一种基于最大熵原理的社交网络用户关系分析模型

肖云鹏*[①] 杨 光[①] 刘宴兵[①] 吴 斌[②]
[①](重庆邮电大学网络与信息安全技术重庆市工程实验室 重庆 400065)
[②](北京邮电大学北京市智能通信软件与多媒体重点实验室 北京 100876)

摘 要：在社交网络的演化和发展过程中，用户之间关系的建立受到多种因素的共同作用。该文通过对社交网络中用户属性以及用户关系数据进行分析，旨在发现影响用户关系建立的关键因素。首先，针对用户关系建立的复杂驱动因素，分别从个人兴趣、好友关系、社团驱动 3 个方面提取影响用户关系建立的因素并定义相应的影响因子函数。其次，针对多种影响因素难以量化以及权值分配不确定等问题，以最大熵原理为基础构建用户关系分析模型，该模型在选择特征时具有不需要依赖于特征之间的关联性等特点，并能够量化各个因素对用户关系建立的驱动强度。从而挖掘影响链接建立的关键因素，分析用户关系发展态势。实验表明，该模型不仅能够量化各因素对链接建立的驱动强度，发现关键影响因素，而且可以对用户关系进行有效预测。

关键词：社交网络；用户关系；关系态势；最大熵原理
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2017)04-0778-07
DOI: 10.11999/JEIT160605

Social Relationship Analysis Model Based on the Principle of Maximum Entropy

XIAO Yunpeng[①] YANG Guang[①] LIU Yanbing[①] WU Bin[②]
[①](*Chongqing Engineering Laboratory of Internet and Information Security*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)
[②](*The Intelligent Communication Software and Multimedia Key Laboratory of Beijing*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)

Abstract: Within the evolution and development of social networks, the establishment of relationships among the users is affected by various factors. By analyzing user behavior data and relationship data in social network, this study tries to detect the key factors that affect the formation of relationship among users. Firstly, considering the complex driving factors for the user relationship establishment, the factors are extracted and the impact factor functions are defined from personal attributes, friendships and community driving. Secondly, in order to quantify driving factors and assign weight, a user relationship analysis model based on the principle of maximum entropy is proposed. The model is, when choosing features, characterized by its independence from the association among features, and can also quantify the strength of various factors that drive users to establish relationship. Furthermore, the key factors that affect the user relationship can be detected and the development trend of user relationship can be analyzed. Experimental results reveal that the proposal model can not only quantify the strength of each factor that drives relationship establishment, it can also predict the user relationship effectively.
Key words: Social network; User relationship; Situation analysis; Principle of maximum entropy

1 引言

随着信息技术的不断进步，在线社交网络得到了蓬勃发展，并且逐渐成为人们生活不可或缺的一部分。在社交网络的研究中，用户关系分析是一个基础问题，近年来受到各个领域越来越多的关注[1,2]。分析用户关系可以帮助人们更加深刻地了解网络的演化模式和发展方向，同时相关研究也可以被广泛地应用于各个领域，例如：电子商务中的商品推荐，以及生物学领域蛋白质相互作用关系的发现中，从而产生巨大的经济效益和社会效益。

现阶段对于社交网络中的关系分析，主要有用户关系强度以及用户关系预测等方面的研究。在用户关系强度的研究中[3,4]，文献[5]运用了监督学习的方法来预测用户关系强度，该方法可以较好地发现网络中所存在的强链接。文献[6]则是提出了一个无监督的模型，该模型可以通过用户的相似度和交互活动来评估用户的关系强度。文献[7]提出一种基于

收稿日期：2016-06-07；改回日期：2016-11-30；网络出版：2017-01-22
*通信作者：肖云鹏 xiaoyp@cqupt.edu.cn
基金项目：国家 973 计划项目(2013CB329606)，国家自然科学基金(61272400)，重庆市青年人才项目(cstc2013kjrc-qnrc 40004)，教育部-中国移动研究基金(MCM20130351)，重庆市研究生研究与创新项目(CYS14146)，重庆市教委科学计划项目(KJ1500425)，重庆邮电大学文峰基金(WF201403)
Foundation Items: The National 973 Program of China (2013CB 329606), The National Natural Science Foundation of China (61272400), Chongqing Youth Innovative Talent Project (cstc2013 kjrc-qnrc40004), Ministry of Education of China and China Mobile Research Fund (MCM20130351), Chongqing Graduate Research and Innovation Project (CYS14146), Science and Technology Research Program of the Chongqing Municipal Education Committee (KJ1500425), WenFeng Foundation of CQUPT (WF201403)

半监督学习的方法，该方法不仅可以量化节点间链接的强度还可以预测出链接的类型。文献[8]通过对社交网络中的线上互动数据进行分析，运用机器学习算法估计亲密度从而识别出现实中的强链接。

