文章编号：1003-0077(2018)04-0130-07

基于主题模型的微博转发行为预测

郭 亚，宫叶云，张 奇，黄萱菁

（复旦大学 计算机科学技术学院，上海 201203）

摘 要：在全部微博内容中，由用户转发而产生的信息占有非常大的比例。同时，内容的转发也是微博中信息传播的主要途径。因此，用户的转发行为有着重要的研究价值，可应用于社交营销、微博检索、热点事件预测等领域中。该文中，我们通过分析所收集的大量真实的新浪微博数据，发现影响用户转发行为的一些因素：微博作者、用户兴趣以及微博热度。基于这些发现，该文提出了一种新颖的基于 LDA 模型的方法，综合利用以上 3 个特征预测用户转发行为。为了对该方法进行评价，我们利用收集的大量的微博数据及对应的社交网络结构模拟真实用户环境。实验表明，该方法的性能优于目前最好的方法，F 值比其他基线方法高出 35%—45%。

关键词：微博转发预测；主题模型；社交网络

中图分类号：TP391 **文献标识码**：A

Retweet Behavior Prediction Using Topic Model

GUO Ya, GONG Yeyun, ZHANG Qi, HUANG Xuanjing

(School of Computer Science, Fudan University , Shanghai 201203, China)

Abstract: In the Microblogging service, retweeting is a key behavior for information diffusion. The task of predicting retweet behavior is an important step for various social network applications, such as social marketing, microblog retrieval, popular event prediction, and so on. We collect a large number of microblogs and the corresponding social networks from Sina Weibo, and discover several factors which affect users' retweet behavior: author of the tweet, user interest and popularity of the tweet. Then we propose a novel retweet behavior prediction method based on LDA model to combine structural, textual and author information. To evaluate the proposed method, we simulate the real user environment on the constructed dataset. Experimental results demonstrate that the proposed method can achieve better performance than state-of-the-art methods. The relative improvement of the proposed over the baseline method is more than 35%—45% in terms of F1-Score.

Key words: retweet prediction; topic model; social network

0 引言

社交媒体发展迅速，已逐渐成为我们文化肌理的一部分。根据 2012 年的社交媒体报告[1]，美国人一个月内花费超过 1 211 亿分钟在社交媒体上。微博服务是一种通过关注机制分享简短实时信息的广播式的社交网络平台，用户可以方便的查看和转发关注用户的微博。微博信息可以通过用户转发迅速从一个社交圈传播到另一个社交圈，这可看作社交网络中的病毒传播[2]。通过对用户转发行为的研究，可以更好的理解用户行为，亦可进一步应用于社交营销[3-4]、微博检索[5]以及热点事件预测[6-7]等领域中。

最近几年，已有很多工作从不同角度对其进行了研究，包括社会影响力[8-9]，文本特征[10]及社交特征[11-13]等。Suh 等人[14]研究了微博内容，Hashtag，URL 以及文本特征对转发行为的影响。通过对转发微博的分析，我们发现用户不仅受到文本等特征的影响，同时，还受到微博本身属性的影响。例如，微博热度、微博作者等。而现有的方法则不能很好的利用这些信息。

收稿日期：2014-12-10 **定稿日期**：2015-11-08

基金项目：国家自然科学基金(61472088、61473092)

为了解决这个问题，我们提出了一种基于 LDA 模型[15]的方法，同时利用文本信息，结构信息和作者信息对用户转发行为进行建模。实验表明该方法的性能显著优于目前最好的方法。

本文的主要贡献有：

(1) 收集大量真实微博数据，包含微博内容、用户信息以及其对应的社交网络。模拟还原用户使用环境。

(2) 通过对数据进行分析，研究发现一些影响用户转发行为的重要因素：用户兴趣、微博热度和作者信息等。

(3) 提出了一个新颖的基于 LDA 模型的方法，该方法同时利用文本信息，结构信息和作者信息对用户转发行为进行建模。实验结果表明该方法的性能优越。

本文结构如下：第一节介绍相关工作以及相关领域最先进的方法；第二节介绍我们如何收集数据和分析数据；第三节介绍本文提出的方法；第四节描述实验方法，实验结果及其分析；第五节为总结部分。

