Journal of Computer Applications
计算机应用,2018,38(6):1682-1690,1697
ISSN 1001-9081
CODEN JYIIDU
2018-06-10
http://www.joca.cn

文章编号:1001-9081(2018)06-1682-09
DOI:10.11772/j.issn.1001-9081.2017122904

基于位置的社交网络中基于时空关系的超网络链接预测方法

胡　敏,陈元会,黄宏程*
(重庆邮电大学 通信与信息工程学院,重庆 400065)
(*通信作者电子邮箱 huanghc@cqupt.edu.cn)

摘　要:针对基于位置的社交网络(LBSN)中因现有方法未能有效融合社会因素、位置因素以及时间因素的综合影响而导致链接预测准确度低的问题,提出了一种LBSN中基于时空关系的超网络链接预测方法。首先,针对LBSN中网络的异构性以及用户间的时空关系特性,将网络划分成"时空-用户-位置-类别"四层超网络,降低影响因素间的耦合性;其次,考虑到边权值对网络的影响,通过挖掘用户影响力、隐关联关系、用户偏好以及节点度信息,对子网的边权值进行定义和量化,构建四层加权超网络模型;最后,在加权超网络模型的基础上,定义超边及加权超边结构,挖掘用户之间的多元关联关系进行预测。实验结果表明,所提方法较基于同构和异构的链接预测方法在准确率、召回率、F1值以及AUC上具有一定的提升,其中AUC指标较基于异构的链接预测方法提升了4.69%。

关键词:链接预测;基于位置的社交网络;超网络;影响力;用户偏好
中图分类号:TP393.02　　**文献标志码**:A

Supernetwork link prediction method based on spatio-temporal relation in location-based social network

HU Min, CHEN Yuanhui, HUANG Hongcheng*
(*School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)

Abstract: The accuracy of link prediction in the existing methods for Location-Based Social Network (LBSN) is low due to the failure of integrating social factors, location factors and time factors effectively. In order to solve the problem, a supernetwork link prediction method based on spatio-temporal relation was proposed in LBSN. Firstly, aiming at the heterogeneity of network and the spatio-temporal relation among users in LBSN, the network was divided into four-layer supernetwork of "spatio-temporal-user-location-category" to reduce the coupling between the influencing factors. Secondly, considering the impact of edge weights on the network, the edge weights of subnets were defined and quantified by mining user influence, implicit association relationship, user preference and node degree information, and a four-layer weighted supernetwork model was built. Finally, on the basis of the weighted supernetwork model, the super edge as well as weighted super-edge structure were defined to mine the multivariate relationship among users for prediction. The experimental results show that, compared with the link prediction methods based on homogeneity and heterogeneity, the proposed method has a certain increase in accuracy, recall, F1-measure (F1) as well as Area Under the receiver operating characteristic Curve (AUC), and its AUC index is 4.69% higher than that of the link prediction method based on heterogeneity.

Key words: link prediction; Location-Based Social Network (LBSN); supernetwork; influence; user preference

0　引言

随着计算机信息技术的不断发展和互联网的迅速普及,在线社交平台已经成为人们生活中不可或缺的一部分,人们可以通过这个平台建立自己的好友关系网,与好友进行即时的交流互动,这很大程度上方便了人们的生活;特别是近年来,基于位置的社交网络(Location-Based Social Network, LBSN)的出现,使得一些位置服务在短时间内受到了大量用户的推崇,获得了极大的成功。在LBSN中,用户可以在他去过的位置进行签到,并向好友分享自己的签到地点,这种签到行为能够真实反映用户的位置活动,使得线上虚拟世界和线下物理世界之间建立起密切联系[1],为社会网络链接预测带来了新的机遇和挑战。

链接预测是社会网络分析与挖掘中的一个重要方法[2],其主要目的是挖掘网络中可能存在的缺失链接或者未来链接。目前研究者们已经提出了许多基于同构网络的链接预测方法,刘思等[3]针对随机游走方法存在较强的随机性问题,利用Deep Walk学习网络中各节点间的潜在网络结构相似性,并指导随机游走过程。Duan等[4]通过引入一种集成方法解决了大规模网络下链接预测问题。这些方法在一定程度上解决了同构网络中的链接预测问题,但这些方法并不适用于异构网络。基于位置的社交网络是一类异构网络,异构网络

收稿日期:2017-12-12;**修回日期**:2018-02-02;**录用日期**:2018-02-11。
基金项目:重庆市科委基础与前沿研究计划项目(cstc2014jcyjA40039)。
作者简介:胡敏(1971—),女,重庆人,副教授,硕士,CCF会员,主要研究方向:通信网体系与协议、大数据;　陈元会(1991—),男,湖北武汉人,硕士研究生,主要研究方向:数据挖掘、复杂网络链接预测;　黄宏程(1979—),男,河南南阳人,副教授,博士,CCF会员,主要研究方向:复杂网络分析、智能信息处理。

是指网络中存在不止一种类型的节点和边(例如:位置节点和用户节点、用户-用户边、用户-位置边等);因此,越来越多的研究者开始研究 LBSN 中的链接预测方法。Scellato 等[5]发现在 Gowalla 中 30% 的新链接发生在有过相同签到位置的用户间,进而通过位置特征、社交特征以及全局特征,提高了链接预测的准确性。该方法虽然通过位置关系缓解了数据稀疏性问题,但只考虑了两跳之内的预测空间,使得预测精度存在一定的瓶颈。Valverde-Rebaza 等[6]探索了用户的移动模式和社会模式,提出了内外共同位置(Within and Outside of Common Places,WOCP)、共同邻居位置(Common Neighbors of Places, CNP)、总共和局部重叠位置(Total and Partial Overlapping of Places, TPOP)三种新特征,并通过实验验证了特征的有效性。Bayrak 等[7]考虑到不同类别的位置对于链接建立的影响程度不同,提出了两种新的基于类别的特征,实验表明,新引入的特征提高了链接预测性能。

以上方法主要是在社会因素的基础上引入了位置因素的影响,通过这种额外的"资源"达到了提高预测性能的目的。时间因素对于链接预测也有着一定的影响。Cheng 等[8]通过签到时间间隔、位置熵以及共同位置信息预测用户之间的朋友关系。Crandall 等[9]发现如果两个用户在相同的时间和相同的地点出现过,即使时空共现的次数比较少,也会极大增加他们之间产生链接的概率;Li 等[10]也得到了相同的结论。然而时间因素的引入,使得本就稀疏的网络变得更加稀疏,因此单独考虑这一种因素显得不够合理。

针对数据的稀疏性问题,刘怡君等[11]首次提出了基于超网络中的链路预测方法,通过引入超三角结构作为相似性指标,能够度量不同层网络对链接产生的影响,提高预测性能。方哲等[12]在此基础上提出了一种加权超网络中的链接预测方法,通过加权超边构建加权超三角形结构,并对节点间的链接关系进行预测,实验证明,权值的引入提高了异构网络中的链接预测性能。文献[13-14]也验证了权值的引入对异构网络预测性能有较大影响。但是在基于位置的社交网络中,边权值不仅存在于同构边之间,同时存在于异构边之间,现有大多数研究仅考虑可观测边权值对网络的影响[14-15](如用户对项目的评分、评论次数等),忽略了网络中不可观测的边权值,难以挖掘整个网络的特性,同时目前基于超网络的方法仅仅考虑了简单的超三角结构,缺乏对更深层次超边结构的发现,无法挖掘更多的隐含关系。

针对以上问题,本文提出了一种 LBSN 中基于时空关系的超网络链接预测方法。该方法可以有效融合时间因素、社交因素、位置因素的影响,并且能够合理量化网络边权重,较好地缓解数据稀疏性问题,同时能够提高网络的解释性以及预测性能。

本文的主要工作如下:

1) 针对 LBSN 中网络的异构性以及时空关系特性,提取时空节点,将网络划分成"时空-用户-位置-类别"四层超网络,通过该模型可以将时间因素通过节点的形式有效融入到超网络中,以一种新颖的方式解决了因时间维度缺失所带来的预测精度缺失问题。

2)为了使网络更加符合实际,本文基于用户影响力、隐关联关系、用户偏好以及节点度信息,量化超网络中的相关边权值,构建四层加权超网络模型,边权值的定义和量化使得网络中节点间的关联关系更加准确,有助于提高模型的可解释性以及预测精度。

3)在加权超网络模型的基础上,定义多种类型的超边结构,提出一种 LBSN 中的超网络链接预测方法,通过该方法捕捉用户之间潜在的多种关联关系,有效解决了数据稀疏性问题,同时提高了预测准确度。

