LBSN中位置信息与网络拓扑相融合的好友预测

潘 果 徐雨明
(湖南大学信息科学与工程学院 长沙 410082)

摘 要 在基于位置的社会网络中,好友预测通常通过相似性标准来衡量用户间的相似性,然后将最相似的用户作为好友推荐给指定用户。传统的用户特征选取没有区分各个特征之间的差异,因而不能很好地代表用户的整体特征。提出了一种位置信息与社会网络拓扑相融合的好友预测方法。首先通过信息增益方法选取出更能代表用户整体特征的3个相关特征,然后对选取的特征进行融合,最后采用分类方法进行好友的预测。实验表明,提出的模型不依赖于具体的分类算法,并且预测性能优于多层好友模型。

关键词 位置,拓扑,推荐,链接预测
中图法分类号 TP319 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.09.022

Friends Prediction Based on Fusion of Topology and Location in LBSN

PAN Guo XU Yu-ming
(College of Information Science and Engineering, Hunan University, Changsha 410082, China)

Abstract In location based social networks, friends prediction usually predicts friends with some similarity metric, and recommends the most similar users to some user. Traditional selection methods of user features don't consider the difference between different features, so they cannot represent the overall features of users. This paper proposed a friends prediction method based on fusion of location information and social topology. First, we selected three relative features that can represent the overall user feature with information gain, then fused the selected relative features, and finally predicted friends with classification method. The experiments show that the proposed method doesn't depend on the concrete classification method, and performs better than the multi-layer friend model.

Keywords Location, Topology, Recommendation, Link prediction

在基于位置的社会网络(Location based Social Networks,LBSN)、基于用户之间的社会关系网络中,用户可以发现好友,搜索带位置标签的内容,以及与附近的好友进行面对面的交流沟通。LBSN除了可以维持用户在虚拟网络中的联系,还可以增强用户与现实世界的联系。近些年,基于位置的社会网络蓬勃发展起来,各大网站都收集了大规模的用户基于位置的状态更新。面对海量的用户位置信息,结合用户间的社会网络关系,研究人员开展了大量的研究工作,如社团挖掘[1]、隐私保护[2]、好友预测[3],以及位置推荐[4]等。

从直观上进行理解,好友预测是根据用户间的相似性对其他用户进行排名,认为与某用户最相似的用户很可能在未来成为该用户的好友,即两个用户越相似,他们在未来成为好友的可能性越大[5]。普遍认为,LBSN中用户间的网络联系以及他们在网上的行为从某种程度上反映了他们在现实生活中的行为。换句话说,那些有着相同兴趣爱好并且在地理位置上离得很近的人,或者有着相同的社会交际圈的人(朋友的朋友)往往很容易成为好友。那么,什么样的度量标准能刻画出上述特征呢?一种可选的方式是选取出这些特征,并对它们进行量化。于是,好友预测的关键变成了如何选取衡量用户相似性的特征。当前,主流的方法是用不同的相似性标准对用户的相似性进行量化,并将具有高相似性的用户对作为好友推荐给对方。然而在实际上,由于相似性定义的局限,相似的用户不一定意味着他们会成为好友。此外,在LBSN中,用户含有很多特征和属性,用户的整体相似性并不是这些特征和属性的简单罗列。在这种方法中,为了提高预测的准确性,研究人员必须为每个选定的特征集合建立详尽的模型。与衡量用户间的相似性不同,本文提出一种基于分类的好友预测方法,该方法对用户的线上和线下特征进行融合,然后完成好友的推荐。

近些年,随着基于位置服务的社交网络的盛行,各大网站,如Facebook Place、FourSquare、Jiepang和Dianping,收集了海量的带位置标签的数据。本文收集了Foursquare和Jiepang两个网站的数据来进行好友预测的实验。普遍认为,用户的行为往往与社会网络拓扑中好友的行为相关。本文对LBSN中的不同特征进行了深度分析,并选取了用户间的网络拓扑和用户的签到(Check-in)特征作为预测好友的相关特

到稿日期:2013-11-03 返修日期:2014-02-05 本文受国家自然科学基金项目(61133005,90715029,61070057,61370095),湖南省科技计划项目(2013GK3082),湖南省教育厅资助项目(08D092)资助。

