基于精细异质信息网络表示学习的文献推荐研究*

李琳娜[1,2]　郭晓琪[1]　张运良[1,2]　王力[1,2]　张晓丹[1]
（1. 中国科学技术信息研究所，北京 100038；2. 富媒体数字出版内容组织与知识服务重点实验室，北京 100038）

摘要：目前针对异质信息网络的文献推荐主要基于已有的文献间引用、共同作者、同期刊发表等信息构建文献之间的关联关系，没有从文献的研究问题、方法模型等研究人员更关心的文献细粒度内容角度构建文献间关系，不能将这些细粒度的关联关系融入文献推荐过程，从而影响最终的推荐效果。将文献的研究问题、方法模型两类细粒度标签加入学术信息网络，并提出了文献推荐模型PRM-FHIN，通过对异质信息网络表示学习方法HECO进行优化学习网络节点的结构向量，通过对SciBERT模型进行微调学习网络节点的内容向量，基于融合的内容向量和结构向量实现最终的文献推荐。从开放学术图谱中抽取2010—2020年计算机科学领域的185万篇论文作为实验数据，实验结果表明：优化后的HECO算法能更好地对网络节点进行嵌入式表示，在异质信息网络中融入研究问题和方法模型等细粒度标签可以丰富文献之间的语义信息，从而提高最终的文献推荐效果。

关键词：精细异质信息网络；异质信息网络表示学习；文献推荐；细粒度标签；计算机科学；研究问题；方法模型

中图分类号：G254　DOI：10.3772/j.issn.1673-2286.2025.05.002

引文格式：李琳娜，郭晓琪，张运良，等. 基于精细异质信息网络表示学习的文献推荐研究[J]. 数字图书馆论坛，2025，21（5）：11-19.

随着科学技术的迅速发展和科技成果的不断涌现，学术资源平台的文献数量迅速增加，海量信息给研究人员带来便利的同时，也让他们陷入信息超载的难题。研究人员在面对海量文献时，越来越难以快速、准确地获取自己想要的信息[1]。如何在大数据环境下高效精准地获取兼具专业性、时效性、权威性的科技文献，已经成为科研人员亟待解决的问题之一[2]。

异质信息网络是包含不同类型节点和关系的网络[3]，学术信息网络是一种典型的异质信息网络，通过利用网络中全面的结构信息和丰富的语义信息，基于学术异质信息网络的文献推荐取得了较好的效果。研究人员在追踪领域相关文献时更关注其他作者采用了哪些模型来解决相应的研究问题，即研究问题、方法模型这些细粒度知识比合著者、发表期刊等信息更能体现文献之间的关联关系。若将这些细粒度的关联关系应用于文献推荐过程，将能较好地提升推荐效果。

1　相关研究

由于同质信息网络在对推荐问题进行建模时没有考虑不同类型对象之间的复杂异质交互关系，大量的信息损失，进而影响最终的推荐效果。为了更加有效地

收稿日期：2024-09-06
*本研究得到中国科学技术信息研究所重点工作项目“面向战略决策的智能情报技术引擎研究及应用”（编号：ZD20250-08）资助。

进行文献推荐，一些研究人员尝试将学术信息网络构建为异质信息网络，利用网络中全面的结构信息和丰富的语义信息得到更好的推荐结果。许侃等[4]通过构建包含作者、论文、主题、会议4种类型节点的学术异质信息网络，采用基于元路径的随机游走学习节点嵌入向量，并将节点嵌入向量融入矩阵分解框架来预测研究人员对文献的评分，最终将评分较高的文献推荐给研究人员。段震等[5]通过构建具有文献、作者、关键词和期刊4种类型节点的学术异质信息网络，采用混合随机游走方法获得节点的嵌入向量，并基于文献节点之间的向量相似度获得文献推荐列表。吴俊超等[6]通过构建包含用户、文献、标签3种类型节点的学术异质信息网络，基于卷积融合文献文本内容信息和网络结构信息，最终向用户推荐其可能感兴趣的文献。Ali等[7]针对文献、作者、主题、关键词4种类型节点之间的关系构建多个加权二分图，学习每个二分图节点间的相似度以生成推荐结果，但该方法只考虑了关系的多样性，没有考虑节点的异质性。Dai等[8]通过构建包含文献、关键词、期刊、作者4种类型节点的学术异质信息网络，构建作者与作者的关联矩阵和文献与文献的关联矩阵，通过矩阵分解实现最终的文献推荐。Ma等[9]构建包含文献、作者、主题、期刊4种类型节点的学术异质信息网络，通过优化两个基于元路径的节点相似度计算方法获得作者和文献的嵌入式表示，最终基于作者和文献的余弦相似度推荐相关文献。Zhu等[10]通过构建包含作者、文献、期刊、机构4种类型节点的学术异质信息网络，分别学习节点的内容向量和结构向量，最终实现融合内容信息和结构信息的文献推荐。这些研究成果主要基于文献的共同作者、文献间的引用关系、文献主题或关键词、文献发表期刊等关联信息构建学术异质信息网络，建立的是较粗粒度的文献间关系。没有从文献的研究问题、方法模型等研究人员更关注的文献细粒度内容角度构建文献间关系，不能将这些细粒度的关联关系融入文献推荐过程，从而影响最终的推荐效果。

