文章编号:1673-0062(2016)04-0100-06

基于用户特征的微博转发预测研究

仇学明,肖基毅*,陈 磊

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:研究微博用户转发行为,预测微博转发概率,确定影响微博转发概率的因素,在热点挖掘、产品营销、舆情监控、谣言控制等方面有重要的现实意义.本文介绍了影响微博转发的用户特征,其中比较典型的有用户影响力、粉丝平均标签数、粉丝活跃度等特征.通过粉丝数-关注数算法、用户标签数算法、粉丝活跃度算法等分析了它们与微博转发之间的关联关系,并确定各个属性的阈值,这些阈值对微博转发预测起到了至关重要的作用.

关键词:微博;用户特征;转发;预测

中图分类号:TP391.1 **文献标志码**:B

Research on Micro-blog Forward Prediction Based on User Characteristics

QIU Xue-ming, XIAO Ji-yi*, CHEN Lei

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: Studying micro-blog user forwarding behavior, micro-blog forward probability, and determining the factors that affect the forwarding probability of micro-blog have important practical significance in the aspects of hot spot mining, product marketing, public opinion monitoring, rumor control and so on. This paper describes the user characteristics affecting micro-blog forward, among which the user influence, the number of fans, the average number of tags, fans active degree and so on are more typical. The micro-blog's relationship with the number of fans, the number of users, the number of tags and the active degree of the fans are analyzed, mico-blog's and the threshold value of each attribute is determined. These thresholds play a crucial role in the forward prediction of micro-blog.

key words: micro-blog; user characteristics; forward; forecast

收稿日期:2016-10-10

基金项目:湖南省哲学社会科学基金资助项目(14YBA335);国家自然科学基金资助项目(61402220)

作者简介:仇学明(1994-),男,宁夏回族自治区平罗人,南华大学计算机科学与技术学院.主要研究方向:智能信息系统与知识管理. * 通讯作者.

0　引　言

随着互联网技术的不断发展,社交网络日益兴起,微博应运而生.微博平台基于用户之间的关注-粉丝关系来实现信息的共享和获取,是一种互动性强和传播速度快的社交媒体.在微博社区中,每个用户都可以被其他用户关注,这些用户被称为粉丝,同时用户也可以关注自己感兴趣的用户,这种行为被称为关注.

自微博出现以来,国内外许多学者从不同方面对用户转发行为进行了大量研究,主要分为两个研究方向:一是通过分析消息的传播路径来建立预测模型;二是通过分析用户结构和内容特征来建立预测模型.

Zaman 等[1]研究了用户影响力对转发的影响,其中包括微博作者是否为某领域的权威、微博作者的粉丝数量已经其粉丝的活跃度,重点分析了微博用户粉丝对传播速率和传播范围的影响,用户与其粉丝的互动强度越大,粉丝的活跃度越高,则该用户的微博被转发的可能性就越大.但是这些方法主要针对了个别主题进行预测,因而不具备普遍性.

相比传统的社交网络和媒体网络,微博社区中的用户关系更加多样,消息传播机制更加复杂,因此,影响用户转发行为的因素也就更多,研究难度相应的也会加大.

1　用户影响力对微博转发的预测

1.1　算法思想

定义 1 用户影响力:影响力是用一种为别人所乐于接受的方式,改变他人的思想和行动的能力[1-2].在社交网络中,用户影响力是指用户进行某种社交行为后,引起其他用户行为变化的能力.就拿微博而言,一个用户的影响力应该可以认为是他的一条博文引发其他用户评论、转发、收藏等的情况.我们用 In $f(u)$ 来表示微博用户影响力.

定义 2 关注:在微博社区中,关注表示不同用户之间的关联关系.如果用户 A 与用户 B 之间存在一条有向边<A,B>,且从 A 指向 B,称为用户 A 关注了用户 B,用户 A 是用户 B 的粉丝.指向用户的有向边称为入度,即用户的粉丝数,用 IND(u) 表示;指出用户的有向边称为出度,即用户的关注数,用 OUTD(u) 表示.

定义 3 转发:当某一用户发布了一条新微博后,其粉丝将这条新微博传递给自己粉丝的行为,称之为转发[3].转发机制是微博平台的重要机制,它使信息呈爆炸式扩散.

