

**摘要**

异构图中的表示学习旨在为每个节点追求有意义的矢量表示，以便于下游应用（例如链接预测，个性化推荐，节点分类等）。然而，此任务具有挑战性，不仅需要合并多种类型的节点和边缘组成的异构结构（图形）信息，也需要考虑与每个节点关联的异构属性或内容（例如，文本或图像）。尽管已经在同构（或异构）图嵌入，属性图嵌入以及图神经网络方面做出了大量努力，但是很少有人能够有效地共同考虑每个节点的异构结构（图）信息以及异构内容信息。在本文中，我们提出了一种异构图形神经网络模型HetGNN，以解决此问题。具体来说，我们首先介绍一种具有重启策略的随机游走，以对每个节点采样固定数量的强相关异构邻居，并根据节点类型对它们进行分组。接下来，我们设计具有两个模块的神经网络架构，以聚合那些采样的相邻节点的特征信息。第一个模块对异构内容的“深度”特征交互进行编码，并为每个节点生成内容嵌入。第二模块汇总不同相邻组（类型）的内容（属性）嵌入，并通过考虑不同组的影响将它们进一步组合以获得最终节点嵌入。最后，我们利用图上下文损失和小批量梯度下降过程以端到端的方式训练模型。在多个数据集上进行的广泛实验表明，HetGNN在各种图挖掘任务（即链接预测，推荐，节点分类和聚类以及归纳节点分类和聚类）中可以胜过最先进的基线。

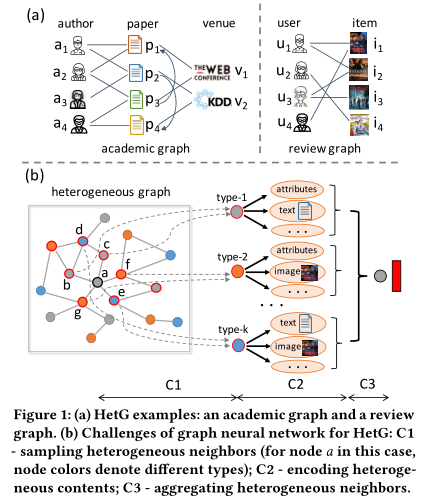
关键词：异构图，图神经网络，图嵌入

**1引言**

异构图[HetG] [26，27]包含丰富的信息，这些信息**包括多类型节点之间的结构关系（边缘）以及与每个节点相关的非结构化内容。**例如，图1（a）中的学术图表示作者和论文（写），论文和论文（引用），论文和地点（发表）之间的关系。此外，该图中的节点带有属性（例如，作者ID）和文本（例如，论文摘要）。另一个示例示出了评论图中的用户-项目关系，并且节点与属性（例如，用户ID），文本（例如，项目描述）和图像（例如，项目图片）相关联。 HetG的这种普遍性导致了对相应图挖掘方法和算法研究的涌入，例如关系推断[2，25，33，35]，个性化推荐[10，23]，节点分类[36]等。

传统上，这些HetG任务中的许多任务都依赖于从手动特征工程任务派生的特征向量。这需要规范和计算有关HetG的不同统计量或属性，作为下游机器学习或分析任务的特征向量。但是，这可能是非常有限的，并且不能一概而论。最近，出现了**表示学习方法以使特征工程任务自动化**，从而可以促进大量下游机器学习或分析任务。**从同质图[6、20、29]开始，图表示学习已扩展到异构图[1、4]，属性图[15、34]以及特定图[22、28]**。例如，最初开发了“浅”模型，例如DeepWalk [20]，以将图上的一组短随机游走馈送到SkipGram模型[19]，以便在这些游走步骤中近似节点共现概率，获取节点嵌入。随后，提出了**语义感知**方法，例如metapath2vec [4]，以解决异构图中的节点和关系异质性。另外，诸如ASNE [15]的**内容感知**方法利用“潜在”特征和属性两者来学习图中的节点嵌入。