1　相关工作

当前很多工作研究不同特征对用户行为的影响，比如文本内容，社交网络和时间信息等。Petrovic 等人[12]对社会特征，包括微博作者和内容进行了研究，他们通过实验说明这个任务确实可行。Naveed 等人[10]使用回归方法，加入高维和低维文本特征来预测转发行为。Luo 等人[13]研究了作者和关注者的历史信息，关注者的社会地位，微博内容和关注者微博内容的相似性。Feng 和 Wang[16]提出了通过历史转发记录来进行个性化的排名。他们使用特征感知的方法结合文本和用户特征对转发行为进行建模。Gupta 等人[6]基于文本内容，时间信息，地理信息和结构属性，将这个看作二分类问题进行研究。同时，他们也使用多分类方法来预测一条微博被转发的次数。Luo 等人[17]介绍了一种基于自回归移动平均模型(ARMA)的方法。其中转发行为被看作一个时间序列，序列值是对应的转发次数或者一段时间内的可能浏览次数。Peng 等人[18]使用条件随机场的方法对用户的发文历史和社交关系进行特征抽取。

与以上这些方法不同，我们提出了一个基于 LDA 的方法来预测转发行为。微博内容、结构信息和作者信息统一到一个模型中。

2　数据收集和分析

我们从新浪微博中收集数据。在新浪微博中，用户只能看到关注用户的微博，我们抓取数据，然后模拟真实的微博网络。下面介绍数据集的构造方法。

首先，随机选取 200 个用户作为核心用户，也是我们的微博网络中的第一层用户。然后抓取这 200 用户的关注列表，将他们所有关注的用户作为微博网络中的第二层用户，这一层共有 82 311 个用户。这样得到了一个两层微博网络。最后我们抓取网络中用户的最新的 2 000 条微博，共约 8 500 万条。具体统计数据见表 1。

表 1　数据集统计数据

	用户数	微博数	转发数	转发评论数
第一层(核心用户)	200	189 070	99 918	27 844
第二层	82 311	84 768 859	50 998 887	16 974 326

从表 1 中可以看出约 60%的微博是转发的，其中约 33%包含评论。这与 Yu 等人[22]统计的结果类似，可以认为这个统计结果能反映不同文化背景的社交媒体的真实情况。

为了便于数据分析，我们对微博进行预处理，去除微博中的标点、URL、表情和图片等无用信息。然后对微博进行分词处理。其中转发的微博分为两类，一类带有评论，另一类没有。对于带评论的转发微博，我们将评论与转发内容当作两条微博处理。

图 1 到图 4 分别统计了第一层的用户微博数分布、用户转发数分布、微博词数分布和用户关注数分布。

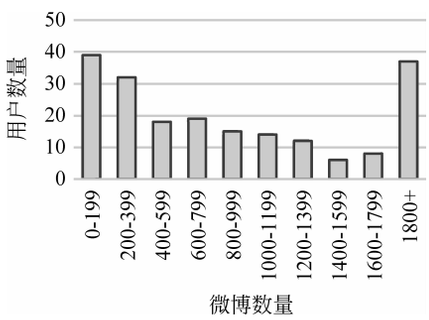


图 1　用户微博数分布

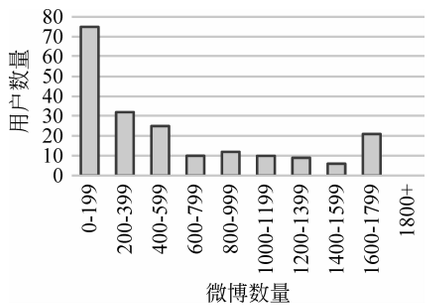


图2 用户转发数分布

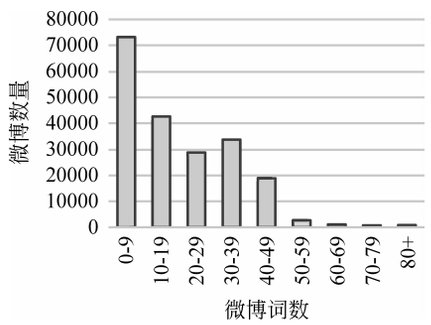


图3 微博词数分布

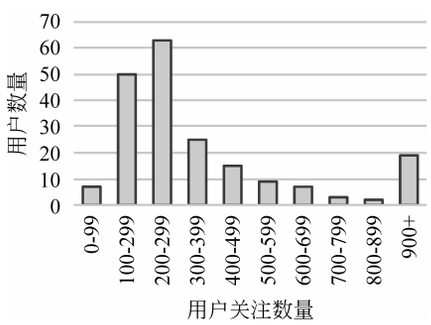


图4 用户关注数分布

由图1和图2可以看出,用户的微博数(转发数)呈现两极分化,微博(转发)数量小于400和大于1 600的用户较多。而从图3可以看出,绝大多数用户发送的微博词数都小于20。图4表明用户关注数量集中于300左右。

