计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009902] [CSTR: 32024.14.csa.009902]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多维特征融合的文献研究领域关联程度量化方法①

韩　进[1], 王　志[1], 石　进[2]

[1](南京信息工程大学 软件学院, 南京 210044)
[2](南京大学 信息管理学院, 南京 210023)
通信作者: 王　志, E-mail: 202212490717@nuist.edu.cn

摘　要: 传统文献特征提取方法通常依赖于单一维度的领域特征, 难以准确预测细化的文献研究领域关联程度. 细化的关联程度预测要求提取极高精度的领域关联特征, 但在多维度提取过程中很容易出现过平滑问题, 进而导致错误的领域关联程度预测, 使得量化精度较低. 为解决上述问题, 本文提出了一种基于多维特征融合的文献研究领域关联程度量化方法. 首先, 在传统 Doc2Vec 模型提取文献语义内容特征的基础上, 构建多个关联维度图并赋予相应权重, 以提高结构关联特征的全面性. 其次, 在图学习模块中引入多通道传播策略和自适应聚合机制, 通过优化节点关联特征的聚合方式, 缓解了传统 GCN 的过平滑问题, 从而实现不同文献间精确的研究领域关联. 最后, 通过构建覆盖学者多维关联特征向量空间的最小 n 维球模型, 定量评估跨领域学者科研能力. 在大规模真实文献数据集上的实验结果表明, 该方法的带误差容限准确率 (tolerance-aware accuracy, TAA) 达到 0.68, 比 Doc2Vec、GCN 和 Sentence-BERT 模型分别高出 0.67、0.08 和 0.02, 且在不同的图神经网络模型中性能波动较小, 证明了所提方法在精度和稳定性方面均优于近年主流的基线模型.

关键词: 多维特征融合; 文献研究; 关联程度; 特征提取; 科研评估

引用格式: 韩进,王志,石进.基于多维特征融合的文献研究领域关联程度量化方法.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9902.html

Quantitative Method for Assessing Relatedness of Literature Research Domains Based on Multi-dimensional Feature Fusion

HAN Jin[1], WANG Zhi[1], SHI Jin[2]

[1](School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: Traditional literature feature extraction methods typically rely on single-dimensional domain features, making it difficult to accurately predict the relatedness of fine-grained literature research domains. The multi-dimensional extraction process often faces the over-smoothing problem, leading to inaccurate predictions of relatedness and lower quantization accuracy. To address these issues, a method is proposed to quantify the relatedness of literature research based on multi-dimensional feature fusion. First, based on the traditional Doc2Vec model for extracting semantic content features from literature, multiple related dimension graphs are constructed and assigned corresponding weights to enhance the comprehensiveness of structural related features. Second, a multi-channel propagation strategy and adaptive aggregation mechanism are incorporated into the graph learning module, which mitigates the over-smoothing problem in traditional GCN by optimizing the aggregation of related node features, thus enabling precise domain-relatedness prediction among different literature. Finally, a minimum n-dimensional sphere model is constructed to cover the multi-dimensional related feature vector space of scholars, enabling the quantitative evaluation of cross-domain scientific research abilities.

① 基金项目: 国家自然科学基金 (L2324126)
收稿时间: 2024-11-10; 修改时间: 2025-01-21; 采用时间: 2025-02-18; csa 在线出版时间: 2025-04-30

Experimental results on a large-scale real literature dataset show that the tolerance-aware accuracy (TAA) of the proposed method reaches 0.68, outperforming Doc2Vec, GCN, and Sentence-BERT models by 0.67, 0.08, and 0.02, respectively. Moreover, the performance fluctuation across different graph neural network models is minimal, demonstrating that the proposed method outperforms mainstream baseline models in terms of both accuracy and stability.

Key words: multi-dimensional feature fusion; literature research; assessing relatedness; feature extraction; scientific research evaluation

文献研究领域的关联度是指在某研究领域内, 不同子研究领域知识的彼此关联程度. 面对海量的学术文献资源, 对文献领域的关联度评估是实现精准信息检索、知识管理和跨学科研究[1]的基础. 不仅可以借此帮助信息专业人员进行信息管理, 提高文献分类的准确性, 提供更加个性化的文献推荐服务给用户, 还可以帮助研究人员快速锁定最相关研究资料, 加速科研进程和推动跨学科研究发展.

目前, 文献的领域关联度评估主要分为基于潜在狄利克雷分布 (latent Dirichlet allocation, LDA) 模型和基于向量空间模型 (vector space model, VSM) 两种. 基于 LDA 模型的评估是将主题建模[2]揭示文献集的主题结构, 根据文献的主题分布得到其特定领域之间的关联, 这是一种粗粒度的关联度评估方法; 基于 VSM 模型的评估主要基于特征工程, 通过从文献的元数据如标题、摘要、关键词等构建文献的特征向量, 利用文献向量的余弦相似度[3]或欧氏距离[4]来计算领域关联度, 该方法使用文本直接获取的领域特征计算领域关联度, 但难以处理深层次语义理解和高维向量表示稀疏[5]问题, 导致结果偏差、缺乏可解释性.

随着词嵌入模型技术和深度学习技术的发展, 研究人员提出了许多新方法, 包括基于 Doc2Vec 和 BERT 等词嵌入模型, 以及卷积神经网络 (convolutional neural network, CNN)、循环神经网络 (recurrent neural network, RNN) 等深度学习方法, 专注于文献内容的领域特征直接提取, 以提高文本内容表示质量. 然而, 这类方法通常忽略了文献间存在的关联结构特征在领域关联度计算中的作用. 近年来, 图神经网络 (graph neural network, GNN) 在文献领域分类[6]、文本相似度分析[7]和推荐系统[8]等方面取得了显著进展, 由于其引入了基于图结构的信息融合机制, 有效结合文本语义内容与关联结构信息, 显著提升了文本嵌入质量. 尽管如此, 当前的方法大多仅采用单图结构, 缺乏对多图特征融合[9]的深入探索, 无法充分挖掘文献之间的多维关联信息. 此外, 在精细化关联度预测、跨学科能力评估等细粒度应用场景中, 现有方法的预测精度和解释性仍存在不足.

为了解决现有技术的上述局限性, 本文提出了一种基于多维特征融合的文献领域关联程度量化方法. 相比现有的最优方法, 本文在多维特征融合、细致关联程度预测以及跨领域学者科研能力评估方面取得了显著提升. 本文主要贡献如下.

(1) 为了解决传统图神经网络模型仅依赖单图结构提取文献领域的直接特征局限性, 本文提出一种多维特征融合方法. 通过整合文献特征的不同维度, 构建多图结构提取更丰富的关联特征. 引入的多通道传播策略和自适应特征聚合机制能够有效缓解传统 GCN 中的过平滑问题, 显著提高关联程度量化的预测精度与稳定性.

(2) 为了解决传统评估方法难以区分细微差异的学科交叉和领域模糊问题, 本文通过构建专门设计的回归模型, 预测更加细致的领域关联程度. 同时, 采用带误差容限的准确率 (TAA) 衡量模型的有效性, 能够在不同精度要求下更精确地评估模型的表现, 进一步提高方法的实用性.

(3) 提出一种跨领域学者科研能力评估方法, 基于领域关联程度量化和最小 n 维球数学模型, 通过在多维特征向量空间中进行分析, 实现对跨领域学者科研能力的定量评估.

在真实数据集上的实验结果表明, 本文方法对比基线方法具有较高的预测正确性, 通过广泛的实验分析证明了该方法的有效性.