为了研究微博转发与用户影响力之间的关系,提出了一种粉丝数-关注数算法,即使用微博用户的粉丝数及关注数来衡量用户的影响力.该算法基于 PageRank 算法[4],主要是通过网页的链接数来评价一个网页,如果网页 A 与网页 B 之间有一个链接,且方向为 A 指向 B,认为网页 A 投给了网页 B 一票,指向网页 B 的网页越多,说明该网页重要性越高,价值越大.在微博平台中,用户的网络结构与万维网中的网页结构类似,将每位微博用户看作一个网络中的节点,将用户间的关注关系看作一条有向边.为了简单起见,把微博用户和用户关系映射到有向图 G 中,$G=<V,E>$,其中 V 是顶点集合,表示微博用户,E 是有向边集合,表示微博用户间的关注关系.$<u,v>$即表示用户 u 与用户 v 之间的关注关系,其中 $u,v\in V$.微博平台中用户网络结构图如图 1.

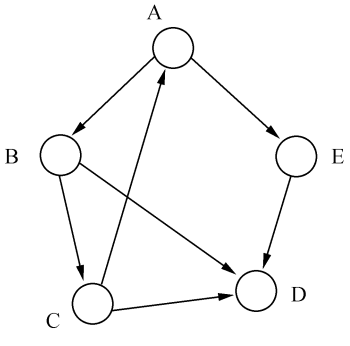

图 1　微博用户网络结构图

Fig.1　Micro-blog user network structure

其中,顶点 ABCDE 表示微博用户,顶点间的有向边表示用户之间的关注.图中,A 关注了 B 和 E,C 关注了 A,A 的入度为 1,出度为 2;B 关注了 C 和 D,A 关注了 B,B 的入度为 1,出度为 2;同样的,C 的入度为 1,出度为 2;D 的入度为 3,出度为 0,即没有关注任何人;E 的入度和出度均为 1.

问题描述:对于微博社区给定的用户集合 $U_n=\{u_1,u_2,u_3,\cdots u_n\}$,每位用户 u_i 都有自己的粉丝数和关注数,使用这两个特征来衡量用户的影响力,并把用户影响力作为一个指标,来研究其与用户的微博转发概率之间的关系.用户的影响力用下面的公示表示:

$$\mathrm{Inf}(u)=\ w_0\mathrm{IND}(u)\ +w_1\mathrm{OUTD}(u),u\in V\ (1)$$

其中,Inf(u) 表示用户的影响力,它代表了用户

的网络结构.IND(u)为用户的入度,即粉丝数,OUTD(u)为用户的出度,即关注数.w_0、w_1 分别为入度与出度的权重.一般情况下,微博用户的粉丝数对用户影响力较大,故 w_0 的取值远远大于的取值 w_1.

1.2 算法

对于粉丝数-关注数算法,分别用入度和出度两个量来描述微博社区中用户的影响力,入度即用户的粉丝数,出度即用户的关注数.将该算法作如下描述:初始化集合 C,集合 C 中存放的是用户的入度和出度,初始值均为 0;粉丝数-关注数算法如图 2.

Algorithm1:粉丝数-关注数算法

```
1:Input:IND(u),OUTD(u)
2:Output:用户影响力 Inf(u)的排序
3:for i=1 to N
4:do
5:将用户 u_i 的粉丝数存入集合 C 中的入度值中,关注数存入到出度值中
6:end for
7:for i=1 to N
8:do
9:对集合 C 按入度降序排列
10:如果入度相同,则按出度降序排列
11:end for
12:return 排序好的集合 C
```

图 2 粉丝数-关注数算法

Fig.2 Algorithm of fan number attention number

1.3 实验分析

实验基于这样的假设:用户的粉丝数越多,则该用户的人际关系越良好,受欢迎程度越高,其微博被转发的概率越大,即权威言论更容易被关注和传播.另外,当用户的粉丝数相同时,考虑用户的关注数,这是由于用户的关注数反映了用户在微博社区的活跃程度.一般情况下,用户关注的人数越多,那么该用户在相关领域中的活跃度越高,与他人互动的频率越大,也就越容易受到别人的关注.活跃度高的用户,其微博被转发的概率同样比较高.

为了简单起见,在本实验中使用粉丝数来表示用户的影响力,另外,提出了一个新的概念:用户的平均被转发率,它表示在某段时间内某一用户的微博转发情况.

$$P(u) = \sum_{i \in t} r_i \ / \ \sum_{i \in t} n_i \qquad (2)$$

其中,r_i 表示用户 u 第 i 天被转发的微博数;n_i 表示用户 u 第 i 天发表的微博总数.选取用户一天之内的微博发布及被转发情况来作为实验数据.