1　问题定义

1.1　相关定义

定义 1　时空子网 $G_T = (V_T, E_T, W_T)$。

时空子网 G_T 是基于时空节点和时空节点间的关联构成的网络。其中:V_T 表示时空节点,如果有两个或两个以上的用户在某个特定的时间段共同访问了某个位置,那么该位置就被称为一个时空节点;E_T 表示时空节点之间的有向连边,即两个时空节点之间的关联关系;W_T 表示时空节点之间的关联权重。

定义 2　用户子网 $G_U = (V_U, E_U, W_U)$。

用户子网 G_U 是基于用户节点和用户节点间的关系构成的网络。其中:V_U 表示用户节点;E_U 表示用户节点之间的有向连边,即两个用户节点之间的社会关系;W_U 表示用户节点之间的关联权重。

定义 3　位置子网 $G_P = (V_P, E_P, W_P)$。

位置子网 G_P 是基于位置节点和位置节点间的关联构成的网络。其中:V_P 表示位置节点;E_P 表示位置节点之间的有向连边,即两个位置节点之间的关联关系;W_P 表示位置节点之间的关联权重。

定义 4　类型子网 $G_C = (V_C, E_C, W_C)$。

类型子网 G_C 是基于类型节点和类型节点间的关联构成的网络。其中:V_C 表示类型节点;E_C 表示类型节点之间的有向连边,即两个类型节点之间的关联关系;W_C 表示类型节点之间的关联权重。

1.2　问题形式化

为了形式化地描述本文研究的科学问题,首先假设 $G = (V, E, W)$ 是本文研究的基于位置的社交网络,以及网络中的用户行为数据 $B = \{(b, v_i) \mid v_i \in V\}$。在上述定义的基础上,量化子网间的关联权重 W_M,则网络可以划分成"时空-用户-位置-类别"四层加权超网络:G_T、G_U、G_P、G_C、W_M,可以利用本文提出的方法预测用户子网中用户之间可能存在的链接关系 E^+。更明确的问题定义表示为:

$$\left.\begin{matrix} G \rightarrow G_T, G_U, G_P, G_C, W_M \\ B \end{matrix}\right\} \Rightarrow f:(G_T, G_U, G_P, G_C, W_M) \rightarrow E^+$$

1.2.1　问题输入

基于上述定义,本文研究内容的输入为:

1)基于位置的社交网络 $G = (V, E, W)$;

2)络中的用户行为 $B = \{(b, v_i) \mid v_i \in V\}$,表示用户 v_i 的行为 b,这里的用户行为包括用户间的跟随行为、用户对位置的签到和评分等。

1.2.2　问题输出

在给定基于位置的社交网络 $G = (V, E, W)$ 以及网络中的用户行为 $B = \{(b, v_i) \mid v_i \in V\}$ 的前提下,解决如下问题:

1) 如何构建网络模型,解决基于位置的社交网络中的链接预测面临的网络异构性、权值定义不完善、未考虑时间因素等问题。通过提取时空节点,将网络划分成"时空-用户-位置-类别"四层超网络,融合时间因素的影响,同时挖掘用户的隐

式行为，计算用户间影响力，利用用户的兴趣偏好以及位置间的潜在关联关系，量化超网络中的相关边权值，构建四层加权超网络模型 $S_G = (G_T, G_U, G_P, G_C, W_M, B)$。

2）如何利用加权超网络模型进行链接预测，并解决网络稀疏性问题。通过定义多种类型的超边结构可以量化用户间不同的关联关系，解决数据稀疏性问题，并引入参数集合 θ，量化每种超边结构的重要程度，求取最优参数集合 $\theta^+ = \arg\max_{\theta} P_\theta(E \mid G)$，利用最优参数集合 θ^+ 对用户间的链接关系 E^+ 进行预测，即 $E^+ = \arg\max_{\theta^+} P_{\theta^+}(E^U \mid G^U)$。

2 基于时空关系的超网络链接预测方法

本文方法主要包括三个阶段：1）构建加权超网络模型，构建时空-用户-位置-类别四层超网络，将时间、社交、位置等因素有效融合，然后基于用户影响力、隐关联关系、用户偏好以及节点度信息量化超网络中的边权值，构建四层加权超网络模型；2）建立加权超边结构，在加权超网络模型的基础上，建立超边结构；3）建立超链接预测，对每种超边结构的重要程度进行量化，从而对用户间的链接关系进行预测。

2.1 构建加权超网络模型

在上述定义的时空子网、用户子网、位置子网以及类别子网的基础上，子网与子网间也有一定的关联，如图 1 所示。通过四种不同的加权方式对子网内以及子网间的边权值进行定义和量化。

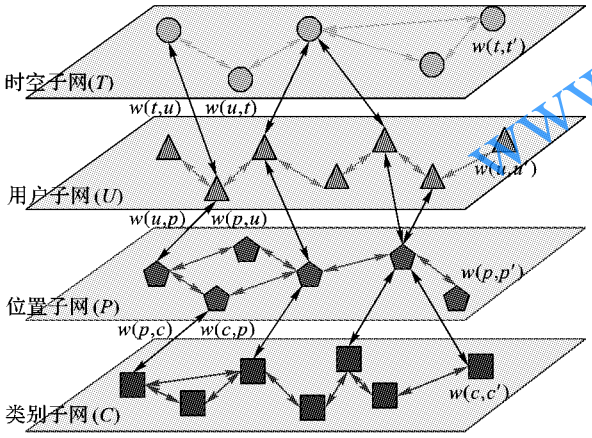


图 1 加权超网络模型

Fig. 1 Weighted supernetwork model

2.1.1 基于用户影响力

在基于位置的社交网络中，每个用户的影响力是不同的。如果某个好友对我们的影响力极低，那么我们很难通过该好友与其他人产生某些行为与联系。所以，通过用户的影响力量化用户-用户边权值 $w(u,u')$ 是提高模型解释性和准确性的一种可行方法。本文主要通过挖掘基于位置的社交网络中的追随行为以及追随网络来度量用户的影响力，如果用户 u' 在其好友 u 签到过的地方进行了一次签到，则认为用户 u' 产生了对用户 u 的追随行为。由追随行为构成的有向网络，称之为追随网络 $G_f = (V_f, E_f)$，其中：V_f 表示追随网络中的用户，E_f 表示追随行为的有向边。与此同时，我们将用户影响力分为用户个体影响力以及用户间影响力，并分别通过追随网络以及追随行为来度量。

1）用户个体影响力。

用户个体影响力 I_u 用于度量用户因自身行为对网络中其他用户产生的影响，是一种全局角度的度量方法。由于个体影响力会随着时间而动态变化，有的用户可能开始比较活跃，其签到行为产生了许多追随边，产生了较大的影响力，之后活跃度下降，则其影响力会逐渐减弱，长久如此，影响力会衰减至一个较低的稳定值。因此，在考虑用户影响力的时候，时间因素的影响极为重要。

本文首先划分 s 个不同的时间切片，将每个时间切片中用户的追随行为构成相应的追随网络 $(G_f^{t0}, G_f^{t1}, \cdots, G_f^{ts})$，用户的最终个体影响力由每一个时域中的个体影响力贡献而来，并且离当前时刻越久远的时域其个体影响力衰减越多。考虑到网络中孤立节点的存在，本文采用 LeaderRank 算法[16] 求解用户个体影响力。该算法通过引入 Ground Node，很好地解决了 PageRank 中因孤立节点而造成的排序结果不唯一的问题，算法的迭代公式如下：

$$I_u(t+1) = \sum_{u' \in N(u)} \frac{a_{u'u}}{k_{u'}^{\text{out}}} I_u(t) \tag{1}$$