在用户关系预测方面主要是关于链接预测的研究，并且通常运用相似性指标对用户关系进行分析[9]，例如：共同邻居、Jaccard 系数、Adamic/Adaic[10]、Katz[11]等。然而这些方法只考虑了网络拓扑结构信息而忽略了其他可以用来提高关系预测准确度的信息。在基于随机游走的方法中[12,13]，文献[14]提出了一种基于网络局部随机游走的相似性指标，区别于全局的随机游走，该方法仅考虑有限步数的随机游走。目前越来越多的研究运用社交理论来对用户关系进行分析，文献[15]运用共同邻居的节点中心性为基础构建用户关系分析模型，包括 3 种中心性：Degree, Closeness 以及 Betweenness，并发现弱连接可以提高预测的准确度。另外在监督学习层面，经常把用户关系预测当作分类问题来处理[16,17]，并将更多的信息应用到用户关系的预测中。其中文献[17]运用机器学习的方法来进行链路预测，并对比不同机器学习算法的预测效果发现 SVM 取得了最优的预测效果。

以上的研究着重于从不同的角度来进行用户关系分析，研究用户关系之间的亲密程度并预测用户关系，而在影响用户关系建立的各个因素之间驱动强度的研究相对较少。针对该问题，本文提出了一种用户关系分析模型，首先通过分析用户关系建立的复杂驱动因素，分别从个人兴趣、好友关系、社团影响 3 个方面提取特征，其次针对多种因素难以量化以及因素权重分配不确定等问题运用最大熵的原理和方法[18]，构建用户关系分析模型。该模型不仅可以量化影响因素的权重，挖掘出影响用户关系建立的关键因素，还可以对用户关系进行预测。

2　问题定义

2.1 相关定义

本文要解决的问题是：通过对当前网络中用户的数据进行分析，提取影响用户之间建立链接的因素，并量化各个因素的驱动强度，然后对用户关系进行预测。首先定义如下几个基本概念：

定义 1　用户关系 $R_{j,k}$

对于任意的用户 v_j 和用户 v_k，若二者之间存在相互关注关系，则认为用户 v_j 和用户 v_k 之间存在用户关系，即 $R_{j,k}=1$；若二者不存在任何关注关系，则认为用户 v_j 和用户 v_k 之间不存在用户关系，即 $R_{j,k}=0$。

定义 2　初始用户关系网络 $G=(V,E)$

其中，G 表示初始用户关系网络。V 是初始用户的集合，$|V|=N$，即初始用户网络中用户的总数。$E\subset V\times V$ 是初始用户群体中的用户关系边，即用户之间是否存在关系。

定义 3　全用户关系网络 $G'=(V',E')$

其中，G' 表示全用户关系网络。V' 是所有用户的集合，$|V'|=N'$，即全网络中用户的总数。$E'\subset V'\times V'$ 表示全网络用户群体中的用户关系边，即用户之间是否存在关系。

2.2 特征提取

本文分别从个人兴趣、好友关系、社团驱动 3 个方面提取影响用户关系的特征，具体特征如下所示：

(1)个人兴趣：通常认为网络中两个节点之间的相似性越大，两个节点之间存在链接的可能性就越大。最简单直接的方法就是运用节点的属性，为了便于描述，定义 X_I 表示个人兴趣特征集合，对于任意的个人兴趣特征 $x_I^i\in X_I$，若用户 v_j 和用户 v_k 满足该条件，则 $x_I^i=1$，反之为 0。本文中分别运用符号 Vip，Location，SamSex，DifSex，Tag，Elite 来表示认证用户、地点、相同性别、不同性别、标签以及精英用户等方面的特征。

相对于普通用户来讲，精英用户拥有更多的链接。本文中运用用户的粉丝特征值[19−21]来选取精英用户，将所得特征值排名前 5%~10%的用户作为精英用户。其中对于用户 v_j 的粉丝特征值 $f(v_j)$ 计算如式(1)：

$$f(v_j)=\varepsilon\left(N_{v_j}^f-N_{v_j}^m\right)+N_{v_j}^m \tag{1}$$

其中，$N_{v_j}^f$ 代表用户 v_j 的粉丝数目，$N_{v_j}^m$ 代表用户 v_j 的互粉好友数目，在本文中选取 $\varepsilon=2$，以缩小用户之间粉丝数量特征值的差距。

(2)好友关系：在社交网络中，用户之间是否建立链接同时也受到好友的影响，若两个人拥有共同好友，那么他们之间建立链接的概率也就更高。因此可以将用户之间的共同粉丝和共同关注作为影响链接建立的特征。为了便于描述，定义 X_U 表示好友关系特征集合，对于任意的特征 $x_U^i\in X_U$，若用户 v_j 和用户 v_k 满足该特征，则 $x_U^i=1$，反之为 0。本文中分别运用符号 Fans，Following 来表示共同粉丝、共同关注等特征。