2 方法设计

本文提出了基于精细异质信息网络表示学习的文献推荐模型（Paper Recommendation Model Based on Fined-Grained Heterogeneous Information Network Representation Learning，PRM-FHIN）。给定一个包含作者、文献等节点的学术异质信息网络和文献的标题、摘要文本等内容信息，该模型会为不同作者推荐其可能感兴趣的文献列表。PRM-FHIN的整体架构（见图1）由精细异质信息网络构建、文本内容嵌入、网络节点预嵌入、精细异质信息网络表示学习和最终推荐5个部分组成。①精细异质信息网络构建模块旨在从文献摘要中抽取研究问题、方法模型两类标签并进行归一化，进而将这些细粒度标签以新增节点的方式融入异质信息网络。②文本内容嵌入模块旨在根据文献的内容信息生成文献节点和作者节点的内容向量。③网络节点预嵌入模块主要利用基础的异质信息网络表示学习方法Metapath2vec对学术网络中的每个节点进行向量初始化，以加快后续精细异质信息网络表示学习算法的学习速度。④精细异质信息网络表示学习模块基于正负样本选择策略对HECO算法进行优化。⑤最终推荐模块首先基于文本内容嵌入和精细异质信息网络表示学习获得文献和作者两类节点的内容嵌入和网络嵌入，随后将这两个向量融合，进而计算融合后向量之间

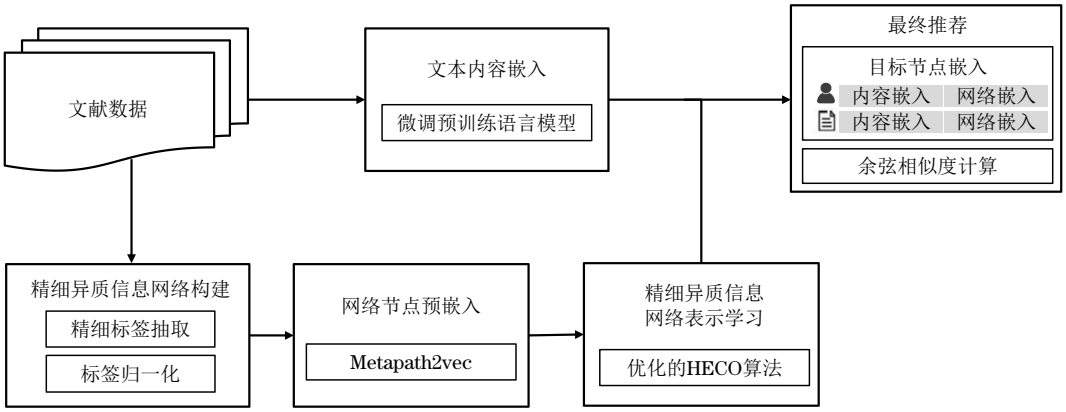

图1 PRM-FHIN整体架构

的余弦相似度，最终将相似度较高的文献列表推荐给作者。

2.1　精细异质信息网络构建

将从文献摘要信息中抽取的研究问题、方法模型两类精细化标签作为新节点加入学术异质信息网络，并与文献节点建立对应的关系。在具体构建时，采用SciAIEngine提供的英文科研实体识别模型NER-SCI-EN的接口[11-12]，从文献的摘要中提取研究问题、方法模型两类精细化标签。该接口采用基于注意力机制的双向LSTM模型，在具有5万条数据的科技论文结构化摘要语料库上，研究问题和方法模型的识别F1值分别为0.950 0和0.940 0[13]。进一步采用大语言模型ChatGLM-6B[14]对抽取的研究问题和方法模型两类标签进行归一化处理。由于大语言模型输入长度的限制，首先将抽取的所有标签存入可更新的标签池，然后从标签池中依次取N个标签作为ChatGLM-6B大语言模型的输入，调整提示指令优化模型输出，直至对N个标签归一化后的标签数量大于$N-a$，最终得到研究问题、方法模型的同义词词典。本文设定每次处理的标签数量N为500、a为10，归一化的指令[15]如下。

