文章编号：1003-0077(2012)04-0109-06

微博客中转发行为的预测研究

张 旸，路 荣，杨 青

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：在微博客中，转发对信息的传播有着至关重要的影响，各种各样的信息正是通过转发得以在微博客上广泛且迅速的传播。另外在很多领域中，例如，市场营销、政治选举和热点提取等，也都需要深入探讨转发的各种特性。该文中，我们以 Twitter 为例，通过预测一条 tweet 是否会被转发，研究微博客中的转发行为。为解决这个问题，我们使用机器学习中的分类算法，并通过对微博上不同特征的重要性进行分析，提出了基于特征加权的预测模型。实验表明，我们的特征加权模型很好的解决了微博客中的转发预测问题，大约 86％的微博能被成功预测。

关键词：微博客；转发；特征加权模型

中图分类号：TP391　　**文献标识码**：A

Predicting Retweeting in Microblogs

ZHANG Yang, LU Rong, YANG Qing

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Retweeting is a key mechanism for information diffusion in Microbloging services such as Twitter. It is the mechanism of retweeting that leads to the fast and wide diffusion of information in Microblogs. In addition, research on the characteristics of retweeting is of vital importance for many different fields such as viral marketing, political campaigns, breaking news detection and so on. In this paper, taking Twitter as an example, we investigate the retweeting mechanism in Microblogs by predicting whether a tweet will be retweeted. We analyze the importance of different features and apply the classification method with weighted features. The experiments show that the proposed method can predict a major fraction of tweets (nearly 86％), out-performing previous researches.

Key words: twitter; retweeting; feature-weighted model

1 引言

微博客（简称微博[①]），是一个基于用户关系的信息分享、传播以及获取平台。用户可以经由SMS、即时通信、电邮、网站或第三方应用发布微博，输入最多 140 字的更新。以前的研究指出[1]，在信息时代，关注已经取代信息本身，成为稀缺资源。特别的，在微博客中，微博通过转发吸引大家的关注，如何准确的预测一条微博是否会被转发是信息传播中的一个基本问题，也是本文研究的重点。

解决这个问题有如下好处。第一，被转发的微博往往反映了用户感兴趣的话题，所以我们的研究可以应用到微博推荐中；第二，Cashmore 等[2]指出，社会化内容的分享（如微博的转发）不是随机的，而是决定于其本身的“传播性”。通过对微博转发行为的研究，我们能更好的理解微博中的“传播性”，并将之应用于市场营销或热点提取等；第三，研究证明，读者更倾向于分享那些能激发他们积极情绪（敬畏，愤怒）的社会化内容，相反的，那些激发读者消极情绪（悲伤）的内容通常不会被分享[3]。所以，通过预测微博的转发，我们可以进一步进行微博上的情感

① 本文提到的微博有两层含义：一是指微博平台如新浪微博，Twitter 等，二是指在平台上发布的状态，比如 Twitter 上的 tweets 等。

收稿日期：2011-09-15　**定稿日期**：2012-03-05

作者简介：张旸（1987—），男，博士研究生，主要研究方向为社交媒体，数据挖掘；路荣（1985—），男，博士，主要研究方向为社交媒体及信息检索；杨青（1970—），男，博士生导师，主要研究方向为社交媒体。

分析及舆情监控。

为了解决这个问题，基于微博中各个特征的重要性的差异，本文提出了特征加权的预测模型，其框架图如图1所示。具体来说，我们的方法分以下四步。第一，通过对大量数据的分析，我们得到微博上不同特征在被转发的微博和没被转发的微博上的差异，得出它们的区分度；第二，利用那些区分度较好的特征，配合有监督的机器学习方法，我们生成了基本的无加权的预测模型；第三，应用特征选择方法，分析哪些特征对转发有着更重要的影响，得到了不同特征的重要性排名；第四，在基本模型中，通过对不同重要性的特征赋予不同的权重，得到最终的特征加权预测模型。

