基于 SVM 的微博转发规模预测方法*

李英乐†，于洪涛，刘力雄
（国家数字交换系统工程技术研究中心，郑州 450002）

摘　要：为了评价微博的传播效果，在分析影响用户转发行为因素的基础上，提出了采用用户影响力、用户活跃度、兴趣相似度、微博内容重要性和用户亲密程度五项特征进行转发行为预测的 SVM 算法，以及基于该算法的转发规模预测算法。最后给出了传播规模预测的评价方法。针对新浪微博用户数据的实验表明，预测精度达到了 86.63%。

关键词：微博；转发行为；转发规模
中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1001-3695(2013)09-2594-04
doi:10.3969/j.issn.1001-3695.2013.09.008

Predict algorithm of micro-blog retweet scale based on SVM

LI Ying-le†, YU Hong-tao, LIU Li-xiong
(*National Digital Switching System Engineering & Technological R&D Center*, *Zhengzhou* 450002, *China*)

Abstract: Based on the analysis of the factors that affect retweet behavior, this paper proposed a predict SVM algorithm with five features: user influence, user activity, interest similarity, the importance of micro-blog content and users closeness. Furthermore, it proposed the predict algorithm of retweet scale on the basis of SVM, also, gave a method to evaluate the predict accuracy. The experiment with Sina micro-blog data shows a good result that the predict accuracy is up to 86.63%.

Key words: micro-blog; retweet behavior; retweet scale

0　引言

微博(micro-blog)是一种基于用户关系的信息分享、传播以及获取平台，用户可以通过 Web、手机等客户端组建个人社区，发布 140 个字左右的文字信息，实现即时分享。2006 年 3 月，互联网上出现了首个微博网站 Twitter。微博的原创性、时效性、草根性、随意性、碎片性等特点给互联网带来了一种全新的社交方式，微博网站及其注册用户的数量呈现出爆炸式的增长。根据中国互联网络信息中心(CNNIC)发布的报告显示，截止到2011 年6 月底，中国微博用户数量已经从年初的6 311 万增加到 1.95 亿，半年增幅高达 208.9%，网民的使用率也从 13.8% 增至 40.2%[1]。作为一种新兴的社会媒体，微博不仅是个人自我表达、人际交流的工具，还日渐发展成为政府、企业、组织用于信息发布、公关营销的手段。

从根本上说，微博仍然是一种传播媒体，最终目的都是向外界传递消息，获得最大的传播效果。而作为新兴的社会媒体，与传统媒体相比又有许多独特的性质。因此，研究如何在新媒体环境下，利用微博进行有效、高效的传播信息显得尤为必要。传播效果是传播学的一个概念，它是指传播活动尤其是报刊、广播、电视等大众传播媒介的活动对受传者和社会所产生的一切影响和结果的总体[2]。传播效果是一个抽象、定性的概念，目前尚没有一个公认的统一的标准来评价传播效果。不同的媒体采用不同的指标来评价其传播效果，如报纸用发行量、电视节目用收视率、电影用票房等。微博是通过转发行为实现了消息的持续传播，转发规模可以作为传播效果的一个重要指标。因此，分析用户的转发行为是预测转发规模的重要途径。

目前，对于社交网络中用户行为研究已经有了一定进展。清华大学的 Tan 等人[3]提出了一种 NTT-FGM(noise tolerant time-varying factor graph model)来模拟和预测社交网络中用户行为。该模型定义了行为偏好因子、朋友影响因子和自相关因子，分别计算这三者对用户行为的影响概率，将预测问题转换为一个条件概率问题来求解。张旸等人[4]针对 Twitter 用户的转发行为提出了一种基于特征加权的预测模型。该模型将 Twitter 数据标记为转发和非转发两类，然后提取了 11 个用户特征和 11 个文本特征，并按重要性进行加权，最后通过 SVM 来训练得到预测模型。该模型在预测转发行为的总体命中率达到了 85.9%。另外还通过信息增益方法对各个特征进行了重要性排名，“用户粉丝数”和“用户被提及数”居于前列，并得出用户特征和文本特征几乎同等重要的结论。加州大学洛杉矶分校的 Bandari 等人[5]提出了一种算法来预测新闻能否在 Twitter 上流行，或者在社交网站上引发热烈讨论。该算法仅仅根据文章的内容就能推断出文章被分享到 Twitter 后获得多少点击和转发。文中提出四个特征，即文章类别、客观程度、提及的人物和地名、文章来源，通过回归算法得到这四个特征与转发量之间的关系式。在预测时，文中将流行度按照转发量分为三个档次，即 1 ~ 20 次为低流行度、20 ~ 100 次为中流行度、100 ~ 2 400 次为高流行度，对这三个档次的预测准确度达到了 84%。

