基于能量优化的微博用户转发行为预测

王 伟,张效尉,任国恒,秦东霞,刘琳琳
(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,在舆情监控、市场营销与政治选举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,本文在 MRF(Markov Random Field)能量优化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、群体转发先验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,而且也受到用户转发行为约束、群体转发先验等因素不同程度的影响.相对于传统算法,本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 **文章编号:** 0372-2112 (2017)09-2987-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.12.022

Predicting Microblog User Retweet Behaviors Based on Energy Optimization

WANG Wei, ZHANG Xiao-wei, REN Guo-heng, QIN Dong-xia, LIU Lin-lin
(*School of Network Engineering, Zhoukou Normal University, Zhoukou, Henan* 466000, *China*)

Abstract: Predicting user retweet behaviors is the basis of building the information diffusion model in micorblog social networks, and also is applied for a variety of fields such as public opinion monitoring, viral marketing, political campaign etc. In order to improve the accuracy of predicting user retweet behaviors, under the MRF (Markov Random Field) framework, the paper comprehensively analyzes the effects caused by user attributes, microblog contents, the constraints between user retweet behaviors and the group retweet priors, and constructs the corresponding energy function based on the logistic regression model to globally predict user retweet behaviors. Experimental results show user retweet behaviors not only depend on user attributes, and micorblog contents, but also are influenced by the constraints between user retweet behaviors and the group retweet priors in varying degrees. Compared to the traditional methods, our proposed method can accurately model user retweet behaviors and thus achieve satisfactory results.

Key words: sina microblog; retweet predicting; energy optimization; logistic regression

1 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.尤其是微博社交网络,由于其信息扩散的快速性、用户操作的便捷性以及荷载媒体的多样性(如文本、图像、视频等),逐渐成为人们分享身边新闻与社会动态的主要渠道.用户在微博社交网络中产生的海量数据蕴含着其潜在的行为模式(如用户对感兴趣主题的评论与转发)与情感因素(如对社会现象表现出愤怒、仇恨等情绪),因而,根据微博社交网络历史数据,有效地分析影响用户转发行为的特征并对其未来的转发行为

收稿日期:2016-06-30;修回日期:2016-11-16;责任编辑:梅志强
基金项目:国家自然科学基金(No. U1404620, No. U1404622);河南省自然科学基金(No. 162300410347);河南省科技攻关项目(No. 172102310727, No. 162102310589, No. 162102210396, No. 162102310590);河南省高校重点科研项目(No. 17A520018, No. 17A520019, No. 15A520116, No. 16B520034, No. 16A520105);周口师范学院高层次人才科研启动基金(No. zknuc2015103)

进行预测,不但有助于挖掘用户的兴趣与情感偏向,从而可为用户提供更准确的推荐服务(如主题、商品推荐),而且有助于理解消息在微博社交网络中的扩散机制以建立可靠的消息扩散模型,这在舆情监控、企业辅助决策等领域也有着广泛的应用.

在对微博用户转发行为进行预测时,除用户属性、微博内容等特征之外,反映用户之间关系的社交网络结构往往也对预测精度产生较大的影响.在此情况下,传统转发行为预测模型通常存在以下问题:(1)仅采用用户属性、微博内容等特征进行预测,而未考虑社交网络结构对预测精度的影响,其精度通常较低;(2)将反映社交网络结构的相关特征(如粉丝数、关注用户数)作为预测用户转发行为时的特征分量,难以体现社交网络结构对用户转发行为预测的实际作用;(3)将用户之间社交关系转化为预测用户转发行为时的约束,但未考虑用户之间的社交关系类型(如单向关注、相互关注等)以及更多用户的转发行为所构成的群体转发先验,其预测精度往往不易得以进一步的提高.

针对以上问题,本文提出了基于 MRF 能量优化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与群体转发先验,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,本文算法可以有效解决传统算法中存在的问题,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 MRF 能量优化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、用户转发行为约束与群体转发先验等信息对用户转发行为进行全局性预测,有效地提高了整体预测精度.

2 相关工作

微博社交网络消息扩散的研究在舆情监控、企业与政府辅助决策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[1~3]的构建具有重要作用.

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[4]探讨了影响用户转发行为的各种因素,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[5]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.张旸等[6]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大小.Hong 等[7]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.Tang[8]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.

