计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190553
57(9): 1928-1938, 2020

多元图融合的异构信息网嵌入

吴 瑶 申德荣 寇 月 聂铁铮 于 戈
(东北大学计算机科学与工程学院 沈阳 110169)
(18642113630@163.com)

Heterogeneous Information Networks Embedding Based on Multiple Meta-Graph Fusion

Wu Yao, Shen Derong, Kou Yue, Nie Tiezheng, and Yu Ge
(*School of Computer Science and Engineering, Northeastern University, Shenyang* 110169)

Abstract Network embedding methods based on meta-structures (such as meta-path or meta-graph) can effectively utilize heterogeneous network structures. Compared with the meta-path, the meta-graph can capture more complex structural information and help improve the accuracy of similar node matching in heterogeneous information networks. However, the existing meta-graph-based embedding method typically has the following limitations: 1) Most of the meta-graph types are specified by experts, and are not applicable in the application environment of large complex networks; 2) Although multiple meta-graphs are integrated for embedding, the weights of meta-graphs are not considered; 3) Some models use the user's expected semantic relationship to generate a combination of meta-graphs that can preserve specific semantics, but such models are over-reliant on meta-pattern selection and samples used to supervise learning, lacking versatility. Based on this, this paper proposes a heterogeneous network embedding method based on multiple meta-graph fusion. The method includes two parts. The first part is graph discovery. The purpose of graph discovery is to mine important meta-graphs representing the current network structure and semantic features. The second part is node embedding based on multiple meta graph fusion. The main content is to propose a general graph similarity measure method based on meta-graphs, and use the neural network to embed the meta-graph features of nodes. Experimental results show that the proposed method has higher accuracy and efficiency compared with other network embedding methods.

Key words heterogeneous information networks; network embedding; meta-graph; meta-path; fusion

摘 要 基于元结构(如元路径或元图)的网络嵌入方法,能够有效地利用异构网络结构.但与元路径相比,元图能够捕获更加复杂的结构信息,更能提升异构信息网中相似节点匹配的准确性.然而,现有的基于

收稿日期:2019-08-15;**修回日期**:2019-12-30
基金项目:国家自然科学基金项目(61672142,U1435216);国家重点研发计划项目(2018YFB1003404);国家自然科学基金联合基金项目(U1811261);中央高校基本科研业务费专项资金(N171606005)
This work was supported by the National Natural Science Foundation of China (61672142, U1435216), the National Key Research and Development Program of China (2018YFB1003404), the National Natural Science Foundation of China Joint Fund Project (U1811261), and the Fundamental Research Funds for the Central Universities (N171606005).
通信作者:申德荣(shendr@mail.neu.edu.cn)

元图的嵌入方法具有如下局限:大多由专家指定元图类型,在大型复杂网络的应用环境中并不适用;虽然融合了多个元图进行嵌入,但并未考虑元图权重的差异性;部分模型利用用户的期望语义关系生成可以保留特定语义的元图组合,但这类模型过分依赖元图选择和用于监督学习的样本,缺乏通用性.基于此,提出一种多元图融合的异构网络嵌入方法,该方法包括2部分:第1部分是元图发现,目的是挖掘代表当前网络结构和语义特征的重要元图;第2部分是基于多元图融合的节点嵌入,主要内容是提出了一种基于元图的通用节点相似度度量方法,同时利用神经网络嵌入节点的元图特征.实验结果表明,与其他网络嵌入方法相比,提出的方法具有较高的准确性和效率.

关键词　异构信息网;网络嵌入;元图;元路径;融合

中图法分类号　TP391

网络表示学习是将网络嵌入到低维空间中,可以将网络中的节点、边或图表示成向量形式.这种表示形式能够更多保留节点的网络结构或者语义信息,因此可以将该向量表示作为特征应用到各种数据挖掘任务当中.

当前已有许多针对同构网络嵌入的研究成果[1-3]和部分面向异构网络嵌入的研究.利用元路径进行网络嵌入是提取异构网络特征的常用方法,但是元路径结构简单,可能会丢失一些重要的信息.元图可以获取到较元路径更复杂的信息,同时考虑节点间的元图和元路径会得到更好的嵌入结果[4].然而,已有方法大多由专家指定元路径和元图,类型较单一,不适用于大型复杂的网络,而且在指定元图和元路径时未考虑同类型网络各自的独特性,影响节点嵌入准确性.虽然已提出了一些针对异构网络元路径的发现算法,但是还没有见到关于元图的发现方法.部分研究利用频繁子图挖掘算法找到当前异构网络的频繁子图,将其作为元图,但是因为频繁子图算法生成的元图数量较大,并且在结构上可能存在很大程度的相似性,导致后续基于元图的相似度计算复杂度过高.

鉴于元路径是一种特殊形式的元图,本文提出了一种元图发现算法.在各种复杂网络的数据挖掘应用场景下,该算法可以发现网络中的适量关键元图,这些元图可以提取原网络的各类重要信息,减少网络嵌入的信息损失.基于元图发现算法,本文提出一种多元图融合的异构网络嵌入方法,该方法利用自动编码器模型,进一步自适应地选择关键元结构并学习权重值,不仅能够更好地获取网络的结构和语义特征,还可能有效降低人为因素对嵌入结果的影响.

本文的主要贡献有4个方面:

1) 提出了一种元图发现算法,该算法可以挖掘代表当前网络结构和语义特征的重要元图,有助于提升后续嵌入的准确性;

2) 提出了一种基于元图的节点相似度度量方法,相比于已有的计算方法,更具有通用性;

3) 利用神经网络嵌入节点的元图特征,通过原始特征的降维和融合,可以根据当前网络的特性计算不同元图的重要性,提高嵌入的准确性;

4) 通过对比实验,证明本文提出的方法在各类下游应用中的执行效果要优于其他网络嵌入算法.

1　相关工作

网络嵌入方法学习网络节点在低维空间的潜在表示,其结果可以作为各种数据挖掘任务的输入特征,例如聚类、分类、检索、链路预测等,因其较好的执行效果和在各领域的通用性,该方法成为近年来的研究热点.DeepWalk[1]是第1个提出使用语言模型skip-gram和无监督方式学习节点表示的方法,具体来说,将单词序列扩展到图,即用节点代替单词,用随机游走抽取的图路径作为单词的上下文,进而学习潜在节点的表示.Node2vec[2]是DeepWalk方法的扩展,它引入偏置随机游走生成序列,并分析了深度优先和广度优先2种游走方式所保留的不同结构信息,同时相比于DeepWalk依据权重的随机游走,该方法则增加了权重的调整参数.LINE[3]保留网络节点的一阶相似度和二阶相似度信息,采用广度优先搜索策略生成上下文节点,并使用负采样优化skip-gram模型.