潘 果(1976—),女,博士,副教授,主要研究领域为并行计算、大数据等;徐雨明(1966—),男,博士,副教授,主要研究领域为并行计算、大数据、网络等。

征。对于用户的签到信息，分析了用户的签到位置和这些位置构成的列表。基于此，提出一种对用户的线上和线下特征进行融合的基于分类的好友推荐算法。

1 相关工作

随着社会网络的发展，好友预测引起了研究人员的广泛关注。Schwartz 和 Wood [6] 提出了一种兴趣距离(Interessst distance)的相似性度量方法，该方法通过对两个用户的好友列表进行相似性分析，并通过相似性将相似的用户聚类成社团。Gregory 等人[8] 通过层次聚类来发现网络中的重叠社团。Grob 等人[7] 应用该算法从相同的聚类中选取用户并迭代式地进行好友推荐。这几种算法都是基于单个用户特征(如社会网络结构或用户兴趣)的好友推荐算法。

在基于多个用户特征的好友推荐中，Özseyhan 等人[13] 基于用户的年龄、位置、收入等特征，应用关联规则挖掘出用户的潜在好友并推荐给用户；Li 等人[10] 研究了用户在现实世界中分享位置的特点，并基于用户的移动性特征、社会网络属性、用户资料，通过对用户间的好友关系进行量化，建立了一个多层的好友模型；Cranshaw 等人[11] 通过分析用户的位置列表，基于与位置相关的特征集合，建立了一个好友预测模型；Sadilek 等人[12] 应用回归决策树对文本相似性系数和相同位置进行分析，以此来提高好友预测的性能。这些方法除了应用用户的社会网络结构数据外，还应用了用户的多个特征来进行相似性分析，它们都假设如果两个用户关于选定的特征相似，那么这两个用户就相互为好友关系。

此外，一些研究人员通过用户的线下交互计算用户间的亲密程度。例如，Guo 等人[13] 根据用户之间在现实世界中会面的次数与持续时间来计算用户的亲密度，并据此进行好友的预测。

上述方法或者采用用户间的网络拓扑结构，或者采用用户的线下交互行为进行好友的预测，本文通过将用户间的网络拓扑和线下特征相融合来进行好友的预测。

2 预测算法设计

2.1 特征选择

众所周知，好友预测的关键是选取合适的特征对用户进行整体的描述。本文假设 LBSN 的社会网络是一个完全图，其中节点表示用户，边表示用户间的关系，实边表示两个用户相互为好友，虚边表示两个用户并非好友关系。令无权无向图 $G_s(U_s,E_s)$ 为 LBSN 的社会网络拓扑图，其中节点 $u_s \in U_s$ 表示用户，边 $(u_1,u_2) \in E_s$ 表示用户 u_1 和 u_2 之间的关系。

由于信息增益是决定某个特征的相关性的重要度量标准[21]，本文使用信息增益来衡量特征在好友预测时的贡献。对于目标特征 X，选定特征 Y，信息增益 $IG(X,Y)$ 等于 X 的信息熵减去通过学习特征 Y 的状态得到的条件熵，即

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

假定 X 有 n 个结果 $\{x_1,x_2,\cdots,x_n\}$，$p(x_i)$ 是结果 x_i 的概率质量函数，X 的信息熵的定义如式(2)所示。

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

令 $p(y_i)$ 是结果 y_i 的概率质量函数，$p(x_i,y_i)$ 是 $X=x_i$ 并且 $Y=y_i$ 的概率，条件熵 $H(X|Y)$ 的定义如式(3)所示。

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

2.1.1 用户社会网络拓扑

在社会网络拓扑结构中，如果两个用户在当前拥有相同的好友，那么他们将比那些没有共同好友的用户在未来更可能成为朋友。如果两个用户在当前是好友，但是他们有很少甚至没有共同的好友，那么他们的好友关系很弱。为了度量这种情况，本文引入了社会距离(Social distance)。社会距离 $a_s(i,j)$ 表示用户 u_i 和 u_j 在网络拓扑中的最短距离，定义如式(4)所示。

$$a_s(i,j)=shorest\_dis(u_i,u_j)\ \text{in}\ G'_s(U_s,E_s-e_{ij}) \tag{4}$$

如果用户 u_i 和 u_j 相互为好友，那么他们之间的社会距离为 1。在计算过程中，如果用户 u_i 和 u_j 相互为好友，首先在网络中移除 e_{ij}，然后计算他们在剩余网络 G'_s 中的最短距离。

