物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于最大熵模型的微博传播网络中的链路预测

李勇军　尹超　于会　刘尊

Link prediction in microblog retweet network based on maximum entropy model

Li Yong-Jun　Yin Chao　Yu Hui　Liu Zun

引用信息 Citation: Acta Physica Sinica, 65, 020501 (2016)　DOI: 10.7498/aps.65.020501
在线阅读 View online: http://dx.doi.org/10.7498/aps.65.020501
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2016/V65/I2

基于最大熵模型的微博传播网络中的链路预测*

李勇军† 尹超 于会 刘尊

(西北工业大学计算机学院, 西安 710072)

(2015 年 6 月 23 日收到; 2015 年 10 月 20 日收到修改稿)

微博是基于用户关注关系建立的具有媒体特性的实时信息分享社交平台. 微博上的信息扩散具有快速性、爆发性和时效性. 理解信息的传播机理, 预测信息转发行为, 对研究微博上舆论的形成、产品的推广等具有重要意义. 本文通过解析微博转发记录来研究影响信息转发的因素或特征, 把微博信息转发预测问题抽象为链路预测问题, 并提出基于最大熵模型的链路预测算法. 实例验证的结果表明: 1) 基于最大熵模型的算法在运行时间上具有明显的优势; 2) 在预测结果方面, 最大熵模型比同类其他算法表现优异; 3) 当训练集大小和特征数量变化时, 基于最大熵模型的预测结果表现稳定. 该方法在预测链路时避免了特征之间相互独立的约束, 准确率优于其他同类方法, 对解决复杂网络中其他类型的预测问题具有借鉴意义.

关键词: 复杂网络, 微博传播网络, 链路预测, 最大熵模型

PACS: 05.10.–a, 89.75.–k, 29.85.–c **DOI:** 10.7498/aps.65.020501

1 引 言

复杂网络是近年来迅速发展的一门新兴交叉学科. 1998 年 Watts 和 Strogatz 提出的描述小世界特性的 WS 模型 [1] 以及 1999 年 Barabási 和 Albert 提出的描述无尺度特性的 BA 模型 [2] 掀起了研究复杂网络的热潮. 在研究过程中, 人们发现在真实网络中存在一些有趣的传播现象, 如计算机病毒在通信网络中的扩散、消息在社交网络中的蔓延等都可以看作是符合某种规律的事件 (或事物) 在复杂网络上的传播行为 [3], 而且这种扩散或蔓延速度迅速. 理解这些事件 (或事物) 的传播机理, 成为复杂网络研究中的热点问题之一. 微博作为复杂网络的一种具体表现形式, 近年来得到了迅速地发展, Twitter 和新浪微博是其中的典型代表. 因为微博兼具社交网络和媒体平台的特性, 其上的信息交互频繁、传播迅速, 成为反映社会舆情的主要场所之一. 理解微博上的信息传播机理、预测用户转发行为, 对研究微博上的舆情形成、产品推广等具有重要意义.

近几年, 微博上的信息传播引起了学者们的关注. 吴腾飞等 [4] 在分析传播网络结构基础上, 使用平均场理论的方法, 推导出传播网络的度分布模型. 王金龙等 [5] 定义了用户之间的相互影响力函数, 提出了一种基于用户相对权重的信息传播模型, 并对信息的传播路径及传播过程进行了分析. Wang 等 [6] 从宏观和微观两个方面对 twitter 中的 hashtags 扩散进行了分析和研究. Zhao 等 [7] 从微博数量和质量两方面研究了向关注者推荐合适微博的问题, 重点考虑了微博内容与关注者兴趣的匹配程度、微博的时效性以及推荐微博的数量. Ding 等 [8] 研究了微博信息传播的广度和深度. 上述工作主要集中在微博中的信息传播机理上, 而未涉及到具体的信息转发预测问题.

微博中的信息转发受多种因素影响, 具有随机性. 准确地预测信息转发是一个富有挑战性的工作. Luo 等 [9] 研究了预测微博转发次数与可能被浏览次数的问题, 并对有关时间序列的行为进行预测估计. Yang 等 [10] 基于社交网络的拓扑结构

* 陕西省自然科学基础研究计划 (批准号: 2014JM2-6104, 2015JM6290) 资助的课题.

† 通信作者. E-mail: lyj@nwpu.edu.cn

和转发记录抽取特征, 预测微博信息是否会被转发. Peng 等[11] 利用随机场理论预测了 twitter 中的 retweet 问题, 考虑了内容、网络结构和时间衰减等三个方面的因素. Zhao 等[12] 从用户转发行为和微博内容信息两方面着手研究了微博的转发过程, 预测微博信息被转发的次数. Hou 等[13] 从微博内容和朋友关系中抽取了九个特征并利用逻辑回归方法预测微博转发. Huang 等[14] 认为用户对微博内容感兴趣才转发, 把微博内容和用户兴趣的匹配程度引入到微博转发预测中. 这些工作主要利用微博特征预测其转发行为, 但存在准确度、复杂度、稳定性和特征广泛性不能兼顾等问题.

在利用机器学习方法预测信息转发时, 因为特征之间存在信息冗余等问题, 增加了计算复杂度, 甚至会降低计算精度. 此问题也引起一些学者的研究兴趣. Wu 等[15] 研究了用户影响力对微博中的信息转发影响, 发现粉丝数量和用户权威性对信息传播的第一步有显著影响. Wang 等[16] 指出时间和空间维度上对信息传播的影响因素, 重点研究了距离与时间对信息扩散范围的影响. Zhang 等[17] 研究了相互关注的用户对信息转发的作用, 发现互相关注的用户之间更容易转发信息. Wu 等[18] 对影响信息传播的特征进行了详细分类与描述, 但其特征范围仅限于文本类特征. 通常来讲, 为提高算法效率和有效性, 特征选择是微博转发预测中不可缺少的环节, 文献 [19, 20] 对此进行了研究.

