

语言模型增强的引文网络连边因子挖掘

王懋懋^{1,2}, 赵洪科^{1,2}, 吴李康^{1,2}, 焦之贤³, 黄振亚⁴

1. 天津大学管理与经济学部, 天津 300072;

2. 天津大学复杂管理系统实验室, 天津 300072;

3. 北京大学光华管理学院, 北京 100871;

4. 中国科学技术大学认知智能国家重点实验室, 安徽 合肥 230026

摘要

GNN可以有效聚合图结构数据中邻近节点的信息以学习节点表征,在引文网络数据挖掘领域展现出了巨大的潜力。然而,现有的大多数GNN在捕获连边信息的驱动因子方面缺乏深入探索,这限制了对节点间复杂关系的深入理解和解释。例如,不同论文之间的引用关系通常由各种研究主题驱动,尽管已有研究尝试通过整合LLM的强文本理解和生成能力来丰富节点与连边的特征表达,但这些方法仍未能有效挖掘连边信息隐含的驱动因子。鉴于此,提出了一个创新的框架——语言模型增强连边因子挖掘,旨在以插件形式提高各种图神经网络对连边关系建模的区分度,探究其在引文网络链接预测场景下的应用价值。粗粒度因子挖掘通过为LLM设计结构化信息提示文本,从引文网络图中提取出显式的类别相关连边因素。细粒度因子挖掘利用K-means聚类算法从图文本数据中捕捉更细致的语义主题级别连边因素。为了验证所提框架的有效性,在3个公开数据集上进行实验。实验结果表明,语言模型增强连边因子挖掘框架在提升各类图神经网络模型性能方面具有显著优势。

关键词

大语言模型; 图神经网络; 连边因子挖掘; 链接预测; 引文网络

中图分类号: TP391.1

文献标志码: A

doi:10.11959/j.issn.2096-0271.2025025

Language model-enhanced edge factor mining in citation network

WANG Minmao^{1,2}, ZHAO Hongke^{1,2}, WU Likang^{1,2}, JIAO Zhixian³, HUANG Zhenya⁴

1. College of Management and Economics, Tianjin University, Tianjin 300072, China

2. Laboratory of Computation and Analytics of Complex Management Systems, Tianjin University, Tianjin 300072, China

3. Guanghua School of Management, Peking University, Beijing 100871, China

4. State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei 230026, China

Abstract

GNN is adept at aggregating information from neighboring nodes in graph-structured data to learn node representations, showing immense potential in the field of citation network data mining. However, most existing GNN lack a deep exploration of the factors driving edge information, which limits a thorough understanding and

interpretation of complex relationships between nodes. For instance, the citation relationships between different papers are often driven by a variety of research topics. Despite attempts to enrich node and edge feature representations by integrating LLM with their strong textual understanding capabilities, these approaches have still not effectively sloved the problem of uncovering the underlying drivers of edge information. In light of this, an innovative framework was proposed—language model-enhanced edge factor mining, aimed to enhance the differentiation of edge relationship modeling in various GNN through a plug-in approach, exploring its application value in citation network link prediction scenarios. Coarse-grained factor mining extracted explicit category-related edge factors from citation network graphs containing documents by designing structured information prompts for LLM. Fine-grained factor mining used the K-means clustering algorithm to capture more detailed semantic topic-level edge factors from graph textual data. To verify the effectiveness of the proposed framework, experiments were conducted on three public datasets. The results demonstrate a significant advantage of language model-enhanced edge factor mining framework in improving the performance of various GNN models.

Key words
large language mode, graph neural network, edge factor mining, link prediction, citation network

0 引言

引文关系挖掘在学术研究和科研管理中扮演着重要的角色^[1]。随着信息技术的发展，不同领域的学术文献数量都呈现爆炸式增长，传统的文献检索和分析方法已经无法满足对海量文献数据进行深入分析的需求，特别是对结构化的引文网络。其依据文献之间的引用关系构建图结构，从而揭示不同文献之间的关联程度，帮助人们理解学术研究领域的发展趋势和知识传播的路径^[2]。

随着图数据挖掘研究的逐渐深入，图神经网络（graph neural network，GNN）已成为图分析和建模的主要工具之一^[3]。GNN通过有效地聚合邻近节点的属性和拓扑结构信息，学习节点的低维稠密表征，为非欧几里得空间中的数据管理带来了新的范式，提供了对节点间交互更详尽和深刻的理解^[4]。由于其优秀的性能和良好的任务适配性，基于图神经网络的技术

也逐渐被应用在引文网络分析中，并取得了令人瞩目的成果，成为主流解决方案之一^[5]。

然而，尽管图神经网络方法在表征引文网络节点向量上取得了显著的研究进展，但在探索图结构数据中节点间关系这一重要研究方向上，连边信息挖掘的潜力在很大程度上仍然未被充分发掘。笔者专注于识别和理解影响节点间连边关系建立的因素，这些因素被学者们称作连边因子，旨在揭示驱动图数据形成背后的隐式结构和暗含连边关系细粒度属性，解释现有链接的成因，以增强链接预测等图数据挖掘主流任务的表现。