收稿日期：2012-12-20；**修回日期：**2013-02-15　　**基金项目：**国家“863”计划资助项目(2011AA010603)

作者简介：李英乐(1985-)，男(通信作者)，硕士研究生，主要研究方向为通信与信息系统(lyl7225@163.com)；于洪涛(1970-)，男，教授，主要研究方向为通信与信息系统；刘力雄(1974-)，男，副教授，主要研究方向为通信与信息系统.

可以看出,对于微博用户转发行为的研究已经有了一定的成果,但文献[2,3]只是针对用户是否转发某条消息开展的研究,只关注了局部的转发行为,而微博的传播过程是由很多个转发行为共同组成的,只有从总体上来审视考量转发规模才能更好地反映微博的传播效果;文献[4]虽然涉及到微博的转发规模,并且根据转发次数将转发规模较为粗糙地划分了三个档次,仅对这三个档次的预测取得了 84% 的精度,没有考虑更为精细的转发规模。因此,本文在前人研究的基础上,提出一种转发规模的预测方法。

1　问题描述

鉴于微博网络的特点,可以用有向无权图 $G=(V,E)$ 来表示,节点 $u_i \in V$ 表示网络中的第 i 个用户,边 $e_{ij} \in E$ 表示用户 i 和 j 的关注关系,其方向表示信息传播的方向,指向粉丝用户一侧。

假设用户 u_i 发布了一条微博消息 w,$y=f(u_i,u_j,w)$ 表示其粉丝 u_j 在看到 w 后采取的行为,当 $y=+1$ 表示转发,$y=-1$表示不转发。因此,转发行为预测问题可以描述为:在已知 u_i、u_j 和 w 的条件下,寻找一个目标函数 $y=f(u_i,u_j,w)$,$y \in \{-1,+1\}$,将 u_i、u_j 和 w 映射到 -1 和 $+1$ 两个类别中,这是一个典型的二分类问题。

对转发行为的预测是针对某一个特定的关注关系而言,是网络 G 中的一个局部预测问题,而对于转发规模的预测则是在网络 G 中的一个全局预测问题。

假设用户 u_0 发布了一条微博 w,沿着这条微博的转发路径可以得到一个如图 1 所示的转发树(retweet tree),该树的根节点为发布用户 u_0,叶子节点为非转发用户,中间节点为转发用户。预测转发规模就是计算转发树中所有除根节点和叶子节点以外的节点数量。

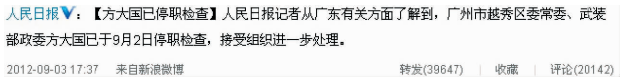
2　特征选取

微博网络是通过用户的关注行为建立起来的一个有向网络,用户之间是关注和被关注关系。被关注者发布的微博消息通过关注者的转发行为在网络中进行传播,因此用户的转发行为是消息得以持续传播的根本动力[6]。一次转发行为主要包含三个方面的因素,即发布用户、接收用户和微博消息。因此本文将从这三个方面提取特征。

2.1　发布用户的影响力

在这里发布用户不是特指微博的原创用户,而是在一次转发行为中相对于接收用户的另一方用户,可以是微博的原创用户,也可以是微博的转发用户。

发布用户的影响力大小会对转发行为产生影响。举例说明,用户名为"人民日报"认证微博于 2012 年 9 月 3 日 17:37 发布了一条微博(http://weibo.com/2803301701/yAbbS9EqJ):



