基于多质心异质图学习的社交网络用户建模

宁上毅，李冠颖，陈琴，黄增峰，周葆华，魏忠钰
复旦大学
syning18@fudan.edu.cn, guanyingli0705@hotmail.com, qin_chen@fudan.edu.cn,
huangzf@fudan.edu.cn, zhoubaohua@yeah.net, zywei@fudan.edu.cn

摘要

用户建模已经引起了学术界和工业界的广泛关注。现有的方法大多侧重于将用户间的人际关系融入社区，而用户生成的内容（如帖子）却没有得到很好的研究。在本文中，我们通过实际舆情传播相关的分析表明，在舆情传播过程中对用户属性进行研究的重要作用，并且提出了用户资料数据的筛选方法。同时，我们提出了一种通过异构多质心图池为用户捕获更多不同社区特征的建模。我们首先构造了一个由用户和关键字组成的异质图，并在其上采用了一个异质图神经网络。为了方便用户建模的图表示，提出了一种多质心图池化机制，将多质心的集群特征融入到表示学习中。在三个基准数据集上的大量实验表明了该方法的有效性。

关键词： 图神经网络 ； 社交网络分析 ； 用户建模

User Representation Learning based on Multi-centroid Heterogeneous Graph Neural Networks

Shangyi Ning, Guanying Li, Qin Chen, Zengfeng Huang, Zhongyu Wei
Fudan University
syning18@fudan.edu.cn, guanyingli0705@hotmail.com, qin_chen@fudan.edu.cn,
huangzf@fudan.edu.cn, zhoubaohua@yeah.net, zywei@fudan.edu.cn

Abstract

User modeling has attracted great attention in both academia and industry. Most of the existing approaches focus on incorporating the personal relationships in communities, while the users' generated content such as posts is not well studied. In this paper, through the analysis of the actual public opinion dissemination, we show that the research on user attributes plays an important role in the process of public opinion dissemination, and propose the screening method of user data. Meanwhile, we propose an approach to capture more diverse community characteristics via heterogeneous multi-centroid graph pooling for user modeling.Specifically, we first construct a heterogeneous graph where the nodes consist of both users and keywords and adopt a heterogeneous GCN on it. To facilitate the graph representation for user modeling, we then propose a multi-centroid graph pooling mechanism,which incorporates the affiliated group features with multiple centroids into representation learning. Extensive experiments on three benchmark datasets show the effectiveness of our proposed approach.

Keywords: Graph Neural Networks , Social Network Analysis , User Modeling

第二十届中国计算语言学大会论文集，第825页-第836页，呼和浩特，中国，2021年8月13日至15日。

1 引言

用户建模以其巨大的优势在学术界和工业界都得到了广泛的关注，具有很高的研究价值。用户建模旨在预测用户的属性，如性别、年龄、教育程度、地区、职业、收入等，这些属性可以用来帮助描绘电子商务，社交网络等各种应用中的潜在用户的特征。

以往的研究主要集中在从用户个人资料和发表内容内容中挖掘有效特征以进行用户建模。Diehl (2019) 利用用户特征，例如用户的粉丝数量和关注数量，以及文本特征，例如用户发文中的话题标签比例等来对用户的特征进行预测。Hasanuzzaman等人(2017)用卷积神经网络（CNN(Krizhevsky et al., 2012)）来检测时间特征，然后用它来预测用户的收入。此外，Gu等人(Gu et al., 2018)还提出了一种心理学方法，将语言风格、自我描述标签、表情符号使用等人口学信息结合起来进行用户个性预测。

尽管上述方法在用户建模方面取得了很大的成功，但它们并没有充分利用对用户特征和发表内容之间的全局语义关系，而这些内容以及他们之间的联系在社交网络分析中起着重要的作用。为了解决这一问题，最近的一些研究转向为用户建模建立各种特征之间的关系。Kanavos和Livieris(2020)建立基于关注关系的用户图，计算用户在社交平台中的影响力。类似地，BACHA和Thi Zin(2018)利用用户-文本二部图来识别有影响力的用户，而不是建模单独的用户或文本的关系，Li等人(2020)构建了一个包含用户图、文本图和用户文本交互图的耦合网络，有效地评价了用户和文本之间的可信度。然而，这些方法仍然是基于粗粒度的浅层关系，而细粒度的关系，如用户内部的社区关系和用户间讨论的主题，还没有得到很好的研究。

如图 1 所示，社交网络平台中有许多用户组成的社区，具有相似兴趣或背景的用户会自发地加入这些社区，他们发送的内容也通常与某些主题相关。在本文中，我们提出了一种异构图卷积网络（HGCN），通过对用户节点和关键词节点之间的异质信息来学习用户和关键词节点的节点表示。我们的方法与单独对用户或者关键词进行建模相比，增强了不同类型节点之间的信息交换。为了获取这些群体的特征来帮助用户建模，我们进一步设计了一种包含多质心图池化和图解耦操作的图上学习的扩充方法。具体来说，多质心图池化模型致力于自动生成一系列节点集群（即社区），如用户组成的社区和关键词组成的主题。图解耦的目的是通过将原始的节点表示与相应的集群中心信息相结合，来获取集群内部的共同特征。最终，这一融合的表示可以被用来进行用户属性预测或用户节点分类任务。

