智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于图卷积神经网络的最短路径距离估计方法

孟祥福, 崔江燕, 邓敏超

引用本文:
孟祥福, 崔江燕, 邓敏超. 基于图卷积神经网络的最短路径距离估计方法[J]. 智能系统学报, 2024, 19(6): 1518-1527.
MENG Xiangfu, CUI Jiangyan, DENG Minchao. Road network shortest distance estimation method based on graph convolutional networks[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1518-1527.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202309006

DOI: 10.11992/tis.202309006
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20240709.0951.002

基于图卷积神经网络的最短路径距离估计方法

孟祥福, 崔江燕, 邓敏超
(辽宁工程技术大学 电子与信息工程学院, 辽宁 葫芦岛 125105)

摘　要: 路网最短路径距离估计问题的关键是提高估计准确度和减少模型训练时间。现有基于嵌入的最短路径距离估计方法要么模型的训练时间较长, 要么通过牺牲估计精度来降低模型训练时间。针对以上问题, 通过分析基于嵌入的最短路径距离估计方法, 提出路网最短路径距离估计编码器−解码器框架, 归纳和整合这类方法的核心过程, 并将核心过程分为嵌入方法、采样方案和模型训练 3 部分。在此基础上, 提出一种基于图卷积网络的路网顶点嵌入方法(road graph convolutional networks and distance2vector, RGCNdist2vec), 用于捕获路网的结构信息。在模型训练样本的采样方面, 设计一种基于图逻辑分区的三阶段采样方法, 能够选取少量优质样本用于模型训练。为验证模型及采样方案的有效性, 在 4 个真实路网数据集上开展实验, 并与现有相关模型进行对比, 结果表明所提模型具有较高的估计准确性, 并且模型训练时间降低为现有基线模型的 1/4。

关键词: 最短路径距离计算; 图神经网络; 数据采样; 表示学习; 图卷积网络; 图分区; 深度学习; 拓扑结构
中图分类号: TP302.7　　**文献标志码**: A　　**文章编号**: 1673−4785(2024)06−1518−10

中文引用格式: 孟祥福, 崔江燕, 邓敏超. 基于图卷积神经网络的最短路径距离估计方法 [J]. 智能系统学报, 2024, 19(6): 1518–1527.
英文引用格式: MENG Xiangfu, CUI Jiangyan, DENG Minchao. Road network shortest distance estimation method based on graph convolutional networks[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1518–1527.

Road network shortest distance estimation method based on graph convolutional networks

MENG Xiangfu, CUI Jiangyan, DENG Minchao
(School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: Improving the accuracy of estimating the shortest path distance while reducing model training time is crucial. Existing methods for embedded shortest path distance estimation often take too long to train or sacrifice accuracy to save time. To solve these problems, an encoder-decoder framework has been designed to estimate the shortest distance in a road network by analyzing existing embedded systems-based shortest-path distance estimation methods. The core process is broken down into three parts: embedding method, sampling method, and model training. A road network vertex embedding method, RGCNdist2vec, leverages road graph convolutional networks to capture the structural information of the road network. For model training, a three-stage sampling method using graph logical partitioning is designed to select a small number of high-quality samples. Experiments conducted on four real road network data sets demonstrate that the proposed model achieves higher estimation accuracy while reducing training time by nearly four times compared to existing baseline models.

Keywords: shortest path distance computation; graph neural networks; data sampling; representation learning; graph convolutional networks; graph partitioning; deep learning; topology

计算路网中两点间的最短路径距离是路网应用场景中的核心问题, 在地图服务和导航系统中

收稿日期: 2023−09−02.　**网络出版日期**: 2024−07−11.
通信作者: 孟祥福. E-mail: marxi@126.com.

有着广泛的应用。例如, 打车软件需要快速准确地为用户匹配最近的车辆, 从而减少用户等待时间。最短路径距离的计算方法可分为准确计算[1-4]和估计计算[5-14]2 种。其中, 估计计算方法计算速度快效率高, 能够为用户提供更为高效的即时反馈。并且, 在大多数情况下采用估计计算方法并不会影响最终结果的准确性。

近年来, 一些研究者提出了基于图嵌入[11-14]的路网最短路径距离估计方法。这些方法运用深度学习技术构建最短路径距离计算的索引, 核心思想是运用图嵌入方法对路网中的每个顶点进行嵌入处理, 使得每个顶点嵌入到欧氏[15] 或双曲[16]空间获得嵌入向量表示, 然后通过计算顶点向量之间的相似距离来估计顶点之间的最短路径距离。然而, 现有基于深度学习的估计方法采用随机游走[11] 或权重矩阵相乘[14] 的方法对顶点进行嵌入处理, 没有充分考虑路网的结构信息, 并且使用路网中所有顶点对作为训练数据, 在估计准确度和模型训练代价方面仍有较大提升空间。文献[17] 对路网最短路径距离估计任务的前沿方法进行了详细的介绍和实验评估, 并着重分析了不同算法的优劣。文献 [18-20] 扩展了基于嵌入方法的应用场景, 为农业和路由规划任务提供技术支持。

图神经网络(graph neural networks, GNN)是一种用于处理图结构数据的深度学习模型。路网的本质是图结构数据, 图神经网络为路网场景下的最短路径距离估计任务提供了新的研究视角。图卷积网络(graph convolutional networks, GCN)[21]是近几年图神经网络的代表之一, 基本思想是利用图的结构信息和顶点的特征信息, 通过卷积操作来更新顶点的嵌入向量表示, 提取图的深层次特征和捕获图的空间关键信息。因此, 本文基于 GCN 构建了一种适用于路网的嵌入方法, 用来提取路网中顶点和边的关联模式。

本文贡献主要包括以下 4 个方面:

1)对基于嵌入方法的核心部分进行分析, 提出了路网最短路径距离估计编码器–解码器(road shortest distance estimation encoder-decoder framework, RSDE)框架, 提炼了路网最短路径距离估计任务的核心过程。

