol{M}_r$ 能分解成 $\boldsymbol{p}$ 和 $\boldsymbol{q}^{\mathrm{T}}$ 的乘积. 通过最小化损失函数 Eq. 3,我们能得到 $\boldsymbol{p}$ 和 $\boldsymbol{q}$.

$$\min_{\boldsymbol{q}^*,\boldsymbol{p}^*} \sum_{(i,j) \in K} (r_{ij} - \boldsymbol{p}_i \boldsymbol{q}_j^{\mathrm{T}})^2 + \lambda(\|\boldsymbol{p}_i\|^2 + \|\boldsymbol{q}_j\|^2) \tag{3}$$

本文中,我们采用随机梯度下降算法. 迭代函数如下.

$$\boldsymbol{p}_i \leftarrow \boldsymbol{p}_i + \gamma(e_{ij} \cdot \boldsymbol{q}_j - \lambda \cdot \boldsymbol{p}_i) \tag{4}$$

$$\boldsymbol{q}_j \leftarrow \boldsymbol{q}_j + \gamma(e_{ij} \cdot \boldsymbol{p}_i - \lambda \cdot \boldsymbol{q}_j) \tag{5}$$

其中 $e_{ij} \xleftarrow{\text{def}} r_{ij} - \boldsymbol{p}_i \boldsymbol{q}_j^{\mathrm{T}}$. γ 是学习速率,它能影响训练时间和结果的收敛性.

通过计算 $\boldsymbol{p}$ 和 $\boldsymbol{q}$,我们能得到基于历史行为的任意粉丝对任意微博的转发意愿,用 $\boldsymbol{M}_p[i,j]$表示.

$$\boldsymbol{M}_p[i,j] = \exp(-(\boldsymbol{p}_i \times \boldsymbol{q}_j^{\mathrm{T}} - 1)^2/\delta) \tag{6}$$

式(6)是一个确保 $\boldsymbol{M}_p[i,j]$在0、1之间的规则化函数. δ 的目的是防止 $\boldsymbol{M}_p[i,j]$太小. 本文中,经过试验 δ 取0.02.

3.2.2　内容相关性

转发行为是建立在浏览行为基础上的. 大部分用户对不能吸引他/她注意力的微博只会一扫而过. 当然也不会转发此条微博. 只有一个微博和他/她的兴趣点相近,用户才会关注它并转发.

我们构建矩阵 $\boldsymbol{M}_c$ 来描述内容相关性. $\boldsymbol{M}_c[i,j]$表示粉丝 F_i 发布的微博集合 E_i 与用户微博 T_j 的内容相关性. 本文采用词袋模型来计算相关性. 它忽略了词的出现顺序,只考虑出现次数.

首先,采用著名的中文分词工具 ansj_seg(https://github.com/NLPchina/ansj_seg)分词,再去掉常见但没意义的停用词,剩下的词组成词袋. W_i 和 W_j 分别表示粉丝的微博集 E_i 和用户微博 T_j 的词袋. 两个词袋间的内容相关性可以通过很多算法计算,比如余弦距离、海明距离等. 我们采用下面的函数计算 $\boldsymbol{M}_c[i,j]$.

$$\boldsymbol{M}_c[i,j] = \frac{|W_i \cap W_j|}{|W_i \cup W_j|} \tag{7}$$

$|W_i \cap W_j|$和$|W_i \cup W_j|$分别表示 W_i 和 W_j 中相同的词和所有的词的个数. 比值表示内容相关性. 内容相关性越大,粉丝 F_i 对微博 T_j 越感兴趣,也就越可能转发.

3.3　转发影响力的计算

除了来自直接粉丝的转发,来自间接粉丝的转发在转发行为中也很重要. 但因为信息总量的指数式增长,我们无法获得间接粉丝的所有信息. 而且还存在两跳粉丝、三跳粉丝及更多跳粉丝. 所以我们选择粉丝的转发影响力来衡量来自间接粉丝的转发量.

