

面向异质信息网络的表示学习方法研究综述

周 慧, 赵中英⁺, 李 超

山东科技大学 计算机科学与工程学院, 山东 青岛 266590

⁺ 通讯作者 E-mail: zzysuin@163.com

摘 要: 网络表示学习旨在为网络中的组件(节点、边、子网络)学习出低维的表征向量,使得这些向量能够在最大程度上保留组件在原网络中的特性。异质信息网络是由多种类型的节点、链接关系以及属性信息组成的网络,具有动态性、大规模和异质性等特点,在我们现实生活中普遍存在。融合多种异质信息进行网络表示学习,能在一定程度上解决数据稀疏问题,同时,有助于训练出具有高区别力和推理能力的表征向量。但与此同时,也面临着如何有效处理复杂数据关系以及平衡异质信息的挑战。近年来,研究者们针对异质信息网络设计了不同的表示学习算法,在很大程度上推动了该领域的发展。针对这些算法,首先设计一个统一的分类框架,接着对各类别下的代表性算法进行概括介绍和比较,分析它们的时间复杂度和优缺点。此外,分类汇总了实验中的常用数据集。最后给出了该领域的挑战和未来可能的研究方向。

关键词: 网络表示学习; 异质信息网络; 网络分析

文献标志码: A **中图分类号:** TP18

周慧, 赵中英, 李超. 面向异质信息网络的表示学习方法研究综述[J]. 计算机科学与探索

ZHOU Hui, ZHAO Zhongying, LI Chao. A Survey on Representation Learning Methods oriented to Heterogeneous Information Network[J]. Journal of Frontiers of Computer Science and Technology

A Survey on Representation Learning Methods oriented to Heterogeneous Information Network^{*}

ZHOU Hui, ZHAO Zhongying⁺, LI Chao

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China

Abstract: Network representation learning aims to learn a series of low-dimensional vectors for the components (node, edge, subgraph, etc.) in a network. Meanwhile, the characters of the components in the original network should be largely retained in these vectors. Heterogeneous information network is the network composed of various types of nodes, link relationships and attribute information. It is characterized by dynamics, large scale and heterogeneity, and is ubiquitous in our real life. Compared with purely relying on

^{*}The National Natural Science Foundation of China under Grant Nos. 61303167, 61433012 (国家自然科学基金); the Ministry of Education in China Foundation for Humanities and Social Sciences under Grant No. 17YJCZH262 (教育部人文社会科学研究项目); the Natural Science Foundation of Shandong Province under Grant No. ZR2018BF013 (山东省自然科学基金); the Innovative Research Foundation of Qingdao under Grant No. 18-2 -2-41-jch (青岛市源头创新计划应用基础研究项目).

network topology information, network representation learning by integrating various heterogeneous information can not only alleviate the problem of data sparsity, but also help to learn the representation vectors with high discriminative and inferential ability. At the same time, it also faces the challenge of dealing with complex data relationships and balancing heterogeneous information. In recent years, researchers have designed different representation learning algorithms for heterogeneous information networks, which have greatly promoted the development of this field. In view of these algorithms, this paper first designs a unified classification framework, then generalizes and compares the representative algorithms in each category, including their time complexities, advantages, etc. In addition, information of the commonly used data sets is summarized into a table. Some challenges and possible researching directions are provided at the end of this paper.

Key words: network representation learning; heterogeneous information network; network analysis

1 引言

网络是表达事物之间关联关系的有效载体，由节点和链接关系（边）组成，在我们日常生活中无处不在。异质信息网络是一种特殊类型的网络，由多种类型的节点、链接关系以及属性信息组成，具有大规模、异质性等特点。随着信息时代的发展，我们所面临的信息网络越来越复杂，各行各业对数据处理的速度和有效性也提出了更高的需求。采用邻接矩阵这种高维稀疏的编码方式来表示网络中的节点，很难被机器学习算法处理。网络表示学习（或网络嵌入）采用低维向量的形式表示网络中的组件，打破了网络固有的节点和边的形态，同时最大程度地保留原网络中的结构信息和特性。由于低维的向量很容易被机器学习算法处理，因而越来越受到学术界和产业界的广泛关注，也有效地运用到了节点分类^{[1][2]}、链接预测^{[3][4]}、社区发现^{[5][6]}和推荐^{[7][8]}等任务中。最初的网络表示学习算法注重对原网络的复现，即最大程度地保留原网络中的信息。相关研究工作包括：DeepWalk^[9]、Large-scale Information Network Embedding (LINE)^[10]、node2vec^[11]和 GraRep^[12]等。虽然这些算法越来越

精确地保留了节点在原网络中的相对位置关系，但它们只是单纯地借助于拓扑结构信息，学得的表现向量缺乏对后续实验任务的区分力和推理能力。现实世界的网络中普遍存在着除拓扑信息之外的异质信息，包括：节点或边上的标签信息、社区信息和属性信息等。充分利用这些异质信息有助于学得更具推理能力和区分力的表征向量。

融合异质信息的网络表示学习工作相继被提出，在很大程度上推动了该领域的发展。对这些算法进行分类梳理有助于了解和掌握该领域的学术思路和动态。在本文中，我们首先创建了一个统一的分类框架，并在每个类别下列举了一些代表性算法。进一步地，我们从横向上对各类别下的代表性算法进行了介绍，包括其主要的理论方法和思想。之后，我们又纵向对比了这些算法的时间复杂度、优缺点和评估实验等。此外，我们整理了实验中一些常用的数据集，并给出了简单的属性介绍和相关链接。在文章最后，我们进一步指出了该领域的挑战和未来可能的研究方向。

1.1 相关定义

1.1.1 异质信息网络

根据石川等人在论文[13]中提出的异质信息网络的定义, 我们进一步对本文所涉及的异质信息网络给出以下形式化的定义。

定义 1(异质信息网络(Heterogeneous Information Network, HIN))用 $G(V, E, T, \phi, \psi, H)$ 表示一个异质信息网络, 其中的每个节点 $v \in V$ 和每条边 $e \in E$ 都对应一种类型映射关系, 分别表示为 $\phi: V \rightarrow T_V$, $\psi: E \rightarrow T_E$, $T = T_V + T_E$; H 表示网络中辅助信息构成的集合, 共包含 W 种信息类型, $H = \bigcup_{i=1}^W X_i$; 此外, 网络需要满足 $|T_V| + |T_E| > 2$ 或 $H \neq \emptyset$ 。

