文章编号：1003-0077(2018)06-0107-07

基于马尔科夫随机场的微博用户转发行为预测

王 宁[1]，高 光[2]，柴争义[1]

（1. 周口师范学院 计算科学与技术学院，河南 周口 466000；
2. 周口师范学院 网络工程学院，河南 周口 466000）

摘 要：微博用户转发行为预测是微博社交网络消息扩散模型构建的基础，在图书阅读推广、舆情监控与市场营销等领域有着广泛的应用。为了提高用户转发行为预测的精度，该文在马尔科夫随机场框架下综合分析了用户属性与微博内容特征、用户转发行为约束等因素对用户转发行为的影响，并在逻辑回归模型的基础上构造了相应的能量函数对用户转发行为进行了全局性的预测。实验结果表明，微博用户转发行为不仅取决于用户属性、微博内容等特征，而且也受到与其相邻用户转发行为的约束。相对于传统算法该文算法可以更准确地对用户转发行为进行建模，因而可获得更好的预测结果。

关键词：新浪微博；转发预测；能量优化；逻辑回归

中图分类号：TP391 **文献标识码**：A

Predicting Microblog User Forwarding Behaviors Based on Markov Random Fields

WANG Ning[1], GAO Guang[2], CHAI Zhengyi[1]

(1. School of Computer Science and Technology, Zhoukou Normal University, Zhoukou, Henan 466000, China;
2. School of Network Engineering, Zhoukou Normal University, Zhoukou, Henan 466000, China)

Abstract: Weibo users' forwarding behavior prediction is the foundation of weibo social network message diffusion model, which can be widely applied in book reading promotion, public opinion monitoring, marketing management and other fields. Under the framework of Markov random field, this paper comprehensively analyzes the effects of user attributes, the weibo contents, and the user forwarding behavior constraints. The logistic regression model is employed to construct the prediction model of the user's forwarding behavior. The experimental results show that the forwarding behavior of weibo users depends on the user attributes, the micro-blog contents, as well as the forwarding behavior of their neighboring users. Compared with the traditional algorithms, the proposed algorithm can model the user's forwarding behavior more accurately.

Key words: Sina weibo; forward prediction; energy optimization; logistic regression

0 引言

随着互联网技术的发展与各种智能终端的普及，微博、论坛等社交网络对人们日常生活的影响日益增大。根据微博社交网络历史数据，有效分析影响用户转发行为的特征并对其未来的转发行为进行预测，在图书阅读推广、舆情监控等领域有着广泛的应用。

为了提高微博用户转发行为预测的精度，本文提出了基于马尔科夫随机场的用户转发行为预测算法，其中的能量函数融合了用户属性、微博内容等特征以及用户转发行为约束，因而可以全局性地对用户转发行为进行预测。实验结果表明，本文算法可以有效克服传统算法的缺点，整体上具有较高的性能。本文贡献主要有以下两点：①对影响用户转发

收稿日期：2017-07-25 **定稿日期**：2017-10-16

基金项目：国家自然科学基金(U1504613,41401463)；河南省高校科技创新团队计划(171RTSTHN009)；河南省知识产权局软科学研究项目(20170106041)；周口师范学院青年基金(zknuB3201601,zknuB315204)

行为的诸多因素(如用户属性、微博内容等)进行了系统的分析,特别对影响用户共同转发行为的特征进行了深入的探讨;②提出了基于马尔科夫随机场的用户转发行为预测模型,综合利用用户属性、微博内容等特征及用户转发行为约束等信息对用户转发行为进行全局性预测,有效提高了整体预测精度。

1 相关工作

微博社交网络消息扩散的研究在图书阅读推广、舆情监控等方面有着广泛的应用。用户转发行为作为消息扩散的原子行为,其预测的可靠性与精度对消息扩散模型的构建具有重要作用[1-3]。