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 Xu[9]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;Yang[10]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;周沧琦等[11]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.

在社交网络中,消息以发布者为中心向四周扩散,与消息相关联(如接收与转发)的用户对消息的转发行为通常主要受与其直接或间接相连的局部用户的影响(即服从 80-20 规律).在此基础上,Peng[12]探索了 Twitter 社交网络用户转发行为中的 Markov 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.该算法与本文算法较为相关,但该算法在转发行为预测模型中仅考虑了两个存在关注关系的用户之间的转发行为约束,而未考虑当前用户的社交圈内更多用户的转发行为所构成的群体转发先验.Zhang[13]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;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[14]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.

在实际中,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(如用户转发行为约束、群体转发先验等)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.与其不同,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与群体转发先验等因素统一在 MRF 能量优化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.

3　问题描述与分析

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(即 $u_i\in U$ 表示用户,$(u_i,u_j)\in E$ 表示用户 u_i 与 u_j 之间的社交关系).对于新发布的微博 m,以 $y_i\in\{0,1\}$ 表示用户 u_i 的转发行为标记(即取值 1 与 0 分别表示转发与不转发微博 m),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 U,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,本文着重从以下几方面展开讨论:

(1)用户转发行为特征

用户转发行为特征主要包括用户属性(如性别、爱好等)、微博内容(如微博内容与用户历史微博的相似度)等基本特征,是采用逻辑回归、支持向量机等传统算法对用户转发行为进行预测的基础.

(2)用户转发行为约束

除用户转发行为特征外,用户转发行为约束往往会对预测精度产生较大的影响.如图 1 所示,用户 u_i 的关注用户 u_1、u_3 与 u_5(灰色圆)均转发了微博 m,则用户 u_i 更可能受其影响而转发微博 m;而在用户 u_i 所有关注用户中,转发微博 m 者的数量越高,则用户 u_i 转发微博 m 的概率也将越高(如用户 u_2 也转发微博 m,则用户 u_i 转发微博 m 的概率将更大).另一方面,由于用户之间兴趣、爱好的差异,用户 u_i 关注用户的转发行为对用户 u_i 转发行为的影响程度也并不相同.如图 1 所示的由不同粗细线条表示的转发行为相似度,相对于用户 u_1 或 u_3,用户 u_5 与 u_i 的转发行为相似度(即转发行为标记连线上的数字)更高,因而,用户 u_5 的转发行为对用户 u_i 的转发行为的影响也将更大.

然而,在传统预测模型中,用户转发行为约束往往并未得以充分的考虑,或者仅将用户 u_i 的关注或粉丝用户的数量作为用户转发行为特征分量对待,这往往不利于刻化这些用户(如 u_1、u_3 与 u_5)的转发行为对用户 u_i 转发行为影响的差异,利用此特征对用户 u_i 的转发行为进行预测的可靠性通常也难以得到保证.

(3)群体转发先验

根据用户转发行为约束,当用户 u_i 转发微博 m 后,用户 u_2 与 u_4 的转发行为也将受到影响.事实上,由于这些用户之间相互关注关系的存在,用户 u_2 与 u_4 的转发行为不仅将受到与其直接相连用户转发行为的影响(如用户 u_4 与 u_3、u_i 直接相连),而且也将受到与其间接相连用户转发行为的影响(如用户 u_4 与 u_1、u_5 间接相连);或者说,微博社交网络中用户转发行为之间具有全局性相互约束的特性(如灰色区域内的转发行为趋于相同).在实际中,此情况是较为常见的,如用户 u_i 尽管对微博 m 的内容并不感兴趣,但由于其社交圈中大量关注与粉丝用户皆转发了微博 m,则用户 u_i 也可能由于社交活动的需要或其他主观原因而转发微博 m.为了方便算法描述,本文将用户 u_i 的社交圈用户(包括直接与更多间接用户)转发行为构成的潜在影响称为群体转发先验.

与用户转发行为特征、用户转发行为约束等因素类似,群体转发先验对用户转发行为的预测也有着重要的影响;然而,传统预测模型在对用户转发行为进行预测时往往忽略了此项因素,因而不易获得较好的预测精度.

