计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200231

基于自我中心网络结构特征和网络表示学习的链路预测算法

赵 曼 赵加坤 刘金诺
西安交通大学电信学部软件学院 西安 710000

摘 要 链路预测是网络分析与挖掘领域中备受关注的研究方向。链路预测算法所预测的网络中的缺失连接实际上是一种数据挖掘的过程,而推断的将来可能产生的连接则与网络的发展演化相关。因此,如何提高链路预测的精确度是一项有意义且具有挑战性的研究。基于自我中心网络分解和社区聚类的最新研究,提出一种基于自我中心网络结构特征和网络表示学习的链路预测算法(Ego-Embedding)。Ego-Embedding将原网络转换成角色图,再结合网络的微观结构信息和上下文信息重构嵌入过程,为每一个节点学习一个或多个向量表示,使向量表示更准确地描述网络节点信息,从而提高链路预测的精确度。在3个公开数据集(Facebook,PPI-Yeast和ca-HepTh)上进行实验仿真,并使用AUC作为评价指标,仿真结果表明,算法Ego-Embedding的表现均优于5个实验对比方法(CN,AA,Node2vec,M-NMF和Splitter),且最高将链路预测的错误率减少了约47%。

关键词:链路预测;自我中心网络;网络表示学习;角色分解;Ego-Embedding
中图法分类号 TP393

Link Prediction Algorithm Based on Ego Networks Structure and Network Representation Learning

ZHAO Man,ZHAO Jia-kun and LIU Jin-nuo
School of Software Engineering,Xi'an Jiaotong University,Xi'an 710000,China

Abstract Link prediction is a research direction that has attracted much attention in the field of network analysis and mining. The link prediction algorithm predicts the missing connections in the network,which is actually a process of data mining,and the inferred possible future connections are related to the development and evolution of the network. Therefore,how to improve the accuracy of link prediction is a meaningful and challenging research. Based on the latest research on ego network decomposition and community clustering,a link prediction algorithm (Ego-Embedding) is proposed,which is based on ego network structure characteristics and network representation learning. Ego-Embedding converts the original network into a persona graph,and then reconstructs the embedding process by combining the microstructure information and context information of the network,learning one or more vector representations for each node,so that the vector representation can describe the node information more accurately,thereby improving the accuracy of link prediction. This paper conducts experimental simulations on three public data sets (Facebook,PPI-Yeast,ca-HepTh),and uses AUC as the evaluation index. The experimental results show that the performance of the algorithm Ego-Embedding is better than the five experimental comparison methods (CN,AA,node2vec,M-NMF,Splitter),and the best link prediction AUC reduces the error by up to 47%.

Keywords Link prediction,Ego network,Network representation learning,Persona decomposition,Ego-Embedding

1 引言

节点及节点之间的边(称为链路)组成的结构称为网络。利用网络结构可以描述现实世界中的复杂系统,例如生活中常见的计算机系统、生物系统和社交系统等。通常,用网络节点来表示现实系统中的实体,将节点之间的关系(即实体对象之间的某种关联)用边进行表示[1]。网络不仅可以用来描述数据,还可以用于对数据进行分析与挖掘。

近年来,在线数据的剧增激发了研究者对网络的挖掘和分析,包括节点分类[2]、用户分析[3]、社区检测[4]和链路预测[5-7]等研究。其中,链路预测因为在理论和现实方面应用广泛而受到越来越多的关注,例如社交网络中的推荐算法(给用户推荐共同好友[8]),对复杂网络构建演化模型[9],网络重组[10]和寻找生物网络中蛋白质之间的相互作用[11]等。因此,提高链路预测的准确度是一项有意义且具有挑战性的研究。链路预测的任务是根据已知网络的节点属性信息和网络拓扑结构来推断网络中丢失的链路或者预测该网络将来可能产生的链路[20]。常见的链路预测的算法主要包括基于节点属性信息、基于网络拓扑结构、机器学习和最大似然这4种。由于我们无法确定节点属性信息的真实性和准确性,且深入挖掘信息又会涉及到用户的隐私问题,所以利用节点属性信息的链路预测准确率难以保证。而基于网络结构信息的方法不需要额外数据,且计算复杂度相对较低,除了小规模网络外,还适用于大型网络,因此受到了广泛关注。

但是,大多数相关算法的研究关注的都是整个网络的结构和功能,而自我中心网络侧重于研究单个节点的性质,因此可以利用自我中心网络去提取整个网络中各个节点的重要结构信息。自我中心网络是指以一个节点为中心,包含其邻居节点以及中心节点与邻居节点之间的边,即包含了该中心节点可以维持的一切社会关系[12]。近期有研究表明,构建网络

通信作者:赵曼(1312435801@qq.com)

中以各个节点为中心节点的自我中心网络，然后再进行聚类分析并应用在朋友推荐算法中的表现更佳[13]，对聚类后的社区节点进行角色分解生成非重叠集群可以获得原网络的重叠聚类[14]，利用角色节点进一步地为其对应的原节点学习多个嵌入表示，使其可以更好地应用在链路预测任务和可视化研究中[15]。