1 相关工作

1.1 基于 LDA 的领域关联度计算

基于 LDA 的计算方法通过对文档集进行主题建模, 得到每篇文档的主题分布和主题的词分布, 捕捉文

档在主题层面的深层次关联. 如 Rortais 等人[10]对媒体语料库使用 LDA 模型发现与语料库相关的主题后, 采用聚类分析来确定文档与特定领域之间的关联. 廖列法等人[11]则是利用 LDA 生成的主题分布计算文本之间的相似度. 这些方法为文献间领域关联的量化分析提供了视角, 但由于主题边界不清晰、主题解释难等主题模糊性[12]问题, 导致模型评估结果缺乏一定的可解释性.

1.2 基于 VSM 的领域关联度计算

通过对文本特征向量在向量空间 (VSM) 中位置的比较, 能够实现特定领域间相关性计算. 当前研究多集中于优化文本特征提取, 以提高文本相似性计算的精度. 例如, Hanifi 等人[13]使用 Doc2Vec 模型提取科学论文的文本特征, 并采用余弦相似度计算不同问题的相关性, 从而加速了科学数据的收集过程. 相较于传统的 TF-IDF 或词袋模型, Doc2Vec 在捕捉文档整体语义方面具有一定优势, 但其主要依赖词汇间的重叠, 忽视了上下文深层次语义关系, 因此难以全面反映不同文档间的语义关联.

近年来, BERT 等预训练模型凭借其对上下文的高度敏感性和多层次语义表达能力, 显著提升了文本特征提取的深度和准确性. 基于实际应用需求, 研究者们进一步改进了 BERT 模型.例如, Sentence-BERT[14]提出了生成固定长度的句子嵌入向量, 并通过距离度量快速计算句子相似度, 大幅提高了处理效率和准确性; 于润羽等人[15]则采用孪生网络架构, 缓解了 BERT 的各向异性问题, 提升了对会议论文间相似度评估的性能. 这些 BERT 改进方法目前已被广泛应用于相似度计算和领域分类任务, 被认为是目前优化文本特征提取的先进技术. 然而, 这些方法大多侧重于单一领域内的文本内容特征提取, 未能深入分析跨领域文献之间的关联模式, 在应对文献间跨领域关联度的精细化量化任务时, 表现仍存在一定的局限性.

1.3 图神经网络

由于 RNN、CNN 等神经网络难以处理长距离依赖和非连续语义问题, 图神经网络通过捕捉整个图结构中的关联关系, 有效学习节点与其邻居特征, 缓解了传统神经网络在关联特征提取上的局限性. Yao 等人[16]利用 GCN 在异构图上学习文档嵌入, 不仅考虑了文本内部的词序信息, 还通过节点连接体现了文档间的引用、主题或语义关联为领域关联特征提取提供了全新视角. 然而这种方法主要依赖单图结构, 忽略了文献特征的不同维度, 在对精细化的关联度进行量化时容易导致关键信息丢失.

此外, 传统的 GCN 学习节点特征时容易出现平滑问题[17], 因此许多学者对其节点特征聚合操作进行了改进, 提出了多种变体 GCN, 如 GraghSAGE[18]和 GAT[19]等. GraphSAGE 通过采样邻居节点并聚合其特征, 实现了对大规模图的有效处理; GAT 则引入了注意力机制, 使模型能够自适应地关注重要的邻居节点, 提高了特征表示的质量. 尽管这些方法在领域分类任务中表现较好, 但更倾向于对研究领域直接特征的提取, 而非有效捕获跨图的关联特征, 限制了模型在精细、复杂的领域关联程度判断上的表现.

1.4 跨领域学者科研能力评估

尽管学科交叉在科学研究中的重要性愈加明显, 但目前针对跨领域学者科研能力评估的研究相对较少. 现有研究往往侧重于引用关系[20]、合著关系[21]或学科标签等单一类型的分析, 难以全面评估学者实际的跨领域科研能力. 例如, 王凯等人[22]通过结合论文引用与学者合著信息, 拓展了跨领域内学术影响力的传播路径, 提高了影响力评价的准确性; 程孟夏等人[23]基于学科标签从统计指标和网络结构上进行交叉领域影响力分析, 在一定程度上反映了学者的跨领域科研情况. 然而, 这类方法通常受到学术热点的影响导致结果不稳定, 且主要关注学者影响力评价, 对跨领域学者科研能力评估研究不足, 难以有效衡量学者在不同学科交叉领域中的贡献.

2 本文方法设计

2.1 方法总体框架

本文提出的方法整体框架如图 1 所示, 主要包括特征提取模块、图学习模块、关联度计算模块和跨领域学者科研能力评估模块 4 个关键部分. 在特征提取方面, 通过构建多个无向图提取多维度结构关联特征, 突破了传统方法中仅依赖单文本的内容对直接领域特征提取的局限. 结合图学习模块, 在 GCN 层引入了多通道传播策略与自适应聚合机制, 进一步增强了文献间领域关联的表征能力. 最终, 通过高质量的文本嵌入与最小 n 维球数学模型的结合, 实现了对跨领域学者科研能力的评估.

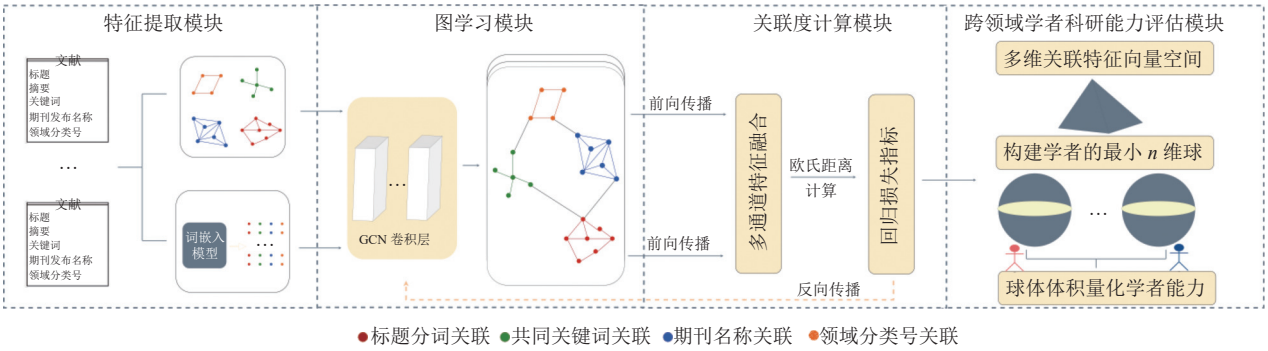


图 1 方法总体框架图

2.2 特征提取模块

在特征提取模块中, 使用 Doc2Vec 提取节点的单个文本内容特征, 并且构建了多种图结构来捕捉文献间的不同维度的关联关系. 这些图结构共同作为 GCN 特征融合的起始输入, 从而丰富了节点的特征表示. 具体来说, 本文分别构建了 4 种关联关系图, 标题分词图、共同关键词图、期刊名称关联图和领域分类关联图并给予不同的边权重初值, 每个图关注文献关联的一个特定方面. 其中每个图$G=(V,E)$定义如下.

节点 V: 代表了文献领域关联度研究中的文献, 每个节点携带文献的基本信息, 如标题、关键词、摘要、期刊信息等. 边 E: 表示文献间的关联, 边的权重则反映了该类型关联的强度.