根据以上分析,为了更可靠地对用户的转发行为进行预测,本文在 MRF 能量优化框架下对相关问题进行求解,其主要原因在于:(1)社交网络在结构上与随机场类似;(2)用户对微博的转发行为具有 Markov 性质[13,14];(3)MRF 可同时利用用户转发行为特征及不同用户转发行为之间的依赖关系对当前用户的转发行为进行预测.

事实上,已知微博 m 的特征 M 及用户集合 U 对应的特征 $X=\{x_i\}(i=1,\cdots,n)$,最优标记集 Y^* 的求取问题可表示为:

$$Y^*=\arg\max_Y P(Y|M,X) \tag{1}$$

由于用户的转发行为通常仅被以其为中心的局部用户(包括关注与粉丝用户)的转发行为所影响[12],因而具有 Markov 性质.根据 Gibbs 分布[15],$P(Y|M,X)$ 可进一步表示为:

$$P(Y|M,X)=\frac{1}{Z}\exp\Big(-\sum_{c\in C}\Phi_c(Y_c)\Big) \tag{2}$$

在式(2)中,C 表示随机场中所有基团的集合,其中的每个基团 $c\in C$ 由多个关注与粉丝用户构成;$\Phi_c(Y_c)$ 为定义在当前基团 $c\in C$ 上的势能函数;Z 为归一化常数.

根据式(2),相应的能量函数可定义为:

$$E(Y)=-\log P(Y|M,X)-\log Z=\sum_{c\in C}\Phi_c(Y_c) \tag{3}$$

因而,式(1)所示问题可转化为以下能量优化问题:

$$Y^*=\arg\min_Y E(Y) \tag{4}$$

由式(3)可知,不同的基团 $c \in C$ 描述了社交网络中不同的局部结构,而相应的势能函数 $\Phi_c(Y_c)$ 则刻化了用户转发行为特征、用户转发行为约束、群体转发先验等因素对当前用户转发行为的影响. 如图 1 所示,对应于单用户的基团及相应的势能函数描述了用户转发行为特征对用户转发行为的作用大小(为方便算法描述,下文将其称为用户的局部转发行为,而由局部转发行为决定的用户转发微博的概率简称为局部转发概率,如 u_1 的局部转发概率为 0.95);对应于两用户的基团及相应的势能函数描述了用户转发行为约束(如 y_i 与 y_1 之间的转发行为相似度为 0.87);对应于更多用户的基团(或社交圈)及相应的势能函数则描述了群体转发先验(如灰色区域内的转发行为趋于相同).

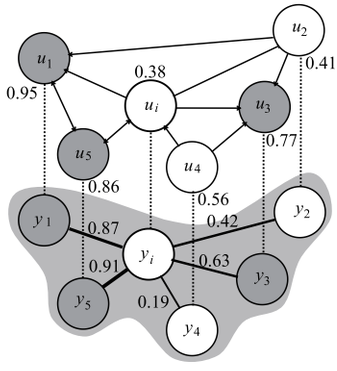


图1 MRF能量优化框架下的用户转发行为预测模型

在 MRF 能量优化框架下对用户转发行为进行预测可以较好地对相关问题进行描述,而能量函数设计与求解则是其中的关键,以下章节着重对其进行阐述.

4 基于能量优化的用户转发行为预测模型

根据以上分析,针对 MRF 能量优化框架下的用户转发行为预测问题,本文定义以下能量函数:

$$E(\boldsymbol{Y}) = \sum_{i=1}^{n} D_m(y_i,u_i) + \lambda_1 \cdot \sum_{j\in N(i)} \varphi_{i,j} \cdot \delta(y_i \neq y_j) + \lambda_2 \cdot \sum_{c\in G_i^\tau} \varphi_c(y_c) \tag{5}$$

在式(5)中,构成能量函数的三部分主要用于描述用户的局部转发行为、用户转发行为约束与群体转发行为先验;$N(i)$ 表示与用户 u_i 存在直接关注关系用户的序号集合;G_i^τ 为用户 u_i 对应 τ-ego 网络[13]的集合,其中的参数 τ 用于控制网络 τ-ego 网络的尺度.