关于社交网络的链路预测算法的研究工作大多是基于整个网络，缺少从单个节点的视角出发来研究该节点在其社交圈中的表现与未来产生连接可能性的关系。针对这个问题，本文提出一种基于自我中心网络结构特征和网络表示学习的链路预测算法(Ego-Embedding)，从自我中心网络级提取网络的结构信息，并且网络表示学习可以将结构特征和节点的语义信息结合起来。首先，构建网络中各个节点的自我中心网络，并通过角色分解生成对应的角色图；其次，计算自我中心网络中各社区之间的桥节点，并改变桥节点与其他节点之间连接的权重；再通过实验找到 Node2vec 模型中的最佳参数 p 和 q；然后，计算节点的中介中心性并将其作为惩罚因子修正节点序列；最后，将上述工作融合在 Skip-Gram 模型中进行训练，学习每一个节点的一个或多个向量表示，并在多个网络数据集上进行链路预测实验，实验结果证实了提出的 Ego-Embedding 算法对预测的准确率有提升并且显著降低了预测的错误率。

2 相关工作

2.1 自我中心网络

随着信息科技的发展，虽然人们沟通和维持社交关系的方式发生了改变，但是人们组织社交关系的方式仍然没有改变。由于人类的注意力有限且社交互动关系之间存在内部优先级，一个人可以同时维持的关系数量在生物学上是有限的，且在数字世界里也没有被超越[16]。1982 年，研究者 Freeman 提出了自我中心网络这个概念[17]，它的网络节点由唯一的一个中心节点(即自我中心)以及中心节点的邻居节点(即联系人)组成，网络的边由自我中心与联系人之间的边以及联系人与联系人之间的边组成。即相对于整个网络，自我中心网络在网络分析挖掘领域主要侧重于研究单个中心节点的性质，它能够帮助我们分析用户节点的特征，比如检测用户分类的社交圈[18]，还能应用于对网络结构和功能的研究[19]等。因此，从自我中心网络结构特征的角度去分析和挖掘网络中节点的整体特征，可以将网络信息描述得更加准确。

2.2 角色分解和网络表示

现实世界中的社区彼此之间都有很强的重叠性，而且外部联系往往比其内部联系更复杂，使得研究人员很难在全局级别上定义社区。但在微观层次上，社区结构比较清晰，例如自我中心网络。Epasto 利用 Schank 算法的简单变体来构建自我中心网络[13]，进而对整个网络中以各个节点为中心节点的自我中心网络进行社区检测，检测结果表明，社区集群与用户自定义的社交圈紧密匹配，并且发现每个节点在其所处的不同社区内扮演的角色不同。例如节点 U 的自我中心网络包含 A，B，C，D，E 共 5 个节点，如图 1 所示，假设 A 和 B 是 U 的亲人，C，D 和 E 是 U 的同学，那么 U 是属于两个社区，我们根据社区信息可以将 U 分割成两个角色节点，即作为家庭角色的 U_1 和作为学生角色的 U_2，如图 2 所示，这样就可以分离开两个重叠的社区。通过该角色分解就可以消除社交网络或者其他现实世界网络中的重叠社区，再对生成的非重叠的角色图进行聚类分区[14]。如此，就可以将复杂的重叠集群问题转化成一个更简单、更易于处理的非重叠分区问题。

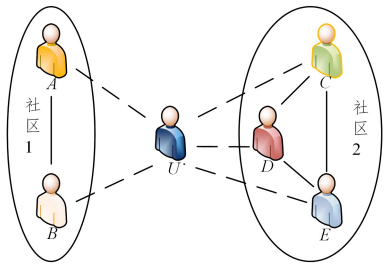


图 1 节点 U 的自我中心网络

Fig. 1 Ego network of node U

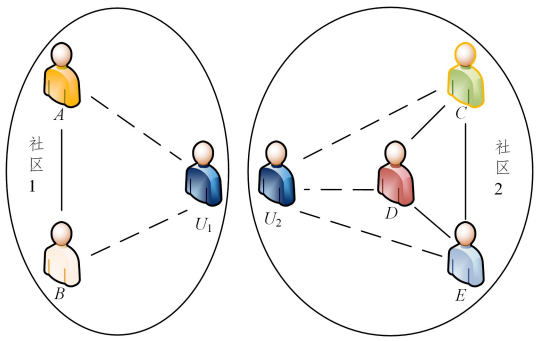


图 2 对节点 U 进行角色分解

Fig. 2 Persona decomposition of node U

网络表示学习属于机器学习领域，该方法就是构建各个节点到向量空间的映射模型(节点嵌入)，将网络中的每一个节点都映射到多维空间中(用向量表示)，即在向量空间内距离近的节点之间其上下文关系也比较相似。而目前大多节点嵌入方法都是将一个节点的嵌入表示为该节点多个语义的均值，这类节点嵌入方法在连续空间中可能没有意义，因为网络中各个节点对应的多个角色代表了其在不同社区中不同的网络行为，所以 Epasto[15] 提出了对一个节点的社交信息的不同分量进行建模，也就是将节点映射为多个向量表示，从而使向量表示更加贴合网络中节点之间的语义关系。