第 i 个粉丝的转发影响力 P_i 表示当粉丝 F_i 转发了该微博后,该微博继续被粉丝 F_i 的粉丝转发的能力. 因为影响力很难计算而且本文的重点在转发模型上,我们用平均转发量来衡量转发影响力. 很明显转发影响力和时间有关,所以直接用所有被粉丝 F_i 转发的微博的平均转发量作为粉丝 F_i 的影响力并不合适. 为了解

决这个问题,我们引入一个时间函数来保证时间的影响. 权重函数如下.

$$P_i = \sum_{k=0}^{|R_i|}\left(\frac{e^{-(t_R(k)-t)^2}}{\sum_{k=0}^{|R_i|}e^{-(t_R(k)-t)^2}} \times N_{R_i}(k)\right) \tag{8}$$

在上面的函数中,k 表示在过去被粉丝 F_i 转发过的第 k 条微博. t 表示预测时间. 通过权重函数,在近期被粉丝 F_i 转发的微博对转发量的贡献更大.

3.4 整体模型

计算 $\boldsymbol{M}_p$ 和 $\boldsymbol{M}_c$ 后,我们能通过下面的公式计算$f(i,j)$.

$$f(i,j) = \boldsymbol{\alpha}_i \times \boldsymbol{M}_p[i,j] + \boldsymbol{\beta}_i \times \boldsymbol{M}_c[i,j] \tag{9}$$

其中 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\beta}_i$ 表示两种信息的权重. 对于每个粉丝 F_i,这两个值是不同的,所以是个性化参数. 加入转发影响力 P_i 后,微博 T_j 的最终转发量预测公式如下.

$$N_p(j) = \sum_{i=0}^{|F|}(\boldsymbol{\alpha}_i \times \boldsymbol{M}_p[i,j] + \boldsymbol{\beta}_i \times \boldsymbol{M}_c[i,j]) \times (1 + P_i) \tag{10}$$

常用的损失函数有很多,如 0-1 损失函数、绝对值损失函数. 本文采用均方误差作为损失函数. 函数如下. 其中 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 表示由 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\beta}_i$ 组成的向量.

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{|T|}\sum_{j=0}^{|T|}(N_r(j) - N_p(j))^2 + \lambda(\|\boldsymbol{\alpha}\|^2 + \|\boldsymbol{\beta}\|^2) \tag{11}$$

接下来问题变成了一个带约束的最优化问题. 优化函数见式(12).

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta}) \quad \text{s.t. } \boldsymbol{\alpha}_i + \boldsymbol{\beta}_i = 1, i = 1,2,\cdots,|F| \tag{12}$$

带约束的最优化问题一般用惩罚函数的方法解决. 然而,我们模型中的约束只是简单的线性约束,所以约束可以通过用一个参数的变形来代替另一个参数的方法抵消掉. 最后优化公式变为

$$\min_{\boldsymbol{\alpha}} \frac{1}{|T|}\sum_{j=0}^{|T|}(N_r(j) - N_p(j))^2 + \lambda(\|\boldsymbol{\alpha}\|^2 + \|1 - \boldsymbol{\alpha}\|^2) \tag{13}$$

因为最优化函数是二次方程式,本文采用随机梯度下降算法. 每一个 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\beta}_i$ 的迭代公式如下.

$$\boldsymbol{\alpha}_i \leftarrow \boldsymbol{\alpha}_i + \gamma\left[\frac{2}{|T|}\sum_{j=0}^{|T|} r_{ij} \times (\boldsymbol{M}_p[i,j] - \boldsymbol{M}_c[i,j]) \times (1 + P_i) - 2\lambda\boldsymbol{\alpha}_i + 2\lambda\right] \tag{14}$$

$$\boldsymbol{\beta}_i \leftarrow 1 - \boldsymbol{\alpha}_i \tag{15}$$

其中 $r_{ij} \xleftarrow{\text{def}} N_r(j) - N_p(j)$,$\gamma$ 是和式(4)和式(5)中一样的学习速率.

详细算法见算法 1.