2)在上述贡献的基础上, 针对路网场景, 提出了一种基于图卷积神经网络的路网最短路径距离估计嵌入方法。

3)在模型的训练样本采样方面, 提出了一种基于图逻辑分区的三阶段采样方法。

4)将所提模型 RGCNdist2vec 在 4 个真实路网上进行实验验证, 结果表明该模型具有显著的效果和性能提升。

1　相关工作

1.1　最短路径距离的计算方法

路网最短路径距离的经典计算方法包括 Dijkstra[1]、Floyd[2] 等。由于这些方法需要访问终点到起点路径范围内所有的点, 计算复杂度高。为了降低计算复杂性, Robert 等[3] 提出了缩短层次网络 (contraction hierarchies, CH) 方法, 该方法通过计算图中某些重要顶点间的最短距离来加速计算和查询。Ouyang 等[4] 提出了 H2H 方法, 该方法通过为每个顶点分配距离标签来提高计算效率。然而, 上述方法都需要大量的预处理时间构建索引, 不能适用于大规模道路网络。

为了适用于大规模路网数据场景, 估计方法被提出用来解决上述问题。基于地标的方法[8-10]是一类典型的估计路网最短路径距离的方法, 该类方法的基本思想是从图中选择部分顶点作为地标顶点, 然后为每个顶点分配距离标签存储该顶点到所有地标顶点的距离。当计算图中任意 2 点的最短路径距离时, 可用 2 个顶点与同一路标顶点的最小距离之和作为结果。虽然基于地标的方法能够降低内存开销, 但由于选择地标顶点是一个 NP-hard(non-deterministic polynomial hard)问题, 所以这类方法的计算准确度无法保证。基于嵌入的方法是对地标方法的一种突破, 该类方法的基本思想是结合图嵌入技术估计大规模路网的最短路径距离。Rizi 等[11] 最早提出运用深度学习技术对图顶点进行嵌入处理, 从而逼近大规模图中的最短路径距离。该方法使用 node2vec[22] 和 Poincare[23] 作为嵌入方法, 在社交网络图上取得了很好的效果, 并在实验中验证 node2vec 不适用于路网结构。Qi 等[12] 提出了一种基于学习的方法来估计路网最短路径距离, 该方法虽然有较高的估计准确度, 但其选取图中所有顶点对作为训练数据, 导致模型训练时间较长。Chen 等[14] 提出了一种基于地标和学习的估计方法, 该模型直接对路网顶点进行编码, 然后运用 MLP(multilayer perceptron)进行训练, 虽然该模型通过减少部分训练轮次的训练数据从而降低了训练时间, 但没有考虑路网结构信息, 并且基于地标的采样具有较大随机性, 进而使得训练结果的估计准确度无法保证。

1.2　图神经网络

近年来, 大量的图神经网络模型被提出, 主要

包括图卷积神经网络[21]、图注意力网络[24]等。大部分图神经网络方法都遵循 Gilmer 等[25]提出的神经信息传递(neural message passing)框架。文献[26-27]详细阐述了 GNN 领域近年来的重要研究成果。Kipf 等[21]提出了 GCN 模型，该模型是一种直推式学习，通过聚合邻居顶点信息产生顶点的嵌入向量表示。Hamilton 等[28]提出了 GraphSAGE 模型，该模型是一种利用顶点的属性信息来高效产生未知顶点嵌入的归纳式学习框架，其核心思想是通过学习一个聚合邻居顶点的表示函数来产生目标顶点的嵌入向量。由于 GraphSAGE 在融合邻居信息时，将各个邻居以同等权重看待，Veličković等[24]提出 GAT 模型，通过添加自注意力机制(self-attention)实现了对不同邻居的权重自适应配置。目前，已有众多现实应用任务采用图神经网络的相关方法，本文提出一个新的视角，将图神经网络方法运用到路网最短路径距离估计任务上，作为提取路网结构的嵌入方法。

2　相关定义与解决方案

2.1　相关定义

定义 1（路网）　路网表示为带权无向图 $G=(V, E, W)$，其中 V 是顶点集合，$v_i \in V$ 为图中的一个顶点；E 是边集合，$e_{v_i,v_j} \in E$ 为顶点 v_i 到顶点 v_j 的边，$w_{v_i,v_j} \in W$ 为边 e_{v_i,v_j} 的权重。

图 1 为路网结构的一个示例，其中有 13 个顶点，16 条边，顶点 v_0 到 v_1 的权重为 2。

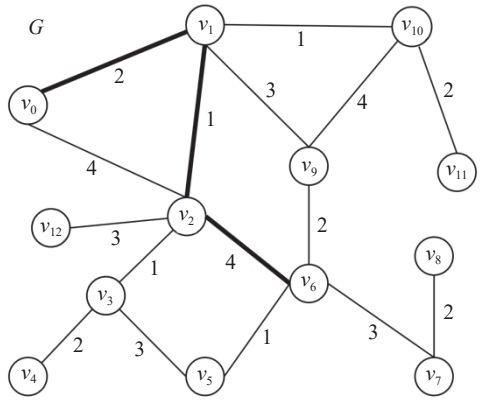


图 1　路网结构示例

Fig. 1　An example of road network structure

定义 2（路网最短路径距离计算）　给定路网中的任意 2 个顶点 v_i 和 v_j，计算它们之间准确距离的方法为 $\varphi(v_i,v_j)=w_{v_i,v_u}+w_{v_u,v_n}+w_{v_n,v_j}$，其中，$v_u$ 和 v_n 为路径上的顶点。

例如，图 1 中标粗线条为 v_0 到 v_6 的最短路径，最短路径距离为 7。

定义 3（路网最短路径距离估计计算）　给定路网 $G=(V, E, W)$，将每个顶点 $v_i \in V$ 嵌入到 d 维空间中，获得其嵌入向量 $\boldsymbol{v}^{(i)}$，$\boldsymbol{H}=\{\boldsymbol{v}^{(1)},\boldsymbol{v}^{(2)},\cdots,\boldsymbol{v}^{(i)}\}$ 为嵌入矩阵。给定任意顶点 v_i 和 v_j，估计距离的计算方法为

$$\hat{\varphi}(v_i,v_j)=|\boldsymbol{H}[i]-\boldsymbol{H}[j]|_1 \tag{1}$$

定义 4（邻接矩阵）　$\boldsymbol{A}_v \in \mathbf{R}^{|V|\times|V|}$ 为路网 G 的邻接矩阵，其中 $\boldsymbol{A}_v(i,j)$ 表示 v_i 到 v_j 的连接性。如果图 G 是无权图且 v_i 和 v_j 有边相连，则 $\boldsymbol{A}_v(i,j)=1$；否则 $\boldsymbol{A}_v(i,j)=0$；如果图 G 是有权图且 v_i 和 v_j 有边相连，则 $\boldsymbol{A}_v^w(i,j)=w_{v_i,v_j}$，否则 $\boldsymbol{A}_v^w(i,j)=0$。