2.3 链路预测

链路预测方法是通过计算网络中两个节点之间潜在关系的强度来进行推荐，其方法主要分为基于相似性的方法和基于机器学习的方法[20]。基于相似性的方法大多通过建立相似性指标，使用评分函数计算网络中目标节点对的相似性得分，不同方法的主要区别体现为相似性指标的不同。该类方法假设得分较高的节点对之间更有可能存在链路，例如基于节点属性信息、基于网络局部信息、基于路径和基于随机游走等方法[20]。而计算机硬件技术的发展使得深度学习被应用在更多的研究领域，其在计算机视觉和自然语言处理等领域的应用也激发了对链路预测领域的研究。Mikolov 提出的模型可以计算大规模数据集中的向量表示[21]，进一步计算相似性得分来完成链路预测任务。由此，基于机器学习的方法主要是通过已知的网络链路结构构建学习模型，使用所学习的模型预测网络中各节点对之间存在潜在链路的可能性，有助

于解决数据分析与挖掘中的常见问题，例如过度拟合、类别不平衡等。该类方法目前主要分为3类研究方向：基于特征分类、基于概率模型以及基于矩阵分解的方法[20]。本文提出的方法Ego-Embedding就是基于概率模型，将自我中心网络的结构特征信息融入到概率模型的构建中，即综合考虑了节点的语义关系和网络特征，能够使训练模型的预测效果更好。

3　问题提出与算法设计

3.1　问题提出

网络表示学习也称为图嵌入，是最近非常活跃的研究领域，并且能够被很好地应用在学习任务中，比如节点的嵌入表示改进了数据挖掘和机器学习任务的结果。但是嵌入方法发展到现在，几乎所有的研究工作最关键的假设均是给图中的每一个节点学习一个向量表示。因此，嵌入方法可以说是定义几何图中每个节点的单个角色或位置，这是一个“平均”或者“全局”位置。在研究初期，图聚类主要集中在无重叠聚类方法的工作，在这些方法中，每个节点都被分配给单个社区。当无重叠聚类问题被更好地理解且被更多地应用以后，最近重叠聚类方法也得到了关注，在这类工作中，每个节点允许参与到多个社区中。除此之外，从全局级别观察真实世界的网络，网络缺乏清楚的社区结构，研究表明社区结构在局部层次上更容易被识别。

因此，本文的Ego-Embedding算法是基于最近发展的自我中心网络分析，特别是基于自我中心网络分解，将自我中心网络与网络表示学习（嵌入）结合起来，综合考虑网络的局部结构信息和上下文信息，为参与到多个社区中的角色节点学习向量表示，即为角色节点对应的每一个原节点学习多个向量表示，使向量表示更准确地描述节点的结构及语义信息，从而提高链路预测的精确度。

3.2　算法设计

通过对自我中心网络结构特征和嵌入学习在链路预测领域应用的分析与研究，本文提出了一个新的链路预测算法框架Ego-Embedding，此方法可以将网络的结构特征和上下文关系结合起来。该框架主要利用自我中心网络的3个结构特征：(1)中心节点的中介中心性(betweenness centrality)；(2)中心节点和其邻居节点的关系；(3)节点的多样化程度，重构了各个网络节点的嵌入方式，使其更好地应用在基于网络表示学习模型的链路预测任务中。

3.2.1　中介中心性

社交网络分析中的“中心性”这一概念是用来分析在社交网络中某个用户或群体的中心位置或者权力。网络中某个节点处于网络中心的程度可以用该节点的中介中心性表示，即中介中心性描述了网络中一个节点的中心程度。网络中的各个节点都可以计算其中介中心性，用来刻画节点特性。Freeman提出了中介中心性[17]这一指标来衡量中心节点与相邻节点之间的紧密程度。其含义是：核心节点就是在网络中位于其他节点对之间的多条最短连通路径上的节点，即该节点作为其他节点之间发生关系的“中介”。一个节点的中介中心性高说明该节点做“中介”的次数多，控制的资源多，就有越多的节点需要通过它才能与其他节点发生联系，进一步说明该节点对邻居节点的网络行为有影响。

中介中心性的计算如下：首先计算网络中的节点对之间最短连通路径的路径数和路径长度；其次，计算网络中每一个节点作为“中介”的所有节点对之间的路径总数，可以用式(1)计算：

$$B_i=\sum_{j\neq k\neq i\in V}\frac{g_{jk}(i)}{g_{jk}} \tag{1}$$

其中，B_i 表示节点 i 的中介中心性；V 是网络中所有节点；分母 g_{jk} 是连接节点 j 和节点 k 的最短路径数目；分子 $g_{jk}(i)$ 是这些最短路径中包含节点 i 的数目。

因此，在根据网络的语义信息进行嵌入表示学习时，如果网络中的某些节点的中介中心性很高，即该节点在网络中充当除其之外的其他节点对的多条最短路径的“中介”，那么在利用随机游走(DeepWalk)生成语料库时该节点很容易与其他节点共现，进而影响模型训练的结果，也就是说中介中心性高的节点会有更高的输出概率。基于上述分析，本算法的第一步是提出方法ego-betweenness，即在Word2vec的嵌入过程中，将各个中心节点的中介中心性作为惩罚因子，对于最终形成的语料库来说，只保留每个节点的 $N*(1-b)$ 条游走序列（N 是某节点的原节点序列总数，b 是该节点在自我中心网络中对应的中介中心性）。