(3)社团驱动：社团也对用户之间链接的建立存在一定的影响，同属于一个社团的用户之间联系更加紧密，更容易产生链接。本文中运用社团分类算

法 CPM 判断用户是否属于同一个社团。为了便于描述，定义 X_G 表示社团特征集合，对于任意的社团特征 $x_G^i \in X_G$，若用户 v_j 和用户 v_k 满足该特征，则 $x_G^i = 1$，反之为 0。本文中运用符号 Community 来表示社团特征。

2.3 问题形式化

在给定初始用户关系网络 $G=(V,E)$，全用户关系网络 $G'=(V',E')$ 的前提下，本文拟解决如下问题：

(1)如何量化因素的驱动强度？构建用户关系分析模型并引入参数集合 θ，通过求取最优参数集合 $\theta^* = \arg\max_\theta P_\theta(Y^s \mid G^s)$，其中 G^s 表示源网络，$Y^s = \{y_1, y_2, \cdots, y_n\}$ 用来表示源网络 G^s 中用户关系是否存在。

(2)怎样通过参数集合 θ 来预测用户关系？利用最优参数集合 θ^* 对用户关系进行预测，即求 $Y^* = \arg\max P_{\theta^*}(Y^t \mid G^t)$，其中 G^t 表示目标网络，Y^t 是用来表示目标网络 G^t 中用户关系是否存在。

3 模型

3.1 用户关系影响因子函数

为解决上述问题，首先定义参数集合并求解最优参数集合，使条件概率最大化 $\theta^* = \arg\max_\theta P_\theta(Y^s \mid G^s)$。根据第 2 节所提到的 3 方面的特征，本文定义的用户关系影响因子函数为

$$f_P^i\left(x_P^i, y_k\right) = \begin{cases} x_P^i, & x_P^i \neq 0 \cap y_k = 1 \\ 0, & \text{其它} \end{cases} \tag{2}$$

其中，y_k 用来表示用户之间是否存在链接，如果存在则 $y_k = 1$，反之为 0。P 是可变参数，取值为 I，U，G 分别表示个人兴趣、好友关系、社团驱动等方面的影响因子函数。例如：$f_I^i(x_I^i, y_k)$ 表示的是用户个人兴趣特征和用户关系的相关性，$x_I^i \neq 0 \cap y_k = 1$ 表示用户之间存在链接，并且满足个人兴趣特征中的第 i 个特征取值不为 0。

3.2 用户关系分析模型

以最大熵原理为基础建立用户关系分析模型，最显著的优点是在选择特征时，不需要独立的假设。对于给定的初始用户关系网络 $G=(V,E)$，全用户关系网络 $G'=(V',E')$，我们的目的是通过网络 G，G' 分别从个人兴趣、好友关系、社团驱动 3 个方面提取特征 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_k, y_k)\}$，$(x_k \in X, y_k \in Y)$，其中 X 表示的是影响用户关系建立的特征，Y 表示所属类别，在这里的意思是是否存在链接。已知的约束条件 1 为所有特征的条件概率总和为 1：

$$\sum_y p(y \mid x) = 1 \tag{3}$$

其中，$p(y \mid x)$ 是条件概率，表示的是在特征 x 出现的情况下，y 出现的概率。另外对于影响因子函数 $f_i(x,y)$，它相对于经验概率分布 $\tilde{p}(x,y)$ 的期望值为

$$E_{\tilde{p}}(f_i) = \sum_{(x,y)} \tilde{p}(x,y) f_i(x,y) \tag{4}$$

影响因子函数 $f_i(x,y)$ 相对于模型的条件概率 $p(y \mid x)$ 的期望值为

$$E_p(f_i) = \sum_{(x,y)} \tilde{p}(x) p(y \mid x) f_i(x,y) \tag{5}$$

$p(y \mid x)$ 是所要求的条件概率，$\tilde{p}(x)$ 是统计概率。因为我们限制在给定的数据集中，那么就可以假设这两个期望值相等，于是得到约束条件 2，即

$$E_p(f_i) - E_{\tilde{p}}(f_i) = 0 \tag{6}$$

以上分析给出了模型的约束条件，根据最大熵原理，此时的条件概率 $p^*(y \mid x)$ 应该是使条件熵 $H(Y \mid X)$ 值最大，此时 $p^*(y \mid x)$ 的表示公式为

$$\begin{aligned} p^*(y \mid x) &= \arg\max H(Y \mid X) \\ &= \arg\max \sum_{(x,y)} \tilde{p}(x) p(y \mid x) \lg \frac{1}{p(y \mid x)} \end{aligned} \tag{7}$$