我们在国内社交媒体数据，即微博上进行了广泛的实验。实验结果表明，我们提出的用户建模方法在预测教育和职业等用户属性方面也明显优于现有的基线方法。此外，我们将我们的节点分类方法扩展到另外两个异质数据集，即DBLP和IMDB，这也验证了我们方法的有效性。本文的主要贡献如下：

- 提出了一种异质图卷积网络，方便了嵌入学习过程中不同类型节点之间的信息交互；
- 提出了一种多质心增强方法，它能够自动吸收用户群体和关键词主题等社会群体的共同特征，增强节点的表示学习；
- 我们在微博数据集和其他两个异构数据集上对实验结果进行了详细的分析，从而更好地解释了我们所提出的方法的有效性。

2 相关工作

2.1 图神经网络

近年来，图神经网络引起了人们越来越多的兴趣。Bruna等人(2013)提出了一种谱图卷积网络，它从连通图中学习卷积层。ChebNet(Defferrard et al., 2016)利用切比雪夫多项式来逼近谱图卷积。Kipf和Welling(2016)对其进行了进一步简化，提出了基于谱图卷积的半监督节点分类和图分类任务的图卷积网络（GCN）。GCN体系结构在节点分类、图分类和推荐等图的研究中取得了很大的进展。其他基于GCN模型的工作如(Li et al., 2018a)、(Chiang et al., 2019)，在节点分类任务上对其进行了改进或简化。关于图神经网络也有一些综合性的综述文章，例如(Zhou et al., 2018)和(Wu et al., 2019)。

异构网络考虑不同类型节点和边的任务。由于异构网络的特殊性，传统的图模型不能直接应用。(Shi et al., 2015)对异构信息网络进行了全面的调研。许多工作将传统的分类扩展到

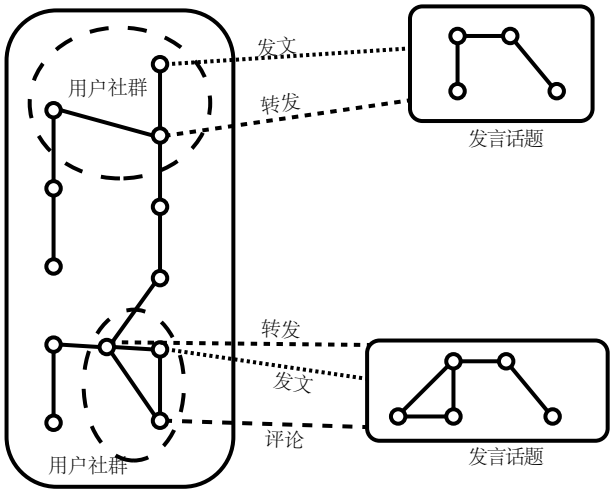


图 1: 社交网络中的用户与言论关系。图中左侧的节点关系表示用户之间的人际关系，相似的用户组成了用户社群；右侧的节点表示用户的言论关键词，不同的关键词组成了不同的话题。用户通过发文、转发和评论与话题词之间产生关系。

异构信息网络。一些工作扩展了传导性学习分类任务，即为给定的未标记数据预测标签。例如，Shi等人(2010a)提出对具有任意网络模式和任意数量的节点或连接类型的信息网络中的链接结构进行建模的方法。Ji等人(2010b)提出通过一种新的元路径选择模型，利用异质网络上的小标记数据进行聚类，Ji(2014)提出了一种通过计算空间中节点的潜在表示来标记不同类型节点的方法，其中两个连接的节点往往具有紧密的潜在表示。一些工作还扩展了归纳分类，即在整个数据空间中构造一个决策函数。例如，Jacob等人(2012)使用无组分异构网络来表示文本文档集合，并提出IMBHN算法来归纳为文本术语分配权重的分类模型。

图数据的池操作主要包括两类：节点聚类和节点采样。Ying等人(2018)通过将节点聚合成更大的超级节点来实现图池化，通过学习分配矩阵，以指定的概率将每个节点软分配到新图中的不同子图中。每个子类下的池化操作保留了节点的所有表示信息，并在新的子图继续进行表示学习。这类方法的一个问题是，子图的训练可能带来过拟合的问题，因为图中边上的权值表示两个节点之间的连通强度。新图中的连通强度可能与原始图中的连通模式有很大的不同。

节点采样方法主要是选取一个固定的节点数目，来形成一个新的子图。在(Zhang et al., 2018)中，模型对每个节点的相同特征进行排序，并选择该特征中具有最大值的k个节点来形成新图。Gao和Ji(2019)使用一个可训练的投影向量，将节点的特征向量投影成标量值，生成排序得分，选择标量值最大的k个节点，形成排序后的子图。

2.2 用户建模

用户建模是建立用户模型的过程，在该过程中，有关用户的不可观察信息是从该用户的可观察信息，例如用户与系统的交互中推断出来的(Zukerman and Albrecht, 2001)。用户模型可以使用用户引导的方法创建，在这之中模型是使用每个用户提供的信息直接创建的；或者使用自动方法创建，其中创建用户模型的过程是由系统控制，并且对用户隐藏的。用户引导方法产生适应性服务和适应性用户模型(Fink et al., 1997)，而自动方法产生获得性服务和获得性用户模型(Brusilovsky and Schwarz, 1997)。一般来说，用户模型将包含一些自适应和自适应元素。理想情况下，适应性因素的集合应该尽可能地减少（如年龄、性别、喜欢的背景颜色等），而其他因素（喜欢的话题、行为模式等）应该通过学习过程来创造。这些概念在文献中也被称为隐式用户模型和显式用户模型习得(Quiroga and Mostafa, 1999)。