链路预测的主要目的是推测网络节点之间存在链路的概率. 因为一条链路连接的两个节点具有一定的相似性[21,22], 目前大多数已有的链路预测算法是利用网络拓扑信息计算节点的相似性, 进而推测链路存在的可能性[23,24]. 链路预测中用到的拓扑信息可分为全局信息、局部信息和半局部信息. 基于全局信息的算法需要整个网络的拓扑结构, 随机游走[25] 是此类算法的典型代表. 共同邻居算法[26] 是基于局部信息的, 仅需要直接邻居的结构信息即可预测链路. 在利用局部随机游走算法[27] 预测链路时, 除了需要直接邻居的拓扑信息外, 还需要多层邻居的拓扑信息. 这类算法需要的信息介于全局信息与局部信息之间, 被称作半局部信息. 获取微博网络的全局拓扑信息比较困难, 而且信息转发主要受信息发布者和其关注者的影响, 微博中的信息转发预测主要基于局部信息. 除了链路预测中提到的拓扑信息外, 还用到微博内容信息和用户个人信息.

在本文中, 微博信息转发预测问题被转化为链路预测问题, 采用基于最大熵模型的分类算法预测用户的转发行为. 与其他模型相比, 最大熵模型在选择特征时具有无需额外的独立假定或内在约束等优点[28]. 本文主要研究基于关注关系的微博传播网络中的链路预测问题, 在分析影响链路预测结果的因素时, 考虑了微博信息、用户自身、用户关系三个方面的因素, 并从微博信息转发记录中抽取了对应特征. 实证研究结果表明: 1) 本文算法在运行时间上具有明显的优势; 2) 与同类其他算法相比, 本文算法在预测结果方面表现优异; 3) 当训练集大小和特征数量变化时, 本文算法的预测结果表现稳定.

2 微博传播网络与问题描述

在微博网络中, 不同信息的扩散路径不尽相同, 每条信息的扩散范围仅是微博网络的一个子集. 为了表述简便, 把微博用户的关注关系网络称为微博用户网络, 消息扩散过程中形成的网络称为微博传播网络.

2.1 微博用户网络

微博用户网络是依靠用户间的关注关系而形成的有向网络. 如图 1 所示, 用户 B 关注用户 A, 就可以看到用户 A 发布或转发的信息. 用户 A 和用户 C 相互关注, 他们发布或转发的信息互相可见.

用 $\boldsymbol{G}_{\text{user}} = (\boldsymbol{V}_{\text{user}}, \boldsymbol{E}_{\text{user}})$ 表示微博用户网络, 其中 $\boldsymbol{V}_{\text{user}} = \{u_1, u_2, \cdots, u_n\}$ 是所有用户的集合, $\boldsymbol{E}_{\text{user}} = \{l_{ij} | 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n\}$, $l_{ij} = 1$ 表示用户 u_j 关注用户 u_i, 否则 $l_{ij} = 0$. 微博信息 msg_1 在 $\boldsymbol{G}_{\text{user}}$ 上的扩散路径构成了 msg_1 的微博传播网络.

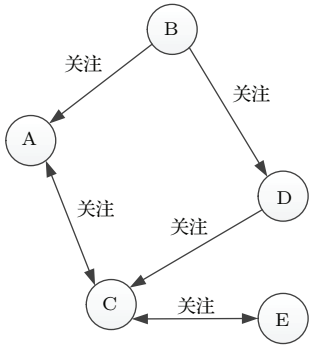

图 1 微博用户网络示例

Fig. 1. An example of user network.

2.2 微博传播网络

在如图1所示的$\boldsymbol{G}_{\rm user}$中, 用户A在$t_1$时刻发布了信息$\rm msg_1$, 用户B在$t_2$时刻转发或评论了$\rm msg_1$, 则表示$\rm msg_1$从用户A传播到了用户B, 在$\rm msg_1$的传播网络中存在一条从用户A到用户B的边. 当用户E评论了用户C转发的$\rm msg_1$时, 用户C又把用户E的评论信息进行了转发或评论时, 则在用户C和用户E之间存在一条双向边. 在研究微博信息传播时, 本文不考虑用户B看到$\rm msg_1$而未转发或评论的情况, 因为此类情况在转发数据中没有体现出来. 图2是一个$\rm msg_1$的传播网络示例. 比较图1和图2 不难发现, 在不考虑边方向的情况下, 图2是图1的一个子集, 由此可见微博用户网络是微博传播网络的基础.

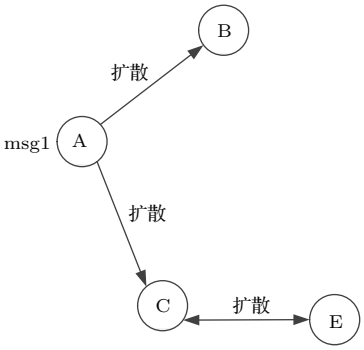

图2 微博传播网络示例

Fig. 2. An example of retweet network.

用$\boldsymbol{G}_{\rm msg}=(\boldsymbol{V}_{\rm msg}, \boldsymbol{E}_{\rm msg}, \boldsymbol{T}_{\rm msg})$表示微博传播网络, 其中$\boldsymbol{V}_{\rm msg}=\{v_1, v_2, \cdots, v_m\}\subseteq \boldsymbol{V}_{\rm user}$是信息msg在$\boldsymbol{G}_{\rm user}$上传播过程所覆盖的用户集合, $\boldsymbol{E}_{\rm msg}=\{e_{ij}|1\leqslant i\leqslant m, 1\leqslant j\leqslant m\}\subseteq \boldsymbol{E}_{\rm user}$, $e_{ij}=1$表示信息msg从用户v_i传播到了用户v_j, 否则$e_{ij}=0$. $\boldsymbol{T}_{\rm msg}=\{t_{ij}|1\leqslant i\leqslant m, 1\leqslant j\leqslant m\}$, t_{ij}表示信息msg首次传播到用户v_j的时刻. 微博传播网络与时间密切相关, 在不同时刻, 同一条微博信息的传播网络也可能存在差异. 当信息被用户转发以后, 微博传播网络中就会增加一条相应的有向边. 本文工作是预测随着时间的变化, 微博传播网络中的哪些链路会出现.

2.3 微博传播网络中链路预测的问题描述

以图3为例, 描述信息$\rm msg_1$的传播网络中的链路预测问题. 图中实线表示在当前时刻, 信息$\rm msg_1$的传播网络; 虚线表示下一时刻可能存在的传播路径. 如图1所示, 当用户D看到用户C分享的信息$\rm msg_1$后, 可能采取两种行为: 转发或评论$\rm msg_1$; 对$\rm msg_1$置之不理. 微博传播网络中的链路预测是预测用户采取前一种行为的概率. 当预测的转发概率值高于给定阈值时, 认为用户会转发或评论$\rm msg_1$, 否则视为对$\rm msg_1$置之不理. 因此, 微博传播网络中的链路预测问题可以进一步转化为二值的分类问题.