于是，现在的问题转化为满足一组约束条件，求解最优解的问题。而求解这个问题经典的方法就是拉格朗日乘子法。首先为约束条件 1 引入拉格朗日乘子 Υ，对于约束条件 2，假设总共有 k 个特征并为每一个特征约束引入拉格朗日乘子 λ_i，则拉格朗日函数的表示形式为

$$\begin{aligned} L(p, \lambda, \Upsilon) = H(Y \mid X) &+ \sum_{i=1}^{k} \lambda_i \left[E_p(f_i) - E_{\tilde{p}}(f_i)\right] \\ &+ \Upsilon\left(\sum_y p(y \mid x) - 1\right) \end{aligned} \tag{8}$$

对式(8)求偏导 $\dfrac{\partial L}{\partial p(y \mid x)}$，当 L 取得最大值时，偏导 $\dfrac{\partial L}{\partial p(y \mid x)} = 0$ 因此求得条件概率的表示为

$$p^*(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i^k \lambda_i f_i(x,y)\right) \tag{9}$$

$$Z(x) = \sum_y \exp\left(\sum_i^k \lambda_i f_i(x,y)\right) \tag{10}$$

因为我们通过个人兴趣、好友关系、社团驱动 3 方面来提取影响链接建立的特征，并定义了相关的影响因子函数。然后分别为各个影响因子函数定义参数集合 $\theta = (\{\alpha\}, \{\beta\}, \{\gamma\})$。所以条件概率 $p^*(y \mid x)$ 又可以表示为

$$p^*(y\mid x)=\frac{1}{Z(x)}\exp\left(\sum_i^{K_I}\alpha_i f_I^i\left(x_I^i,y_k\right)+\sum_i^{K_U}\beta_i f_U^i\left(x_U^i,y_k\right)+\sum_i^{K_G}\gamma_i f_G^i\left(x_G^i,y_k\right)\right) \quad (11)$$

$$Z(x)=\sum_y\exp\left(\sum_i^{K_I}\alpha_i f_I^i\left(x_I^i,y_k\right)+\sum_i^{K_U}\beta_i f_U^i\left(x_U^i,y_k\right)+\sum_i^{K_G}\gamma_i f_G^i\left(x_G^i,y_k\right)\right) \quad (12)$$

其中，$Z(x)$是归一化因子，确保概率为 1。$f_I^i(x_I^i,y_k)$，$f_U^i(x_U^i,y_k)$，$f_G^i(x_G^i,y_k)$分别代表从个人兴趣、好友关系、社团驱动 3 个方面所定义的影响因子函数。k_I，k_U，k_G分别代表每类特征的数目。α_i，β_i，γ_i代表各个影响因子函数的权值，即该特征对用户关系建立的驱动强度的大小。

条件概率如式(11)所示，但实际上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，在本文中采用 GIS(Generalized Iterative Scaling）算法来进行求解。以参数集合$\{\alpha\}$为例，可得到参数更新梯度η，其中参数更新的公式为

$$\alpha_{ne}=\alpha_{pre}+\eta,\quad \eta=\frac{1}{\mathrm{c}}\lg\frac{E_{\tilde{\mathrm{p}}}[f_i(x_I^i,y_k)]}{E_{\mathrm{p}}[f_i(x_I^i,y_k)]} \quad (13)$$

常数 c 是训练样本中最大的特征个数。其中α_{pre}，α_{ne}分别代表更新前的参数和更新后的参数，$E_{\tilde{\mathrm{p}}}[f_i(x_I^i,y_k)]$，$E_{\mathrm{p}}[f_i(x_I^i,y_k)]$分别代表经验分布$\tilde{p}(x,y)$的期望值和模型$p(y\mid x)$的期望值。最后，判断是否收敛。收敛条件可以有不同的方法，本文中采用收敛方式为：每个参数的变化值都小于某个阈值。若收敛转到输出，若不收敛，代入更新后的参数集合，继续迭代直至收敛。

对于问题 1，影响因素驱动强度大小依据参数的变化而不同，运用模型学习算法所获取的最优参数集合θ^*，可以定量反映出各个因素对用户关系建立的影响。

对于问题 2，因为用户关系的预测受到多种因素的影响，把影响因素组成向量$\boldsymbol{X}$，然后运用已经训练好的模型计算在该条件下用户v_j和用户v_k产生链接的概率$p_{jk}=p(y\mid x)$。并且仅当p_{jk}的值大于指定阈值ξ时，y取值为 1；否则为 0。

$$y=\begin{cases}1, & p(y\mid x)\ge\xi\\ 0, & p(y\mid x)<\xi\end{cases} \quad (14)$$

3.3 模型学习算法

用户关系分析模型的学习过程就是求解最优参数的过程，即求解参数集合$\theta=(\{\alpha\},\{\beta\},\{\gamma\})$，其模型完整的参数学习过程如表 1 所示。