在实际中,微博用户转发行为预测问题通常可视为两类分类(即转发与不转发)问题进行求解;其中,参与分类的样本特征对分类的结果有着重要的影响。在相关工作中,Suh 等探讨了影响用户转发行为的各种因素[4],并采用广义线性模型分析了影响因素(如 URL、关注用户数等)与转发行为之间的关系。曹玖新等以新浪微博为研究对象,对各种可能影响用户转发行为的因素进行了统计与分析,并利用用户属性、社交关系与微博内容等特征分别采用逻辑回归、贝叶斯网络等方法对用户转发行为进行了预测[5]。张旸等为了提高用户转发行为预测的精度,采用特征加权的方式以强调不同特征对用户转发行为预测的作用[6]。Hong 等在对微博内容与主题信息、网络结构、微博发布时间等因素进行分析的基础上,分别采用两类分类与多类分类方法对用户转发行为与微博转发范围进行了预测[7]。Tang 为了突出用户转发行为的个性差异,在传统逻辑回归模型的基础上将不同用户的转发行为定义为不同的任务进行处理,进而提高了预测精度[8]。

为了进一步提高用户转发行为预测的精度,近年来,研究者也对更多类型的算法进行了深入的探索,如 Zaman 等人采用协同过滤方法预测用户的转发行为[9];Petrovic 通过调查人类对微博的转发意向,采用被动攻击算法对用户的转发行为进行全局性的预测[10];Xu 根据用户转发行为被突发新闻、朋友的发布、用户自身的兴趣等因素所影响的特点,采用一种混合型的隐主题模型对用户转发行为进行预测并获得了较好的结果[11];Yang 等人对影响用户转发行为的因素(如用户兴趣、微博内容、转发时间等)进行了分析,采用因子图模型对单级用户转发行为以及微博被转发的范围进行了预测,发现概率图模型在微博转发范围的预测中表现出较好的性能[12];周沧琦等根据行为周期时长差异性与昼夜作息规律对兴趣可变的人类行为动力学模型进行了改进,同时也兼顾了影响用户转发行为的内在与外在因素,进而构建了相对可靠的用户转发行为模型[13]。

社交网络中的消息通常以当前发布者为中心向四周扩散,由于当前发布者的所有关注与粉丝用户具有较高聚集性,因而,这些用户对消息的转发行为往往也主要受与其直接或间接相连的较小部分用户的影响(即服从 80-20 规则)。在此基础上,Peng 等人探索了 Twitter 社交网络用户转发行为中的马尔科夫性质,并将微博内容、用户关系等特征融合于条件随机场框架下对用户的转发行为进行预测[14]。Zhang 等人在社交网络中以当前用户为中心的局部区域内,针对其他用户转发行为对当前用户转发行为的影响进行了研究,发现当前用户的转发行为往往更易于受到其直接关注用户所构成的局部社交网络结构的影响[15];基于此,在不需要刻意构造用户或微博特征的情况下,仅利用传统逻辑回归方法即可对用户转发行为进行较好的预测。Wang 等人在对社交网络消息扩散最大化问题的研究中,在相关模型中考察了活动用户与被通知用户(即相邻用户中存在至少 1 个活动用户)之间的相互影响,以及对消息扩散的影响,进而获得了较好的效果[16]。

以上算法尽管在特定条件下可获得较好的效果,但相关模型通常不能全局性地对社交网络中所有用户的转发行为或者影响用户转发行为的关键因素进行描述,其可靠性与精度在很多情况下往往难以得到保证。本文算法将用户属性、微博内容等特征以及用户转发行为约束等因素统一在马尔科夫随机场框架下,以对用户的转发行为特征进行描述,可以更可靠地对用户的转发行为进行预测。

2 基于马尔科夫随机场的用户转发行为预测模型

已知微博社交网络 $G=<U,E>$,其中 $U=(u_i\}(i=1,\cdots,n)$ 与 $E=\{(u_i,u_j)\}(i,j=1,\cdots,n)$ 分别表示用户的集合与用户之间社交关系的集合。对于新发布的微博 T,以 $y_i\in\{1,0\}$ 表示用户 u_i 的转发行为(即 y_i 取值 1 与 0,分别表示转发与不转发),本文定义用户转发行为预测问题为:对于网络 G 中的用户集合 $U=\{u_i\}(i=1,\cdots,n)$,如何全局性地确定相应的转发行为标记集 $Y=\{y_i\}(i=1,\cdots,n)$。