近些年,随着应用系统复杂性的提升和用户需求的多样性发展,简单的同构网络难以表示现实世界,而异构信息网的异质性恰好可以解决复杂关系的建模问题,因此针对异构信息网的网络嵌入方法也相继出现.PME[5]是一种基于度量学习的异构信息网络

嵌入模型,以统一方式捕获一阶和二阶邻近关系,并在单独的对象空间和关系空间中构建对象和关系嵌入,同时提出一种损失感知自适应采样方法用于模型优化.异构信息网中的异质性在引入丰富信息的同时也引入了潜在的不兼容语义,为了保留网络嵌入中丰富但可能不兼容的信息,Shi 等人[6]提出了 HEER 算法,该算法通过将边缘表示和异构度量相结合,解决异构信息网络的综合转录问题.为解决网络中多类型节点、关系的结构信息和非结构化属性、文本的信息融合问题,Zhang 等人[7]提出了 SHNE 异构网络嵌入模型,该模型通过 skip-gram 和深度语义编码的联合优化,捕获节点之间的异构结构接近度和非结构化语义关系.Huang 等人[8]提出一个用于大规模网络异构信息学习的通用嵌入框架,加速拓扑结构与节点属性、二阶相似性、链路方向性等信息的联合学习,同时将复杂的建模和优化过程分解为许多简单的独立子问题,以分布式方式完成节点相似度的评估.

元路径和元图作为异构信息网的不同层次的网络模式,可以代表网络中特定的语义关系,因此在大量研究中都得到了充分的应用.TransPath[9]利用知识图中转换机制的概念,将元路径视为从第一个节点到最后一个节点的转换操作.此外还提出了一种用户引导的元路径采样策略,该策略以用户的偏好为指导,可以更精确地探索路径的语义,同时通过避免其他噪声和无意义的元路径提高了模型效率.Ji 等人[10]提出了一个基于元路径融合的关注机制模型,用于异构信息网络的嵌入.该模型首先利用元路径从原始异构网络中抽取出多个同构网络,然后使用共同注意机制融合从多个同构网络中学习的节点嵌入.Sun 等人[4]提出基于元图的网络嵌入模型,利用耦合张量-矩阵分解的方法获得节点的联合嵌入,即元图和其嵌入元路径的公共潜在特征.Zhang 等人[11]引入多对齐属性异构网络的概念建模网络结构,将社交网络中的元路径进行分类,通过异构链接和属性信息定义了用户间各种类型的交互关系,然后将基于元路径的相似度作为深度自动编码器的输入特征,学习用户节点的低维表示.

为了将网络的元路径或者元图特征更好地应用到网络嵌入模型当中,少数研究集中于关键元路径或元图的发现和基于二者的特征度量.Huang 等人[12]提出 3 种基于元结构的相关性度量方法,由于这些度量的计算复杂性较高,进一步设计一种支持以上数据结构的度量算法.Meng 等人[13]研究自动发现元路径的方法,根据用户提供的具有较高相似度的节点对,生成能够解释节点对语义关系的元路径集合.同时提出贪婪树算法选择最相关的元路径以降低计算的时间复杂度.Fang 等人[14]提出基于元图集合的相似度,利用监督方法自动学习元图集合中相似度的正确形式以适应期望的关系类型,同时设计双阶段训练和基于对称的匹配算法加速元图匹配的过程.

2 问题定义

针对异构信息网嵌入问题,我们首先给出相关定义,然后进行问题描述.

定义 1. 异构信息网.异构信息网是带有对象类型映射函数 $\phi:V\to L$ 和链接类型映射函数 $\psi:E\to R$ 的有向图 $G=(V,E)$,图中的每个对象 V 属于一种对象类型 $\phi(V)$,每条边 E 属于一种链接类型 $\psi(E)$.

定义 2. 元图[12].元图 S 是定义在网络模式 $T_G=(L,R)$ 上具有单一源节点 n_{sou}(入度为 0)和单一目标节点 n_{tar}(出度为 0)的有向无环图,可以表示为 $S=(N,M,n_{\text{sou}},n_{\text{tar}})$,其中 N 是节点的集合,M 是边的集合.

定义 3. 元路径[15].给定一个异构信息网(heterogeneous information network, HIN),一条元路径是定义在网络模式图 $T_G=(L,R)$ 中的一条路径,$A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots\xrightarrow{R_l}A_{l+1}$ 定义了类型 A_1 和 A_{l+1} 之间的复合关系,$R=R_1\circ R_2\circ\cdots\circ R_l$ 表示关系之间的复合操作.在本文中,我们认为元路径是特殊结构的元图.

我们研究异构信息网基于元图的节点嵌入方法,即给出一个异构信息网络 $G=(V,E)$,根据生成的元图 $M\_G=(V_{\text{sou}},V_{\text{tar}},E)$ 计算节点相似度 $s(V_{\text{sou}},V_{\text{tar}})$ 作为原始特征 M_{ij},利用神经网络模型得到所有节点的低维表示 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$.

本文提出的多元图融合的异构网络嵌入模型(heterogeneous network embedding model based on multiple meta-graph fusion, HE-MGF)包括元图发现和基于多元图融合的节点嵌入 2 部分.模型的基本思想如图 1 所示,首先利用频繁元图发现算法得到当前异构信息网的一系列频繁元图;然后通过聚类算法,选择具有代表性的元图集合;接下来计算节点间基于元图集合的相似性分数;最后利用神经网络模型对元图集合进行融合嵌入.