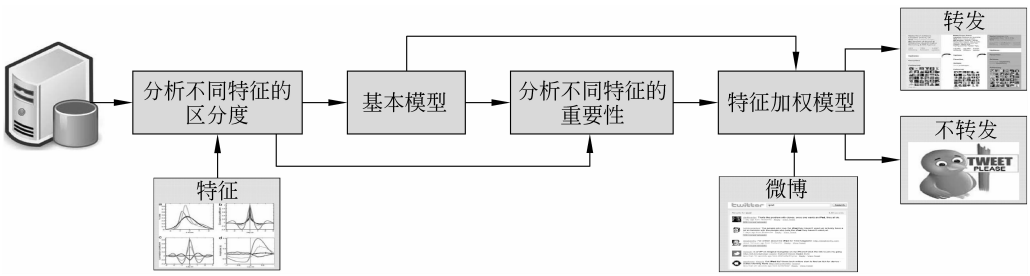

图1 特征加权模型的框架图

实验结果表明，本文提到的方法很好地解决了微博上转发行为的预测问题，能正确预测约86%的微博。

本文的组织结构如下，第二节介绍微博的相关工作；第三节描述了我们的数据集，并分析了不同特征的区分度；我们在第四节介绍特征加权的预测模型。第五节给出了实验的结果和分析；第六节是总结与展望。

2 相关工作

研究表明，微博在很多领域都发挥着重要作用，例如，政治选举、市场营销、突发灾难及日常生活中[4-7]。Tumasjan等[4]发现，微博能有效地反映现实社会中选民的倾向。Bollen等[5]发现从微博中收集的情感信息，与道琼斯指数紧密相关。Sakaki等[6]通过Twitter来迅速找出地震的震中，而Qu等[7]则通过研究2010年中国玉树地震前后，新浪微博①内容的变化，进一步指出微博在灾难面前所发挥的巨大且积极的作用。

当前的很多研究都集中在微博的各种特性和转发这一重要机能上[8-10]。Kwak等[8]分析了Twitter的拓扑特征，指出微博是一种新的信息分享媒介。Boyd等[9]对Twitter的转发功能retweet做了细致的分析，探讨人们如何转发，为什么转发以及转发什么的问题。Suh等[10]分析了影响转发的各种因素，发现文本特征和社会化关系都对转发有一定影响。

哪些用户更容易被转发？这个问题近来也吸引了越来越多研究者的关注[11-13]。例如，Cha等[11]通过粉丝的数量、以前被转发的数量等特征去衡量一个用户的影响力。Weng等[12]则利用用户之间的相互关注关系和在不同话题上的相似程度，去寻找有影响力的用户。Welch等[13]则改进了Weng等的做法，通过用户之间的转发关系代替关注关系。

近来，也有少数工作涉及了微博的转发预测问题[14-16]。Hong等[14]尝试通过分类去解决这个问题，而Zaman等[15]则引入了协同过滤的算法，可是他们的方法都不能取得令人满意的预测结果。Petrovic等[16]通过人工实验证明了这个问题的可行性，然后利用改进的passive-aggressive算法预测转发，可是也仅仅能正确预测46.6%的微博。

3 数据描述

本文中，我们以Twitter为例，研究微博上的转发预测问题。本节我们介绍了在Twitter上收集数据的方法，并统计分析了微博上的不同特征在转发微博和非转发微博上的区别，以期找到那些具有良好区分度的特征。

3.1 数据收集

通过Twitter上提供的Streaming API，从2011年3月11号到3月14号，我们抓取了Twitter上四天的公共微博数据，共收集了6 003 346条微博，

① http://weibo.com/.