该实验的数据基于新浪微博用户的粉丝数、关注数、微博数及被转发数,因此,使用新浪微博开放平台 API 返回数据中的 followers_count、friends_count、statuses_count 及 reposts_count 字段值.获取了 5000 个用户的相关数据,将这些数据分为两个集合:在第一个集合中,用户按照影响力排序(基于入度和出度);在第二个集合中,将用户的平均被转发率进行排序,平均转发率为被转发数与总微博数的比值.就得到两个排序的数据集,数据集一体现了用户的影响力,数据集二则体现了用户的微博转发概率,将这两个集合的数据相对应,如果数据集一中的用户 A 的平均被转发率大于紧随其后的 B 的平均被转发率,则用户的影响力对微博转发概率是有作用的.该实验的部分数据见表 1.

表 1 按用户影响力排序结果

Table 1 Sort results by user influence

用户名	粉丝数	关注数	微博数	被转发数	被转发率
乖乖 ESSE	37	60	212	35	0.165 094 33
xX 笑笑	35	60	301	32	0.106 241 69
容 wxqbaby	30	58	174	22	0.126 002 29
东小小虾米鱼儿	30	50	187	17	0.090 618 33
糖菓吔眼淚 Henry	23	53	153	30	0.195 822 45
because 在心中	22	53	100	22	0.195 822 45
上官梓仪	22	40	80	13	0.162 500 00
戇憨憨的 Aka	21	46	535	18	0.033 594 624
天荃地草	20	37	61	10	0.162 337 662
Baiyqkqpibaid	15	50	85	13	0.152 582 159

在这 5 000 组数据中,有 3 230 组是满足假设的,正确率在 0.646.实验证明,用户的影响力对预测微博转发概率有很重要的作用.用户影响力与微博转发之间的关系,如图 3.

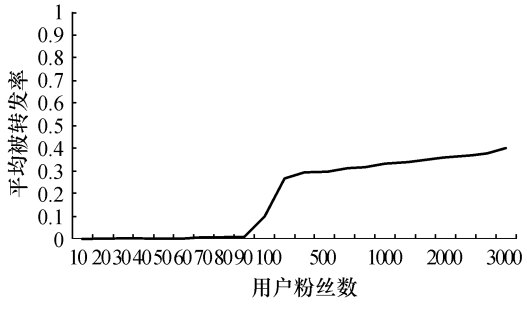

图 3　用户粉丝数与用户微博被转发率关系曲线

Fig.3　Relationship curve of the number of fans and the number of users micro-blog forward rate

由图 3 可知,当用户的粉丝数达到 100 左右时,粉丝数对微博转发的贡献明显提高.因此,把粉丝数 100 作为阈值,并将用户粉丝数作为一个特征属性加入到模型中.用户的粉丝数大于等于 100 时,该属性值设置为 1;反之,将其设置为 0.

2　粉丝平均标签数对微博转发的预测

2.1　算法思想

在国外微博社区 twitter 中,粉丝可以将自己关注的用户添加到一个特定的分组,并且可以对关注对象进行描述,这些信息都是公开的,所有用户都可以看到其他用户的分组描述信息[5-6].研究者可以通过收集用户粉丝的分组描述信息来研究该用户的角色,例如,如果一名用户的粉丝在对其进行描述时,大部分都含有“体育明星”这项描述,那么就可以认为该用户是体育领域内的专家.研究者在用户粉丝对他的描述中,选择出现频率较高的几个词汇作为对该用户的描述,并为其建立一个描述文档,当需要查找某个领域的话题人物时,就可以利用这些信息来进行检索[7-9].

新浪微博也提高了类似的功能,在新浪微博中,用户也可以为自己关注的对象添加分组和特征描述,但这些信息是不公开的,因此研究者无法获取这些信息.但新浪微博允许用户为自己添加标签,这个功能允许用户为自己添加最多 10 个关键词来对自己进行描述,用户可以通过标签来寻找特定领域的人从而进行关注行为.

定义 4 标签:在新浪微博中,标签是指添加描述自己职业、兴趣爱好等方面的词语,让更多的人找到你,让你找到更多同类.

在微博社区中,每个用户都有属于自己的用户标签.用户标签反映了用户关心的领域和感兴趣的内容.如果一个用户觉得某一领域或话题是自己研究的方向,或者是自己所擅长的部分,他就会添加该标签以便随时接受最新的内容.另外,如果用户仅仅是出于兴趣爱好,那他也可以添加自己感兴趣的标签来与好友讨论相关话题.该算法的基本思想是:添加了同一标签的用户具有相同的兴趣爱好或擅长领域,用户在选择关注对象的时候,一般会选择那些和自己标签相同的用户,进行转发的时候也会更加容易转发一些与自己标签内容相关的微博消息.因此,一个用户的标签内容及数量会影响该用户的转发行为.可以假设,标签数量越多的用户,其兴趣范围越广泛,转发他人微博的能力也越强.