定义 5（划分图）　给定图 G，使用集合 $\Delta(G)=\{G_1,G_2,\cdots,G_n\}$ 表示图 G 的划分结果。将图 G 划分为 n 个子图，则 $\Delta(G)$ 中的元素为 $G_i=(V_i, E_i, W_i)$，$(i\in[1, n])$。每个子图满足以下 3 个条件：1）$V=\mathrm{U}_{i\in[1,n]}V_i$；2）如果 $i\neq j, V_i\cap V_j=\varnothing$；3）每个子图 G_i 为 G 的诱导子图，即子图 G_i 的属性与 G 一致。

给定图 G 和区域阈值 k_f，经过划分后，图 G 被划分为近似大小的 k_f 个子图($|V_1|\approx|V_2|\approx\cdots\approx|V_{kf}|$)。本文采用多层图划分框架 METIS[29]，目的是将图划分为 k_f 个顶点个数相近的子图。

如图 2 所示，图 G 被分为 4 个子图 $\Delta(G)=\{G_1, G_2, G_3, G_4\}$，区域阈值 $k_f=4$。

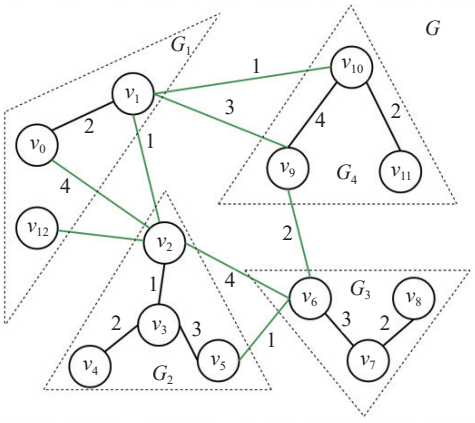


图 2　图 G 的划分结果

Fig. 2　Partition results for G

定义 6（顶点的度）　图顶点的度表示依附于某个顶点的边的数目，用 $D(v_i)$ 表示 v_i 的度。如图 2 所示，$D(v_3)=3$。

定义 7（边界）　给定图 $G=(V, E, W)$，令 $G_x=(V_x, E_x, W_x)$ 为图 G 的一个子图，如果存在一条边 $e_{v_i,v_j}\in E$ 但 $e_{v_i,v_j}\in E_x$，则这样的边称为子图的边界，该边所连接的点是子图的边界点。$B(G_x)$ 定义为 G_x 的边界点，则图 2 中 $B(G_2)=\{v_2, v_5\}$，$B(G_3)=\{v_6\}$。

2.2　解决方案

对现有的基于图嵌入的路网最短路径距离估计方法进行分析和归纳, 可将其核心过程概括为 3 部分: 1)以某种方式进行采样, 并使用传统方法计算所采样数据的最短路径距离, 与采样顶点对拼接作为训练数据; 2)使用图嵌入方法对路网顶点进行嵌入处理, 获得顶点嵌入表示向量。3)选用恰当的度量方式/函数(范数或神经网络等)训练模型来优化顶点的向量表示并输出顶点特征矩阵。

基于上述分析, 本文提出了路网最短路径距离估计编码器解码器框架 (RSDE framework), 该框架组织和提炼了基于嵌入方法中的核心思想和步骤, 方便后续改进和理解各部分的作用。RSDE 框架如图 3 所示, 该框架分为 2 部分, RSDE−编码器和 RSDE−解码器, 解码器层又进一步细分为样本选择过程和图嵌入过程。下文将在此框架的基础上具体阐述各部分的改进策略, 以提升路网最短路径距离估计的准确性和效率。

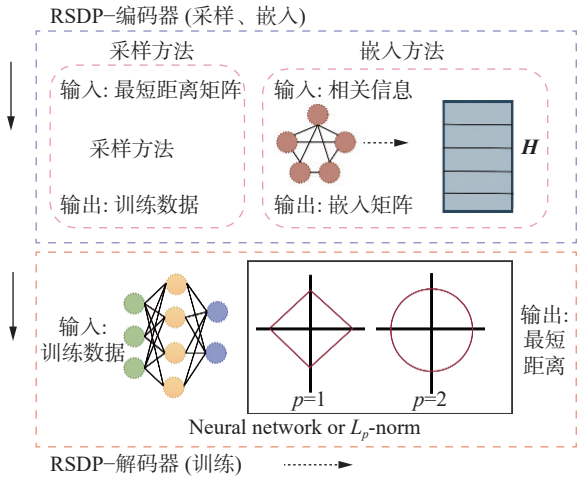


图 3　RSDE 框架结构

Fig. 3　Overall frame structure of RSDE

3　RGCNdist2vec 模型

3.1　模型总体框架

RGCNdist2vec 模型结构如图 4 所示, 主要包括 2 个部分: 1)RGCNdist2vec−编码器层: 构建路网图卷积网络(road graph convolutional networks, RGCN), 通过融合路网权重和结构信息, 获得每个顶点的嵌入向量; 2)RGCNdist2vec−解码器层: 接收编码器层的输出和训练数据, 选择合适的度量方式和损失函数优化编码器层中的可学习权重矩阵, 获得最优的嵌入矩阵, 用于计算路网 2 点间的最短路径距离。

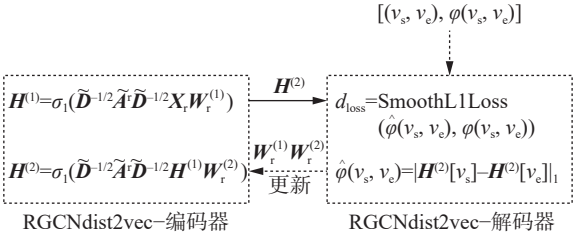


图 4　RGCNdist2vec 模型结构

Fig. 4　Architecture of RGCNdist2vec

3.2　RGCNdist2vec 编码器

图嵌入的作用是整合路网中的信息, 为路网中的每个顶点生成低维嵌入向量表示, 这是 RGCNdist2vec 模型的核心部分。路网被建模为带权无向图, 图信息主要保存在边集合上。而现有嵌入方法并没有充分考虑顶点的邻域信息, 不能为每个顶点学习更丰富的向量表示, 进而影响最终的估计准确性。而 GCN 在提取图空间特征方面具有独特优势, 通过卷积操作提取顶点和邻居之间的关联模型, 从而优化顶点的嵌入表达, 因此本文将其作为最短路径距离估计任务的嵌入方法, GCN 的消息传递计算方法为