针对以上问题，本文在马尔科夫随机场框架下对其进行求解，相应的能量函数如式(1)所示。

$$E(Y)=\sum_{i=1}^{n}\left(D_m(y_i,u_i)+\lambda_1\cdot\sum_{j\in N(i)}\psi_{i,j}\delta(y_i\neq y_j)\right) \tag{1}$$

其中，构成能量函数的两部分主要用于描述用户的局部转发行为、用户转发行为约束；$N(i)$表示与用户 u_i 存在直接关注关系用户的序号集合；λ_1 为用户转发行为约束权重。

2.1　局部转发行为

局部转发行为度量了用户仅根据用户转发行为特征对当前微博进行转发的行为(为方便算法描述，相应的转发概率简称为局部转发概率)，相应的代价通常与用户的局部转发概率相关，即：当用户的局部转发概率较高时，则其实际转发微博的概率就越高，相应的代价则越低；否则实际转发微博的概率就越低，相应的代价则越高。因而，相应的度量 $D_m(y_i,u_i)$ 定义为式(2)。

$$D_m(y_i,u_i)=|y_i-P(u_i,m)| \tag{2}$$

其中，$y_i\in\{1,0\}$ 为用户 u_i 可能被分配的转发标记，$P(u_i,m)$ 为用户 u_i 对微博 m 的局部转发概率。

以下对局部转发概率 $P(u_i,m)$ 的求取方法进行描述。

2.1.1　用户转发行为特征

用户属性、微博内容等用户转发行为特征是用户转发行为预测的基础，近年来有大量文献对此报道，在此不再赘述。以下仅对本文算法采用的用户属性与微博内容等特征进行介绍。实验中发现，本文算法由于融合了用户转发行为约束，因而对用户转发行为特征的提取并不敏感。

(1) 用户属性

用户属性特征主要包括关注用户数、粉丝数、是否认证、发布微博数、被转发微博数、转发活跃度等项，通常可从微博数据中直接获取。

在这些特征中，关注用户数表明当前用户利用微博社交网络获取信息的偏向性或其社交活跃性，该值越大，则用户在主观上转发微博的意愿越强。相对地，粉丝数与是否认证在用户转发行为预测上区分性较小，但由于其是微博社交网络中用户影响力判别的重要特征，根据当前求解问题的性质，此处也将其考虑在内。发布微博数度量了用户在微博社交网络中的活跃度，而被转发微博数则是度量其在社交网络中影响力的基本标准，两者对用户转发行为的预测均具有重要的影响。此外，在本文中，用户转发活跃度定义为用户在一定时期内转发微博的数量与其所发布微博总数的比例，度量了用户转发微博的倾向或积极性；其值越高，则用户转发微博的可能性越大。

(2) 微博内容

根据微博社交网络的特征，用户更偏向于转发其感兴趣的微博或与当前热点话题相关的微博，因而，微博内容对消息在微博社交网络中的持续扩散具有重要影响。为了度量用户 u_i 对微博 m 的内容感兴趣的程度，本文将其历史原创与转发的微博汇集成文档 d_i，然后采用 LDA(latent dirichlet allocation)模型[17]分别计算文档 d_i 与微博 m 在预定的 50 个主题(如教育、军事等)上的概率分布，最后利用余弦距离确定相应的主题相似度，如式(3)所示。

$$L(d_i,m)=\frac{\mathrm{LDA}(d_i)\cdot\mathrm{LDA}(m)}{\|\mathrm{LDA}(d_i)\|\|\mathrm{LDA}(m)\|} \tag{3}$$

除微博的主题特征之外，微博被转发的次数、微博内容长度以及微博内容中是否包含 URL 或@信息等也对用户转发行为产生一定的影响，因此本文也将其考虑为预测用户转发行为的部分特征。