其中：N_u 表示用户 u 的邻居节点；$k_{u'}^{\text{out}}$ 表示 u' 的出度。

在稳定状态下，LeaderRank 将 Ground Node 的分数均匀分布到所有其他节点，因此节点的最终得分可以表示为：

$$I_u = I_u(t_d) + I_g(t_d)/N \tag{2}$$

其中：$I_g(t_d)$ 是 Ground Node 在稳定状态下的分数；N 表示用户节点个数。

随着时间的推移，用户的影响力会随之递减，所以本文定义衰减函数为：

$$W_{ut_i} = \exp(-\ln 2 \times (t_c - t_i)/t_m) \tag{3}$$

其中：t_c 表示当前时刻；t_i 表示第 i 个时间片；t_m 表示影响力减小的半衰期。则用户 u 在当前时刻个体影响力总值 I_u 为：

$$I_u = \sum_{i=0}^{k} I_{ut_i} \cdot W_{ut_i} \tag{4}$$

其中 I_{ut_i} 表示第 t_i 个时间片用户 u 的个体影响力。

2）用户间影响力。

用户间影响力 $I_i(u,u')$ 用于度量用户 u 对用户 u' 的影响力大小，是一种局部视角的度量方法。通常情况下，两个用户之间交互次数越多，则他们之间的影响力会越大。在本文中将追随行为视为用户间的交互并以此来度量用户间影响力。提出追随地点比例 I_p 和追随签到比例 I_c 这两种衡量指标：

$$I_p(u,u') = M_{u',u}/Position_u \tag{5}$$

$$I_c(u,u') = K_{u',u}/Checkin_u \tag{6}$$

其中：$M_{u',u}$ 表示用户 u' 追随用户 u 的签到地点数；$Position_u$ 表示用户 u 的签到位置总数；$K_{u',u}$ 表示用户 u' 追随用户 u 的总签到次数；$Checkin_u$ 表示用户 u 的签到总次数。

综合考虑用户间影响力 $I_i(u,u')$ 和用户的个体影响力 I_u，则用户影响力 $I(u,u')$ 为：

$$I(u,u') = \partial_1 \cdot I_u + \partial_2 \cdot I_p(u,u') + \partial_3 \cdot I_c(u,u') \tag{7}$$

其中：I_u 为用户个体影响力；I_p 为追随地点比例；I_c 为追随签到比例；$\partial_1 + \partial_2 + \partial_3 = 1$，本文中 $\partial_1 = 0.4$，$\partial_2 = \partial_3 = 0.3$。

基于用户影响力，可以定义用户-用户边权值 $w(u,u')$，对于节点对 $u-u'$，如果 u' 对 u 的用户间影响力高，则对应权值 $w(u,u')$ 也应当高，所以用户间的边权值为：

$$w(u,u') = I(u',u) \Big/ \sum_{v \in U_u} I(v,u) \tag{8}$$

2.1.2　基于隐关联关系

隐关联关系指的是无法直接通过用户的签到信息观察到的关系,例如两个位置之间的关系以及两个类别之间的关系。曹玖新等[17]将两个位置之间的这种关系定义为兴趣点相关,本文中的位置-位置边权值是在此定义的上进行改进的。

1)位置-位置边权值 $w(p,p')$。

如果某个用户在一定的时间阈值内连续访问了某两个位置,那么这两个位置就存在一定的隐含关联关系,同时由于大多数用户喜欢访问与曾经去过的位置相近的位置[18],所以本文通过距离以及隐含关联关系定义位置与位置之间的边权值:

$$w(p,p') = \left(1 - \frac{geodist_{p,p'}}{\sum_{\forall z \in P} geodist(p,z)}\right) \cdot \frac{Count(p,p') + 1}{Count_p + 1} \tag{9}$$

其中:$geodist(p,p')$ 表示位置 p 和 p' 间的距离;$Count_p$ 表示所有被关联位置次数的最大值;$Count(p,p')$ 表示两个位置被关联的次数。

2) 时空-时空边权值 $w(t,t')$。

时空节点之间的边权值是在位置边权值的基础上增加时间距离,其定义如下:

$$w(t,t') = \left(1 - \frac{geodist_{t,t'}}{\sum_{\forall z \in T} geodist(t,z)}\right) \cdot \frac{Count(t,t') + 1}{Count_t + 1} + \left(1 - \frac{timedist(t,t')}{\sum_{\forall z \in T} timedist(t,t')}\right) \tag{10}$$

3) 类别-类别边权值 $w(c,c')$。

如果两种类别属性同时出现在多个位置,那么这两种类别之间也就存在一定的隐含关联关系,例如从数据统计中可以发现类别 Bars 与类别 Nightlife 经常出现在多个位置的类别属性中,隐含地表明这两种类别之间存在一定的相关性。本文通过这种相关性定义类别-类别边权值:

$$w(c,c') = (Count(c,c') + 1)/(Count_c + 1) \tag{11}$$

其中:$Count(c,c')$ 表示同时属于类别 c 和类别 c' 的地点个数;$Count_c$ 表示同时属于类型 c 和其他某种类型的地点数的最大值。

2.1.3　基于用户偏好

用户偏好反映了用户对位置的喜好程度,如果两个用户对同一个兴趣点表现出浓厚的兴趣,那么这两个用户有很大的可能会产生链接关系,所以为了有效融入偏好信息,提高超网络模型的可解释性,本文从用户评分入手定义和量化用户-位置边权值 $w(u,p)$ 以及用户-时空边权值 $w(u,t)$。

在基于位置的社交网络中,用户对位置的评分属性能够直观地反映出用户对这个位置的偏好程度。比如,用户 u_1 在 p_1、p_2、p_3 三个位置进行过评分,并分别给出了 5、3、1 的评分值,如果不考虑用户对这个位置的评分属性,那么会给每条用户-位置边赋值 1/3,但实际上这是不准确的,如果用户 u_1 对 p_3 的评分为 1 分,表明用户对这个地方是不满意的,这个时候,应当增加 $u_1 - p_1$ 的边权值,减小 $u_1 - p_3$ 的边权值。因此,应当给用户偏好高的位置更高的权值,本文参考文献[19]的方法通过指数函数定义用户-位置边权值,计算式如下:

$$w(u,p) = e^{r(u,p)} \Big/ \sum_{\forall p' \in P_u} e^{r(u,p)} \tag{12}$$

其中 $r(u,p)$ 表示用户 u 在位置 p 处的评分。

基于同样的道理,用户-时空边权值也可以相应地通过指数函数进行定义,计算式如下:

$$w(u,t) = e^{r(u,t)} \Big/ \sum_{\forall t \in T} e^{r(u,t)} \tag{13}$$

其中 $r(u,t)$ 表示用户 u 在时空节点 t 处的评分。

需要注意的是如果用户对一个位置进行过多次评分,本文仅将用户的最后一次的评分作为标准。

2.1.4　基于节点度信息

通过节点度信息定义和量化边权值类似于资源分配的原理,如果节点出度较多,则每个出度节点获得的资源就会相对较少。

1)时空-用户边权值 $w(t,u)$:

$$w(t,u) = 1/|U_t| \tag{14}$$

其中 $|U_t|$ 表示时空节点包含的用户个数。

2) 位置-用户边权值 $w(p,u)$:

$$w(p,u) = n(p,u) \Big/ \sum_{\forall u \in U_p} n(p,u) \tag{15}$$

其中:$n(p,u)$ 表示位置 p 被用户 u 访问的次数;$\sum_{\forall u \in U} n(p,u)$ 表示位置 p 被所有用户访问的总次数;U_p 表示访问了位置 p 的用户集合。

3) 位置-类别边权值 $w(p,c)$:

$$w(p,c) = 1/|C_p| \tag{16}$$

其中 $|C_p|$ 表示位置 p 所属的类别个数。

4) 类别-位置边权值 $w(c,p)$:

$$w(c,p) = 1/|P_c| \tag{17}$$

其中 $|P_c|$ 表示类别 c 包含的位置个数。

2.2　建立加权超边结构

2.2.1　加权超边结构相关定义

在加权超网络模型中,存在着多种类型的超边,比如用户节点与位置节点间构成的边可称为一条超边,用户节点与时空节点间构成的边也可称为一条超边。由于不同的超边,其包含的异构节点个数不同,因此,定义三种类型的超边。

定义 5　一类超边 SE_{I}。一类超边是指只包含一种类型节点的超边,在超网中属于一类特殊的超边。例如,两个用户节点构成的一条超边称之为一类超边。一类超边表明了同层子网中节点之间的关联关系,例如对于用户子网,指的是用户之间的好友关系。

定义 6　二类超边 SE_{II}。二类超边是指相邻两层子网之间的节点对构成的边,其特点是只包含两种异构节点。例如,用户与位置节点或者用户与时空节点之间构成的超边,称之为二类超边。

定义 7　三类超边 SE_{III}。三类超边是指相邻三层子网构成的边,其特点是包含三种异构节点。例如,用户、位置和类别节点构成的超边,称之为三类超边。

如图 2~3 所示,图 2 为相邻的两层网络,其中:$(t_1 - t_2)$ 构成了一条一类超边,记为 $SE_{\mathrm{I}}(t_1 - t_2)$;$(u_1 - t_1)$ 构成了一条二类超边,记为 $SE_{\mathrm{II}}(u_1 - t_1)$;$(u_3 - t_1)$ 也构成了一条二类超边,记为 $SE_{\mathrm{II}}(u_3 - t_1)$。图 3 为相邻的三层网络,其中:$(u_1 - p_1 - c_1)$ 组成了一条三类超边,记为 $SE_{\mathrm{III}}(u_1 - p_1 - c_1)$,$(c_1 - p_3 - u_3)$ 也组成了一条三类超边,记为 $SE_{\mathrm{III}}(c_1 - p_3 - u_3)$。