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{2}$$

式中: $\tilde{\boldsymbol{A}}=\boldsymbol{A}_v+\boldsymbol{I}_{|V|}$是附加自连接的邻接矩阵, $\boldsymbol{I}_{|V|}$是大小为$|V|$的单位矩阵, $\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$为度矩阵, $\boldsymbol{W}^{(l)}$为第 l 层的可学习权重矩阵, $\boldsymbol{R}$ 为第 l 层的顶点特征矩阵, $\sigma(\cdot)$为 ReLU 激活函数。

由于路网的邻居顶点之间存在连接权值, 而 GCN 融合邻居和自身的顶点信息是无偏好的, 所以仅采用 GCN 从路网邻域关系的角度考虑并不合理。为此, 提出了 RGCN 模型, 主要由 2 部分组成: 一是使用路网带权邻接矩阵替换无权邻接矩阵, 以路网的权重信息为评分, 不同偏好的融合邻域信息; 二是选用邻域均值作为顶点的自连接性, 因为邻域均值不仅与邻居权重在同一个值域区间, 而且能够反映一组数据的集中水平。综上, RGCN 的消息传递规则计算方法为

$$\boldsymbol{H}^{(l+1)}=\sigma_1\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}^{\mathrm{r}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}_{\mathrm{r}}^{(l)}\right) \tag{3}$$

式中: $\tilde{\boldsymbol{A}}_{ij}^{\mathrm{r}}=\boldsymbol{A}_v^w+\mathrm{eye}\ (\mathrm{mean}(w_{v_i,v_j}))$ 为经过权重均值自连接后的邻接矩阵, $\mathrm{eye}(\mathrm{mean}(w_{v_i,v_j}))$, $w_{v_i,v_j}\in\boldsymbol{W}$ 表示邻居权重均值对角矩阵; $\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}^{\mathrm{r}}$为度矩阵; $\boldsymbol{W}_{\mathrm{r}}^{(l)}$ 表示第 l 层可学习的权重矩阵; $\sigma_1(\cdot)$为激活函数 LeakyReLU(·), 用来表示路网最短路径距离和嵌入向量之间的非线性关系。

随着消息传递阶数的增加, GCN 会存在过平

滑的问题。这是因为 GCN 可看作一个低通滤波器，其特性会使信号变得平滑，多次进行消息传递会使得信号趋于一致，导致顶点失去表示的多样性。考虑到 RGCN 的底层原理与 GCN 一致，所以 RGCN 采用 2 阶消息传递机制。

3.3　RGCNdist2vec 解码器

$$d_{\text{loss}} = \text{SmoothL1Loss}(\varphi(v_i,v_j),\hat{\varphi}(v_i,v_j)) =$$

$$\begin{cases} \dfrac{0.5}{N}\displaystyle\sum_{i=1}^{N}(\varphi(v_i,v_j)-\hat{\varphi}(v_i,v_j))^2, & |\varphi(v_i,v_j)-\hat{\varphi}(v_i,v_j)|<1 \\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(|\varphi(v_i,v_j)-\hat{\varphi}(v_i,v_j)|-0.5), & \text{其他} \end{cases} \tag{4}$$

解码器层首先接收编码器层的输出，即嵌入矩阵 $\boldsymbol{H}^{(2)}$ 和训练数据 $[(v_s,v_e),\varphi(v_s,v_e)]$。然后，通过计算式(1)，迭代更新权重矩阵来优化模型的表达能力。顶点嵌入向量度量方式的选取对模型的训练结果有很大影响，本文选用 L_1-norm 作为其度量方式，文献 [13] 给出了选择 L_1-norm 的原因以及相关证明。在训练过程中，采用 SmoothL1Loss 作为损失函数评估回归任务的精度，损失函数是 L1Loss 和 L2Loss 的结合，如式(4)所示，其同时拥有两者的优点，当预测值和真实值之间的绝对值之差小于 1 时，选用 L2Loss，此时有助于模型收敛；反之则选用 L1Loss，由于 L1Loss 对异常值不敏感，因此容易控制梯度的量级，从而达到更好的训练效果。实验采用 Adam 优化器来训练模型参数，该优化器可以在训练中自动更新学习率，使得模型的收敛速度更快。

3.4　数据采样方案

训练数据的选择对于学习路网中各顶点的嵌入向量表示至关重要，文献 [12] 选择图中所有顶点对(包含 $n(n-1)$ 个顶点对，n 为顶点个数)进行训练，文献 [14] 在首轮训练过程中选择所有顶点对，其余轮次根据地标选择 $l(n-l)$，l=(15~20)%n，其中 l 为地标顶点个数，n 为路网的顶点个数。如果训练数据量过大，训练数据构造和模型训练的时间都会变长。其次，如果在路网中进行随机采样，样本的选择概率和选择范围将是不确定的，这种方式也会因随机性而导致模型训练结果变差。

为了兼顾模型训练开销和估计准确性，设计了一种基于图逻辑分区的三阶段采样方法。该采样方式分为 3 个层次：子图内采样、子图间采样和全图内采样。

1)子图内采样方式：对路网顶点进行划分，得到顶点个数相对均匀的子图；如果计算属于 2 个不同子图的 2 点之间的距离，那么对应子图的边界点是必经之点。

证明 1　使用反证法。给定来自不同子图的 2 个顶点 $u\in B(G_1)$ 和 $v\in B(G_2)$，u 到 v 的最短路径为 $p = uv_1v_2\cdots v_xv$。假定 v_1，v_2，…，v_x 并不是 $B(G_2)$ 中的边界点。如果 $v_x\notin G_2$，那么 v 一定是边界点，这与假设相互矛盾。如果 $v_x\in G_2$，那么找到 G_2 中下标最大的点(如 v_i)。这样，$v_{i-1}\notin G_2$ ($v_0=u$)。基于边界点的定义，v_i 是一个边界点，这也与假设相矛盾。因此，引理得证。

图中顶点的度是用来描述图中心性的一项重要指标。因此，以子图中度值较大的点和子图的边界点为切入点对其进行采样。如果采取确定的图划分方法，则子图边界点是固定且已知的，将其称为边界点集合，用 Set(B_n) 表示子图 G_n 的边界点集合。然而，对于图中度值较大顶点的选择是不确定的。将度值较大的顶点集合简称为度集合，用 Set(D_n) 表示子图 G_n 的度集合。为了选取恰当的度集合，本文对多个真实路网数据集的顶点度值进行了正态性检验，发现顶点的度值存在尖峰分布现象，为保证采样数据的数量，将峰值对应的度值作为度集合采样的阈值。

