

采用位置信息的半监督链接预测方法

朱乔亚,陈可佳,方 彪

(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:链接预测是社会网络分析领域的一个关键问题,如何从网络的已知信息有效预测网络的未知信息面临巨大的挑战。为了有效利用网络中大量未连接的节点及节点对信息,文中将节点的位置信息(签到信息)加入到线社交网络中,并将节点的位置信息引入基于半监督的链接预测方法(LB-SSLP 方法),根据用户之间的关系以及位置签到信息预测用户未来可能的签到位置,同时与传统的 SSLP 方法和 SLP 方法进行对比。在现实数据集 Gowalla 中的实验结果表明,位置信息的引入以及半监督学习的使用均能有效提高链接预测方法的准确率。

关键词:基于位置的网络;链接预测;半监督学习;社会网络分析

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2015)07-0063-04

doi:10.3969/j.issn.1673-629X.2015.07.014

A Semi-supervised Link Prediction Method Using Place Features

ZHU Qiao-ya, CHEN Ke-jia, FANG Biao

(College of Computer, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China)

Abstract: Link prediction is a key issue in the research area of social network analysis, aiming to predict unknown links from the known network. In order to make full use of the information of nodes and node pairs with no connection, add the place features to the online social networks (check-ins) and join it to a semi-supervised link prediction method (LB-SSLP) so that the future possible location of a user can be predicted according to the relationship with other users and his check-ins, at the same time, compare with the traditional SSLP method and SLP method. The experimental results in a real dataset Gowalla show that both the use of place features and semi-supervised learning make the proposed method perform a higher prediction accuracy.

Key words: location-based network; link prediction; semi-supervised learning; social network analysis

0 引 言

随着信息技术的不断发展,形形色色的网络中聚集了海量的数据。对网络数据的挖掘,尤其是对链接的挖掘^[1](link mining)成为一个新兴的研究领域,在商品推荐、生物信息学、在线社交网络、学术论著与引用分析等领域均有广泛的应用。作为链接挖掘的一个分支,链接预测(link prediction)研究从当前已观察到的网络中预测未观察到的网络部分或者预测在未来某时刻可能出现的新链接。不同于传统的数据挖掘任务,链接预测不仅要考虑节点的内容属性还需要考虑节点之间的关系属性。

现实网络具有海量性、稀疏性等特点,这给链接预测问题带来了巨大挑战^[1]。例如,在线社交网络(on-

line social network)一般拥有至少数百万个节点,而链接数往往最多只有数千万个,这导致了巨大而高度不平衡的链接预测空间^[2]。如何有效地缩小预测空间是一个值得探讨的问题。此外,针对网络密度稀疏的现象,如何有效利用网络中大量未连接的节点对的信息帮助预测链接也是一个值得研究的问题。

已有研究表明^[2],在基于位置的社交网络(location-based social network)中,大约 30% 的新链接(即朋友关系)是在访问过同一地点的用户之间形成的。对基于位置的在线社交网络 Gowalla 数据的分析中发现,根据位置朋友(place-friends)选择数据可以缩小约 15 倍的预测空间,而链接数仍占整体链接数的 66%^[2]。因此,文中尝试将位置信息(例如签到信息)

收稿日期:2014-08-20

修回日期:2014-11-21

网络出版时间:2015-06-23

基金项目:国家自然科学基金青年基金项目(61100135,61302158)

作者简介:朱乔亚(1989-),女,硕士研究生,研究方向为机器学习、社会网络分析等;陈可佳,副教授,研究方向为机器学习、数据挖掘、信息检索等。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150623.1028.011.html>

作为节点的重要描述特征,加入并改进已有的链接预测方法。

目前,机器学习已成为数据挖掘与数据分析的核心技术。从机器学习的角度,链接预测可视为:在监督(半监督)学习下,根据已有的网络数据训练出一个模型,再通过该模型对未连接的节点对预测其存在链接的可能性^[3]。由于网络中存在大量未连接的节点对,文中采用了半监督学习(semi-supervised learning)^[4]方法,挖掘未连接节点对中潜在的有用信息,期望进一步提高链接预测的准确率。