在标题分词图中, 如果两篇文献的标题分词后存在交集, 即它们之间至少有一个分词关联, 则在对应节点间构建边. 将边的初始权重设置为两篇文献标题的相同分词个数, 这不仅量化了标题的相似性, 也提取了隐含的文献间领域关联性信息; 在共同关键词图中, 基于文献的关键词列表, 当两篇文献出现同一个关键词时, 将它们在图中通过边相连. 边的权重设置为相同关键词的个数, 借此表示关键词在领域关联表达的重要性. 同上, 在期刊名称关联图和领域分类关联图中, 若文献发表在同一学术期刊或属于同一个中图分类号时则建立边且权重设置为 1, 以此反映文献所属期刊、领域分类信息对领域关联的贡献度.

2.3 图学习模块

该模块提供一种多图特征融合框架作用于 GCN 模型, 旨在有效提升文献领域关联特征的学习能力. 该框架通过引入多通道特征传播机制和自适应特征聚合策略, 提升模型在处理复杂领域关联时的学习效果.

与传统 GCN 模型通常依赖单一图结构进行节点特征传播的做法不同, 本文设计的图学习模块通过引入多个图结构 (即多通道) 并行传播特征, 允许信息在不同图结构之间互补与融合. 这种多通道特征传播机制使得模型能够在多层次上获取文献间的多维度信息, 并逐步提取更高阶的领域关联特征. 特别是在处理跨学科领域的知识融合时, 该机制能够展示其显著优势. 基于该机制下, 本文提出计算节点特征向量的方法, 如式 (1) 所示.

$$H_{\text{fusion}}^{(l+1)}=\Phi\left(f_1^{(l)},f_2^{(l)},f_3^{(l)},f_4^{(l)}\right) \tag{1}$$

其中, $f_i^{(l)}$表示第i个图的 GCN 节点特征的更新过程, 计算方法如式 (2) 所示[24].

$$f_i^{(l)}=ReLU\left(\tilde{D}_i^{-1/2}\tilde{A}_i\tilde{D}_i^{-1/2}H^{(l)}W^{(l)}\right) \tag{2}$$

在第l层, 节点特征矩阵$H^{(l)}$通过信息传播与更新过程从多个图结构中得到, $\tilde{A}$为加上自环的邻接矩阵, $\tilde{D}$为$\tilde{A}$的度矩阵, $W^{(l)}$为权重矩阵, $ReLU$为激活函数. 每个图结构的节点更新过程由$f_i^{(l)}$表示, 而Φ是用于特征聚合的操作符, 最终在第$l+1$层进行特征融合, 生成新的特征矩阵$H_{\text{fusion}}^{(l+1)}$.

为了进一步缓解在精细化关联程度量化任务中, 由特征范围变化导致的节点特征聚合过平滑问题, 本文系统性地对比了多种聚合策略, 最终选择求和聚合操作以自适应平衡不同图结构特征的关联贡献. 与平均、最大和最小池化等方式相比, 求和聚合在保留不同图结构特征贡献方面更具优势. 其不仅有效规避了特征幅度缩小及对异常值的敏感性问题, 还确保了各图结构特征在不同关联维度上的均衡贡献. 同时, 求和聚合在特征融合过程中保持了特征的完整性和多样性, 形成了一种自适应特征聚合策略, 使得该多图特征融合框架具有广泛适用性, 能够灵活应用于多种图神经网络架构, 满足跨学科领域知识关联与特征聚合的需求.其多图特征融合框架的节点学习过程如图 2 所示.

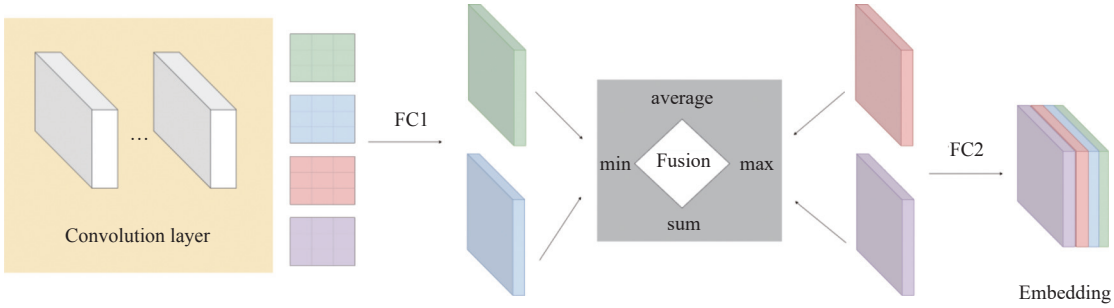


图 2 多图特征融合学习不同领域关联特征

2.4 关联度计算模块

该模块通过处理图学习模块输出的领域关联特征的文本嵌入, 并基于此计算文献间的研究领域关联度. 具体而言, 多维特征融合下的 GCN 在每一轮训练后输出最终特征聚合后的文档向量, 利用得到的文档高质量嵌入向量, 采用欧氏距离计算文献间的领域相关性, 通过模型逐轮迭代训练优化文献特征向量, 实现对文献领域关联程度更加精确地量化分析. 其中每对文献向量之间的欧氏距离的计算公式如下:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{3}$$

其中, x_i和y_i是空间中两个点在第 i 个维度上的坐标. 将$d(x,y)$作为文献间关联程度量化的预测值, 用于和实际关联程度进行比较, 为后续分析提供依据.

2.5 跨领域学者科研能力评估模块

该模块在关联度量化方法基础上, 结合最小 n 维球数学模型, 提供一种新的跨领域学者科研能力评估方法. 将学者发表的所有文献的特征向量整合成一个集合$V=\{v_1,v_2,\cdots,v_m\}$, 其中每个特征向量v_i表示一篇文献的特征. 使用特征向量均值$\bar{v}$作为最小 n 维球的球心, 由于这些向量是经过迭代优化后的关联特征增强向量, 即$\bar{v}$在多维关联特征向量空间中表示学者所有文献的中心位置, 反映了学者在不同领域之间的交叉研究情况. 其特征向量均值计算公式如下:

$$\bar{v}=\frac{1}{m}\sum_{i=1}^{m}v_i \tag{4}$$

通过计算所有特征向量到特征向量均值的最大欧氏距离作为半径 R, 代表其覆盖整个多维关联特征向量空间的最小 n 维球. 其半径 R 计算如下:

$$R=\max(d(v_i,\bar{v})) \tag{5}$$

最后, 对每位学者抽取所有发表的文献重复上述步骤, 生成每位学者的最小 n 维球模型. 通过直观比较不同学者的最小 n 维球体积, 体积越大, 表明该学者的科研领域越广泛, 跨领域科研能力越强. 其体积计算如下:

$$V_n=\frac{\pi^{n/2}R^n}{\Gamma(n/2+1)} \tag{6}$$

其中, Γ为伽马函数, 当 n 为整数时可转化为阶乘函数. 通过从多维关联特征向量空间构建最小 n 维球对学者能力进行定量评估. 值得注意的是, 这种方法适用于研究领域分布符合典型特征的学者. 具体来说, 为了确保特征向量均值和最小 n 维球模型的有效性, 以下关键假设被提出并讨论.

(1) 数据分布的代表性: 学者的研究成果在多维关联特征向量空间中的分布应具有代表性, 能够反映其主要研究领域的特征.

(2) 异常值的影响较小: 大多数文献特征向量应集中在合理的范围内, 仅有少量或没有极端异常值, 以避免这些异常值对模型结果产生显著影响.