1.1.2 网络表示学习

定义 2(网络表示学习^[14](Network Representation Learning, NRL))给定一网络 $G(V, E)$, 其中 V 表示网络 G 中的节点集, E 表示网络 G 中的边集。目标是为网络中的节点 $v \in V$ (或边、子图等) 学习一映射关系 $f: v \rightarrow r_v \in R^d$, 其中 r_v 是为节点 v 学得的低维稠密向量, $d \ll |V|$, 转换函数 f 用于捕获定义在原网络中的相似度信息。

结合定义 1, 我们进一步给出异质信息网络上的表示学习的形式化定义, 如下所示。

定义 3(异质信息网络表示学习) 给定一异质信息网络 $G(V, E, T, \phi, \psi, H)$, 目标是为网络中的节点 $v \in V$ (或边、子图等) 学习一映射关系 $f: v \rightarrow r_v \in R^d$, $d \ll |V|$, r_v 是为节点 v 学得的低维稠密向量, 其中包含了节点在原网络中的异质信息。

1.2 符号和解释

我们进一步总结了论文中用到的主要符号及其

解释, 如表 1 所示。

Table 1 The notations used in this paper.

表 1 论文中的常用符号

Notations [◊]	Descriptions [◊]
G°	The original network for embedding [◊]
V°	Set of nodes in network G°
$ V ^{\circ}$	Number of nodes [◊]
E°	Set of edges in network G°
$ E ^{\circ}$	Number of edges [◊]
$ Y ^{\circ}$	Number of labels [◊]
T°	Object and link type set [◊]
A°	Adjacent matrix [◊]
d°	Dimension of the representational vector [◊]
k°	Number of community categories [◊]
c°	Number of convolution layers [◊]
m°	Dimension of attribute information vector [◊]
Q°	Number of negative samples [◊]
N_A°	Number of nonzero elements in matrix A°
t°	Number of workers working in parallel [◊]

2 算法分类介绍

针对融合异质信息的网络表示学习算法, 我们首先设计了一个统一的分类框架, 并在每个类别下列举了一些代表性算法。如图 1 所示, 该框架首先以信息类型作为分类依据将算法分成了四大类, 即: 标签/社区信息、属性信息、多种类型信息以及异构网络, 其中前三类面向的是带有辅助信息的网络, 第四类面向的是异构网络。接下来, 在每个类别下进一步划分成有监督和无监督这两个子类, 并列举了相应的代表性算法。此外, 框图的最下方是这些算法用到的基础理论方法, 包括矩阵分解、神经网络、自定义损失函数以及其它的综合框架, 我们采用不同的颜色进行标注, 并与相应算法的颜色相对应。在下文中, 我们将对一些代表性算法进行分类介绍。

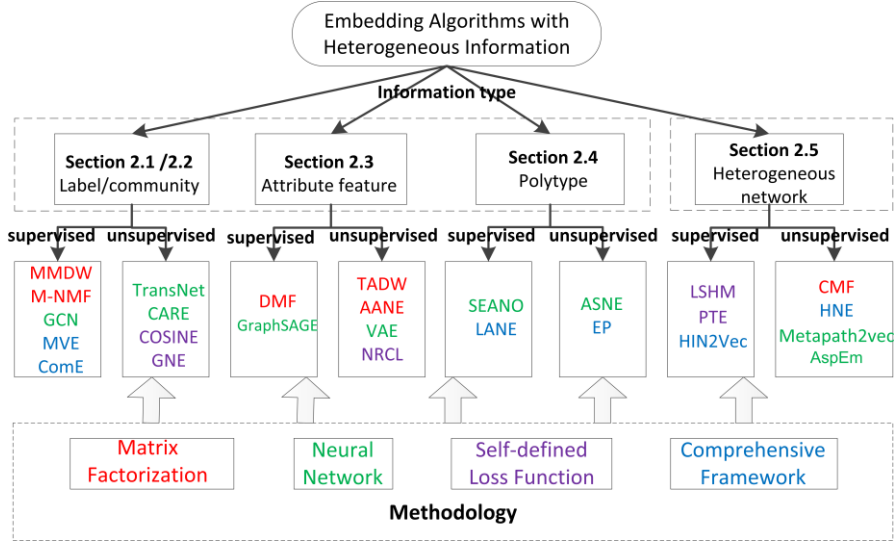


Fig.1 The classification framework for the algorithms (color of the algorithm is corresponding to the methodology).

图 1 算法分类框架（算法的颜色与理论相对应）

2.1 标签信息辅助的网络表示学习

2.1.1 基于矩阵分解的方法

Max-margin deepwalk (MMDW)^[15]是基于矩阵分解的有监督的表示学习模型。该模型首先从理论上证明了 DeepWalk 相当于矩阵分解；然后将矩阵分解得到的表征向量输入到支持向量机^[16](Support Vector Machine, SVM)中进行训练,并根据分类边界产生有偏置的梯度方向；之后根据该方向进行向量更新,以增大表示向量在分类器中的分类间隔。总的来说,该模型共同优化最大间隔分类器和目标矩阵分解模型,使得学得表征向量具有对后续机器学习任务的区分力。

与之类似的算法还包括 Discriminative Deep Random Walk (DDRW)^[17]、TLIN^[18]以及 Semi-Supervised Network Embedding (SemNE)^[19]等。这些算法共同的优点是能够直接优化分类损失,使学得的表征向量具有区分力,但这些算法容易产生过拟合。

2.1.2 基于神经网络的方法

Thomas 等人提出了 Graph Convolution Networks (GCN)^[20]算法,运用卷积神经网络来处理图结构数据。该算法首先将图的拉普拉斯矩阵的特征向量矩阵通过傅立叶变换转化到谱空间上,然后进行卷积操作。然后通过部分带标签的节点对模型进行训练,从而实现半监督学习。此外,作者还设计了一种简单有效的层级传播规则,通过这种方式可快速有效地处理图数据上的半监督分类问题。特别地,在不借助任何外部信息的情况下,该模型所获得的图节点向量表示和 DeepWalk 算法结果是极为相似的,这也表明了图卷积神经网络应用在网络表示学习任务中的有效性。更多关于图神经网络的内容可参阅文献[21]。

TransNet^[22]将自编码器模型与转换机制^{[23][24]}相结合,对网络中边上的标签信息进行建模。主要方法是对边上的标签信息进行自编码,并对自编码器中间层上的边向量进行转换操作,将节点