3.2.2　中心节点和邻居节点的关系

人类学和社会学的相关研究表明，人脑的认知极限限制了个体能够积极维持的社会关系的数量。实际上，保持社会关系的活跃需要认知资源，而认知资源又受到自然的限制。邓巴提出的大脑假说理论，认为一个人只能维持大约150个朋友[22]；并且陆续有研究证明，新的通信技术和设备也无法改变人们组织社交关系的方式，在在线和真实世界的社交网络中，用户的自我中心网络的结构相似度很高，并且其平均规模也符合邓巴数定律[16]。一旦自我中心网络超过临界值（邓巴数），情感亲密度将显著下降，自我中心与社交网络的平衡关系也将崩溃，强关系比例下降[23]。

个体的自我中心网络表现为一系列具有特征性的联系人社交圈[19]，这些社交圈根据不断降低的亲密程度排列在一个有层次的同心圆中，如图3所示。

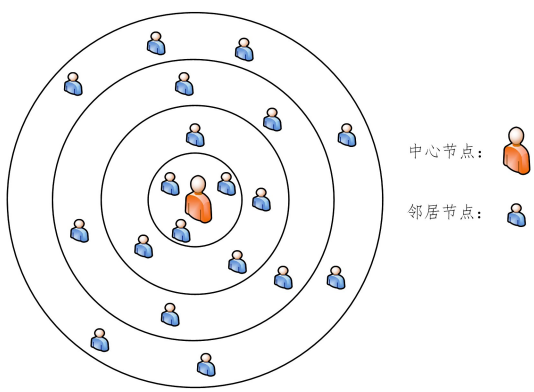


图3　自我中心网络的中心节点与邻居节点之间的亲密程度

Fig. 3　Degree of intimacy between the ego and alters in the ego network

这些社交圈由中心节点和邻居节点之间的联系频率来定义，一般有4层，且社交圈中的邻居节点随着亲密程度的增加而减少。中心节点与邻居节点的关系越紧密，越需要更多的

资源(时间和精力等)来维系亲密关系,这与人类有限的注意力紧密相关。

因此,在每个节点的自我中心网络中,中心节点与其邻居节点的关系亲密度比整个社交网络中的其它他节点高,并且中心节点和邻居节点之间的亲密程度在接近中心节点的层次上(在同心圆中半径小的)比那些较外部的更强烈。

Node2vec 算法[24]是在 DeepWalk 算法[25]的基础上由 Grover 等人提出的,他借鉴 Word2vec 算法[26]提出了有偏随机游走,优化了 DeepWalk 算法,如图 4 所示。其核心思想是采用更灵活的选择邻居节点的游走方法——利用参数 p 和 q 来控制在产生游走序列时的广度优先搜索(BFS)和深度优先搜索(DFS)偏重(式(2)),并且设计了一个二阶转移概率公式(式(3))来计算游走时选择下一个节点的转移概率。假设从节点 t 出发进行随机游走,该时刻到达节点 v,选择下一个节点 x 的几种情况如图 4 所示。

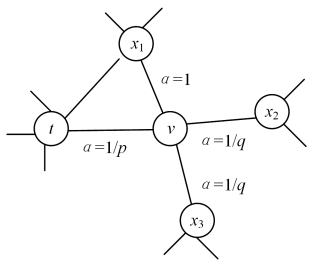


图 4 游走序列的选择

Fig. 4 Choice of walking sequence

$$\alpha_{pq}(t,x)=\begin{cases}\dfrac{1}{p}, & \text{if } d_{tx}=0\\ 1, & \text{if } d_{tx}=1\\ \dfrac{1}{q}, & \text{if } d_{tx}=2\end{cases} \tag{2}$$

其中,$\alpha_{pq}(t,x)$是节点 t 到 x 的游走偏向;d_{tx}代表图 4 中节点 t 到节点 x 路径的最短长度;$d=0$ 表示从节点 v 又回到了 t 本身,$d=1$ 表示节点 x 是 t 的邻居节点,$d=2$ 表示节点 x 与 t 不相邻。所以 p 用于控制重复访问已访问节点的概率,q 用于控制游走是偏广度还是深度优先搜索。

$$\pi_{vx}=\alpha_{pq}(t,x)\cdot w_{vx} \tag{3}$$

其中,π_{vx}是节点 v 到节点 x 的转移概率;w_{vx}是边(v,x)的权重。p,q,w 的值可以根据具体的场景选择。

由上述分析可知,网络中不同位置的节点与中心节点的亲密程度不同,会对中心节点造成不同程度的影响,所以在网络表示学习中,当以某个节点为开始节点生成游走序列时,应该优先选择以该节点为中心节点的自我中心网络中的邻居节点。基于此,本算法的第二步是提出方法 ego-n2v,使用有偏游走 node2vec 算法,在产生游走序列时,通过选择两个参数 p 和 q 来控制下一个节点的选择,即目标是偏向广度优先搜索,优先选择亲密度高的邻居节点。

3.2.3 节点的多样性

网络的不同部分以不同的方式发展,对网络的不同部分应用不同的方法可能会提升预测效果,例如将网络中的社区信息应用在链路预测任务中。从社区级别可以观察到,对于同一社区来说,节点之间的连接比不同社区的节点之间的连接更加密集。从网络的结构来看,如果一个节点连接到多个社区,那该节点就被认为比主要连接到一个社区的节点具有更加多样性的特点,也可以说是该节点的兴趣广泛,更容易接受新事物的影响,也就更有可能形成新的链路。目前已经有很多通过集成社区信息来提高链路预测准确性的研究,例如 Gupta 提出了社区桥节点概念[27],设置节点的度阈值和最大社区主导比率(MCDR)筛选过滤桥节点,然后通过将所选桥节点的相似性得分加倍来提高链路预测准确性。