用户建模，包括教育和职业分类，被社会学家和数据科学家广泛研究。Gu等人(2018)运用皮尔逊相关度分析法对微博用户的五大人格特征进行了研究。Preoţiuc-Pietro等人(2015)公布了一个推特数据集，其中包含了拥有职位信息和历史推文的推特用户。Hasanuzzaman等人(2017)用时间方向的分类方法研究。现有的社会网络论文缺乏对用户话语和社会网络关系的挖掘。其中大多数只是数据分析和分类，或者高度相关标签的集成。

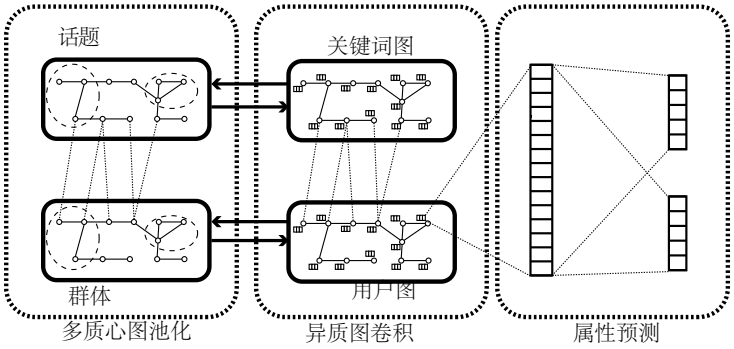


图 2: 本文提出的基于异质图网络的用户建模模型。图中的图网络提取自是从社交网络数据。从图网络中学习到的向量表示可以用于进行用户的教育和职业分类。

3 基于异质图网络的用户建模

在本章中，我们将介绍我们提出的使用多质心池化图模型进行用户建模的方法，该方法使用异构质心池化图模型对用户的社交网络关系和发表的内容进行建模。在第 3.1 节中，我们将介绍如何构造一个包含两种类型节点的异构图，其中两种节点分别代表用户和关键词。在第 3.2 节中，我们提出了异质图卷积网络（HGCN）来对图中节点进行表示学习，并且考虑了社交网络中的聚类效应，将多质心图池化模型应用于图中的聚类节点。模型的总体框架如图 2所示，其中包括异质图卷积、多质心图池化和属性预测三个模块。

3.1 异质网络构建

为了预测用户的受教育程度和职业类别，我们构建了一个包含用户关注关系和历史数据的异构图网络。网络中的节点由用户节点和关键词节点构成，因此存在两种不同类型的节点。在网络中，两种不同类型的节点形成了三个不同的子图。我们需要构造的这样的三个子图，即用户图、关键词图以及用户与关键词之间的二部图。用户图是最容易构造的，每个节点代表了一个用户，如果一个用户在社交网络中关注了数据集中的另一个用户，那么它们之间就会有一条边。实验结果表明，对称无向图的性能是优于有向图的，因此我们没有使用有向的关注关系，而是如果两个用户是互相关注的，那么他们之间边的权重会被设置为2。关键词图和二部图中引入了一种新的节点，即关键词节点。关键词节点是从用户的历史推文中提取出来的。我们使用中文分词框架jieba对所有用户的历史微博进行了分词操作，找出了前10, 000个最常用词。这些词被设置为关键词的节点。在关键词图中，边的权重被定义为两个词在一条社交网络文本中同时出现的次数。在用户-关键词二部图中，权重被定义为用户在发布的内容中提及这些关键字的次数。在IMDB和DBLP数据集中，我们使用类似的方法构造图结构。如果用户与同一篇论文或是同一部电影有联系，那么用户和论文、用户和电影之间就会有边相连。在DBLP数据集中，具有相同关键词的论文被认为是相邻的。在IMDB数据集中，具有相同导演的电影是相邻的。

3.2 异质图上的表示学习

为了处理异质图网络，我们提出了一种异构图卷积网络（HGCN），它能同时捕获用户级别和文本级别的特征。

Kipf和Welling(2016)提出了图卷积网络的逐层传播规则:

$$H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)}\right) \tag{1}$$

这一规则可表示为

$$H^{(l+1)} = \sigma\left(\left(\sum_{u\in N(v)\cup v}\hat{A}_{v,u}x_u\right)W^{(l)}\right). \tag{2}$$