的向量表示和对应的边向量训练成平移关系。由此, 可根据节点向量直接推出边向量, 通过解码操作即可获得边上的标签信息。该模型被成功运用在社会关系提取任务中。

2.1.3 综合框架方法

在现实网络中, 节点之间存在着不同类型的相似性, 使得网络呈现出多种视图(Multi-View)。传统网络表示学习模型大都是学习网络中的单一视图

信息, 唐建等人提出了一种可整合网络中多视图信息的鲁棒性表示学习模型 MVE^[25]。该模型基于注意力机制学习各视图的权重并进行信息整合, 最终学得具有高鲁棒性的表征向量。

我们进一步将本节中介绍的算法进行综合比较, 包括时间复杂度、理论基础、有无监督以及评估方法等, 对比结果如表 2 所示。

Table 2 The comprehensive comparison on the algorithms assisted by label information

表 2 标签信息辅助的算法综合比较

Category	Algorithms	Methodology	Supervised	Time complexity	Evaluating methods
Label information	MMDW ^[15]	Matrix factorization	Yes	$O(V ^2)$	Node classification Visualization
	DDRW ^[17]	Deep neural network	Yes	$O(V \lg V d)$	Node classification
	GCN ^[20]	Graph convolution neural network	Yes	$O(E d^2)$	Node classification
	TransNet ^[22]	Deep neural network	No	$O(E d)$	Social relation extraction
	MVE ^[25]	Comprehensive framework	Yes	$O(E Q d)$	Node classification Link prediction

2.2 社区信息辅助的网络表示学习

2.2.1 基于矩阵分解的方法

Modularized Nonnegative Matrix Factorization (M-NMF)^[26]算法对网络中节点分别从微观和中观的角度学习向量表示, 同时考虑节点之间的相似度和节点所属的社区信息。微观层面上, 该算法主要编码节点间的一阶和二阶拓扑相似度; 中观层面上, 该模型运用基于模块性最大化的社区发现方法来建模网络中的社区结构。最终学得的表征向量同时包含了节点在原网络中的拓扑信息以及所属的社区信息, 在后续节点聚类和分类任务中表现优越。

2.2.2 综合框架方法

除了利用社区信息对网络中的单个节点进行

表示学习之外, Sandro 等人设计了 Community Embedding (ComE)^[27]模型, 可直接对整个社区进行表示学习。模型中节点嵌入、社区发现以及社区嵌入这三部分之间循环进行, 并且相互促进。节点嵌入有助于增强社区发现, 用于输出更好的社区, 由此能够进一步拟合出更好的社区嵌入。实验证明这对社区层面的任务有很大的帮助, 包括社区发现、社区可视化和社区推荐等。结合社区信息进行表示学习的算法还包括 COSINE^[28]、CARE^[29]和 Galaxy Network Embedding (GNE)^[30]等无监督学习算法。

我们进一步将本节中介绍的算法进行综合比较, 总结如表 3 所示。

Table 3 The comprehensive comparison on the algorithms assisted by community information

表 3 社区信息辅助的算法综合比较

Category	Algorithms	Methodology	Supervised	Time complexity	Evaluating methods
Community information	M-NMF ^[26]	Matrix factorization	Yes	$O(V ^2 d + V ^2 k)$	Node classification
					Community discovery
	ComE ^[27]	Comprehensive framework	Yes	$O(V k + E)$	Node classification
					Community discovery
					Visualization

2.3 属性信息辅助的网络表示学习

2.3.1 基于矩阵分解的方法

刘知远等人在 DeepWalk 算法基础之上进行改进, 提出了 Text-Associated DeepWalk (TADW)^[31] 算法。该算法将节点的属性信息加入到矩阵分解的过程中, 主要思想如图 2 所示。其目标是对矩阵 M ($M = (A + A^2)/2$) 进行矩阵分解, 使得 $M \approx W^T H T$, 其中 T 矩阵中包含了节点的属性信息。对应的目标函数如式(1)所示, 最小化目标函数并更新矩阵 W 和 H 。节点最终的向量表示由矩阵 W 和 $H T$ 对应的列向量拼接得到的。DMF^[32]对 TADW 算法进一步作了改进, 通过增加一线性分类器进行有监督学习, 从而使学得表征向量更具区别力。

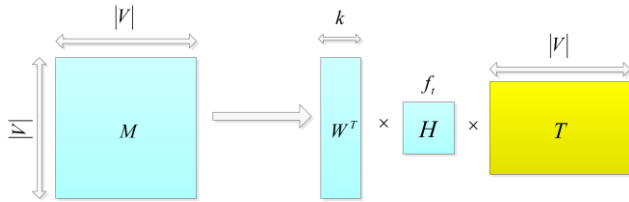


Fig.2 The framework of TADW model (from [31]).

图 2 TADW 模型框架(来源于文献[31])

$$\min_{W, H} \sum_{(i, j) \in E} (M_{ij} - (X^T W^T H Y)_{ij})^2 + \frac{\lambda}{2} (\|W\|_F^2 + \|H\|_F^2) \quad (1)$$

Accelerated Attributed Network Embedding (AANE)^[33]模型同样基于矩阵分解将网络中的拓扑信息以及属性信息整合到一起。特别地, 该模型将优化过程分解成了多个子问题进行并行工作, 通过这种方法大大提升了模型的工作效率。Li 等人引入矩阵摄动理论, 进一步设计了能够适应网络动态变化的表示学习模型 Dynamic Attributed Network

Embedding (DANE)^[34], 在网络拓扑结构和属性信息产生变化后, 该模型能够在较短时间内对网络中的节点生成新的向量表示。

2.3.2 基于神经网络的方法

李航等人提出了一深度学习模型 Variational Autoencoder (VAE)^[35]来学习网络中节点的嵌入。该模型整合了 doc2vec^[36]和深度自编码器, 将属性信息和拓扑信息进行整合并映射到同一语义空间中。其中 doc2vec 用于将属性信息向量化, 深度自编码器用于整合向量化的属性信息和拓扑信息(邻接矩阵)。特别地, 在编码区的最后一层会产生两种信息新的分布, 以提取两者的主要特征。最终, 模型将自编码器的中间层作为节点的向量表示。学得的表征向量同时捕获了节点在原网络中的属性信息以及高度非线性的拓扑信息。

GraphSAGE^[37]在传统图卷积神经网络上做了延伸, 通过汇聚多阶邻居节点的信息来生成向量表示。特别地, 该模型对于不同阶层的邻居节点训练不同的聚合器, 并可通过后续的任务进行有监督的学习。最终学得的节点的向量表示一方面汇聚了多阶邻居节点的属性信息, 另一方面也捕获了原网络中的拓扑信息。Embedding Propagation (EP)^[38]算法同样是通过汇聚邻居信息来进行表示学习。不同之处在于, EP 算法属于无监督学习, 且独立学习每种不同类型标签的嵌入。