(3) 对于符合上述假设条件的学者, 最小 n 维球体积能够有效反映其科研领域的广泛性和跨领域科研能力. 然而, 对于研究领域分布极不均匀或存在大量异常值的学者, 可能需要进一步的数据预处理或必要时采用人工判断方法来确保结果的准确性和可靠性.

2.6 量化方法步骤

本文提出的关联度量化方法具体步骤如算法 1 所示.

算法 1. 基于多维特征融合的领域关联度量化方法

输入: 文献数据集 N.
输出: 领域关联度 S.

1) 样本选择: 对 N 随机采样保证文献的领域关联度标签数据均衡分布;
2) 基于 Doc2Vec 学习节点初始化特征;
3) 构建领域关联关系图$G=(V,E)$, 其中 V 代表文献节点, E 代表不同领域关联关系的边;
4) **for** (α,β) of N **do**:
5) 分别将单篇文献元数据内容初始向量化为 α_1,β_1;
6) **if** (α,β) in E_1,E_2,E_3,E_4:
7) $(\alpha_{11},\beta_{11}),(\alpha_{12},\beta_{12}),(\alpha_{13},\beta_{13}),(\alpha_{14},\beta_{14})=GCN(G,\alpha_1,\beta_1)$

8) α_2=*Aggregating*(α_{11}, α_{12}, α_{13}, α_{14}), β_2=*Aggregating*(β_{11}, β_{12}, β_{13}, β_{14})
9) 基于 GCN 的多图特征聚合学习最终文档嵌入 α_2, β_2;
10) 通过欧氏距离计算两篇文献 (α_2, β_2) 的领域关联度 S;
11) **end for**
12) 与标签对比, 评估量化方法有效性并通过网络反向传播更新参数, 得到增强后的关联特征向量;
13) 基于上述步骤得到学者的文献特征向量, 构建覆盖每位学者多维关联特征向量空间的最小 n 维球模型并计算体积 V_n;
14) 直观比较最小 n 维球体积, 评估学者的跨领域科研能力大小.

3 实验结果与分析

3.1 数据集

实验所使用的数据集为万方收集的学术文献数据. 数据共包含 241 351 篇文献, 23 个类别 (A–W). 通过对文献数据进行一系列预处理操作, 包括特征选择相关字段, 去空值等保留文献数为 24 591, 在此基础上选择特定领域 G 类别下的文献, 然后为保证评估的公平性, 通过随机采样得到子类别 G1–G8 中图分类号前缀的文献解决样本不平衡问题. 实验评估以文献对为单位, 最终实验数据共 1 120 条, 共 626 640 对, 本文按照 8:2 分配训练数据和测试数据, 详细的数据情况分布如图 3 所示.

3.2 评价指标

考虑大量人工标注的主观性且效率低等局限性, 通过算法利用文献的中图分类号的层级关联关系将领域关联度标签自动划分为 0.25、0.5、0.75、1 代表文献对之间不同的关联程度, 数值越高代表文献间的领域关联程度越强, 本文设计的回归模型预测值为 0–1 之间的连续值, 采用 MSE 和 MAE 回归指标衡量模型有效性.

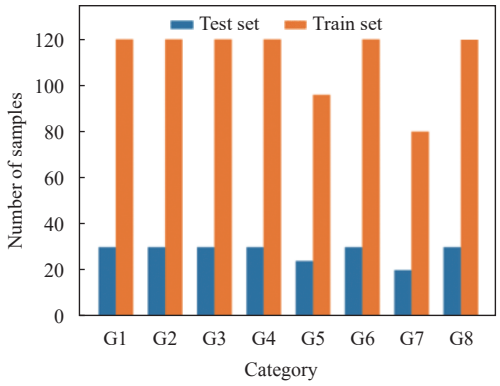


图 3 样本比例分布图

由于这些标签表示为连续值范围内的离散等级, 传统的回归指标 MSE 和 MAE 虽然能够提供模型预测与真实标签之间的偏差信息, 但它们无法直接反映出模型对于这些特定等级分类的准确性. 此外, 考虑到领域关联度量任务的实际应用场景, 往往并不需要绝对精确到小数点后几位的预测值, 而是更关心预测值是否落在了正确的等级范围内. 例如, 在实际应用中, 如果一个文献对的真实关联度为 0.75, 而模型预测为 0.78, 如果误差范围为 0.1, 则这个预测应被视为正确, 因为它足够接近于 0.75 这一等级.

这种在实际应用场景下使用容错意识的评估方式已经在许多研究中广泛应用, 如 Halpern 等人[25]引入了“容忍等级 (tolerance tiers)”的概念, 通过选择合适的容忍等级, 用户可以在保持可接受的准确性的前提下显著降低机器学习云服务 API 的延迟, 而 Johnson 等人[26]利用容错意识的仿真结果来比较设计空间中的不同容错水平, 以提高系统的可靠性和性能. 因此为了更好地描述方法的准确性, 本文采用 *TAA* (tolerance-aware accuracy) 一个带误差容限的准确率来评估模型的性能, 作为衡量关联度量化任务更灵活的评估指标. *TAA* 计算如式 (7) 所示:

$$TAA = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i, y_i, \varepsilon) \tag{7}$$

假设有 N 个样本, 每个样本都有一个预测值$\hat{y}$和一个实际标签y, 允许的预测误差为ε. 那么, 对于每一个样本 i, 可以定义一个指示函数 I 来检查预测值是否在允许的误差范围内. 在函数 I 中, 只有当$\hat{y}_i$和y_i的误差范围在ε中才视为一次正确.

3.3 实验设置

实验采用 Doc2Vec 词向量表示初始文献特征, 维度为 100 生成的嵌入向量. 将 GCN 层数设置为 2 层, 训练的 epoch 的次数设置为 300 次, 隐藏层设置为 128 维, 输出维度为 64 维, 学习率设置为 0.000 1, 正则化系数设置为 0.000 5, dropout 参数为 0.2, 采用 Adam 优化网络, 其余参数根据数据集而定.

3.4 实验结果

(1) 与不同模型的实验对比

由于文献领域关联度量化是一个相对较新的研究方向, 为了更全面地评估本文方法的有效性, 将目前几种主流的文本特征提取模型作为基线, 包括词嵌入模型 Doc2Vec 和未经过预训练的多图 GCN 以及预训练模型 Sentence-BERT. 此外, 在本文提出的多维特征融合框架下, 还采用了 GraphSAGE 和 GAT 等图神经网

络模型作为参照, 进一步验证本文方法的表现. 表 1 展示了本文方法与这些基线模型的实验效果对比.

表 1 与不同模型的实验效果对比

Model	MSE	MAE	*TAA*
Doc2Vec	0.098±0.002	0.255±0.001	0.011±0.005
GCN*	0.036±0.001	0.134±0.002	0.600±0.010
Sentence-BERT	0.060±0.001	0.168±0.001	0.660±0.002
GAT	0.047±0.004	0.150±0.001	0.667±0.001
GraphSAGE	0.045±0.002	0.143±0.002	0.672±0.002
本文方法	0.034±0.001	0.122±0.003	0.679±0.001

注: “*”表示文本向量为随机初始化

根据实验数据, 本文提出的方法在关联度量化任务中的 MSE、MAE 和 *TAA* 指标上均表现出显著效果. 具体而言, 均方误差 (MSE) 从最高约 0.1 降低至 0.03 左右, 平均绝对误差 (MAE) 从最高约 0.25 降至 0.12 左右. 同时本文方法的 *TAA* 达到 0.68, 较 Doc2Vec、GCN 和 Sentence-BERT 分别提升了 0.67、0.08 和 0.02. 尽管 MSE 和 MAE 可以衡量模型预测值与真实值之间的误差大小, 但这些连续值指标不足以直接反映模型在不同领域关联程度等级上的识别能力, 而 *TAA* 基于误差容限的概念, 不仅衡量了模型预测值与真实值之间的误差, 还考虑了在一定范围内的误差可接受性, 适用于不同精度评估需求, 提供了更具实用性的性能衡量标准.