Your task is to perform the following actions：

1. Classify words that represent the same meaning in the following vocabulary lists

2. Each class choose one as the key and the rest as the value

Use the following format：

Words1：[words2，words2，…]

得到的研究问题和方法模型同义词词典示例如表1所示。

表1　研究问题和方法模型标签的同义词词典示例

研究问题标签同义词词典	方法模型标签同义词词典
‘image comparison’：[‘image matching’，‘image similarity’]	‘CFS’：[‘Completely Fair Scheduler’]
‘dynamic interaction’：[‘interactive dynamics’，‘real-time interaction’]	‘VMMs’：[‘Virtual Machine Monitors’]
‘MapReduce application’：[‘distributed computing application’]	‘Ceph’：[‘distributed storage system’]
‘vector graphics’：[‘vector-based graphics’，‘scalable graphics’]	‘locator/id separation protocol（lisp）’：[‘LISP’]
‘Stereoscopic rendering’：[‘3D rendering’，‘stereoscopic display’]	‘VLSD’：[‘Variable Length Subnet Mask’]

2.2　文本内容嵌入

文本内容嵌入模块旨在将文献的内容信息融入文献推荐过程，提高仅利用网络结构信息的推荐算法性能。此模块对SciBERT预训练语言模型[16]进行微调，从微调后的SciBERT模型中提取隐藏层输出的128维向量作为文献内容的向量化表示。SciBERT模型与BERT模型都是基于Transformer的双向编码器获取文本向量表示的语言模型。相比传统的深度学习模型，BERT模型可以捕捉更长距离的依赖关系。BERT模型使用WordPiece算法标记输入文本并构建词汇表，而SciBERT使用SentencePiece库构建关于科学文本语料的词汇表。相较于BERT，SciBERT更加适用于科技论文方向的自然语言处理任务[17]。考虑到SciBERT是采用生物医学和计算机科学领域共计114万篇科技论文预训练得到的模型，且预训练过程中生物医学领域文献数据占比较高，本文采用实验数据集对SciBERT模型进行领域自适应预训练，使模型更好地适用于计算机领域数据的向量表示。

2.3　网络节点预嵌入

网络节点预嵌入模块旨在将异质信息网络中的每个节点预先映射到一个低维向量空间中，从而有助于后续的精细异质信息网络表示学习。选择Metapath2vec算法[18]实现网络节点预嵌入，该算法在异质信息网络中使用元路径随机游走策略来捕获不同类型节点和关系的结构以及语义相关性。

给定一个异质信息网络$G=(V, E, T, \phi, \varphi)$，其中：$V$是网络中的节点集合；$E$是网络中边的集合；$T=(T_V, T_E)$，$T_V$和$T_E$分别代表节点和边的类型集合且满足$|T_V|+|T_E|>2$；$\phi$和$\varphi$分别是每个节点和每条边的类型映射函数。元路径[19] $\vartheta=V_1\xrightarrow{R_1}V_2\xrightarrow{R_2}\cdots\xrightarrow{R_{l-1}}V_l$，其中$R$表示节点之间的路径组合，$l$表示元路径长度，即边的数量。Metapath2vec算法采用元路径随机游走的方式生成节点的邻域上下文，这种随机游走方式可以同时捕捉不同类型节点之

间的语义关系和结构关系。针对指定的元路径，第i步转移概率的计算公式如式（1）所示。

$$p\left(v^{i+1}|v_t^i,\vartheta\right)=\begin{cases}\dfrac{1}{\left|N_{t+1}\left(v_t^i\right)\right|},\left(v^{i+1},v^i\right)\in E,\phi\left(v^{i+1}\right)=t+1\\0,\left(v^{i+1},v^i\right)\in E,\phi\left(v^{i+1}\right)\neq t+1\\0,\left(v^{i+1},v^i\right)\notin E\end{cases}\tag{1}$$

式中：$v_t^i\in V_t$，$N_{t+1}(v_t^i)$表示第i步类型为t的节点v_t^i的t+1类型的邻域节点集合。即根据元路径定义的节点类型变化进行游走，该随机游走是在元路径上的有偏游走，保证不同类型节点之间的语义关系可以适当融入Skip-Gram模型。Metapath2vec算法通过在节点v的邻域$N_t(v)$上最大化条件概率学习异质信息网络的节点嵌入，相应的目标函数为$\operatorname{argmax}_{\vartheta}\sum_{v\in V}\sum_{t\in T_V}\sum_{c_t\in N_t(v)}\log p(c_t\mid v;\vartheta)$，其中$\log p(c_t\mid v;\vartheta)$表示给定节点$v$和元路径$\vartheta$，$c_t$是$v$的邻域节点的概率。