2.2　算法

为了研究用户的标签数与微博转发之间的关系,采用最简单的方法,将用户的标签数与用户的转发率一一对应,观察二者之间是否存在正相关的关系.需要得到一个排序好的用户标签数集合,具体步骤如下:

1.初始化集合 C,集合 C 中存放的是用户的标签数量,初始值均为 0;

2.将第一个用户的标签数存入集合 C 中;

3.重复步骤 2,直到遍历数据集中的所有用户;

4.对集合 C 按标签数量降序排列;

5.返回.

用户标签数算法如图 4.

Algorithm2:用户标签数算法

```
1:Input:用户 U_i 的标签数 L_i
2:Output:用户标签数排序集合 C
3:for i=1 to N
4:do
5:将用户 U_i 的标签数存入集合 C 中
6:end for

7:for i=1 to N
8:do
9:if L_{i+1} ≥ L_i
10:对调 U_i 和 U_{i+1} 的位置
11:end for
12:return 排序好的集合 C
```

图 4　用户标签数算法

Fig.4　User tag number algorithm

2.3　实验分析

为了分析用户标签与用户转发行为之间的关系,提出了平均转发率的概念,即用户转发的微博数占所发布的总微博数的比值.如果用户 A 比用

户 B 的标签数量多,而且 A 的转发率也比 B 的高,就可以断定,用户标签对用户转发微博的行为是有正向影响的.同样使用了 5 000 个用户的数据来进行实验.对这 5 000 个用户的两个特征进行排序,一个是用户标签数量的排序集合,另一个是用户平均转发率的排序集合,部分数据见表 2.

表 2 按用户标签排序结果

Table 2 Sort results by user Tags

用户名	用户标签数	用户平均转发率
泳泳 Vujfljfah	8	10.7%
Queen_Yuri	8	8.1%
蜡笔小旧_sakula	7	7.6%
Number_Ninety_five	7	8.0%
WENJTNG-	7	3.7%
褒头 de 嘉静	6	5.8%
逗麻团儿	6	2.1%
yyj001 祈	5	6.6%
奈菜子 nanako	4	13.2%
呦呦呦游小包	4	3.2%

实验选取了用户一周之内的微博转发情况作为实验数据,为了减小误差,去除了标签过多或过少的用户.按照实验前的设定,在 5 000 个数据中,有 2 650 个数据满足的预期,正确率在 53%.实验结果表明,用户标签对用户的转发行为产生了较大的影响.用户粉丝的平均标签数越多,说明其粉丝越活跃,越有可能转发用户的微博.将所有用户粉丝的平均标签数与该用户微博的被转发率在曲线上表示,如图 5 所示.

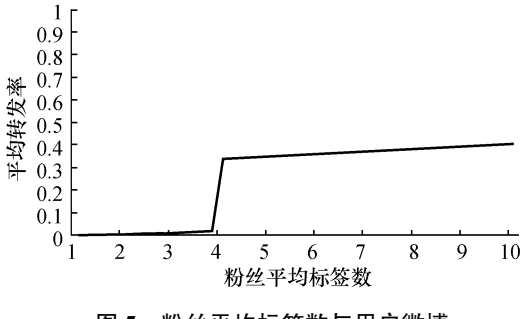

图 5 粉丝平均标签数与用户微博被转发率关系曲线

Fig.5 The relationship curve between the average number of fans and the number of the users micro-blog forward rate

由图 5 可知,当用户粉丝的平均标签数达到 4 个左右时,标签数对微博转发的贡献明显提高.因此,把粉丝平均标签数为 4 作为阈值,并将粉丝标签数作为一个特征属性加入到模型中.用户粉丝的平均标签数大于等于 4 时,将该属性值设置为 1;反之,将其设置为 0.

3 粉丝活跃度对微博转发的预测

3.1 算法思想

活跃粉丝对微博转发有着至关重要的作用,随着微博的兴起,网络上出现了一个特殊的群体:水军及僵尸粉."网络水军"大多是受雇于网络公关公司,为他人发帖回帖造势的网络人员.为客户发帖回帖造势常常需要成百上千个人共同完成,那些临时在网上征集来的发帖人被叫做"网络水军".版主把主帖发出去后,获得最广大的"网民"的注意,进而营造出一个话题事件,所有网络公关公司都必须雇佣大批的人员来为客户发帖回帖造势."网络水军"也有专职和兼职之分.在微博社区中,这类用户对微博的转发起不到任何作用.僵尸粉一般是指微博上的虚假粉丝,花钱就可以买到"关注",有名无实的微博粉丝,它们通常是由系统自动产生的恶意注册的用户.手机用户注册时,僵尸粉是由系统自动产生的关注.一些大 V 用户通过购买数量庞大的虚假账号来冒充自己的粉丝以提高人气,这类不活跃用户并不能对微博转发起到作用.

