# 基于网络拓扑和地理特征融合的朋友关系预测模型

罗 惠 郭 斌 於志文 王 柱 封 云
(西北工业大学计算机学院 西安 710072)

**摘 要** 朋友关系预测已成为基于位置的社交网络(LBSN)的主要研究方向之一。提出一种基于网络拓扑特征和地理融合的面向LBSN的朋友关系预测方法。首先,利用信息增益评估不同特征对朋友关系的影响,最终选取3种重要特征:用户社交拓扑、用户签到地点类型和用户签到地点。然后,提出基于这3种特征融合的朋友关系预测方法,分别采用随机森林、支持向量机和朴素贝叶斯3种分类算法建模实现朋友关系推理。最后通过Foursquare和街旁的实际签到数据验证了特征选取的有效性和朋友关系预测的准确性。

**关键词** 基于位置的社交网络,朋友关系预测,信息增益,特征融合
**中图法分类号** TP39 **文献标识码** A

## Friendship Prediction Based on Fusion of Network Topology and Geographical Features

LUO Hui GUO Bin YU Zhi-wen WANG Zhu FENG Yun
(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

**Abstract** Friendship prediction has become one of the major studies of location based social network (LBSN). This paper proposed an approach for predicting friendship, which fuses the topology network and geographical features of LBSN. We first adopted the information gain to measure the contribution of different features to human friendship, and chose three key features: user social topology, the category of the location where people check in, and check in points. We then presented the friendship prediction method based on the fusion of the selected features. Three different classification models, including Random Forests, Support Vector Machine (SVM), and Naive Bayes, were selected to predict human friendship. Experimental results on the real collected data from Foursquare and JiePang verify the efficacy of the selected features and the accuracy of friendship prediction.

**Keywords** Location-based social network, Friendship prediction, Information gain, Feature fusion

## 1 引言

基于位置的社交网络(LBSN)通过时间序列、行为轨迹和地理位置的信息标记组合,帮助用户与外部世界建立更加广泛和密切的联系,增强社交网络与地理位置的关联性。LBSN在国外起步较早,国外著名的LBSN服务有Twitter、Foursquare等。国内起步较晚,主要有街旁、在哪、大众点评等。基于位置共享的社交网络的流行使得大规模采集用户的历史位置成为可能,引起研究者的广泛兴趣,研究主题包括社交拓扑分析[1,2]、社区挖掘[3,4]、社区结构理解[5]、社交隐私保护[6]和社会关系预测[7-11]等。对于社会关系预测,目前主流的方法是选取用户特征属性计算用户的相似度,将相似度高的用户预测为朋友。然而用户是否相似与其是否是朋友之间并没有必然的联系,且用户有很多方面的特征,选取的一些方面的特征相似并不表示用户整体上相似。该类方法存在的问题是需要为选取的特征属性设计专门的建模方法或特征转换方法以提高模型预测的精度。

本文提出一种基于网络拓扑特征融合的面向LBSN的朋友关系预测方法,将用户社交拓扑网络视为完全图,网络中节点为用户,边为用户关系,虚线表示非朋友,实线表示朋友。针对拓扑网络中的用户边,从用户社交拓扑网络特征和签到行为两个角度选取特征属性,以信息增益评估不同特征属性对用户朋友关系的影响,最终选取3种重要特征:用户社交拓扑、用户签到地点类型和用户签到地点。然后,提出了基于这3种特征融合的朋友关系预测方法,即从社交拓扑网络中提取朋友边和非朋友边,采用分类算法建立模型,分别采用随机森林、支持向量机和朴素贝叶斯3种分类模型实现朋友关系推理。最后通过国内外两个具有代表性的基于位置的社交网

到稿日期:2013-08-10 返修日期:2013-11-02 本文受国家重点基础研究发展计划(973计划)(2012CB316400),国家自然科学基金(61222209,61103063),教育部"新世纪优秀人才支持计划"(NCET-12-0466),教育部高等学校博士学科点专项科研基金(博导类)(20126102110043),陕西省自然科学基础研究计划项目(2012JQ8028),西北工业大学基础研究基金(JC20110267)资助。