算法 1 转发量预测模型

输入:

$U = (F,T,N_r,\boldsymbol{M}_r), F_i = (E_i,R_i)$

输出:

N_p

1. 用 $\boldsymbol{M}_r$ 计算 $\boldsymbol{M}_p[i,j]$
2. 用 T 和所有 E_i 计算 $\boldsymbol{M}_c[i,j]$
3. 用 R_i 计算 P_i
4. 初始化 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$
5. 循环 未收敛
6. 用式(10)计算 N_p
7. 用式(14)更新 $\boldsymbol{\alpha}$
8. 用式(15)更新 $\boldsymbol{\beta}$
9. 结束循环
10. 返回 $N_p(j) = \sum_{i=0}^{|F|}(\boldsymbol{\alpha}_i \times \boldsymbol{M}_p[i,j] + \boldsymbol{\beta}_i \times \boldsymbol{M}_c[i,j]) \times (1 + P_i)$

4 实验

本节中,我们先介绍从新浪微博得到的数据集. 然后验证模型的有效性和个性化参数的效果. 最后做了一个实例研究.

4.1 数据集

我们从新浪微博获得数据集. 新浪微博是中国的最大微博服务之一. 而且它提供 API 给所有用户. 通过这些 API,我们能得到包括微博内容、时间、转发状况等所有信息. 在我们的数据集中,共有 9,535 个用户. 这些用户涵盖了转发量巨大的大 V 用户以及转发量很小的普通用户. 为了确保转发量的稳定性,我们删除最新一个月的微博,因为它们可能仍在被转发. 一共有 745,919 条微博和 326,180 个粉丝. 转发总量为 18,108,061 次.

为了验证微博的选取是否有代表性,对微博的转发量的分布进行分析. 转发量分布见图 2. 其中横坐标为一条微博的转发量,纵坐标为这个转发量的微博条数,坐标轴均为对数刻度. 从图 2 中可见,转发量从 0 到 100,000 以上均有覆盖,大部分微博的转发量较低,随着转发量增多,微博越来越少,符合长尾分布. 真实的微博转发情况也应是如此,大部分微博的转发量都很低,少数微博的转发量特别高,可见选取的微博还是有覆盖性的.

4.2 对比实验

本节中,我们通过和几种方法作对比来验证提出的方法. 我们选择了如下 4 种方法及 3 种模型变形来对比. 基本方法中所用的部分特征见表 2.

(1)多元线性回归(MLR)[13]是普通线性回归的一般化,考虑了多个独立变量的情况.

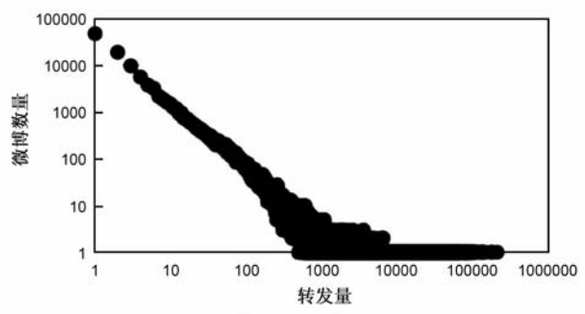


图2　转发量分布

(2)多重感知机(MP)[14]是一种带有前向结构的人工神经网络. 它能将输入向量映射到输出向量.

(3)M5P[15]模型是决策树和多元线性回归的结合. 每一个叶结点是一个线性回归模型,所以 M5P 能用于连续值的回归问题.

(4)两阶段模型(TP)[11]将转发量预测问题分成两个阶段. 第一阶段,他们将微博基于潜在的转发量分成几类. 在第二阶段,在每类中做回归.

(5)我们的模型(BCI)及模型变形. BCI 使用了两种信息计算转发意愿,所以我们通过只用一种信息的方式能得到两种变形. 模型 $BCI_{\backslash C}$ 只使用过去历史行为而模型 $BCI_{\backslash B}$ 只使用内容相关性. 模型 $BCI_{\backslash BC}$ 则不使用过去历史行为和内容相关性,直接用转发影响力来预测转发量. 对应的函数如下.