子图的顶点集 Set(G_n) 被划分成 3 部分，即 Set(E_n)= Set(G_n) − Set(B_n) − Set(D_n)、Set(B_n) 和 Set(D_n)。首先，选择 $v_i\in$ Set(E_n) 和 $v_j\in$ Set(D_n)∪Set(B_n)；其次，再分别选择 $v_i\in$ Set(D_n) 和 $v_j\in$ Set(B_n)。最后，将所选数据进行拼接构成子图 G_n 采样结果。

假设子图 G_2 的度集合阈值为 3，图 5 为子图 G_2 的图内采样过程。

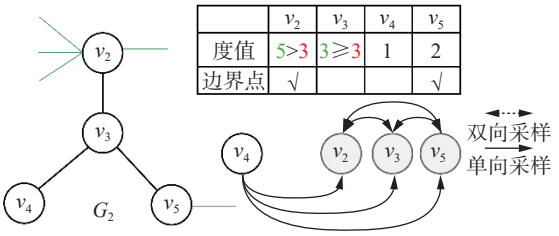


图 5　子图内采样过程

Fig. 5　Process of intra-subgraph sampling

2)子图间采样方式：对于每个子图边界点，选取其到除自身以外其他子图的边界点。图 6 给出了子图 G_2 的边界点 v_2 的采样过程。上述方法是针对子图和子图间的采样，目的是学习子图细粒度的结构信息。

3)全图内采样方式：全图内采样是针对整个图层次的采样，目的是学习图的全局信息(即粗粒度学习)。根据第 1 阶段和第 2 阶段可知，已对图中顶点的所属子图进行采样，对于其余子图，

每个顶点随机选取除自身子图以外其他子图顶点数的 10%~15%。

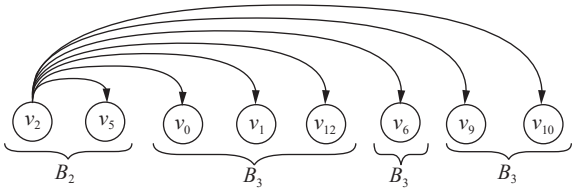


图 6　子图间采样过程

Fig. 6　Process of inter-subgraph samplings

4　实验及结果分析

4.1　数据集及实验环境

为了验证 RGCNdist2vec 模型的有效性, 本文在 4 个不同规模的路网数据集上对所提模型的效果和性能进行评估。在预处理阶段, 提取路网的最大连通分量输入到模型中。表 1 为所选路网的具体信息, 属性分别为顶点数量、边数量、顶点度值、平均权重。图 7 是 4 个路网数据集的可视化结果。

表 1　路网数据集

Table 1　Road network datasets

数据集	地区	顶点数量	边数量	度值
SU	Surat	2 508	3 591	7 182
DG	Dongguan	7 658	10 542	21 084
HA	Harbin	10 132	14 185	28 370
AH	Ahmedabad	12 747	18 117	36 243

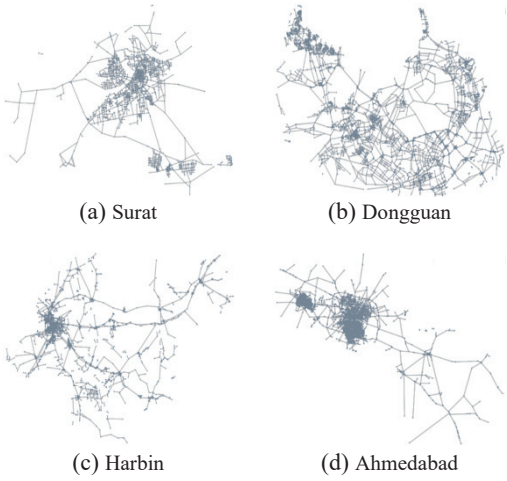


图 7　数据集的可视化结果

Fig. 7　Visualization results of the datasets

RGCNdist2vec 模型使用 Python 语言进行开发, 深度学习框架选择 PyTorch, 所有训练与测试实验的模型均运行在 2 张 RTX2080Ti、12 GB 显卡、操作系统为 Ubuntu 16.04、运行内存 32 GB 的主机上。

4.2　评价指标

本文采用平均绝对误差 E_{MA}(MAE)和平均相对误差 E_{RE}(MRE)衡量估计距离 $(\hat{\varphi}(v_i,v_j))$ 与真实距离 $(\varphi(v_i,v_j))$ 之间的误差, 误差越小则表示模型效果越好, MAE 和 MRE 的定义分别为

$$E_{\mathrm{MA}}=\frac{1}{N}\sum\left|\hat{\varphi}(v_i,v_j)-\varphi(v_i,v_j)\right| \tag{5}$$

$$E_{\mathrm{RE}}=\frac{1}{N}\sum\frac{\left|\hat{\varphi}(v_i,v_j)-\varphi(v_i,v_j)\right|}{\varphi(v_i,v_j)} \tag{6}$$

式中 N 表示验证数据的数量。训练时间(precomputation/training time, PT)和查询时间(query time, QT)用来衡量模型的训练与查询效率。

4.3　参数设置

对于本文提出的 RGCNdist2vec 模型, 在实验过程中采用 2 阶消息传递, 其共有 4 个超参数, 分别为 1 阶消息传递中的隐藏层维度 hid、模型的嵌入维度 d、学习率 l_r、训练轮次 epoch。为了获得更好的训练效果, 不同的路网设置不同的超参数值。模型最优情况下的参数值设置如下: 对于数据集 SU 和 DG, 设置隐藏层维度为 512, 嵌入维度为 32, 学习率为 0.05; 而对于数据集 HA 和 AH, 设置隐藏层维度为 1 024, 嵌入维度为 64, 学习率为 0.05。对于所有实验(除对比模型), 将三阶段采样方法获得的训练数据打乱后输入到网络, 验证数据由路网中随机选取的 10^6 个顶点对构成。路网顶点的初始特征采用 one-hot 编码构建。

4.4　消融实验

为了测试 RGCNdist2vec 模型中各个模块对模型整体估计准确度的影响, 通过选取最优嵌入维度及参数设置, 对采样方法和嵌入方法的组合选择进行了模型的自对比。