这里，$\hat{A}$是具有附加自连接的图的归一化邻接矩阵，也就是

$$\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}. \tag{3}$$

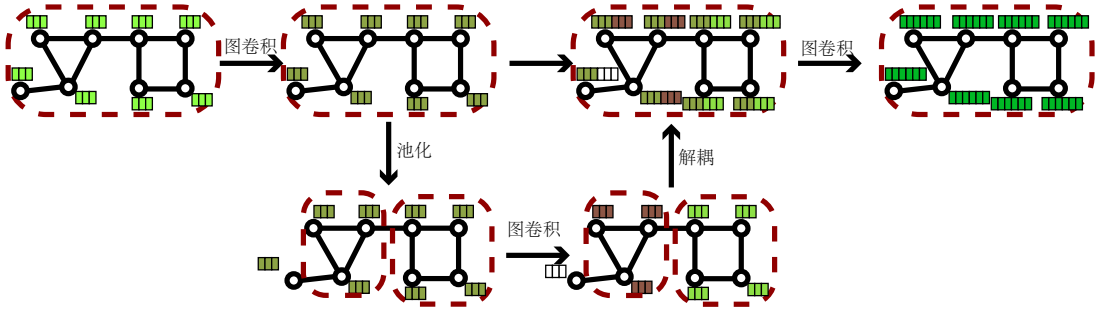


图 3: 我们提出的多质心异质图节点表示模型。图中的颜色变化表示节点的表示发生了改变。

其中，$\tilde{A} = A + I_N$。I_N是N维单位阵，$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$。需要注意的是，变换$\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$规定了$\hat{A}$任何行或列的总和为1。这可以看作是一种归一化变换。也就是说，$\hat{A}$是由自循环矩阵$\tilde{A}$归一化得到的，这与PageRank随机游走的邻接矩阵形式类似。

Duan等人(2012)提出了一种异质社交网络图上的主题摘要模型。如果仅考虑用户级别和内容级别信息，则个性化页面层可表示为：

$$X^{(l+1)} = \alpha_1 \tilde{A} X^{(l)} + \beta_1 \tilde{C} Y^{(l)} + \gamma_1 \tilde{I}_N X^{(l)}, \tag{4}$$

$$Y^{(l+1)} = \alpha_2 \tilde{B} Y^{(l)} + \beta_2 \tilde{D} X^{(l)} + \gamma_2 \tilde{I}_M Y^{(l)}, \tag{5}$$

其中，A, B是任意两个用户或关键词之间边的权重矩阵，C, D分别是从关键次到用户和从用户到关键次的边的权重矩阵，$\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}$是$A, B, C, D$进行归一化后得到的矩阵。$\tilde{I}_N$和$\tilde{I}_M$和是个性化矩阵，其中包含来自目标节点的消息，这些矩阵也是归一化后得到的。$X^{(l)}, Y^{(l)}$是第l层的节点表示。$\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2$是和权重相关的超参数。

假设权重矩阵为归一化邻接矩阵，个性化矩阵为单位矩阵且$\alpha_1 = \gamma_1$，公式4变为

$$X^{(l+1)} = \alpha_1 \hat{A} X^{(l)} + \beta_1 \tilde{C} Y^{(l)}, \tag{6}$$

同样地，在类似的假设下，对于内容级别表示的更新，我们有

$$Y^{(l+1)} = \alpha_2 \hat{B} Y^{(l)} + \beta_2 \tilde{D} X^{(l)}. \tag{7}$$

因为在我们的图中，用户和关键字之间的边是对称的，所以C等于D的转置。

在我们的异质图中，聚合层如公式6-7所示。将它们与一个线性层结合，我们可以得到具有两种节点的异质图的图卷积层。

$$X^{(l+1)} = \sigma\left(\left(\alpha_1 \hat{A} X^{(l)} + \beta_1 \tilde{C} Y^{(l)}\right) W_1^{(l)}\right), \tag{8}$$

$$Y^{(l+1)} = \sigma\left(\left(\alpha_2 \hat{B} Y^{(l)} + \beta_2 \tilde{D} X^{(l)}\right) W_2^{(l)}\right). \tag{9}$$

为了使用户节点和关键字节点的表示携带相同的信息，我们要求所有层中的$W_1 = W_2$为简化计算，公式 8-9可表示为

$$\begin{pmatrix} X^{(l+1)} \\ Y^{(l+1)} \end{pmatrix} = \sigma\left(\begin{pmatrix} \alpha_1 \hat{A} & \beta_1 \tilde{C} \\ \beta_2 \tilde{D} & \alpha_2 \hat{B} \end{pmatrix} \begin{pmatrix} X^{(l)} \\ Y^{(l)} \end{pmatrix} W^{(l)}\right). \tag{10}$$

因此，图卷积层可以在异质图网络上通过一个改进的矩阵归一化来实现。

3.3 异质图上的多质心图池化模型

在本节中，我们将介绍基于图池化和图解耦(Gao and Ji, 2019)操作的多质心图池化机制，并将其扩展到异质图上，如图 3 所示。Gao和Ji(2019)介绍了图池化（gPool）和图解耦（gUnpool）层的工作机制。在图池化层，我们会从原始图中选择一个子图，并且希望所选择

的子图中的节点能尽可能地代表原始节点。在这里，我们需要一个选择向量来表示对整张图的中间位置进行表示，并计算图中所有节点的投影，

$$y_i = \mathbf{X}_i\mathbf{p}/\|\mathbf{p}\|. \tag{11}$$

我们贪心地选择最大的y_i，相应的节点就是子图中选定的节点。将第l层所选节点表示为$V^{(l)}$，子图定义就是

$$G^{(l+1)} = \left(V^{(l+1)}, E^{(l+1)}\right), \tag{12}$$

$$E^{(l+1)} = \bigcup_{v_i, v_j \in V^{(l)}} (v_i, v_j) \in E^{(l)}. \tag{13}$$

在图池化操作之后，我们会把图卷积层应用于所选的子图，将得到子图中每个节点的新的表示。之后的图解耦操作将这些新的表示释放回原始的图中。在图解耦层，我们将所选节点的新表示与前一个层的表示相连接。在这种操作之后，表示向量的长度会被改变，因此我们需要调节那些未选择的节点的表示。Gao和Ji(2019)提出的图上的单元网络是由多个图池化层和图解耦层组成的。在原始实验环境下，选取子图中的重要节点进行放大表示。在实验中，我们发现未选择的节点往往具有相同的标签，换句话说，图池化和图解耦的重要性采样在标签分布上是不平衡的。从这个角度出发，我们希望用类似的方法对节点进行聚类。