表 1 模型学习算法

输入：初始用户关系网络$G=(V,E)$；全用户关系网络$G'=(V',E')$。

输出：$\theta^*=\arg\max_\theta P_\theta(Y^s\mid G^s)$；$Y^*=\arg\max P_{\theta^*}(Y^t\mid G^t)$。

步骤 1　从初始用户关系网络$G=(V,E)$中，选取正样本和负样本构建源网络$G^s=(V^s,E^s)$和目标网络$G^t=(V^t,E^t)$；

步骤 2　由影响因子函数分别统计出$\tilde{p}(x,y)$和$\tilde{p}(x)$；

步骤 3　由$\tilde{p}(x,y)$计算出经验期望$E_{\tilde{\mathrm{p}}}(f_i)$；

步骤 4　初始化参数集合θ，在这里初始化为 0；

迭代过程：

步骤 5　通过式(5)计算模型期望$E_p(f_i)$；

步骤 6　通过式(13)更新每一个参数，直至收敛；

步骤 7　运用收敛后所获取的最优参数集合$\theta^*=(\{\alpha\},\{\beta\},\{\gamma\})$计算条件概率$p^*(y\mid x)$；

步骤 8　通过条件概率$p^*(y\mid x)$对目标网络G^t中的用户关系进行预测。

模型学习算法首先是根据所定义的影响因子函数，进行统计得到$\tilde{p}(x,y)$，$\tilde{p}(x)$，然后初始化参数，计算条件概率和模型中的期望并利用 GIS 算法更新参数；不停地迭代直至收敛。其中构建源网络和目标网络的时间复杂度为$O(|V|^2)$。设N为源网络中样本数目，k为本文所选择的特征的个数，若算法经过p次迭代后收敛，则在迭代阶段算法的时间复杂度$T_{re}=O(pkN)$；设M为目标网络中样本数目，n为所属类别数，则在预测阶段算法的时间复杂度$T_{pr}=O(nkM)$。

4 实验结果与分析

4.1 数据集

本文实验数据采自腾讯微博社交平台。腾讯微博是中国目前最主流的在线社交网络平台之一，早在 2013 年，其在线用户已超过 2 亿。我们通过抓取局部网络的用户，以这些用户为起始点构建用户关系网，来对我们所提出的用户关系分析模型的效果进行分析验证。并且利用社团分类算法 CPM 对用户关系网中的用户进行了社团划分。3 个数据集 Data A, Data B, Data C 的统计情况如表 2 所示。

4.2 实验结果分析

首先，从初始用户关系网络$G=(V,E)$中找出顶点对构建样本集合，根据 2.1 节相关定义中关于

表2 数据集统计情况

数据	初始用户数	粉丝	关注	社团
Data A	3504	898126	858110	132
Data B	7022	879780	878203	266
Data C	9626	534459	534420	90

用户关系的定义，若该顶点对存在用户关系即 $R=1$，则构成正样本；若该顶点对不存在任何关注关系即 $R=0$，则构成负样本。接下来介绍源网络 $G^s=(V^s,E^s)$ 和目标网络 $G^t=(V^t,E^t)$ 的具体构建过程：(1)构建正样本集合。从初始用户关系网络 $G=(V,E)$ 找到所有正样本的顶点对，假设我们从 $G=(V,E)$ 中找到的所有正样本的数量为 N_1；(2)构建负样本集合。从初始用户关系网络 $G=(V,E)$ 中随机选取顶点对，若该顶点对不存在任何关注关系则构成一个负样本。本文中选取负样本的数量也为 N_1，也就是选择相同数量的负样本和正样本；(3)当正样本和负样本都选取好之后，构成最初的样本集合，对样本集合采用十折交叉验证从而产生 10 组源网络 $G^s=(V^s,E^s)$ 和目标网络 $G^t=(V^t,E^t)$。