通过与基线模型对比, 发现采用多维特征框架下的 GCN 模型 (带有“*”表示文本向量为随机初始化) 在训练的初期, *TAA* 值约为 0.6, 表现出不错的准确率. 然而, 随着训练的深入, 随机初始化的文本向量未能有效捕捉文献节点的实际领域特征, 导致 *TAA* 值持续下降. 在同一实验中, Sentence-BERT 模型展现了强大的文本嵌入提取能力, 取得了较高的 *TAA* 值为 0.66, 相比之下, 传统的 Doc2Vec 词嵌入模型仅取得了 0.01 的 *TAA* 值. 这一显著差异表明, Doc2Vec 模型仅能提取单一文本特征, 难以有效捕捉文本之间的上下文关系, 关联度量化精度受到严重制约.

在实验中, 本文方法还通过对比在多维特征融合框架下的图学习模块中, 采用不同的图神经网络模型训练的表现. 发现其在 *TAA* 指标上表现接近, GCN 模型为最高 0.68, GraphSAGE 和 GAT 模型的 *TAA* 指标分别为 0.672 和 0.667. 这样的稳定性表明, 本文方法的多维特征融合框架在相同参数设置下, 模型的鲁棒性较强, 准确率在不同网络结构间变化较小, 从而显著提升关联度计算的精度与稳定性.

(2) 图神经网络在不同特征融合方式下的实验对比

将 GCN 的多图特征融合方式分别为求和、平均、最大和最小值池化时的实验, 测试集准确率随着训练轮数的变化如图 4 所示.

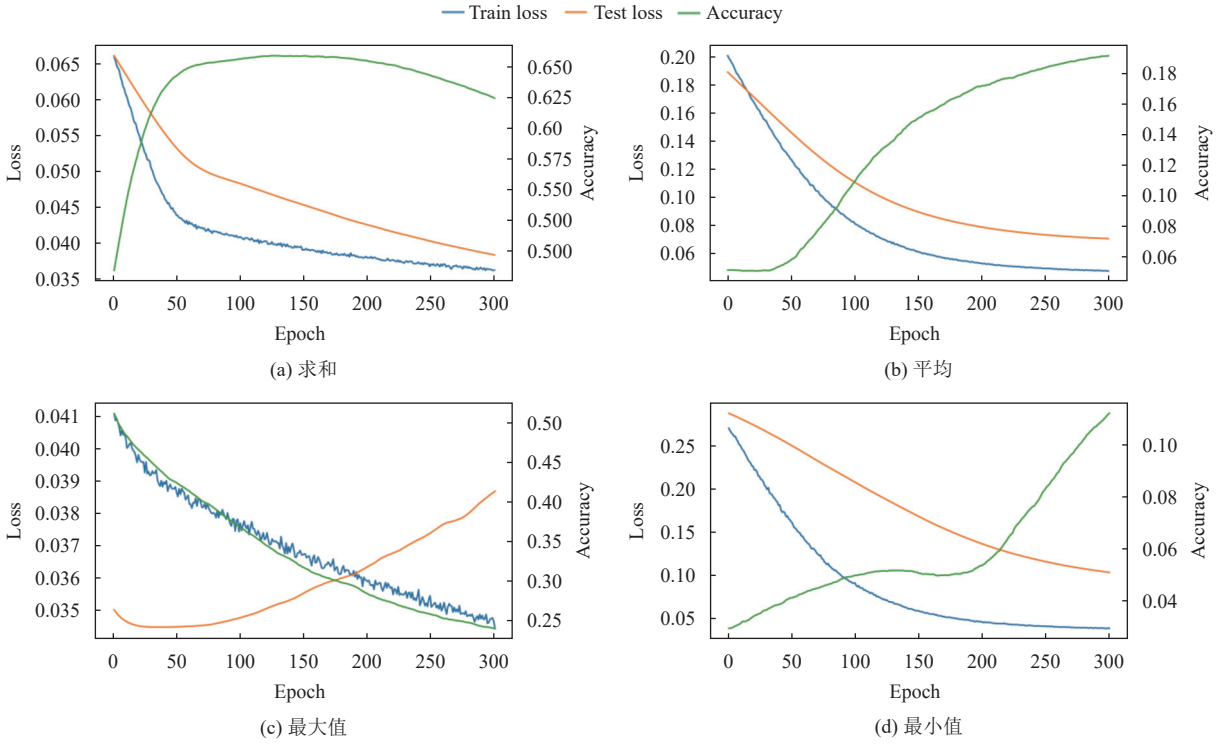


图 4 不同特征融合方式下的模型性能

从图 4 可以看出, 采用图 4(a) 求和作为多图特征融合策略的准确率最高, 这是因为当 GCN 聚合不同图结构特征时使用平均方式减小了特征聚合的动态范围, 导致节点不能学习到充分的关联特征, 而当使用最大、最小值作为特征融合方式时, 由于对关联特征数值峰值的高度敏感性, 图节点特征不对齐或者不相关则会导致其效果不佳.

从图 4(b) 可以看出, 模型在训练集和测试集上的初始损失 loss 值较高、正确率较低, 这是因为在该特征融合方式下特征范围扩大较慢, 可以观察到图中模型在前 50 轮正确率上升并不明显, 在经过一段时间训练后模型正确率快速上升, 但最终不超过 0.2. 值得注意的是, 在图 4(c) 和图 4(d) 中使用最大、最小值来融合多通道节点特征效果有着显著的区别, 最大值融合虽然在训练初期正确率相对较高, 但该方式导致模型出现了严重的过拟合现象, 随着训练轮次增加正确率持续下降, 这是因为模型在训练初期误将局部峰值、异常值等局部特征当作全局特征. 相反, 由于最小值融合对异常值的敏感度较低, 更关注于数据中的泛化模式, 在训练初期表现出较低的正确率, 但随着训练的进行正确率在稳步上升.

(3) 模型在不同ε下的实验对比

为选用合适的ε误差作为模型评估的方式, 根据标签中不同关联度等级间差异度为 0.25, 考虑到模型输出精度的有效性, 将 0.125、0.1、0.075 分别作为模型预测正确允许的误差范围的实验, 测试集正确率随着训练轮数的变化如图 5 所示.

随着允许的误差范围的增加, 显然模型的准确率也在提高, 提升幅度在前 50 轮较为明显, 在 50–150 轮正确率相对稳定, 150 轮往后模型开始过拟合正确率有所下降.

(4) 方法敏感性分析

为了探究采用不同方法进行关联度计算的有效性, 分别采用了欧氏距离和余弦相似度的度量方法进行实验, 训练过程的损失变化及准确率随训练轮数的变化如图 6 所示.