该微博共被转发了 39 647 次,本文利用北京大学 PKUVIS 微博可视分析工具(http://vis.pku.edu.cn/weibova/weiboevents/)跟踪了这条微博的转发情况,直接转发该微博的共有5 272 个用户。转发的同时,也就相当于这些用户作为发布用户时发表了一条内容相同的微博。这条微博在不同用户中转发次数排名情况如表 1 所示(只列出了前十位用户)。

表 1　不同传播主体发布相同内容的被转发情况

排名	用户名	转发数	占比/%
1	姚晨	487	1.20
2	Vista 看天下	253	0.62
3	光远看经济	149	0.37
4	石扉客	74	0.18
5	但斌	67	0.16
6	许单单—互联网分析师	55	0.14
7	迟夙生律师	55	0.14
8	老徐时评	33	0.08
9	黎绮雯 GDTV	30	0.07
10	江南 Ricardo	28	0.07

从表 1 中可以看出,有"微博女王"之称的"姚晨"认证微博得到了最多的转发量,占到了总转发量的 1.2%。可见,发布用户不同所得到的转发量也不尽相同,发布用户的影响力[7]是影响转发行为的一个因素。

在这里用 PageRank 算法来评价用户的影响力[8]。PageRank算法是在搜索引擎中确定网页重要性的方法。该算法计算用户影响力的基本思想是:a)被重要用户关注得越多,该用户越重要;b)该用户关注的用户越少越重要。用户的 PageRank值可以用式(1)计算:

$$P(u_i)=(1-d)+d\sum_{j\in \mathrm{in}(u_i)}\frac{P(u_j)}{\mathrm{out}(u_j)} \tag{1}$$

其中:$P(u_i)$ 表示节点 u_i 的影响力;$P(u_j)$ 表示节点 u_j 的影响力;d 为 0~1 之间的一个阻尼系数,表示从一个给定用户转移到另一个随机用户的概率,在实际应用中常设为 0.85;$\mathrm{in}(u_i)$ 表示所有指向节点 u_i 的节点数(即 u_i 的关注用户数);$\mathrm{out}(u_j)$ 表示 u_j 所有指向的节点数(即 u_j 的粉丝用户数)。

2.2　接收用户的活跃度

接收用户就是通常所说的粉丝用户,是转发行为的直接产生者,不同的用户会有不同的行为特征。有些用户目的是向外发布信息来表达自己的观点和心情,其发表的多为原创微博,很少关注他人,也很少转发他人微博;而有些用户目的是获取信息,其偏爱浏览他人的微博,而不喜欢发表或转发微博;还有些用户为了吸引粉丝,其热衷于转发各类微博,即使是自己不感兴趣的内容,也不遗余力地进行转发。显然,那些转发微博比较活跃的用户更容易产生转发行为。用户发表微博的行为包含原创和转发两种类型,因此用户包含两个层次的活跃度:

a)发表活跃度(publish activity)。它表示用户在一段时间 t 内发表微博的频繁程度,用单位时间内微博的发表数量来表示:

$$PA=n/t \tag{2}$$

b)转发活跃度(retweet activity)。它表示一段时间 t 内(按天)用户转发微博占所有发表微博的比值,可以表示为

$$RA=\frac{\sum_{i\in t} r_i}{\sum_{i\in t} p_i} \tag{3}$$

其中:n 表示用户在时间 t 内发表的微博数量;r_i 表示用户第 i 天转发的微博数量;p_i 表示用户第 i 天发表的微博数量。当接收用户发表活跃度和转发活跃度达到一定程度时,才更容易发生转发行为。

2.3　接收用户的兴趣相似度

除了活跃度之外,微博消息内容能否引起接收用户的兴趣也是影响其转发的一个因素。通常接收用户只对某一方面或

某几方面的微博消息感兴趣,因此可以从一段时期内该用户发表的所有微博中提取出用户的兴趣空间,然后比较新微博与该用户兴趣空间之间的相似度,即可判断其对此微博的感兴趣程度。具体方法如下:

a)兴趣采集。收集一段时期内用户发表的所有 m 条微博,组成该用户的语句级兴趣空间 $I_s = \{s_i, 0 < i < m\}$,其中 s_i 表示第 i 条微博。