即平均每天随机抓取了约 150 万条微博。为了便于处理，剔除那些非英文的微博，最终得到 4 242 405 条微博。因为是通过分类预测一条微博是否会被转发，我们还需要把收集到的数据标记为两类，被转发和未被转发。根据微博数据集中一条微博被转发的次数，我们得到了 556 402 条被转发的微博，3 686 003 条未被转发的微博。选取前三天的数据作为训练集，余下的作为测试集。

3.2　特征的区分度

Twitter 上的特征一般分为用户特征和文本特征，本节将分别分析它们在被转发和未被转发两类微博上的区分度。

3.2.1　用户特征

用户特征用来描述用户的行为，例如，发微博的频率，社交关系以及在 Twitter 上的影响力等。图 2 描述了用户的粉丝数及被列表数这两个特征在被转发的微博和未被转发微博上的累计分布函数(CDF)。图 2 直观的反映了这两个特征的区分度，由图 2 可以看出，用户的粉丝数和列表数都能很好的区分微博的转发与否。例如，被转发的那些微博的作者平均有 277 421 名粉丝，而未被转发的微博的作者只有 831 名。再如，在被转发的作者中，被列表超过 15 次的用户占 71%，而在未被转发的作者中，这一比例仅为 23%。当然，除了上述两个特征外，还有其他许多具有良好区分度的用户特征。例如，被转发的微博的作者平均关注 5 541 个用户，发布 10 040 条微博，而未被转发的作者平均只关注 722 个用户，发布 7 558 条微博。约有 11%的被转发微博的作者通过了认证，而只有 0.09%的未被转发作者通过认证。

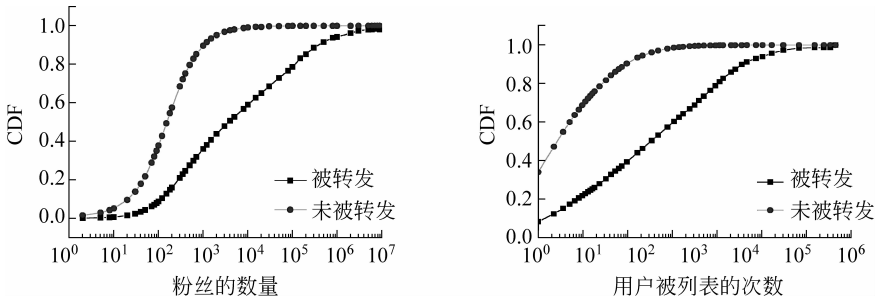

图 2　两个用户特征在转发和非转发上的分布

3.2.2　文本特征

文本特征描述了用户发布微博的方式，例如，是否包含 URL，是否包含 hashtag，一条微博的长度等。表 1 给出了一些文本特征在转发微博和非转发微博上的对比。可以看出，这些特征都能较好的分辨转发和非转发，特别是是否提及他人、是否为回复、以前是否被转发这三个特征。有一点值得注意，如果一条微博以前曾被转发，则它很难再被转发(在被转发中的比例远小于在未被转发中的比例)，这一现象从侧面反映了大部分的微博只被转发一次[8]。总之，通过本小节和上一小节的分析，我们发现，不管是用户特征还是文本特征，都能较好的区别被转发和未被转发微博。

表 1　文本特征在转发和非转发微博上的对比

	转发	非转发
包含 URL 的比例	0.22	0.17
包含 hashtag 的比例	0.31	0.15
提及他人的比例	0.24	0.54
回复的比例	0.08	0.31
微博长度(单词数)	14.4	11.3
以前被转发的比例	0.08	0.21

4　特征加权的预测模型

本节将介绍特征加权模型。我们通过分类的方法实现预测，在我们的方法中，每条微博通过一组数值来表示，其中每个数值对应于一个特征，利用前文提到的具有良好区分度的特征，并配合机器学习算法，我们训练得到预测模型。4.1 节首先介绍了无加权的基本预测模型，4.2 节分析了不同特征的重要性，最后 4.3 节在无加权模型的基础上，通过对不同重要性的特征给予额外的赏罚，得到最终的特征加权模型。