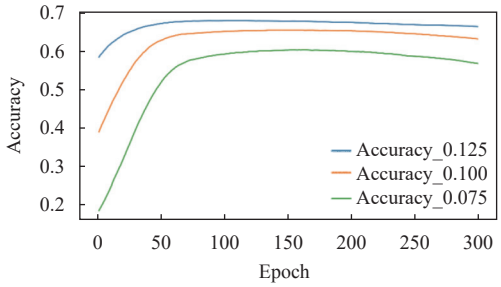


图 5 不同ε下的模型性能

在对模型的有效性进行评估过程中, 当采用余弦相似度作为度量标准时, 测试集的损失下降后仍有 0.3 左右, 并且其准确率低于 0.1. 相较之下, 使用欧氏距离度量能够提升模型性能, 达到较高正确率水平. 这是因为尽管余弦相似度和欧氏距离均可用于衡量文献向量间的领域关联程度, 但二者在相关性量化机制上存在本质区别: 前者通过向量间夹角的余弦值评估方向一致性, 越相近则表明向量表示的领域间越相关; 后者则是通过衡量两篇文献向量在多维特征空间中的距离量化相关性, 距离越小则表明领域关联程度越高. 两种度量方法主要区别为数值大小的敏感性. 通过实验对比发现, 采用欧氏距离方法更适合本数据集的数据特点, 具体表现为考虑特征向量模长及其在多维空间中的精确位置上, 模型的有效性显著优于仅考虑向量方向的一致性.

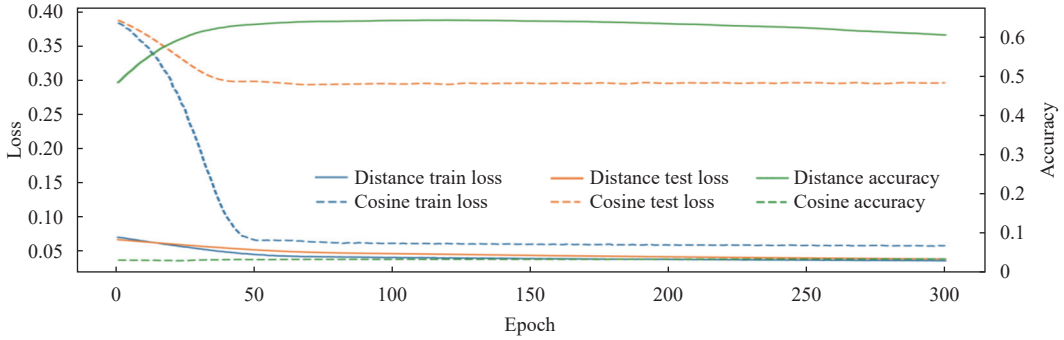


图 6 不同关联度计算方法下的模型性能

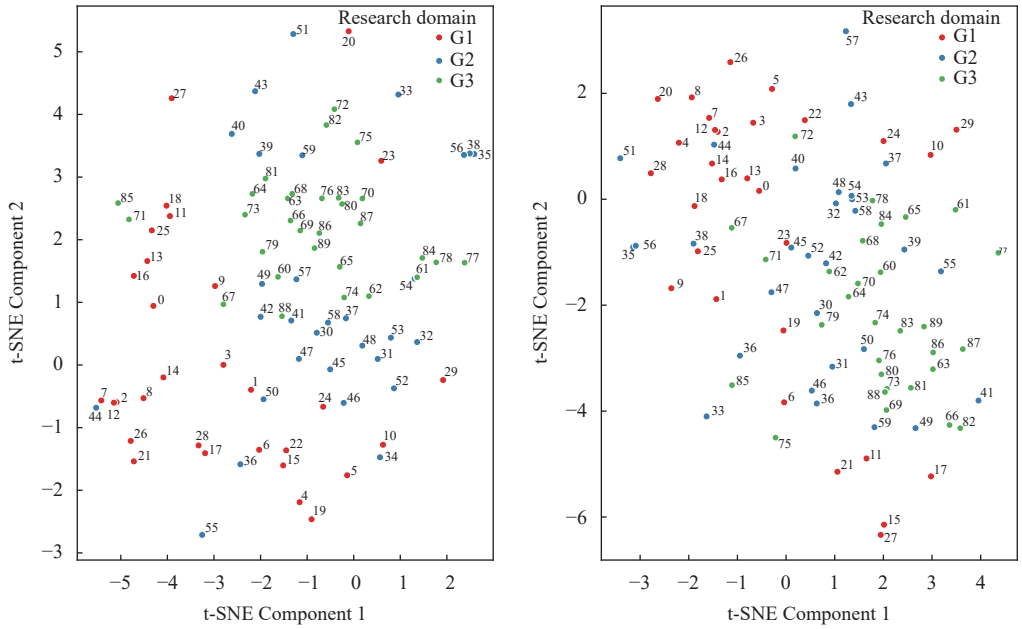
3.5 可视化分析

本文对真实文献的研究结果进行了可视化分析. 实验案例为测试集中的 90 篇文献, 共涉及 3 个研究领域.

(1) t-SNE 降维可视化

采用 t-SNE 降维方法对文献特征向量进行分析, 直观呈现其研究领域的分布情况. 通过对比训练前后关联特征的分布变化, 如图 7 所示, 验证了所提方法在优化文献关联特征方面的有效性.

图 7 展示了这些文献通过 t-SNE 降维后的文献特征向量在二维空间中的分布, 其中每个点表示一篇文献, 编号代表文献 ID. 图中的不同颜色代表文献所属不同的研究领域 (G1、G2 和 G3). 图 7(a) 显示了未经过关联度量化方法训练的文献特征向量的空间分布情况, 发现文献大部分按研究领域聚集在一起, 但不同领域的文献相对分散. 说明在未经过关联度量化方法训练时, 文献研究领域之间的关系区分度较低, 尽管语义特征提取能够部分区分不同领域的文献, 领域内文献之间的关系较为明显, 但不同领域的文献缺乏有效的关联特征提取.



(a) t-SNE Visualization (features_0) (b) t-SNE Visualization (features_1)

图 7 关联特征训练前后分布变化

图 7(b) 则展示了经过关联度量化方法训练后的结果. 相比图 7(a), 文献分布变得更加复杂: 同一领域的文献依旧大部分聚集在一起, 但不同领域的文献之间开始出现交集, 部分文献聚集在一起更为明显, 表明不同领域之间有了更多的关联. 该现象进一步表明, 关联度量化方法不仅能够有效捕捉同领域文献的关联特征, 还能够促进跨领域文献之间的相互关联, 从而识别学者的跨领域研究能力.

(2) 相关系数

为进一步具体展示关联度量化方法的有效性, 本实验采用皮尔逊相关系数分析文献特征向量, 提供一种与传统精度和损失指标不同的验证方式. 通过可视化同一领域和不同领域文献间的皮尔逊相关系数热力图, 结合真实文献的部分元数据如表 2 所示, 展现所提方法是否有效捕捉和区分文献间的关联关系. 图 8 展示了来自 3 个研究领域的 9 篇文献间相关系数热力图. 其中, 图 8(a)–(c) 展示了同一研究领域组内的关联性, 而图 8(d)–(f) 则展示了不同研究领域组间的关联性.

从图 8(a) 可见, G1 文献组内部的相关系数较高 (0.81–0.97), 例如文献《中国文化产业技术效率度量研究: 2000–2011 年》与文献《非物质文化遗产产业化中的权利归属研究——以手工技艺类为例》在标题和关键词上均集中体现了“文化”相关主题, 使其相关系数达到最高值 0.97, 展示了方法在捕捉同领域特征关联性方面的能力. 图 8(b) 显示 G2 文献组相关系数在 0.85–0.92 之间, 组内文献一致性较高, 例如 ID 为 30 与 31 的文献同发布在“中国科技期刊研究”期刊, 并在关键词“传播”和“对策”上具有语义关联, 这也表明了方法能够捕获文本特征多种维度的关联性. 而图 8(c) 中

G3 文献组的相关系数处于稍低水平 (0.77–0.87), 这与该组文献的多学科交叉属性有关, 例如 ID 为 60 与 61 的文献尽管在“知识”与“反馈”层面具有交集, 但其多样化的领域背景导致相关性略低.