涂存超等人指出现实网络中节点与不同邻居节点交互时所展现的主题会有所不同, 进而设计了 Context-Aware Network Embedding (CANE)^[39]算法。该算法通过一深度架构将节点的属性信息和拓扑

信息映射到同一语义空间中，并引入注意力机制来捕获节点与不同邻居交互时所侧重的主题。

2.3.3 基于自定义损失函数的方法

许林川等人认为现实世界网络中的链接信息和属性信息大都是部分可见的，由此提出了 Noise-resilient Representation Consensus Learning (NRCL)^[40]算法为具有部分链接和属性信息的网络学习向量表示，并应用到链接预测任务中。该模型

将网络中的节点分为三类，分别为具有连边的节点集、具有属性信息的节点集以及同时具有这两种信息的节点集，网络中的节点属于这三类中的一类或多类。对前两类节点集分别进行建模，然后通过第三类节点集相联系，由此使得链接信息与属性信息相互补充，学得向量表示具有高鲁棒性。

我们进一步将本节中介绍的算法进行综合比较，总结如表 4 所示。

Table 4 The comprehensive comparison on the algorithms assisted by attribute information
表 4 属性信息辅助的算法综合比较

Category	Algorithms	Methodology	Supervised	Time complexity	Evaluating methods
Attribute information	TADW ^[31]	Matrix factorization	No	$O(V ^2)$	Node classification
	AANE ^[33]	Matrix factorization	No	$O(V N_A + V ^2/t)$	Node classification
	VAE ^[35]	Deep neural network	No	$O(V d + V \lg V)$	Node classification
	GraphSAGE ^[37]	Graph convolution neural network	Yes	$O(V c)$	Node classification
	NRCL ^[40]	Self-defined loss function	No	$O(E Q)$	Link prediction

2.4 多类别信息辅助的网络表示学习

以上介绍的表示学习模型都只结合了一种异质信息，并不适用于多种异质信息的有效融合。而如何有效地融合和平衡多种异质信息是一个极具挑战的任务。已有一些研究者针对此问题各自提出了不同的解决方法。在下文中，我们将对这些研究成果分别进行详细介绍。

2.4.1 基于神经网络的方法

Attributed Social Network Embedding (ASNE)^[41]运用深层神经网络对不同类型的节点信息（离散+连续）进行综合的学习，模型的整体架构如图 3 所示。首先，该模型对节点的离散属性（例如：id）进行 one-hot 编码生成 W^{id} ，对于连续属性（例如：文本）进行 Term Frequency-Inverse Document Frequency (TF-IDF)编码生成 W^{att} ；然后，将这两类向量输入

到嵌入层（由两个全连接层组成）进行降维；之后将降维后的向量进行拼接，并输入到隐含层中进行非线性映射；最后，在输出层上计算出目标节点与其它节点连接的概率，由此将拓扑信息与多种类别的属性信息联系起来。总的来说，该模型通过深层神经网络建模不同类型信息之间的复杂的关联关系，在后续的节点分类和链接预测的任务中表现优越。

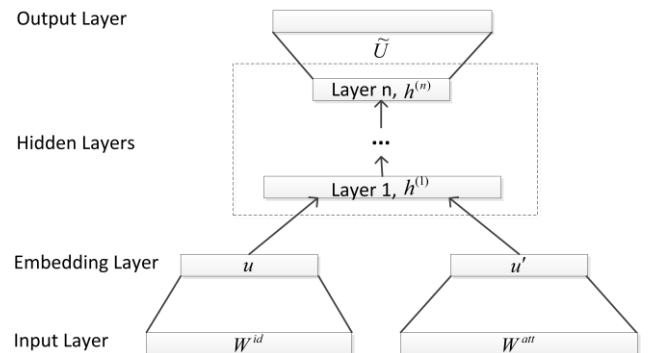


Fig.3 The framework of ASNE model (from [41]).

图 3 ASNE 模型框架(来源于文献[41])

Jacobs 等人提出了一种半监督模型 SEANO^[42] 为具有部分标签和属性信息的网络学习节点的向量表示。该模型通过深度学习架构将节点的拓扑结构信息、属性信息和标签信息关联起来。模型的整体框架如图 4 所示，它包含了两个输入层和两个输出层，中间通过非线性映射层将异构信息相连。其中，输入层中包含了目标节点和其邻居节点的属性信息向量；左侧的输出层作为该模型的有监督学习部分，通过带有标签信息的节点进行训练；右侧的输出层作为模型的无监督学习部分用于捕获节点在原网络中的拓扑信息。最后，将模型嵌入层中的向量作为节点最终的向量表示。

2.4.2 综合框架方法

Huang 等人提出了一种半监督的综合框架 Label informed Attributed Network Embedding (LANE)^[43] 来汇聚多种异构信息。该模型通过将标签信息、属性信息和拓扑信息映射到相同的语义空间中，并寻找三者之间的关联关系来获得节点的向量表示。其中，对于拓扑信息和属性信息的嵌入主要

运用了谱方法；对于标签信息的嵌入，主要根据同质性原则，使具有相同标签的节点在向量空间中距离相近。最后通过相关投影将学得的这三类向量嵌入到一个新的维度空间中，最大化三者在新空间上的关联性，进而得出最终的向量表示。

我们将本节介绍的算法及带有辅助信息的网络表示学习方法进行综合对比，从算法的理论基础、有无监督、时间复杂性、评测任务等方面进行比较，分别如表 5 和表 6 所示。

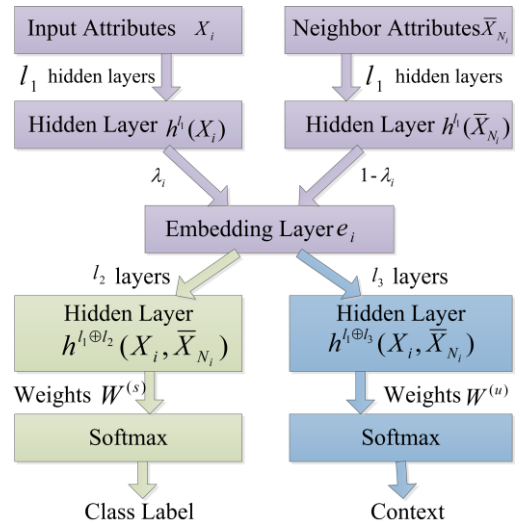


Fig.4 The framework of SEANO model (from [42]).