4.4.1　分区阈值 k_{f} 的设置策略与采样方法的影响

该实验的目的是通过对三阶段采样方法中的分区阈值 k_{f} 取不同值, 来观察分区阈值 k_{f} 如何影响训练数据质量, 进而影响模型的估计准确度。实验策略为, 调整分区阈值 k_{f} 在 [2,10] 范围内以 1 为步长进行变化, 在 SU 数据集和 DG 数据集上的开展实验, 实验结果如图 8 和图 9 所示。

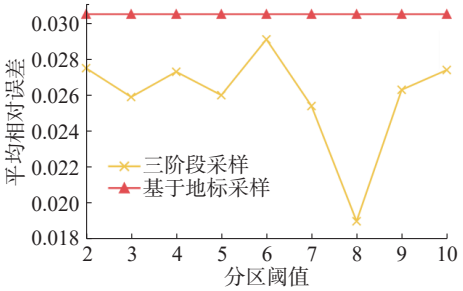


图 8　SU 数据集分区阈值变化下的效果对比

Fig. 8　Performance comparison under the change of k_{f} of SU dataset

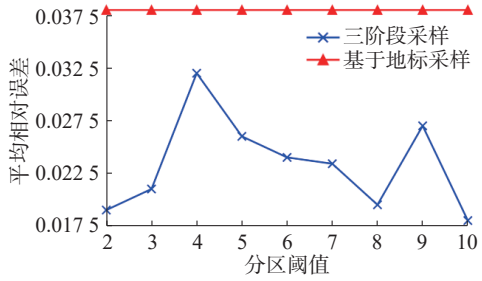


图 9　DG 数据集分区阈值变化下的效果对比

Fig. 9　Performance comparison under the change of k_f of DG dataset

从实验结果来看，不同路网结构的最优结果处的分区阈值 k_f 不同。为使模型的估计准确度最佳，本文采用一种自适应优化算法获得不同数据集的最优分区阈值 k_f[30]，该优化算法结合了模拟退火算法和粒子群优化算法，旨在为不同优化任务寻取最优解。表 2 为本文所用数据集的最优阈值。

表 2　不同数据集下的最优分区阈值

Table 2　Optimal partitioning thresholds for different datasets

数据集	最优分区阈值
SU	8
DG	10
HA	13
AH	15

图 8 和图 9 也显示了在 SU 数据集和 DG 数据集上基于地标采样与三阶段采样的对比结果。实验结果表明，在不同的分区阈值 k_f 下，三阶段采样方法的估计准确度均优于基于地标采样的估计结果。这是由于三阶段采样方法偏向于选取路网中重要的顶点对构成训练数据，相较于基于地标的采样对路网结构和权重的学习更有针对性。

4.4.2　消息传递阶数对模型估计准确度的影响

该实验目的是验证 RGCN 是否存在由于过深的消息传递阶数产生过平滑问题，在此分别对 RGCN 设置 2、3、4 阶消息传递，用来评估消息传递阶数对模型准确度的影响。实验结果如图 10。

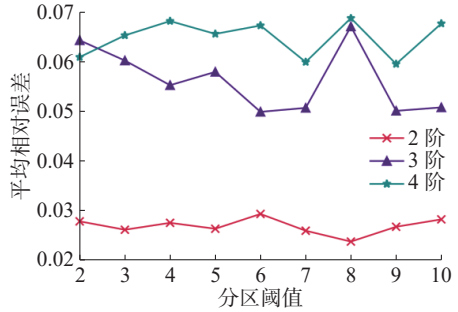


图 10　不同消息传递阶数的效果对比 (SU)

Fig. 10　Performance comparison of different messaging stages (SU)

从实验结果可以看出，在 2 阶消息传递变换到 3、4 阶消息传递的过程中，模型的平均相对误差出现了大幅度上升，由此可见，RGCN 和 GCN 相同，不适合过深阶的消息传递。

4.5　效果实验

为了评估 RGCNdist2vec 模型的估计准确度，使用以下几种主流模型进行对比实验。

1）ndist2vec[14]：该模型通过深度优先算法选取一定比例的顶点构成地标顶点，从地标顶点和其余顶点中采样获得训练数据，从而训练神经网络模型逼近两顶点间的最短路径距离。

2）vdist2vec-S[12]：该模型通过运用权重矩阵嵌入图顶点，然后再训练多层感知机逼近路网最短距离，模型的训练采用路网所有顶点对作为训练数据。

3）node2vec-Sg[11]：运用 node2vec[22] 或 Poincare[23] 作为嵌入方法，通过训练神经网络用于预测社交网络的最短距离。在此，将其用于路网场景以对比模型性能。

通过对上述基线方法的介绍可知，ndist2vec 和 vdist2vec-S 是基于多层感知机的估计计算方法，二者的区别主要在于采样方案的不同，前者根据地标采样，后者是进行全图采样。node2vec-Sg 是一种主要用于无权图的估计方法。上述方法均没有充分考虑路网的拓扑结构。而 RGCNdist2vec 分别在采样方案和嵌入方法 2 个层次进行了创新，对路网结构进行了细化分析并融合了权重信息，进一步提升了最短路径距离估计任务的估计效果和效率。

4.5.1　与基线算法在估计准确度方面的对比结果

通过设置最优的参数对比 4 个模型在 4 个真实路网数据集上的估计准确度。表 3 为 4 个模型在估计准确度方面的实验结果。图 11 为实验结果的可视化分析。

由对比结果可知，在 3 个不同规模的真实路网上（除了 HA），本文所提模型 RGCNdist2vec 的 MAE 和 MRE 均优于其他 3 个模型。这是由于其他模型的图嵌入方法采用基于随机游走 (node2vec) 或简单与权重矩阵相乘的方法，相较于本文所提的图嵌入方法，不能很好地学习路网的结构和权重信息。对于 HA 数据集来说，RGCNdist2vec 的准确率略低于 vdist2vec-S，但高于 ndist2vec 和 node2vec-Sg。通过分析 Harbin 数据集的可视化结果 (图 7)，发现 HA 相较于其他 3 个数据集分布稀疏且权重差距较大，会削弱模型融入邻居特征的能力。

表 3　不同模型的效果对比

Table 3　Comparison on effects of different models

模型		SU		DG		HA		AH	
		MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE
基线方法	ndist2vec	99.74	0.034	278.41	0.046	256.47	0.0548	146.32	0.033
	vdist2vec-S	87.22	0.028	144.32	0.030	**158.60**	**0.0351**	100.98	0.021
	node2vec-Sg	642.09	0.171	2421.34	0.203	3646.78	0.2300	3711.44	0.258
所提方法	RGCNdist2vec	**60.92**	**0.0227**	**127.65**	**0.0174**	179.85	0.0409	**98.36**	**0.0196**