在本节中，我们将介绍用于节点集合的多质心池化机制。由于图上的单元网络子图的节点是不平衡的，我们希望每个子图都能代表一个数据集中的标签类。因此，我们需要几个不同的子图。假设有m个子图，则每个子图都分配有一个中心向量，$\mathbf{p}_i, i = 1, 2, \ldots, m$。如公式 11-13，我们将子图被提取出来。对于第n个子图图，我们计算

$$y_i^{(n)} = \frac{\mathbf{X}_i\mathbf{p}_n}{\|\mathbf{p}_n\|}, v_i \notin V_j, j < n, \tag{14}$$

然后选择其中前k_n大的$y_i^{(n)}$，他们对应的节点就是我们选定的第n个子图中的节点。

图解耦的操作与(Gao and Ji, 2019)中的类似。在同一层生成的所有表示都分配在节点表示的同一位置。例如，如图 3所示，图中要提取两个子图要，也就是$m = 2$。我们设$k_1 = 4, k_2 = 3$。在第一个子图中，我们选择并提取右侧的四个节点。在第二个子图中，提取中间的三个节点。在图池化操作之后，子图中提取的节点通过图卷积层表示。它们的输入是原始图上图卷积层输出的表示，并输出一个相同长度的向量。在图解耦步骤中，新的输出表示被附加回原始向量的后面，这样一来，节点的表示向量长度比以前长一倍。另外，图中未被选定节点的表示会用零填充。

在这一节中，我们将介绍我们在异质图上的多质心图池化的优化。由于我们的任务是一个只在用户节点上有标签的半监督节点分类问题，所以在用户网络和网络上应该分别考虑池和图卷积关键字网络。在图池化操作中，我们分别从用户网络和关键词网络中提取节点。当我们从用户网络中提取节点时，关键词网络中的所有节点都被保留。提取这些节点后，子图上的图卷积层将更新提取的用户节点和所有关键字节点。

在图解耦操作期间，只有用户节点会被连接回来。另外，我们会同时提取关键字节点。在这个子图中，关键词节点的表示被更新并附加到表示向量。因为真实值只标注在用户向量上，我们可以使用一些技巧可以减少训练的层数。例如，在提取用户节点的子图中，关键词节点的表示不需要在最后一层更新。在提取关键词节点的子图中，只有有真实标签的节点才需要更新，从而能完成反向传播。

3.4 特征预测

我们提出的框架以三个图的邻接矩阵为输入，输出用户节点的分类标签。在每一层中，都有一个多质心的图池化和图解耦，抽出的子图用图卷积网络来进行表示。在所有这些子图都被解耦之后，我们会得到一个得到最终表示的图卷积层。我们使用多任务分类器，用逻辑回归和

交叉熵完成职业和教育分类，

$$p(w_j) = \frac{1}{n}\sum_{j=1}^{n} s_j \tag{15}$$

$$p(e_j) = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{16}$$

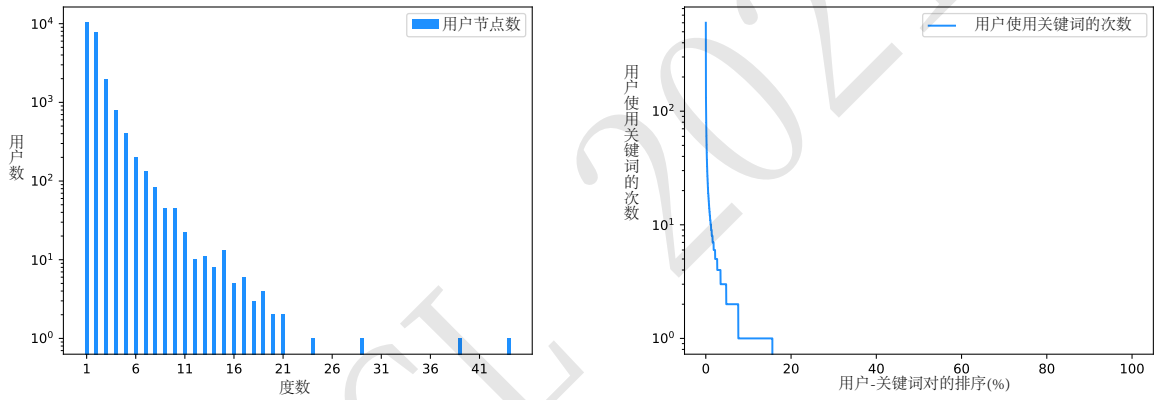
其中，s_j, t_i 是职业和教育分类的输出分别。另外，最终的损失函数中还有一个关于子图的损失，使同一类型的节点尽量出现在同一个子图中。最终的损失函数定义为

$$L = -\sum_{i,j}(I(j = W_i)\log p(w_j)) - \sum_{i,k}(I(k = E_i)\log p(e_k)) - \sum_{l}\frac{1}{n}\|x_l - p\|, \tag{17}$$

其中W_i, E_i 是教育和职业分类的真实值，$p(w_k), p(e_k)$是教育和职业分类各标签的预测概率。x_i是第i个节点的表示向量，n子图中的节点数量，p是聚类中心的中心向量。