罗 惠(1987-),男,硕士,CCF学生会员,主要研究方向为普适计算,E-mail:brouselh@gmail.com;郭 斌(1980-),男,副教授,CCF会员,主要研究方向为普适计算、社会智能、移动社会网络;於志文(1977-),男,博士,教授,CCF杰出会员,主要研究方向为普适计算、智能信息处理、个性化技术等;王 柱(1983-),男,博士生,CCF学生会员,主要研究方向为普适计算;封 云(1990-),女,硕士,CCF学生会员,主要研究方向为普适计算。

络 Foursquare 和街旁上的实际签到数据验证了特征选取的有效性和朋友关系预测准确性。3 种分类模型在 Foursquare 签到数据中的预测精度分别为 64.31%、77.23%和 64.32%，召回率分别为 75.81%、66.68%和 82.82%。在街旁签到数据的预测精度分别为 98.99%、98.95%和 98.95%，召回率分别为 72.73%、74.93%和 74.1%。实验结果证明，基于网络拓扑提取的特征既不需要建立专门的建模方法，也不需要进行特征的转换，采用通常的分类算法也能取得较高的朋友关系预测的准确率。

本文第 1 节介绍研究背景和本文的研究思路；第 2 节介绍基于社交网络中用户关系预测的相关工作；第 3 节详细介绍本文特征选取的方法；第 4 节介绍朋友关系预测方法；最后是实验结果与分析。

## 2 相关工作

目前已经有一些研究人员利用签到数据进行用户关系预测。用户关系预测的关键是采用合适的特征表征用户之间的关系，Schwartz 和 Wood[8] 通过“兴趣距离”计算两个用户之间的相似度，通过分析用户之间的兴趣将相同兴趣的用户聚类在一起。Gregory 等[11] 提出了一种扩展格文-纽曼算法用以发现网络中重叠的社交结构，Grob 采用该方法在相同的社交团体中给出朋友推荐[9]，这些方法都只用到单一属性，比如社交网络或用户兴趣。

文献[10,12-14]选取多种属性实现朋友关系预测。Ozseyhan 等[12] 利用社交网络用户历史数据，采用关联规则算法对用户交友做出推荐，从匹配用户的分数高于不匹配用户的分数做出有效推荐。Li 等[13] 通过分析 MSN 上用户签到数据，选取用户社交拓扑、用户标签和签到距离 3 个属性，建立三层的朋友关系模型对用户关系进行预测。Cranshaw 等[14] 提出了基于社交网络特征以及用户移动模式属性的提取方法。Sadilek 等[10] 利用回归决策树对用户标记和共同地点两个特征属性进行转换，以提高朋友关系预测精度和召回率。该类方法从用户相似度的角度选取了多个特征进行用户的关系预测，他们将相似度高的用户预测为朋友。为了提高预测的正确率，研究者们不得不为他们选取的特征建立专有的模型。

随着图论研究的兴起，人们利用复杂网络链路预测[15] 挖掘网络中潜在的边以达到关系预测的目的，通过选取节点和网络结构的特征属性来预测那些已经存在但尚未被发现或未来可能出现的边。Liben-Nowell 等[16] 分析了在社交网络中进行链路连接存在的问题。Zhou 等[17] 利用 10 种基于节点局部信息的相似性指标对 6 个实际网络进行预测，取得较好的结果。Leskovec 等[18] 研究多个数据集，以较高的精度预测出用户的朋友关系和对立关系。该类方法将用户放在社交拓扑网络中考虑用户之间的关系，利用图论研究网络的演化达到关系预测的目的。

## 3 特征选取

预测朋友关系的关键在于选取合适的特征来衡量用户之间的关系。本文基于用户社交拓扑网络进行用户关系预测，将拓扑网络视为完全图，借鉴复杂网络链路预测的思想选取用户边的特征属性，通过选取节点（用户签到行为）和网络结构（用户社交拓扑网络）的特征属性来预测那些已经存在但尚未被发现或未来可能出现的边（朋友关系）。