2.1.2 位置列表

用户在 LBSN 中的签到列表从某种程度上反映了他们的行为或者爱好。多个用户在不同的位置偶遇具有不同的含义。例如，如果用户经常在私人场所见面，那么他们成为好友的可能性要比那些经常在公共场所见面的可能性要大。为了量化用户的偏好与签到位置的语义之间的关系，本文定义了用户签到的位置列表。用户 u_i 的位置列表为 $(t_{i1},t_{i2},\cdots,t_{iT})$，该用户在位置列表中相应位置的签到次数为 $(c_{i1},c_{i2},\cdots,c_{iT})$，用户 u_i 的总的签到次数为 C_i。如果总的用户数为 L，$p_i(k)$ 表示用户 u_i 在位置 t_{ik} 下的概率，那么位置信息熵的定义如式(5)所示。

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

式中，如果多个用户在位置 l_1 签到过，并且他们都在 l_1 签到过多次，那么 l_1 很可能是一个公共场所，并且 l_1 的信息熵很大。那些有着共同签到位置的用户在未来很可能成为朋友。以用户 A 的家为例，A 在自己家里签到的次数最多，如果用户 B 有时也在 A 的家里签到，那么 B 很可能是 A 的朋友。如果 A 和 B 都在公共场所签到，那么很可能是巧合。对于每对用户 i 和 j，位置列表属性 a_t 的定义如式(6)所示。

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

2.1.3 签到位置

那些有着较多相同签到位置的用户往往比那些有着较少签到位置的用户更容易成为朋友。当共同签到位置的次数占有很大的比例时，这种情况更为明显。假设用户 u_i 的签到位置为 $(l_{i1},l_{i2},\cdots,l_{iH})$，$Dist(l_{im},l_{jn})$ 为用户 u_i 在签到位置 l_{im} 和用户 u_j 在签到位置 l_{jn} 之间的距离。设想如下情况，如果两个签到位置的距离小于 300 米，那么这两个签到很可能是同一地点，因此本文忽略了 300 米以上的距离，该特征的描述如式(7)所示。

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

2.2 好友预测

在 LBSN 中，以用户为节点，用户之间的关系为边，构成了一个社会拓扑图。图中的每个用户有一个签到列表，并且

每个签到位置都有自己的分类(如医院、学校、餐馆等)。本文从3个不同的层次对用户间的关系进行量化,每对用户间包含用户社会网络拓扑、位置列表和签到位置3个特征。在这3个特征中,用户社会网络拓扑是在线特征,其它两个特征是线下特征。

本文对所选特征进行融合,将社会网络拓扑中的边分为好友和非好友两种,并将该模型用于好友的推断。为了表明本文所选的特征与具体的分类算法无关,本文采用随机森林、支持向量机和朴素贝叶斯3种分类算法,其中随机森林和朴素贝叶斯算法在Weka上实现,支持向量机采用LibSVM实现。

3 实验

本文的实验采集了Foursquare和Jiepang两个LBSN上的部分数据。Foursquare数据集的采集从2011年10月24日到12月15日,共包含720000个匿名用户和3百万个地点。Jiepang数据集的采集从2012年4月1日到2012年6月1日,共包含89936个匿名用户和2百万个签到信息。这两个数据集都包含社会网络拓扑签到时间、地理坐标(经纬度)和签到地点的分类信息。

3.1 信息熵验证

实验选取了Foursquare数据集上的2731个用户,以及这些用户的状态更新,并观察相关特征的信息增益。在选取的数据中,好友的节点对为5590,非好友的节点对为3722225。本实验从社会网络拓扑和地理特征上选取了5个用户特征,表1列出了所选取的用户特征及其信息增益。从实验的结果可以看出,本文选取的社会网络拓扑、位置列表和签到位置是信息增益相对较高的用户特征,从而更能表达用户的整体特征。

表1 各个特征的信息增益

特征	信息增益
用户社交网络	0.0055
位置列表	0.0021
签到位置	0.0012
用户签到的位置对	0.0004
签到次数	0.0002

3.2 特征提取

令$|AE_s|$表示网络中总的节点对个数,$|E_s|$表示好友的数目,$P_f(a_v)$表示特征a_v关于好友对的概率,$P_{nf}(a_v)$表示特征a_v关于非好友对的概率。由于社会网络拓扑的好友对是稀疏的,即$|E_s| << |AE_s|$,可得$P(a_v)=P_{nf}(a_v)$。由贝叶斯理论的公式可知,用户对是好友的概率为:

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

令$F(a_v)=\frac{P_f(a_v)}{P_{nf}(a_v)}$,假设选取的特征之间是不相关的,并忽略$O(P^2)$和$O(P^3)$,于是可得

$$\begin{aligned}P(u_1u_2|a_sa_ta_l)&=P_1+P_2+P_3\\&=(F_1+F_2+F_3)\times\frac{|E_s|}{|AE_s|}\end{aligned} \tag{9}$$

本实验在Foursquare中选取了具有状态更新的2731个用户,根据式(9)建立贝叶斯模型,并应用ROC曲线对算法的排序性能进行评价。通过对比上述3个特征的F值,对用户的好友关系进行排序,结果如图1所示。用户社会网络拓扑的ROC面积是0.9421,这意味着该模型可以很好地对社会网络拓扑进行重建。从图中可以看出,本文所选取的特征可以有效地描述用户的特征。

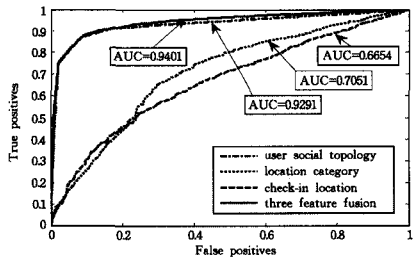

图1 多特征好友排序的ROC曲线

3.3 好友预测

在Foursquare数据集中,选取了在巴黎签到的用户及相关数据,在Jiepang数据集中,选取了在北京签到的用户及相关数据。在得到的子数据集中由于好友关系非常稀疏,对网络拓扑中节点度低的节点进行移除,进一步得到好友关系占所有节点对的比例为1∶3的子数据集。

挖掘潜在的好友关系:在巴黎和北京两个数据集中,分别随机移除了5%和10%的好友关系,并应用分类模型进行好友的预测,结果如图2所示。从图中可以看出,即使移除了10%的好友关系,采用随机森林、支持向量机和朴素贝叶斯3种分类方法在好友预测时都有较好的预测性能。在5%和10%两种情况下,相应的预测性能差距并不大,这说明地理特征在某种程度上可以很好地弥补网络拓扑的不足。

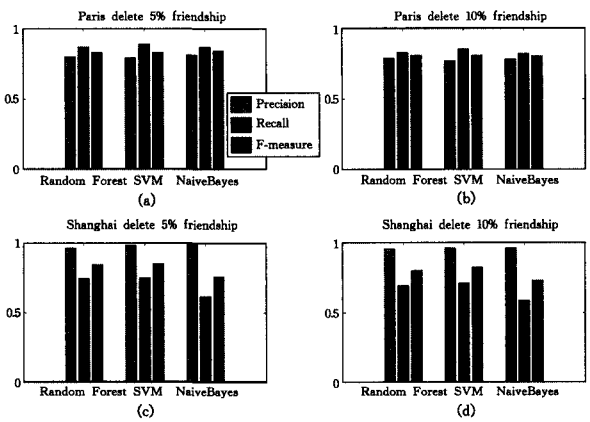

图2 挖掘潜在好友的性能对比

双重交差验证:本实验应用双重交差验证来验证好友预测的可靠性。在Foursquare数据集中,选取了在巴黎签到的2731个用户,并用D1表示;选取了在伦敦签到的5655个用户,并用D2表示。首先应用D1对算法进行训练,应用D2进行验证;然后应用D2进行训练,并应用D1进行验证。在Jiepang数据集中,选取了在北京签到的3656个用户作为D3,选取了在上海签到的5275个用户作为D4。先应用D3进行训练用D4验证,然后用D4进行训练用D3进行验证。实验结果如图3所示。在Foursquare数据集中,SVM的性能最好,朴素贝叶斯次之。在Jiepang数据集中,3种算法都有着较好的性能。这表明本文提出的网络拓扑与位置相融合的