这些方法直接学习节点“潜在”嵌入，但在捕获丰富的邻域信息方面受到限制。图神经网络（GNN）使用深层神经网络来聚合相邻节点的特征信息，这使聚合得到的嵌入在功能上更加强大。另外，GNN可以自然地应用于涉及训练期间中不存在的节点的归纳任务。例如，GCN [12]，GraphSAGE [7]和GAT [31]分别采用卷积运算符，LSTM体系结构和自注意机制来聚集相邻节点的特征信息。 GNN的进步和应用主要集中在同构图上。当前最先进的GNN并不能很好地解决HetG面临的以下挑战，我们将在本文中解决这些挑战。



（C1）**HetG中的许多节点可能未连接到所有类型的邻居。另外，相邻节点的数量因节点而异。**例如，在图1（a）中，任何作者节点都没有与场所节点的直接连接。同时，在图1（b）中，节点a具有5个直接邻居，而节点c仅具有2个。大多数现有的GNN仅聚合直接（一阶）邻居节点的特征信息，并且特征传播过程可能会削弱更远邻居的影响。此外，弱关联的邻居（“噪声”邻居）削弱了“集线器”节点的嵌入生成，并且由于有限的邻居信息而无法充分表示“冷启动”节点的嵌入。因此挑战1是：如何针对HetG中每个节点的嵌入生成过程去采样与其紧密相关的异构邻居，如图1（b）中的C1所示？

（C2）**HetG中的节点可以携带非结构化的异构内容**，例如属性，文本或图像。另外，**与不同类型的节点相关联的内容可以不同**。例如，在图1（b）中，类型1节点（例如，.b或c）包含属性和文本内容，类型2节点（例如，其中的f或g）包含属性和图像，类型-k个节点（例如d或e）与文本和图像相关联。当前GNN的直接串联操作或线性变换无法对节点异构内容之间的“深度”交互进行建模。而且，不适用于所有节点类型的相同特征转换功能，因为它们的内容彼此不同。因此，挑战2是：如何设计节点内容编码器以解决HetG中不同节点的内容异质性，如图1（b）中的C2所示？

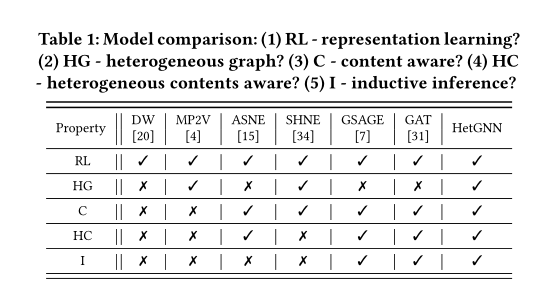
（C3）**不同类型的邻居对HetG中节点嵌入的贡献不同**。例如，在图1（a）的学术图中，作者和论文邻居应该对作者节点的嵌入产生更大的影响，因为场所节点包含各种主题，因此具有更一般的嵌入。当前的大多数GNN主要关注同质图，并且不考虑节点类型的影响。因此，挑战3是：如何通过考虑不同节点类型的影响来聚合异构邻居的特征信息，如图1（b）中的C3所示。

为了解决这些挑战，我们提出了HetGNN，这是一种用于HetG中表示学习的异构图神经网络模型。**首先，**我们设计了一种基于重新启动的随机行走策略，对HetG中每个节点的固定大小的强相关异质邻居进行采样，并根据节点类型对它们进行分组。**接下来，**我们设计了一个具有两个模块的异构图神经网络架构，以聚合在上一步中采样到的邻居的特征信息。第一个模块采用循环神经网络对异构内容的“深度”特征交互进行编码，并获得每个节点的内容嵌入。第二个模块利用另一个循环神经网络来聚合不同相邻组的内容嵌入，然后通过注意力机制将其进一步组合，以测量异构节点类型的不同影响并获得最终节点嵌入。**最后，**我们利用图上下文损失和小批量梯度下降过程来训练模型。总而言之，我们工作的主要贡献是：