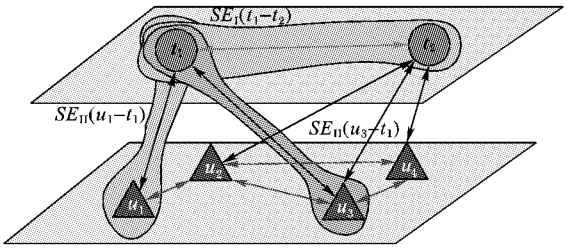


图 2 二层超网络

Fig. 2 Two-layer supernetwork

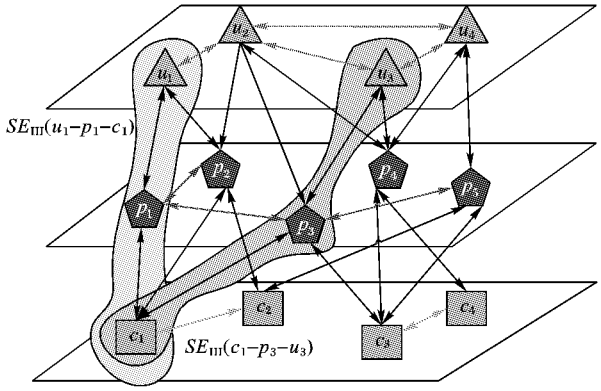


图 3 三层超网络

Fig. 3 Three-layer supernetwork

定义 8 超边权重。超边权重是指每条超边所具有的权值，可通过超边中包含的边权值计算得到。例如，图 2 中的二类超边 $SE_{II}(u_1-t_1)$，其超边权重 $W_{SE_{II}}(u_1-t_1)=w(u_1,t_1)$；图 3 中的三类超边权重 $W_{SE_{III}}(u_1-p_1-c_1)=w(u_1,p_1)\cdot w(p_1,c_1)$，$W_{SE_{III}}(c_1-p_3-u_3)=w(c_1,p_3)\cdot w(p_3,u_3)$。

基于这三种类型的超边，本文提出加权超边结构，通过加权超边结构解决用户与用户间的超链接预测问题。在以往的方法中，主要是通过加权超三角形结构来计算节点之间的关联程度，其主要思想是通过不同超边之间的共现节点将两条超边关联起来，从而得到超三角形结构用于度量节点间的相似性。这种方法适用于异构网络，能够简单高效地捕捉两个节点之间的额外关联，在缓解数据稀疏问题的同时也提高了预测准确度。然而超网络不仅可以描述同构节点之间的关联，同时能够描述异构节点之间的关联，因此考虑的网络层次越深，关联链越长，则越可以反映出节点间细粒度的隐性关联。本文通过构造多种类型的加权超边结构挖掘节点间的隐含语义关系。

在本文的方法中主要包含三类加权超边结构，分别为加权超三角形结构、加权超矩形结构以及加权超混合结构。下面通过图 2～3 对这三类结构进行举例说明。

2.2.2 加权超三角形结构

1）单加权超三角形结构。

如图 2 所示，可以通过时空节点 t_1 与用户节点 u_1、u_3 构成的单层加权超三角形结构来计算 u_1、u_3 间的相似性。如果包含 u_1、u_3 的单加权超三角形结构个数越多，权值越大，则认为它们之间的关联性也越大，越可能产生链接。该超三角形结构包含两条二类超边 $SE_{II}(u_1-t_1)$ 和 $SE_{II}(t_1-u_3)$，超三角形结构的权重为对应超边权重之积，计算式如下：

$$W_{S_3}(u_1u_3)=W_{SE_{II}}(u_1-t_1)\cdot W_{SE_{II}}(t_1-u_3) \tag{18}$$

需要强调的是本文定义的超边结构均为闭环结构，具有方向性，所以 $W_{S_3}(u_1u_3)\neq W_{S_3}(u_3u_1)$，后文同理。

2）双加权超三角形结构。

双加权超三角形结构包含两个连续的单加权超三角形结构，例如，图 2 中 $SE_{II}(u_1-t_1)$ 和 $SE_{II}(t_1-u_2)$ 组成了一个单加权超三角形结构 $W_{S_3}(u_1u_2)$，$SE_{II}(u_2-t_2)$ 和 $SE_{II}(t_2-u_3)$ 也构成了一个单加权超三角形结构 $W_{S_3}(u_2u_3)$，这两个单加权超三角形结构可以组合成一个双加权超三角形结构，用来度量 u_1、u_3 之间的相似度，该结构的语义信息为用户 u_1 和 u_3 都喜欢与用户 u_2 在相同时间相同的位置活动。该双加权超三角形结构权值为对应的两个单加权超三角形结构权值之积，所以权值为：

$$W_{S_6}(u_1u_3)=W_{S_3}(u_1u_2)\cdot W_{S_3}(u_2u_3) \tag{19}$$

3）三层加权超三角形结构。

三层加权超三角形结构是指由两条三类超边组成的三角形结构。如图 3 所示，超边 $SE_{III}(u_1-p_1-c_1)$ 和超边 $SE_{III}(u_1-p_3-c_3)$ 就组成了一个三层加权超三角形结构。其权值为两条三类超边权值之积，所以权值为：

$$W_{S_8}(u_1u_3)=W_{SE_{III}}(u_1-p_1-c_1)\cdot W_{SE_{III}}(c_1-p_3-u_3) \tag{20}$$

2.2.3 加权超矩形结构

如图 2 所示，超边 $SE_{II}(u_1-t_1)$，$SE_I(t_1-t_2)$，$SE_{II}(t_2-u_3)$ 可以构成一个加权超矩形结构，该加权超矩形结构中包含了 u_1、u_3 两个节点，可用于度量 u_1、u_3 之间的相似性，该结构的语义信息为用户 u_1 和 u_3 喜欢在两个相关的时空节点处活动，其权值为对应超边权值之积，所以权值为：

$$W_{S_{10}}(u_1u_3)=W_{SE_{II}}(u_1-t_1)\cdot W_{SE_I}(t_1-t_2)\cdot W_{SE_{II}}(t_2-u_3) \tag{21}$$

2.2.4 加权超混合结构

1）加权超混合 I 结构。

混合 I 结构是指在单三角形结构的基础上增加一条一类超边而组成的结构。如图 2 所示，由超边 $SE_{II}(u_1-t_1)$，$SE_{II}(t_1-u_2)$，$SE_I(u_2-u_3)$ 组成的结构属于混合 I 结构，该结构表达的语义信息为 u_3 的好友 u_2 喜欢与 u_1 在相同的时间相同的位置活动，其权值为对应的单加权超三角形结构权值与一类超边权值之积，计算式如下：

$$W_{S_{12}}(u_1u_3)=W_{S_3}(u_1u_2)\cdot W_{SE_I}(u_2-u_3) \tag{22}$$

2）加权超混合 II 结构。

混合 II 结构是指在矩形结构的基础上增加一条一类超边而组成的结构。如图 2 所示，由超边 $SE_{II}(u_1-t_1)$，$SE_I(t_1-t_2)$，$SE_{II}(t_2-u_2)$，$SE_I(u_2-u_3)$ 组成的结构属于混合 II 结构，其权值为对应的加权超矩形结构权值与一类超边权值之积，计算式如下：

$$W_{S_{14}}(u_1u_3)=W_{S_{10}}(u_1u_2)\cdot W_{SE_I}(u_2-u_3) \tag{23}$$

可以看出层次越深，关联链路越长，超边结构越丰富。本文列出了其中 19 种加权超边结构，如表 1 所示。

从前文分析可知，不同的加权超边结构具有不同的语义信息，例如 S_2 结构体现了位置熵[20]的含义，位置熵是指如果两个用户在一个很多人去过的地方有过共同签到，很难预测这两个人之间存在好友关系，因为这有可能是一种巧合，但是如果两个用户经常在一个很少人去过的地方进行签到，则表明它们之间很可能存在一定的关系。所以一个位置的受欢迎