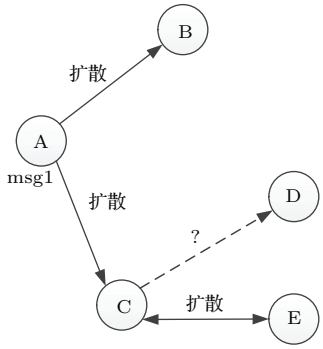

图3 微博传播网络的链路预测

Fig. 3. Link prediction in retweet network.

假设当前时刻为t, 在时间段$(t-\Delta t)$内, $\boldsymbol{G}_{\rm user}$上有$n$条信息在传播, 信息$\rm msg_{\it i}$的传播网络为$G_{\rm msgi}$. 所有信息的微博传播网络集合记为$\Lambda$. 用户$v_i$转发或评论了$n$条中的部分信息, 这些信息对应的微博传播网络集合记为Λ_i; 同理, 用户v_j参与转发的信息对应的传播网络集合记为Λ_j. 在$\boldsymbol{G}_{\rm user}$中, $l_{ij}=1$. 在时刻t用户v_i分享了信息$\rm msg_{\it k}$. 在时间段$(t+\Delta t_1)$内, 用户v_j转发或评论$\rm msg_{\it k}$的概率问题可抽象为$\boldsymbol{G}_{\rm msg\it k}$中的$e_{ij}$是否存在的预测问题. 用$p_{ij}$表示边$e_{ij}$出现的概率, 则有

$$p_{ij}=P(e_{ij}=1|\Lambda_i, \Lambda_j, G_{{\rm msg}k}, l_{ij}=1). \quad (1)$$

假设指定阈值为ζ, 引入指示函数$f(x)$可以将上述预测的概率问题转化为链路是否存在的二值分类问题, 如下所示:

$$e_{ij}=f(p_{ij})=\begin{cases}1, & p_{ij}\geqslant\zeta,\\ 0, & p_{ij}<\zeta.\end{cases} \quad (2)$$

根据微博用户的历史转发记录和关注关系, 分析影响用户转发信息的因素并提取对应特征, 然后利用机器学习方法计算转发链路存在的概率p_{ij}, 再依据(2)式判断用户是否会转发微博信息.

3 微博传播解析与特征提取

一条微博信息在微博用户网络是如何传播的? 微博用户网络和微博传播网络上有些哪些可用的数据? 在这些数据中隐含了哪些影响微博传播的因素? 以新浪微博数据为例, 但不局限于新浪微博, 围绕上述问题进行分析, 提取影响微博转发的特征.

3.1 微博用户网络和传播网络的数据解析

微博传播网络由多条传播路径组成, 如图2所示的$\boldsymbol{G}_{\rm msg1}$由(A→B)和(A→C→ E)两条传播路径组成. 这里以微博传播网络中的一条传播路径为例解释其所蕴含的信息. 一条典型的微博传播路径可以用表1中的数据格式描述.

每条微博转发数据由多个字段组成. 每个字段用$|x|$表示, $|$是字段的分界符, x是字段内容, $\#$是字段间的分隔符. x采用$m{:}n$形式表示, 其中m是字段的关键字, n是关键字的值. 表2列出了表1中各个字段的含义. 其中, 在$v_1\$v_{11}\$v_{12}\backslash tv_2\$v_{21}\backslash tv_3$中, 转发信息的用户之间用$\backslash t$分割, 涉及到三个用户, 分别是$v_3$, v_2和v_1. 用户在转发信息中提及的用户紧跟在该用户名后用$\$$分隔. 如$v_1\$v_{11}\$v_{12}$, 用户$v_1$在转发的微博信息中提及两位用户分别是$v_{11}$和$v_{12}$.

表1 微博传播路径的数据格式示例

Table 1. Data formation of retweet path.

|time:t|#|Mid:mid|#|uid:$v_1\$v_{11}\$v_{12}\backslash tv_2\$v_{21}\backslash tv_3$|#|isContainLink:$bl$|#|eventList:elist|#|rtTime:rtt|# |rtMid:rtm|#|rtUid:rtu|#|rtIsContainLink:rtbl|#|rtEventList:rtelist

表2 微博转发数据格式的字段含义

Table 2. the means of fields in data formation.

关键字	值	含义
time	t	微博的转发时间
Mid	mid	微博ID
isContainLink	bl	微博是否含链接
rtTime	rtt	微博发布的时间
rtMid	rtm	发布时的微博ID
rtUid	$v_{\rm c}$	发布微博的用户ID
rtIsContainLink	rtbl	微博发布时是否含链接
uid	$v_1\$v_{11}\$v_{12}\backslash tv_2\$v_{21}\backslash tv_3$	微博转发路径($v_{\rm c}\rightarrow v_3\rightarrow v_2\rightarrow v_1$)
eventList	elist	微博转发时提及的事件名或关键字
rtEventList	rtelist	微博发布时提及的事件名或关键字

从上述数据解析中不难看出, 微博转发数据中不仅包含了微博内容的信息, 而且包含了与时间相关的、用户间互动的信息.

微博用户网络主要描述用户之间关注关系, 通常采用如"UserID\tFolloweeID"的数据格式描述, 其中, UserID是用户ID, FolloweeID是关注UserID的用户ID. 微博用户网络含有两方面与微博转发相关的数据, 一是用户个人信息, 如关注和被关注的用户数; 二是用户之间的关注关系.

3.2 链路预测中的特征提取

微博传播网络的中链路预测受多种因素的影响, 且各种因素的影响力大小也不尽相同. 分析和提取这些影响因素是用机器学习方法进行链路预测的基础, 在机器学习中这些因素又称为特征. 基于上述对微博用户网络和微博传播网络的数据解析, 可将这些因素归为三类: 与微博相关的、与用户个人信息相关的和与用户间的关系相关的. 下文中

分别简称为微博特征、用户特征和关系特征.

微博特征主要反映微博自身的属性对转发行为的影响. 如包含热门关键字的微博被转发的概率大一些, 而发布时间较久的微博被转发的可能性会低. 此类特征仅与微博信息相关.