4.1 无加权基本模型

首先介绍无加权模型所使用的特征。在3.2节我们已经指出,微博上的特征分为用户和文本两大类,每类特征都能很好的区分被转发和未被转发微博。我们共选择了以下22个特征来训练模型。用户特征(11个):**用户的粉丝数、用户的关注数、用户的被列表数、注册的天数、发布的微博总数、用户名的长度、喜爱的微博的数目、是否被认证、用户每天发布的微博数、平均每条微博带来的粉丝数、平均每天拥有的粉丝数**。文本特征(11个):**是否包含URL、是否包含hashtag、是否提及他人、是否为回复、URL的数量、hashtag的数量、提及他人的次数、微博的长度、微博中的字母数、发布的时间、以前是否被转发**。

然后我们通过支持向量机(SVM)来训练无加权模型。SVM的目标是在N维空间中找到一个最优的超平面,能够分开训练集中的两类数据,并使它们有最大的间距。在本文中,我们使用基于径向基核函数的SVM模型,并借助开源工具LibSVM的帮助,训练得到基本的预测模型。

4.2 特征的重要性

Suh等[10]指出,在对微博转发与否的影响上,不同的特征发挥的作用大不相同。为了定量评测各个特征的重要性,我们使用了一种广泛使用的特征选择算法,信息增益算法(IG)。相对于其他的特征选择算法,例如,互信息,它往往更加简洁有效[17]。一个特征的信息增益值越大,说明该特征越重要。

通过本节的分析,我们发现,在用户特征中,用户的粉丝数和被列表数最为重要,而用户发布微博的频率和总数对转发的影响则相对较小。在文本特征中,相比是否包含URL或hashtag,一条微博是否为回复及是否提及他人更能影响该微博的转发,以前是否被转发这个特征则不像预想中的那么重要。我们将在实验结果中给出各个特征重要性的具体排名。

4.3 特征加权模型

4.2节指出不同特征对于一条微博是否会被转发有着显著不同的影响,我们在设计预测模型时就应该考虑不同特征的差异,赋予各个特征以不同的权重,以期获得更好的结果。为了得到特征加权模型,在无加权模型的基础上,我们为每个特征引入权重参数,该参数将作为一个额外的赏罚因子作用到无加权模型中(本文中即为SVM)。4.2节中,我们通过IG值定量的描述了各个特征的重要程度,自然的,我们也依据不同特征的IG值为不同的特征选择合适的权重参数,IG值越大,则该特征的权重参数就越大。我们首先计算出所有特征的平均IG值IGMEAN,然后依据式(1)为每个特征 f 选择权重:

$$weight(f) = \sqrt{\mathrm{IG}(f)/\mathrm{IG}_{\mathrm{MEAN}}} \qquad (1)$$

其中,$weight(f)$表示特征 f 的权重,$\mathrm{IG}(f)$表示特征 f 的信息增益值。可以看到,如果某特征对微博的转发与否有着重要影响,则它的IG值就会更高,对应的权重参数就大于1,该特征相应的在加权模型中就会发挥更大的作用。式(1)中的开平方是为了缓和该特征加权机制的影响。

5 实验结果与分析

5.1 评测指标

预测结果以表2中混淆矩阵的形式表示。为了评价预测模型的效果,我们选用信息检索的标准指标,包括准确率、查全率、总体命中率、F_1值。准确率是一类中被正确预测的微博占预测的属于该类的全部微博的比例,例如,被转发的微博这一类中,准确率P为a/(a+c)。查全率为一类中被正确预测的微博占该类实际的全部微博的比例,例如还是被转发的微博这一类中,查全率R为a/(a+b)。总体命中率是各个类中所有被正确预测的微博占总数的比例,表2中的命中率为(a+d)/(a+b+c+d)。F_1值折中考虑了准确率和查全率,表中被转发微博这一类的F_1值为2PR/(P+R)。