程度也对链接预测有着影响,而通过 S_2 结构能够有效捕捉到这种影响。S_3 结构能够挖掘出用户的一种短期兴趣,此处的短期兴趣解释为用户可能只在某一个时间段才具有的兴趣,例如每周五晚上7点去电影院看电影。这种兴趣只会发生在特定的时间段,但更能体现出用户的个性。S_{11} 结构反映了两个用户虽然没有去过同一个位置,但却经常去相同类别的地方,这体现了两个用户拥有相同的类别偏好,有助于挖掘用户间的关联关系。

表 1　加权超边结构

Tab. 1　Weighted super-edge structures

结构类型	结构序号	细化类型	组成超边
加权超三角形结构	S_1	单加权超三角形结构 $S_1(u_iu_j)$	$SE_{\mathrm{I}}(u_i-u_k)-SE_{\mathrm{I}}(u_k-u_j)$
	S_2	单加权超三角形结构 $S_2(u_iu_j)$	$SE_{\mathrm{II}}(u_i-p_k)-SE_{\mathrm{II}}(p_k-u_j)$
	S_3	单加权超三角形结构 $S_3(u_iu_j)$	$SE_{\mathrm{II}}(u_i-t_k)-SE_{\mathrm{II}}(t_k-u_j)$
	S_4	反向双加权超三角形结构 $S_4(u_iu_j)$	$SE_{\mathrm{II}}(u_i-t_k)-SE_{\mathrm{II}}(t_k-u_d)-SE_{\mathrm{II}}(u_d-p_e)-SE_{\mathrm{II}}(p_e-u_j)$
	S_5	反向双加权超三角形结构 $S_5(u_iu_j)$	$SE_{\mathrm{II}}(u_i-p_k)-SE_{\mathrm{II}}(p_k-u_d)-SE_{\mathrm{II}}(u_d-t_e)-SE_{\mathrm{II}}(t_e-u_j)$
	S_6	同向双加权超三角形结构 $S_6(u_iu_j)$	$SE_{\mathrm{II}}(u_i-t_k)-SE_{\mathrm{II}}(t_k-u_d)-SE_{\mathrm{II}}(u_d-t_e)-SE_{\mathrm{II}}(t_e-u_j)$
	S_7	同向双加权超三角形结构 $S_7(u_iu_j)$	$SE_{\mathrm{II}}(u_i-p_k)-SE_{\mathrm{II}}(p_k-u_d)-SE_{\mathrm{II}}(u_d-p_e)-SE_{\mathrm{II}}(p_e-u_j)$
	S_8	三层加权超三角形结构 $S_8(u_iu_j)$	$SE_{\mathrm{III}}(u_i-p_k-c_d)-SE_{\mathrm{III}}(c_d-p_e-u_j)$
加权超矩形结构	S_9	二层加权超矩形结构 $S_9(u_iu_j)$	$SE_{\mathrm{II}}(u_i-p_k)-SE_{\mathrm{I}}(p_k-p_d)-SE_{\mathrm{II}}(p_d-u_j)$
	S_{10}	二层加权超矩形结构 $S_{10}(u_iu_j)$	$SE_{\mathrm{II}}(u_i-t_k)-SE_{\mathrm{I}}(t_k-t_d)-SE_{\mathrm{II}}(t_d-u_j)$
	S_{11}	三层加权超矩形结构 $S_{11}(u_iu_j)$	$SE_{\mathrm{III}}(u_i-p_k-c_d)-SE_{\mathrm{I}}(c_d-c_e)-SE_{\mathrm{III}}(c_e-p_f-u_j)$
加权超混合结构	S_{12}	加权超混合 I 结构 $S_{12}(u_iu_j)$	$SE_{\mathrm{II}}(u_i-p_k)-SE_{\mathrm{II}}(p_k-u_d)-SE_{\mathrm{I}}(u_d-u_j)$
	S_{13}	加权超混合 I 结构 $S_{13}(u_iu_j)$	$SE_{\mathrm{II}}(u_i-t_k)-SE_{\mathrm{II}}(t_k-u_d)-SE_{\mathrm{I}}(u_d-u_j)$
	S_{14}	加权超混合 I 结构 $S_{14}(u_iu_j)$	$SE_{\mathrm{I}}(u_i-u_k)-SE_{\mathrm{II}}(u_k-p_d)-SE_{\mathrm{II}}(p_d-u_j)$
	S_{15}	加权超混合 I 结构 $S_{15}(u_iu_j)$	$SE_{\mathrm{I}}(u_i-u_k)-SE_{\mathrm{II}}(u_k-t_d)-SE_{\mathrm{II}}(t_d-u_j)$
	S_{16}	加权超混合 II 结构 $S_{16}(u_iu_j)$	$SE_{\mathrm{II}}(u_i-p_k)-SE_{\mathrm{II}}(p_k-p_d)-SE_{\mathrm{II}}(p_d-u_e)-SE_{\mathrm{II}}(u_e-u_j)$
	S_{17}	加权超混合 II 结构 $S_{17}(u_iu_j)$	$SE_{\mathrm{II}}(u_i-t_k)-SE_{\mathrm{II}}(t_k-t_d)-SE_{\mathrm{II}}(t_d-u_e)-SE_{\mathrm{II}}(u_e-u_j)$
	S_{18}	加权超混合 II 结构 $S_{18}(u_iu_j)$	$SE_{\mathrm{II}}(u_i-u_k)-SE_{\mathrm{II}}(u_k-p_d)-SE_{\mathrm{II}}(p_d-p_e)-SE_{\mathrm{II}}(p_e-u_j)$
	S_{19}	加权超混合 II 结构 $S_{19}(u_iu_j)$	$SE_{\mathrm{II}}(u_i-u_k)-SE_{\mathrm{II}}(u_k-t_d)-SE_{\mathrm{II}}(t_d-t_e)-SE_{\mathrm{II}}(t_e-u_j)$

2.3　超链接预测

对于训练集中任意两个用户 (u,v),均存在上述的 19 种加权超边结构特征,这些特征的集合可以表示为 $W(u,v)=\{W_{S_1}(u,v),W_{S_2}(u,v),\cdots,W_{S_{19}}(u,v)\}$。本文通过逻辑回归计算 u 对 v 产生链接的概率:

$$P(y_{u\to v}\mid W)=\frac{\exp\left(\sum_{i=1}^{19}\theta_i^{+}W_{S_i}\right)}{1+\exp\left(\sum_{i=1}^{19}\theta_i^{+}W_{S_i}\right)} \tag{24}$$

其中 θ_i 表示第 i 个加权超边结构对链接建立的影响程度,利用梯度下降算法来更新参数值,参数更新过程如下:

$$\theta_{i\text{-new}}=\theta_{i\text{-old}}+\lambda\left(y_{u-v}-\sum_{i=1}^{19}\theta_iW_{S_i}\right)\times W_{S_i} \tag{25}$$

其中:λ 表示学习步长;y_{u-v} 表示用户间是否存在链接。当每个参数的变化值都小于某个阈值时,参数更新已收敛,得到最优参数集合 θ^{+}。

最终对于测试集中的用户对(u,u'),其产生链接的概率为:

$$P(y_{u-u'}\sim W)=[P(y_{u\to u'}\mid W)+P(y_{u'\to u}\mid W)]/2 \tag{26}$$

当 $P(y_{u-u'}\sim W)$ 大于等于阈值 ξ 时,$y_{u-u'}$ 取值为 1,认为用户间存在链接;否则 $y_{u-u'}$ 取值为 0,认为用户间不存在链接。从而找出了用户间可能存在的链接关系 E^{+}。$y_{u-u'}$ 定义式如下:

$$y_{u-u'}=\begin{cases}1, & P(y_{u-u'}\sim W)\geqslant\xi\\ 0, & \text{其他}\end{cases} \tag{27}$$

2.4　算法描述

本文输入为基于位置的社交网络 $G=(V,E,W)$ 以及网络中的用户行为 $B=\{(b,v_i)\mid v_i\in V\}$,输出为用户子网中的可能存在的链接关系 E^{+}。通过构建加权超网络模型,建立多种加权超边结构计算用户节点之间产生链接的概率,从而预测用户间可能存在的链接关系。其算法描述如算法 1 所示。