•我们将涉及图结构异质性和节点内容异质性的异构图表示学习问题形式化

•我们提出了一种创新的异构图神经网络模型，即HetGNN，用于在HetG上进行表示学习。 HetGNN能够捕获结构和内容的异质性，并且对于转导和归纳任务都非常有用。表1总结了HetGNN的主要优势，与包括同构，异构，属性图模型和图神经网络模型在内的许多最新模型相比。

•我们在几个公共数据集上进行了广泛的实验，我们的结果证明了HetGNN在许多图挖掘任务（包括链接预测，推荐，节点分类和聚类以及归纳节点分类和聚类）上均优于最先进的基准。



**2问题定义**

在本节中，我们介绍将在本文中使用的与内容相关的异构图的概念，然后正式定义异构图表示学习的问题。

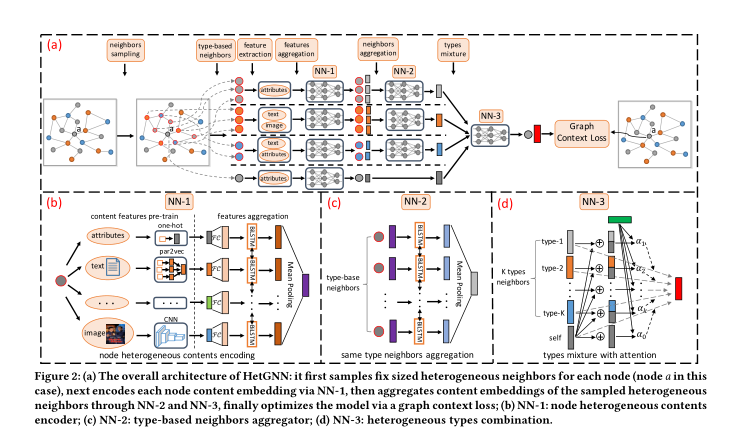
定义2.1。内容相关的异构图。与内容相关的异构图（C-HetG）被定义为具有多种类型的节点V和链接E的图G =（V，E，OV，RE）。OV和RE分别表示对象类型集和关系类型集。。另外，每个节点都与异构内容，例如属性，文本或图像相关联。

图1（a）中的学术图是C-HetG。节点类型OV包括作者，论文和地点。链接类型RE包括作者撰写论文，论文引用论文和论文发表地点。此外，作者或场所节点都与作者所撰写的或包含在场所中的论文摘要相关联，并且论文节点包含摘要、参考文献以及场所。图1（a）中的二分评价图也是C-HetG，为| OV | + | RE | ≥3，其中OV包括用户和物品，则关系RE表示审阅行为。用户节点与用户撰写的评论相关联，物品节点包含标题，描述和图片。

问题1.异构图表示学习。给定具有节点内容集C的C-HetG G =（V，E，OV，RE），任务是设计一个具有参数Θ的模型FΘ，以学习三维嵌入E∈R | V |×d（d 《 | V |）。 ），它们能够同时编码异质结构紧密度和异质非结构化内容。节点嵌入可用于各种图挖掘任务中，例如链接预测，推荐，多标签分类和节点聚类

**3 HetGNN**

在本节中，我们正式提出HetGNN以解决第1节中描述的三个挑战。HetGNN包括四个部分：（1）对异构邻居进行采样； （2）编码节点异构内容； （3）聚集异类邻居； （4）制定目标，设计模型训练程序。图2说明了HetGNN的框架。



**3.1采样异构邻居（C1）**

大多数图神经网络（GNN）的关键思想是聚合来自节点的直接（一阶）邻居的特征信息，例如GraphSAGE [7]或GAT [31]。但是，将这些方法直接应用于异构图可能会引起一些问题：

•它们无法直接从不同类型的邻居中捕获特征信息。例如，作者没有直接与图1（a）中的局部作者和会场邻居联系，这可能会导致代表不足。

•它们被各种邻居大小削弱。有些作者写了许多论文，而有些人在学术界只写了几篇论文。有些项目被许多用户审查，而有些则在审查图表中收到的反馈很少。弱关联的邻居可能会损害“集线器”节点的嵌入，并且“冷启动”节点的嵌入可能无法充分体现。