第一层的网络由200核心用户构成,我们认为他们的浏览历史包含了他们转发一条微博的时间点到这条被转发微博的发送时间点之间的所能看到的微博。通过恢复用户的发送和浏览历史,可以观察到下面的现象:

1) 第一层200核心用户共关注了82 311个用户。其中被核心用户转发过微博的用户有52 177个,占总关注用户的63.3%。此外,被转发超过1次的只占17.8%。所以用户一般只会转发某一部分关注用户的微博。

2) 当用户浏览看到多条同样的微博时,不一定转发首次看到的那一条微博。根据统计大约37.4%的转发行为属于这类情况。从而说明用户的转发行为不仅受到微博内容的影响还受到微博作者的影响。

3) 每条微博在被转发之前,可能已经被其他关注用户转发过多次,我们称之为微博热度。统计每个用户转发的微博的热度分布,可以发现,不同用户的转发热度分布不同,即有些人偏好转发那些很火的微博,而有些则不然。后文我们将这个特征称为结构特征。

4) 用户更偏好转发自己感兴趣的微博,而不同用户有不同的兴趣爱好,我们使用用户微博的话题分布来表示用户的兴趣。

综上所述,用户的转发行为受到用户兴趣、微博作者和微博热度的影响,分别称之为内容影响、用户影响和结构影响。因此,我们假设用户 a 是否转发一条微博由以下因素决定:1)谁发送这条微博;2)用户 a 关注的用户中有多少人发送或转发了这条微博;3)微博的内容;4)用户 a 的兴趣。

3 用户行为预测模型

本节中,首先简要介绍一下LDA模型,然后详细介绍我们提出的预测转发行为的方法。

3.1 LDA主题模型

Latent Dirichlet Allocation(LDA)模型由Blei等人[15]在2003年提出,LDA是一种主题模型,可以将文档集中每篇文档的主题按照概率分布的形式给出。LDA也是一种非监督学习方法,可用于识别大规模文档集中潜藏的主题信息,目前广泛应用于文本挖掘等领域。

LDA采用词袋(bag of words)方法,认为词之间没有顺序关系。文档是由词构成的集合,文档包含多个主题,文档中每一个词都由其中的一个主题生成。

3.2 ASC-LDA

通过第二节介绍我们可以知道影响用户行为的关键因素:用户影响、结构影响和内容影响。通过扩展LDA模型,利用这三个因素对用户行为进行建模。

用户影响(A):由第二节的统计数据可知,用户

可能只转发几个特定用户的微博。因此，对于用户 u，我们假设他转发每个关注用户 p_{fe_i} 的微博的概率 fe_i 服从二项分布，这个二项分布以 $Beta$ 分布为先验分布。

结构影响(S)：一些用户可能比较喜欢转发那些已经被很多用户转发过的微博，即热度高的微博，另一些用户则相反。因此，我们假设每个用户 u 对应一个转发热度分布。我们首先对每条微博的转发次数做归一化处理，使其取值范围为 0 到 1 之间，归一化后的值用 x_d 表示。最后使用 $Beta$ 分布对其进行模拟。

内容影响(C)：内容影响通过隐含的主题进行建模。我们使用基于 LDA 的主题模型来完成这一任务。通过使用 Gibbs 采样估计隐含变量，微博 d 的生成概率如式(1)所示。

$$p_c(w_d \mid \mathbf{z},\mathbf{l}) = \prod_{n=1}^{N_d} f(w_{dn} \mid \phi^{z_{dn},l_d}) \qquad (1)$$

式(1)中，w_d 是微博 d 中的词，N_d 是微博 d 中的词数，w_{dn} 表示微博 d 中的第 n 个词，z_{dn} 表示微博 d 中第 n 个词的主题，l_d 是微博 d 的转发标记，$f(w_{dn} \mid \phi^{z_{dn},l_d})$ 是在当前转发标记 l_d 下生成词 w_{dn} 的似然函数。符号说明见表 2。