图 4 SEANO 模型框架(来源于文献[42])

Table 5 The comprehensive comparison on the algorithms assisted by polytype information

表 5 多类型信息辅助的算法综合比较

Category	Algorithms	Methodology	Supervised	Time complexity	Evaluating methods
	LANE ^[43]	Comprehensive framework	Yes	$O(V ^2 d + V ^2 m)$	Node classification
Polytype information	ASNE ^[41]	Deep neural network	No	$O(V \lg V d)$	Node classification Link prediction
	SEANO ^[42]	Deep neural network	Yes	$O(V Q d)$	Node classification

Table 6 A comparison of NRL algorithms in terms of theoretical foundation.

表 6 网络表示学习算法的理论基础比较

Methodology	Algorithms	Advantage	Disadvantage
Matrix Factorization	MMDW ^[15] , M-NMF ^[26] , TADW ^[31] , DMF ^[32] , AANE ^[33]	Capture high order node similarity	High time and space complexity
Neural Network	DDRW ^[17] , SemiNE ^[19] , GCN ^[20] , TransNet ^[22] , CARE ^[29] , VAE ^[35] , GraphSAGE ^[37] , ASNE ^[41] , SEANO ^[42]	Capture nonlinear relation between nodes	High computational cost in optimization process
Self-defined Loss Function	TLNE ^[18] , MVE ^[25] , COSINE ^[28] , GNE ^[30] , NRCL ^[40]	Relatively efficient	Only local structure information can be captured
Comprehensive Framework	ComE ^[27] , EP ^[38] , LANE ^[43]	High robust and discriminate ability	High time, space complexity and computational cost

2.5 面向异构网络的表示学习

2.5.1 基于自定义损失函数的方法

Jacob 等人在论文[44]中提出了一种半监督表示学习算法 LSHM，用于对异构网络中的不同类型的节点学习向量表示。该算法同时对相同类型和不同类型的节点进行平滑性约束，并将所有类型的节点映射到同一潜在空间中。此外，对于每类节点，都会进一步学习一分类函数，并对带标签的节点进行预测。最终学得向量表示同时包含了节点在原网络中的拓扑信息和标签信息。

Predictive Text Embedding (PTE)^[45] 算法在 LINE 算法的基础上做了改进，用于在异构文本网络中学习文本的嵌入。异构文本网络中包含了单词、文档和标签这三种类型的节点。在建模过程中，作者首先将异构文本网络划分成了三个子网络，分别为“单词-单词”、“单词-文本”和“单词-标签”子网络。然后采用改进后的 LINE 算法分别对这些子网络建模，捕获不同类型信息之间的相似性。最终文本的向量表示由相应单词的向量表示取平均得到。

以上介绍的算法都是将异构数据拆分成两两交互的类型分别进行建模，这种方法的缺点就是无法捕获一些强类型对象(strongly-typed objects)产生的共同作用效果。Gui 等人提出了 HyperEdge-Based Embedding (HEBE)^[46] 算法用于解决这个问题。该算

法将一类事件产生的交互集合（即强类型对象）看作一个超边，并将每个超边作为一个整体进行建模。在建模过程中，将超边中的一特定类型的节点设定为预测目标，并用该超边中剩余节点对该目标节点进行预测，从而捕获了基于事件的相似性关系。

2.5.2 基于矩阵分解的方法

CMF^[47] 是一种基于矩阵分解的异构网络表示学习算法，用于同时学习维基百科中实体、类别和词的向量表示。该算法通过矩阵分解的方式分别建立实体之间、实体与类别之间、实体与词之间的联系，总的损失函数如式 2 所示。通过这种方式，一方面可以克服网络稀疏问题，另一方面也方便增加更多的关系矩阵，具有可扩展性。

$$\begin{aligned}
 J = & \sum_{(i,j) \in N_x} (X_{ij} - E_{H_i} E_{L_j}^T)^2 + \sum_{(i,j) \in N_y} (Y_{ij} - C_{H_i} E_{H_j}^T)^2 \\
 & + \sum_{(i,j) \in N_z} (Z_{ij} - W_{H_i} E_{H_j}^T)^2 + \lambda (\|E_H\|_F^2 + \|E_L\|_F^2) \quad (2) \\
 & + \gamma \|C_H\|_F^2 + \delta \|W_H\|_F^2,
 \end{aligned}$$

其中 $X \in R^{|V_e| \times |V_e|}$, $Y \in R^{|V_c| \times |V_e|}$, $Z \in R^{|V_e| \times |V_e|}$ 分别表示实体之间、实体与类别、实体与词之间的链接关系矩阵， $|V_e|$, $|V_c|$, $|V_e|$ 分别为维基百科异构网络中实体节点、类别节点和词节点的个数； $E_H \in R^{|V_e| \times k}$, $C_H \in R^{|V_c| \times k}$, $W_H \in R^{|V_e| \times k}$ 分别为实体、

类别和词节点的向量表示； N_x, N_y, N_z 分别表示 X, Y, Z 矩阵中的非零项； λ, γ, δ 为正则化因子。

2.5.3 基于神经网络的方法

metapath2vec^[48]算法基于随机游走和 skip-gram 模型来学习异构网络中节点的向量表示。该算法设计了基于元路径的随机游走方法，通过定义的对称的 meta-path 来控制随机游走过程。产生的节点序列可以直接输入到 skip-gram 模型中进行训练，不区分节点的类别。作者将 skip-gram 进一步作了改进，改进后的模型 metapath2vec++使得不同类型的节点能够在输出层上被区别开来。Zhang 等人在论文[49]中指出了基于 meta-path 的方法只能严格按照设定的路径类型选取节点建立联系，并不能捕获那些距离较远的节点间的相似性，而且只有一少部分节点符合设定的元路径，从而导致在训练过程中数据的稀疏。由此，作者设计一种基于元图(metagraph)的方法来引导随机游走的过程。由于每个元图中包含了多种元路径，因而在随机游走过程中能够实现更加灵活的匹配，从而捕获更加复杂的节点关系。

在论文[50]中，作者提出了一种无监督的异构网络表示学习框架 ASPEM，为异构节点学习汇聚多方面信息的向量表示。如图 5 所示，(a)图例可以分解成(b)中两个方面的子图。文中设计了一种不兼容信息度量方法来提取这种多方面子图。该模型进一步将不同子图通过 skip-gram 分别进行训练，最后将同一节点的来自不同子图的向量表示整合在一起。

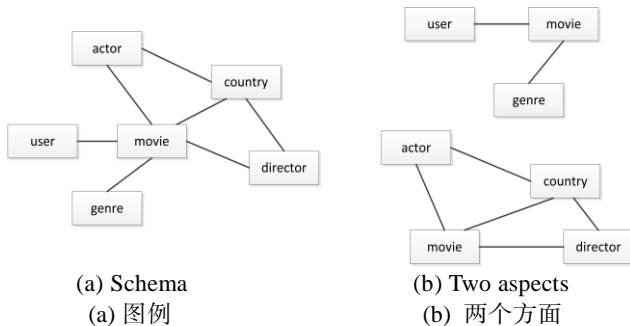


Fig.5 The schema and two aspects of an HIN (from [50]).