b)分词。对于英文微博,可以按照英文单词之间的空格直接进行分词;对于中文微博,可以采用中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS[9] 进行分词。得到用户的词语级兴趣空间 $I_w = \{w_i\}$,其中 w_i 表示第 i 个词语。

c)去除停用词(stop word)。停用词包含两类,一类是使用十分广泛,甚至过于频繁的单词,如"i""is""the""我""是"等;另一类是出现频率很高,但实际意义不大的词,如语气助词、副词、介词、连词等。可通过比对 CSDN(2010)提供的停用词列表来去除,该列表包含 900 多个停用词。去除停用词后得到用户的兴趣空间为 $I = \{w_i\}$。

d)对新微博 w 进行步骤 b)c)的处理,得到其特征空间 $J = \{w_j\}$。

e)计算 I 和 J 的相似度。由于 I 和 J 中都是特征词语,而 Jaccard 系数主要用于计算符号度量或布尔值度量的个体间的相似度[10],只关心个体间共同具有的特征是否一致。因此可以采用 Jaccard 相似系数(Jaccard coefficient)来计算其相似度:

$$S = \frac{I \cap J}{I \cup J} \tag{4}$$

2.4 微博消息内容的重要性

通常包含重要信息或者热门信息的微博更容易受到转发,微博内容是影响转发行为的又一因素。本文采用文本分类领域计算词语权重的 TF-IDF(term frequency inversed document frequency)算法来计算微博的重要性[11]。该算法认为:特定文档中的一个词语在本文档中出现的频率越高,说明它在该文档中越重要(TF);在其他文档中出现的范围越广,说明它在文档中的重要性越低(IDF)。微博 $w \in W$ 中词语 d 的 TF-IDF 值可用式(5)计算:

$$\mathrm{tf}(d) = n_w \times \log \frac{N}{n_d} \tag{5}$$

其中:n_w 表示 d 在 w 中出现的次数;N 表示微博集合 W 中包含的微博总数;n_d 表示微博集合 W 中包含词语 d 的微博数。

微博 w 的 TF-IDF 值就可以用该微博中所有词语的 TF-IDF 值之和来表示:

$$\mathrm{tf}(w) = \sum_j \mathrm{tf}(d_j) \tag{6}$$

2.5 用户亲密程度

与人类社会类似,微博网络的用户也有关系的亲密,越亲密的用户越容易得到认同,其发表的微博消息越容易得到转发。用户之间的亲密程度可以交互频度来表示,交互越频繁,说明用户的关系越亲密。

微博用户可以通过转发、评论、提及(@)和私信四个方面进行交互。私信的数据为用户隐私,通常无法获得。因此,用一定时期 t 内其他三个指标的均值来表示用户关系的亲密程度:

$$Q = \frac{1}{6}(r_{ij} + c_{ij} + a_{ij} + r_{ji} + c_{ji} + a_{ji}) \tag{7}$$

其中:r_{ij} 表示用户 i 转发 j 的微博消息数;c_{ij} 表示用户 i 评论 j 的微博消息数;a_{ij} 表示用户 i 提及用户 j 的次数。

3 转发预测

在第 2 章中提取了发布用户影响力、接收用户活跃度、接收用户兴趣相似度、微博消息内容重要性和用户亲密关系五个特征。本章将使用机器学习的分类算法来解决转发行为预测问题。

3.1 数据采集

目前,国内主要有新浪微博、搜狐微博、腾讯微博和网易微博四大微博平台。其中新浪微博的注册用户最多(已经接近 2 亿),影响较大。本文利用新浪微博 API 接口函数,从某一个用户出发,根据其关注列表逐层爬取用户信息数据,获得了 9 258条记录。然后爬取了 2012 年 8 月 1 日 ~2012 年 8 月 7 日这些用户发表的全部微博,共获得了 226 950 条微博数据。

3.2 转发行为预测

1)提取训练数据集

新浪微博用户的首页上以分页方式显示了其所有关注用户发表的微博消息。这些消息按照时间倒序排列,用户最先看到的是最晚发表的微博。实际上能够被用户浏览的仅仅是前几页的微博,还有一部分由于排序太靠后未被用户浏览到。因此,训练数据集应该从用户浏览到的微博中提取。