表2 以混淆矩阵形式表示的预测结果

		预测	
		被转发	未被转发
实际	被转发	a	b
	未被转发	c	d

5.2 基本模型的预测结果

表3显示了基本模型的预测结果。可以看到,通过基本模型,大约72%的被转发微博能被正确预测,而有将近92%的未被转发微博被正确预测。

表 3　基本模型的预测结果

		预测	
		被转发/%	未被转发/%
实际	被转发	72.1	27.9
	未被转发	8.2	91.8

5.3　特征的重要性分析

5.3.1　特征重要性排名

表 4 通过信息增益的方法，给出了各个特征的重要性排名。我们可以看到在表 4 的顶部(前 4)全是用户特征，例如，粉丝数、列表数。是否提及他人及是否为回复也同样重要，排在第 5 和第 7 位。值得注意的是，有一些特征尽管具有良好的区分度，但对一条微博是否转发却发挥很小的作用。例如，关于 URL 的那些文本特征，如前所述能很好的区分被转发和未被转发，但它们却全都位于该表的底部(第 20，第 21)，再如，关于 hashtag 的那些特征也仅仅位于表的中间。我们还发现，尽管像粉丝数等用户特征非常重要，也存在一些用户特征对转发与否的影响很小，例如，该用户的微博总数和发微博频率，分别位于第 17 位和第 22 位。这说明相较于用户的影响力，用户的活跃程度对他/她的微博的转发与否影响很小。表 4 中也有一些有趣的现象，例如，用户的关注数不像粉丝数那么重要，该微博以前是否被转发过也不像我们预期的那样重要。

表 4　不同特征的重要性排名

排名	特　　征
1	用户粉丝数
2	用户被列表数
3	平均每天积累的粉丝数
4	平均每条微博积累的粉丝数
5	该微博是否提及他人
6	该微博中提及他人的次数
7	该微博是否为回复
8	用户是否被认证
9	该微博中 hashtag 的数量
10	该微博是否包含 hashtag
11	用户注册的天数
12	该微博以前是否被转发
13	该微博发布的时间
14	该微博的单词数量
15	该微博的字母数量
16	用户的关注数
17	用户喜欢的微博的数量
18	用户总共发布的微博数量
19	用户名的长度
20	该微博是否包含 URL
21	该微博中 URL 的数量
22	该用户平均每天发布的微博数

5.3.2　用户特征和文本特征

我们将在本小节探讨用户特征和文本特征谁更重要的问题。表 5 列出了在特征的重要性排名中，前 10 名及前 22 名中用户特征和文本特征的数目。可以看出，不管是在前 10 名中还是前 22 名中，这两组特征的数目都是相等的，所以我们设想，用户特征和文本特征对转发行为的影响是相似的。为了证明我们的设想，在基本模型中，我们分别使用用户特征和文本特征去预测微博的转发。图 3 给出了预测结果的对比。

表 5　用户特征和文本特征的数目

	用户特征	文本特征
前 10	5	5
前 22	11	11

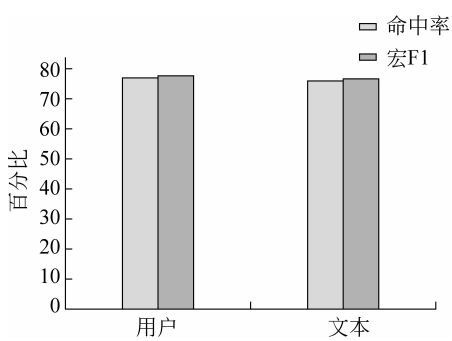

图 3　分别利用文本和用户特征得到的预测结果

从图 3 中可以看出，分别使用用户特征和文本特征，所得的预测结果非常相近(77.6%和78.1%)。这说明了在对微博转发的影响上，用户特征和文本特征几乎同等重要。这个结论在现实中有着重要的意义，例如，如果一个用户希望自己发布的微博能尽