图 5 异构网络的图例和两个方面(来源于文献[50])

王鸿伟等人提出了 Signed Heterogeneous Information Network Embedding (SHINE)^[51]模型用于

预测异构情感网络中用户的潜在情感倾向。文中通过建立情感网络、社会网络和资料网络这三大网络来汇聚异构信息，并通过三个深度自编码器分别对三个网络进行信息嵌入，之后将节点在不同网络下得到的向量表示进行汇聚，产生节点最终的向量表示。最后可通过计算用户表征向量之间的相似性来进行情感预测。该模型充分借助异构信息，有效地解决了冷启动问题，进而可有效应用到链接预测和推荐任务中。

2.5.4 综合框架方法

为了整合异构网络中不同类型的信息，Chang 等人设计了一深度架构 Heterogeneous Network Embedding (HNE)^[52]，将不同类型的信息映射到相同的语义空间中。以包含文本和图片信息的异构网络为例，该模型对不同类型的节点采用不同的处理方法（图片-卷积神经网络，文本-全连接神经网络），并将它们映射到同一潜在空间中。进一步地，该模型根据原网络中的拓扑信息对不同类型的节点对（“图片-图片”、“图片-文本”和“文本-文本”）进行相似性建模。由此也证明了不同模态的数据类型能够通过这种深层架构在同一潜在空间中建立联系。

Keikha 等人提出了 HIN2Vec^[53]模型基于神经网络同时学习异构网络中的节点和元路径的向量表示。与上文介绍的基于 metapath 的算法不同，HIN2Vec 是以完全随机游走的方式选取训练集，并通过预测相邻节点间的特定关系来建模异构网络中节点的关联关系。这种多任务学习方法使得表示学习过程得以高效进行。

3 常用数据集

本章中主要介绍上述论文实验中常用的数据集，包括数据集的来源(Link)、节点数($|V|$)、边数($|E|$)、标签类别数($|y|$)、是否有属性信息(Attributes, Att.)等。同时对这些数据集进行分类，类别包括：社交网络(Social Network)、引文网络(Citation Network)、合著网络(Collaboration Network)、网页链接网络

(Webpage Network)和生物学网络(Biological Network) 等。我们将数据集的信息总结如表 7 所示。

Table 7 Data Sets and Descriptions
表 7 数据集及其介绍

Category	Dataset	Directed	$ V $	$ E $	$ y $	Att.	Link
Citation Network	Cora	Yes	2708	5429	7	Yes	
	Citeseer	Yes	3312	4732	6	Yes	https://linqs.soe.ucsc.edu/data
	Pubmed	Yes	19,717	44,338	3	Yes	
	DBLP	Yes	13,184	48,018	5	No	https://github.com/andompesta/ComE/tree/master/data/DBLP
	Citeseer-M10	Yes	10,310	77,218	10	Yes	http://citeseerx.ist.psu.edu/
	PubMed	Yes	19,717	44,338	3	Yes	http://linqs.cs.umd.edu/projects/lbc/index.html
Collabo- ra-tion Network	Arnet-S	No	187,939	1,619,278	100	No	
	Arnet-M	No	268,037	2,747,386	500	No	https://github.com/thunlp/transnet
	Arnet-L	No	945,589	5,056,050	500	No	
Webpage Network	Wiki	Yes	2405	17,981	19	Yes	https://linqs.soe.ucsc.edu/data
	Political blog	Yes	1222	16,715	2	No	http://www-personal.umich.edu/~mejn/netdata/
	WebKB	Yes	877	1608	5	No	http://linqs.cs.umd.edu/projects/projects/lbc/
	Flickr	No	80,513	589,988	195	Yes	http://dmml.asu.edu/users/xufei/datasets.html
	BlogCatalog	No	10,312	333,983	39	Yes	http://www.public.asu.edu/~ltang9/
	YouTube	No	1,138,499	2,990,443	47	Yes	http://socialcomputing.asu.edu/datasets/YouTube2
Social Network	Facebook	No	4039	88,234	4	Yes	https://snap.stanford.edu/data/egonets-Facebook.html
	Twitter	Yes	456,631	14,855,875	No	No	https://snap.stanford.edu/data/higgs-twitter.html
	Political blog	Yes	1222	16,715	2	Yes	http://www-personal.umich.edu/~mejn/netdata/
	Karate club	Yes	34	78	2	No	https://networkdata.ics.uci.edu/data.php?id=105
	Yelp	Yes	249,012	1,779,803	11	No	https://www.yelp.com/dataset_challenge/dataset
	PPI	No	4777	184,812	40	No	https://thebiogrid.org/
Biological Network							

4 总结与展望

本文为融合异质信息的网络表示学习算法设计了一个统一的分类框架，并对一些代表性算法进行了概括介绍和分类对比。总的来说，复杂网络中丰富的信息越来越充分地得到挖掘和运用，这也使模型学得表征向量更能反映网络的真实形态。随着社会的发展，现实世界网络会变的更加复杂，如何有效地汇聚不同类型的信息来辅助表征向量的学习，使其更具推理能力和区别力，同时降低整个模型的

复杂度是今后仍需研究的开放性问题。我们进一步对此领域未来可能的研究方向做了如下总结：

(1)适应大规模复杂网络。现有的融合异质信息的网络表示学习算法仅适用于小规模数据网络，难以适用到现实应用场景中。因此，设计具有可扩展性的算法值得未来进行深入研究。

(2)适应网络的动态变化。现实网络具有动态性，节点、链接关系以及复杂的异质信息都会不断发生变化，而现有的网络表示学习算法大都不适用于在线更新。设计能够有效融合异质信息的在线网