$$N_p(j) = \sum_{i=0}^{|F|} \boldsymbol{M}_p[i,j] \times (1 + P_i) \tag{16}$$

$$N_p(j) = \sum_{i=0}^{|F|} \boldsymbol{M}_c[i,j] \times (1 + P_i) \tag{17}$$

$$N_p(j) = \sum_{i=0}^{|F|} (1 + P_i) \tag{18}$$

表 2　基本方法所用部分特征

微博发布者的特征		微博的特征	
特征	含义	特征	含义
GD	微博发布者的性别	HI	是否含有 Hashtags
VR	是否为验证用户	HC	Hashtags 的数量
VT	验证类型	AI	是否含有@
ED	已经注册的天数	AC	@ 的数量
NL	昵称长度	HL	是否含有链接
FON	粉丝数	LC	链接的数量
FRN	关注数	TM	微博创建时间
…	…	…	…

4.3　衡量标准

我们用平均绝对误差 MAE 和相对绝对误差 RAE 来衡量结果. 在统计学中,它们是常用的标准之一,用来衡量预测值和真实值的差距. 其定义如下.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{p}_i - r_i| \tag{19}$$

$$\mathrm{RAE} = \sum_{i=1}^{n} |p_i - r_i| \Big/ \sum_{i=1}^{n} |r_i - r_m| \tag{20}$$

其中 p_i 是测试集中第 i 条微博的预测转发量,r_i 是真实转发量. r_m 是测试集的平均转发量. MAE 和 RAE 的值越小,模型越有效. 它表示预测值与真实值更接近.

习惯上,我们更喜欢用准确率[11]来衡量结果. 单一的值更容易给我们以直观印象. 但对于一个预测转发量问题,要得到绝对正确的值太过严苛. 所以我们定义一个范围来衡量预测结果. 定义的范围如下.

$$\left[N_r(j) - \frac{10^{\lceil \log_{10} N_r(j) \rceil} - 10^{\lfloor \log_{10} N_r(j) \rfloor}}{m},\ N_r(j) + \frac{10^{\lceil \log_{10} N_r(j) \rceil} - 10^{\lfloor \log_{10} N_r(j) \rfloor}}{m} \right] \tag{21}$$

m 是控制范围的参数. 如果 $N_p(j)$ 落在范围内,我们认为得到正确的结果,否则错误的. 然后我们用准确率来衡量预测结果. 准确率 Acc(accuracy)定义如下.

$$\mathrm{Acc} = \frac{\text{正确预测的数量}}{\text{所有预测数量}} \tag{22}$$

4.4　实验设置

实验共有 γ 和 λ 两个参数. 其中 γ 是学习速率. γ 的大小不仅会影响训练时间,也会影响结果的收敛性. γ 的值越大,学习速度越快,但可能结果无法收敛. γ 的值越小,学习速度越慢,结果收敛性更好. 一般都会把 γ 取的很小,在 0.001 这个量级. 本文的模型是对每一个用户 U 建立的,所以要计算多次模型. 根据尝试,γ 设置为 0.002,对于大部分用户数据已经可以收敛. 对于无法收敛的用户数据,将 γ 缩小,直到所用用户的数据都收敛.

λ 是正则化系数,目的是防止模型过拟合. 一般会将 λ 的值设置在 0.01 这个量级. 由于模型要计算多次,每一次都确定一次 λ 过于复杂,所以 λ 的参数实验建立在整体结果上. 对于每个用户,随机选取数据集中的 60% 作为训练集,剩下的作为测试集,进行参数 λ 的实验. 结果见图 3. 可见 λ 对实验结果有影响,但不是很明显. 最终将 λ 设置为 0.02.

4.5　有效性实验

首先,我们做有效性实验来验证模型的效果. 对于每个用户,随机选取数据集中的 60%、70%、80% 和 90% 作为训练集,剩下的作为测试集,采用 4.2 节提到的方法和 4.3 节提的衡量方法做实验. 表 3 中展示的是 MAE 和 RAE 的结果,准确率结果 Acc 在图 4 中.