用户特征在一定程度上反映某位用户的属性对微博被转发的影响力. 例如被关注数量多的用户, 可能是公众人物, 其微博被转发的概率会大些. 如果一个用户极少转发或评论被关注者的微博, 那么呈现在该用户面前的微博被转发的概率就会很低. 关系特征是指微博用户网络中用户之间与关注关系相关的特征, 是预测微博被转发的关键特征. 如通常来讲, 只有粉丝用户才可能转发被关注者的微博. 互动较为频繁的用户间转发微博的概率也会大些.

表 3 特征分类及特征说明

Table 3. Feature and its description.

特征类别	特征序号	特征说明
微博特征	1	微博中提及的用户数量 (@ 的数量)
	2	微博是否包含有链接或关键字
	7	微博中提及的粉丝数量
	8	微博中提及的被关注者数量
	18	微博发布的时间
	19	微博发布时间的区间值
用户特征	5	用户发布的微博中出现链接或关键字的次数
	6	用户发布微博的数量
	10	用户粉丝数量的区间值
	11	用户的粉丝数量
	12	用户关注的用户数量的区间值
	13	用户关注的用户数量
	14	用户被提及的次数
	15	用户提及其他用户的次数
	16	用户的微博被转发的次数
	17	用户发布/分享微博的数量
	20	用户微博被转发的几率
关系特征	3	两个用户拥有共同粉丝的数量
	4	两个用户是否共同提及某个用户
	9	两个用户是否互为好友

表 3 中列出了从微博用户网络和微博传播网络中提取的所有特征. 其中特征 18 和 19 都表示微博发布时间对链路预测的影响, 然而具体时间的数值较大, 在机器学习过程会产生偏差. 为了克服此类问题, 依据时间数值的分布情况, 把时间值分在不同的区间内, 同一区间内的时间认为是无差异的. 特征 10 和 11、特征 12 和 13 的情况相同.

4 基于最大熵模型的微博传播网络中的链路预测

在 $\boldsymbol{G}_{\rm msg}$ 中, 假设 $e_{12}=1$, 则在 $\boldsymbol{G}_{\rm user}$ 中必有 $l_{12}=1$. $l_{12}=1$ 是预测链路 e_{12} 是否存在的必要条件, 但不是充分条件. 微博传播网络中的链路预测解决的问题就是, 当 $l_{12}=1$ 时, $e_{12}=1$ 还是 $e_{12}=0$? 如 (1) 和 (2) 式所述. 为描述方便, 我们用随机变量 y 表示链路预测结果, 当 $e_{ij}=1$ 时, $y=1$; 否则 $y=0$.

链路预测的结果受诸多因素影响, 如表 4 中所列的特征. 把所有这些影响因素组成向量, 记为 $\boldsymbol{X}'$. 把向量 $\boldsymbol{X}'$ 看作链路预测的输入, 则 (1) 式可变换为

$$p_{ij}=P(y|\boldsymbol{X}',l_{ij}), \tag{3}$$

l_{ij} 作为 y 的必要条件, 也是 y 取值的影响因素之一, 因此也可归并到向量 $\boldsymbol{X}'$ 中, $\boldsymbol{X}=\boldsymbol{X}'\cup\{l_{ij}\}$. (3) 式可以进一步简化为

$$p_{ij}=P(y|\boldsymbol{X}). \tag{4}$$

当 $P(y|\boldsymbol{X})$ 的值大于指定阈值时, y 值取 1; 否则为 0. (2) 式可简化为下式的形式:

$$y=\begin{cases}1, & P(y|\boldsymbol{X})\geqslant\zeta,\\ 0, & P(y|\boldsymbol{X})<\zeta.\end{cases} \tag{5}$$

基于最大熵模型链路预测的思想是, 从已有的微博传播网络集合 Λ 中获取样本数据集 $\{(\boldsymbol{X}_i, y_i)\}$, 利用最大熵模型对链路预测进行建模, 然后基于样本 $\{(\boldsymbol{X}_i, y_i)\}$ 学习每种特征对预测结果的影响权重. 为便于表示各种特征对预测结果的影响, 引入

$$f(x,y)=\begin{cases}1, & \text{如果}(x,y)\text{满足特定条件},\\ 0, & \text{否则},\end{cases} \tag{6}$$

所示的指示函数, 也称为特征函数, 其中 $x \in \boldsymbol{X}$ 表示某种特征.

为描述方便, 在不混淆的情况下, 也简称特征函数 f 为特征. 利用最大熵模型预测链路状态时, (4) 式所示的条件概率可改写为

$$P(y|\boldsymbol{X}) = Z_\lambda(x)\exp\left(\sum_i \lambda_i f_i(x,y)\right), \tag{7}$$

$$Z_\lambda(x) = \frac{1}{\sum\limits_y \exp\left(\sum\limits_i \lambda_i f_i(x,y)\right)}, \tag{8}$$

其中, Z_λ 是归一化因子; λ_i 是权重因子, 表示特征函数 f_i 的重要性.

影响链路预测的三类特征, 微博特征、用户特征和关系特征分别记为 $f_{\rm T}(x, y)$, $f_{\rm U}(x, y)$ 和 $f_{\rm R}(x, y)$. 每类特征中的特征数量分别记为 $k_{\rm T}$, $k_{\rm U}$ 和 $k_{\rm R}$, $k= k_{\rm T} + k_{\rm U} + k_{\rm R}$. 基于上述特征定义, (7) 式可以进一步表示为

$$\begin{aligned} P(y|\boldsymbol{X}) =& Z_\lambda(x)\exp\left\{\sum_i^{k_{\rm T}} \lambda_i f_{\rm T}^i(x,y)\right. \\ &+ \sum_i^{k_{\rm U}} \lambda_i f_{\rm U}^i(x,y) \\ &\left.+ \sum_i^{k_{\rm R}} \lambda_i f_{\rm R}^i(x,y)\right\}. \end{aligned} \tag{9}$$

参数集合 $\{\lambda_i, i = 1, 2, \cdots, k\}$ 可利用已有的样本数据集合进行训练求解. 在求解过程中, 为防止参数被过度学习, 通常为参数假设一个先验分布, 本文采用的分布是高斯分布.

$$P(\lambda_i) = \frac{1}{\sqrt{2\pi}\delta}\exp\left\{-\frac{(\lambda_i-\mu)^2}{2\delta^2}\right\}. \tag{10}$$