定义社交网络 $G_s(U_s,E_s)$，节点 $u_s$ 表示用户，若两用户 $u_i,u_j$ 是朋友则在 $G_s$ 用一条边 $e_{ij}$ 进行连接，由现实网络中朋友关系通常是对称的可知 $G_s$ 为无向图。以信息熵[20] 表示社交拓扑网络中朋友关系的信息量，由于信息增益[19] 能从整体上评估特征属性对目标属性（朋友关系）的贡献，因此选用信息增益来选取合适的特征属性。对于目标特征 $X$ 和待评估的特征 $Y$，信息增益 $IG(X,Y)$ 的值为 $X$ 的信息熵 $H(X)$ 减去 $X$ 关于 $Y$ 的信息期望[20] $H(X|Y)$，定义如式(1)所示。

$$IG(X,Y)=H(X)-H(X|Y) \tag{1}$$

假设 $X$ 有 $n$ 个值 $\{x_1,\cdots,x_n\}$，$p(x_i)$ 是关于 $x_i$ 的概率统计函数，$p(x_i,y_i)$ 是 $X=x_i$ 且 $Y=y_i$ 的概率，信息期望定义如下：

$$H(X)=\sum_{i=1}^{n}-p(x_i)\log^{p(x_i)} \tag{2}$$

假设 $Y$ 有 $m$ 个值 $\{y_1,\cdots,y_m\}$，$p(y_j)$ 是关于 $y_j$ 的概率统计函数，信息熵定义如下：

$$H(X|Y)=\sum_{i,j}p(x_i,y_j)\log\frac{p(y_j)}{p(x_i,y_j)} \tag{3}$$

本文 Foursquare 数据集采集了从 2011 年 10 月开始连续 8 个星期的数据，共 72 万个匿名用户大约 1200 万条签到数据。由于 Foursquare 的 API[21] 仅提供有限的授权，本文通过 Twitter 的 API[22] 来采集公开的签到消息，采集用户和签到信息的元数据，包括每个用户在 Twitter 上的个人资料以及每个签到地点在 Foursquare 上的信息。本文的街旁数据集通过公开 API[23] 采集了 2012 年 4 月到 6 月的数据，包括 89936 个匿名用户共 2 百多万条的签到数据。

以选取 Foursquare 中巴黎签到的用户数据为例，其中用户 2731 人，朋友边 5590，非朋友边 3722225，目标属性（用户关系）信息熵为 0.0162。从用户的在线活动和离线行为数据中选取的 5 个属性如表 1 所列。可以看出用户签到地点对和用户签到次数的信息增益太小，对朋友预测的影响几乎可以忽略，因此在本文中选取 3 个重要属性：用户社交拓扑、用户签到地点类型和用户签到地点。

表 1 不同属性的信息增益

| 属性 | 信息增益 |
|---|---|
| 用户社交拓扑 | 0.0055 |
| 用户签到地点类型 | 0.0021 |
| 用户签到地点 | 0.0012 |
| 用户签到地点对 | 0.0004 |
| 用户签到数目 | 0.0002 |

### 3.1 用户社交拓扑

两个人是否是朋友与他们的社交网络的属性有很强关系[24]。在用户拓扑网络中，两个用户当前不是朋友，但他们有共同的邻居，则在未来一段时间他们成为朋友的可能性比没有共同邻居的用户要大。同样，若两个用户虽然当前是朋友，但他们没有任何的共同邻居，和有共同邻居的用户相比，他们之间的朋友边变为非朋友边的可能性更大。为了量化这种关系，本文引入社交距离的概念，在社交网 $G_s$ 中，去掉图 $G_s$ 中用户 $u_1,u_2$ 之间的边（若用户是朋友的话），用户 $u_1,u_2$ 拓扑网络中的最短距离即为他们在用户拓扑网络中的社交距离。设用户 $i$ 和 $j$ 之间的用户边在社交网络结构中的社交距离为属性 $a_s$，则：