从表 3 中,可以观察到提出的模型 BCI 在所有情况下都是最好的,都有相对明显的提升. 在 60% 和 70% 的情况下,我们的模型相比 MLR 提升超过了 100%. 在 4

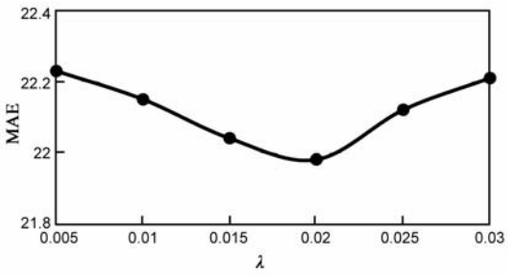


图3 参数λ的选取

种对比方法中,除了90%的情况TP均取得了最好的结果,同时MLR效果总是最差的. $BCI_{\backslash C}$的结果不理想,结果甚至比一些对比方法更糟. 但$BCI_{\backslash B}$的结果优于对比方法,接近模型BCI. 而$BCI_{\backslash BC}$结果很低,可见转发意愿的计算是有意义的,只用转发影响力的话结果会大很多. 我们可以推断出,用户是否转发微博更主要的取决于兴趣. 历史转发行为表示一种习惯,在大规模统计结果中有效果. 但对于一条微博,兴趣更重要.

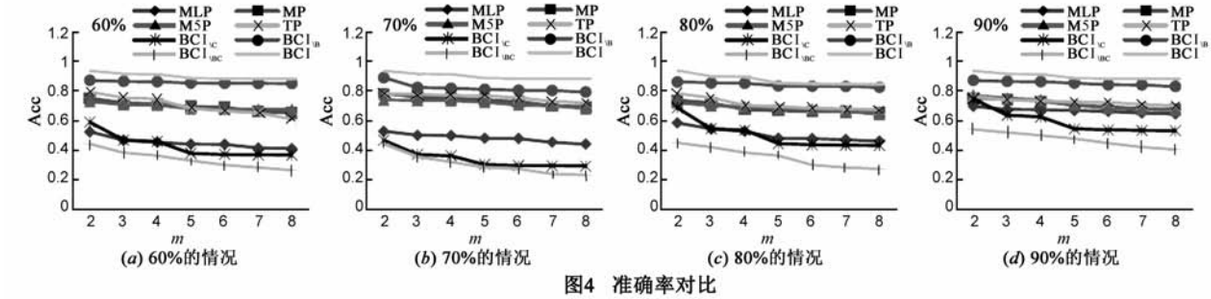
从图4中,我们能得到更多的信息. 横坐标随着式(20)中的m值的变化而变化. m值越大,准确率应该越低. 然而,因为转发量有很多是0,预测值在此时更容易正确,所以下降趋势不明显. 结果分成了3个层次,特别是在60%的情况下. 相对来讲,$BCI_{\backslash C}$和$BCI_{\backslash BC}$的下降趋势最明显. $BCI_{\backslash C}$使用了$\boldsymbol{M}_r$中的历史转发信息. 经过矩阵分解,$\boldsymbol{M}_r$中的0值将被填上. 所以$BCI_{\backslash C}$的结果相对来说离0比较远,趋势更明显. $BCI_{\backslash BC}$的下降趋势和$BCI_{\backslash C}$类似.

4.6 个性化参数的效果实验

我们的模型中,每一个粉丝都有其特殊的$\boldsymbol{\alpha}_i$和$\boldsymbol{\beta}_i$. 接下来,我们测试模型中$\boldsymbol{\alpha}_i$和$\boldsymbol{\beta}_i$的有效性. 这两个参数的目的是整合两种信息:历史转发信息和内容相关性信息,它们对于每个粉丝是不同的.

表3 有效性对比

Method	60%		70%		80%		90%	
	MAE	RAE	MAE	RAE	MAE	RAE	MAE	RAE
MLR	46.27	0.2439	45.37	0.2417	43.83	0.2337	37.75	0.1961
MP	34.83	0.1845	33.40	0.1777	34.43	0.1825	33.03	0.1771
M5P	36.54	0.1926	38.24	0.2037	36.45	0.1931	29.06	0.1557
TP	30.05	0.1592	31.82	0.1693	30.31	0.1607	32.81	0.1705
$BCI_{\backslash C}$	40.23	0.2111	51.51	0.2747	36.18	0.1918	33.87	0.1815
$BCI_{\backslash B}$	22.68	0.1195	24.87	0.1352	25.76	0.1353	25.04	0.1342
$BCI_{\backslash BC}$	58.41	0.3065	55.19	0.2989	57.64	0.3015	53.14	0.2903
BCI	**21.98**	**0.1159**	**22.97**	**0.1224**	**23.08**	**0.1223**	**22.11**	**0.1185**