1 研究背景

1.1 链接预测

链接预测研究网络图中任意一对未连接的节点之间存在链接的可能性。网络可形式化地表示为图 $G = \langle V, E \rangle$ 。其中, $V = (v_1, v_2, \dots, v_N)$ 表示由 N 个节点组成的节点集 V , 边 $e_{i < \alpha, \beta >} \in E$ 表示节点 v^α 和节点 v^β 在时间 t 存在链接^[5]。

同时,定义向量 $\hat{x}^{(\alpha, \beta)} = \{x_1, x_2, \dots, x_n\}$, 表示任意节点对的特征。该节点对既可以表示用户之间的朋友关系 $\langle v^\alpha, v^\beta \rangle$, 又可以表示用户与位置之间的签到关系 $\langle v^\alpha, \text{loc}_\alpha \rangle$ 。其中, 每个 x_i 既可以描述给定网络中节点 v^α 和 v^β 之间的关系属性, 又可以描述节点 v^α 或 v^β 自身的位置属性。定义 $F = \{f\}$ 表示用户之间或用户与位置之间是否存在链接^[6]的标记集合。

$$f^{(\alpha, \beta)} = \begin{cases} 1, & \langle v^\alpha, v^\beta \rangle \in E \\ 0, & \langle v^\alpha, v^\beta \rangle \notin E \end{cases} \quad (1)$$

链接预测可形式化地描述为在机器学习框架下的二分类问题: 通过训练学习模型 M , 对于任一未连接的节点对 $\langle v^\alpha, v^\beta \rangle$ 计算其链接权(存在链接的概率) $\text{score}^{(\alpha, \beta)} = M(\hat{x}^{(\alpha, \beta)})$, 从而判断链接存在的可能性。

需要说明的是, 文中实现的链接预测方法主要针对基于位置的网络, 用于预测用户节点与位置节点之间未来可能建立的链接。

1.2 节点的位置特征

在基于位置的社会网络中, 链接预测任务在很大程度上受到朋友的朋友或位置朋友的影响^[7], 围绕位置的活动与交互可能因位置的特性而引起个体的社会关系及行为的改变。然而, 如何利用用户的签到位置帮助预测新的链接是一个挑战性较大的问题。首先, 并非所有的位置对不同的用户都有相同的重要性; 其次, 并非所有的位置对于新链接的建立都有相同的作用。

目前, 已有一些研究者基于用户的签到数据来分析其在社交网络中的行为方式^[8]。还有一些工作^[9-10]

采用地理位置信息研究用户的移动性。文中的研究目标为在基于位置的网络中如何为用户预测可能的新的位置, 根据节点对当中一位用户签到位置的属性和频率以及节点对之间的潜在关系属性(例如共同邻居、最短路径等)学习出节点对中另一位用户在同一位置签到的概率。

不同位置对不同用户的重要性也不相同。一个被大量签到的位置对用户形成新链接的影响, 并不一定比被少数用户签到的位置的影响更大。例如, 旅游景点、机场、火车站等公共场所可被大量签到, 而私人住宅、体育场馆、办公室等只被少数用户签到, 但后者对于形成新链接的意义更大。因此, 文中采用了“位置熵(place-entropy)^[11]”作为网络中位置的描述特征, 用来度量社会网络中位置的重要性。

在基于位置的网络 $G = \langle V, E \rangle$ 中, 集合 $P = \{p_1, p_2, \dots, p_L\}$ 表示 L 个不同位置的集合。令 C_k 表示在位置 p_k 签到的用户总数, c_{ik} 表示用户 v_i 到目前为止在位置 p_k 签到的次数, $q_{ik} = c_{ik}/C_k$ 表示用户 v_i 在位置 p_k 的签到总次数与位置 p_k 的所有签到次数的比值。 $\{q_{1k}, q_{2k}, \dots, q_{Nk}\}$ 表示位置 p_k 在整个预测空间上签到概率的离散型概率分布。 Φ_k 表示在位置 p_k 签到的所有用户的集合。从而可以计算位置 p_k 的熵 E_k :

$$E_k = - \sum_{u_i \in \Phi_k} q_{ik} \log q_{ik} \quad (2)$$

一般来说, 在位置 p_k 签到的用户与他人成为朋友的概率与 E_k 的值成反比。例如, 一个仅被用户临时访问的位置不太可能影响其朋友或者与其相关的其他用户在此签到, 对促进用户与位置之间建立新链接的意义有限。文中将位置熵作为一个重要特征, 判断该位置是否有助于新链接的形成。