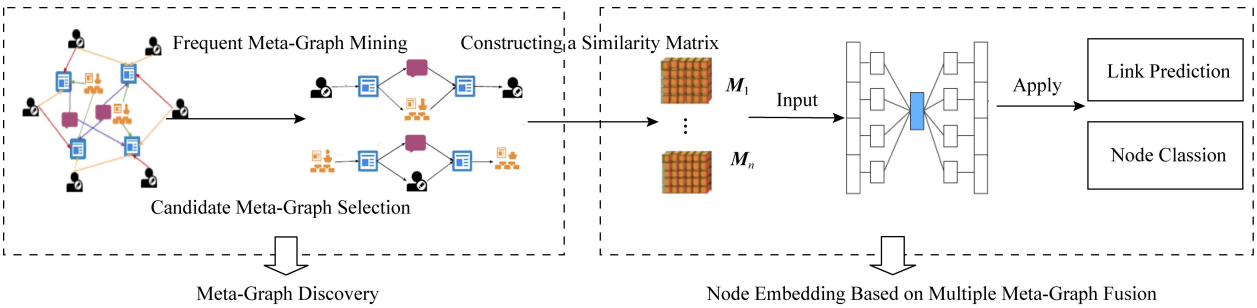

Fig. 1　Framework of HE-MGF

图 1　基于元图的异构信息网嵌入模型的框架

3　元图发现

元图代表特定的结构和语义信息,不同网络中不同类型元图的重要性程度也存在一定差异.因此,需要根据特定网络的特性选择频繁元图集合,以涵盖网络的重要语义特征.所以本文提出频繁元图发现方法挖掘不同网络的关键信息.

3.1　频繁元图挖掘

元图是一种有向无环图,具有单一起点和终点,是网络中子图的子集;而包含节点数量过多的元图对计算节点对的相似度并无重要的作用;大多应用场景要求元图的起点和终点的类型要一致等.为此,本文利用 GRAMI[16] 算法,并进行部分修正,称为 FMGM 算法,该算法能更好地适用于自动生成元图的应用场景.FMGM 包括 3 部分:FrequentMeta-graphMining(见算法 1)是算法的整体流程(其中行⑦～⑮是 FMGM 增加的元图抽取过程),Subgraph-Extension(见算法 2)为子图扩展过程(其中行⑤～⑭是 FMGM 增加的剪枝部分),IsFrequentCsp[16](见算法 3)用于解决子图同构问题.

算法 1. FrequentMeta-graphMining.

输入:$G=(V,E)$、频繁度阈值 t;

输出:所有频繁度超过 t 的元图集合 M.

① M,S 置为空集;

② 构建 G 的频繁边集合 $fedgs$;

③ for e of $fedgs$ /* 选择仅包含目标类型节点的频繁边 */

④　$S \leftarrow S \cup SubgraphExtension$; /* 子图扩展 */

⑤　remove e from G and $fedgs$; /* 从原图和频繁边集合中删除当前边 */

⑥ end for

⑦ for each m of S /* 抽取频繁子图中的元图 */

⑧　$Mg \leftarrow mpath(tp)$; /* 根据设定的起点和重点类型 tp,选择当前子图中满足要求的路径 */

⑨ end for

⑩ for each $mpath1, mpath2$ of Mg

⑪　if $mpath1.sc == mpath2.sc$ and $mpath1.tg == mpath2.tg$

⑫　$M \leftarrow mpath1 \& mpath2$; /* 相同起始点的路径可以构成组合成狭义的元图 */

⑬　end if

⑭　else $M \leftarrow mpath1, mpath2$;

⑮ end for

⑯ return M.

算法 1 的行①②初始化结果集,并构建频繁边集合;行③～⑥选择起始边并执行子图扩展,在此次扩展结束后删除当前边;行⑦～⑮抽取频繁元图.

算法 2. SubgraphExtension.

输入:G 的子图 S、元图图节点数量的最大值 m、t、$fedgs$;

输出:所有扩展 S 的频繁元图 M.

① $M \leftarrow s$, $candidateset$ 置为空; /* 当前子图放入结果集中,候选集置为空 */

② for u in s and e in $fedgs$ /* 针对子图中的每个节点和每条频繁边 */

③　if e 可以被用于扩展 u

④　$ext = s + e$; /* 利用频繁边扩展子图 */

⑤　if 当前子图 ext 未生成过并且其节点量小于阈值 m

⑥　$candidateset \leftarrow ext$;

/* 加入待计算图集 */
⑦ end if
⑧ else if *ext* 中的节点数量为 m
⑨ if *ext* is meta-graph
⑩ *candidateset*←*ext*;
/* 判断 *ext* 中是否有满足元图要求的子图 */
⑪ end if
⑫ end if
⑬ end if
⑭ end for
⑮ for each c of *candidate*
⑯ if $IsFrequentCsp(c)>t$
⑰ $M \leftarrow M \cup SubgraphExtension$; /* 当前元图的频繁度高于阈值,执行子图扩展 */
⑱ end if
⑲ end for
⑳ return M.

算法 2 的行②～④依据频繁边进行子图扩展;行⑤～⑭对扩展结果进行初步筛选;行⑮～⑲计算频繁度.

算法 3. IsFrequentCsp.

输入:s,G,t;

输出:if s 是频繁元图,返回 true,否则返回 false.

① consider the *CSP* of s in G;
② 应用节点和边的一致性规则,去掉矛盾组合;
③ if 任意领域的节点数量小于 t
④ return false; /* 任意领域节点数量小于频繁阈值 */
⑤ end if
⑥ for each *sol* /* *CSP* 的一个解决方案 */
⑦ 在对应领域 *domains* 中标记 *sol* 所有节点;
⑧ if 所有 *domains* 至少有 t 个节点
⑨ return true;
⑩ end if
⑪ end for
⑫ return false.

算法 3 的行①～⑤计算子图的 *CSP*,删除不满足频繁度条件的子图;行⑥～⑪遍历 *domains* 删除不满足条件的子图.算法 3 中的 *CSP* 表示为三元组形式(X,D,C),X 是子图节点(变量)的有序集合,D 是域集合(*domains*),每个 X 都对应一个包含 G 中所有节点的域,C 是 X 之间的约束条件集合 *CSP* 的一个可行的解决方案(*sol*),就是一个满足约束条件 C 的节点分配方案,即向 X 中的每个变量分配 *domains* 中满足约束条件的节点.

3.2 候选元图选择

由于生成的元图数量较多,且大部分结构相似,而元图匹配的计算代价又非常高,所以我们提出基于 k-means 聚类的思想,选择具有代表性的关键元图.具体思路是将生成的频繁元图根据其结构相似度聚成 k 个类别,每个类别再根据其组内的结构相关性和组间的结构差异性,投票选择出最能代表本类别元图的某一个或者某几个元图.