2.4 精细异质信息网络表示学习

精细异质信息网络表示学习旨在利用网络的结构、节点内容等信息将节点嵌入低维的向量空间。为了使学习到的表征更有意义和价值，研究人员提出了不同的表征学习方法，如Metapath2vec、HGT[20]、HECO[21]等。HECO是一种基于交叉视图对比学习的异质信息网络表示学习方法，采用跨视图机制将网络模式和元路径两种视图角度下的节点信息进行融合，捕获异质信息网络中的局部和高阶结构信息。

HECO算法不依赖于大量的标注数据，而是基于自监督学习的方式，利用网络自身的结构信息来指导学习过程。但HECO方法在对比学习时认为，如果两个节点被多个元路径关联，则它们是正样本，否则是负样本。该方式计算复杂，且元路径数量并不能直接反映节点之间的相关性。HGT模型于2020年提出，利用异质信息网络的元关系来参数化异质互注意力、异质信息传递和聚合，保持节点和边类型相关的表示，捕捉网络动态性。HGT算法主要由3个步骤组成。①异质互注意力计算。计算目标节点v_t与源节点v_s的注意力权重，如式（2）所示。对每条边(v_s，v_t)计算h个头注意力并将其拼接为一个h维的重要性向量。通过Softmax函数将其转换为一个概率分布，来评价源节点v_s对目标节点v_t的重要性。②异质信息传递。计算每个源节点v_s对目标节点v_t的信息贡献程度，如式（3）所示。对每条边(v_s，v_t)同样计算h个头注意力并将其拼接为一个贡献度向量。③将异质互注意力和异质信息传递向量从源节点聚合到目标节点，获取目标节点的嵌入，如式（4）所示。$N(v_t)$表示节点v_t的所有源节点。在计算异质互注意力的过程中，目标节点v_t的源节点v_s的注意力是经过Softmax操作的，故所有源节点的注意力权重和为1。本步骤直接将注意力作为权重，对所有源节点的异质信息传递向量加权求和，得到目标节点的更新向量$\boldsymbol{H}$。

$$\mathrm{ATT}_{\mathrm{HGT}}\left(v_{\mathrm{s}},v_{\mathrm{t}}\right)=\mathrm{Softmax}\left(\|_{i\in[1,h]}\ \mathrm{ATT_head}^i\left(v_{\mathrm{s}},v_{\mathrm{t}}\right)\right)\tag{2}$$

$$\mathrm{MSG}_{\mathrm{HGT}}\left(v_{\mathrm{s}},v_{\mathrm{t}}\right)=\|_{i\in[1,h]}\mathrm{MSG_head}^i\left(v_{\mathrm{s}},v_{\mathrm{t}}\right)\tag{3}$$

$$\boldsymbol{H}=\mathop{\oplus}_{v_{\mathrm{s}}\in N(v_{\mathrm{t}})}\left(\mathrm{ATT}_{\mathrm{HGT}}\left(v_{\mathrm{s}},v_{\mathrm{t}}\right)\cdot\mathrm{MSG}_{\mathrm{HGT}}\left(v_{\mathrm{s}},v_{\mathrm{t}}\right)\right)\tag{4}$$

受HGT方法中注意力权重的启发，两个节点之间的注意力权重更能反映节点之间的相关性，故将HGT方法作为预训练模型，使用其在异质信息网络上得到的注意力权重作为HECO方法正样本选择的条件。相比基于元路径数量划分正负样本的策略，基于预训练模型的注意力权重更能反映节点间的相关性。

2.5 最终推荐

针对精细异质信息网络中的文献节点，其最终的向量化表示是文本嵌入向量与异质信息网络表示学习向量的拼接。针对精细异质信息网络中的作者节点，其最终的向量化表示是其所有论文类型的邻域节点预嵌入向量的均值与异质信息网络表示学习向量的拼接。若P为网络中的文献节点，其相应的文本嵌入向量为(P_{c1}，P_{c2}，…，P_{ci})、网络表示学习向量为(P_{g1}，P_{g2}，…，P_{gj})，则最终的向量为(P_{c1}，P_{c2}，…，P_{ci}，P_{g1}，P_{g2}，…，P_{gj})。若A为网络中的作者节点，网络表示学习向量为(A_{g1}，A_{g2}，…，A_{gj})，有k个直接相连的文献节点，第k个文献节点的文本嵌入向量为($P_{k,c1}$，$P_{k,c2}$，…，$P_{k,ci}$)，那么节点A的最终向量化表示为(A_{c1}，A_{c2}，…，A_{ci}，A_{g1}，A_{g2}，…，A_{gj})，$A_{ci}=\frac{1}{k}\sum_{m=1}^{k}P_{m,ci}$。