$$a_s(i,j)=shortestDis(u_i,u_j)in\quad G_s'(U_s,E_s-e_{ij}) \tag{4}$$

### 3.2 用户签到地点类型

用户的签到地点的类型反映用户的偏好，两个用户签到的地点虽然不同，但可能是同一类型。为了便于处理，本文选取 400 种类型的地点，去掉那些不能分类于这 400 类地点的签到地点。定义用户 $u_i$ 签到地点类型为 $(t_{i1}, t_{i2}, \cdots, t_{iN})$，每个地点签到的次数分别为 $(c_{i1}, c_{i2}, \cdots, c_{iN})$，总的签到次数为 $C_i$，设共有 $L$ 个用户，每个用户在地点 $i$ 签到的次数为 $(T_{i1}, T_{i2}, \cdots, T_{iL})$，定义用户 $k$ 在地点 $t_{ik}$ 签到的概率为 $p(k)$。引入地点信息熵[14]的概念，定义如式(5)所示，则如果某个地点访问的用户数越多，每个用户访问的概率就越小，该地点的信息熵就越大，说明该地点越开放。

$$E(t_i)=\sum_{i=1}^{L}-p_i(k)\log^{p_i(k)} \tag{5}$$

定义用户签到地点类型属性为 $a_t$，用户在地点信息熵小的地方有共同签到的人更有可能成为朋友，如用户 $a$ 的住宅，用户 $a$ 签到的次数较多，其他用户签到次数较少，这个地点的地点信息熵小，为私密地点，若用户 $b$ 也在该地点签到，则用户 $b$ 很可能是用户 $a$ 的朋友或者用户 $b$ 成为用户 $a$ 的朋友的概率更大。在地点公开程度较高的地方，如公共汽车站，两个用户虽然同时在该地方签到，但偶然性的概率非常大。本文忽略地点信息熵大于 5 的地点记录。则对每个用户 $i$ 和 $j$，其对应的用户边为：

$$a_t(i,j)=\sum_{m=1}^{M}\sum_{\substack{n=1\\ t_{im}=t_{jn}\,且\,E(t_{im})<5}}^{N}(c_{im}+c_{jn})/(C_i+C_j) \tag{6}$$

### 3.3 用户签到地点

定义用户 $u_i$ 签到的地点序列 $(l_{i1}, l_{i2}, \cdots, l_{iN})$，每个地点签到的次数分别为 $(c_{i1}, c_{i2}, \cdots, c_{iN})$，总的签到次数为 $C_i$。$Dist(l_{im}, l_{jn})$ 表示用户 $i$ 的第 $m$ 个签到地点和用户 $j$ 的第 $n$ 个签到地点之间的距离，本文认为签到距离在 0.3km 以内为同一个地方。在距离相同时签到的次数越多成为朋友的可能性越大。如果两个用户经常签到的地点相近，则表明他们是邻居或在相同的地方工作，否则表明他们只是偶然碰见。定义不同用户之间签到地点的距离为属性 $a_l$，同时考虑减少签到次数多的用户的影响，则对用户 $i$ 和 $j$，其对应的用户边为：

$$a_l(i,j)=\sum_{m=1}^{M}\sum_{\substack{n=1\\ Dist(l_{im},l_{jn})<0.3}}^{N}(c_{im}+c_{jn})/(C_i+C_j) \tag{7}$$

## 4 预测方法

本文将用户社交拓扑网络视为完全图，图中的节点表示用户，节点间的连线表示用户之间的关系，实线为朋友关系，虚线为非朋友关系。将用户之间的社交关系置于用户社交拓扑网络中考虑，提取出的社交拓扑网络中的边用选取的 3 个特征属性表示，用分类算法融合 3 个特征建立模型，以拓扑网络中已经存在的朋友边和非朋友边数据训练模型，把要预测的用户关系作为测试数据，将边分为朋友边和非朋友边，以达到朋友关系预测的目的。

本文选取 3 种典型的分类算法：随机森林[25]、支持向量机[26]和朴素贝叶斯[27]建立模型，其中随机森林和朴素贝叶斯采用 Weka 工具[28]实现，支持向量机采用 LibSVM 工具[29]实现。

## 5 实验结果与分析

分类算法只有在选取的特征属性较好地表征了用户的社会关系时才能做出有效的预测。本文采用信息增益选取特征，在用分类算法建模前选取 Bayes 模型验证选取特征的有效性，之后用分类算法融合选取的特征在两个真实的社交网络数据集上进行朋友关系预测。