方法不依赖于具体的分类算法，且在不同算法下都有着较好的预测性能。此外，由于采用的是交差验证，在一个地区的训练数据可用于用户其它地区的好友预测。

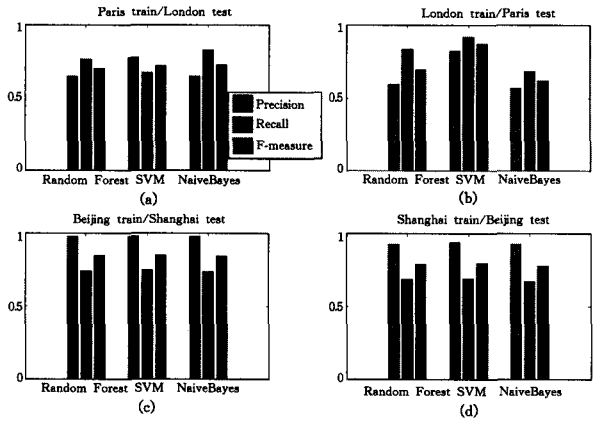

图 3　签到数据的交差验证

3.4　与其它模型的对比

本实验将本文提出的算法与文献[8]的多层好友模型进行对比。在 Foursquare 数据集中，本算法采用与 Multi-model 相同的特征。实验结果如图 4 所示，本文提出的模型有着高的准确率、召回率和 F 值。对于 Multi-model，当 F 值增大时，召回率几乎为 0，这说明该模型能更好地用于非好友关系的预测。

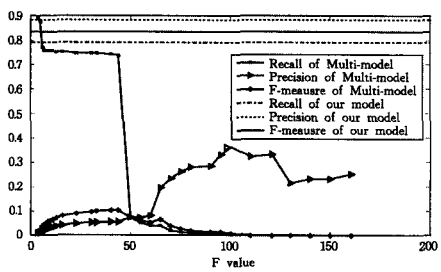

图 4　与多层好友模型的对比图

3.5　特征组合对性能的影响

最后，实验通过对 3 个特征进行不同的组合，进而观察本文提出的模型的预测性能。令 a 表示社会网络拓扑，b 表示位置列表，c 表示签到位置，$a+b$ 表示 a 和 b 的组合，实验结果如图 5 所示。从图中可以看出，该模型在应用社会网络拓扑的在线特征时，模型的预测性能好于其它两个线下特征，这说明用户的在线性能更能代表他们的整体特征。

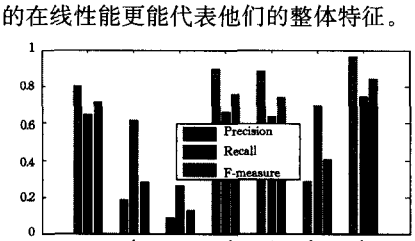

图 5　特征组合的性能对比图

结束语　好友预测是社会网络的重要研究内容之一。常用的好友预测方法往往通过相似性标准衡量用户间的相似性，然后将最相似的用户作为好友推荐给指定用户。本文提出了一种位置信息与社会网络拓扑相融合的好友预测方法。通过信息增益方法，选择出更能代表用户特征的社会网络拓扑、位置列表和签到位置，最后采用分类方法对 3 个特征进行融合从而进行好友的预测。实验表明，本文提出的模型不依赖于具体的分类算法，并且预测性能优于多层好友模型。

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. ACM, 2008: 695-704

[2] 潘晓，郝兴，孟小峰. 基于位置服务中的连续查询隐私保护研究[J]. 计算机研究与发展，2010，47(1)：121-129

[3] Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks[C]// Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 1082-1090

[4] Ye M, Yin P, Lee W C. Location recommendation for location-based social networks[C]// Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2010: 458-461

[5] 邓爱林，朱扬勇，施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报，2003，14(9)

[6] Schwartz M F, Wood D M. Discovering shared interests using graph analysis [C] // Communications of the ACM. August 1993: 78-89

[7] 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. May 2009: 81-90

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

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

[10] Li N, Chen G. Analysis of a Location-Based Social Network [C]// Proceedings of the 2009 International Conference on Computational Science and Engineering. 2009: 263-270

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

[12] 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

[13] Guo B, Zhang D, Yu Z, et al. Hybrid SN: Interlinking Opportunistic and Online Communities to Augment Information Dissemination[C]// The 9th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC'12). Fukuoka, Japan 2012

[14] 荆栋，肖刚. 网络拓扑发现算法[J]. 重庆理工大学学报：自然科学版，2012，26(9)：90-95