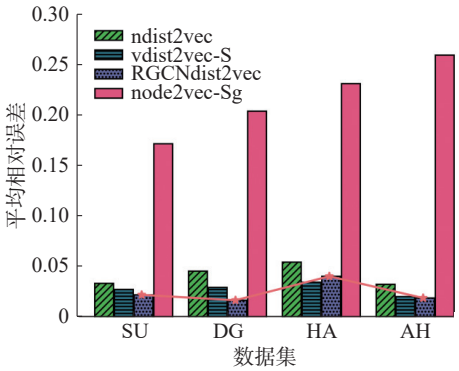


图 11　不同模型的效果对比

Fig. 11　Comparison on effects of different models

4.5.2　不同嵌入方法在估计准确度方面的对比结果

为了验证 GCN 的不同改进策略对模型估计准确度的影响, 本文针对图嵌入方法进行模型自对比(如图 12 所示)。在对 RGCN 改进策略的不同组合下, 进行效果对比实验。GCN、GCN-1 和 GCN-2 分别对应了 GCN、GCN 将自连接矩阵更改为邻域权重均值自连接矩阵和 GCN 将无权邻接矩阵更改为带权邻接矩阵。

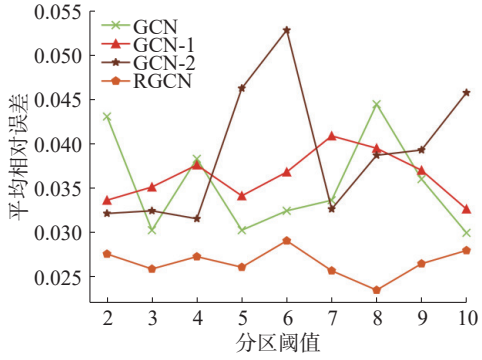


图 12　不同嵌入组合下的效果对比(SU)

Fig. 12　Performance comparison under different embedding combinations (SU)

从表 4 可以看出, 在 SU 数据集中, RGCN 模型相较于 3 个组合模型平均绝对百分比误差(mean absolute percentage error, MAPE)降低了约 1.2 百分点, 这说明 RGCN 的改进方案对估计结果起到了促进作用。MAPE 的定义为

$$E_{\mathrm{MAPE}}=\frac{1}{N}\sum\frac{\left|\hat{\varphi}(v_i,v_j)-\varphi(v_i,v_j)\right|}{\varphi(v_i,v_j)}\times 100\% \tag{7}$$

综上, 对 GCN 进行改进使其根据邻居权重融合邻域信息是有效的。

表 4　不同嵌入组合下的效果对比

Table 4　Performance comparison under different embedding combinations　%

模型	平方绝对百分比误差
GCN	3.55
GCN-1	3.65
GCN-2	3.92
RGCN	2.66

4.6　效率实验

表 5 为 RGCNdist2vec 模型与基线模型在效率方面的对比实验结果。通过对实验结果进行分析, 本文所提 RGCNdist2vec 模型最大的优势体现在 PT 指标上, 其在 4 个数据集上均优于对比模型。虽然基于嵌入的方法相较于传统的方法在训练时间方面已经有明显的提升, 但图神经网络的相关方法对比于该任务以往方法具有更快的训练速度。同时, RGCNdist2vec 模型在训练时并没有选取路网中所有顶点对作为训练数据, 这也会使训练时间大幅减少。对于查询时间 QT, 通过计算路网 $n(n-1)$ 对顶点对的查询均值来获得, 实验结果表明查询时间和路网嵌入维度成正比关系。对于数据集 SU 和 DG, 本文所提模型 RGCNdist2vec 设置 d=32, ndist2vec 设置 d=50, vdist2vec-S 和 node2vec-Sg 分别设置 d 为 0.02n 和 128。因 RGCNdist2vec 模型设置的维度均小于其余 3 个模型, 所以查询时间最短。而对于数据集 HA 和 AH, 由于 RGCNdist2vec 模型设 d=64, 其余模型的设置值不变, 则查询时间略高于 ndist2vec, 而低于 vdist2vec 和 node2vec-Sg。基于嵌入类的方法均可在线性时间复杂度内获得查询结果, 这也是该类方法的优点之一, 能够满足数百万查询同时发起时的及时反馈[31]。

表 5　不同模型的效率对比

Table 5　Efficiency comparison of different models

模型		SU		DG		HA		AH	
		PT/h	QT/μs	PT/h	QT/μs	PT/h	QT/μs	PT/h	QT/μs
基线方法	ndist2vec	0.10	7.79	0.48	16.95	1.20	**23.45**	2.50	**25.56**
	vdist2vec-S	0.41	8.11	2.40	17.12	3.7	37.48	12.00	40.37
	node2vec-Sg	0.35	8.02	1.37	18.06	7.45	36.64	9.74	35.69
所提方法	RGCNdist2vec	**93.399 s**	**7.52**	**0.14**	**15.47**	**0.43**	28.34	**1.22**	30.48

5　结束语

为了兼顾路网最短路径距离估计任务的估计准确性和训练时间，本文通过分析图神经网络的相关方法，提出了路网最短路径距离估计编码器−解码器框架。其次，改进图卷积神经网络将其运用到路网场景上，作为路网最短路径距离估计的嵌入方法，捕获顶点的邻域关系。同时，为了进一步提高训练效率，本文对路网分区处理，设计了一种基于图逻辑分区的三阶段采样方法，用于选取优质训练样本。最后，在 4 个真实的数据集上开展实验，表明所提方法在估计准确度和训练效率上均有显著的效果。在路网场景中，有向的距离计算、路网顶点嵌入的可扩展性以及路网最短路径恢复工作也非常的重要和有意义，在接下来的研究中，将进一步开展上述研究工作。

参考文献：

[1] DIJKSTRA E W. A note on two problems in connexion with graphs[J]. Numerische mathematik, 1959, 1(1): 269−271.

[2] FLOYD R W. Algorithm 97: Shortest path[J]. ACM, 1962, 5(6): 345.

[3] GEISBERGER R, SANDERS P, SCHULTES D, et al. Contraction hierarchies: faster and simpler hierarchical routing in road networks[C]//International Workshop on Experimental and Efficient Algorithms. Berlin: Springer, 2008: 319−333.