在爬取用户微博数据时,新浪 API 函数提供了一个标记"RetweetedStatus"来指示该微博是否是转发微博,因此转发微博集很容易获取。其余微博中包含两种类型:用户浏览到但未进行转发、用户未浏览到的微博。用户在浏览其时间轴时,转发了其中的一条微博,与其时间越接近的微博,被浏览到的概率就越大。可以选择与转发微博时间最接近的 2 条微博作为用户的非转发数据集。但用户通常是利用碎片时间来浏览微博,因此这 2 条微博与转发微博的时间间隔不能太大,本文选择小于 1 h。

最终从数据集中共提取到 126 950 条数据,包含 60 658 条转发记录、66 292 条非转发记录。然后提取前文提出的 5 个特征组成训练数据集。

2)分类算法选择

本文在数据挖掘软件 WEKA 平台上,采用 10 次交叉验证(10 cross validation)方式,验证常用分类算法的性能。分类结果用表 2 所示的混淆矩阵表示。

表 2 预测结果混淆矩阵

分类器	转发预测	实际转发	实际非转发
Naïve Bayes	预测转发	50 953	9 705
	预测非转发	8 992	57 300
KNN	预测转发	51 559	9 099
	预测非转发	8 393	57 899
C4.5	预测转发	55 199	5 459
	预测非转发	6 133	60 159
SVM	预测转发	57 019	3 639
	预测非转发	4 292	62 000

本文采用分类任务中常用的精度、召回率和 F_1 值来评价分类效果(表 3)。从表 3 中可以看出,SVM 分类器的分类精度达到 94%,召回率达到 93%,分类效果最好。因此本文选择 SVM 算法。

表 3　分类器性能比较

分类器	精度/%	召回率/%	F_1
Naïve Bayes	0.855	0.864	0.860
KNN	0.864	0.873	0.869
C4.5	0.917	0.907	0.912
SVM	0.945	0.935	0.940

3.3　转发次数预测

1)预测算法

在实际的微博网络中,用户的转发量由直接粉丝转发、间接粉丝转发和非粉丝用户转发三部分组成。其中,来自非粉丝用户的转发主要是用户在随便看看的时候产生的。本文统计了 1 000 条微博中这三部分的占比(图 2),从中发现,来自非粉丝用户的转发量只占了很小的比例,对转发规模的预测可以忽略不计。因此,只需要对直接粉丝和间接粉丝的转发行为进行预测。

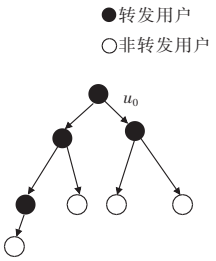


图 1　转发树

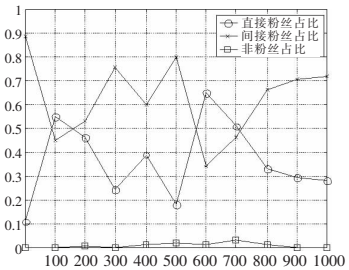


图 2　转发量来源占比

预测时,从微博的初始发布用户出发,沿着其粉丝路径对每一次转发行为进行预测,直到不再转发为止。假设用户对同一条微博只会转发一次,并且不会转发自己发布的微博。预测模型可以用如下的伪代码来表示:

```
输入:发布用户 U0,微博内容 T。
输出:转发次数 C。
初始化:U0 加入用户队列 Q;C = 0;
  while Q 不为空
    用户 Ui 从 Q 中出队;
    获取 Ui 粉丝列表 L(N)(N 为粉丝总数,且 N≥0);
    for i = 1 : N
      if 用户 L(i) 转发微博 T then
      用户入队 Q;
      C = C + 1;
      end if;
    end while;
```

本算法是一个迭代算法。为了防止迭代过度,本文跟踪了 400 个用户的 9 910 条原创微博,统计了从最初发布用户开始每一跳(跳:用户到最初发布用户的最短距离)用户的平均转发量占比。同时计算了在转发预测正确率为 0.93 时,每一跳的转发预测正确率。如图 3 所示,大部分的转发都集中在前 4 跳,而且 4 跳之后的预测准确率降到了 70% 以下,无法满足要求。因此,本算法只对 4 跳之内的用户进行预测。