训练参数的正则对数似然函数为

$$\begin{aligned} L(\lambda) =& \sum_{x,y}\{P'(x,y)\cdot\log P(y|x)\} - \sum_i \log P(\lambda_i) \\ =& \sum_{x,y}\left\{P'(x,y)\left(\sum_{i=1}^{k_{\rm T}}\lambda_i f_{\rm T}^i(x,y)\right.\right. \\ &+ \sum_{i=1}^{k_{\rm U}}\lambda_i f_{\rm U}^i(x,y) + \sum_{i=1}^{k_{\rm R}}\lambda_i f_{\rm R}^i(x,y) \\ &\left.\left.- \log Z_\lambda(x)\right)\right\} \\ &- \sum_{i=1}^{k}\frac{(\lambda_i-\mu)^2}{2\delta^2} - k\log\sqrt{2\pi}\delta, \end{aligned} \tag{11}$$

其中 $P'(x,y)$ 是样本数据 (x, y) 的统计概率. 似然函数 $L(\lambda)$ 是凸函数, 利用

$$\frac{\partial L(\lambda)}{\partial\lambda} = 0, \tag{12}$$

求解最优参数值, 但实际中很难找到一个解析解, 一般采用基于梯度的数值优化算法进行求解, 目前常用的算法是 L-BFGS 算法 [29].

基于最大熵模型的微博传播网络中的链路预测算法主要步骤描述如下.

输入 $\Lambda = \{\boldsymbol{G}_{\rm msgi}|\ i = 1, 2, \cdots, k-1\}$, $\boldsymbol{G}_{\rm user}$, $f_{\rm T}$, $f_{\rm U}$ 和 $f_{\rm R}$, 输出特征函数的权重参数 $\{\lambda_i, i = 1, 2, \cdots, k\}$.

步骤 1 从样本数据集 Λ 和 $\boldsymbol{G}_{\rm user}$ 按照特征函数 $f_{\rm T}$, $f_{\rm U}$ 和 $f_{\rm R}$ 的定义, 计算样本数据的特征值;

步骤 2 计算每个特征的统计概率 $P'(x,y)$;

步骤 3 获取的样本特征值 $\{(\boldsymbol{X}_i, y_i)\}$ 及其相应的统计概率 $P'(x, y)$ 代入 (11) 式和 (12) 式中, 利用 L-BFGS 算法求解 $\{\lambda_i, i = 1, 2, \cdots, k\}$;

步骤 4 输出 $\{\lambda_i, i = 1, 2, \cdots, k\}$.

得到模型参数 λ=$\{\lambda_i, i = 1, 2, \cdots, k\}$ 后, 基于最大熵模型的微博传播网络的链路预测模型训练完成, 可用于预测微博信息被转发的概率. 给定阈值 ζ 后, 利用 (5) 式可以计算变量 y 的值. 当 $y = 1$ 时, 微博传播网络中的对应两个用户之间会产生一条有向链路, 即该微博信息被转发; 否则, 不会被转发.

5 实例验证

选用 2009—2012 年的新浪微博转发记录作为实验数据集, 该数据集包含了 100 万条转发记录, 涉及到 5.8×10^6 个新浪微博用户. 基于实验数据集构建微博传播网络和微博用户网络. 为检验基于最大熵模型的链路预测算法 (下文中简记为 ME) 的性能, 并使得结果比较具有广泛性, 选用支持向量机 (SVM)、朴素贝叶斯分类器 (NBC)、决策树 (DT)、随机森林 (RF)、K 最近邻分类 (KNN)、感知器 (Pre) 六种常用的分类算法进行对比实验. 在训练过程中, SVM 模型采用 SMO 算法, 核函数为 RBF; NBC 利用经典贝叶斯公式; DT 采用 C4.5 算法、RF 使用多个决策树 C4.5 算法, 为避免过度拟合, 设置置信度为 0.25 来估计剪枝后的误差; KNN 采用 K 最近邻算法, K 值设置为 1000, 距离度量采

用欧式距离; Pre采用前向传播算法, 权值参数通过训练进行自适应更新直到最优. 利用训练好的模型计算微博被转发的概率, 选取概率值大的对应类别作为输出结果. 利用10倍交叉方法验证预测结果. 在准确率(Prec)、召回率(Rec)、精确度(Acc)和F1值(F1)四个指标上对比预测结果.

5.1 不同预测算法的结果比较

七种不同分类算法基于表3中所列的20个特征在数据全集上的实验结果如图4所示.

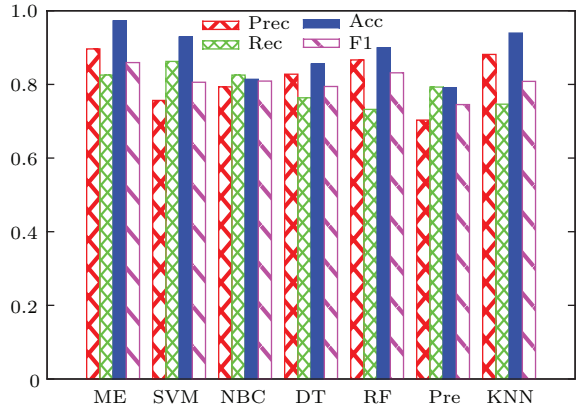

图4 (网刊彩色) 不同预测方法的结果比较

Fig. 4. (color online) Results comparison on seven different methods.

从图4的对比结果中不难看出, ME的准确率、精确度和F1值均优于其余六种预测算法, 说明利用ME算法预测链路的结果最为精确. 在召回率上, ME仅次于SVM, 但高于其他五种算法, 说明ME算法在查全方面表现也很优异. 与SVM算法比较, ME算法的准确率明显优于SVM算法的, 在精确度和F1值上也优于SVM算法.

ME, SVM, NBC和Pre四种分类算法预测结果的ROC曲线如图5所示, 其对应的AUC值分别为: 0.863,0.845,0.823和0.809. 从图5和AUC值上不难看出, ME算法优于其他分类算法.

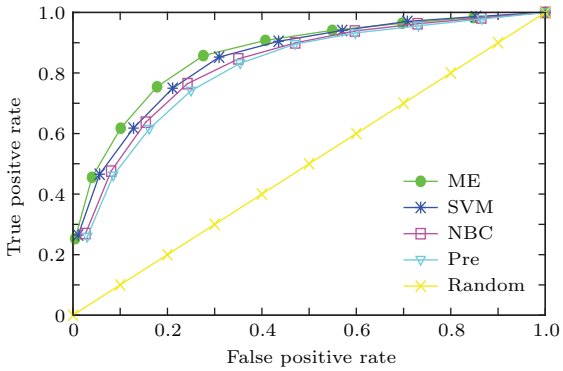

图5 (网刊彩色) 不同预测方法结果的ROC曲线对比

Fig. 5. (color online) ROC comparison on different methods.