•它们不适用于聚合具有不同内容特征的异构邻居。异构邻居可能需要不同的特征转换才能处理不同的特征类型和尺寸。

鉴于这些问题并解决挑战C1，我们设计了一种基于随机游走并重启（RWR）的异构邻居采样策略。它包含两个连续的步骤：

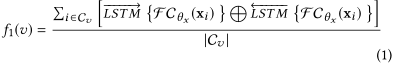
•步骤1：采样固定长度的RWR。我们从节点 v∈V开始随机游走。该游走以概率p迭代行进到当前节点的邻居或返回到起始节点。 RWR运行直到成功收集固定数量的节点（表示为RWR（v））。请注意，RWR（v）中不同类型节点的数量受到限制，以确保所有节点类型都被采样。

•步骤2：对不同类型的邻居进行分组。对于每个节点类型t，我们根据频率从RWR（v）中选择前k个节点t，并将它们作为节点v的t型相关邻居的集合。

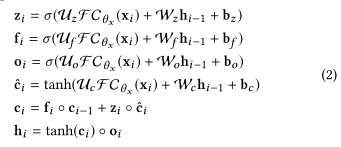
由于以下原因，该策略能够避免上述问题：（1）RWR为每个节点收集所有类型的邻居； （2）每个节点的采样邻居大小是固定的，并选择访问最频繁的邻居； （3）将相同类型（具有相同内容功能）的邻居划分为一组，以便可以应用基于类型的聚合。接下来，我们设计具有两个模块的异构图神经网络架构，以聚合每个节点的采样异构邻居的特征信息。

**3.2编码异构内容（C2）**

为解决挑战C2，我们设计了一个模块，用于从节点v∈V中提取异构内容Cv，并通过神经网络f1将其编码为固定大小的嵌入。具体而言，我们用xi∈Rdf×1中表示Cv中第i个内容的特征表示（df：内容特征维）。请注意，关于不同类型的内容可以使用不同的技术对xi进行预训练。例如，我们可以利用Par2Vec [13]来预训练文本内容，或者使用CNN [17]来预训练图像内容。与以前的将不同内容特征直接或线性转换为统一矢量的模型[7，31]不同，我们设计了一种基于双向LSTM（Bi-LSTM）[9]的新架构来捕获“深度”特征交互并获得更大的表达能力。形式上，v的内容嵌入形式如下：



其中f1（v）∈Rd×1（d：内容嵌入维），FCθx表示特征转换器，可以是恒等式（无变换），具有参数θx的全连接神经网络等。运算符表示串联。 LSTM公式为：



其中是第i个内容的输出隐藏状态，◦表示Hadamard乘积，和是可学习的参数，zi、fi和oi分别是忘记门矢量，输入门矢量和第i个内容特征的输出门矢量。更具体地说，上述架构首先使用不同的FC层来转换不同的内容特征，然后使用Bi-LSTM捕获“深层”特征交互并累积所有内容特征的表达能力，最后在所有的隐藏状态上应用均值池层以获得v的一般内容嵌入，如图2（b）所示。请注意，Bi-LSTM在无序内容集Cv上运行，这受先前工作[7]聚合无序邻居的启发。此外，我们使用不同的Bi-LSTM来聚合不同类型节点的内容特征，因为它们的内容互不相同。

该编码体系结构具有三个主要优点：（1）具有简洁的结构，具有较低的复杂度（较少的参数），使得模型的实现和调整相对容易； （2）能够融合异构内容信息，表现出较强的表达能力； （3）添加额外的内容功能很灵活，使模型扩展方便

**3.3聚合异构邻居（C3）**

为了聚合每个节点的异构邻居的内容嵌入（从第3.2节获得）并解决挑战C3，我们设计了另一个模块，该模块是基于类型的神经网络。它包括两个连续步骤：（1）同一类型的邻居聚合； （2）类型组合。