算法 1　基于时空关系的超网络链接预测方法。

输入　基于位置的社交网络 $G=(V,E,W)$;网络中的用户行为 $B=\{(b,v_i)\mid v_i\in V\}$。

输出　用户子网中可能存在的链接关系 E^{+}。

// 划分四层超网络

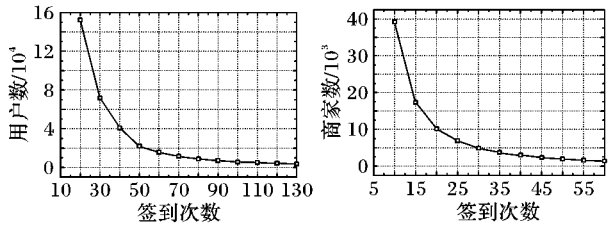
1) 基于用户的签到位置 p 和时间 *time* 构建时空节点 T;
2) 构建时空子网 G_T、用户子网 G_U、位置子网 G_P 和类别子网 G_C;
3) 基于用户影响力、隐关联关系、用户偏好以及节点度信息定义网络边权值,并依据式(1) ~ (17) 对网络边权值进行量化;
4) 构建三种类型的超边 SE_{I}、SE_{II}、SE_{III},并计算超边权重 W_{SE};
5) 基于三类超边,建立加权超三角形结构、加权超矩形结构、加权超混合结构,并依据式(18) ~ (23) 计算加权超边结构的权重 W_{S_i};
6) For $j\leftarrow 1$ to $|M|$ do　　// $|M|$ 为总节点对个数
7)　For $i\leftarrow 1$ to 19 do　　//19 种超边结构
8)　　计算每一个超边结构下的 W_{S_i};
9)　End for
10)　依据式(24) 计算链接产生概率,并加入候选集合 *Ms*;
11) End for
12) 根据式(26) 计算 *Ms* 集合中的 $P(y_{u-u'}\sim W)$;
13) 根据式(27) 判断用户间可能存在的链接关系 E^{+}。

3 仿真实验与分析

3.1 数据集

本文采用 yelp 数据集进行实验，数据集来源于 yelp 挑战赛的开放数据集，包含了用户表、商家表以及评论表。用户表中有每个用户的好友列表、评论次数等信息，商家表中有所属类别、经纬度以及一些详细的属性标签。评论表中有用户对商家的评论、评分以及评分时间。

实验中，删除总评论数少于 10 次的不活跃用户以及被评论总次数少于 5 次的商家以减小稀疏数据的影响，处理后的用户签到分布以及商家被签到分布情况如图 4 所示。图 4(a)中横坐标为 20 的坐标点表示用户签到次数在[10，20)的用户数。从图 4 可以看出用户签到以及商家被签到次数均符合幂律分布特性。



(a) 用户签到分布 (b) 商家签到分布

图 4 用户和商家签到分布

Fig. 4 Check-in distribution of users and businesses

接下来对预处理后各层子网以及相邻子网间的情况进行统计说明。

1)四层子网。针对时空子网，由于 yelp 数据集中评论时间只能精确到天，考虑到相隔时间太长会影响时空节点的本质作用，所以本文以同一天为标准提取时空节点，实验中总共提取的时空节点个数为 850 437 个，所以该层节点数为 850 437 个，理论上任意两个时空节点之间均存在连边(通过权值控制连边强弱)；针对用户子网，由于剔除了总评论次数少于 10 次的不活跃用户，所以该层保留的有效节点数为 426 656 个，相比于原始用户数据占比 0.414，好友关系数为 2 703 594 条，相比于原始用户关系占比 0.461；针对位置子网，由于剔除了被评论次数少于 5 次的商家，所以该层的剩余节点数为 95 190 个，相比于原始商家数占比为 0.661，理论上任意两个商家位置之间均存在连边；针对类别子网，由于部分商家被踢除，导致总体类别数下降，由原来的 1 192 类商家减少到 1 007 类，所以该层的节点个数为 1 007 个，相比于原始类别数据占比为 0.845，理论上任意两个类别节点之间均存在连边。

2)相邻子网。由于剔除了部分用户以及商家，使得用户-位置边减少(评论签到关系)，由原来的 4 153 150 条边较少为 2 943 312 条边，占比为 0.709；用户-时空节点之间的边数统计为 4 295 788 条，从数据统计结果可以看出，对于大多数时空节点来说，仅包含两个用户，随着时空节点包含用户数的增多，相应时空节点个数按幂律分布减少，如图 5 所示，位置-类别之间的边数统计为 351 952 条。

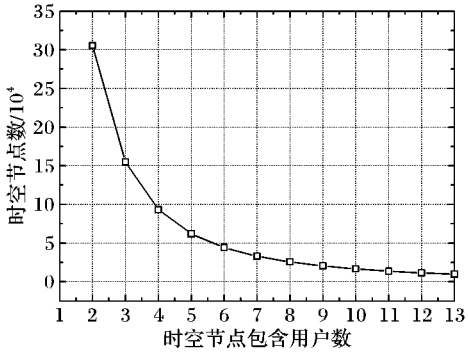


图 5 时空节点分布

Fig. 5 Distribution of spatio-temporal nodes

3.2 评价指标

本文中，主要使用准确率(precision)、召回率(recall)、F1 指标(F1-measure)、受试者工作特征曲线下面积(Area Under the receiver operating characteristic Curve，AUC)来评估算法的质量。

1)准确率。

令 V_L 为得分最高的 L 个节点对，如果在 V_L 中只有 m 个节点对存在连边，则准确度可以表示为：

$$precision = m/L \tag{28}$$

2)召回率。

令 $M = |E|$，表示网络中存在的边，假设算法预测出来的缺失边个数为 m，则召回率可通过如式(29)求得：

$$recall = m/M \tag{29}$$

3)F1 指标。

通过准确率和召回率，可以得到 F1 综合指标，其定义如下：

$$F1\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \tag{30}$$

4)AUC。

AUC 是评估链接预测结果质量的常用指标。AUC 指标可以描述为在测试集中随机选择一条存在连边的分数值比随机选择一条不存在连边的分数值高的概率。这样独立重复比较 n 次，在 n 次中，如果有 n' 次在测试集中存在连边的分数比不存在连边的分数值高，有 n'' 次在测试集中存在连边的分数值与不存在连边的分数值相等，则 AUC 值可以定义为：

$$AUC = (n' + 0.5n'')/n \tag{31}$$

一般来说，AUC 值越高代表预测性能越好，完美结果的 AUC 值是 1.0，而随机预测结果的 AUC 是 0.5。

3.3 实验结果与分析

在实验中，采用不同的比例(0.5～0.9)来划分数据集以准确评估算法的性能。首先对比不同超边结构及其组合下的准确率、召回率、F1 值以及 AUC。令：K_1 表示加权超三角形结构，K_2 表示加权超矩形结构，K_3 表示加权超混合结构。

实验结果如图 6 所示，其中横坐标表示训练集占整个数据集的比例。从图 6 可以看出，随着训练集边数占总边数的比例加大，所有指标均呈现上升状态，表明预测的效果在不断提升。从图 6(d)中 AUC 指标可以看出，训练集占比为 0.9 时，各特征的预测性能达到最优。其中：K_1 的预测效果是最好的，其次是 K_3，最后是 K_2。其原因是 K_1 结构能够捕捉到较多重要的信息，比如，其中的 S_1 结构能够捕捉两个用户的共同好友信息，S_2 结构能够捕捉两个用户的共同签到地点信息，S_3 结构能够捕捉两个用户的共同时空节点信息，而这些都是预测用户之间链接较为重要的信息。