4.1 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,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,相关度量 $D_m(y_i,u_i)$ 定义为:

$$D_m(y_i,u_i) = |y_i - P(u_i,m)| \tag{6}$$

其中,$y_i \in \{0,1\}$ 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,其求取方法描述如下.

4.1.1 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与群体转发先验,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[5],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,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[16]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,即:

$$L(d_i,m) = \frac{\mathrm{LDA}(d_i) \cdot \mathrm{LDA}(m)}{\|\mathrm{LDA}(d_i)\| \cdot \|\mathrm{LDA}(m)\|} \tag{7}$$

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@信息等信息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,即:

$$f' = \frac{f - f_{\min}}{f_{\max} - f_{\min}} \tag{8}$$

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,如表 1 所示,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.

表 1　用户转发行为特征

特征序号	特征类别	特征名称
1	用户属性	关注用户数
2	用户属性	粉丝数
3	用户属性	是否认证
4	用户属性	发布微博数
5	用户属性	被转发微博数
6	用户属性	转发活跃度
7	微博内容	内容相似度
8	微博内容	被转发的次数
9	微博内容	微博内容长度
10	微博内容	是否包含 URL
11	微博内容	是否包含@

4.1.2　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 $P(u_i, m)$ 为:

$$P(u_i, m) = \frac{1}{1 + \exp(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)} \tag{9}$$

其中,$\boldsymbol{w}$ 为特征权重向量,在本文中通过梯度下降算法最小化特定损失函数获取,即:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} (L(y_i, P(u_i, m)) + \lambda_3 \cdot \|\boldsymbol{w}\|_2^2) \tag{10}$$

在式(10)中,损失函数 $L(\cdot)$ 采用交叉熵损失函数,N 为样本数量,$\|\cdot\|$ 为 L_2 范式正则化项,λ_3 则为控制正则化强度的参数.

4.2　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将越高,实际的转发行为更可能相同;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 $\varphi_{i,j}$ 定义为:

$$\varphi_{i,j} = \exp(P(u_i, u_j)/\sigma) \tag{11}$$

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,其求取方法描述如下.

4.2.1　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(7),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 表示相互关注关系,而取值 0 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,则两者更可能同时转发相同的微博,相应的特征度量为:

$$S_{ij} = \frac{U_i \cap U_j}{U_i \cup U_j} \tag{12}$$

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为:

$$R_{ij} = \max\left(\frac{T_{ij}}{T_i}, \frac{T_{ji}}{T_j}\right) \tag{13}$$

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 转发微博的总数.

式(13)表明,当 R_{ij} 越高时,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为:

$$M_{ij} = \frac{T_i \cap T_j}{T_i \cup T_j} \tag{14}$$

其中,T_i 的定义与式(13)相同.

4.2.2 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(8)对其进行了规范化处理以生成5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了求取.

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,而最终用户的转发行为将由式(5)所示预测模型中的三部分共同决定.

4.3 群体转发先验

微博 m 在社交网络转发过程中,当前用户往往会在以其为中心的局部范围内发现一个或多个已转发微博 m 的用户(包括其直接与更多间接用户);根据群体转发先验(参见第3节),这些用户的转发行为将潜在地对当前用户转发微博 m 的主观意愿产生一定的影响. 为了可靠地对用户的转发行为进行建模,对于描述群体转发先验的能量项 $\varphi_c(y_c)$,本文采用了鲁棒的 P^n-Potts 模型[17]:

$$\varphi_c(y_c)=\begin{cases}\dfrac{\rho}{Q}\cdot\lambda_{\max}, & \rho\leqslant Q\\ \lambda_{\max}, & \text{otherwise}\end{cases} \tag{15}$$

其中,常数 $\lambda_{\max}$ 为群体转发先验惩罚量,ρ 为网络 G_i^τ 中两两用户转发行为相似度概率 $P(u_i,u_j)$ 的均值,Q 为网络 G_i^τ 中所有对微博 m 局部转发概率小于指定阈值 ζ 的用户所占比例,即:

$$Q=\frac{\sum_{v\in G_i^\tau}\delta(P(v,m)<\zeta)}{|G_i^\tau|} \tag{16}$$

在式(16)中,$|G_i^\tau|$ 为网络 G_i^τ 中用户数量;$\delta(\cdot)$ 为指示函数,参数为真时取值1,否则取值0.