为了排除这些虚假账号和虚假交互行为带来的干扰,引入了用户粉丝活跃度[10] 这个特征属性.在微博社区中,有些用户的目的是向外界表达自己的观点以及抒发自己的情感,因此他们会经常性的发表自己的原创微博,很少去转发他人的微博;有些用户是为了获取信息和学习知识,这类用户会关注自己感兴趣的话题,浏览相关用户的微博,但很少发布原创微博及转发微博;还有一部分用户为了提高知名度,吸引更多关注,会尽可能多的转发微博,这类用户便是活跃用户.显然,只有那些热衷于转发微博的活跃用户才有可能转发被关注者所发布的微博.因此使用转发活跃度来描述用户粉丝的活跃度.

转发活跃度表示在特定时间内,用户转发微博的频率,即在时间 t 内,用户转发的微博数与发布的总微博数的比值.

$$R(u) = \sum_{i \in t} r_i \ / \ \sum_{i \in t} p_i \qquad (3)$$

其中,r_i 表示用户第 i 天转发的微博数;p_i 表

示用户第 i 天发表的总微博数.只有用户的转发活跃度达到一定值时,才可能产生转发行为.

3.2　实验分析

选取了一周中用户粉丝的转发活跃度作为自变量来研究用户微博被转发的情况,微博用户活跃度与微博转发之间的关系如图 6 所示.

通过该曲线可以看出,当用户的粉丝平均转发的微博数达到发布总微博数的 30%左右时,该用户的微博更易被转发.因此,将 0.3 设置为该特征属性的阈值,用户的粉丝活跃度大于或等于 0.3 时,将该属性值设置为 1;反之,将其设置为 0.

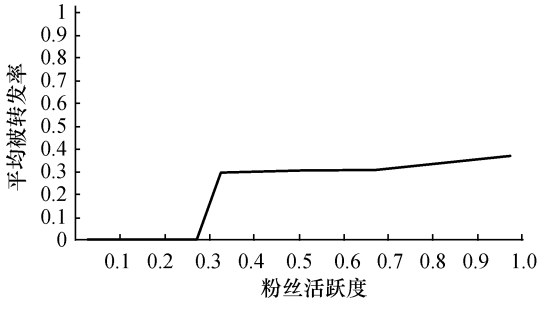

图 6　用户粉丝活跃度与用户微博被转发率关系曲线

Fig.6　The relationship curve of fan's activity and the user's micro-blog forward rate

4　结　论

本文介绍了三个影响微博转发的用户特征:用户影响力、粉丝标签数和粉丝活跃度.分析了各个属性与微博转发之间的关系,并通过实验,验证了这些因素对微博消息转发概率的影响大小,实验结果符合预期.

参考文献:

[1] ZAMAN T,HERBRICH R,VAN G.Predicting information spreading in Twitter[C]//Workshop on Computational Social Science and the Wisdom of Crowds.Whistler,Canada:NIPS,2010:17599-17601.

[2] 张昊,刘功申,苏波.一种微博用户影响力的计算方法[J].计算机应用与软件,2015,32(3):41-44.

[3] 张亚,阮彤,丁军.面向领域微博权威性人物分析技术与研究[J].计算机应用研究,2014,31(10):2907-2911.

[4] SUH B,HONG L,PIROLLI P,et al.Want to be Retweeted? Large scale analytics on factors impacting retweet in twitter network[C]//International Conference on Social Computing.Minneapolis,USA:IEEE,2010:177-184.

[5] KWAK H,LEE C,PARK H,et al.What is Twitter,a social network or a news Media? [C]//International World Wide Web Conference.Raleigh,USA:ACM,2010:591-600.

[6] WENG J,LIM E,JIANG J,et al.Twitter rank:finding topic-sensitive influential twitterers[C]//International Conference on Web Search and Data Mining.New York,USA:ACM,2010:261-270.

[7] 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(6):1809-1812.

[8] 谢婧,刘功申,苏波,等.社交网络中的用户转发行为预测[J].上海交通大学学报,2013,47(4):584-588.

[9] 邹青,张莹莹,陈一帆,等.社交网络中一种快速精确的节点影响力排序算法[J].计算机工程与科学,2014,36(12):2346-2354.

[10] SUNDAR S.To Tweet or to Retweet? That is the question for health professionals on Twitter[J].Health Communication,2013,28(5):509-524.