络表示算法具有很高的实用价值。

(3)寻找更多的应用场景。目前网络表示学习的应用场景相对局限,主要包括节点分类、聚类、社区发现、链接预测等。挖掘出更多复杂网络中的应用场景,并设计针对特定任务的网络表示学习算法是未来值得探索的方向。

References:

- [1] Tang J, Aggarwal C, Liu H. Node classification in signed social networks[C]//Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, May 5-7, 2016. Philadelphia: SIAM, 2016: 54-62.
- [2] Zhou D, Huang J, Schölkopf B. Learning with hypergraphs: clustering, classification, and embedding[C]//Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, December 4-7, 2006. Cambridge: MIT Press, 2006: 1601-1608.
- [3] Cao B, Liu N N, Yang Q. Transfer learning for collective link prediction in multiple heterogenous domains[C]//Proceedings of the 27th International Conference on Machine Learning, Haifa, June 21-24, 2010. Madison: Omnipress, 2010: 159-166.
- [4] Singh A P, Gordon G J. Relational learning via collective matrix factorization[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, August 24-27, 2008. New York: ACM, 2008: 650-658.
- [5] Zhao Z Y, Li C, Zhang X J, et al. An incremental method to detect communities in dynamic evolving social networks[J]. Knowledge-Based Systems, 2019, 163: 404-415.
- [6] Zhao Z Y, Zheng S Q, Li C, et al. A comparative study on community detection methods in complex networks[J]. Journal of Intelligent and Fuzzy Systems, 2018, 35(1): 1077-1086.
- [7] Han X, Shi C, Zheng L, et al. Representation learning with depth and breadth for recommendation using multi-view data[C]//Web and Big Data - Second International Joint Conference, Macau, July 23-25, 2018. Springer, 2018: 181-188.
- [8] Shi C, Hu B, Zhao W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge & Data Engineering, 2019, 31(2): 357-370.
- [9] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, August 24-27, 2014. New York: ACM, 2014: 701-710.
- [10] Tang J, Qu M, Wang M, et al. LINE: large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web, Florence, May 18-22, 2015. New York: ACM, 2015: 1067-1077.
- [11] Grover A, Leskovec J. node2vec: scalable feature learning for Networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, August 13-17, 2016. New York: ACM, 2016: 855-864.
- [12] Cao S, Lu W, Xu Q. Grarep: learning graph representations with global structural information[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, October 19-23, 2015. New York: ACM, 2015: 891-900.
- [13] Shi C, Li Y, Zhang J, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(1): 17-37.
- [14] Zhang D, Yin J, Zhu X, et al. Network representation learning: a survey[J]. IEEE Transactions on Big Data, 2018, 10.1109/TBDATA.2018.2850013.
- [15] Tu C, Zhang W, Liu Z, et al. Max-margin deepwalk: discriminative learning of network representation[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, July 9-15, 2016. IJCAI Press, 2016: 3889-3895.
- [16] Hearst M A, Dumais S T, Osman E, et al. Support vector machines[J]. IEEE Intelligent Systems, 1998, 13(4): 18-28.
- [17] Li J, Zhu J, Zhang B. Discriminative deep random walk for network classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, August 7-12, 2016. Stroudsburg: Association for Computer Linguistics, 2016: 1004-1013.
- [18] Zhang X, Chen W, Yan H. TLINE: scalable transductive

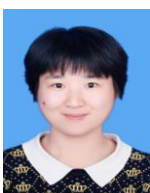
- network embedding[C]//Information Retrieval Technology - 12th Asia Information Retrieval Societies Conference, Beijing, November 30 - December 2, 2016. Springer, 2016: 98-110.
- [19] Li C, Li Z, Wang S, et al. Semi-supervised network embedding[C]//Database Systems for Advanced Applications - 22nd International Conference, Suzhou, March 27-30, 2017. Springer, 2017: 131-147.
- [20] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. *Clinical Orthopaedics and Related Research*, 2016: abs/1609.02907.
- [21] Zhou J, Cui G, Zhang Z, et al. Graph Neural Networks: A Review of Methods and Applications[J]. *Clinical Orthopaedics and Related Research*. 2018, abs/1812.08434.
- [22] Tu C, Zhang Z, Liu Z, et al. TransNet: translation-based network representation learning for social relation extraction[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, August 19-25, 2017. ijcai.org, 2017: 2864-2870.
- [23] Bordes A, Usunier N, Weston J, et al. Translating embeddings for modeling multi-relational data[C]//In Advances in Neural Information Processing Systems, Lake Tahoe, December 5-8, 2013: 2787-2795.
- [24] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. *Clinical Orthopaedics and Related Research*, 2013, abs/1301.3781.
- [25] Qu M, Tang J, Shang J, et al. An attention-based collaboration framework for multi-View network representation learning[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, November 06-10, 2017. New York: ACM, 2017: 1767-1776.
- [26] Wang X, Cui P, Wang J, et al. Community preserving network embedding[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, February 4-9, 2017. Menlo Park: AAAI Press, 2017: 203-209.
- [27] Cavallari S, Zheng V W, Cai H, et al. Learning community embedding with community detection and node embedding on graphs[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, November 06-10, 2017. New York: ACM, 2017: 377-386.
- [28] Zhang Y, Lyu T, Zhang Y. COSINE: community-preserving social network embedding from information diffusion cascades[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, February: 2-7, 2018. Menlo Park: AAAI Press, 2018: 2620-2627.
- [29] Keikha M M, Rahgozar M, Asadpour M. Community aware random walk for network embedding[J]. *Knowledge-Based Systems*, 2018, 148: 47-54.
- [30] Du L, Lu Z, Wang Y, et al. Galaxy network embedding: A hierarchical community structure preserving approach[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, July 13-19, 2018. ijcai.org, 2018: 2079-2085.
- [31] Yang C, Zhao D, Zhao D, et al. Network representation learning with rich text information[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, July 25-31, 2015. Menlo Park: AAAI Press, 2015: 2111-2117.
- [32] Zhang D, Yin J, Zhu X, et al. Collective classification via discriminative matrix factorization on sparsely labeled networks[C]//Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, October 24-28, 2016. New York: ACM, 2016: 1563-1572.
- [33] Huang X, Li J, Hu X. Accelerated Attributed Network Embedding[C]//Proceedings of the 2017 SIAM International Conference on Data Mining, Houston, April 27-29, 2017. New York: ACM, 2017: 633-641.
- [34] Li J, Dani H, Hu X, et al. Attributed network embedding for learning in a dynamic environment[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, November 06-10, 2017. New York: ACM, 2017: 387-396.
- [35] Li H, Wang H, Yang Z, et al. Variation autoencoder based network representation learning for classification[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, July 30 - August 4, 2017. Stroudsburg: Association for Computational Linguistics, 2017: 56-61.
- [36] Le Q V, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of the 31th International Conference on Machine Learning, Beijing, June 21-26, 2014. Cambridge: JMLR.org, 2014: 1188-1196.
- [37] Hamilton W L, Ying R, Leskovec J. Inductive representa-