基于此,本算法的第三步是提出方法 ego-bridge,先使用社区检测算法对节点的自我中心网络划分社区,然后选择出桥节点集合,并假设更具多样性的桥节点比社区结构较单一的节点更容易产生新链路。ego-bridge 通过改变网络中桥节点与其他节点之间已知链路的权重,重构网络表示学习模型,在产生游走序列时选择桥节点的概率增大,即桥节点与其他节点共现的概率更大,更容易被推荐。

综上所述,Ego-Embedding 算法框架的流程如算法 1 所示。

算法 1 Ego-Embedding 算法

输入:网络图,包括训练集中的所有节点和所有已知链路,即原图

输出:网络图中每个节点对应的一个或多个向量表示

Step1 构建原图中以各个节点作为中心节点的自我中心网络。

Step2 对自我中心网络进行社区检测,构建各个节点的角色节点集合,即进行角色分解;再根据角色节点和原节点的对应关系生成角色图。

Step3 计算自我中心网络中社区之间的桥节点,并改变桥节点对应的角色节点在角色图中与其他节点之间链路的权重 w,即 ego-bridge。

Step4 通过 grid search 得到最佳的参数 p 和 q,并使用 node2vec 模型对角色图中每个角色节点生成有偏游走序列,即 ego-n2v。

Step5 计算各个中心节点的中介中心性,并将该值作为节点游走序列的惩罚因子,即 ego-betweenness。

Step6 将最终得到的节点序列集合看作语料库,对节点序列使用 skip-gram 模型进行概率建模(计算每个窗口内的节点对的共现概率),使用梯度下降算法训练模型,从而将每一个角色节点表示为多维空间中的嵌入向量,即原图中的每个节点都被表示为一个或多个向量表示。

本算法的复杂度主要受到生成角色图和训练模型两个步骤的影响。假设一个聚类算法分析 m 条边网络的复杂度是 $T(m)$,那么在最坏情况下,一个有 m 条边的网络生成角色图的复杂度是 $O(m^{\frac{3}{2}}+\sqrt{m}T(m))$。角色图中每个角色节点的嵌入表示的复杂度是 $O(m\gamma twd\log(m))$,其中 γ 是随机游走执行次数,t 是游走长度,w 是窗口大小,d 是嵌入维度。

4 实验结果与分析

本算法选取了 3 个公开网络数据集进行实验,且 3 个数据集分属不同领域,通过对本文所提出的算法框架和现存的多种链路预测算法进行对比,进一步验证了本算法的可行性和有效性。

4.1 实验数据集

本文选用公开实验数据集中的 3 个不同领域的数据集。

(1)Facebook 社交网络:节点表示用户,边表示两个用户之间的好友关系。

(2)ca-HepTh 物理学研究合著者网络:节点表示论文作

者，边表示两个作者之间的合著关系。

(3)PPI-Yeast 蛋白质相互作用网络：节点表示蛋白质，边表示两种蛋白质之间的相互作用。

3 个数据集的有关网络拓扑信息如表 1 所列，其中 n 表示网络总节点数，e 表示网络总边数，c 表示网络的聚类系数，d 表示网络模块度。

表 1　实验数据集的基本统计信息

Table 1　Basic statistics of the experimental data sets

数据集	n	e	c	Q
Facebook	4039	88234	0.6055	0.835
Ca-HepTh	9877	25998	0.4714	0.762
PPI-Yeast	2617	11855	0.3870	0.573

在本文的实验中，数据集的划分是按照文献[15]中的方法：随机且均匀地移除网络图中的边(保证图的连通性)，将图划分为两个相等的边集合，分别作为训练集和测试集。

4.2　实验对比方法和实验设置

本文选用了 5 个比较具有代表性的链路预测算法进行实验对比。

基于网络结构相似性的非嵌入方法：选择了共同邻居(Common Neighbors，CN)指标[28]和 AA(Adamic Adar)指标[28]。其中，CN 指标考虑两个节点之间的共同邻居节点数，AA 指标考虑节点对共同邻居节点的重要程度。指标定义的评分函数如下：

$$\mathrm{CN}: S_{xy}=|N(x)\cap N(y)| \tag{4}$$

$$\mathrm{AA}: S_{xy}=\sum_{z\in N(x)\cap N(y)}\frac{1}{\log k(z)} \tag{5}$$

其中，$N(x)$表示训练集中节点 x 的邻居节点的集合，S_{xy}表示节点 x 和节点 y 的相似性得分，$k(z)$表示节点 z 的度。假设得分越高，表示这两个节点之间越容易产生链路。

基于学习的嵌入方法：选择了 Node2vec 算法[24]、M-NMF 算法[30]和 Splitter 算法[15]。Node2vec 算法是通过对在训练网络上执行有偏随机游走生成的节点序列进行概率建模，为网络中的每一个节点学习一个向量表示。M-NMF 算法使用基于模块化的社区检测模型来优化网络中每个节点的社区划分和嵌入表示，也是将每一个节点映射成一个向量表示。Splitter 算法是先对网络中的节点进行角色分解，再使用 Deepwalk 模型将每一个角色节点映射为一个向量表示。