3.3.1同类型邻居聚合。

在第3.1节中，我们使用基于RWR的策略为每个节点采样不同节点类型的固定大小的邻居集。因此，我们将v∈V的t型采样邻居集表示为Nt（v）。然后，我们使用神经网络汇总v'∈Nt（v）的内容嵌入。形式上，v的聚集t型邻居表示如下：



其中（d：聚合内容嵌入维数），是3.2节中模块生成的v'的内容嵌入，是t型邻居聚合器，可以是全连接神经网络，卷积神经网络，循环神经网络等。在这项工作中，我们使用Bi-LSTM，因为它在实践中会产生更好的性能。因此，我们重新构造如下：



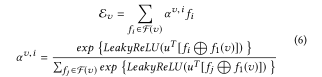
其中LSTM模块的公式与公式（2）除了输入和参数设置之外相同 。显然，我们采用Bi-LSTM聚合所有t型邻居的内容嵌入，并使用所有隐藏状态的平均值表示一般的聚合嵌入，如图2（c）所示。我们使用不同的Bi-LSTM区分邻居聚合的不同节点类型。请注意，Bi-LSTM在无序邻居集上运行，这受GraphSAGE [7]的启发。

3.3.2类型组合

上一步生成| OV |（OV：图中的节点类型集）个节点v的聚合嵌入。为了将这些基于类型的邻居嵌入与v的内容嵌入结合起来，我们采用了注意力[31]机制。其动机是不同类型的邻居将对v的最终表示做出不同的贡献。因此，输出嵌入的公式为：



其中（d：输出嵌入维数），αv，\*表示不同嵌入的重要性，是从3.2节获得的v的内容嵌入，是从第3.3节获得的基于类型的聚合嵌入。我们将嵌入集表示为，并将v的输出嵌入重新表示为：



其中LeakyReLU表示整流线性单元的Leaky形式，是关注参数。图2（c）给出了此步骤的图示。

在此框架中，为了使嵌入维一致并且易于模型调整，我们在第3.2节中使用相同的维d进行内容嵌入，在第3.3节中使用聚合的内容嵌入，而在第3.3节中使用输出节点嵌入。

**3.4目标和模型训练**

为了执行异构图表示学习，我们使用参数Θ定义以下目标：



其中相对于t型上下文节点的集合，例如图中的一阶/二阶邻居[29]或短暂随机游走[20]中的本地邻居。条件概率定义为异构softmax函数：



其中是图中t型节点的集合，是输出节点嵌入,由所提出的具有参数Θ的图神经网络（公式6）表示的。我们利用负采样技术（NS）[19]优化方程（7）中的目标。具体而言，通过将NS应用于等式（8）中softmax函数的构造，我们可以将的对数近似为：



其中M是负样本大小，是关于t型节点的预定义的噪声分布 。在该模型中，我们将M设置为1，因为当M> 1时，它的影响很小。 因此公式（9）退化为交叉熵损失：



换句话说，对于v的每个上下文节点，我们根据对负节点进行采样。因此，我们可以重新公式化公式（7）中的目标如下：



其中，表示通过图上的步行采样收集的三元组的集合。类似于DeepWalk [20]，我们采用随机游走来生成。具体来说，首先，我们在异构图中统一生成一组随机游走S。然后，对于遍历中的每个节点v，我们收集满足以下条件的上下文节点：，即在中距离节点v在τ以内的。最后，对于每个，我们根据对与具有相同类型的节点进行负采样，其中是在S中的频率。为了优化HetGNN的模型参数，在每次迭代时，我们首先在中对三元组进行小批次采样，然后根据等式（11）累积目标。然后，我们通过Adam优化器[11]更新模型参数。我们重复训练迭代，直到两次连续迭代之间的变化足够小为止（有关此训练过程的伪代码，请参见第A.1节）。使用学习到的模型参数，我们可以推断出各种图挖掘任务的节点表示形式，如第4节所示。