### 5.1 特征有效性验证

信息增益虽然能从整体上评估特征属性的贡献，但是，用户社交拓扑网络有很多方面的特征，穷举所有的特征是不现实的。本文利用从社交拓扑网络中提取的 3 个特征属性，建立 Bayes 模型进行用户社交拓扑网络重现。若利用现有的社交拓扑网络采集的数据经过 Bayes 模型后能较完整地重现拓扑网络，则证明选取的特征属性能够很好地表征社交拓扑网络。

#### 5.1.1 Bayes 模型

定义 $|E_s|$ 表示朋友边总数，$|AE_s|$ 表示用户边总数，$a_v$ 是用户边的某一个属性值，$P_f(a_v)$ 表示朋友边中 $a_v$ 出现的概率，$P_{nf}(a_v)$ 表示非朋友对中 $a_v$ 出现的概率。在线社交网络中整个社交图中用户之间的朋友边是很稀疏的即 $|E_s| \ll |AE_s|$，因此有 $p(a_v)\approx p_{nf}(a_v)$，由贝叶斯定理公式可知，属性值为 $a_v$ 时用户 $u_1, u_2$ 之间的边为朋友边的概率为：

$$P(u_1u_2|a_v)\approx\frac{P(a_v|u_1u_2)\times P(u_1u_2)}{P_{nf}(a_v)}=\frac{P_f(a_v)}{P_{nf}(a_v)}\times\frac{|E_s|}{|AE_s|} \tag{8}$$

定义 $F(a_v)=\dfrac{P_f(a_v)}{P_{nf}(a_v)}$，针对本文选取的 3 个特征属性，单独计算朋友概率为 $P_1, P_2, P_3$，假设选取的 3 个属性是相互独立的，忽略 $O(P^2)$ 和 $O(P^3)$，则有：

$$P(u_1u_2|a_sa_ta_l)=P_1+P_2+P_3=(F_1+F_2+F_3)\times|E_s|/|AE_s| \tag{9}$$

#### 5.1.2 有效性验证结果

本文选取 Foursquare 中巴黎的数据，从由用户组成的社交拓扑网络中提取特征，根据选取的特征，利用 Bayes 模型重现社交拓扑网络这样的方法来验证所选取特征的有效性。

随机选取 Foursquare 中在巴黎签到的 2731 位用户的数据，以当前的社交拓扑网络为基准值，根据式(9)建立 Bayes 概率模型，可知两个用户之间的边是否是朋友边的概率正比于其用户对 3 个属性 $F$ 值之和。图 1 所示为 Bayes 模型预测用户关系的 ROC 曲线图，其面积为 0.9401，由数据选取的随机性说明选取的 3 个特征属性能较好地重现用户社交拓扑网络，从而证明本文选取的 3 个特征属性很好地表征了社交拓扑网络的特征。

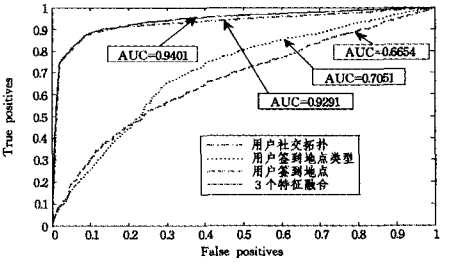


图 1 Bayes 模型预测 ROC 曲线

为了验证所选特征在其它社交网络中的适用性，另选择国内知名社交网络——街旁进行了测试。本文随机选取街旁数据中在北京有签到数据的 2271 位用户的朋友关系数据及其签到数据，对每个用户计算相应的 3 个特征值，根据式(9)

建立 Bayes 概率模型，得出对应 ROC 曲线图，如图 2 所示，其 ROC 曲线面积为 0.8762，说明本文提出的 3 个特征较好地重现了街旁数据中的用户社交拓扑网络，同时也说明本文选取的特征在不同社交网络上具有相同的适用性。