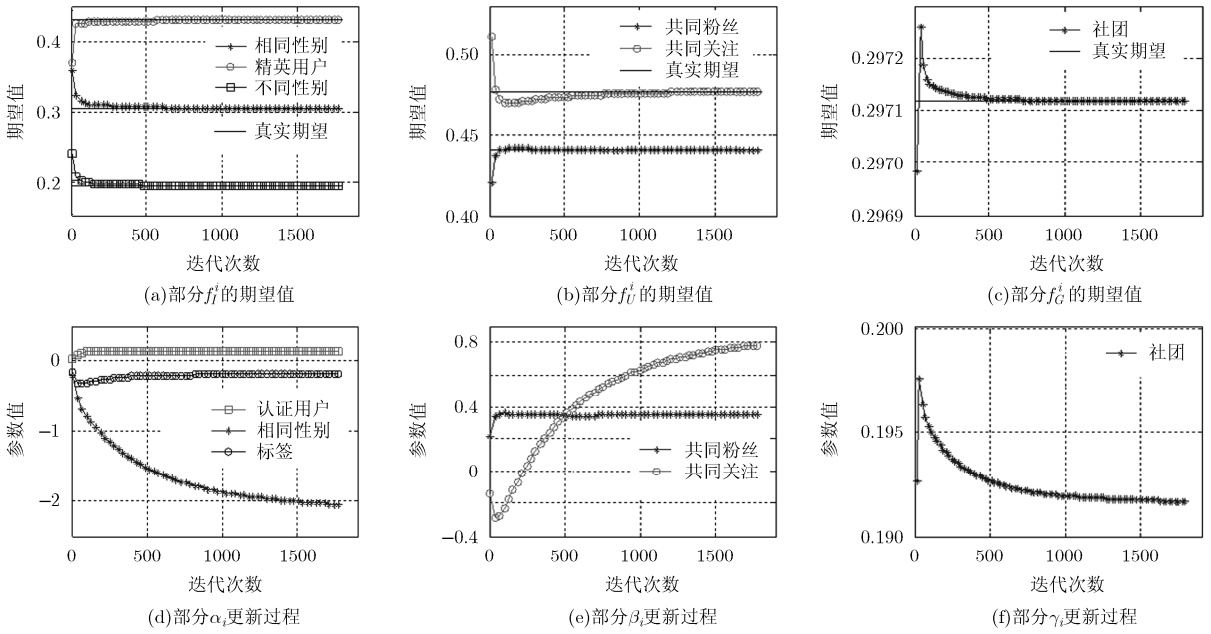
在源网络 $G^s=(V^s,E^s)$ 中，分别从个人兴趣、好友关系、社团驱动 3 个方面选取影响用户关系建立的特征，然后在模型中进行训练，选取 Data A 的部分参数来演示模型的迭代过程。如图 1 所示，对于图 1(a)，图 1(b)，图 1(c)，其横坐标代表迭代的次数，纵坐标代表模型的期望值；同样的对于图 1(d)，图 1(e)，图 1(f)，其横坐标代表迭代的次数，纵坐标代表参数值。

从图 1(a)，图 1(b)，图 1(c)可以看出不同的影响因素期望值也是不一样的，但是随着迭代次数的增加，各个影响因素的模型期望都越来越接近真实的期望值。从图 1(d)，图 1(e)，图 1(f)中我们看到在迭代的初始阶段参数的变化很不稳定，迭代的步伐较大，但随着迭代次数的增加参数越来越趋于稳定。从图 1 (e)中可以发现部分参数值为负值，这说明在该数据集中，该特征对用户关系的建立起到了相反的作用。当参数的变化小于某个阈值的时候可以认为参数收敛了。当每一个参数都迭代至收敛时，获取最优参数集合，从而量化各个影响因素的驱动强度值。

本文从个人兴趣、好友关系、社团驱动提取对用户关系有影响的特征，因此在这里从这 3 个方面出发，分别选取对用户关系建立影响最大的关键因素进行展示。如图 2 所示，其中横坐标代表实验的次数；纵坐标代表关键影响因素的驱动强度量化值。

从图中可以很直观的观察到 3 个数据集中关键影响因素的变化情况。在 Data A 中驱动强度值最大的是共同关注；在 Data B 中驱动强度值最大的是精英用户；在 Data C 中驱动强度值最大的是社团。虽然在不同的实验数据下关键影响因素以及驱动强度值略有不同，但就总体而言：精英用户、共同关注、共同粉丝、社团等均成为关键影响因素，也就是对用户关系的影响占据优势地位，这与我们的日常感知也是趋于一致的。

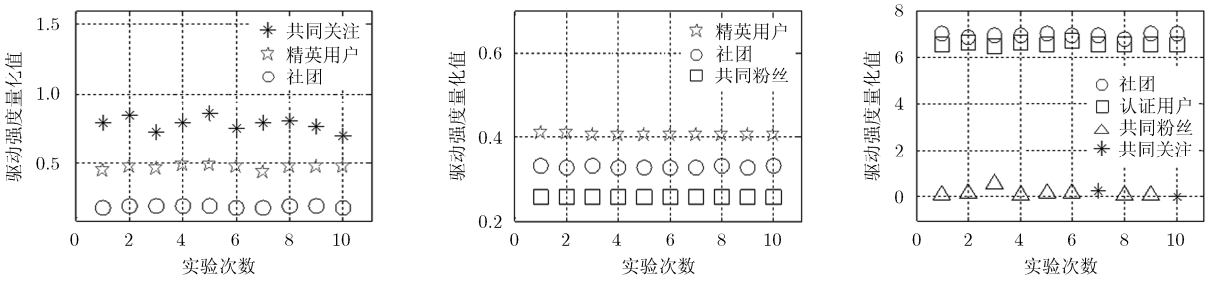
最后，利用十折交叉验证的方法，对目标网络 $G^t=(V^t,E^t)$ 中的用户关系进行预测展示模型的预测效果，并分别计算准确率、精确度和召回率，求取平均值作为最终的结果。本文采用支持向量机(SVM)、朴素贝叶斯(Naive Bayes)与用户关系分析模型进行比较对照。我们随机抽取 100%, 75%, 50% 的数据组成不同的数据集，测试本模型在不同大小数据集上的预测效果。实验效果如表 3～表 5 所示。