由式(15)可知,当网络 G_i^τ 中的用户转发行为相似度不变时,局部转发概率较小的用户越多,则应给予较小的惩罚量以鼓励网络 G_i^τ 中的用户取不同的转发标记;否则则应给予较大的惩罚量以鼓励网络 G_i^τ 中用户取相同的转发标记;另一方面,当网络 G_i^τ 中用户局部转发概率不变时,用户转发行为相似度越高,应给予较大的惩罚量以鼓励网络 G_i^τ 中的用户取相同的转发标记,否则则应给予较小惩罚以鼓励网络 G_i^τ 中用户取不同的转发标记.

4.4 能量函数的求解

在实际中,式(5)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[17]获取其近似最优解. 需要注意的是,如果微博社交网络中的用户数量较多,式(5)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[18]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. 由于此部分内容非本文工作重点,在此不再赘述.

5 实验与分析

为了测试本文算法(简称 ERBP, Energy-based Retweet Behavior Predicting)的可行性与有效性,本文主要进行了以下实验:(1)用户转发行为约束、群体转发先验、τ-ego 网络尺度等因素对用户转发行为预测精度的影响;(2)用户转发行为预测结果分析与算法比较;(3)微博转发路径预测结果分析与算法比较. 为方便实验分析,下文将局部转发概率称为 LERBP(Local ERBP),将仅包含局部转发行为与转发行为约束的用户转发行为预测模型称为 PERBP(Pairwise ERBP).

5.1 实验数据

本文首先采用了文献[13]公开的数据集(简称 $D1$)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1000个微博以及用户之间的社交关系. 此外,为了进一步验证算法的有效性,本文也通过新浪微博的 API 开放接口获取了以微博转发深度为特征的数据集(简称 $D2$);在此过程中,相应的爬虫程序首先随机选择10,000条热门微博作为种子,然后采取深度优先的方式持续抓取每个种子微博的所有转发用户以及每个转发用户的粉丝与关注用户,最终共获取1,132,145个用户的基本信息与社交关系.

为了缓解用户数量过多引起的计算复杂度过高的问题(参见第4.4节),在对每个数据集对应的用户转发行为进行预测时,本文首先采用文献[18]将相应的社交网络划分为多个子社交网络(数据集 $D1$ 与 $D2$ 对应的社交网络分别划分为43与31个子社交网络),然后在每个子社交网络完成用户转发行为的预测,而最终原社交网络预测结果则为所有子社交网络预测结果的平均.

5.2 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 $F1$ 度量等3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 $F1$ 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).

5.3 参数设置

本文算法在所有实验中采用相同的参数配置. 其中,参数 σ 用于控制实际具有相似转发行为的两用户在能量优化过程未分配相同转发标记时的惩罚量(参

见 4.2 节),我们在实验中发现,其在区域[1,10]中取值时,预测结果基本不变,因而将其设置为 5. 在 P^n-Potts 模型(参见 4.3 节)中,局部转发概率阈值 ζ 与群体转发先验惩罚量 λ_{max} 共同决定了相应能量项的取值(特别地,当两者同时增大或减小时,相应能量项的取值可能保持不变),我们因此将阈值 ζ 设置为概率值的中间值 0.5 后再确定 λ_{max} 的取值. 对于群体转发先验惩罚量 λ_{max}、用户转发行为约束与群体转发先验权重 λ_1 与 λ_2,本文建议采用学习的方法确定其最优值;而为了简化实验过程与分析不同因素对预测精度的影响(参见第 5.4 节),本文采用遍历方式确定其近似最优值. 具体而言,我们在区域[1,10]对参数 λ_{max} 取值后再在区域[0,1]分别对 λ_1 与 λ_2 的可能取值进行遍历,同时统计相应的预测精度,最终将最高预测精度对应的参数取值作为其最优值. 实验中发现,参数 λ_1、λ_2 与 λ_{max} 分别取值为 0.6、0.3 与 4 时本文算法在多个子社交网络中均表现出较好的性能.

5.4　实验结果

根据以上实验设置,相应的实验结果与分析如下.

5.4.1　不同因素对预测精度的影响

为了分析用户转发行为约束、群体转发先验、τ-ego 网络尺度等因素对当前用户转发行为的影响,本文通过改变能量函数中的参数取值以观察数据集 *D*1 对应用户转发行为预测精度的变化.