得到作者和文献的最终嵌入向量之后，计算作者向量与所有文献向量之间的相似度，并根据相似度值从高到低排序，选择相似度排名前k的文献推荐给作者。向量之间相似度的计算公式如式（5）所示。

$$\mathrm{Sim}\left(\boldsymbol{A},\boldsymbol{P}\right)=\frac{\sum_{x=1}^{n}\left(A_x\times P_x\right)}{\sqrt{\sum_{x=1}^{n}\left(A_x\right)^2}\times\sqrt{\sum_{x=1}^{n}\left(P_x\right)^2}}\tag{5}$$

式中：A_x、P_x分别为作者、文献的最终向量在x维上的值；n为向量维度。

3　实验及分析

3.1　实验数据集

从开放学术图谱（Open Academic Graph，OAG）[22]中抽取2010—2020年计算机科学领域的论文作为实验数据。将计算机科学下的34个二级领域作为领域字段过滤条件，如computational science、distributed computing、natural language processing等，同时删除论文标题、作者姓名、机构名称、期刊名称、论文参考文献等主要字段为空的数据，将该数据集以OAG-CS命名。该数据集涉及1 852 225篇论文（标记为“P”）、2 248 205个作者（标记为“A”）、120 992个领域（标记为“F”）、13 747个机构（标记为“I”）、11 177种期刊（标记为“V”），关系类型有P-A、A-I、P-V、P-P和P-F，共36 372 642个关系，以此构建相应的学术异质信息网络。

为了进一步构建精细异质学术信息网络，基于OAG-CS数据集中文献的领域字段选取18个领域。同时，筛选摘要字数大于0（摘要文本不为空）且小于1 050（NER-SCI-EN模型对输入文本的长度限制）的文献共282 493篇。将这些文献的摘要文本分别输入NER-SCI-EN模型，获取研究问题、方法模型两类标签。共抽取不重复的研究问题641 474个、方法模型214 639个。由于只在一篇论文中出现的标签只与该论文有关系，不会参与异质信息网络表示学习的过程，将其归入无意义的标签。剔除无意义的标签后共得到不重复的研究问题标签18 238个（标记为“Q”）、方法模型标签9 824个（标记为“M”）。为了验证NER-SCI-EN模型在本文实验数据集上的效果，分别随机抽取100个研究问题和方法模型标签，让3位相关领域专家人工评价抽取结果是否准确，至少两位专家评价结果为准确时才认为抽取结果正确。计算得到研究问题标签的识别F1值为0.870 5，方法模型标签的识别F1值为0.843 1。为了进一步确保抽取结果的质量和准确性，对研究问题、方法模型两种标签采用大语言模型ChatGLM-6B进行归一化，并将归一化后的标签与OAG-CS中的文献节点关联，构建精细异质信息网络数据集OAG-CS-Tags。

3.2　评价方法及评价指标

选取精确率、召回率及F1值作为评价指标[23]，每个评价指标的计算方法如式（6）~式（8）所示。

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}} \tag{6}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{7}$$

$$F_1=2\times\frac{P\times R}{P+R} \tag{8}$$

式中：N_{TP}为在作者参考文献列表中的推荐文献数量；N_{FP}为不在作者参考文献列表中的推荐文献数量；N_{FN}为在作者的参考文献列表中但没有被推荐的文献数量。精确率P表示在作者参考文献列表中的被推荐文献数量与总推荐文献数量的比值；召回率R表示推荐成功的文献数量与作者所有参考文献数量的比值；F1值同时考虑精确率和召回率，用来评估推荐算法的综合性能。

目前针对文献推荐的评价方法主要分为人工评价、基于用户行为信息的评价和基于引文信息的评价3类。由于人工评价不够客观，且无法适用于数据规模较大的情况，结合本文实验数据情况，采用基于引文信息的评价方法，即认为作者引用过的文献是其可能感兴趣的文献。考虑到数据集中不同作者的发表文献数量和引用文献数量分布不均衡对推荐结果的影响，基于Metapath2vec方法在OAG-CS数据集上对不同目标作者的推荐效果进行评价。根据发表文献数量和引用文献数量将目标作者分成4组。第一组是发表文献数量大于10篇且引用文献数量等于20篇的作者，符合此条件的作者共有443位。第二组是发表文献数量大于10篇且引用文献数量等于10篇的作者，符合此条件的作者共有258位。第三组是发表文献数量大于10篇且引用文献数量为11~19篇的作者，符合此条件的作者共有3 113位。第四组是发表文献数量大于10篇且引用文献数量为21~60篇的作者，符合此条件的作者共有20 621位。分别为每位作者推荐10篇文献（记为“Top10”）或者20篇文献（记为“Top20”），4组目标作者选择实验结果如表2所示。