表 2　模型中主要参数说明

D	训练集
W	语料中词典
L	微博转发标记
K	主题数目
A	关注用户集(微博作者)
w_{dn}	第 d 条微博中的词
z_{dn}	第 d 条微博第 n 个词的主题
l_d	第 d 条微博的转发标记
$\phi^{z,l}$	标记 l 下主题 z 的词分布
θ_d	微博 d 的主题分布
x_d	正则化后的一条微博被邻居转发的次数
ψa	用户 a 的被转发的标记 l 的分布

这里使用 D 表示用户 u 的浏览历史微博。D 中第 d 条微博包含一个词序列 $w_d = \{w_{dn}\}_{n=1}^{N_d}$，其中 N_d 是第 d 条微博的字数，w_{dn} 是字典 W 中的一个字。A_d 表示第 d 条微博的作者。给定一个用户，一条微博以及它的作者，那么任务就是判断该用户是否会转发这条微博。

模型的生成过程如图 5：

(1) 用户关注的每个用户 $a \in A$

—生成 $\psi^a \sim Beta(\lambda)$

(2) 对于每一个主题 $z \in K$，和转发标记 l，根据 $\phi^{z,l} \sim Dir(\delta^l)$，得到主题词分布 $\phi^{z,l}$。

(3) 对于每一条微博 $d \in D$

a) 生成转发标记 $l_d \sim Binomial(\psi^a)$

b) 生成正则化后的转发次数 $x_d \sim Beta(\eta_{l_d})$

c) 根据 $\theta_d \sim Dir(\alpha)$，得到主题分布 θ°_d

d) 微博中每一个词 $n = 1,\ldots,N_d$

—根据分布 $z_{dn} \sim Mult(\theta_d)$，得到主题 z_{dn}

—根据分布 $w_{dn} \sim Mult(\phi^{z_{dn},l_d})$，得到词 w_{dn}

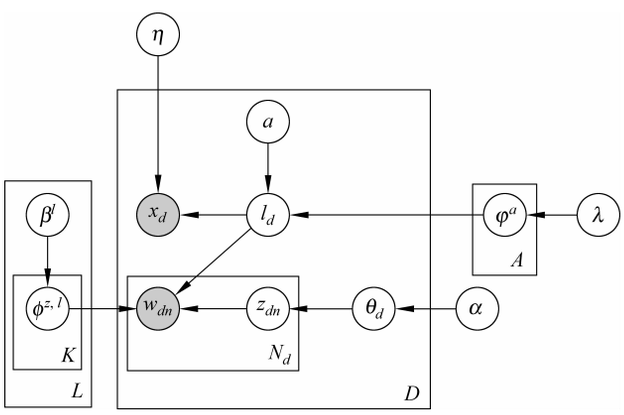


图 5　ASC-LDA 图模型

3.2.1　模型推断

我们使用 *Gibbs* 采样学习模型的参数，采样过程分为对每个词的主题 z 采样和对微博的转发标记 l 进行采样。

对 z 采样：微博中的每一词 w_{dn} 在转发标记 l 下，分配主题 $z_{dn} = k$ 条件概率：

$$p(z_{dn} = k \mid \mathbf{z}^{\neg dn}, l_d) \propto (m_k^{d,\neg dn} + \alpha_k) \frac{n_{w_{dn},\neg dn}^{k,l_d} + \beta^l_d}{n_{(.),\neg dn}^{k,l_d} + \beta^l_d} \qquad (2)$$

式(2)中，$\neg w_{dn}$ 表示去除当前词 w_{dn} 的情况。$m_k^{d,\neg dn}$ 表示主题 k 分配给文档 d 中的词的次数，$n_{w_{dn},\neg dn}^{k,l_d}$ 微博标记为 l_d 并且主题为 k 的条件下，词 w_{dn} 出现的次数，$n_{(.),\neg dn}^{k,l_d}$ 微博标记为 l_d 并且主题为 k 的所有词的个数，$\neg dn$ 表示在计算时不考虑当前位置的词。

对 l 采样：当给定每个词的主题 z 的情况下，对第 d 篇文档的转发标记利用式(3)进行采样：

$$p(l_d=l\mid \mathbf{z},w_d,l^{\neg d},x_d)\propto \prod_{n=1}^{N_d}\frac{n^{z_{dn},l_d}_{w_{dn},\neg dn}+\beta^{l_d}}{n^{z_{dn},l_d}_{(.),\neg dn}+\beta^{l_d}}\cdot\frac{N^l_a+\lambda_1}{N_a+\lambda_1+\lambda_2}\cdot\frac{(1-x_d)^{\eta_{l1}-1}x_d{}^{\eta_{l2}-1}}{B(\eta_{l1},\eta_{l_2})} \tag{3}$$