在两两组合的预测效果中，$K_2 + K_3$ 的预测性能最差，预测效果低于 K_1 特征，$K_1 + K_3$ 的预测性能最好，可以看出组合

特征的效果要优于组合中单个特征的效果。$K_1+K_2+K_3$ 由于有效融合了每组结构特征的信息,所以其预测性能最优。

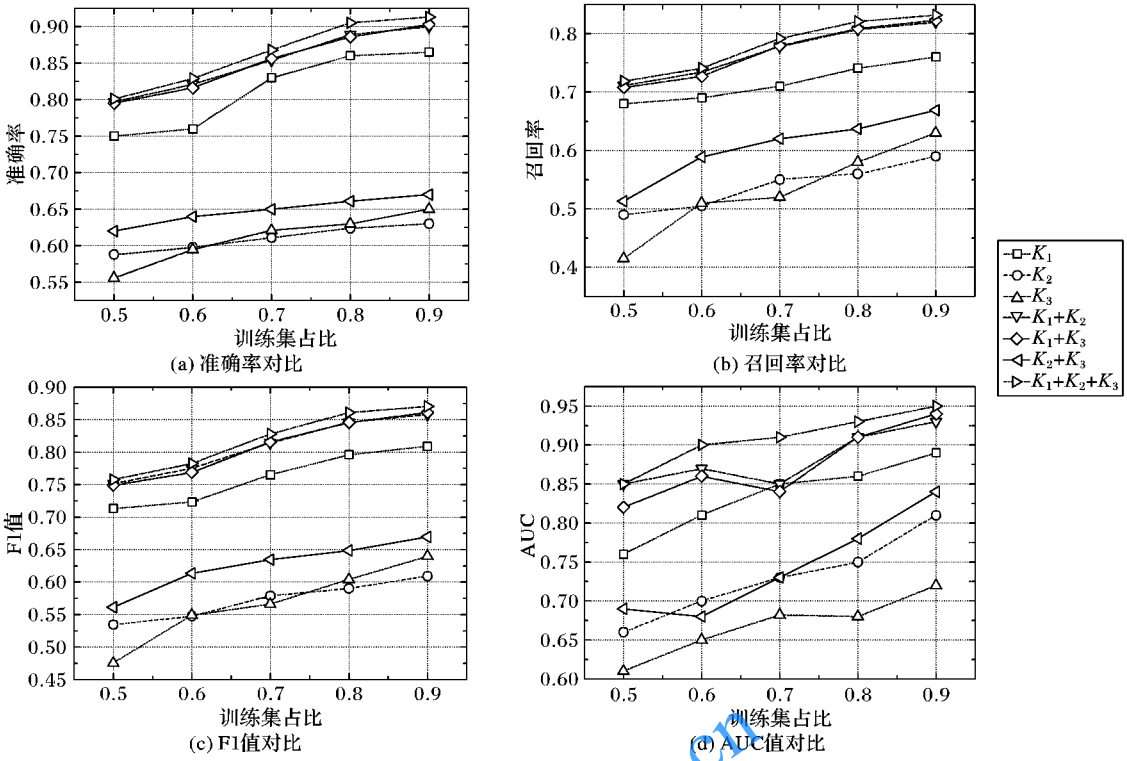


图 6　不同特征组合下各指标对比

Fig. 6　Comparison of different indicators under different combinations of features

由于训练集占比为 0.9 时,算法性能最优,所以后面的对比实验均采用 0.9 来划分数据集。为了验证时空节点的引入对实验结果的影响,本文对比包含时空节点层的加权超网络模型与不包含时空节点层的加权超网络模型(用户-位置-类别三层),并分别提取模型中的加权超边结构,计算节点间的相似性,实验结果如图 7 所示,其中横坐标表示不同的指标,纵坐标表示这些指标下的测试值。

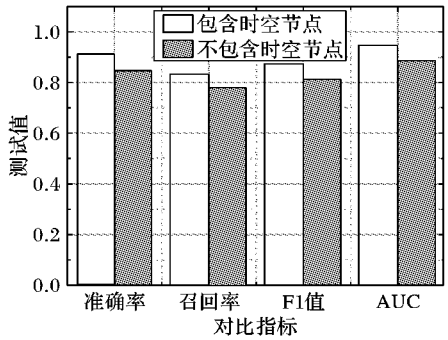


图 7　时空节点对比

Fig. 7　Comparison of spatio-temporal nodes

从图 7 中可以看出,引入时空节点层之后,算法的准确率、召回率、F1 值以及 AUC 都有一定程度的提升,其原因是时空节点不仅能够捕捉两个用户在访问位置上的兴趣相似性,同时还能捕捉其在访问时间上的相似性。因此,即使两个用户共同签到位置数较少,但只要他们在这些地方的签到时间比较接近,也能够通过时空节点层准确捕捉他们之间的相似性,从而对链接关系进行有效预测,而这些是无法通过位置节点做到的。时空节点层的引入使得我们可以兼顾位置因素的影响以及时空因素的影响,通过对这两类信息的有效利用,本文的算法性能得到了一定的提升。

最后对比不同方法之间的性能优劣,本文选取同构网络中的方法以及异构网络中的方法与本文的方法进行对比。同构网络中的对比方法选择经典的共同邻居指标(Common Neighbor, CN)[21]、资源分配指标(Resource Allocation, RA)[22],其中 CN 是由 Liben-Nowell 等[21]提出的解决链接预测问题的基本方法,是一种基于邻居结构的标准度量方法;RA 则是由 Zhou 等[22]在 CN 的基础上提出的改进方法,并验证了其在多个网络中的有效性,这两种方法属于同构网络中解决链接预测问题的代表性方法。异构网络中选择文献[8]中的方法Model-II、文献[12]中的超链接预测方法超网络杰卡德指标(Supernetwork JACCARD, SJACCARD),以及文献[13]中的 PathPredict 方法,其中 Model-II 方法基于两个用户所有的共现位置信息提出了五种相关特征:加权共现次数、加权共同位置数、平均时间间隔、最小时间间隔、最大时间间隔,并验证了算法的优越性;SJACCARD 方法为加权超网络下的一种方法,该方法通过加权超三角结构预测节点之间的链接关系,取得了一定的效果;PathPredict 方法为异构网络链接预测中的基准方法,该方法通过路径数、随机游走等方式度量异质节点之间的语义关联关系,适用于无权网络。本文所有结果均采用 10 折交叉验证得到,六种方法的性能对比结果如表 2 所示。

由表 2 可以看出,所有的异构方法均要优于同构方法,反映出异构方法对信息的利用率更高。CN 和 RA 方法仅仅利用到了同构网络中的社交关系,所以其各指标均相对较低,RA 方法由于通过节点的度信息量化了不同邻居节点的贡献,所以其性能要优于 CN。在异构方法中由于 SJACCARD

方法仅仅利用到了两种异构加权超三角形结构,未能利用到同构边关系,所以其 AUC 值仅比 CN 高出 0.049,PathPredict 方法虽然通过多种元路径特征较好地捕捉了异构关系,但其没能有效利用网络中的多种权值信息,所以其 AUC 值低于本文方法,但高出 CN 方法 0.154,Model-II 方法的性能优于 PathPredict,AUC 值达到了 0.915,较好地证明了时间因素的引入能够提高预测精度。本文方法通过引入时空节点以及多种异构结构,同时较好地结合了网络边权值,预测性能最优,AUC 值达到 0.958,相比于异构网络方法中的 Model-II 方法,AUC 值提升了 4.69%。

表 2 不同方法的性能对比

Tab. 2 Performance comparison of different methods

算法	精确率	召回率	F1 值	AUC
CN	0.634	0.511	0.565	0.713
RA	0.685	0.538	0.602	0.759
PathPredict	0.857	0.708	0.787	0.867
SJACCARD	0.714	0.664	0.688	0.762
Model-II	0.889	0.863	0.876	0.915
本文方法	0.913	0.832	0.871	0.958

4 结语

链接预测研究可以用于解决网络中的链路缺失问题,在推荐领域具有重要的应用前景。针对 LBSN 链接预测中网络权值利用不完善以及数据稀疏性等问题,本文提出了一种可行方法。该方法通过用户影响力、隐关联关系、用户偏好以及节点度信息对超网络中的边权值进行定义和量化,并通过多种加权超边结构来捕捉用户间的多种关联关系,缓解数据稀疏性问题的同时提高了链接预测性能。实验结果表明,本文方法相比现有其他方法有着更加准确的预测精度,更加适用于复杂的异构网络推荐领域,可提高用户对平台的忠诚度以及黏度,从而令平台具备更加长久的发展空间。

本文仅针对一个社交网络数据进行研究,而大多数用户会活跃在多个社交平台,例如 QQ 和微博,如何获取相应数据并合理构建多维度的超网络模型来预测网络中的连边关系是我们下一步要解决的关键问题。

参考文献 (References)

[1] 龚卫华,陈彦强,裴小兵,等. LBSN 中融合多维关系的社区发现方法[J]. 软件学报, 2018, 29(4): 1163 - 1176. (GONG W H, CHEN Y Q, PEI X B, et al. Community detection of multi-dimensional relationships in location-based social networks [J]. Journal of Software, 2018, 29(4): 1163 - 1176.)

[2] LÜ L Y, ZHOU T. Link prediction in complex networks: a survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150 - 1170.

[3] 刘思,刘海,陈启买,等. 基于网络表示学习与随机游走的链路预测算法[J]. 计算机应用, 2017, 37(8): 2234 - 2239. (LIU S, LIU H, CHEN Q M, et al. Link prediction algorithm based on network representation learning and random walk [J]. Journal of Computer Applications, 2017, 37(8): 2234 - 2239.)

[4] DUAN L, MA S, AGGARWAL C, et al. An ensemble approach to link prediction [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(11): 2402 - 2416.

[5] SCELLATO S, NOULAS A, MASCOLO C. Exploiting place features in link prediction on location-based social networks [C]// Proceedings of the 2011 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1046 - 1054.