从实验结果可以看出第四组的推荐效果较不理

表2　目标作者选择实验结果

指标	第一组		第二组		第三组		第四组	
	Top10	Top20	Top10	Top20	Top10	Top20	Top10	Top20
精确率	0.167 9	0.129 0	0.126 4	0.096 1	0.175 3	0.124 4	0.205 7	0.148 1
召回率	0.126 5	0.150 6	0.150 8	0.206 6	0.140 7	0.199 0	0.082 3	0.109 4
F1值	0.144 3	0.139 0	0.137 5	0.131 2	0.156 1	0.153 1	0.117 6	0.125 8

想，这是因为为该组每位作者推荐10篇或者20篇文献，而每位作者的引用文献数量为21~60篇，二者之间数量差别较大。第一组与第二组的推荐效果与第三组相近，但是这两组中的作者数量较少，实验结果不具有代表性。因此，选择第三组作者作为推荐的目标。

3.3　实验结果及分析

（1）OAG-CS数据集上文献推荐实验。为验证本文设计的推荐模型PRM-FHIN的效果，在OAG-CS数据集上分别以Metapath2vec、HECO为异质信息网络表示学习的基准方法（实验中分别记为PRM-Metapath2vec、PRM-HECO），与PRM-FHIN进行文献推荐对比实验。各方法的参数设置如表3所示，其中GCN为图卷积神经网络，GAT为图注意力网络。①PRM-Metapath2vec：实验中设置的元路径为{APVPA，PAVPA，VPAPV，FPAPF，IAPAI}。②PRM-HECO：元路径视图中设置元路径为{PVP，PAVPA，PAP}，编码元路径特征的GCN层数为1；网络模式视图中只考虑节点与其他类型节点的一跳邻居间的交互。模型训练中的学习率为0.001，对比损失的平衡系数为0.5。③PRM-FHIN：元路径视图中设置元路径为{PVP，PAVPA，PAP}，网络模式视图中只考虑节点与其他类型节点的一跳邻居间的交互。模型训练中的学习率为0.001，正样本选择策略中的注意力头数量为8，对比损失的平衡系数为0.5。

表3　OAG-CS数据集上实验参数设置

参数	PRM-Metapath2vec	PRM-HECO	PRM-FHIN
节点特征维度	128	128	128
隐藏层		GCN	GCN+GAT
隐藏层维度		64	64
学习率		0.001	0.001
迭代次数	10	200	200

最终的实验结果如表4所示。从实验结果可以看出PRM-FHIN模型在OAG-CS数据集上为每位目标作者推荐10篇文献时的F1值比PRM-Metapath2vec模型约提高0.154 9、比PRM-HECO约提高0.055 6，为每位目标作者推荐20篇文献时的F1值比PRM-Metapath2vec模型约提高0.155 3、比PRM-HECO约提高0.055 2。这是因为PRM-FHIN在实现文献推荐时，不仅利用了异质信息网络的结构信息进行表示学习，也融合了文献的内容信息。同时本文基于HGT训练节点间注意力权重，优化HECO算法对于正负样本的选择，比基于元路径数量的启发式规则选择方法能更好地获得表示学习的正负样本。

（2）OAG-CS-Tags数据集上文献推荐实验。为验证在学术异质信息网络中融入研究问题、方法模型两类细粒度标签是否提升了文献推荐的效果，分别基于HECO、优化后的HECO方法在OAG-CS和OAG-CS-Tags两个数据集上进行实验。即用HECO和优化后的HECO方法分别在基于OAG-CS数据集构建的异质信息网络和基于OAG-CS-Tags数据集构建的精细异质信息网络上学习节点的结构向量，基于优化后的SciBERT模型学习节点的内容向量，将结构向量和内容向量拼接作为节点的向量，并基于作者和文献节点的余弦相似

表4　OAG-CS数据集上实验结果

指标	PRM-Metapath2vec		PRM-HECO		PRM-FHIN	
	Top10	Top20	Top10	Top20	Top10	Top20
精确率	0.175 3	0.124 4	0.285 6	0.247 0	0.328 4	0.312 7
召回率	0.140 7	0.199 0	0.230 9	0.259 8	0.295 3	0.304 2
F1值	0.156 1	0.153 1	0.255 4	0.253 2	0.311 0	0.308 4