5.2 特征选择算法对预测结果的影响

预测算法的结果受多种因素的影响, 且不同因素的影响大小也不尽相同. 基于信息增益(IG)、卡方检验(CHI)和信息增益-卡方检验(IG-CHI)三种方法对表3中所列的20种特征进行筛选. 基于筛选的特征集合, 利用不同的分类算法进行链路预测. 筛选后的特征数量相对原始特征数量减少了, 直观上分类算法的运行时间也会缩短. 为便于对比, 所有分类算法运行在如表4所示的配置机器上.

表4 实验硬件环境

Table 4. Hardware configuration in experiment.

处理器	内存	操作系统	软件开发环境
Pentium Dual-CoreCPU E5300 2.60 GHz CPU	1.99 GB 内存	Microsoft Windows XP	VC++ 6.0

利用IG, CHI和IG-CHI三种不同特征选择方法筛选后的特征集合见表5, 原始特征集合记为OR, 三种特征选择的结果集合分别记为IG, CHI和IG-CHI. 不同预测算法在三种特征集合上预测结果的准确率见表6.

从预测结果的准确率上不难发现, ME算法的运行时间显著低于SVM, NBC和KNN算法, 且运行结果准确率也明显高于SVM和NBC. 在预测结果的准确性上, ME算法略低于KNN, 但在运行时间上明显优于KNN近三个数量级. 基于IG, CHI和IG-CHI特征集合的四种预测算法的准确率与OR特征集合上的预测结果相比, 均有明显提高.

表 5 IG, CHI 和 IG-CHI 特征集合

Table 5. Feature set of IG, CHI and IG-CHI.

特征序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
OR	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
IG	■	■			■	■				■	■	■	■	■	■	■	■	■		■
CHI	■	■				■	■		■	■	■		■	■	■	■	■	■		■
IG-CHI	■	■									■		■	■	■	■	■	■		■

表 6 IG, CHI 和 IG-CHI 三种特征集合在各种机器学习算法的运行结果

Table 6. the results of different methods on three feature sets.

机器学习算法	原始特征数	被选择特征数			特征分类结果准确率%			运行时间/s		
		IG	CHI	IG-CHI	IG	CHI	IG-CHI	IG	CHI	IG-CHI
ME	20	14	14	10	93.84	93.69	93.36	0.23	0.23	0.19
SVM	20	14	14	10	93.19	93.19	93.36	25.68	25.68	15.93
NBC	20	14	14	10	84.87	84.87	84.36	1019.12	1019.12	692.43
KNN	20	14	14	10	96.75	96.75	97.36	134.37	134.37	112.64

5.3 不同特征组合对预测结果的影响

为检验预测算法在不同特征数量下的运行稳定性, 从表 3 所列的特征中利用 5.2 节中的 IG-CHI 方法按照计算结果大小分别选取前 5, 10, 15 个特征, 三种不同的特征组合见表 7.

七种不同预测算法在表 7 所示的三种不同特征集合上的预测结果如图 6 所示.

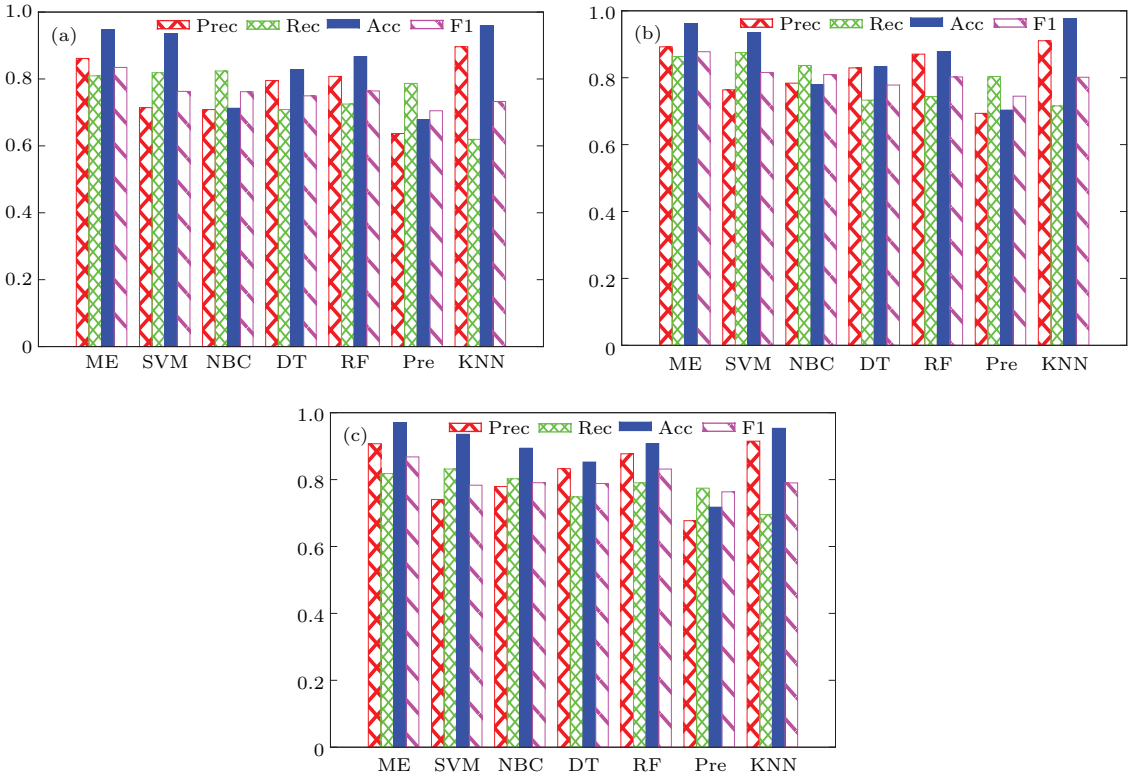

图 6 (网刊彩色) 不同特征组合在不同预测方法下的结果比较 (a) 5 个特征的运行结果; (b) 10 个特征的运行结果; (c) 15 个特征的运行结果

Fig. 6. (color online) Result comparison of seven methods on different feature set: (a) Result of 5-feature set; (b) result of 10-feature set; (c) result of 15-feature set.