首先介绍元图相似度的计算方法,如果 2 个元图共享一些公共的表示,则两者的结构可能是相似的,代表的语义关系可能也是相似的.基于最大公共子图(MCS)计算元图的结构相似度是一种理想的方法,二者的 MCS 越大,它们的结构相似度就越高.元图结构相似度的计算公式为:

$$S(M_i,M_j)=\frac{(|V_M|+|E_M|)^2}{(|V_{M_i}|+|E_{M_i}|)\times(|V_{M_j}|+|E_{M_j}|)},\tag{1}$$

因为计算最大公共子图是 NP-Hard 问题,所以本文提出一种近似的简化替代算法,每个元图都可以表示为$(\boldsymbol{x},\boldsymbol{y})^{\mathrm{T}}$ 的矩阵形式,其中 $\boldsymbol{x}=(nt_1,nt_2,\cdots,nt_k)$,$\boldsymbol{y}=(et_1,et_2,\cdots,et_k)$,$nt_k$ 表示图的节点类型,et_k 表示图的边类型,两者的取值分别表示特定类型节点和边的数量.因为余弦相似度在计算文本相似性时有较好的效果,所以本部分利用余弦相似度(式(2))计算 2 图的结构相似度,利用式(3)计算类心$(m_x,m_y)^{\mathrm{T}}$:

$$SS(a,b)=\cos(a,b)=\left(\frac{\sum_{i=1}^{k}(x_{ai}\times x_{bi})}{\sqrt{\sum_{i=1}^{k}(x_{ai})^2}\times\sqrt{\sum_{i=1}^{k}(x_{bi})^2}}+\frac{\sum_{i=1}^{k}(y_{aj}\times y_{bj})}{\sqrt{\sum_{i=1}^{k}(y_{aj})^2}\times\sqrt{\sum_{i=1}^{k}(y_{bj})^2}}\right)\times\frac{1}{2}=\left(\frac{\boldsymbol{x}_a\cdot\boldsymbol{x}_b}{\|\boldsymbol{x}_a\|\times\|\boldsymbol{x}_b\|}+\frac{\boldsymbol{y}_a\cdot\boldsymbol{y}_b}{\|\boldsymbol{y}_a\|\times\|\boldsymbol{y}_b\|}\right)\times\frac{1}{2},\tag{2}$$

$$m\_x_i=\frac{1}{n}\sum_{j=1}^{n}x_{ji},m\_y_i=\frac{1}{n}\sum_{j=1}^{n}y_{ji}.\tag{3}$$

下面简单描述元图聚类的方法.

算法 4. Meta-graphSelection.

输入:簇数目 k、频繁元图集合 D;

输出:候选元图集合 R.

① 在D 中随机选择 k 个对象构建初始类心集合 w_1;

② for d of D and c of w

③ $S_{dc}=SS(d,c)$; /* 计算 d 与类心的结构相似度 */

④ end for

⑤ for d of D

⑥ $m_i \leftarrow \max(S_{dc})$; /* 将 d 归类到相似度分数最高的类心所在的聚类 m_i 中 */

⑦ end for

⑧ 计算新类心 w_t;

⑨ if $w_t \neq w_{t-1}$

⑩ repeat ②～⑧; /* 若 2 次类心不同,则重复行②～⑧ */

⑪ end if

⑫ for $(d,*)$ of m_i

⑬ $SI_d=\sum_d SS(d,*)$; /* 对同一类别中的 $SS(d,*)$求和,作为类内元图相似度 */

⑭ end for

⑮ for d of m_i and w_k of D

⑯ $SE_d=-SS(d,w_k)$; /* 类间元图相似度 */

⑰ end for

⑱ $Q_d=SI_d+SE_d$; /* 每个元图的类内相似度分数与类间相似度分数求和 */

⑲ for m_i of D

⑳ $R \leftarrow \max(Q_d)$ in m_i; /* 每个类别中选择分数最高的元图放入结果集中 */

㉑ end for

㉒ return R.

算法 4 的行②～⑪完成元图聚类;行⑫～㉑根据类内相似度和类间差异性选择代表元图.

4　基于多元图融合的网络嵌入

本节利用一种适用于非对称元图的相似度度量方法 HeteMGSim 计算节点的相似度矩阵,并将该矩阵作为后续嵌入模型的原始特征.

4.1　基于元图的节点相似度计算方法——HeteMGSim

现有的基于元图的相似度计算方法,包括 StructCount,SCSE[12],GraphSim[10]这 3 种.这些方法都在不同方面存在缺陷,GraghSim 使用的前提是元图必须为对称形式,StructCount 没有考虑起始节点的活跃度对相似度的影响程度,SCSE 需要使用元结构层的概念,但是对于非标准形式的元图而言,很难界定节点所处的结构层次.受 HeteSim[17]启发,我们提出一种适用非对称元图的节点相似度计算方法 HeteMGSim,该方法的通用性要优于以上计算方法.

HeteSim 的基本思想是将查找元路径实例问题转化为节点对的随机游走问题,并将源节点和目标节点的相似度定义为二者在特定元路径的中点相遇的概率.

因为元图是具有相同起始点的元路径的组合,从给定的元图中可以抽取到若干条元路径,所以计算 2 点在不同元路径相遇概率的乘积可以作为节点在元图中的相遇概率.给定一个 HIN,2 点沿元路径 P 的可达矩阵 $\boldsymbol{C}_p$ 为:

$$\boldsymbol{C}_p=\boldsymbol{U}_{A_1A_2}\times\boldsymbol{U}_{A_2A_3}\times\cdots\times\boldsymbol{U}_{A_lA_{l+1}}, \tag{4}$$

其中,$\boldsymbol{U}_{A_iA_{i+1}}$ 是邻接矩阵的行归一化结果,表示 $A_i\rightarrow A_{i+1}$ 的转移概率矩阵,同时,邻接矩阵的按列归一化 $\boldsymbol{V}_{A_iA_{i+1}}$ 表示 $A_{i+1}\rightarrow A_i$ 的转移概率.式(5)表示 2 点沿元路径在中点类型 M 下相遇的概率:

$$HeteSim(A_1,A_{l+1}|P)=\boldsymbol{U}_{A_1A_2}\times\boldsymbol{U}_{A_2A_3}\times\cdots\times\boldsymbol{U}_{mid-1\,M}\times\boldsymbol{V}_M\times\boldsymbol{V}_{mid+1}\times\boldsymbol{V}_{A_lA_{l+1}}, \tag{5}$$