(a)部分f_I^i的期望值 (b)部分f_U^i的期望值 (c)部分f_G^i的期望值

(d)部分α_i更新过程 (e)部分β_i更新过程 (f)部分γ_i更新过程

图 1 模型迭代过程



(a) Data A驱动强度量化值　(b) Data B驱动强度量化值　(c) Data C驱动强度量化值

图 2 不同影响因素驱动强度比较

表 3　Data A 上平均预测效果

算法	100%			75%			50%		
	准确率	精确率	召回率	准确率	精确率	召回率	准确率	精确率	召回率
支持向量机	91.36	91.50	91.40	91.54	91.60	91.50	91.79	91.90	91.80
朴素贝叶斯	91.78	91.90	91.80	92.03	92.10	92.00	92.27	92.40	**92.30**
本文模型	**92.00**	**92.39**	**91.96**	**92.42**	**92.82**	**92.23**	**92.32**	**93.37**	91.67

表 4　Data B 上平均预测效果

算法	100%			75%			50%		
	准确率	精确率	召回率	准确率	精确率	召回率	准确率	精确率	召回率
支持向量机	70.00	70.40	**70.00**	69.80	70.20	69.80	**69.75**	70.20	69.70
朴素贝叶斯	69.94	70.20	69.90	69.93	70.20	**69.90**	69.70	70.00	69.70
本文模型	**70.12**	**73.94**	66.43	**70.12**	**73.96**	66.12	66.37	**70.64**	**71.63**

表 5　Data C 上平均预测效果

算法	100%			75%			50%		
	准确率	精确率	召回率	准确率	精确率	召回率	准确率	精确率	召回率
支持向量机	94.45	95.00	**94.50**	95.30	95.70	**95.30**	96.22	96.50	**96.20**
朴素贝叶斯	92.56	92.60	92.60	92.45	92.50	92.40	92.44	92.60	92.40
本文模型	**94.56**	**99.43**	89.37	**95.42**	**97.50**	93.40	**96.41**	**99.44**	92.97

在表 3~表 5 中，我们看到 Data B 上的预测效果要整体的低于 Data A, Data C 的预测效果，可能是由于 Data B 中数据不均匀、稀疏性问题所造成的。对于 Data A，从表 3 中可以看出本文的模型除了在 50%的数据上的召回率不是最高的，在其他的指标上的表现都是最优的。对于 Data B，从表 4 看出本文的模型在召回率上相对较低，另外在 50%的数据上准确率较高的是 SVM，而在其他的指标上表现最好的是用户关系分析模型。对于 Data C，从表 5 可以看出来在召回率上表现较好的是 SVM，但是在准确率以及精确度上表现最好的依然是用户关系分析模型。另外从表 5 可以看到在 50%的数据上实验效果最好，这种情况可能由以下原因引起：(1)算法的效果与数据密切相关，当数据量较少的情况下，实验效果具有偶然性；(2)可以看到用于对比的两种算法 SVM 和朴素贝叶斯算法也都是在 50%的数据上收到了较好的效果，其中 SVM 更是在 3 项指标中都取得了最大值，可以发现该数据可能更加适合此场景下的应用。

5　结论

本文通过对影响用户关系建立的复杂驱动因素进行分析，分别从个人兴趣、好友关系、社团驱动 3 个方面定义相关的影响因子函数，提取影响用户关系的因素。然后以最大熵理论为基础，构建了用户关系分析模型。该模型与传统的 SVM，朴素贝叶斯等算法相比，不仅可以挖掘出用户关系建立的关键影响因素，还可以有效地提高预测的效果。

参 考 文 献

[1] WANG P, XU B W, WU Y R, *et al*. Link prediction in social networks: The state-of-the-art[J]. *Science China Information Sciences*, 2015, 58(1): 1–38. doi: 10.1007/s11432-014-5237-y.

[2] LÜ L and ZHOU T. Link prediction in complex networks: A survey[J]. *Physica A*: *Statistical Mechanics and Its Applications*, 2011, 390(6): 1150–1170. doi: 10.1016/j.physa.2010.11.027.