式(3)中，N_a 表示用户 u 可以看到用户 a 的微博数。其中用户 u 转发了 N^l_a 条。$B(\eta_{l1},\eta_{l_2})$ 是以 η_{l1} 和 η_{l2} 参数的 *Beta* 分布，每一轮迭代采样后，更新 η 的取值，公式如式(4)～(5)所示。

$$\eta_{l1}=\bar{x}^l_d\left(\frac{\bar{x}^l_d(1-\bar{x}^l_d)}{\delta^2_l}-l\right) \tag{4}$$

$$\eta_{l2}=(1-\bar{x}^l_d)\left(\frac{\bar{x}^l_d(1-\bar{x}^l_d)}{\delta^2_l}-l\right) \tag{5}$$

式(4)～(5)中，$\bar{x}^l_d$ 和 δ^2_l 是分别是在标记 l 时训练集结构特征的均值和方差。对于那些转发次数小于20次的用户，为了避免稀疏引起的问题，我们设置 $\eta_{l1}=\eta_{l2}=0.5$。

3.2.2 转发预测

给定一条用户看到的未标记的微博 d，首先通过迭代采样，直到隐含变量稳定后，计算得到该微博的主题分布，然后通过式(6)计算这条微博被用户转发的概率：

$$p(l_d=l\mid w_d,\mathbf{z}^{\neg dn},l^{\neg d},x_d)\propto p(w_{dn}\mid w_d)$$
$$\sum_{z=1}^{K}\prod_{n=1}^{N_d}\frac{n^{z_{dn},l_d}_{w_{dn},\neg dn}+\beta^{l_d}}{n^{z_{dn},l_d}_{(.),\neg dn}+\beta^{l_d}}\cdot p(z_{dn}\mid w_d,\mathbf{z}^{\neg dn},l)\cdot\frac{N^l_a+\lambda_1}{N_a+\lambda_1+\lambda_2}\cdot\frac{(1-x_d)^{\eta_{l1}-1}x_d{}^{\eta_{l2}-1}}{B(\eta_{l1},\eta_{l2})} \tag{6}$$

式(6)中 $p(w_{dn}\mid w_d)$ 是词 w_{dn} 在微博 d 中的权重，权重值通过 TD-IDF 计算；$p(z_{dn}\mid w_d,\mathbf{z}^{\neg dn},l)$ 是转发标记 l 时生成主题 z_{dn} 的概率。

4 实验

4.1 实验设置

在第二节中介绍了数据集的收集，通过恢复核心用户的浏览历史，我们可以模拟用户的实际使用环境。每一个用户我们将浏览历史中的70%作为训练集，剩下30%作为测试集，统计信息见表3。

表3 实验数据集统计信息

	微博数	词数
训练集	140 239	1 253 568
测试集	38 375	344 942

实验中使用精度(P)、召回率(R)和F1-score($F1$)来评价模型效果。其中F1-score是精度和召回率的调和平均数。模型进行500次迭代采样。在基于LDA的模型中，α 设为 $50/K$，$\beta=0.1$。其中 K 是主题个数，模型中参数 λ_1 和 λ_2 均设为0.1，通过试验，我们将所有基于LDA的模型的主题个数设为20。

实验中将我们的方法与以下几个 baseline 方法进行比较：

(1) **Naïve Bayes**：转发预测任务被看作一个二分类问题，每条微博转发与不转发标记代表两类，通过朴素贝叶斯模型计算给定一条微博各个标记的后验概率。