其中,$P_1P_2,\cdots,P_n$ 是构成元图的元路径集合,所以两者在中点类型 M 下相遇的概率表示为

$$\begin{aligned}&HeteMGSim(A_1,A_{l+1}|S)=\\&HeteSim(A_1,A_{l+1}|P_1)\times\\&HeteSim(A_1,A_{l+1}|P_2)\times\cdots\times\\&HeteSim(A_1,A_{l+1}|P_k)=\\&HeteSim(A_1,A_{l+1}|P_{1L}P_{1R})\times\cdots\times\\&HeteSim(A_1,A_{l+1}|P_{kL}P_{kR}).\end{aligned} \tag{6}$$

4.2　利用神经网络嵌入节点

在异构信息网中,每个元图都可以获取特定的语义关系信息,利用关键元图的信息将其融合,可以很大程度上保留原网络的全部结构和语义信息,所以利用基于元图的节点相似度进行网络嵌入将会是非常有效的方法.

我们由此提出一种多元图融合的无监督网络嵌入方法,利用自动编码器模型[11]首先对基于元图的相似度特征进行降维,然后学习不同元图的权重,最后融合不同元图下节点的向量表示.

深度自动编码器模型可以通过一系列非线性映射操作将网络的原始特征映射到一个低维的特征空间中.自动编码器包括编码和解码 2 部分,编码部分映射原始特征向量到目标特征空间,同时解码部分恢复潜在特征表示到重构空间.模型的目的是保证原始特征与重构特征尽可能相似,以减少降维过程造成的信息损失.

本文扩展传统的自动编码器模型用于解决基于元图的网络节点嵌入问题,该模型以多个元相似矩阵作为输入,比如在社交网络中,如果想保留所有用户之间的相似度,那么矩阵的行就可以表示特定元图下的某一用户与其他用户的相似度分数.每个元图对应的相似度矩阵,都需要执行一系列独立的编码和解码操作,同时为了融合不同元图的信息,增加了编码阶段潜在特征整合的隐藏层和解码阶段潜在特征分解的隐藏层.自动编码器的结构如图 2 所示:

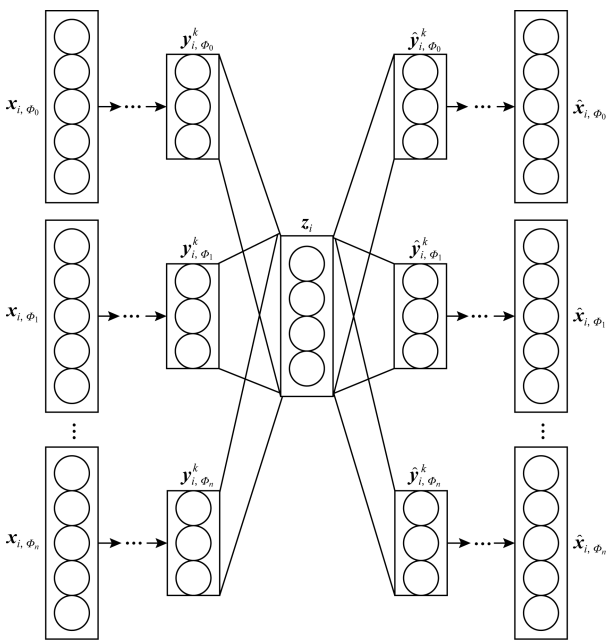

Fig. 2 Auto-encoder for node embedding

图 2 自动编码器嵌入节点

在编码部分,$\boldsymbol{x}_i$ 代表节点的原始特征,$\boldsymbol{y}_i$ 是隐藏层的潜在特征,在目标空间中的编码结果表示为 $\boldsymbol{z}_i$,式(7)~(9)表示这些变量间的关系:

$$\boldsymbol{y}_{i,\Phi_j}^{1}=\sigma(\boldsymbol{W}_{\Phi_j}^{1}\boldsymbol{x}_{i,\Phi_j}+\boldsymbol{b}_{\Phi_j}^{1}), \tag{7}$$

$$\boldsymbol{y}_{i,\Phi_j}^{k}=\sigma(\boldsymbol{W}_{\Phi_j}^{k}\boldsymbol{y}_{i,\Phi_j}^{k-1}+\boldsymbol{b}_{\Phi_j}^{k}),\ k=2,3,\cdots,K, \tag{8}$$

$$\boldsymbol{z}_i=\sigma\Big(\sum_{\Phi_j\in\{\Phi_0,\Phi_1,\cdots,\Phi_n\}}\boldsymbol{W}_{\Phi_j}^{K+1}\boldsymbol{y}_{i,\Phi_j}^{K}+\boldsymbol{b}_{\Phi_j}^{K+1}\Big). \tag{9}$$

在解码部分,输入是潜在特征 $\boldsymbol{z}_i$,输出是重构向量,$\hat{\boldsymbol{y}}_i$ 表示隐藏层的潜在向量.式(10)~(12)表示这些变量间的关系:

$$\hat{\boldsymbol{y}}_{i,\Phi_j}^{K}=\sigma(\hat{\boldsymbol{W}}_{\Phi_j}^{K+1}\boldsymbol{z}_i+\hat{\boldsymbol{b}}_{\Phi_j}^{K+1}), \tag{10}$$

$$\hat{\boldsymbol{y}}_{i,\Phi_j}^{k-1}=\sigma(\hat{\boldsymbol{W}}_{\Phi_j}^{k}\hat{\boldsymbol{y}}_{i,\Phi_j}^{k}+\hat{\boldsymbol{b}}_{\Phi_j}^{k}),\ k=2,3,\cdots,K, \tag{11}$$

$$\hat{\boldsymbol{x}}_{i,\Phi_j}=\sigma(\hat{\boldsymbol{W}}_{\Phi_j}^{1}\hat{\boldsymbol{y}}_{i,\Phi_j}^{1}+\hat{\boldsymbol{b}}_{\Phi_j}^{1}). \tag{12}$$

该模型的目标函数是最小化网络中所有实例的原始特征向量与重构特征向量之间的编码损失.由于输入向量是极其稀疏的,即相似度矩阵中零元素的数量要远远超过非零元素的数量,直接将其放入模型中会增加零元素的编码解码操作,为克服这个问题,在定义损失函数时,为非零特征的损失分配较高的权重值,进一步对每个元图的损失求和,最终的损失函数表示为

$$L=\sum_{\Phi_k\in\{\Phi_0,\Phi_1,\cdots,\Phi_n\}}\sum_{u_i\in V}\left\|(\boldsymbol{x}_{i,\Phi_j}-\hat{\boldsymbol{x}}_{i,\Phi_j})\odot\boldsymbol{b}_{i,\Phi_j}\right\|_2^2. \tag{13}$$

5 实验与分析

在数据集上测试本文提出的网络嵌入算法.