度实现最终推荐。①PRM-HECO：在元路径视图中设置元路径为{PVP，PAVPA，PAP，PQP，PMP}，GCN层数为1；网络模式视图中只考虑节点与其他类型节点的一跳邻居间的交互。模型训练中学习率设置为0.001，迭代次数为200，对比损失的平衡系数为0.5。②PRM-FHIN：在元路径视图中设置元路径为{PVP，PAVPA，PAP，PQP，PMP}，网络模式视图中只考虑节点与其他类型节点的一跳邻居间的交互。模型训练中设置学习率为0.001，正样本选择策略中的注意力头数量为8，对比损失的平衡系数为0.5，迭代次数为200。

实验结果如表5所示。可以看出PRM-HECO模型在OAG-CS-Tags数据集上为每位目标作者推荐10篇文献时的F1值比在OAG-CS数据集上约提高0.039 9，为每位目标作者推荐20篇文献时的F1值约提高0.030 2。PRM-FHIN模型在OAG-CS-Tags数据集上为每位目标作者推荐10篇文献时的F1值比在OAG-CS数据集上约提高0.020 6，为每位目标作者推荐20篇文献时的F1值约提高0.009 6。PRM-FHIN模型在OAG-CS-Tags数据集上为每位目标作者推荐10篇文献时的F1值比PRM-HECO模型约提高0.035 7，为每位目标作者推荐20篇文献时的F1值约提高0.034 6。这说明在学术异质信息网络中融入文献的研究问题、方法模型细粒度标签不但能丰富网络中包含的语义信息，而且能够提升文献推荐方法的效果；在OAG-CS和OAG-CS-Tags两个数据集上PRM-FHIN模型均比PRM-HECO取得了更好的效果，说明对HECO算法的优化是有效的。

表5　不同数据集上实验结果

模型	指标	OAG-CS		OAG-CS-Tags	
		Top10	Top20	Top10	Top20
PRM-HECO	精确率	0.285 6	0.247 0	0.313 2	0.270 5
	召回率	0.230 9	0.259 8	0.279 4	0.297 7
	F1值	0.255 4	0.253 2	0.295 3	0.283 4
PRM-FHIN	精确率	0.328 4	0.312 7	0.347 9	0.315 8
	召回率	0.295 3	0.304 2	0.316 7	0.320 1
	F1值	0.311 0	0.308 4	0.331 6	0.318 0

4　结论

近年来，研究人员越来越意识到异质信息对于推荐的重要性——基于网络中丰富的语义信息能更精准地形成用户画像，提高推荐性能。但是目前基于异质信息网络的文献推荐主要基于引用信息构建文献之间的关系，没有基于文献内容特征构建文献之间的深层语义关系，影响了最终的推荐效果。本文在对每篇文献抽取研究问题、方法模型两类科研实体的基础上构建精细学术异质信息网络，优化异质信息网络表示学习算法HECO，从而更好地学习网络中作者节点和文献节点的向量化表示，通过融合文献内容向量和网络结构向量获得文献和作者的最终向量化表示，并基于相似度计算进行文献推荐。在OAG-CS数据集上的实验结果验证了优化后的HECO算法能更好地学习节点的向量化表示，在OAG-CS-Tags数据集上的实验表明在学术异质信息网络中增加细粒度标签可有效提升文献推荐的效果。

由于科研实体抽取采用中国科学院文献情报中心SciAIEngine提供的英文科研实体识别模型，受到接口调用数量的限制，只探讨了在学术异质信息网络中增加研究问题、方法模型两类精细化标签的推荐效果。今后将探索自主可控的科研实体抽取方法，将更多精细化标签（如评价指标、实验数据等）融入学术异质信息网络，进一步验证本文所提文献推荐模型的效果。

参考文献

[1]　祝婷．融合网络表示学习与文本信息的学术文献推荐方法[J]．情报工程，2022，8（3）：81-92.

[2]　杜若鹏，张洁，寇远涛．基于共现词分析的专业科技信息平台用户画像主题标注方法改进[J]．数字图书馆论坛，2023，19（9）：58-63.

[3]　石川，王睿嘉，王啸．异质信息网络分析与应用综述[J]．软件学报，2022，33（2）：598-621.

[4]　许侃，刘瑞鑫，林鸿飞，等．基于异质网络嵌入的学术论文推荐

方法[J]. 山东大学学报（理学版），2020，55（11）：35-45.

[5] 段震，余豪，赵姝，等. 基于异质信息网络表示学习的引文推荐方法[J]. 小型微型计算机系统，2021，42（8）：1591-1597.

[6] 吴俊超，刘柏嵩，沈小烽，等. 卷积融合文本和异质信息网络的学术论文推荐算法[J]. 计算机应用研究，2022，39（5）：1330-1336.

[7] ALI Z，QI G L，MUHAMMAD K，et al. Paper recommendation based on heterogeneous network embedding[J]. Knowledge-Based Systems，2020，210：106438.