在提取以上特征之后，为了消除不同特征之间数值类型(如离散与连续型)与取值范围的差异，本文对其进行了规范化处理，如式(4)所示。

$$f'=\frac{f-f_{\min}}{f_{\max}-f_{\min}} \tag{4}$$

其中，f 与 f' 分别为初始特征与规范化后的特征，$f_{\min}$ 与 $f_{\max}$ 分别为所有用户当前特征的最小值与最大值。

最终，本文将规范化后的特征构成 11 维的特征向量，用于局部转发概率 $P(u_i,m)$ 的求取。

2.1.2　局部转发概率的求取

逻辑回归模型是一种线性分类模型，可以获取样本属于不同类别的概率。由于用户局部转发行为仅有两种状态(即转发与不转发)，因而逻辑回归模型可用于求取用户的局部转发概率。具体而言，已知用户 u_i 对应的规范化特征向量 $\boldsymbol{x}_i$，则其对微博 m 的局部转发概率 $P(u_i,m)$，如式(5)所示。

$$P(u_i,T)=\frac{1}{1+\exp(-w^T x_i)} \tag{5}$$

其中，w 为特征权重向量，在本文中通过梯度下降算法最小化特定风险函数获取，如式(6)所示。

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} (L(y_i, P(u_i, T)) + \lambda_2 \cdot \|w\|_2^2) \tag{6}$$

式(6)中,损失函数 $L(\cdot)$ 采用交叉熵损失函数,n 为样本数量,$\|\cdot\|_2$ 为 L_2 范式正则化项,λ_2 则为控制正则化强度的参数。

2.2 用户转发行为约束

用户转发行为约束度量了具有关注关系的两用户转发行为之间的依赖性。当两用户具有相似的转发行为时,两者共同转发相同微博的概率越高,实际的转发行为也应该相同;或者说,当两用户具有相似的转发行为时,如果在能量优化过程中为其分配不同的转发标记,则应给予较大的惩罚量,否则应给予较小的惩罚量,以鼓励其获取不同的转发标记。因而,相应的惩罚量 $\psi_{i,j}$ 定义为式(7)。

$$\psi_{i,j} = \exp(P(u_i, u_j)/\sigma) \tag{7}$$

其中,参数 σ 用于控制惩罚强度,$P(u_i, u_j)$ 为根据用户转发行为相似性特征(如兴趣偏好、共同关注等)确定的用户 u_i 与 u_j 的转发行为相似度。

以下对转发行为相似度 $P(u_i, u_j)$ 的求取方法进行描述。

2.2.1 用户转发行为相似性特征

在微博社交网络中,用户转发行为之间的相似关系是微博得以转发与消息得以扩散的基础,而转发行为相似性特征则用于确定用户转发行为之间的相似度。本文所采用的转发行为相似性特征如下。

(1) 主题偏好

用户 u_i 与 u_j 主题偏好越相近,则两者对微博 m 所蕴含的主题同时感兴趣的可能性越高,相应的转发行为也更可能相同(即均转发或不转发微博 m)。为了度量用户 u_i 与 u_j 之间的主题偏好相似性,类似于式(3),本文将两者的历史微博(包括原创与转发)汇集成文档,然后将相应的 LDA 模型主题分布向量之间的余弦距离值作为两者的主题偏好相似度。

(2) 相互关注与共同关注

用户 u_i 与 u_j 是否相互关注是用户之间社交关系强弱的重要体现,如果两者相互进行了关注,则两者更可能转发相同的微博。在本文中,相应的特征取值 1 时表示相互关注关系,而取值 0 时则表示单向关注关系。

此外,当两者共同关注的用户数量较多时,则两者也更可能同时转发相同的微博,相应的特征度量如式(8)所示。

$$S_{ij} = \frac{U_i \cap U_j}{U_i \cup U_j} \tag{8}$$

其中,U_i 表示用户 u_i 的所有关注用户。

(3) 相互转发与共同转发

用户 u_i 与 u_j 相互转发对方的微博较多,则表明两者更可能具有相似的主题偏好性,因而也更可能同时转发相同的微博,相应的特征度量定义如式(9)所示。

$$R_{ij} = \max(T_{ij}/T_i, T_{ji}/T_i) \tag{9}$$

其中,T_{ij} 表示用户 u_i 转发用户 u_j 的微博数,T_i 表示用户 u_i 转发的微博总数。

式(9)表明,当 R_{ij} 越高,用户 u_i 或 u_j 原创或转发的微博被用户 u_j 或 u_i 转发的可能性越高,因而两者对相同微博同时进行转发的概率也越高。