5.1 数据集

如表 1 所示,本文使用 3 个数据集,DBLP1 包含 339 篇文章以及文章的作者、来源、参考文献 3 种属性信息.为降低稀疏性,我们在数据集中选择发表论文 25 篇以上的作者集合,以及论文数量在 50 篇以上的会议和期刊,利用论文及其属性信息构建异构网络.网络链接包括 author-paper, paper-venue, paper-paper 这 3 种类型.我们用 DBLP1 完成节点的链路预测任务.DBLP2[17] 是带有标签的数据集,与 DBLP1 相比,DBLP2 增加了论文的主题信息和各类节点的标签信息,其中的标签类型代表 4 类研究领域.因为 DBLP2 丢失了论文间的引用关系,而该信息对 paper, author 类型节点的链路预测有很重要的作用(如预测论文未来的引用关系、预测合作者关系),所以我们仅利用 DBLP2 完成节点分类任务.DEG 是包含学位信息的数据集,其构建的异构网络包括 degree, person, school, term 这 4 种类型的节点以及 person-degree, degree-term, degree-school 这 3 种类型的连接关系.我们用 DEG 完成节点的链路预测任务.

Table 1 Statistics of Datasets

表 1 数据集统计

Dataset	Size/MB	# Type	# Node	# Link	Label
DBLP1	2.1	3	90 857	70 122	0
DBLP2	3.5	4	37 791	17 079	4
DEG	1.4	4	20 542	23 756	0

5.2　实验设置及对比实验

本文将 HE-MGF 算法与 4 种算法进行比较.

1) AD.我们用人工指定关键元图代替 HE-NGF 模型的元图生成部分.图 3 所示的元图是 DBLP1,DEG 的常用元图.

2) Dual-stage[14].一种利用样本进行监督学习的方法,具体思路是利用种子元图,通过结构相似度启发式获取候选元图,再根据特定语义下选择的训练样本进行双向训练得到有效的元图和相应的权重值.

3) DeepWalk[1].忽略网络异构性,利用 skip-gram 模型的网络表示学习方法.

4) Metapath2vec[18].基于元路径的随机游走重构节点的异质邻居,并用 skip-gram 模型进行节点嵌入的方法.

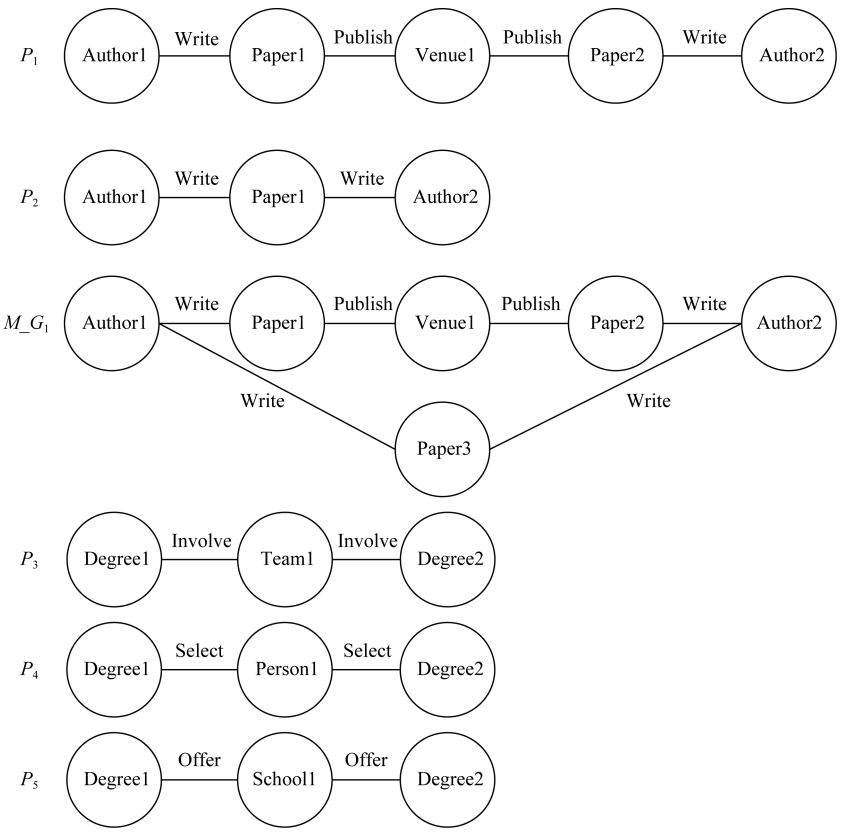

Fig. 3　Meta-graphs

图 3　常用元图

设置 AD 和 Dual-stage 对比实验的目的是评估元图生成算法的有效性.根据本文的频繁元图挖掘和聚类算法得到当前网络中的部分关键元图,将其嵌入结果与常用的 DBLP 元图以及利用样本进行监督学习的结果进行分析比较,以评估元图生成和聚类算法的优越性.设置 DeepWalk 和 Metapath2vec 对比实验的目的是评估利用元图信息相比于元路径信息,或者不考虑异构信息进行嵌入的优越性.

实验中,将 4 种算法的嵌入结果分别应用于 DBLP 的链路预测和节点分类的挖掘任务中,因为所选数据集包含时间属性信息,所以可以实现 2 种链路预测任务,包括当前网络中未发现链路的推理和未来时点的链路预测.在链路预测任务中,随机选择作者类型的节点对构建测试集,包括当前存在链接(合作者关系)的正样本和不存在链接的负样本,目标是预测网络中的合作者关系.在节点分类任务中,我们利用 DBLP2 的类标签训练 SVM 模型,测试节点关于研究领域分类的准确性.在 DEG 链路预测任务中,随机选择 degree 类型的节点对,根据节点对是否处于同一语义类别划分正样本和负样本.用准确率、召回率、$F1$ 值度量不同算法的执行效果.