4 社交网络中的用户数据

本文中，我们使用了从新浪微博上抓取的用户关系与用户言论数据来构建异质网络。



(a) 固定第二层对第一层聚类中心数的影响。不同形状的线表示第二层中使用的不同数量的聚类中心。

(b) 用户发表关键词的频率，关键词是根据词频进行排序的。

图 4: 微博数据集中节点的分布

我们在新浪微博上搜集了超过10万名用户的用户信息、历史微博和关注信息关系数据集。我们在2018年10月爬取了这些数据，其中微博的范围覆盖了2009年至2018年之间。经过一些预处理后，我们删除了其中一些信息不完整或发文太少的微博用户。最后，我们的数据集中共有35830个用户。根据用户自己填写并提交的信息，我们将他们的教育和职业划分为不同的类别。

本文主要从教育程度和职业类别两个方面进行研究。对于教育程度分类，我们使用用户最终教育大学的平均录取分数。需要注意的是，在预处理步骤中，仍然保留在数据集中的所有用户都填写了他们的本科学校信息。我们手动将录取分数分成不同的类别，形成了不同的教育水平的分类。对于职业分类，我们使用了用户引用的公司和工作岗位信息。我们使用预先训练好的中文词向量工具(Li et al., 2018b) 来生成用户职业表示的向量表示。然后我们使用KMeans聚类将用户分为五个不同的标签。通过对集群中心的观察，我们将这些集群命名为咨询、艺术、管理、科学和文化，以反映这些集群中用户的职业。表2显示了用户排序的一些统计分析数据集。

用户节点之间的连接是通过用户之间的关注关系来实现的，因此不同用户之间的连接边数是不同的，这是一个值得研究的问题。图 4(a)显示了用户图中用户节点的度分布。在用户图

中，节点的边由用户的关注关系决定。如果用户A关注了用户B，那么A和B之间会有一条连变优势，反之亦然。因此，一个节点的度数表示用户图中的关注这一用户的用户数量。图 4(b)显示大多数用户在网络中的关注和粉丝之和少于10人。直观地说，在进行节点表示时，用户的节点很容易被他所关注的或关注他的人所影响。

5 实验结果

5.1 实验设置

在本文中，我们主要在微博数据集上进行了实验，另外，我们还将该模型与其他几种异构数据集上的一些最新模型进行了比较，如DBLP和IMDB数据集。为了与其他节点分类方法进行比较，我们对常见的异构图进行了实验，包括DBLP和IMDB数据集。对于这两个数据集，我们遵循了(Wang et al., 2019)中的数据集设置。

对于异质图节点分类，我们采用了2层多质心图池化框架，其中第一层用户节点包含6个池化中心，比例为[0.4,0.25,0.15,0.1,0.05,0.05]。在这一步中，我们将第一层的比例调整范围设为0.05，池化中心数从3*3到8*8不等。第二层包含8个池化层，每个池层的大小相同。两层关键字节点分别包含4个和5个池化中心，每个子图的大小相同。初始GCN的输入输出维数为32，因此最终输出维数为32×3=96。在第一个多质心图化池层中，我们在池的开始和结束使用了两层GCN框架，在其他层中，每个池层上只有一层GCN。我们在模型中使用ReLU作为激活函数，在GCN的最后一层使用Sigmoid作为激活函数。我们使用Adam(Kingma and Ba, 2014)优化器训练模型，学习率为0.01，每10^4次迭代递减因子为0.9，随机失活率设置为0.1。对于DBLP和IMDB数据集，我们使用了一个2层多质心图池框架，其中池质心的结构从3*3到7*7不等。在验证集上的实验表明，3*5和5*5的结构性能最好。所有这些子图都具有相同数量的节点。初始GCN的输入和输出维度在DBLP中为16，在IMDB中为32。学习率分别为0.01和0.05。其他超参数与微博异质图相同。

用于实验对比的模型有

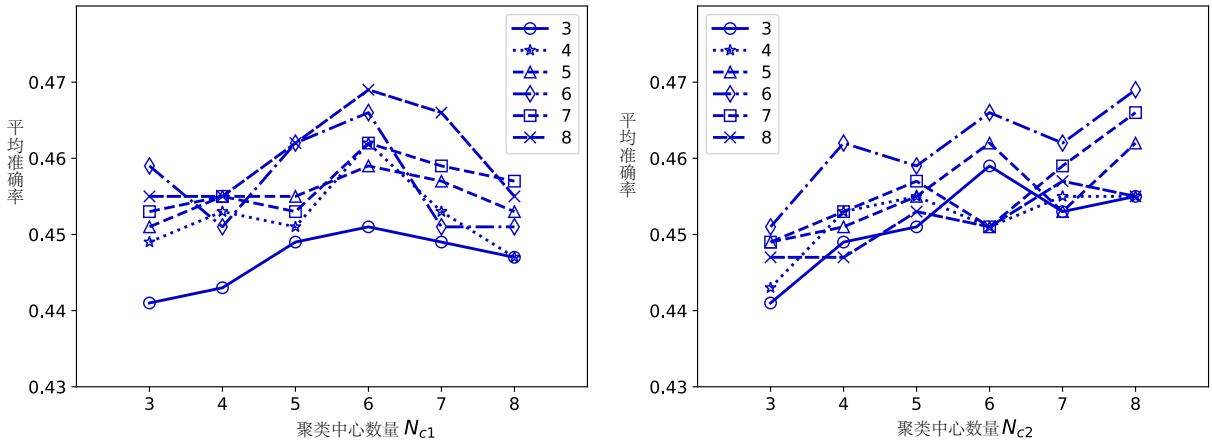
- **GCN**是(Kipf and Welling, 2016)中提出的图神经网络模型。
- **HAN**是(Wang et al., 2019)中提出的异质图神经网络模型。
- **GCN+MUCA**是(Kipf and Welling, 2016)中的神经网络模型与多质心图池化模型的结合，没有使用异质网络的信息。
- **HGCN**是本文提出的异质网络信息传递机制。
- **HGCN+MUCA**是本文提出的多质心图池化的异质网络。