此外,用户 u_i 与 u_j 共同转发的微博数也是度量两者对相同微博表现相同转发行为的重要标识。与微博的相互转发不同,由于用户 u_i 或 u_j 所关注的用户可能较多,两者所转发的微博并不仅来自于用户 u_j 或 u_i;因而,两者共同转发的微博数量较好地度量了两者共同的兴趣与转发倾向,相应的特征度量定义如式(10)所示。

$$M_{ij} = \frac{T_i \cap T_j}{T_i \cup T_j} \tag{10}$$

其中,T_i 的定义与式(9)相同。

2.2.2 用户转发行为相似度的求取

在提取用户转发行为相似度特征之后,本文仍采用式(4)对其进行规范化处理,以生成五维规范化的特征向量,并采用与局部转发概率相似的求取方法对用户转发行为相似度进行求取。

需要注意的是,用户转发行为相似度的大小仅表明了用户之间社交关系的强度与转发行为的相似程度,而最终的用户转发行为将由式(1)预测模型中的两部分共同决定。

2.3 能量函数的求解

在实际中,式(1)所示能量函数的求解属于 NP-hard 问题,本文因此采用 Graph Cuts 算法获取其近似最优解[18]。事实上,由于用户转发行为约束与群体转发先验对应的能量项采用了 Potts 模型,因而其求解过程较为可靠。

需要注意的是,如果微博社交网络中的用户数

量较多，式(1)所示能量函数的求解复杂度可能会很高。为了解决此问题，本文采用快速的社区发现算法[19]将尺度较大的社交网络划分为多个尺度较小的子社交网络，然后再针对每个子社交网络进行相应能量函数的求解，其结果则进行合并以作为原社交网络的求解结果。由于此部分内容不是本文工作重点，在此不再赘述。

3　实验与分析

为了测试本文算法(energy-based retweet behavior predicting，ERBP)的可行性与有效性，本文主要进行了以下三方面的实验：①用户转发行为约束对预测精度的影响；用户转发行为预测结果分析与算法比较；②微博转发路径预测结果分析与算法比较。为了方便实验分析，下文将局部转发概率称为 LERBP(Local ERBP)。

本文首先采用了文献[15]公开的数据集对算法的可行性进行验证。该数据集共包含 1 787 443 个新浪微博用户的基本信息(如姓名、性别、粉丝数等)、用户最新发布的 1 000 个微博，以及用户之间的社交关系结构。

3.1　评价指标

为了评价用户转发行为预测模型的性能，本文选用了信息检索中的查全率(召回率)、准确率与 F_1 度量等三项评价标准。其中，查全率(召回率)为所有被预测为“转发”与“未转发”的用户中被正确预测为“转发”的用户所占比例；准确率为所有被预测为“转发”的用户中被正确预测为“转发”的用户所占比例；而 F_1 度量则是一个综合性指标，即：准确率×召回率×2/(准确率＋召回率)。

3.2　实验结果

根据以上实验设置，相应的实验结果与分析如下。

3.2.1　用户转发行为约束对预测精度的影响

表 1 列出了 ERBP 在不同权重 λ_1 时的预测精度。从中不难发现，当权重 λ_1 过大时，用户转发行为的个性化特征则不利于凸显，因而预测精度将受到一定的影响；否则，过于降低其关注或粉丝用户的转发行为的影响，则不利于反映社交网络中用户之间的社交关系特征，因而也不易获得较高的预测精度。

表 1　ERBP 在不同参数(λ_1)下的预测结果

结果	0.4	0.5	0.6	0.7	0.8	0.9
召回率	0.831 2	0.856 0	0.875 4	0.831 0	0.802 2	0.756 7
准确率	0.800 1	0.831 8	0.843 0	0.819 3	0.790 2	0.766 1
F_1 度量	0.815 4	0.843 7	0.858 9	0.825 1	0.796 2	0.761 4