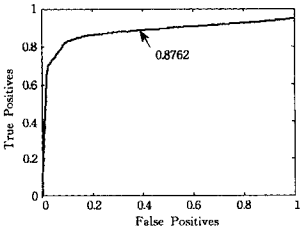


图 2　街旁数据预测 ROC 曲线

## 5.2　朋友关系预测

### 5.2.1　挖掘潜在朋友关系

朋友关系预测的一个重要衡量标准就是能挖掘社交拓扑网络中潜在的朋友关系，在已有的社交拓扑网络中先随机删除部分朋友边，然后根据删除朋友边后的社交拓扑网络计算边的特征属性值，将删除的朋友边数据以及随机选取的部分数据作为测试数据，随机选取一定的朋友边和非朋友边数据，采用分类算法建立模型，根据模型对测试数据分类的结果检测模型挖掘社交拓扑网络中潜在朋友关系的性能。

本文分别采用 Foursquare 和街旁两个社交网络的数据来验证挖掘潜在朋友关系的准确性。社交拓扑网络一个相同的特点是朋友边相对很稀疏，以 Foursquare 中在巴黎签到的数据为例，签到用户 2731 人，朋友边为 5590，非朋友边为 3722225，数据的不平衡度[30]大于 1∶700。文献[31]研究了训练数据集的类分布与分类性能的关系，结果表明，相对平衡的类分布会得到较理想的结果。因此，本文选择训练数据中朋友边与非朋友边比例为 1∶3。分别选取 Foursquare 中巴黎签到数据和街旁中上海签到数据，以当前的用户社交网络作为基准值，分别随机删除 5%和 10%的朋友边，验证选取的特征属性挖掘拓扑网络中潜在用户关系的性能。采用欠抽样[30]的方法随机选取训练数据，分别采用随机森林、SVM 和 Naïve Bayes 3 种分类算法建立模型，对测试数据进行测试。采用重抽样[30]的方法，重复多次随机选取训练数据，建立模型对测试数据进行测试，测试模型预测的平均精度和平均召回率。

实验结果如表 2 和表 3 所列，当删除朋友边高达 10%时该模型依然取得较高的精度和召回率，结果表明以社交网络中的用户边为研究对象进而对用户关系进行预测取得了较好的预测结果。Foursquare 和街旁数据相比，朋友预测精度低而召回率高，说明选取的特征属性在巴黎签到数据中能挖掘较多的潜在用户关系，但可信度不如选取的街旁数据的高。总的来说，选取的特征属性建立的分类模型能很好地挖掘出用户拓扑网络中潜在的用户关系。

表 2　巴黎签到数据潜在用户关系挖掘结果

| Foursquare 中巴黎签到数据 | | 精度 | 召回率 | F 值 |
|---|---|---|---|---|
| 随机删除 5%朋友对 | RandomForest | 0.8012 | 0.8724 | 0.6463 |
| | SVM | 0.7902 | 0.8882 | 0.6467 |
| | NaïveBayes | 0.8137 | 0.871 | 0.6486 |
| 随机删除 10%朋友对 | RandomForest | 0.7895 | 0.9138 | 0.6508 |
| | SVM | 0.7728 | 0.9359 | 0.6506 |
| | NaïveBayes | 0.7919 | 0.8831 | 0.6461 |

表 3　上海签到数据潜在用户关系挖掘结果

| 街旁中上海签到数据 | | 精度 | 召回率 | F 值 |
|---|---|---|---|---|
| 随机删除 5%朋友对 | RandomForest | 0.9632 | 0.7482 | 0.6489 |
| | SVM | 0.9833 | 0.7521 | 0.6528 |
| | NaïveBayes | 0.9852 | 0.6191 | 0.6165 |
| 随机删除 10%朋友对 | RandomForest | 0.9732 | 0.74 | 0.6483 |
| | SVM | 0.7728 | 0.9916 | 0.76159 |
| | NaïveBayes | 0.9927 | 0.6187 | 0.61674 |

### 5.2.2　交叉验证

采用交叉验证的方法是以一个社交拓扑网络中的数据建立模型，以另一个社交拓扑网络中的数据做测试，然后反过来以作测试的社交拓扑网络为训练数据，以训练的社交拓扑网络为测试数据。

对于交叉验证，选取 Foursquare 中巴黎签到的数据(2731 位用户)和伦敦签到的数据(5665 位用户)以及街旁上北京签到的数据(3656 位用户)和上海签到的数据(5275 位用户)进行交叉验证。由于社交拓扑网络中朋友边都相对比较稀疏，提取用户边数据的不平衡度较大，将选取的朋友边与非朋友边的比例定为 1∶3，使训练数据保持相对的平衡，减小数据不平衡对分类结果的影响[31]。