可能多的被转发，他不必苦恼于自己的粉丝较少等很难在短时间内改变的用户特征，完全可以通过改进发布微博的方式来达成这一目标。

5.4 特征加权模型

图4对比了无加权模型与特征加权模型的预测结果。我们可以看出，通过引入特征加权模型，预测的效果得到进一步的提升。特征加权模型的总体命中率为85.9%，优于基本模型的81.9%。另外，因为加权是按照特征重要性的排名进行的，图4的结果反过来证明了特征重要性排名的可信度，这也为微博上信息传播的控制指明了方向。例如，如果我们想让一条微博传播的更远，我们应该专注于以下几个方面。第一，我们需要吸引尽可能多的粉丝；第二，我们尽量不要在该微博中提及他人；第三，该微博最好不是一条回复。而其他一些方面，例如，该微博是否包含URL或hashtag，或者该作者是否活跃等，都不那么重要。

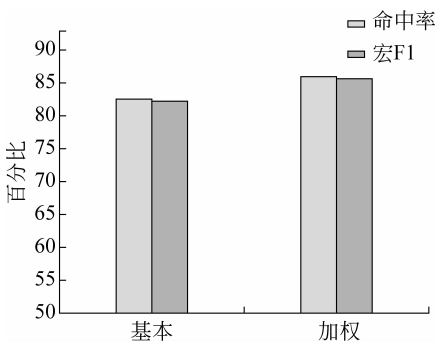

图4 特征加权模型与无加权模型的预测结果比较

6 总结与展望

微博的转发预测问题是理解信息在微博客上如何传播的关键，也是本文研究的重点。在本文中，考虑到微博中各个特征的重要性的显著差异，我们提出了特征加权的预测模型。通过特征加权机制，那些重要的特征就能在模型中得到进一步加强，反之则减弱。实验表明，我们的模型很好的解决了转发预测问题，大约有86%的微博被成功预测。

尽管如此，我们的工作仍存在一些需要改进的地方，这也是以后工作的方向。第一，我们忽视了各个特征之间的联系，例如，粉丝数和关注数就有很大的关联，这对我们的模型有一定程度的影响，将来的工作需要详细讨论特征之间的相互联系。第二，关于各个特征的权重的选取，除了本文用到的根据信息增益的相对关系，还可以尝试其他一些方法，例如，卡方分布等。

参考文献

[1] R. Lahan. The Economics of Attention[M]. University of Chicago Press, 2006.

[2] Pete Cashmore. YouTube: Why Do We Watch? [DB/OL]. http://editin. cnn. com/2009/TECH/12/17/cashmore. youtube/ index. html, 2010.

[3] J. Berger, K. L. Milkman. Social Transmission, Emotion, and the Virality of Online Content[R]. Wharton Research Paper, 2010.

[4] A. Tumasjan, T. O. Sprenger, P. G. Sandner, et al. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment[C]//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. ICWSM'10, 2010.

[5] J. Bollen, H. Mao, A. Pepe. Determining the public mood state by analysis of microblogging posts[C]//Proceedings of the Alife XII Conference MIT Press, 2010.

[6] T. Sakaki, M. Okazaki, Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors[C]//Proceedings of WWW'10, 2010.

[7] Y. Qu, C. Huang, P. Zhang, et al. Microblogging after a Major Disaster in China: A Case Study of the Yushu Earthquake[C]//Proceedings of CSCW2011, 2011.

[8] H. Kwak, C. Lee, H. Park, et al. What is Twitter, a Social Network or a News Media[C]//Proceedings of WWW'10, 2010.

[9] D. Boyd, S. Golder, G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter[C]//Proceedings of 43rd Hawaii International Conference on System Sciences, 2010.

[10] B. Suh, L. Hong, P. Pirolli, et al. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network[C]//Proceedings of IEEE 2nd International Conference on Social Computing (SocialCom), IEEE. 2010:177-184.

[11] M. Cha, H. Haddadi, F. Benevenuto, et al. Measuring User Influence in Twitter: The Million Follower Fallacy[C]//Proceedings of AAAI'10, 2010.

[12] J. Weng, E-P. Lim, J. Jiang, et al. Twitterrank: Finding topic-sensitive influential twitterers[C]//Proceedings of WSDM'10, 2010.

(下转第121页)