5.2 总体实验结果

模型	教育	职业	平均
HAN	0.374	0.483	0.424
GCN	0.383	0.493	0.433
GCN + MUCA	0.393	0.501	0.442
HGCN	0.409	0.519	0.459
HGCN + MUCA	**0.410**	**0.528**	**0.469**

表 1: 微博数据集上职业和教育分类的准确度。

我们报告了微博数据集中不同模型的用户教育和职业分类模型的结果。评价指标为分类准确度，结果见表 1。结果表明，我们提出的HGCN模型比目前最优的图神经网络（如HAN和GCN）的性能提高了8.24%，这说明我们在节点表示学习过程中融合异构信息的有效性。此外，当我们将多中心聚类整合到模型中时，性能可以进一步提高。具体而言，GCN+MUCA和HGCN+MUCA模型的平均准确度分别比GCN和HGCN提高了0.01左右。这一结果验证了我们的假设，也就是说将社会社区特征建模为多个中心有助于增强对用户的建模。



(a) 固定第二层对第一层聚类中心数的影响。不同形状的线表示第二层中使用的不同数量的聚类中心。

(b) 固定第一层对第二层聚类中心数的影响。不同形状的线表示第一层中使用的不同数量的聚类中心。

图 5: 多中心聚类方法中的质心数目的影响。N_{c1} 和N_{c2} 分别表示第一层和第二层的聚类中心数量。

模型	IMDB	DBLP
HAN	0.576	0.902
GCN	0.554	0.882
GCN + MUCA	0.568	0.895
HGCN	0.582	0.902
HGCN + MUCA	**0.596**	**0.910**

表 2: IMDB 和DBLP数据集上的准确度。

如表1所示，我们验证了我们提出的方法对于用户节点建模是有效的。由于我们的异构图可以包含其他类型的节点，因此我们将我们的方法扩展到其他节点分类任务中，以便进行更全面的评估。我们分别为DBLP构造了一个以研究者和论文节点为节点的异构图，为IMDB分别构造了一个以用户和电影为节点的异构图。在这两个数据集中，我们分别对研究者的研究领域和电影类别进行分类，结果见表 2。我们在这组实验中观察到了相似的结果，我们的HGCN模型和多中心聚类方法都有助于增强节点分类，这表明我们的方法对于不同类型的节点都是有效的。

5.3 消融实验

图	教育	工作	平均
$g_u + g_w + g_b$	**0.410**	**0.528**	**0.469**
- g_u	0.282	0.460	0.371
- g_w	0.388	0.496	0.442
- g_u - g_w	0.266	0.433	0.349
- g_w - g_b	0.361	0.477	0.408

表 3: 微博数据集中不同子图的作用。g_u, g_w 和g_b 分别表示用户图，关键词图和二部图。

到为了进一步研究我们的异质图对用户建模的有效性，我们进行了图上的消融实验，结果如表 3所示。我们观察到，用户图在用户建模中起着最重要的作用，当从完全异构图中迁移时，平均准确率下降了20.9%。关键词图和二部图也有助于提高教育和职业分类的准确性。因此，在实际操作中，我们有必要构造一个包含综合三类图的信息的异质图来进行用户建模。

6 实验分析

为了进一步研究我们的多质心异质图模型，我们在微博数据集上进行了进一步的实验，研

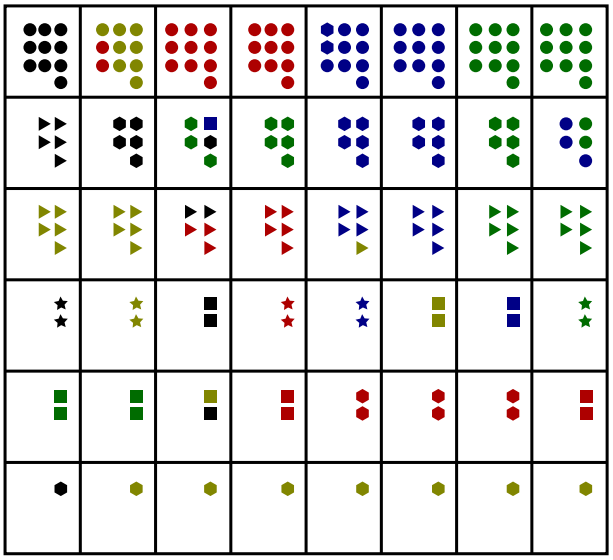


图 6: 多中心聚类中的用户群体。每一行代表第一层多中心聚类中的一个簇，每一个块代表最后一层的簇。不同形状的符号代表不同的职业类别，不同的颜色代表不同的教育程度。