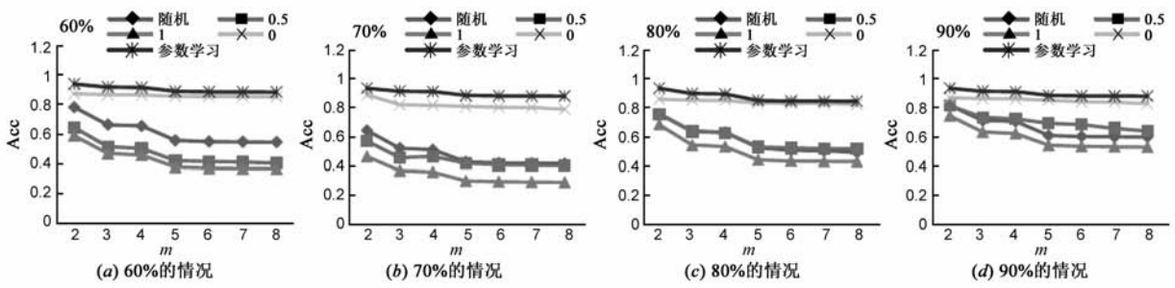


(a) 60%的情况 (b) 70%的情况 (c) 80%的情况 (d) 90%的情况

图4 准确率对比

表4 参数的效果

$\boldsymbol{\alpha}$	60%		70%		80%		90%	
	MAE	RAE	MAE	RAE	MAE	RAE	MAE	RAE
随机	30.42	0.1605	36.62	0.1952	28.63	0.1518	26.42	0.1416
0.5	35.74	0.1886	37.75	0.1961	28.60	0.1516	26.35	0.1412
1	40.23	0.2111	51.51	0.2747	36.18	0.1918	33.87	0.1815
0	22.68	0.1195	24.87	0.1352	25.76	0.1353	25.04	0.1342
参数学习	21.98	0.1159	22.97	0.1224	23.08	0.1223	22.11	0.1185



(a) 60%的情况　(b) 70%的情况　(c) 80%的情况　(d) 90%的情况

图5　随α_i变化准确率的变化

表4 和图5 中的 $\boldsymbol{\alpha}$ 值设置为随机、0.5、1、0 和参数学习的值. $\boldsymbol{\beta}$ 的值是 $1-\boldsymbol{\alpha}$. 表4 中的衡量标准还是 MAE 和 RAE,图5 中为准确率 Acc.

从表4,我们可以看出参数学习取得了最好的效果. $\boldsymbol{\alpha}=1$ 和 $\boldsymbol{\alpha}=0$ 就是 $\mathrm{BCI}_{\backslash C}$和 $\mathrm{BCI}_{\backslash B}$. $\mathrm{BCI}_{\backslash C}$和 $\mathrm{BCI}_{\backslash B}$的结果比我们的模型 BCI 的结果差. 这表明两种信息都有效果. 在大多数情况,$\boldsymbol{\alpha}$ = random 和 $\boldsymbol{\alpha}=0.5$ 的结果相似且在 $\mathrm{BCI}_{\backslash C}$和 $\mathrm{BCI}_{\backslash B}$的结果之间. 这种现象表明尽管两种信息都有用,但还是需要一个有效的整合方法来利用它们. 所以我们模型中的参数学习是有必要的.

在图5 中我们同样发现下降趋势. 同时,$\boldsymbol{\alpha}$ = random、$\boldsymbol{\alpha}=0.5$ 和 $\boldsymbol{\alpha}=1$ 的结果的下降趋势相似. 正相反,$\boldsymbol{\alpha}=0$ 的下降趋势不明显. 参数学习的结果曲线在所有曲线的上方,它的下降趋势和 $\boldsymbol{\alpha}=0$ 的下降趋势相似但也不一样. 可见,参数学习能有效地整合两种信息,相比一种信息有所提高.