5.3　实验结果

1) 预测、分类性能比较

如表 2～4 所示,在实验参数选择最优的情况下,与 AD 和 Dual-stage 算法相比,HE-MGF 模型发现的元图能更全面地获取节点间的结构和语义关系.而与 DeepWalk 和 Metapath2vec 算法相比,HE-MGF 模型充分考虑了节点类型的异构信息,以及网络中

代表不同语义关系的复杂图结构,使学习到的节点向量保留了除节点邻接关系之外的其他重要信息,进而达到较好的链路预测效果.

Table 2 Link Prediction Performance in Network Embedding(DBLP1)

表 2 网络嵌入算法的链路预测执行效果(DBLP1)

Algorithm	Precision	Recall
AD	0.445 6	0.234 6
Dual-stage	0.6816	0.374 5
DeepWalk	0.666 7	0.200 0
Metapath2vec	0.873 4	0.052 3
HE-MGF	0.877 4	0.449 5

Table 3 Node Classification Performance in Network Embedding (DBLP2)

表 3 网络嵌入算法的节点分类执行效果(DBLP2)

Algorithm	Micro-$F1$	Macro-$F1$
AD	0.442 5	0.429 6
Dual-stage	0.587 4	0.554 1
DeepWalk	0.577 3	0.553 4
Metapath2vec	0.288 0	0.247 8
HE-MGF	0.623 4	0.580 0

Table 4 Link Prediction Performance in Network Embedding(DEG)

表 4 网络嵌入算法的链路预测执行效果

Algorithm	Precision	Recall
AD	0.810 5	0.724 3
Dual-stage	0.807 4	0.743 9
DeepWalk	0.675 4	0.342 8
Metapath2vec	0.822 9	0.703 1
HE-MGF	0.851 6	0.784 9

2) 频繁度阈值的影响

元图发现算法生成的关键元图数量取决于子图频繁度的设置,较大的频繁度可以获得最关键的元图集合,而较小的阈值设定可以得到更多类型的元图,增加结构和语义信息的多样性.通过设置不同的阈值挖掘频繁元图,进而完成节点嵌入和节点分类,可以发现适用于当前数据集的阈值大小.我们比较了在不同嵌入维度 d 下,频繁度阈值 s 对节点分类效果的影响,实验结果如图 4 所示,最优值为:$d=150$,$s=60$.

3) 节点的向量表示维度的选择

HE-MGF 中自动编码器的嵌入维度会很大程度上影响节点表示学习的效果,所以需要对嵌入维度的敏感性进行测试和分析.在 4 种对比算法中,因为 DeepWalk 和 Metapath2vec 与 HE-MGF 采用的是完全不同的嵌入方式,而另外 2 种对比算法与 HE-MGF 除了在选择元图的部分存在差异外,它们使用的网络嵌入模型都是编码器,所以嵌入维度的敏感性是相似的.因此本文选择对比算法中的 DeepWalk 和 Metapath2vec,观测维度参数的变化对嵌入结果的影响.实验结果如图 5 所示,可以发现随着嵌入维度的增大,节点分类的效果得到了显著的提升,直到达到最优值,之后嵌入维度的增加反而会降低节点分类效果.HE-MGF 的嵌入维度最佳值为 $d=150$.

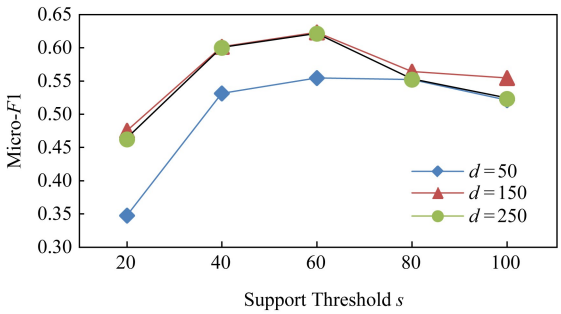

Fig. 4 Influence of support threshold

图 4 频繁度阈值的影响

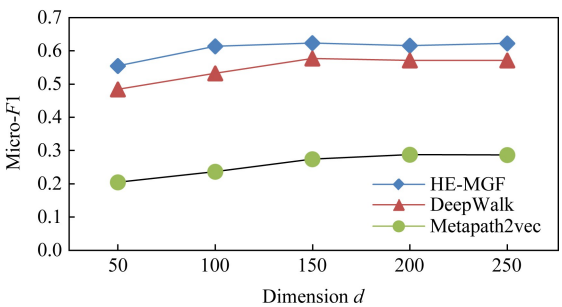

Fig. 5 Influence of node embedding dimension

图 5 节点嵌入维度的影响

4) 聚类数量 k 的影响

较多的元图能抓取更加全面的信息,但是考虑到后续节点嵌入模型具有较高的复杂度,如果选择过多的元图,则需要大量计算节点在不同结构下的相似度,进而构建目标节点的相似度矩阵.根据实验结果发现,过多的元图对于嵌入效果并没有显著的提升,所以考虑到效率问题,应该在保证较高嵌入质量的前提下选择较少的元图.我们比较了在不同频繁度阈值 s 下,k 值对链路预测效果的影响.我们在这部分展示 DBLP1 与 DEG 的 k 值选择结果,实验

结果如图 6 所示,DBLP1 的最佳值为 $s=60$,$k=8$.得到的有效元图包括 APA,APAPA,APPPA,APVPA,APPA,APA(APPPA),APVPAPA,APA(APPPVPA).如图 7 所示,DEG 的最佳值为 $s=20$,$k=8$.得到的有效元图包括 DTD,DPD,DSD,DTDTD,DPDPD,DTDPD,DTD(DSD),DPD(DSD).与常用元图相比,HE-MGF 算法发现的元图不但类型更全面,而且应用于不同网络的灵活性较高.