究聚类中心数量对结果的影响，并对聚类中心分类效果进行可视化。

6.1 聚类中心数量的影响

在这一节中，我们研究了用于多中心聚类方法中的质心数目的影响。我们的实验中使用了两个多中心聚类层，第一层和第二层的质心数从3到8不等。图 5用不同的质心数绘制实验结果。我们观察到，在大多数情况下，当第一层的质心数量增加时，性能会提高，如图 5(a)所示。具体地说，当第一层（N_{c_1}）的质心数增加到6时，性能最佳。考虑到第二层（N_{c_2}）质心数的影响，性能随N_{c_2}先增大后下降，如图 5(b)所示。当N_{c_2}设置为8时，将获得最佳性能。总的来说，在我们的实验中，第一层和第二层的质心数建议设置为6和8作为一个默认的设置。

6.2 聚类中心可视化

图 6显示了多中心聚类中的用户群体。不同形状的符号代表不同的职业类别，不同的颜色表示不同的教育程度。我们观察到，在第一层中代表一个簇的每一行主要包含同一个符号形状，它对应于一个职业类别。此外，在最后一层中代表一个群集的每个块都有几乎相同的符号颜色，这与教育程度一致。也就是说，在多中心聚类过程中生成的用户集群可能会反映某个用户属性，这表明了通过多个聚类中心来建模用户组特征对于提升用户建模效果的有效性。

7 总结与展望

在本文中，我们提出了一个异质多质心图模型用于进行社交网络上的用户建模。我们首先构造了一个由用户图、关键词图和用户-关键词二部图组成的异质图。为了方便不同类型节点之间的信息交互，我们提出了一种异构图形进化网络用于嵌入学习。此外，我们还设计了一种多质心图池化方法，该方法通过允许节点获得其所属子图的质心信息来捕获所属集群的特征。在三个数据集上的实验结果表明了我们提出的方法的有效性。通过对不同子图的分析发现，对于用户节点建模任务来说，用户之间的子图比其他子图在建模中起着更重要的作用。此外，我们对多质心图池化模型进行了详细的分析，解释了我们提出的方法的有效性。

今后，我们将从两个方向继续研究。首先，我们将致力于用户建模的数据集构建，从而包含更多我们感兴趣的用户属性。第二，我们可以研究更多的方法来构建用户建模的社区特征。

参考文献

R. E. BACHA and T. Thi Zin. 2018. Ranking of influential users based on user-tweet bipartite graph. In *2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pages 97–101.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Peter Brusilovsky and Elmar Schwarz. 1997. User as student: Towards an adaptive interface for advanced web-based applications. In *User Modeling*, pages 177–188. Springer.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. 2019. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. *arXiv preprint arXiv:1905.07953*.

Michal Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering.

Frederik Diehl. 2019. Edge contraction pooling for graph neural networks. *arXiv preprint arXiv:1905.10990*.

Yajuan Duan, Zhumin Chen, Furu Wei, Ming Zhou, and Heung-Yeung Shum. 2012. Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of COLING 2012*, pages 763–780.

Josef Fink, Alfred Kobsa, and Andreas Nill. 1997. Adaptable and adaptive information access for all users, including the disabled and the elderly. In *User Modeling*, pages 171–173. Springer.

Hongyang Gao and Shuiwang Ji. 2019. Graph u-nets. *CoRR*, abs/1905.05178.

Haiqian Gu, Jie Wang, Ziwen Wang, Bojin Zhuang, and Fei Su. 2018. Modeling of user portrait through social media. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Mohammed Hasanuzzaman, Sabyasachi Kamila, Mandeep Kaur, Sriparna Saha, and Asif Ekbal. 2017. Temporal orientation of tweets for predicting income of users. Association for Computational Linguistics (ACL).

Yann Jacob, Ludovic Denoyer, and Patrick Gallinari. 2014. Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 373–382.

Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. 2010a. Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer.

Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. 2010b. Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer.

Andreas Kanavos and Ioannis E Livieris. 2020. Fuzzy information diffusion in twitter by considering user's influence. *International Journal on Artificial Intelligence Tools*, 29(02):2040003.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2018a. Adaptive graph convolutional neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018b. Analogical reasoning on chinese morphological and semantic relations. *CoRR*, abs/1805.06504.

Peiyao Li, Weiliang Zhao, Jian Yang, Quan Z Sheng, and Jia Wu. 2020. Let's corank: trust of users and tweets on social networks. *World Wide Web*, pages 1–25.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

Luz Marina Quiroga and Javed Mostafa. 1999. Empirical evaluation of explicit versus implicit acquisition of user profiles in information filtering systems. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 238–239.

Rafael Geraldeli Rossi, Thiago de Paulo Faleiros, Alneu de Andrade Lopes, and Solange Oliveira Rezende. 2012. Inductive model generation for text categorization using a bipartite heterogeneous network. In *2012 IEEE 12th international conference on data mining*, pages 1086–1091. IEEE.

Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. 2015. A survey of heterogeneous information network analysis. *CoRR*, abs/1511.04854.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, Philip S. Yu, and Yanfang Ye. 2019. Heterogeneous graph attention network. *CoRR*, abs/1903.07293.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2019. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. 2018. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, pages 4800–4810.

Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. 2018. An end-to-end deep learning architecture for graph classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.

Ingrid Zukerman and David W Albrecht. 2001. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11(1):5–18.