4.7　实例研究

本节中,我们具体地展示个性化参数. 我们是对每一个用户建模. 每一个用户有很多粉丝,粉丝数从 1 到数百不等. 所以我们选择一个适当的用户作为例子,该用户有 94 个粉丝. 由于空间限制,表5 中只列出 5 对有代表性的 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\beta}_i$. 同时列出 $\boldsymbol{M}_p[i,j]$,$\boldsymbol{M}_c[i,j]$和 P_i 帮助理解. 然后我们还需要一个预测结果很好的微博. 我们找到一个真实转发量为 11 的微博,它的预测转发量为 12.

表5　$\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\beta}_i$ 的实例

No.	$\boldsymbol{\alpha}_i$	$\boldsymbol{M}_p[i,j]$	$\boldsymbol{\beta}_i$	$\boldsymbol{M}_c[i,j]$	P_i
1	0.1635	0.00005	0.8364	0.0017	0.3771
2	0.1939	0.3188	0.8060	0.0010	1.5458
3	0.5225	0.4357	0.4774	0.0027	0.3422
4	0.7448	0.8780	0.2551	0.0037	0.2735
5	0.8890	0.9999	0.1109	0.0026	0.0024
…	…	…	…	…	…

在表5 中,$\boldsymbol{\alpha}_i$ 的值递增. 这个结果反映了不同粉丝的区别. 通过分析数据,$\boldsymbol{M}_p[i,j]$ 的值越大,$\boldsymbol{\alpha}_i$ 的值越大. 一个转发过微博的粉丝有更大的 $\boldsymbol{\alpha}_i$. 如果一个粉丝转发过微博,未来中他/她更可能转发微博. 所以 $\boldsymbol{M}_p[i,j]$的值更大. 为了利用 $\boldsymbol{M}_p[i,j]$的信息,$\boldsymbol{\alpha}_i$ 就要更大. 因为 $\boldsymbol{M}_c[i,j]$比大部分 $\boldsymbol{M}_p[i,j]$都小,$\boldsymbol{\alpha}_i$ 的值主要受 $\boldsymbol{M}_p[i,j]$影响. 同时,$\boldsymbol{\alpha}_i$ 的值与 P_i 相互独立.

上面的结果表明,$\boldsymbol{M}_p[i,j]$更加占主导地位,然而利用 $\boldsymbol{M}_p[i,j]$的 $\mathrm{BCI}_{\backslash C}$的效果要比 $\mathrm{BCI}_{\backslash B}$差. 经过分析发现,$\mathrm{BCI}_{\backslash C}$的预测结果一般偏大,可见只依靠 $\boldsymbol{M}_p[i,j]$会使结果比较大,偏离真实值,经过较小的 $\boldsymbol{M}_c[i,j]$的修正,结果向真实值靠拢,但结果还是 $\boldsymbol{M}_p[i,j]$占主导. 因为数据集中大部分转发量较小,而 $\mathrm{BCI}_{\backslash B}$预测的结果与 $\mathrm{BCI}_{\backslash C}$恰好相反,预测结果偏小,预测值与真实值更加接近,结果比 $\mathrm{BCI}_{\backslash C}$好.

5　总结

转发是微博网络中信息传播的核心手段之一. 转发量是转发传播影响力的一种衡量方法,而且具有很大的实际意义. 我们提出一个基于粉丝转发意愿和影响力的模型. 用历史转发行为、内容相关性两种信息来计算转发意愿. 新浪微博数据集上的实验表明我们的模型效果优于一般的预测模型.

未来,我们可以继续提高模型的效果. 一方面,我们的模型可以扩展到使用更多种信息. 理论上,我们能用任意数量的矩阵来计算转发意愿. 另一方面可以利用更复杂的特征,比如微博的主题. 更多的使用那样的复杂特征,模型会得到更好的效果. 同时,也可以根据转发意愿来研究微博的实际转发路线,而不再只是计算转发量的结果.