(2) ***SVM^{rank}***：我们实现 Luo 等人[13]提出的方法，该方法利用微博内容，粉丝的身份信息、关注时间以及兴趣等特征来完成这一任务。

(3) **SC-LDA**：同样基于LDA模型实现，在完整模型的基础上去除作者信息的影响进行训练。在得到每篇微博的主题分布后，对于用户 u，给定他看到的一篇微博，转发标记打分计算如式(7)所示。

$$p(l_d=l\mid w_d,\mathbf{z}^{\neg dn},l^{\neg d},x_d)\propto p(w_{dn}\mid w_d)$$
$$\sum_{z=1}^{K}\prod_{n=1}^{N_d}\frac{n^{z_{dn},l_d}_{w_{dn},\neg dn}+\beta^{l_d}}{n^{z_{dn},l_d}_{(.),\neg dn}+\beta^{l_d}}\cdot p(z_{dn}\mid w_d,\mathbf{z}^{\neg dn},l)\cdot\frac{\lambda_1}{\lambda_1+\lambda_2}\cdot\frac{(1-x_d)^{\eta_{l1}-1}x_d{}^{\eta_{l2}-1}}{B(\eta_{l1},\eta_{l2})} \tag{7}$$

- **AC-LDA**：该方法忽略结构信息的影响，转发标记打分计算如式(8)所示。

$$p(l_d=l\mid w_d,\mathbf{z}^{\neg dn},l^{\neg d},x_d)\propto p(w_{dn}\mid w_d)$$
$$\sum_{z=1}^{K}\prod_{n=1}^{N_d}\frac{n^{z_{dn},l_d}_{w_{dn},\neg dn}+\beta^{l_d}}{n^{z_{dn},l_d}_{(.),\neg dn}+\beta^{l_d}}\cdot p(z_{dn}\mid w_d,\mathbf{z}^{\neg dn},l)\cdot\frac{N^l_a+\lambda_1}{N_a+\lambda_1+\lambda_2} \tag{8}$$

4.2　实验结果

我们将从两个方面对提出的方法进行评估分析：

1)与其他当前最好方法进行比较。

2)评估实验参数对实验结果的影响。

表 4 展示了各种方法的实验结果。通过结果可以看出：1)Naïve Bayes 实验效果最差。2)我们提出的方法效果明显好于其他方法。3)各个特征都对实验结果有影响。

表 4　实验对比结果

方法	精度(P)	召回率(R)	F1-score(F1)
Naïve Bayes	0.362	0.617	0.456
SVMrank	0.485	0.450	0.467
C-LDA	0.474	0.507	0.490
AC-LDA	**0.752**	0.545	0.632
SC-LDA	0.446	**0.650**	0.529
ASC-LDA	0.694	0.636	**0.664**

其中 C-LDA 是只考虑文本特征的实验结果，但也比 Naïve Bayes 和 SVM 方法要好。将它分别与 AC-LDA 和 SC-LDA 比较可以发现，作者信息有助于提高精确度，而结构信息对召回率有较大影响。比较 C-LDA 和 ASC-LDA 的 F1-score 可以发现：在作者信息和结构信息同时作用下实验结果提高大概 35%，效果明显。

图 6 中将用户根据微博数分为五组，分析了微博数对实验的影响。由图 6 可见，用户发送的微博越多实验结果越好。同时通过图 1 可知，微博数量超过 1 000 条的用户占用户总数的 38.5%，所以这部分用户对实验结果有较大影响。

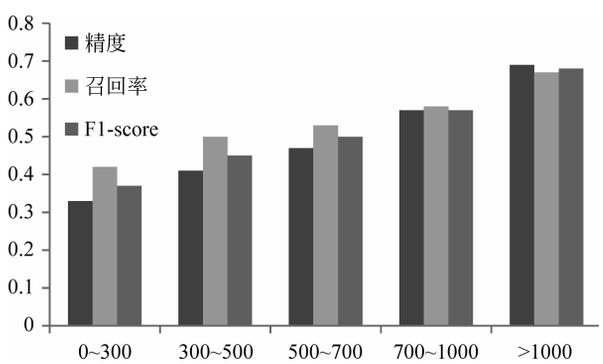


图 6　用户微博数对实验结果的影响

模型中有一些超参数，其中主题数是最重要的参数之一。表 5 中展示了主题数目对实验效果的影响，从表中可以看出，在主题数目为 20 时效果最好。

表 5　主题数目对 ASC-LDA 方法实验结果的影响

主题数	精度(P)	召回率(R)	F1-score(F1)
10	0.681	0.625	0.651
20	**0.694**	**0.634**	**0.664**
30	0.692	0.629	0.659
50	0.659	0.608	0.633
70	0.645	0.593	0.618
100	0.630	0.578	0.603

5　总结

本文首先从真实的社交网络中收集了大量的微博数据以及网络信息，重构了用户的实际使用环境。然后通过大量的数据观察和分析，发现了影响用户转发行为的重要因素：作者信息、用户兴趣和微博热度。基于此，我们提出一个新颖的预测微博转发行为的方法 ASC-LDA。该方法基于 LDA 模型，同时利用结构信息、作者信息和文本信息对用户行为进行建模。实验表明，结构信息、作者信息和文本信息都对实验结果有影响。我们的方法效果优于当前最好的方法，F 值比其他 Baseline 方法高出 35%—45%。

参考文献

[1] State of the Media: The Social Media Report 2012 [DB/OL]. http://www.nielsen.com/us/en/reports/2012/state-of-the-media-the-social-media-report-2012.html,2012.