此外，本文也对包含不同用户数量子社交网络的预测结果进行了统计。从表 2 所示的结果可知，随着用户数量的增加，ERBP 的预测精度相差不大，整体上未表现出规律性的变化。实验表明，在社交网络规模引起的计算复杂度可控的情况下(本文实验中用户数量不超过 6 万)，ERBP 具有较好的适应性。

表 2　ERBP 在不同子社交网络(用户数量)的预测结果($\lambda_1=0.6$)

结果	4 679	9 098	13 976	29 301	36 884	50 034
召回率	0.853 6	0.876 6	0.862 3	0.847 9	0.863 7	0.852 2
准确率	0.834 5	0.880 1	0.840 9	0.870 1	0.852 5	0.879 7
F_1 度量	0.843 9	0.878 3	0.851 5	0.858 9	0.858 1	0.865 7

3.2.2　用户转发行为预测

对于社交网络中用户发布的微博，本文采用不同的算法对其他用户的转发行为进行了预测。在表 3 中，SVM 与 LERBP 分别为 SVM(support vector machine)与 LERBP 采用用户属性、微博内容等基本特征的预测结果。结果表明，由于 SVM 与 LERBP 未考虑用户之间的社交关系，其整体预测精度普遍偏低，而 SVM 表现出了相对较好的性能。另一方面，将反映用户之间社交关系的粉丝数与关注用户数作为用户转发行为预测的特征分量，并不能准确地描述用户之间社交关系特征，因而其预测精度并没有本质性的提高。

表 3　用户转发行为预测($\lambda_1=0.6$)

结果	SVM	文献[16]	文献[8]	LERBP	ERBP
召回率	0.621 7	0.701 7	0.612 1	0.562 3	0.842 6
准确率	0.613 1	0.639 1	0.862 8	0.505 4	0.811 9
F_1 度量	0.631 5	0.668 9	0.716 1	0.532 3	0.827 0

相对地，文献[15]考虑了局部范围内用户之间的转发行为约束与社交网络结构的影响，因而可以更好地对用户的转发行为进行预测。事实上，在微博社交网络中，用户之间社交关系通常会引起相应

转发行为之间的相互影响，其结果甚至会改变用户自身的偏好与兴趣，导致用户转发行为趋于局部一致性。然而，文献[15]由于未考虑用户转发行为约束的全局性特征，因而难以获得更好的预测精度。同样，文献[8]虽然采用多任务学习方法与用户转发行为相似性特征，以突出不同用户转发行为的个性化差异，但由于未考虑更多用户转发行为之间的影响，因而也未能获得更高的预测精度。

相对而言，ERBP 在马尔科夫随机场框架下能较好地融合用户转发行为特征与用户转发行为约束，不但有利于突出不同用户转发行为的个性化差异，而且有利于刻化社交网络中更多用户转发行为的共同特性，进而可获取全局最优化的预测结果。

3.2.3 微博转发路径预测

为了验证本文算法在社交网络消息扩散中的性能，本文也对微博转发路径进行了预测。在本文中，对于当前微博，在被用户逐级转发直至遇到不转发情况时所经历的所有用户定义为一条微博转发路径，相应的转发用户数量则定义为该微博转发路径的长度。对于当前微博转发路径的预测，只有当其中每个用户的转发行为以及最后不转发微博的用户的行为均被准确预测时，则表明被预测成功。

如表 4 所示，相对于传统采用级联方式对微博转发路径进行预测的算法[5]，ERBP 由于综合考虑了用户属性、微博内容等特征以及用户转发行为约束的影响，因而获得了更好的预测结果。

表 4 微博转发路径预测($\lambda_1=0.6$)

路径长度	文献[5]	ERBP
2	0.506 3	0.791 6
3	0.347 6	0.611 8
4	0.248 1	0.544 4
5	0.194 6	0.447 5
6	0.107 3	0.324 4

总体上，基于马尔科夫随机场的用户转发行为预测模型可以较好地对用户属性与微博内容等特征、用户转发行为约束等因素进行描述，整体上具有较高的性能。

4 结论

为了提高微博用户转发行为预测的精度，本文将用户转发行为预测问题转化为马尔科夫框架下能量优化问题进行求解。其中，能量函数综合描述了用户转发行为特征、用户转发行为约束等因素对用户转发行为的影响，不但可突出不同用户转发行为的个性化差异，而且可刻化社交网络中更多用户转发行为的共同特性及微博转发的本质特点。实验结果表明，本文算法在对用户转发行为与微博转发路径的预测中均表现出较好的性能。