[8] DAI T，GAO T Y，ZHU L，et al. Low-rank and sparse matrix factorization for scientific paper recommendation in heterogeneous network[J]. IEEE Access，2018，6：59015-59030.

[9] MA X，WANG R R. Personalized scientific paper recommendation based on heterogeneous graph representation[J]. IEEE Access，2019，7：79887-79894.

[10] ZHU Y F，LIN Q K，LU H，et al. Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks[J]. Knowledge-Based Systems，2021，215：106744.

[11] 中国科学院文献情报中心. 文献知识人工智能引擎[EB/OL]. [2024-12-12]. http://sciengine.las.ac.cn/.

[12] 张智雄，刘欢，于改红. 构建基于科技文献知识的人工智能引擎[J]. 农业图书情报学报，2021，33（1）：17-31.

[13] 张智雄，刘欢，丁良萍，等. 不同深度学习模型的科技论文摘要语步识别效果对比研究[J]. 数据分析与知识发现，2019，3（12）：1-9.

[14] DU Z X，QIAN Y J，LIU X，et al. GLM：general language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2022：320-335.

[15] 郭晓琪. 基于精细异质图表示学习的科技文献推荐方法研究[D]. 北京：中国科学技术信息研究所，2024.

[16] BELTAGY I，LO K，COHAN A. SciBERT：a pretrained language model for scientific text[EB/OL]. [2024-12-12]. https://arxiv.org/abs/1903.10676v3.

[17] 吕琦，上官燕红，张琳，等. 基于文本内容自动分类的跨学科测度研究[J]. 数据分析与知识发现，2023，7（4）：56-67.

[18] DONG Y X，CHAWLA N V，SWAMI A. Metapath2vec：scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2017：135-144.

[19] 崔鸿飞，冯子函，张靖雨. 科研兴趣空间挖掘与多任务推荐研究：基于异质信息网络表示学习[J]. 情报学报，2023，42（10）：1224-1237.

[20] HU Z N，DONG Y X，WANG K S，et al. Heterogeneous graph transformer[C]//Proceedings of the Web Conference 2020. New York：ACM Press，2020：2704-2710.

[21] WANG X，LIU N，HAN H，et al. Self-supervised heterogeneous graph neural network with co-contrastive learning[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York：ACM Press，2021：1726-1736.

[22] ZHANG F J，LIU X，TANG J，et al. OAG：toward linking large-scale heterogeneous entity graphs[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York：ACM Press，2019：2585-2595.

[23] 王日花. 基于时间兴趣因子的网络表示学习图书推荐模型研究[J]. 情报工程，2023，9（1）：118-127.

作者简介

李琳娜，女，博士，副研究员，研究方向：推荐系统。
郭晓琪，女，硕士，工程师，研究方向：推荐系统。
张运良，男，博士，研究员，通信作者，研究方向：情报学，E-mail：zhangyl@istic.ac.cn。
王力，女，硕士，助理研究员，研究方向：情报学。
张晓丹，女，博士，研究员，研究方向：推荐系统。

Literature Recommendation Based on Representation Learning of Fine-Grained Heterogeneous Information Network

LI LinNa[1,2] GUO XiaoQi[1] ZHANG YunLiang[1,2] WANG Li[1,2] ZHANG XiaoDan[1]
(1. Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China; 2. Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038, P. R. China)

Abstract: Current research on literature recommendation based on heterogeneous information network are primarily based on existing citations, co-authorships, and publications in the same journals to build the relationship between literature. These methods do not build relationships from the perspectives of the research questions and methodological models of the literature, which are the finer-grained contents that researchers are more concerned about. This limitation prevents the integration of these fine-grained relationships into the literature recommendation process, thereby affecting the final recommendation effectiveness. This paper incorporates two types of fine-grained labels, namely research questions and methodological models of literature, into an academic information network and proposes a literature recommendation model called PRM-FHIN. It optimizes the HECO approach for heterogeneous information network representation learning to acquire structural vectors of network nodes and fine-tunes the SciBERT model to learn content vectors of network nodes. The final literature recommendation is achieved based on the integration of these content vectors and structural vectors. Using 1.85 million papers in the field of computer science from 2010 to 2020 extracted from the Open Academic Graph as experimental data, the results show that the optimized HECO algorithm can better embed the network nodes. Incorporating fine-grained labels such as research questions and methodological models into the heterogeneous information network enriches the semantic information between the literature, thereby improving the final literature recommendation effectiveness.

Keywords: Fine-Grained Heterogeneous Information Network; Representation Learning of Heterogeneous Information Network; Literature Recommendation; Fine-Grained Label; Computer Science; Research Question; Methodological Model

（责任编辑：王玮）