1.3 半监督学习的引入

半监督学习是基于传统的监督学习与无监督学习之间的一种机器学习方法。其训练样本中既包含已标记样本也包含未标记样本。在半监督学习过程中, 学习器首先根据已标记样本训练出一个粗糙的模型 M , 再用这个模型预测部分未标记样本的标记, 并加入已标记训练样本集中, 在此基础上重新训练模型。这一过程可以重复进行。在条件独立分布的前提下, 初始的分类器通过在训练中使用新样本而得到性能的改变^[12], 具有更好的泛化性。

两种基础的半监督学习范式是由 Nigam 和 Ghahramani^[12]提出的自我训练(self-training)范式和由 Blum 和 Mitchell^[13]提出的协同训练(co-training)范式。前者根据已标记的数据构造一个分类器, 然后将预测概率较高的未标记数据进行标记并加入样本集重新训练直至所有的未标记数据均被标记为止。后者从已标记数

据中划分出两个独立的属性集(或称属性划分),然后用这两个学习器相互标记新样本来进行学习和改进。然而,在现实网络海量数据中,充分独立的属性集较难获得,Goldman 和 Zhou^[14]在此基础上改进了现有的协同训练方法,将样本空间划分成不同等价类,再让两个学习器协同学习。

为了能够使用网络数据中的潜在信息(即大量未连接节点对的信息),文中在基于机器学习的链接预测问题中采用了自我训练的半监督学习技术并进行了初步实验。

2 文中提出的链接预测方法

扩展了节点对二元组,将节点 v^α 、 v^β 以及 v^α 签到的位置 loc_α 组成一个三元组样本。在样本中,如果节点 v^α 在位置 loc_α 已签到并且节点 v^β 在下一时刻也在同一位置 loc_α 签到,则该样本视为正例样本,组成集合 L 。而节点 v^β 暂未在 loc_α 签到的样本则视为未标记样本,组成集合 U 。从 L 和 U 中选出训练集 T 和测试集 C 。其中,最初的训练集 T 中的 $n/2$ 个正例和 $n/2$ 个反例分别以随机取样的方法从 L 和 U 中获得。文中采用 SVM 分类器实现学习模型 M 。

链接预测算法的伪代码描述如下:
输入: $G = \langle V, E \rangle$: 网络图;
 L : v^β 已在与 v^α 签到的位置签过到的样本集合
 U : v^β 暂未在与 v^α 签到的位置签过到的样本集合
 T : 训练集
 C : 测试集
 M : 学习模型
过程:

(1)从 L 中选择 $n/2$ 个样本作为正例,从 U 中选择 $n/2$ 个样本作为反例,一同组成训练集 T
(2)从 T 中抽取训练集样本的所有属性特征,包括网络结构特征和新增加的位置特征
(3)采用 SVM 分类器在 T 上进行训练
(4) repeat until 结果满意
for each $\langle v^\alpha, v^\beta \rangle \in V$ do
计算预测值 $\text{score}(\langle v^\alpha, v^\beta \rangle)$;
end for $i \in 1, \dots, k$ do
 $u \leftarrow \text{argmaxAbs}(\text{score}(\langle v^\alpha, v^\beta \rangle))$;
 $\text{Label}(u) \leftarrow \text{sign}(\text{score}(\langle v^\alpha, v^\beta \rangle))$;
 $U' \leftarrow U - \{u\}$;
 $T' \leftarrow T \cup \{u\}$;
end
在 T' 上重新训练
end
输出: $G = \langle V, E' \rangle$: 预测后的新网络

3 实验

3.1 实验数据与特征抽取

文中在最具代表性的基于位置的在线社交网络 Gowalla 数据集上进行实验。表 1 统计了 2009 年 2 月至 2010 年 10 月期间 Gowalla 网络数据集的朋友关系网络和用户签到的统计信息。经过初步分析,发现超过 99% 的用户签到次数在 50 以下,而 90% 左右的用户签到次数在 10 以下。鉴于数据规模过于庞大,文中的初步实验从原始的数据集中随机抽取了 2010 年 5 月和 6 月各 10 000 名用户的签到信息及其朋友关系网络信息作为实验数据。