2)预测结果评价

本文按照数量级来划分数量规模,并将数量规模定义为:

假设正整数 a、b、n,满足 $a<b$ 且 $10^a<n<10^b$,n 的数量规模为以 n 为中点、左右各扩展所在数量级长度一半的区域,即

$$S_n \in \left[n-\frac{10^b-10^a}{2}, n+\frac{10^b-10^a}{2}\right] \tag{8}$$

因此,当预测值满足式(9)时判定为预测正确。

$$n_p \in \left[n_r-\frac{10^{\lceil \log_{10}(n_r)\rceil}-10^{\lfloor \log_{10}(n_r)\rfloor}}{2}, n_r+\frac{10^{\lceil \log_{10}(n_r)\rceil}-10^{\lfloor \log_{10}(n_r)\rfloor}}{2}\right] \tag{9}$$

其中:n_p 表示预测值;n_r 表示实际值;$\lceil\cdot\rceil$ 表示向上取整;$\lfloor\cdot\rfloor$ 表示向下取整。

然后对 400 个用户的 9 910 条原创微博进行了转发规模预测,并计算了每个用户的预测准确率(图 4)。

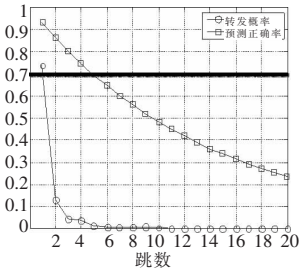


图 3　微博转发跳数统计

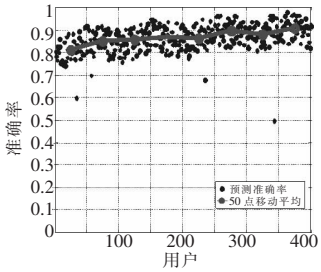


图 4　转发规模预测准确率

从预测结果可以看出,对于不同用户的不同微博,本文提出预测算法都能够达到 80% 以上的准确率,预测的总体准确率为 86.63%,能够很好地预测出微博的转发规模。

4　结束语

微博客的转发预测问题是研究微博信息传播的关键,而转发规模是衡量微博信息传播效果的重要指标。本文分析了转发行为影响因素,提出了发布用户影响力、接收用户活跃度、接收用户兴趣相似度、微博内容重要性和用户亲密程度五种特征,采用 SVM 分类算法实现对转发行为的预测。通过对新浪微博数据的实验表明对转发行为具有较好的预测性能,准确率达到了 93% 以上。然后在此基础上提出了一种转发规模预测算法,通过对粉丝用户的迭代计算实现对转发规模的预测。最后给出了相同数量规模的评价指标,在迭代次数为 4 时,预测的总体准确率达到了 86.63%。本文对转发次数的预测还不够准确,但转发规模已经能够对传播效果作出一个客观的评价,具有一定的实际意义。

参考文献:

[1] 中国互联网络信息中心. 第 28 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,2012.

[2] 郭庆光. 传播学教程[M]. 北京:中国人民大学出版社,1999.

[3] TAN Chen-hao,TANG Jie,SUN Ji-meng,*et al*. Social action tracking via noise to tolerant time-varying factor graphs[C]//Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press,2010:1049-1058.

[4] 张旸,路荣,杨青. 微博客中转发行为的预测研究[C]//全国信息检索会议. 2011.

[5] BANDARI R,ASUR S,HUBERMAN B. The pulse of news in social media:forecasting popularity[C]//Proc of AAAI Conference. 2012.

[6] KAPLAN A M,HAENLEIN M. The early bird catches the news:nine things you should know about micro-blogging[J]. Business Horizons,2011,54(2):105-113.

[7] CHA M,HADDADI H,BENEVENUTO F,*et al*. Measuring user influence in twitter:the million follower fallacy[C]//Proc of AAAI Conerence on Weblogs and Social Media. 2010.

[8] 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.

[9] 张华平,刘群. 中文自然语言处理开发平台[EB/OL]. (2002). http://www.nlp.org.cn.

[10] 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.

[11] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(Z1):167-170.