表 7 利用 IG-CHI 选择的前 5, 10, 15 个特征列表

Table 7. 5-feature list, 10-feature list and 15-feature list selected by IG-CHI.

特征序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
前 5 个										■		■		■	■					■
前 10 个	■	■								■		■	■	■	■	■			■	■
前 15 个	■	■		■	■				■	■	■	■	■	■	■	■	■		■	■

从图 6 中不难看出, 在准确率和精确度上 ME 算法略低于 KNN 算法, 但在召回率和 F1 值上优于 KNN 算法. 相对于其余算法, ME 算法在四种指标上均具有优势. 在不同特征数量上, ME 算法与其余六种算法相比, 性能较稳定.

5.4 训练集大小对预测结果的影响

训练集大小对预测算法的结果有明显的影响. 为检验各种预测算法在不同大小数据集上的性能, 从原始数据集中随机抽取 25%, 50% 和 75% 组成三个不同大小的数据集. 七种预测算法在三种不同数据集上的结果如图 7 所示.

从图 7 中可以看出, ME 算法在精确度上要低于 KNN 算法, 在召回率上低于 SVM 算法, 在其余方面均具有优势. 随着数据集的逐渐变大, ME 算法相对于 KNN 算法和 SVM 算法的劣势逐渐缩小, 在数据全集上 (如图 4 所示), ME 算法优于 KNN. 说明 ME 相对其他算法在不同大小数据集上的表现较好.

综合上述实验分析结果可以得出, 在训练集合足够大的情况下, ME 算法除了在召回率上略低于 SVM 算法外, 在其余方面均具有优势, 尤其在综合指标精确度和 F1 值方面表现优异. 在运行时间方面, 相对其他算法具有明显的优势. ME 算法的运行结果在不同特征数量和不同训练集大小上的表现也相对稳定.

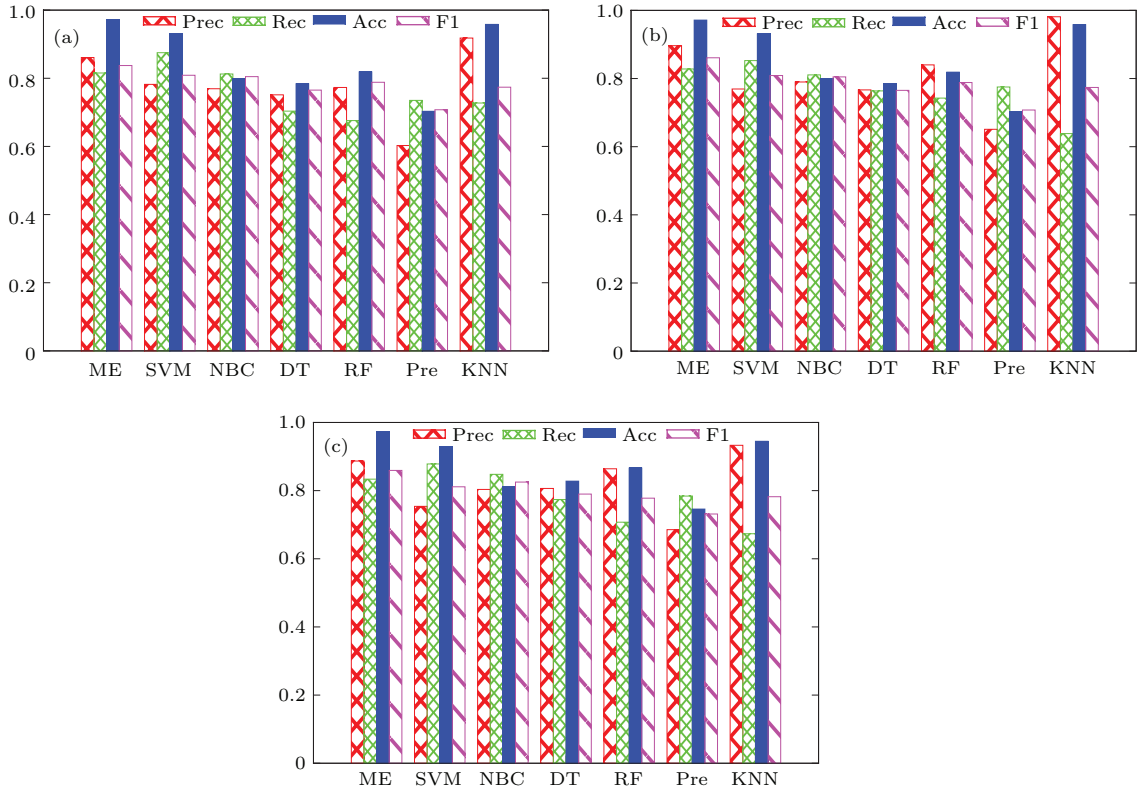

图 7 (网刊彩色) 不同大小训练集在不同预测方法下的运行结果比较 (a) 25% 的数据; (b) 50% 的数据; (c) 75% 的数据

Fig. 7. (color online) Result comparison of seven methods on different size training set: (a) 25% of original training set; (b) 50% of original training set; (c) 75% of original training set.

6 结 论

本文研究了微博网络中的信息转发问题, 并将该问题抽象为微博传播网络中的链路预测问题. 以微博信息转发数据为基础, 从微博信息、用户本身、用户关系三个方面分析了影响微博信息转发的因素, 并从微博转发记录中抽取对应的特征. 建立基于最大熵模型的微博传播网络中的链路预测模型, 预测微博被转发的概率. 利用新浪微博的数据集在特征数量、训练集大小、特征选择等方面与六种不同预测算法进行实例对比. 实验表明: 1) 在运行时间上, ME 算法具有明显的优势; 2) 在综合指标精确度和 F1 值方面 ME 算法相对其余 6 中预测算法表现优异; 3) 在不同大小数据集和不同特征数量集合上的测试结果表现稳定.