表 1 Gowalla 数据集的统计数据

节点数	边数	签到数	签到数≥1的用户数	用户平均签到次数	用户最大签到次数	位置平均签到次数	位置最大签到次数	位置总数
196 591	950 327	6 442 890	107 092	33	2 175	5	5 811	1 280 957

算法中使用了 3 个网络结构特征和 1 个位置特征用以描述节点对样本的属性(见表 2)。

表 2 Gowalla 数据集样本属性的描述

特征	描述
共同邻居数目	与两个节点均相连接的节点总数
最短路径	两个节点间所有路径长度中的最短值
Jaccard 系数	$J(v^\alpha, v^\beta) = \frac{ \Gamma(v^\alpha) \cap \Gamma(v^\beta) }{ \Gamma(v^\alpha) \cup \Gamma(v^\beta) }$ $\Gamma(v^\alpha)$ 表示 v^α 直接邻居的数目
位置熵	$E_k = - \sum_{u_i \in \Phi_k} q_{ik} \log q_{ik}$

网络结构特征的选择基于启发式的认识。一般来说,节点对的共同邻居数目越多,用户之间形成链接的可能性越大,一个用户对另一用户的签到位置的影响

就越大^[15];网络中节点大多以较短路径相连,距离越短的节点,用户之间形成链接的可能性也就越大,未来访问该样本中位置的可能性也越大^[16];网络中的一个节点用户在一定时间间隔内签到了某个位置,该位置熵的值越小,该位置对节点用户的影响就越大,与该用户相关的其他用户未来在此位置形成链接的可能性也越大。图 1 反映了位置熵对链接预测准确率的影响。该图是在实验结果的基础上对位置熵进行统计得到的。

3.2 实验设置

实验中,共实现了三种链接预测方法:未采用位置特征的基于监督学习的链接预测方法(SLP)、未采用位置特征的基于半监督学习的链接预测方法(SSLP),以及采用位置特征的半监督链接预测方法(LB-SS-

LP)。SLP 方法仅采用共同邻居数目、最短路径和 Jaccard 系数这三种结构特征来描述样本且并未采用未标记样本的信息。若 SLP 预测 v^a 与 v^b 是朋友,则认为 v^b 将可能在 v^a 签到过的位置签到;SSLP 方法是在 SLP 方法的基础上用部分未标记样本扩充了训练样本集;LB-SSLP 方法是在 SSLP 方法的基础上扩充了样本的描述特征,加入了位置信息即位置熵。

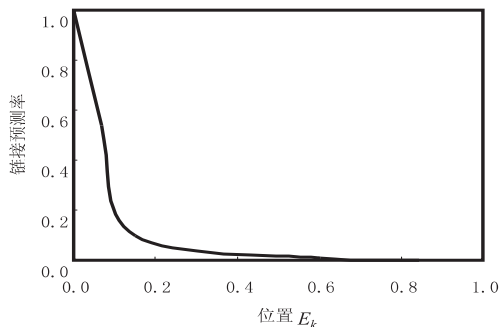


图 1 Gowalla 数据集位置熵与链接预测的关系

实验过程中,以时间标签将数据集分为互不重叠的两个部分。将 2010 年 5 月的数据作为训练样本,2010 年 6 月的数据作为测试样本。因此,出现在测试集而未出现在训练集中的链接应预测为正例,在整个数据集中均未出现的链接应预测为反例。实验分别实现了方法 SLP、SSLP 和 LB-SSLP,并得到了用户与位置存在链接可能性的预测结果。