表 2　真实文献的元数据

文献ID	标题	关键字	期刊	所属领域
0	中国文化产业技术效率度量研究: 2000–2011年	“随机前沿分析”“文化产业”“技术效率”“内生性”	中国软科学	G1
1	文化寄生: 一种跨文化传播的变异范式——《西游记》	“文化寄生”“西游记”“英国广播公司”“原生文化”“寄生文化”	东南学术	G1
2	非物质文化遗产产业化中的权利归属研究——以手工技艺类为例	“非物质文化遗产”“文化产业”“权利归属”“权利流转”“手工技艺”	东岳论丛	G1
30	如何提高高校学报学术质量	“高校”“学报”“问题”“对策”	中国科技期刊研究	G2
31	新经济时代增强科技期刊传播力度的思考	“新经济科技期刊传播”	中国科技期刊研究	G2
32	论编辑分析	“编辑”“文稿”“分析”“复杂性”	中国科技期刊研究	G2
60	上海科技发展的学科结构与绩效评价	“科技发展”“学科”“主成分分析法”	中国科技论坛	G3
61	微笑曲线的知识论释义	“微笑曲线”“附加值”“嵌入编码知识”“非嵌入编码知识”“隐性知识”	东南大学学报(哲学社会科学版)	G3
62	一种基于网络的同伴写作评改方法	“合作学习”“同伴互评”“评改反馈”“有效性”	中国外语	G3

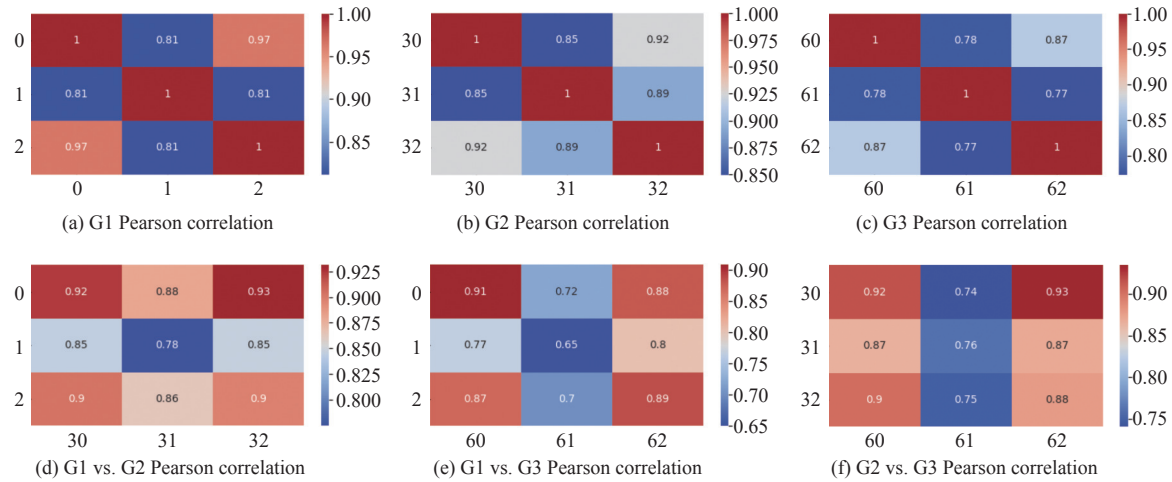


图 8　文献间的皮尔逊相关系数

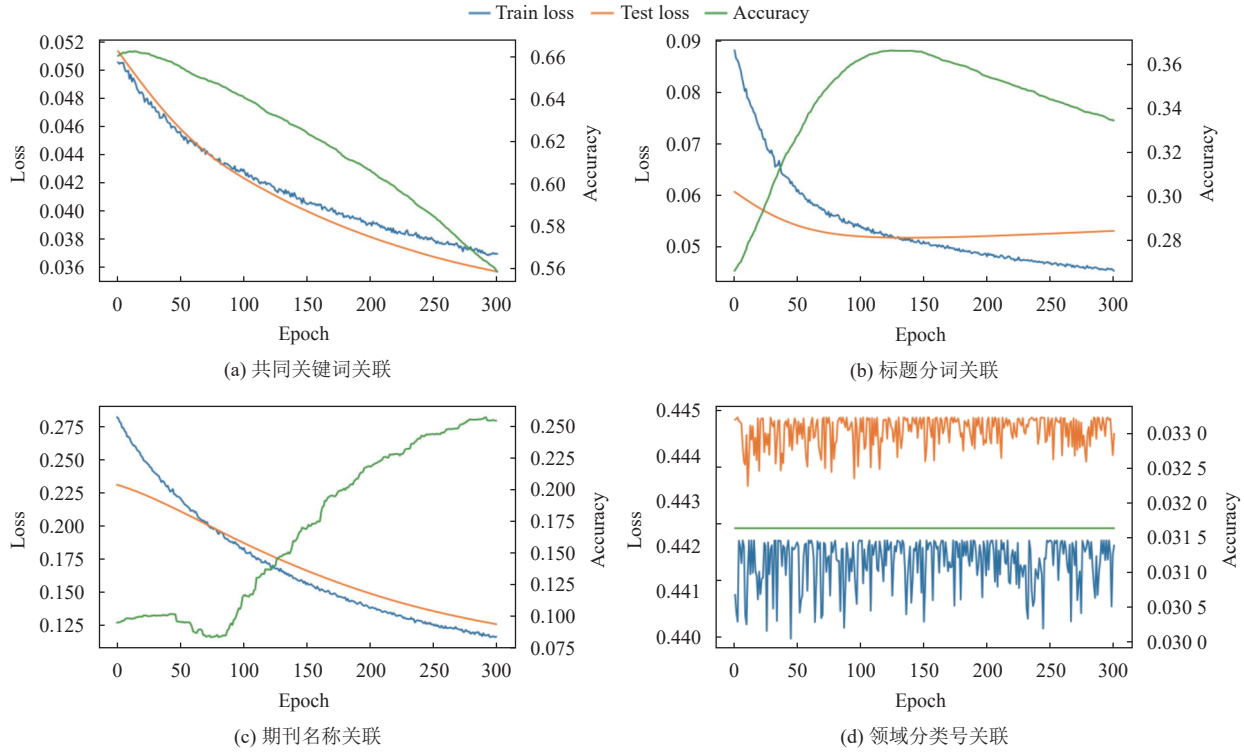
对于跨领域文献间相关性, 图 8(d) 表明 G1 与 G2 之间的相关性范围为 0.78–0.92, 文献间整体表现出高度相关性, 例如 ID 为 1 与 31 的文献分别探讨“文化”与“科技期刊”不同主题下的传播, 但两者的传播模式和基于背景影响的研究方法有显著相似性, 使得其相关系数达到了 0.92. 然而, 随着文献领域的交叉与差异的增大, 如图 8(e) 和图 8(f) 所示, 相关系数呈现一定的波动. 这表明, 尽管该方法在处理跨领域文献时能够捕捉到领域间潜在关联性的差异性, 但在某些复杂的跨领域关联关系中, 例如 ID 为 31 与 62 的文献在基于文献对应的元数据进行对比时, 表面上难以直接推断出两者在不同研究领域中的显著关联性, 但其相关系数却显示较高的关联性值, 与实际的关联程度不一致. 这也表明了方法在处理某些复杂的跨领域关联关系中, 尤其是那些样本的研究领域关联程度难以明确界定时, 会出现一定程度的预测偏差.