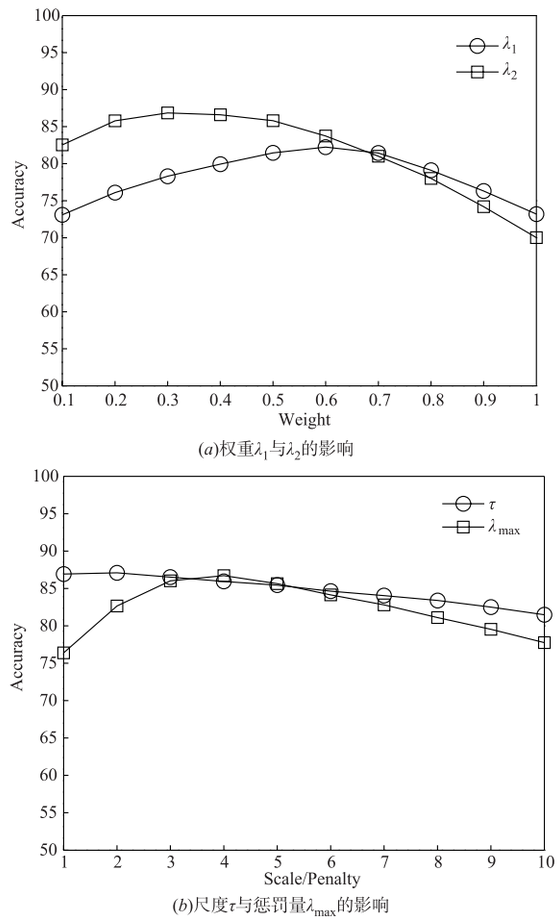
(1)用户转发行为约束的影响

本文首先在 LERBP 的基础上考察了用户转发行为约束对预测精度的影响. 用户转发行为约束表达了社交网络中具有关注关系的两用户之间的转发行为依赖性. 如图 2(*a*)所示,当权重 λ_1 逐渐增加时,预测精度呈现出先增大后降低的变化趋势,其原因主要在于:开始阶段的预测精度主要由 LERBP 确定,而随着用户转发行为约束的增大,用户转发行为的预测将趋于全局最优化,因而预测精度得以较大的提高;而当用户转发行为约束过大时,LERBP 的作用力度则相对较弱,预测精度因此而降低.

(2)群体转发先验的影响

在 PERBP 的基础上,本文考察了群体转发先验对预测精度的影响. 与用户转发行为约束类似,如图 2(*a*)与 2(*b*)所示,当权重 λ_2 或惩罚量 λ_{max} 逐渐增加时,预测精度也呈现出先增大后降低的变化趋势. 此情况表明,群体转发先验对用户转发行为的预测具有一定的影响,但相对于用户转发行为特征与用户转发行为约束等因素的影响则较弱.

表 2 列出了 ERBP 在不同权重 λ_1 与 λ_2 时的预测精度. 从中不难发现,用户转发行为约束与群体转发先验对最终预测精度都存在不同程度的影响. 然而,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. 同样,当权重 λ_2 过大或过小时,当前用户社交圈内的用户转发行为对其转发行为的潜在影响则得不到充分的体现,预测精度也将受到影响.



(*a*)权重λ_1与λ_2的影响

(*b*)尺度τ与惩罚量λ_{max}的影响

图2　不同因素对预测精度的影响

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 3 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(实验中用户数量不超过 6 万),ERBP 具有较好的适应性.

表 2　ERBP 在不同参数(λ_1,λ_2)下的预测结果(τ=1)

结果	(0.4,0.1)	(0.5,0.2)	(0.6,0.3)	(0.7,0.4)	(0.8,0.5)	(0.9,0.6)
召回率	0.8312	0.8560	0.8754	0.8310	0.8022	0.7567
准确率	0.8001	0.8318	0.8430	0.8193	0.7902	0.7661
*F*1 度量	0.8154	0.8437	0.8589	0.8251	0.7962	0.7614

表 3 ERBP 在不同子社交网络(用户数量)的预测结果($\tau=1$、$\lambda_1=0.6$、$\lambda_2=0.3$)