3.3 实验结果与分析

文中采用 ROC 曲线以及 ROC 下的 AUC 值^[17]来衡量各个方法在实验中的性能。

图 2 绘制了三种方法的 ROC 曲线图。

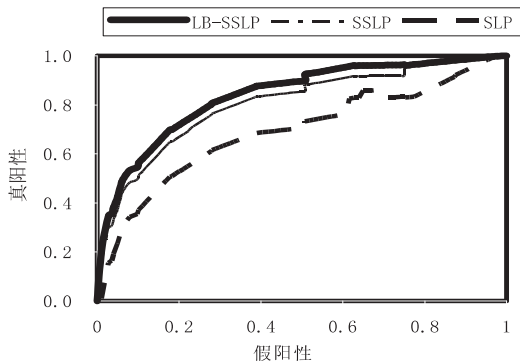


图 2 Gowalla 数据集中三种方法的 ROC 曲线

从图中可以看出,在 Gowalla 数据集上 LB-SSLP 实验结果的 ROC 曲线的位置总体高于 SLP 方法和 SSLP 方法。表明半监督学习方法与传统的监督方法相比可在一定程度上提高链接预测的性能,而在半监督学习方法的基础上加入社交网络的位置特性亦可进一步提高链接预测的性能。此外,计算了 LB-SSLP、SSLP、SLP 三种方法的 AUC 值,分别为:0.84、0.80、0.69。LB-SSLP 方法得到的 AUC 值比 SSLP 方法高出约

4%,比 SLP 高出约 15%。这表明,半监督学习的引入能较大提高链接预测的准确率,而对于基于位置的神经网络,样本的描述中加入了位置信息对于预测也有一定的帮助。

此外,如图 2 所示的三条曲线均在某些数据点出现转折或断链,可能的原因包括:随机取样方法导致了数据分布的改变而产生了噪音数据,以及数据选择的局限性造成正反例数目的不平衡。此外,实验只选取了 Gowalla 数据集中 5 月与 6 月的数据进行了一次实验的结果,采用多次实验取平均值的方法可能会避免或减少转折点数据或断链的现象。

4 结束语

文中提出并评价了一个基于位置的半监督链接预测方法。该方法使用了网络中未连接用户的潜在关系信息以及用户的签到位置信息,达到改进学习器预测性能的目的。在 Gowalla 大规模数据集集中的实验表明, LB-SSLP 在不同程度上优于基准的 SSLP 方法和 SLP 方法。使用未连接节点对数据及在训练样本的描述中加入位置特征均有助于链接预测性能的提高。

文中方法可以在如下方面得到改进:

- (1) 目前 LB-SSLP 方法中的半监督过程仅采用了简单的自我训练范式,未来将尝试更优的半监督方法;
- (2) 仅在 Gowalla 数据集中进行了实验,未来将应用在其他数据集中,观察是否能得到类似的结果;
- (3) 实验中训练样本的选择仅采用了随机取样的方法,未来将采用能保持数据分布的选择取样方法。

文中的链接预测方法主要是预测用户与签到位置之间的关系,在将来的工作中,可将该方法扩展到用户与用户的关系预测之中,考察位置特征是否有助于用户之间朋友关系的预测。

参考文献:

- [1] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.
- [2] Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks[C]//Proceedings of 17th ACM SIGKOD international conference on knowledge discovery and data mining. San Diego: ACM, 2011: 1046-1054.
- [3] 陈可佳, 韩京宇, 郑正中, 等. 主动学习在通信网络推荐系统中的应用[J]. 计算机应用, 2012, 32(11): 3038-3041.
- [4] Cohen I, Cozman F G, Sebe N, et al. Semi-supervised learning of classifiers: theory, algorithms, and their application to human-computer interaction[J]. IEEE Transactions on Pattern

swarm theory [C]//Proceedings of the sixth international symposium on micro machine and human science. Nagoya: IEEE,1995:39-43.

[2] Kennedy J, Kennedy J F, Eberhart R C. Swarm intelligence [M]. [s. l.]:Morgan Kaufmann,2001.

[3] Shi Y, Eberhart R. A modified particle swarm optimizer[C]//Proc of the 1998 IEEE international conference on evolutionary computation. Anchorage: IEEE,1998:69-73.

[4] Eberhart R C, Shi Y. Comparing inertia weights and constriction factors in particle swarm optimization[C]//Proceedings of the 2000 congress on evolutionary computation. [s. l.]: IEEE,2000:84-88.

[5] Clerc M, Kennedy J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space [J]. IEEE Transactions on Evolutionary Computation,2002,6(1): 58-73.

[6] 高 鹰,谢胜利. 基于模拟退火的粒子群优化算法 [J]. 计算机工程与应用,2004,40(1):47-50.