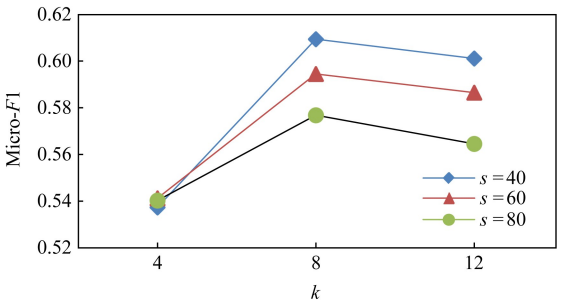

Fig. 6　Influence of k in DBLP1

图 6　DBLP1 中 k 值的影响

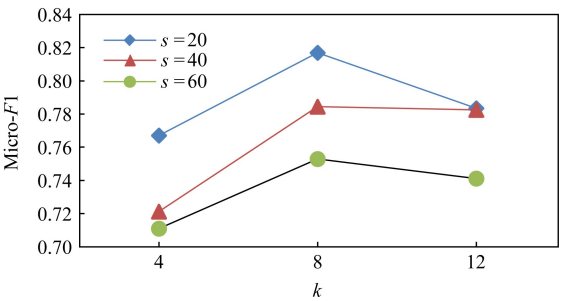

Fig. 7　Influence of k in DEG

图 7　DEG 中 k 值的影响

5) HE-MGF 算法的执行效率

因为频繁元图发现算法中频繁度参数的设定对 HE-MGF 执行时间有重要影响,所以在这部分我们比较了在不同频繁度阈值 s 下,数据集规模对算法性能的影响,实验结果如图 8 所示:

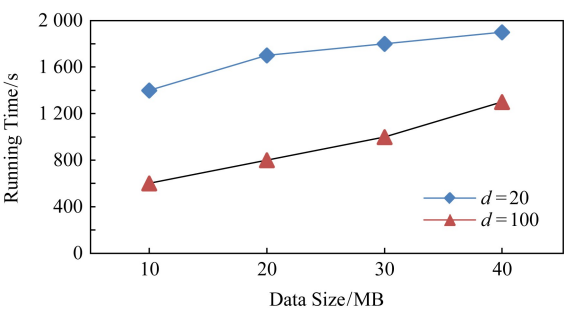

Fig. 8　Running time of different data sizes

图 8　不同数据集规模的运行时间

6　总　　结

本文提出一种多元图融合的网络节点表示模型.其中,利用频繁元图挖掘算法和进一步元图聚类实现了元图的自动发现任务,避免通过用户指定和监督学习生成元图造成的信息缺失问题.同时,基于多元图的网络节点嵌入方法,利用无监督的自动编码器结构,可以自动学习网络中的关键元路径以及相应的权重值.另外,对基于元路径计算节点相似度方法的改进可以较好地适应元图的应用场景,且具有更强的通用性.实验证明,本文提出的方法在准确率和召回率上都有较好的效果.接下来,将本文的网络嵌入方法扩展到跨网络或者动态网络的应用场景中,以解决更广泛的异构信息网的数据挖掘问题.

参　考　文　献

[1] Perozzi B, Alrfou R, Skiena S. DeepWalk: Online learning of social representations [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701-710

[2] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855-864

[3] Tang Jian, Qu Meng, Wang Mingzhe, et al. LINE: Large-scale information network embedding [C] //Proc of the 24th Int Conf on World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015: 1067-1077

[4] Sun Lichao, He Lifang, Huang Zhipeng, et al. Joint embedding of meta-path and meta-graph for heterogeneous information networks [C] //Proc of 2018 IEEE Int Conf on Big Knowledge (ICBK). Piscataway, NJ: IEEE, 2018: 131-138

[5] Chen Hongyu, Yin Hongzhi, Wang Weiqing, et al. PME: Projected metric embedding on heterogeneous networks for link prediction [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 1177-1186

[6] Shi Yu, Zhu Qi, Guo Fang, et al. Easing embedding learning by comprehensive transcription of heterogeneous information networks [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 2190-2199

[7] Zhang Chuxu, Swami A, Chawla N V. SHNE: Representation learning for semantic-associated heterogeneous networks [C] //Proc of the 12th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2019: 690-698

[8] Huang Xiao, Li Jundong, Zou Na, et al. A general embedding framework for heterogeneous information learning in large-scale networks [J]. ACM Transactions on Knowledge Discovery from Data, 2018, 12(6): 70:1-70:24

[9] Fang Yang, Zhao Xiang, Tan Zhen, et al. TransPath: Representation learning for heterogeneous information networks via translation mechanism [J]. IEEE Access, 2018, 6: 20712-20721

[10] Ji Houye, Shi Chuan, Wang Bai. Attention based meta path fusion for heterogeneous information network embedding [C] //Proc of the Pacific Rim Int Conf on Artificial Intelligence: Trends in Artificial Intelligence. Berlin: Spriner, 2018: 348-360

[11] Zhang Jiawei, Xia Congying, Zhang Chenwei, et al. BL-MNE: Emerging heterogeneous social network embedding through broad learning with aligned autoencoder [C] //Proc of 2017 IEEE Int Conf on Data Mining (ICDM). Piscataway, NJ: IEEE, 2017: 605-614

[12] Huang Zhipeng, Zheng Yudian, Cheng Reynold, et al. Meta structure: Computing relevance in large heterogeneous information networks [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1595-1604

[13] Meng Changping, Cheng Reynold, Maniu S, et al. Discovering meta-paths in large heterogeneous information networks [C] //Proc of the 24th Int Conf on World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015: 754-764

[14] Fang Yuan, Lin Wenqing, Zheng V W, et al. Semantic proximity search on graphs with metagraph-based learning [C] //Proc of the 32nd IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2016: 277-288

[15] Sun Yizhou, Han Jiawei, Yan Xifeng, et al. PathSim: Meta path-based top-k similarity search in heterogeneous information networks [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003

[16] Elseidy M, Abdelhamid E, Skiadopoulos S, et al. GRAMI: Frequent subgraph and pattern mining in a single large graph [J]. Proceedings of the VLDB Endowment, 2014, 7(7): 517-528

[17] Shi Chuan, Kong Xiangnan, Huang Yue, et al. HeteSim: A general framework for relevance measure in heterogeneous networks [J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(10): 2479-2492

[18] Dong Yuxiao, Chawla Nitesh V, Swami A. Metapath2vec: Scalable representation learning for heterogeneous networks [C] //Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 135-144

Wu Yao, born in 1994. Master candidate. Her main research interests include heterogeneous information networks and network embedding.

Shen Derong, born in 1964. PhD, professor. Her main research interests include Web data processing and distributed database.

Kou Yue, born in 1980. PhD, associate professor. Her main research interests include entity resolution and Web data management.

Nie Tiezheng, born in 1980. PhD, associate professor. His main research interests include data quality and data integration.

Yu Ge, born in 1962. PhD, professor. His research interests include data stream, data mining, distributed database.