结果	4,679	9,098	13,976	29,301	36,884	50,034
召回率	0.8536	0.8766	0.8623	0.8479	0.8637	0.8522
准确率	0.8345	0.8801	0.8409	0.8701	0.8525	0.8797
*F*1 度量	0.8439	0.8783	0.8515	0.8589	0.8581	0.8657

(3)τ-ego 网络尺度的影响

群体转发行为通常涉及到当前用户社交圈内更多用户之间的社交关系,从图 2(b)所示实验结果中可以发现,当 τ 值取 1 与 2 时,相应的预测精度几乎不变,而之后则呈线性下降低趋势变化. 此情况表明,群体转发先验在一定程度上反映了用户转发微博的主观意愿,而其在发布或转发微博时,在很多情况下仅能发现与其直接相连用户的转发行为,因而,在小范围内考虑群体转发先验有利于提高预测精度,否则可能导致较大的预测误差.

5.4.2 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 4 中,SVM_1 与 LERBP_1 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,而 SVM_2 与 LERBP_2 则为 SVM 与 LERBP 采用不包含粉丝数与关注数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.

相对地,文献[13]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,因而可以较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而导致用户转发行为趋于局部一致性. 然而,文献[13]由于未考虑用户转发行为约束的全局性特征,因而不易获得更好的预测精度. 同样,文献[8]中算法虽然采用多任务学习方法与用户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.

与文献[13]与文献[8]不同,PERBP 在 MRF 能量优化框架下较好地融合了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. ERBP 由于在 PERBP 的基础上通过融合群体转发先验,进一步描述了用户社交圈更多用户转发行为的影响,因而更准确地反映了用户转发行为的本质特征,相应的预测精度从而可得以进一步的提高.

5.4.3 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.

表 4 用户转发行为预测($\tau=1$、$\lambda_1=0.6$、$\lambda_2=0.3$)

数据集	结果	SVM_1	SVM_2	文献[8]	文献[13]	LERBP_1	LERBP_2	PERBP	ERBP
*D*1	召回率	0.6217	0.6510	0.6121	0.7017	0.5623	0.5443	0.8426	0.8754
	准确率	0.6131	0.5944	0.8628	0.6391	0.5054	0.5289	0.8119	0.8430
	*F*1 度量	0.6315	0.6077	0.7161	0.6689	0.5323	0.5365	0.8270	0.8589
*D*2	召回率	0.5923	0.6134	0.6094	0.6743	0.5267	0.5779	0.8291	0.8621
	准确率	0.5896	0.5739	0.8512	0.6398	0.5163	0.5342	0.8057	0.8293
	*F*1 度量	0.5830	0.6013	0.7103	0.6566	0.5214	0.5552	0.8172	0.8454

在本文中,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确定了该微博的转发路径,相应的转发用户数量则定义为该微博转发路径的长度. 对于该微博转发路径,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 5 所示,相对于文献[5]中采用传统级联方式对微博转发路径进行预测的算法,ERBP 除采用用户属性、微博内容等特征外,同时也综合考虑用户转发行为约束及群体转发先验,因而获得了更好的预测结果.

表 5 微博转发路径预测($\tau=1$、$\lambda_1=0.6$、$\lambda_2=0.3$)

路径长度	预测精度			
	*D*1		*D*2	
	文献[5]	ERBP	文献[5]	ERBP
2	0.5063	0.7916	0.4693	0.7760
3	0.3476	0.6118	0.2894	0.6472
4	0.2481	0.5444	0.1749	0.4988
5	0.1946	0.4475	0.1034	0.4297
6	0.1073	0.3244	0.0795	0.2911

总体上,基于 MRF 能量优化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、群体转发先验等因素进行描述,整体上具有较高的性能.

6　总结与展望

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 MRF 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、群体转发先验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 模型与 P^n-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, Crowcroft J, Conti M. Why do I retweet it? an information propagation model for microblogs[A]. Proceedings of the 5th International Conference on Social Informatics[C]. Kyoto, Japan: Springer, 2013. 360 – 369.

[2] Yang J, Leskovec J. Modeling information diffusion in implicit networks[A]. IEEE International Conference on Data Mining[C]. Sydney, Australian: IEEE Press, 2010. 599 – 608.