实验结果如表 4 和表 5 所列，采用 Foursquare 数据在进行交叉验证时 3 种算法的结果各有差异，SVM 效果最好，Naïve Bayes 效果较差。街旁数据中的实验结果表明 3 种算法预测的结果相差不大，精度都很高，召回率和模型预测的正确率也可以接受。通过两组交叉验证结果看出，尽管不同的分类算法的结果存在一定差别，但总体的预测精度已达到预期。说明本文提出的基于网络拓扑特征融合的朋友关系预测方法能较好地预测用户之间的朋友关系。

表 4　巴黎签到数据和伦敦签到数据交叉验证

| Foursquare 签到数据 | | 精度 | 召回率 | F 值 |
|---|---|---|---|---|
| Paris train /London test | RandomForest | 0.6431 | 0.7581 | 0.5899 |
| | SVM | 0.7723 | 0.6668 | 0.5982 |
| | NaïveBayes | 0.6432 | 0.8282 | 0.6017 |
| London train /Paris test | RandomForest | 0.5942 | 0.8304 | 0.5885 |
| | SVM | 0.8225 | 0.9205 | 0.6591 |
| | NaïveBayes | 0.5666 | 0.6798 | 0.5559 |

表 5　北京签到数据和上海签到数据交叉验证

| 街旁签到数据 | | 精度 | 召回率 | F 值 |
|---|---|---|---|---|
| Beijing train /Shanghai test | RandomForest | 0.9906 | 0.7453 | 0.6522 |
| | SVM | 0.9906 | 0.7496 | 0.6532 |
| | NaïveBayes | 0.9906 | 0.741 | 0.6511 |
| Shanghai train/ Beijing test | RandomForest | 0.9899 | 0.7273 | 0.6475 |
| | SVM | 0.9895 | 0.7493 | 0.6529 |
| | NaïveBayes | 0.9895 | 0.741 | 0.6509 |

## 5.3　其他模型比较与特征分析

本文选取 Foursquare 数据中巴黎签到的数据来选取本文所述的 3 个特征，采用 Li 等[13]方法建立多层朋友关系模型，模型预测结果如图 3 所示。从图可知，随着 $F$ 值的增大，预测精度始终没有超过 0.4，而召回率几乎减为 0，显然，本文的模型预测结果要比多层朋友关系模型好。

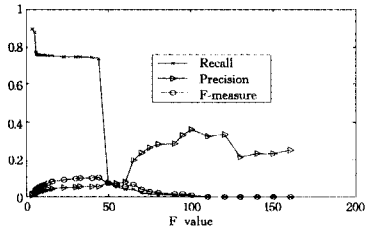


图 3　多层朋友关系模型预测结果

为了验证各个特征对朋友关系预测的影响，本文选取不同的特征组合建立模型，结果如图4所示，其中a表示用户社交拓扑特征，b表示用户签到地点类型，c表示用户签到地点。由图可知，用户的社交拓扑特征对用户的关系影响更大，说明用户的在线行为能较真实地反映用户之间的社交关系。

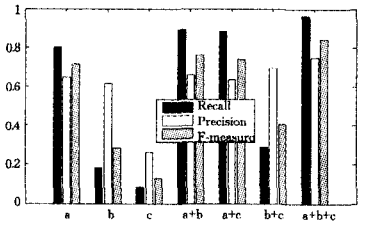


图4 各个特征组合对朋友关系预测的影响

**结束语** 本文提出了基于用户拓扑网络提取用户线上和线下特征属性的方法，并利用Bayes模型重现社交拓扑网络验证特征属性选取具有代表性。分别选取了国内和国外具有代表性的基于位置的社交网络Foursquare和街旁中实际的签到数据对特征属性的有效性和朋友关系预测的准确性进行了验证，均取得了较好的实验结果。与传统的方法相比，本文方法不需要建立专有的模型，同时能取得较高的预测正确率。