参考文献

[1] Watts D J, Strogatz S H 1998 *Nature* **393** 440
[2] Barabási A L, Albert R 1999 *Science* **286** 509
[3] Pastor S R, Vespignani A 2001 *Phys. Rev. Lett.* **86** 3200
[4] Wu T F, Zhou C L, Wang X H, Huang X X, Chen Z Q, Wang R B 2014 *Acta Phys. Sin.* **63** 240501 (in Chinese) [吴腾飞, 周昌乐, 王小华, 黄孝喜, 谌志群, 王荣波 2014 物理学报 **63** 240501]
[5] Wang J L, Liu F A, Zhu Z F 2015 *Acta Phys. Sin.* **64** 050501 (in Chinese) [王金龙, 刘方爱, 朱振方 2015 物理学报 **64** 050501]
[6] Wang Y Z, Zheng B H 2014 *Proceedings of 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* Beijing, China, Aug. 17–20, 2014, p285
[7] Zhao X Q, Tajima K 2014 *Proceedings of 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies* Warsaw, Poland, Aug. 11–14, 2014, p282
[8] Ding H Y, Wu J 2015 *Proceedings of 2015 IEEE International Conference on Multimedia Big Data* Beijing, China, Apr. 20–22, 2015, p56
[9] Luo Z L, Wang Y, Wu X T 2012 *Proceedings of the 13th International Conference on Web Information System Engineering* Paphos, Cyprus, Nov. 28–30, 2012, p777
[10] Yang Z, Guo J Y, Cai K K, Tang J, Li J Z, Zhang L, Su Z 2010 *Proceedings of the 19th ACM conference on information and knowledge management* Toronto, Canada, Oct. 26–30, 2010, p1633
[11] Peng H K, Zhu J, Piao D Z, Yan R, Zhang Y 2011 *Proceedings of IEEE 11th International Conference on Data Mining Workshops* Vancouver, Canada, Dec. 11, 2011, p336
[12] Zhao H D, Liu G, Shi C, Wu B 2014 *Proceedings of 2014 IEEE International Conference on Data Mining Workshop* Shenzhen, China, Dec. 14, 2014, p952
[13] Hou W, Huang Y, Zhang K 2015 *Proceedings of IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing* Beijing, China, Jul. 6–8, 2015, p255
[14] Huang D X, Zhou J, Mu D J, Yang F S 2014 *Proceedings of 7th International Symposium on Computational Intelligence and Design* Hangzhou, China, Dec. 13–14, 2014, p30
[15] Wu Y, Hu Y, He X H, Deng K 2014 *Chin. Phys. B* **23** 060101
[16] Wang F, Wang H Y, Xu K 2012 *Proceedings of IEEE ICDCS Workshop on Peer-to-Peer Computing and Online Social Networking* Macau, China, Jun. 18–21 2012, p133
[17] Zhang L M, Pei J, Jia Y, Zhou B, Wang X 2014 *Proceedings of 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* Beijing, China, Aug. 17–20, 2014, p208
[18] Wu Z X, Liao J X, Zhang L J 2013 *Proceedings of 5th IEEE Conference on Broadband Network & Multimedia Technology* Guilin, China, Nov. 17–19, 2013, p119
[19] Suh B, Hong L C, Pirolli P, Chi E D H 2010 *Proceedings of The 2010 IEEE International Conference on Privacy, Security, Risk and Trust* Minneapolis, USA, Aug. 20–22, 2010, p177
[20] Xu Z H, Yang Q 2012 *Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* Istanbul, Turkey, Aug. 26–29, 2012, p46
[21] Leicht E A, Holme P, Newman M E J 2006 *Phys. Rev. E* **73** 026120
[22] Liben-Nowell D, Kleinberg J 2007 *J. Am. Soc. Inf. Sci. Tec.* **58** 1019
[23] Lü L Y, Zhou T 2011 *Physica A* **390** 1150
[24] Bai M, Hu K, Tang Y 2011 *Chin. Phys. B* **20** 128902
[25] Brin S, Page L 1998 *Comput. Netw. ISDN Syst.* **30** 107
[26] Lorrain F, White H C 1971 *J. Math. Soc.* **1** 49
[27] Liu W P, Lü L Y 2010 *Europhys. Lett.* **89** 58007
[28] Berger A L, Pietra S, Pietra V 1996 *Comput. Linguist.* **22** 39
[29] Byrd R H, Nocedal J, Schnabel R B 1994 *Math. Program.: Series A and B* **63** 4

Link prediction in microblog retweet network based on maximum entropy model*

Li Yong-Jun† Yin Chao Yu Hui Liu Zun

(*School of Computer, Northwestern Polytechnical University, Xi'an 710072, China*)

(Received 23 June 2015; revised manuscript received 20 October 2015)

Abstract

Microblog is a social media platform, based on the follower-followee relationship, that enables users to share real-time information, by which the information propagation is characterized as rapid, explosive, and immediate. The research on the information propagation and retweet prediction is very important for public sentiment analysis and product promotion. A majority of existing works adopt several traditional prediction methods to predict the future information retweet based on the features extracted from existing retweet behaviors, which are hard to reconcile accuracy, complexity, robustness and feature extensiveness. To overcome the above mentioned shortcomings in existing works, we propose in this paper a link prediction algorithm based on maximum entropy model to predict retweet behavior on microblog. In our proposed approach, firstly we abstract the retweet prediction problem to a link prediction problem. Then we analyze the retweet behaviors on microblog and determine the factors influencing the retweet behavior. We extract the features from the retweet behaviors based on these factors in the next step. Now based on these features, the retweet behavior could be predicted by the proposed approach. However, information redundancy and other issues may exist among these features. These issues will cause an increase in computational complexity or a decrease in computational accuracy. To solve the above problems, we selecte the features dominating the retweet behavior with feature selection methods such as Information Gain, IG-CHI. The proposed model requires no further independent assumption in features or intrinsic constraints, and omits the processing in relation to features, which is usually the prerequisite of other prediction methods. We take the Sina Weibo retweet records in a time span from 2009 to 2012 as an example to test the effectiveness and efficiency of our link prediction algorithm. Results show that: 1) the proposed algorithm has incomparable advantages in running time; 2) as for the predicted result, the proposed algorithm is better than other algorithms in performance evaluations; 3) the proposed algorithm runs stably for different sizes of training sets and feature sets; 4) the accuracy of the predicted results remains stable based on the selected features. The proposed approach avoids the independent restriction among features and shows better accuracy than other similar methods, thus it has reference values for resolving other prediction problems in complex networks.

Keywords: complex network, microblog network, link prediction, maximum entropy

PACS: 05.10.–a, 89.75.–k, 29.85.–c **DOI:** 10.7498/aps.65.020501

* Project supported by the Shaanxi Provincial Natural Science Foundation, China (Grant Nos. 2014JM2-6104, 2015JM6290).

† Corresponding author. E-mail: lyj@nwpu.edu.cn