在利用上述方法进行实验时，使用节点对 u 和 v 的相似性得分来对链路(u,v)产生的可能性进行排序，该相似性得分是使用评分函数计算的。其中，非嵌入方法主要是输入节点的邻域信息，采用式(4)、式(5)计算指标结果；嵌入方法是将节点的向量表示作为输入，其评分函数定义如下：node2vec 使用二元分类逻辑回归算法来模拟两个节点特征的 Hadamard 乘积的模型；M-NMF 和 node2vec 相似；Splitter 是将两个节点各自对应的多个角色节点的向量表示任意两两组合计算点积，点积的最大值就是这两个节点之间的相似性得分。

实验中的参数都按照上述方法推荐的默认参数进行设置。

本文的算法框架 Ego-Embedding 为网络中的每个节点输出一个或多个向量表示，因此需要从可能的多个角色节点的成对相似性得分中提取单个得分，所以将节点对的相似性得分定义为它们任意角色组合之间的最大点积。另外，实验中的参数设置如下：将网络中已知边的权重初始化为 $w=1$；计算桥节点的筛选过程中用于量化桥节点与社区连接比例的值 $R=0.7$，更新桥节点与其他节点之间已知边的权重 $w=2$；游走长度 $t=40$，游走次数 $\gamma=10$；窗口大小 $ws=5$；学习率 $\alpha=0.025$；嵌入维度 $d=128$。

4.3　评价指标

为了评估实验结果的准确性，本算法采用评价链路预测算法精确度的常用指标——AUC(Area Under the receiver operating characteristic Curve)，AUC 是 ROC 曲线下的面积[31]。本实验中使用不存在的边集作为负样本。AUC 的评分过程可理解为在测试集中随机选择一条边，再从负样本中随机选择一条边，如果在测试集中取得边的分数大于负样本中的边的分数，记 1 分；如果两者相等，记 0.5 分；前者小于后者则不记分。AUC 越高，代表链路预测的精确度越高。假设一共比较 n 次，若有 n'次测试集所取边分数高，n''次两者分数相等，则 AUC 可以计算为：

$$AUC=\frac{n'+0.5n''}{n} \tag{6}$$

4.4　实验结果分析

为了得到 Ego-Embedding 算法最佳的预测结果，首先在 3 组实验数据集上利用 4.2 节中所述的参数设置针对不同的参数 p 和q 进行实验。在 Facebook 网络上的实验结果如表 2 所列，在 Ca-HepTh 网络上的实验结果如表 3 所列，可以发现当 $p=q=2$ 时这两个数据集的预测 AUC 值最高。而在 PPI-Yeast 网络上，当 $p=1.5$，$q=2$ 时其预测效果最佳，如表 4 所列，不过 $p=q=2$ 的实验结果与最佳结果相差并不大。因此，在 Ego-Embedding 算法中将参数设置为 $p=q=2$。

表 2　在 Facebook 数据集上设置不同参数值的 AUC 值

Table 2　AUC of different parameter values on the Facebook dataset

	$q=0.5$	$q=1.0$	$q=1.5$	$q=2.0$
$p=0.5$	0.8901	0.8832	0.8892	0.8812
$p=1.0$	0.8862	0.8894	0.8935	0.9125
$p=1.5$	0.8894	0.9067	0.9168	0.9158
$p=2.0$	0.8933	0.9122	0.9046	**0.9234**

表 3　在 Ca-HepTh 数据集上设置不同参数值的 AUC 值

Table 3　AUC of different parameter values on the Ca-HepTh dataset

	$q=0.5$	$q=1.0$	$q=1.5$	$q=2.0$
$p=0.5$	0.9015	0.9031	0.9033	0.9095
$p=1.0$	0.9127	0.9064	0.9187	0.9213
$p=1.5$	0.9076	0.9106	0.9151	0.9275
$p=2.0$	0.9013	0.9122	0.9236	**0.9306**

表 4　在 PPI-Yeast 数据集上设置不同参数值的 AUC 值

Table 4　AUC of different parameter values on the PPI-Yeast dataset

	$q=0.5$	$q=1.0$	$q=1.5$	$q=2.0$
$p=0.5$	0.9167	0.9288	0.9187	0.9233
$p=1.0$	0.9189	0.9376	0.9206	0.9315
$p=1.5$	0.9215	0.9384	0.9358	**0.9426**
$p=2.0$	0.9261	0.9409	0.9455	**0.9413**

为了比较的一致性，本文使用的实验对比方法中的参数设置都是其默认参数，且对数据集的划分方法也相同。本文对每个数据集都分别使用每种方法进行了 30 次独立实验，每次实验都重新按照上述方法划分训练集和测试集，并且通过