本文目前只考虑用户社交拓扑、用户签到地点类型和用户签到地点3个属性，下一步将融入签到时间等特征，采用更全面的信息建立模型来对用户关系进行预测。另外，对用户未来可能签到的地点进行预测，也将是下一步工作的重点。

## 参考文献


[1] Leskovec J, Lang K J, Dasgupta A, et al. Statistical properties of community structure in large social and information networks [C]// Proceedings of the 17th international conference on World Wide Web. 2008:695-704

[2] Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks[C]// Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. 2007: 29-42

[3] Wakita K, Tsurumi T. Finding community structure in Mega-scale social networks[C]// Proceedings of the 16th International Conference on World Wide Web. 2007:1275-1276

[4] Kwak H, Choi Y, Eom Y H, et al. Mining communities in networks: A solution for consistency and its evaluation[C]// Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. 2009:301-314

[5] 窦炳琳，李澍淞，张世永. 基于结构的社会网络分析[J]. 计算机学报，2012(4):741-753

[6] 谈嵘，顾君忠，杨静，等. 移动社交网络中的隐私设计[J]. 软件学报，2010:298-309

[7] Cho E, Myers S A, Leskovec J. Friendship and mobility: User movement in location-based social networks[C]// KDD 2011. 2011:1082-1090

[8] Schwartz M F, Wood D M. Discovering shared interests using graph analysis[J]. Communications of the ACM, 1993, 36(8): 78-89

[9] Grob R, Kuhn M, Wattenhofer R, et al. Cluestr: Mobile social networking for enhanced group communication[C]// Proceedings of the ACM 2009 International Conference on Supporting Group Work. 2009:81-90

[10] Sadilek A, Kautz H, Bigham J P. Finding Your Friends and Following Them to Where You Are[C]// Fifth ACM International Conference on Web Search and Data Mining. 2012:723-732

[11] Gregory S. An algorithm to find overlapping community structure in networks[C]// PKDD 2007. 2007:91-102

[12] Ozseyhan C, Badur B, Darcan O N. An Association Rule-Based Recommendation Engine for an Online Dating Site[C]// Communications of the IBIMA. 2012

[13] Li N, Chen G L. Multi-Layered Friendship Modeling for Location-Based Mobile Social Networks[C]// Int'l. Conf. Mobile and Ubiquitous Systems: Computing, Networking and Services. Toronto, Canada, July 2009:1-10

[14] Cranshaw J, Toch E, Hong J, et al. Bridging the gap between physical location and online social networks[C]// Ubicomp. 2010:119-128

[15] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报，2010，39(5)：651-661

[16] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks[C]// Journal of the American Society for Information Science and Technology. 2007:1019-1031

[17] Zhou T, Lv L Y, Zhang Y. Predicting missing links via local information[J]. the European Physical Journal B, 2009(10):623-630

[18] Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks[C]// Proceedings of the 19th International Conference on World Wide Web. 2010:641-650

[19] Mitchell T M. Machine Learning[M]. The Mc-Graw-Hill Companies, Inc., 1997

[20] Shannon C E. A Mathematical Theory of Communication [J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1):3-55

[21] [Online]. https://developer.foursquare.com/docs

[22] [Online]. https://dev.twitter.com/docs

[23] [Online]. http://dev.jiepang.com

[24] Leskovec J, Backstrom L, Kumar R, et al. Microscopic evolution of social networks[C]// Proceeding of the ACM KDD. August 2008:462-470

[25] Breiman L. Random forests[M]// Machine Learning. 2001:5-32

[26] Cortes C, Vapnik V. Support-vector network [M]// Machine Learning. 1995:273-297

[27] Zhang H. The Optimality of Naive Bayes [C]// Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference. AAAI Press, 2004

[28] Holmes G, Donkin A, Witten I H. Weka: A machine learning workbench[C]// Proceedings of the 1994 Second Australian and New Iealand Conference on Intelligent Information System. Brisbane, QID, 1994:357-361

[29] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011(4)

[30] 翟云，杨炳儒，曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学，2010，37(10)：27-31

[31] Weiss G, Provost F. Learning when training data are costly: The effect of class distribution on tree induction[J]. Journal of Artificial Intelligence Research, 2003, 19:315-354