当前，本文算法的缺点与改进之处在于：当社交网络规模较大时，本文仅采用社区发现技术将其划分为多个子社交网络；然而，这些子社交网络规模的可控性以及相应能量函数求解的效率仍需要做进一步探讨。此外，本文能量函数中仅采用了易于求解的 Potts 模型，其对用户转发行为特征的描述可能存在一定的局限性，因而如何构造更有效且易于求解的能量函数也是一个值得深入探讨的问题。

参考文献

[1] Pezzoni F, An J, Passarella A, et al. Why do I retweet it? An information propagation model for microblogs [J]. Social Informatics. Springer International Publishing, 2013, 8238: 360-369.

[2] Yang J, Leskovec J. Modeling information diffusion in implicit networks [C]//Proceedings of IEEE International Conference on Data Mining. IEEE Computer Society, 2010: 599-608.

[3] Huang D, Zhou J, Yang F, et al. Understanding retweeting behaviors in twitter[J]. Journal of Computational Information Systems, 2015, 11 (13): 4625-4634.

[4] 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 the Social Computing, IEEE International Conference on Privacy, Security, Risk and Trust, 2010: 177-184.

[5] 曹玖新，吴江林，石伟，等. 新浪微博网信息传播分析与预测[J]. 计算机学报，2014，37(4)：779-790.

[6] 张旸，路荣，杨青. 微博客中转发行为的预测研究[J]. 中文信息学报，2012，26(4)：109-114.

[7] Hong L, Dan O, Davison B D. Predicting popular messages in twitter [C]//Proceedings of the 20th International Conference Companion on World Wide Web, 2011: 57-58.

[8] Tang X, Miao Q, Quan Y, et al. Predicting individual retweet behavior by user similarity: A multi-task learning approach [J]. Knowledge Based Systems, 2015, 89(C): 681-688.

[9] Zaman T R, Herbrich R, Gael J V, et al. Predicting in-

formation spreading in twitter [J]. Computational Social Science and the Wisdom of Crowds Workshop, 2010,104(45): 17599-17601.

[10] 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, AAAI Press, 2011: 586-589.

[11] Xu Z,Zhang Y,Wu Y,et al. Modeling user posting behavior on social media[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development In information Retrieval. ACM, 2012: 545-554.

[12] Yang Z,Guo J,Cai K,et al. Understanding retweeting behaviors in social networks. [C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010: 1633-1636.

[13] 周沧琦,赵千川,卢文博. 基于兴趣变化的微博用户转发行为建模[J]. 清华大学学报(自然科学版), 2015,55(11): 1163-1170.

[14] Peng H K,Zhu J,Piao D,et al. Retweet modeling using conditional random fields [C]. Proceedings of the Data Mining Workshops,2011: 336-343.

[15] Zhang J,Tang J,Li J,et al. Who influenced you? predicting retweet via social influence locality [J]. ACM Transactions on Knowledge Discovery from Data, 2015,9(3): 1-26.

[16] Wang Z,Chen E,Liu Q,et al. Maximizing the coverage of information propagation in social networks [C]//Proceedings of International Conference on Artificial Intelligence. AAAI Press,2015: 2104-2110.

[17] Blei D M,Ng A Y,Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[18] Kohli P,Ladick L,Torr P H S. Robust higher order potentials for enforcing label consistency [J]. International Journal of Computer Vision, 2009, 82(3): 302-324.

[19] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review e Statistical Nonlinear and Soft Matter Physics, 2004, 69(6): 066133.

王宁(1982—),硕士,助教,主要研究领域为大数据,网络化嵌入式计算。
E-mail: wnpet@qq.com

高光(1987—),硕士,助教,主要研究领域为大数据与云存储。
E-mail: 1041220021@qq.com

柴争义(1978—),博士,副教授,主要研究领域为大数据网络化嵌入式计算。
E-mail: 1551960034@qq.com