[3] ZHU B and XIA Y. Link prediction in weighted networks: A weighted mutual information model[J]. *PloS One*, 2016, 11(2): e0148265. doi: 10.1371/journal.pone.0148265.

[4] YUAN N J, ZHONG Y, ZHANG F, *et al*. Who will reply to/retweet this tweet?: The dynamics of intimacy from online social interactions[C]. Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, California, USA. 2016: 3–12. doi: 10.1145/2835776.2835800.

[5] KAHANDA I and JENNIFER N. Using transactional information to predict link strength in online social networks[C]. International AAAI Conference on Weblogs and Social Media, California, USA, 2009: 74–81.

[6] XIANG Rongjing, NEVILLE J, and ROGATI M. Modeling relationship strength in online social networks[C]. Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, 2010: 981–990. doi: 10.1145/1772690.1772790.

[7] KASHIMA H, KATO T, YAMANISHI Y, *et al*. Link propagation: A fast semi-supervised learning algorithm for link prediction[C]. Proceedings of the SIAM International Conference on Data Mining, Nevada, USA, 2009, 9: 1099–1110. doi: 10.1137/1.9781611972795.94.

[8] JONES J J, SETTLE J E, BOND R M, *et al*. Inferring tie strength from online directed behavior[J]. *PloS One*, 2013: e52168. doi: 10.1371/journal.pone.0052168.

[9] SHAO C, DUAN Y, and WANG B. Attractive density: a new node similarity index of link prediction in complex networks[C]. IEEE International Conference on Information Science and Technology, Changsha, China, 2015: 74–78. doi: 10.1109/icist.2015.7288943.

[10] ADAMIC L A and ADAR E. Friends and neighbors on the web[J]. *Social Networks*, 2003, 25(3): 211–230. doi: 10.1016/s0378-8733(03)00009-1.

[11] KATZ L. A new status index derived from sociometric analysis[J]. *Psychometrika*, 1953, 18(1): 39–43. doi: 10.1007/bf02289026.

[12] CHEN Y X and CHEN L. Random walks for link prediction in networks with nodes attributes[C]. Network Security and Communication Engineering: Proceedings of the 2014 International Conference on Network Security and Communication Engineering, Hong Kong, 2015: 291. doi: 10.1201/b18660-64.

[13] XIE X, LI Y, ZHANG Z, *et al*. A joint link prediction method for social network[C]. International Conference of Young Computer Scientists, Engineers and Educators, Harbin, China, 2015: 56–64. doi: 10.1007/978-3-662- 46248-5_8.

[14] LIU W and LU L. Link prediction based on local random walk[J]. *Europhysics Letters*, 2010, 89(5): 58007–58012. doi: 10.1209/0295-5075/89/58007.

[15] LIU H, HU Z, HADDADI H, *et al*. Hidden link prediction based on node centrality and weak ties[J]. *Europhysics Letters*, 2013, 101(1): 18004–18009. doi: 10.1209/0295-5075/101/18004.

[16] AHMED C, LKORANY A, and BAHGAT R. A supervised learning approach to link prediction in Twitter[J]. *Social Network Analysis and Mining*, 2016, 6(1): 1–11. doi: 10.1007/s13278-016-0333-1.

[17] AHEMD C and ELKORANY A. Enhancing Link prediction in twitter using Semantic user attributes[C]. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 2015: 1155–1161. doi: 10.1145/2808797.2810056.

[18] BERGER A L, PIETRA V J D, and PIETRA S A D. A maximum entropy approach to natural language processing [J]. *Computational Linguistics*, 1996, 22(1): 39–71. doi: 10.1007/springerreference_179232.

[19] PAKZAD F and ABHARI A. Characterization of user networks in Facebook[C]. Proceedings of the 2010 Spring Simulation Multiconference. Society for Computer Simulation International, Orlando, FL, USA, 2010: 105. doi: 10.1145/1878537.1878647.

[20] PARK B, LEE K, and KANG N. The impact of influential leaders in the formation and development of social networks[C]. Proceedings of the 6th International Conference on Communities and Technologies ACM, Munich, Germany, 2013: 8–15. doi: 10.1145/2482991.2483004.

[21] TANG X and YANG C C. Ranking user influence in healthcare social media[J]. *ACM Transactions on Intelligent Systems and Technology*, 2012, 3(4): 565–582. doi: 10.1145/2337542.2337558.

肖云鹏： 男，1979 年生，副教授，硕士生导师，研究方向为大数据、移动互联网、信息安全.

杨 光： 男，1992 年生，硕士生，研究方向为社交网络分析.

刘宴兵： 男，1971 年生，教授，博士生导师，研究方向为网络分析与网络安全.

吴 斌： 男，1969 年生，教授，博士生导师，研究方向为智能信息处理、图数据挖掘、云计算.