[7] Liang J J, Qin A K, Suganthan P N, et al. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions [J]. IEEE Transactions on Evolutionary Computation,2006,10(3):281-295.

[8] 韩 飞,杨春生,刘 清. 一种改进的基于梯度搜索的粒子群优化算法 [J]. 南京大学学报:自然科学,2013,49(2): 196-201.

[9] 王 立,郑 昊. 粒子群遗传混合算法在点状注记配置中的应用 [J]. 计算机与现代化,2012(10):30-33.

[10] 唐贤伦. 混沌粒子群优化算法理论及其应用研究 [D]. 重庆:重庆大学,2007.

[11] 刘林炬. 引入禁忌搜索的双种群粒子群算法及其应用研究 [D]. 无锡:江南大学,2008.

[12] Li C, Yang S, Nguyen T. A selflearning particle swarm optimizer for global optimization problems [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42 (3):627-646.

[13] 汤继涛,戴月明. 内嵌区域震荡搜索的粒子群优化算法 [J]. 计算机工程与应用,2013,49(21):33-36.

[14] 王莉荣,祁云嵩. 基于函数最优解问题的粒子群算法改进 [J]. 计算机技术与发展,2013,23(2):49-51.

[15] Trelea I. The particle swarm optimization algorithm: convergence analysis and parameter selection [J]. Information Processing Letters,2003,85(6):317-325.

+++++

(上接第 66 页)

Analysis and Machine Intelligence, 2004, 26 (12) : 1553 - 1566.

[5] Getoor L, Diehl C P. Link mining: a survey [J]. ACM SIGKDD Explorations Newsletter, 2005, 7 (2) : 3-12.

[6] Hasan M A, Chaoji V, Salem S, et al. Link prediction using supervised learning [C]//Proceedings of SDM workshop on link analysis, counterterrorism and security. [s. l.]: [s. n.], 2006.

[7] Eagle N, Pentland A, Lazer D. Inferring friendship network structure by using mobile phone data [J]. PNAS, 2009, 106 (36):15274-15278.

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

[9] Eagle N, Bettencourt L M A, de Montjoye Y. Community computing: comparisons between rural and urban societies using mobile phone data [C]//Proceedings of CSE'09. [s. l.]: [s. n.], 2009:144-150.

[10] Nanavati A A, Singh R, Chakraborty D, et al. Analyzing the structure and evolution of massive telecom graphs [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20 (5):703-718.

[11] Cranshaw J, Toch E, Hong J, et al. Bridging the gap between physical location and online social networks [C]//Proceedings of the 12th ACM international conference on ubiquitous computing. Copenhagen: ACM, 2010:119-128.

[12] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training [C]//Proceedings of the 9th international conference on information and knowledge management. [s. l.]: [s. n.], 2000:86-93.

[13] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training [C]//Proceedings of the 11th conference on computational learning theory. [s. l.]: [s. n.], 1998:92-100.

[14] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data [C]//Proceedings of ICML' 2000. [s. l.]: [s. n.], 2000:327-334.

[15] Adamic L A, Adar E. Friends and neighbors on the web [J]. Social Networks, 2003, 25(3):211-230.

[16] Kleinberg J. Navigation in a small world [J]. Nature, 2000, 406 (6798):845-845.

[17] Hand D J, Till R J. A simple generalisation of the area under the ROC curve for multiple class classification problems [J]. Machine Learning, 2001, 45(2):171-186.

采用位置信息的半监督链接预测方法

作者：[朱乔亚](#)，[陈可佳](#)，[方彪](#)，[ZHU Qiao-ya](#)，[CHEN Ke-jia](#)，[FANG Biao](#)
作者单位：[南京邮电大学 计算机学院](#)，[江苏 南京](#)，[210003](#)
刊名：[计算机技术与发展](#)[ISTIC](#)
英文刊名：[Computer Technology and Development](#)
年，卷(期)：[2015 \(7\)](#)

引用本文格式：[朱乔亚](#)，[陈可佳](#)，[方彪](#)，[ZHU Qiao-ya](#)，[CHEN Ke-jia](#)，[FANG Biao](#) [采用位置信息的半监督链接预测方法](#)

[期刊论文]-[计算机技术与发展](#) 2015 (7)