20000 次随机抽样测试集与不存在边集中的边来比较其相似性得分，以此计算相应的 *AUC*。实验结果如表 5 所列，取 30 次独立实验的平均值。

表 5 各方法在各数据集上的 *AUC* 值

Table 5 *AUC* of each method on each dataset

	Facebook	Ca-HepTh	PPI-Yeast
CN	0.6861	0.7648	0.8513
AA	0.7661	0.7842	0.8659
Node2vec	0.8229	0.8610	0.9031
M-NMF	0.8327	0.8785	0.9105
Splitter	0.8563	0.9092	0.9187
ego-betweenness	0.8724	0.9023	0.9218
ego-n2v	0.9190	0.9231	0.9402
ego-bridge	0.8601	0.9117	0.9279
Ego-Embedding	0.9234	0.9306	0.9413

通过对表 5 所列的实验结果进行分析发现，本文提出的算法框架 Ego-Embedding 相比其他 5 种对比方法都有提升，且最高将链路预测的错误率减少了约 47%。

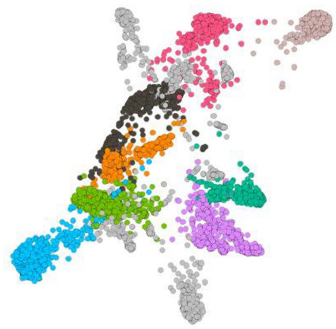
从表 5 可以看到，本文的算法实验分为以下 4 步。(1)ego-betweenness是将网络中各个节点作为中心节点计算其中介中心性，然后将该值作为惩罚因子，按照 3.2.1 节中介绍的方法优化节点序列，得到新的节点序列集合。(2)ego-n2v 是将 Node2vec 作为网络表示学习模型(通过上述 grid search 计算得到最优参数 $p=q=2$，且在这一步骤中将网络中所有已知连边的权重 w 均初始化为 1)学习各个节点的向量表示。在此过程中，先将原网络通过角色分解转换成角色图，再将角色图作为网络表示模型的输入网络，为每一个角色节点学习一个向量表示(利用原网络的节点向量表示初始化角色节点)，最后将角色节点对应到原网络中的节点，即会给每一个节点学习一个或者多个向量表示。(3)ego-bridge 是使用贪婪模块化方法 Louvain 进行社区检测并筛选过滤出桥节点集合，然后修改桥节点所在连边的权重为 2，增大桥节点在游走序列产生时的选择概率。(4)Ego-Embedding 是将前 3 步整合起来，即按照算法 1 进行实验所得最终结果。

上述实验步骤对应的实验结果如表 5 所列。可以看到，ego-betweenness 对预测精确度的提升较低，甚至在 ca-HepTh 数据集中的表现没有对比方法 Splitter 好，其原因可能是该网络属于稀疏网络，其节点的中介中心性的值影响较小；ego-bridge 对预测精确度也有提升但较低，其原因是 Facebook 和 Ca-HepTh 网络的模块度较高，社区结构比较清晰，从而社区之间的桥节点数量相对较少，因此仅提升了桥节点相关链路的预测结果，对整体网络的精确度提升不是很明显；ego-n2v 是这 3 个方法中精确度提升最多的，因为该模型在构建节点序列时，对自我中心网络的结构信息和语义信息充分进行了利用；Ego-Embedding 将上述 3 个方法整合，所得结果在 3 个数据集上的表现都优于其他 5 个对比方法，该算法在 Facebook 网络上的预测精确度表现最好，将错误率减少了 47%，因为该网络节点数量和连边数量较多，各个节点的自我中心网络中邻居节点较多，就近邻居的结构信息和上下文信息的充分利用优化了模型，使得预测结果提升较高。

综上，本文提出的算法框架 Ego-Embedding 在不同类型和不同规模的网络上的预测精确度都相对较优。

4.5 可视化

为了体现对网络中的节点根据其社区身份进行角色分解后对重叠社区聚类的有效改进，本小节利用可视化结果来表明与原网络相比，通过角色分解得到的角色图能够划分出更清晰的社区结构。如图 5 所示，图 5(a)是 Facebook 数据集随机划分的训练数据集的社区检测结果，图 5(b)是该训练数据集通过角色分解得到的角色图上的社区划分结果(使用具有相同设置的可视化工具 Gephi，其采用的社区检测算法是 Louvain 算法)。图中节点的颜色对应于使用贪婪模块化优化检测的社区，可以观察到，图 5(a)检测到的社区有很多重叠部分，而图 5(b)检测的社区将重叠区域更好地分离开，能够将社区结构识别得更加清楚。



(a)在原网络上进行社区检测的结果

(b)利用自我中心网络进行角色分解后社区检测的结果

图 5 角色分解前后进行社区检测的结果对比(电子版为彩色)

Fig. 5 Community detection before and after persona decomposition

结束语 本文介绍了一种新的链路预测算法框架 Ego-Embedding，该方法以自我中心网络分析和重叠聚类的最新研究为基础，利用网络表示学习将自我中心网络的结构特征和网络上下文特征结合起来，为网络中的每个节点学习一个或多个向量表示。从网络的结构信息和语义信息推断节点之间的链路可能性，突破了其他方法的某些局限性。实验验证，在不同类型的真实网络中，该算法的预测精确度较其他方法均有所提升；并且，本算法主要基于网络中以各个节点为中心节点的自我中心网络，所以也可应用于大规模网络。

后续研究主要关注以下挑战：(1)将该方法扩展到有向网络中，在本算法的基础上分析网络节点之间的交互；(2)利用节点之间的互动频次来刻画节点对之间的亲密关系；(3)将该嵌入方法应用在分类任务中。

参 考 文 献

[1] LI L, FANG S, BAI S, et al. Effective Link Prediction Based on Community Relationship Strength[J]. IEEE Access, 2019, 7: 43233-43248.