[4] OUYANG Dian, QIN Lu, CHANG Lijun, et al. When hierarchy meets 2-hop-labeling: efficient shortest distance queries on road networks[C]//Proceedings of the 2018 International Conference on Management of Data. New York: ACM, 2018: 709−724.

[5] GEISBERGER R, SCHIEFERDECKER D. Heuristic contraction hierarchies with approximation guarantee [C]// Proceedings of the Third Annual Symposium on Combinatorial Search. Menlo Park: AAAI, 2010: 31−38.

[6] CHECHIK S. Approximate distance oracles with improved bounds[C]//Proceedings of the forty-seventh annual ACM on symposium on Theory of Computing. Portland: ACM, 2015: 1−10.

[7] TANG Liying, CROVELLA M. Virtual landmarks for the internet[C]//Proceedings of the 2003 ACM SIGCOMM Internet Measurement Conference. Miami Beach: ACM, 2003: 143−152.

[8] ZHAO Xiaohan, ZHENG Haitao. Orion: shortest path estimation for large social graphs[C]//3rd Workshop on Online Social Networks. Boston: USENIX, 2010: 9.

[9] ZHAO Xiaohan, SALA A, ZHENG Haitao, et al. Efficient shortest paths on massive social graphs[C]//Proceedings of the 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing. Orlando: IEEE, 2011: 77−86.

[10] GUBICHEV A, BEDATHUR S, SEUFERT S, et al. Fast and accurate estimation of shortest paths in large graphs [C]//Proceedings of the 19th ACM international conference on Information and knowledge management. Toronto: ACM, 2010: 499−508.

[11] QI J, WANG W, ZHANG R, et al. A learning based approach to predict shortest-path distances[C]//Proceedings of the 23rd International Conference on Extending Database Technology. Copenhagen: OpenProceedings, 2020: 367−370.

[12] RIZI F S, SCHLOETTERER J, GRANITZER M. Shortest path distance approximation using deep learning techniques[C]//2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Barcelona: IEEE, 2018: 1007−1014.

[13] HUANG Shuai, WANG Yong, ZHAO Tianyu, et al. A learning-based method for computing shortest path distances on road networks[C]//2021 IEEE 37th International Conference on Data Engineering. Chania: IEEE, 2021: 360−371.

[14] CHEN Xu, WANG Shaohua, LI Huilai, et al. Ndist2vec: node with landmark and new distance to vector method for predicting shortest path distance along road networks[J]. ISPRS international journal of geo-information, 2022,

11(10): 514.

[15] DARMOCHWAL A. The euclidean space[J]. Formalized mathematics, 1991, 2(4): 599−603.

[16] KLEINBERG R. Geographic routing using hyperbolic space[C]//IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications. Anchorage: IEEE, 2007: 1902−1909.

[17] ANIRBAN S, WANG Junhu, ISLAM M S. Experimental evaluation of indexing techniques for shortest distance queries on road networks[C]//2023 IEEE 39th International Conference on Data Engineering. Anaheim: IEEE, 2023: 624−636.

[18] PACINI F, GUNBY-MANN A, COHEN S, et al. ANEDA: adaptable node embeddings for shortest path distance approximation[C]//2023 IEEE High Performance Extreme Computing Conference. Boston: IEEE, 2023: 1−7.

[19] LUO Jiaqi, DAI Baisheng, CHANG Penghao, et al. Determination of rice leaf midrib deflection in field environment by using semantic segmentation and shortest distance algorithm[J]. Computers and electronics in agriculture, 2023, 215: 108326.

[20] LEI Yi, SHAO Hu, WU Ting, et al. An accelerating algorithm for maximum shortest path interdiction problem by upgrading edges on trees under unit Hamming distance[J]. Optimization letters, 2023, 17(2): 453−469.

[21] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of the 5th International Conference on Learning Representations. France: ICLR, 2017: 1−14.

[22] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[EB/OL]. (2016−07−03)[2021−01−01]. https://arxiv.org/abs/1607.00653.

[23] NICKEL M, KIELA D. Poincaré embeddings for learning hierarchical representations[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. New York: Curran Associates, 2017: 6338−6347.

[24] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//Proceedings of the International Conference on Learning Representations. Vancouver: ICLR, 2018: 566−577.

[25] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum chemistry[C]// Proceedings of the 34th International Conference on Machine Learning. New York: PMLR, 2017: 1263−1272.

[26] ZHOU Jie, CUI Ganqu, HU Shengding, et al. Graph neural networks: a review of methods and applications[J]. AI open, 2020, 1: 57−81.

[27] WU Zonghan, PAN Shirui, CHEN Fengwen, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2021, 32(1): 4−24.

[28] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. New York: Curran Associates, 2017: 1024−1034.

[29] KARYPIS G, KUMAR V. Analysis of multilevel graph partitioning[C]//Proceedings of the 1995 ACM/IEEE conference on Supercomputing. San Diego: ACM, 1995: 29.

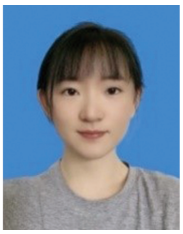
[30] 闫群民, 马瑞卿, 马永翔, 等. 一种自适应模拟退火粒子群优化算法[J]. 西安电子科技大学学报, 2021, 48(4): 120−127.
YAN Qunmin, MA Ruiqing, MA Yongxiang, et al. Adaptive simulated annealing particle swarm optimization algorithm[J]. Journal of Xidian University, 2021, 48(4): 120−127.

[31] 孟祥福, 赖贞祥, 崔江燕. 集合空间关键字内聚组查询方法[J]. 智能系统学报, 2024, 19(3): 707−718.
MENG Xiangfu, LAI Zhenxiang, CUI Jiangyan. Cohesive group query approach for collective spatial keywords[J]. CAAI transactions on intelligent systems, 2024, 19(3): 707−718.

作者简介:

孟祥福, 教授, 博士生导师, 主要研究方向为 top-k 查询、时空大数据。主持国家自然科学基金项目 2 项、辽宁省各类基金项目 4 项, 发表学术论文 20 余篇, 出版学术专著 2 部。E-mail: marxi@126.com。

崔江燕, 硕士研究生, 主要研究方向为图嵌入、最短路径距离计算。E-mail: 1315249764@qq.com。



邓敏超, 硕士研究生, 主要研究方向为异构图嵌入、社交网络分析。E-mail: 1093523593@qq.com。