3.6　消融实验

为验证多图特征融合模块在文献的领域关联特征提取上的有效性, 从而提高模型输出的准确性与稳定性, 在本文数据集上进行了消融实验. 在去除多图特征融合模块后分别采用 4 种不同的单图结构进行文献的领域关联度任务评估, 训练过程的损失变化及准确率随训练轮数的变化如图 9 所示.

由图 9 可见, 只采用单图结构提取关联特征, 模型准确性有明显下降, 这是因为单一类型的图结构无法充分捕捉到所有的相关信息, 从而信息损失导致模型

预测准确率下降. 同时单图结构下模型难以学习泛化的关联特征, 使得输出结果缺乏可解释性, 可以明显注意到图 9(a) 中模型出现了过度拟合以及图 9(d) 中模型只能学习到特定于该图的关联模式导致完全不适用于该评估任务, 而在多个文本关联图上训练可以提升模型对文本间不同领域关联模式的适应能力, 保证了模型具有一定的稳定性. 因此, 验证了采用多图特征融合方式在文献领域关联度评估任务上的有效性.



(a) 共同关键词关联 (b) 标题分词关联

(c) 期刊名称关联 (d) 领域分类号关联

图 9 单图关联结构下的模型性能

4 结论

本文针对传统文献领域关联度计算方法存在预测精度不足、缺乏解释性问题, 结合图神经网络提出一种基于多维特征融合的文献研究领域关联程度量化方法. 本文方法以新的视角量化了文献研究领域关联程度, 采用特征融合方式提取更高质量的领域关联特征的文本嵌入. 在实验数据集上的实验结果表明, 本文方法在传统回归指标上实现了显著下降, 并且在误差容限深度学习评估方式下的准确率达到了 0.68, 比 Doc2Vec、GCN 和 Sentence-BERT 模型分别高出 0.67、0.08 和 0.02, 证明了所提方法在精度和稳定性方面均优于近年主流的基线模型. 但本文也存在一定的局限性, 虽然自动化处理标签可以显著减少人工标注所需的时间、人力以及主观误判, 从而适用于大规模真实数据集上实验, 但是基于该方式得到的某些标签可能产生不正确的实际关联等级; 另一方面, 面对文献间难以确定领域相关程度的样本, 方法很难给出精确的预测值, 这是未来要研究的问题. 同时, 目前由于缺乏带标签的跨领域学者科研能力评估数据集, 难以进行有效的数据验证和结果对比, 因此本文提出的跨领域学者科研能力的定量评估方法目前主要停留在理论层面. 尽管方法在理论上科学合理, 但缺少实际数据支持仍限制了其验证和应用的广泛性. 这些局限性在未来工作中需要进一步优化改进.

参考文献

1 Zezario RE, Fu SW, Chen F, *et al*. Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 54–70. [doi: 10.1109/TASLP.2022.3205757]

2 Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993–1022.

3 Thongtan T, Phienthrakul T. Sentiment classification using document embeddings trained with cosine similarity.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence: ACL, 2019. 407–414. [doi: 10.18653/v1/P19-2057]

4 Liu ZG, Fu YM, Pan Q, *et al*. Orientational distribution learning with hierarchical spatial attention for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8757–8772.

5 Malali N, Keller Y. Learning to embed semantic similarity for joint image-text retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 10252–10260. [doi: 10.1109/TPAMI.2021.3132163]

6 Liu TF, Hu YL, Wang BY, *et al*. Hierarchical graph convolutional networks for structured long document classification. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(10): 8071–8085. [doi: 10.1109/TNNLS.2022.3185295]

7 Qian SS, Xue DZ, Fang Q, *et al*. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4794–4811.

8 Wu SW, Sun F, Zhang WT, *et al*. Graph neural networks in recommender systems: A survey. ACM Computing Surveys, 2023, 55(5): 97.

9 Zhou T, Zhou Y, Gong C, *et al*. Feature aggregation and propagation network for camouflaged object detection. IEEE Transactions on Image Processing, 2022, 31: 7036–7047. [doi: 10.1109/TIP.2022.3217695]

10 Rortais A, Barrucci F, Ercolano V, *et al*. A topic model approach to identify and track emerging risks from beeswax adulteration in the media. Food Control, 2021, 119: 107435. [doi: 10.1016/j.foodcont.2020.107435]

11 廖列法, 勒孚刚, 朱亚兰. LDA 模型在专利文本分类中的应用. 现代情报, 2017, 37(3): 35–39. [doi: 10.3969/j.issn.1008-0821.2017.03.007]

12 Zhang YX, Calyam P, Joshi T, *et al*. Domain-specific topic model for knowledge discovery in computational and data-intensive scientific communities. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2): 1402–1420. [doi: 10.1109/TKDE.2021.3093350]

13 Hanifi M, Chibane H, Houssin R, *et al*. Problem formulation in inventive design using Doc2Vec and cosine similarity as artificial intelligence methods and scientific papers. Engineering Applications of Artificial Intelligence, 2022, 109: 104661. [doi: 10.1016/j.engappai.2022.104661]

14 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 3980–3990. [doi: 10.18653/v1/D19-1410]

15 于润羽, 李雅文, 李昂. 融合领域特征的科技学术会议语义相似性计算方法. 智能系统学报, 2022, 17(4): 737–743. [doi: 10.11992/tis.202203050]

16 Yao L, Mao CS, Luo Y. Graph convolutional networks for text classification. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 7370–737.

17 黄学坚, 刘雨飏, 马廷淮. 基于改进型图神经网络的学术论文分类模型. 数据分析与知识发现, 2022, 6(10): 93–102. [doi: 10.11925/infotech.2096-3467.2022.0071]

18 Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1024–1034.

19 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. Proceedings of the 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.

20 Jiang XD, Zhu RH, Ji RS, *et al*. Co-embedding of nodes and edges with graph neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7075–7086. [doi: 10.1109/TPAMI.2020.3029762]

21 Wu HR, Yan YG, Ng MKP. Hypergraph collaborative network on vertices and hyperedges. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3245–3258. [doi: 10.1109/TPAMI.2022.3178156]

22 王凯, 刘鹏. 融合论文引用与学者信息的交叉领域学术影响力分析. 情报理论与实践, 2024, 47(4): 143–151. [doi: 10.16353/j.cnki.1000-7490.2024.04.019]

23 程孟夏. 基于单篇论文学科分类号的跨学科研究——以人文社会科学为例 [硕士学位论文]. 南京: 南京理工大学, 2021. [doi: 10.27241/d.cnki.gnjgu.2021.001989]

24 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

25 Halpern M, Boroujerdian B, Mummert T, *et al*. One size does not fit all: Quantifying and exposing the accuracy-latency trade-off in machine learning cloud service APIs via tolerance tiers. Proceedings of the 2019 IEEE International Symposium on Performance Analysis of Systems and Software. Madison: IEEE, 2019. 34–47.

26 Johnson T, Lam H. Incorporating fault-tolerance awareness into system-level modeling and simulation. Proceedings of the 11th IEEE/ACM Workshop on Fault Tolerance for HPC at Extreme Scale. St. Louis: IEEE, 2021. 31–40. [doi: 10.1109/FTXS54580.2021.00008]

(校对责编: 张重毅)