[2] LIAO H, MARIANI M S, MEDO M, et al. Ranking in evolving complex networks[J]. Physics Reports, 2017, 689: 1-54.

[3] PEROZZI B, SKIENA S. Exact Age Prediction in Social Net-

works[C]// the 24th International Conference. ACM,2015.
[4] SAID A,ABBASI R A,MAQBOOL O,et al. CC-GA:A Clustering Coefficient based Genetic Algorithm for Detecting Communities in Social Networks[J]. Applied Soft Computing,2017,63:59-70.
[5] ABU-EL-HAIJA S,PEROZZI B,AL-RFOU R. Learning Edge Representations via Low-Rank Asymmetric Projections[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM'17). NewYork: ACM,2017:1787-1796.
[6] YUAN R,SONG Y R,MENG F R. A Link Prediction Method Based on Weighted Network Topology Weight[J]. Computer Science,2020,47(5):273-278.
[7] YANG X H,YU J,ZHANG D. Link prediction algorithm based on local community and node correlation[J]. Computer Science,2019,46(1):155-161.
[8] MA C,ZHOU T,ZHANG H F. Playing the role of weak clique property in link prediction:A friend recommendation model[J]. Scientific Reports,2016.
[9] ZHANG Q M,XU X K,ZHU Y X,et al. Measuring multiple evolution mechanisms of complex networks[J]. Scientific Reports,2015,5(1):10350.
[10] MARTÍNEZ V,BERZAL F,CUBERO J C . A Survey of Link Prediction in Complex Networks[C]// ACM Computing Surveys (CSUR). 2016.
[11] CANNISTRACI C V,ALANIS-LOBATO G,RAVASI T . From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks[J]. Scientific Reports,2013,3(1):1613.
[12] HANNEMAN R A,RIDDLE M. Introduction to social network methods. Riverside[D]. CA:University of California,Riverside,2005.
[13] EPASTO A,LATTANZI S,MIRROKNI V,et al. Ego-net community mining applied to friend suggestion[J]. VLDB,2015,9(4):324-335.
[14] EPASTO A,LATTANZI S,LEME R P . Ego-splitting Framework:from Non-Overlapping to Overlapping Clusters[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM,2017.
[15] EPASTO A,PEROZZI B . Is a Single Embedding Enough? Learning Node Representations that Capture Multiple Social Contexts[C]// WWW'19. 2019:394-404.
[16] GONCALVES B,PERRA N,VESPIGNANI A,et al. Modeling Users' Activity on Twitter Networks:Validation of Dunbar's Number[J]. Plos One,2011,6(8):e22656.
[17] FREEMAN L C . Centered graphs and the structure of ego networks[J]. Mathematical Social Sciences,1982,3(3):291-304.
[18] MCAULEY J,LESKOVEC J. Learning to Discover Social Circles in Ego Networks[C]// NIPS. 2012:539-547.
[19] ARNABOLDI V,CONTI M,PASSARELLA A,et al. Analysis of Ego Network Structure in Online Social Networks[C]// Ase/ieee International Conference on Social Computing & Ase/ieee International Conference on Privacy. IEEE,2012.
[20] ZHANG Y X,FENG Y X. Overview of Link Prediction Methods and Development[J]. Measurement & Control Technology,2019,324(2):12-16.
[21] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient Estimation of Word Representations in Vector Space[C]// International Conference on Learning Representations. 2013.
[22] DUNBAR R I M. The Social Brain Hypothesis[J]. Evolutionary Anthropology Issues News & Reviews,1998,6(5):178-190.
[23] WANG Q,GAO J,ZHOU T,et al. Critical size of ego communication networks[J]. Europhysics Letters,2016(114):58004.
[24] GROVER A,LESKOVEC J . node2vec:Scalable Feature Learning for Networks[C]// the 22nd ACM SIGKDD International Conference. ACM,2016.
[25] PEROZZI B,AL-RFOU R,SKIENA S. DeepWalk:Online Learning of Social Representations[C]// Knowledge Discovery and Data Mining. 2014.
[26] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed Representations of Words and Phrases and their Compositionality[C]// Proceedings of NIPS. 2013.
[27] GAO F,MUSIAL K,GABRYS B. A Community Bridge Boosting Social Network Link Prediction Model[C]// the 2017 IEEE/ACM International Conference. ACM,2017.
[28] NEWMAN M E J. Clustering and preferential attachment in growing networks[J]. Phys Rev E Stat Nonlin Soft Matter Phys,2001,64(2):025102.
[29] ADAMIC L A,ADAR E. Friends and neighbors on the Web[J]. Social Networks,2003,25(3):211-230.
[30] WANG X,CUI P,WANG J,et al. Community Preserving Network Embedding[C]// The 31st AAAI Conference on Artificial Intelligence. 2017.
[31] HANELY J A,MCNEIL B J. The meaning and use of the area under a receiver operating characteristic(ROC) curve[J]. Radiology,1982,143(1):29-36.

ZHAO Man, born in 1996, postgraduate. Her main research interests include complex network, link prediction and data analysis.