参考文献

[1] Ma H, Qian W, Xia F, et al. Towards modeling popularity of microblogs[J]. Frontiers of Computer Science, 2013, 7(2): 171 – 184.

[2] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks[A]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management[C]. Toronto, Ontario, Canada: ACM,

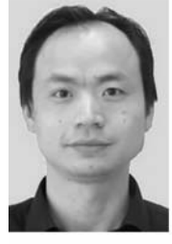
2010. 1633 – 1636.
[3] Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter[A]. Proceedings of the International AAAI Conference on Weblogs and Social Media [C]. Washington, USA: AAAI, 2010. 355 – 358.
[4] Gao Q, Abel F, Houben G J, et al. A comparative study of users' microblogging behavior on Sina Weibo and Twitter [A]. User Modeling, Adaptation, and Personalization[C]. Montreal, Canada: Springer, 2012. 88 – 101.
[5] Stieglitz S, Dang-Xuan L. Political communication and influence through microblogging-an empirical analysis of sentiment in twitter messages and retweet behavior[A]. System Science (HICSS), 2012 45th Hawaii International Conference on[C]. Hawaii: IEEE, 2012. 3500 – 3509.
[6] Morchid M, Dufour R, Bousquet P M, et al. Feature selection using principal component analysis for massive retweet detection[J]. Pattern Recognition Letters, 2014, 49: 33 – 39.
[7] Peng H K, Zhu J, Piao D, et al. Retweet modeling using conditional random fields[A]. Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on [C]. Vancouver, British Columbia, Canada: IEEE, 2011. 336 – 343.
[8] Cui P, Wang F, Liu S, et al. Who should share what?: item-level social influence prediction for users and posts ranking [A]. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Beijing, China: ACM, 2011. 185 – 194.
[9] Lu X, Yu Z, Guo B, et al. Modeling and predicting the repost behavior in SinaWeibo[A]. Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing[C]. Beijing, China: IEEE, 2013. 962 – 969.
[10] 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.
Y Li, H Yu, L Liu. Predict algorithm of micro-blog retweet scale based on svm[J]. Application Research of Computers, 2013, 30(9): 2594 – 2597. (in chinese)
[11] Liu G, Shi C, Chen Q, et al. A two-phase model for retweet number prediction[A]. Web-Age Information Management[C]. Macau, China: Springer International Publishing, 2014. 781 – 792.
[12] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30 – 37.
[13] Breiman L, Friedman J H. Predicting multivariate responses in multiple linear regression[J]. Journal of the Royal Statistical Society, 1997, 59(1): 3 – 54.
[14] Ruck D W, Rogers S K, Kabrisky M, et al. The multilayer perceptron as an approximation to a Bayes optimal di44scriminant function [J]. Neural Networks, IEEE Transactions on, 1990, 1(4): 296 – 298
[17] . Frank E, Wang Y, Inglis S, et al. Using model trees for classification[J]. Machine Learning, 1998, 32(1): 63 – 7

作者简介

赵惠东 男, 1990 年 11 月出生, 辽宁沈阳人, 2013 年在北京邮电大学获得学士学位, 现为北京邮电大学计算机学院硕士研究生, 研究方向为数据挖掘.
E-mail: zhaohuidong1121@foxmail.com

刘 刚 男, 1989 年 5 月出生, 辽宁沈阳人, 2012 年在北京邮电大学获得学士学位, 2015 年在北京邮电大学获得工学硕士学位, 研究方向为数据挖掘.

石 川 男, 1978 年 4 月出生, 湖北洪湖人, 教授、博士生导师、IEEE/ACM/CCF 会员. 2001 年在吉林大学获得学士学位, 2004 年在武汉大学获得硕士学位, 2007 年在中国科学院计算技术研究所获得博士学位. 2007 年加入北京邮电大学, 研究方向包括机器学习、数据挖掘和演化计算.



吴 斌 男, 1969 年 11 月出生, 湖南长沙人, 教授、博士生导师, 2002 年中国科学院计算技术研究所博士毕业, 现在北京邮电大学计算机学院工作, 主要从事复杂网络、数据挖掘、海量数据并行处理、可视分析、电信客户关系管理等方面的研究工作.