- tion learning on large graphs[J]. *Clinical Orthopaedics and Related Research*, 2017: abs/1706.02216.
- [38] Garcia-Duran A, Niepert M. Learning graph representations with embedding propagation[J]. *Clinical Orthopaedics and Related Research*, 2017, abs/1710.03059.
- [39] Tu C C, Liu H, Liu Z Y, et al. CANE: context-aware network embedding for relation modeling[C]//*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, July 30 - August 4, 2017. Stroudsburg: Association for Computational Linguistics, 2017: 1722-1731.
- [40] Wei X, Xu L, Cao B, et al. Cross view link prediction by learning noise-resilient representation consensus[C]//*Cross View Link Prediction by Learning Noise-resilient Representation Consensus*, Perth, April 3-7, 2017. New York: ACM, 2017: 1611-1619.
- [41] Liao L, He X, Zhang H, et al. Attributed social network embedding[J]. *IEEE transactions on knowledge and data engineering*, 2018, 30(12): 2257-2270.
- [42] Liang J, Jacobs P, Sun J, et al. Semi-supervised embedding in attributed networks with outliers[C]//*Proceedings of the 2018 SIAM International Conference on Data Mining*, San Diego, May 3-5, 2018. Philadelphia: SIAM, 2018: 153-161.
- [43] Huang X, Li J, Hu X. Label informed attributed network embedding[C]//*Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, Cambridge, February 6-10, 2017. New York: ACM, 2017: 731-739.
- [44] Jacob Y, Denoyer L, Gallinari P. Learning latent representations of nodes for classifying in heterogeneous social networks[C]//*Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, New York, February 24-28, 2014. New York: ACM, 2014: 373-382.
- [45] Tang J, Qu M, Mei Q. PTE: predictive text embedding through large-scale heterogeneous text networks[C]//*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, August 10-13, 2015. New York: ACM, 2015: 1165-1174.
- [46] Gui H, Liu J, Tao F, et al. Large-scale embedding learning in heterogeneous event data[C]//*IEEE 16th International Conference on Data Mining*, Barcelona, December 12-15, 2016. Piscataway: IEEE, 2016: 907-912.
- [47] Sun M, Zhao Y, Liu Z. Representation learning for measuring entity relatedness with rich information[C]//*Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, Buenos Aires, July 25-31, 2015. Menlo Park: AAAI Press, 2015: 1412-1418.
- [48] Dong Y, Chawla N V, Swami A, et al. metapath2vec: scalable representation learning for heterogeneous networks[C]//*Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, August 13-17, 2017. New York: ACM, 2017: 135-144.
- [49] Zhang D, Yin J, Zhu X, et al. MetaGraph2Vec: complex semantic path augmented heterogeneous network embedding[C]//*Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference*, Melbourne, June 3-6, 2018. Springer, 2018: 196-208.
- [50] Shi Y, Gui H, Zhu Q, et al. AspEm: embedding learning by aspects in heterogeneous information networks[C]//*Proceedings of the 2018 SIAM International Conference on Data Mining*, San Diego Marriott Mission Valley, May 3-5, 2018. Philadelphia: SIAM, 2018: 144-152.
- [51] Wang H, Zhang F, Hou M, et al. SHINE: signed heterogeneous information network embedding for sentiment link prediction[C]//*Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Marina Del Rey, February 5-9, 2018. New York: ACM, 2018:592-600.
- [52] Chang S, Han W, Tang J, et al. Heterogeneous network embedding via deep architectures[C]//*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, August 10-13, 2015. New York: ACM, 2015: 119-128.
- [53] Fu T Y, Lee W C, Lei Z. HIN2Vec: explore meta-paths in heterogeneous information networks for representation learning[C]//*Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, November 06-10, 2017. New York: ACM, 2017: 1797-1806.



ZHOU Hui was born in 1995. She is an M.S. candidate at Shandong University of Science and Technology. Her research interests include network representation learning and social network analysis.

周慧(1995–), 女, 山东鄄城人, 山东科技大学计算机科学与工程学院硕士研究生, 主要研究领域为网络表示学习与社交网络分析。



ZHAO Zhongying was born in 1983. She received the Ph.D. in computer science, from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2012. From 2012 to 2014, she was an assistant professor in Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS). She is currently an associate professor in College of Computer Science and Engineering, Shandong University of Science and Technology (SDUST). Her research interests include social network analysis and data mining. She has published 26 papers in international journals and conference proceedings. As the Principal Investigator(PI), she is in charge of more than 10 research fundings, such as the NSF of China and NSF of Shandong Province. She was awarded as Young Scientist for Excellence (SDUST) in 2015, and High-level talent of Huangdao District in 2016.

赵中英(1983–), 女, 山东聊城人, 2012年毕业于中科院计算技术研究所, 获工学博士学位。2012至2014年, 在中科院深圳先进技术研究院任助理研究员, 现任山东科技大学计算机科学与工程学院副教授。主要研究领域为社交网络分析和数据挖掘, 在具有重要影响力的国际期刊和会议上发表论文近30篇, 主持国家自然科学基金、山东省自然科学基金等科研项目十余项。2015年被评为“优秀青年科技工作者”(SDUST), 2016年入选青岛市西海岸新区“高层次一紧缺人才”。



LI Chao was born in 1984. He received the Ph.D. degree in computer application technology from the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, in 2014. From 2014 to 2015, he was a Visiting Scholar with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology. He has published more than ten papers in international journals and conferences. His research interests include big data analysis, social network analysis and data mining. As the Principal Investigator(PI), he is in charge of a series of research fundings, such as the NSF of China.

李超(1984–), 男, 河北唐山人, 2014年毕业于中国科学院深圳先进技术研究院, 获工学博士学位。2014至2015年, 赴香港科技大学计算机科学与工程学院做访问学者。在国内外期刊及会议上发表论文10余篇。研究方向为大数据分析, 社交网络分析和数据挖掘。作为负责人, 主持国家自然科学基金、教育部人文社科基金等。