[6] VALVERDE-REBAZA J, ROCHE M, PONCELET P, et al. Exploiting social and mobility patterns for friendship prediction in location-based social networks [C]// Proceedings of the 2016 23rd International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2526 - 2531.

[7] BAYRAK A E, POLAT F. Examining place categories for link prediction in location based social networks [C]// Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway, NJ: IEEE, 2016: 976 - 979.

[8] CHENG R, PANG J, ZHANG Y. Inferring friendship from check-in data of location-based social networks [C]// Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway, NJ: IEEE, 2015: 1284 - 1291.

[9] CRANDALL D J, BACKSTROM L, COSLEY D, et al. Inferring social ties from geographic coincidences [J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(52): 22436 - 22441.

[10] LI R H, LIU J Q, YU J X, et al. Co-occurrence prediction in a large location-based social network [J]. Frontiers of Computer Science, 2013, 7(2): 185 - 194.

[11] 刘怡君,唐先一,李倩倩,等. 超链路预测[J]. 管理评论, 2012, 24(12): 137 - 145. (LIU Y J, TANG X Y, LI Q Q, et al. Super-link prediction [J]. Management Review, 2012, 24(12): 137 - 145.)

[12] 方哲,游宏梁,薛非,等. 专家知识协作加权超网络模型及其超链路预测研究[J]. 科研管理, 2017(s1): 251 - 258. (FANG Z, YOU H L, XUE F, et al. A research on the weighted expert knowledge collaboration super-network model and super-link prediction method [J]. Science Research Management, 2017(s1): 251 - 258.)

[13] SUN Y Z, BARBER R, GUPTA M, et al. Co-author relationship prediction in heterogeneous bibliographic networks [C]// Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. Washington, DC: IEEE Computer Society, 2011: 121 - 128.

[14] SHI C, ZHANG Z Q, LUO P, et al. Semantic path based personalized recommendation on weighted heterogeneous information networks [C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2015: 453 - 462.

[15] CUI Y W, ZHANG L L, WANG Q D, et al. Heterogeneous network linkage-weight based link prediction in bipartite graph for personalized recommendation [J]. Procedia Computer Science, 2016, 91: 953 - 958.

[16] LÜ L Y, ZHANG Y C, YEUNG C H, et al. Leaders in social networks, the delicious case [J]. PloS One, 2011, 6(6): e21202.

[17] 曹玖新,董羿,杨鹏伟,等. LBSN 中基于元路径的兴趣点推荐[J]. 计算机学报, 2016, 39(4): 675 - 684. (CAO J X, DONG Y, YANG P W, et al. POI recommendation based on meta-path in LBSN [J]. Chinese Journal of Computers, 2016, 39(4): 675 - 684.)

[18] YUAN Q, CONG G, MA Z Y, et al. Time-aware point-of-interest recommendation [C]// Proceedings of the 2013 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2013: 363 - 372.

(下转第 1697 页)

点的选择比一般分簇的簇头的选择速度要快;而且对基于密度的均衡分簇,当网格内节点数少于某阈值就不需要选择投影节点了,开销更为减少。因此,本文建簇的方法更为节能。

4　结语

针对投影节点随机选择、位置不均衡的问题,本文提出了基于均衡投影的压缩数据收集方法,并针对均匀分布节点 WSN,提出基于空间位置的均衡分簇法,以大小相同的网格划分来实现,保证了投影节点的位置均衡;针对不均匀节点分布 WSN,提出基于节点密度的均衡分簇法,同时考虑了位置和密度因素,减少了孤立点的能耗,均衡了能量,延长了网络寿命。通过与随机投影节点方法的仿真比较,本文方法运行状态优良,网络生存期显著延长。但对于非均匀 WSN,按密度分簇的阈值需要根据实际应用中网络规模、节点整体分布密度和应用需求等综合因素来选择,并需要先进行实验和验证。

参考文献(References)

[1] BAJWA W U, HAUPT J, SAYEED A M, et al. Compressive wireless sensing [C]// Proceedings of the 2006 Fifth International Conference on Information Processing in Sensor Networks. Piscataway, NJ: IEEE, 2006: 134 – 142.

[2] HAUPT J, BAJWA W U, RABBAT M, et al. Compressed sensing for networked data [J]. IEEE Signal Processing Magazine, 2008, 25(2): 92 – 101.

[3] LUO C, WU F, SUN J, et al. Compressive data gathering for large-scale wireless sensor networks [C]// Proceedings of the 15th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2009: 145 – 156.

[4] HAUPT J, NOWAK R. Signal reconstruction from noisy random projections [J]. IEEE Transactions on Information Theory, 2006, 52(9): 4036 – 4048.

[5] WIMALAJEEWA T, VARSHNEY P K. Wireless compressive sensing over fading channels with distributed sparse random projections [J]. IEEE Transactions on Signal and Information Processing over Networks, 2015, 1(1): 33 – 44.

[6] RAN R, OH H. Adaptive sparse random projections for wireless sensor networks with energy harvesting constraints [J]. EURASIP Journal on Wireless Communications and Networking, 2015, 2015: 113 – 122.

[7] ABBASI-DARESARI S, ABOUEI J. Toward cluster-based weighted compressive data aggregation in wireless sensor networks [J]. Ad Hoc Networks, 2016, 36(P1): 368 – 385.

[8] 乔建华, 张雪英. 基于压缩感知的无线传感器网络数据收集研究综述[J]. 计算机应用, 2017, 37(11): 3261 – 3269. (QIAO J H, ZHANG X Y. Compressed sensing based data gathering in wireless sensor networks: a survey [J]. Journal of Computer Applications, 2017, 37(11): 3261 – 3269.)

[9] EBRAHIMI D, ASSI C. Compressive data gathering using random projection for energy efficient wireless sensor networks [J]. Ad Hoc Networks, 2014, 16: 105 – 119.

[10] XIAO F, GE G W, SUN L J, et al. An energy-efficient data gathering method based on compressive sensing for pervasive sensor networks [J]. Pervasive and Mobile Computing, 2017, 41: 343 – 353.

[11] DONOHO D L. Compressed sensing [J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289 – 1306.

[12] BARANIUK R G. Compressive sensing [J]. IEEE Signal Processing Magazine, 2007, 24(4): 118 – 121.

[13] CANDÈS E J, WAKIN M B. People hearing without listening: an introduction to compressive sampling [J]. IEEE Signal Processing Magazine, 2008, 25(2): 21 – 30.

[14] CANDÈS E J. The restricted isometry property and its implications for compressed sensing [J]. Comptes Rendus Mathematique, 2008, 346(9/10): 589 – 592.

[15] HEINZELMAN W R, CHANDRAKASAN A, BALAKRISHNAN H. Energy-efficient communication protocol for wireless microsensor networks [C]// Proceeding of the 33rd Annual Hawaii International Conference on System Sciences. Washington, DC: IEEE Computer Society, 2000: 8020.

[16] 尚凤军. 无线传感器网络通信协议[M]. 北京: 电子工业出版社, 2011: 95. (SHANG F J. Communication Protocol for Wireless Sensor Networks [M]. Beijing: Publishing House of Electronics Industry, 2011: 95.)

This work is partially supported by the Natural Science Foundation Project of Shanxi Province (2013011019-1).

QIAO Jianhua, born in 1975, Ph. D. candidate, associate professor. Her research interests include wireless sensor network, compressed sensing.

ZHANG Xueying, born in 1964, Ph. D, professor. Her research interests include speech signal processing, multimedia communication, Internet of things.

(上接第 1690 页)

[19] VAHEDIAN F, BURKE R, MOBASHER B. Weighted random walk sampling for multi-relational recommendation [C]// Proceedings of the 2017 25th Conference on User Modeling, Adaptation and Personalization. New York: ACM, 2017: 230 – 237.

[20] CRANSHAW J, TOCH E, HONG J, et al. Bridging the gap between physical location and online social networks [C]// Proceedings of the 2010 12th ACM International Conference on Ubiquitous Computing. New York: ACM, 2010: 119 – 128.

[21] 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.

[22] ZHOU T, LÜ L Y, ZHANG Y C. Predicting missing links via local information [J]. European Physical Journal B, 2009, 71(4): 623 – 630.

This work is partially supported by the Foundation and Frontier Research Project of Chongqing Science and Technology Commission (cstc2014jcyjA40039).

HU Min, born in 1971, M. S., associate professor. Her research interests include communication network system and protocol, big data.

CHEN Yuanhui, born in 1991, M. S. candidate. His research interests include data mining, link prediction in complex network.

HUANG Hongcheng, born in 1979, Ph. D., associate professor. His research interests include complex network analysis, intelligent information processing.