[2] Rodrigues T, Benevenuto F, Cha M, et al. On word-of-mouth based discovery of the web[C]//Proceedings of SIGCOMM '11, 2011.

[3] Castellanos M, Dayal U, Hsu M, et al. Lci: a social channel analysis platform for live customer intelligence [C]//Proceedings of SIGMOD '11, 2011.

[4] Homan D L, Fodor M. Can you measure the roi of your social media marketing[C]//Proceedings of MIT Sloan Management Review, 2010:41-49.

[5] Chang J, Kim H J. Twitter search methods using retweet information[C]//Proceedings of BUSTECH '12, 2012:67-71.

[6] Gupta M, Gao J, Zhai C, et al. Predicting future pop-

ularity trend of events in microblogging platforms [C]//Proceedings of the American Society for Information Science and Technology, 2012:1-10.

[7] Hong L, Dan O, Davison B D. Predicting popular messages in twitter[C]//Proceedings of WWW '11, 2011.

[8] Liu L, Tang J, Han J, Jiang M, et al. Mining topic-level influence in heterogeneous networks[C]//Proceedings of CIKM '10, 2010.

[9] Zhang J, Liu B, Tang J, et al. Social influence locality for modeling retweeting behaviors[C]//Proceedings of IJCAI'13, 2013.

[10] Naveed N, Gottron T, Kunegis J, et al. Bad news travel fast: A content-based analysis of interestingness on twitter[C]//Proceedings of Web Science Conf., 2011.

[11] Zaman T R, Herbrich R, Van Gael J, et al. Predicting information spreading in twitter[C]//Proceedings of Workshop on Computational Social Science and the Wisdom of Crowds, NIPS, 2010.

[12] Petrovic S, Osborne M, Lavrenko V. Rt to win! predicting message propagation in twitter[C]//Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,2011.

[13] Luo Z, Osborne M, Tang J, et al. Who will retweet me?: Finding retweeters in twitter[C]//Proceedings of SIGIR '13, 2013.

[14] Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network[C]//Proceedings of SocialCom' 10, 2010.

[15] Blei D M, Ng A Y and Jordan M L. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research 2003: 993-1022.

[16] Feng W, Wang J. Retweet or not?: personalized tweet re-ranking[C]//Proceedings of the sixth ACM international conference on Web search and data mining, 2013:577-586.

[17] Luo Z, Wang Y, Wu X. Predicting retweeting behavior based on autoregressive moving average model [C]//Proceedings of Web Information Systems Engineering-WISE 2012, 2012:777-782.

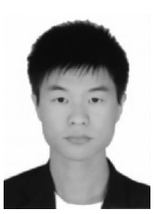
[18] Peng H K, Zhu J, Piao D, et al. Retweet modeling using conditional random fields[C]//Proceedings of ICDMW '11, 2011.

[19] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter[C]//Proceedings of HICSS '10, 2010.

[20] Nagarajan M, Purohit H, Sheth A P. A qualitative examination of topical tweet and retweet practices [C]//Proceedings of the ICWSM, 2010.

[21] Letierce J, Passant A, Decker S, et al. Understanding how twitter is used to spread scientific messages [C]//Proceedings of Web Science Conference, 2010.

[22] Yu L L, Asur S, Huberman B A. Artificial inflation: The real story of trends and trend-setters in sina weibo[C]//Proceedings of Social Com-PASSAT '12, 2012.



郭亚(1990—),硕士,主要研究领域为自然语言处理和信息检索。
E-mail: gyalife@gmail.com

张奇(1981—),博士,副教授,主要研究领域为自然语言处理和信息检索。
E-mail: qi_zhang@fudan.edu.cn

宫叶云(1987—),博士,主要研究领域为自然语言处理和信息检索。
E-mail: ye163yun@163.com