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

[4] Suh B, Hong L, Pirolli P, Chi E H. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network[A]. IEEE International Conference on Privacy, Security, Risk and Trust[C]. Minneapolis, USA: IEEE Press, 2010. 177 – 184.

[5] 曹玖新,吴江林,石伟,刘波,郑啸,罗军舟. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014(4): 779 – 790.
CAO Jiu-xin, WU Jiang-lin, SHI Wei, LIU Bo, ZHENG Xiao, LUO Jun-zhou. Sina microblog information diffusion analysis and prediction[J]. Chinese Journal of Computers, 2014(4): 779 – 790. (in Chinese)

[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.
ZHANG Yang, LU Rong, YANG Qing. Predicting retweeting in microblogs[J]. Journal of Chinese Information Processing, 2012, 26(4): 109 – 114.

[7] Hong L, Dan O, Davison B D. Predicting popular messages in twitter[A]. International Conference on World Wide Web[C]. Hyderabad, India: ACM Press, 2011. 57 – 58.

[8] Tang X, Miao Q, Quan Y, Tang J, Deng K. Predicting individual retweet behavior by user similarity: a multi-Task learning approach[J]. Knowledge-Based Systems, 2015, 89(C): 681 – 688.

[9] Xu Z, Zhang Y, Wu Y, Yang Q. Modeling user posting behavior on social media[A]. International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Portland, USA: ACM Press, 2012. 545 – 554.

[10] Yang Z, Guo J, Cai K, Tang J, Li J, Zhang L, Su Z. Understanding retweeting behaviors in social networks[A]. ACM International Conference on Information and Knowledge Management[C]. Toronto, Canada: ACM Press, 2010. 1633 – 1636.

[11] 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.
ZHOU Cang-qi, ZHAO Qian-chuan, LU Wen-bo. Modeling of the forwarding behavior in microblogging with adaptive interest[J]. Journal of Tsinghua University (Science and Technology), 2015, 55(11): 1163 – 1170. (in Chinese)

[12] Peng H K, Zhu J, Piao D, Yan R, Zhang Y. Retweet modeling using conditional random fields[A]. International Conference on Data Mining Workshops[C]. Vancouver, Canada: IEEE Press, 2011. 336 – 343.

[13] Zhang J, Tang J, Li J, Liu Y, Xing C. Who influenced you? predicting retweet via social influence locality[J]. Acm Transactions on Knowledge Discovery from Data, 2015, 9(3): 1 – 26.

[14] Wang Z, Chen E, Liu Q, Yang Y, Ge Y, Chang B. Maximizing the coverage of information propagation in social networks[A]. International Conference on Artificial Intelligence[C], Buenos, Argentina: AAAI Press, 2015. 2104 – 2110.

[15] Koller D, Friedman N. Probabilistic Graphical Models-Principles and Techniques[M]. Cambridge, USA: The MIT Press, 2009.

[16] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation

[J]. Journal of Machine Learning Research, 2003, 3: 993 – 1022.
[17] Kohli P, Ladický L, Torr P H S. Robust higher order potentials for enforcing label consistency[J]. International Journal of Computer Vision, 2009, 82(3): 302 – 324.
[18] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 69(6): 066133 – 066133.

作者简介

王 伟 男,1976 年 10 月生,河南周口人. 周口师范学院副教授, 主要研究方向为计算机视觉、社交网络分析等.
E-mail: wangwei@ zknu. cn

张效尉 男,1982 年 12 月生,河南开封人. 2009 年毕业于郑州轻工业学院,获硕士学位. 目前为周口师范学院讲师,主要研究方向为数据挖掘、社交网络分析.

任国恒 男,1982 年 4 月生,河南周口人, 2011 年 6 月毕业于西安工业大学,硕士学位. 目前为周口师范学院讲师,研究方向为社交网络分析.

秦东霞 女,1983 年 7 月生,河南周口人. 2009 年毕业于重庆大学,获硕士学位. 目前为周口师范学院讲师,主要研究方向为数据挖掘、社交网络分析.

刘琳琳 女,1986 年 11 月生,河南洛阳人. 2013 年毕业上海大学,获硕士学位. 目前为周口师范学院助教. 主要研究方向为数据挖掘、社交网络分析.