在现实世界的应用场景中，图结构数据中包含的复杂连边关系解构往往不在大多数GNN的处理范畴^[6]。例如，在引文网络中，不同类型的引用连接（如同一作者关联、相似领域关联等）可能具有极大的差异性，这些差异性对于理解整个网络的结构和功能至关重要。然而，传统GNN的邻居连边统一化处理方法往往忽略或低估了更细致连边信息的重要性，这限制了它

们在处理复杂图数据或者区分属性相似节点时的性能表现。因此，挖掘连边因子以增强已有的GNN性能显得尤为重要。一些研究尝试开发利用连边信息的图神经网络来应对这一挑战^[7-8]，但这些只在向量映射空间中进行计算的模型往往对于挖掘的子连边所代表的语义解释不足。还有一些研究尝试开发主题模型以挖掘可解释连边因子，然而这些模型的增强效果仅局限于单一模型，可拓展性不足，且模型增强效果也相对有限^[9-10]。近年来，随着大语言模型（large language model, LLM，简称大模型）的兴起^[11]，一些研究开始探索利用LLM来增强节点或连边的特征表达^[12-13]。这些模型具有生成文本特征的能力，为图数据分析提供了新的可能性。然而，目前这些尝试未能深入探索对连边因子的挖掘，并且对特定模型的依赖程度也较高，还未有不依赖于特定模型并且能够容易嵌入图神经网络进行表征增强的方法。实现该需求对于提升已有图模型性能上界至关重要。

鉴于上述需求，本文提出了一种新型的、针对图数据广泛适用的模型增强框架——语言模型增强连边因子挖掘，专注于提高引文网络中的连边关系辨别能力，并挖掘构成节点间连接的因素。语言模型增强连边因子挖掘框架由两部分构成。一是粗粒度因子挖掘，将图文本库中的显式类别作为LLM框架的文本提示，结合图结构信息，以识别连边因子。二是细粒度因子挖掘，通过K-means算法^[14]对文档中的文本信息进行语义聚类分析，以挖掘细致、直观可见的连边因子。其通过将这两层次的因子权衡合并，以及进行高性能的预训练模型处理，以向量形式呈现连边连接的构成因素，并融入图神经网络算法的信息聚合过程，从而提升图学习模型在链接预测任务上的性能。

1 相关文献工作

在引文网络分析中，链接预测的主要任务是预测网络中未被直接观测到的或可能即将发生的引用关系，其主要可以分为两大类：基于启发式的方法和基于表征学习的方法。基于启发式的方法依赖于网络中已知的结构特征来预测潜在的链接。例如，共同邻居指数^[15]基于以下假设：如果两个节点共享多个邻居，那么它们之间很可能存在链接。Adamic-Adar指数^[16]通过给予连接较少的共同邻居更大的权重来预测链接。Katz指标^[17]则通过加权考虑所有长度的路径来评估节点对的连接可能性，其中短路径具有更高的权重。PageRank算法^[18]通过模拟随机游走在网络中的运动路径来评估节点的重要性。这些基于启发式的方法通常基于简化的假设，并且在处理具有复杂特征和动态变化的大规模网络时，其预测精度和泛化能力可能受限。

近年来，GNN的发展为引文网络分析带来了新的范式，这代表了处理图结构数据能力的一次重大飞跃。GNN能够捕捉引文网络中错综复杂的关系和模式，提高了处理复杂系统的精确性^[19-20]。随着GNN的不断演进，GNN在处理大规模图数据和适应动态图结构方面取得了显著进展^[21]。图卷积网络（graph convolutional network, GCN）^[22]的引入标志着一个重要的进步，通过利用局部一阶近似的频谱图卷积，使节点表示的学习更有效。图注意力网络（graph attention network, GAT）^[23]对GNN进一步创新和改进，采用自注意力机制动态地分配权重，使模型能够更加专注于图中最相关的部分，从而提高从复杂关系数据中学习的能力。同时，归纳学习框

架 GraphSAGE (graph sample and aggregate) 等的出现扩展了 GNN 的适用性, 便于在动态环境和现实世界应用中被使用^[24], 这些技术同样也在社交网络分析等领域发挥了关键作用。基于表征学习的方法如变分图自编码器 (variational graph auto-encoder, VGAE)^[25]通过学习节点的潜在向量表征提高了链接预测的准确性和模型的泛化能力^[26]。此外, 一些研究通过结合上下文嵌入信息提高了对重要引用的识别能力, 并在图数据中实现了更好的表达^[27]。

在引文网络分析中, GNN 已被广泛用于预测学术论文的未来引用趋势, 显示出在学术领域的巨大潜力^[28]。随着技术的不断进步, GNN 在应对引文网络中的各种挑战上也取得了重要进展。例如, 异构图神经网络 (high-order graph neural network, HGNN) 被开发出来, 以更有效地处理不同类型的节点和边^[29]。此外, 时序图神经网络 (temporal graph neural network, T-GNN) 考虑了时间动态性的增加, 有助于模型捕获引文的动态变化趋势^[30]。进一步地, 研究表明, 采用如图变换网络 (graph transformer network, GTN) 这类高级深度学习技术可以自动识别图数据中的关键拓扑结构, 从而提高预测的准确性^[31]。随着 Transformer 技术的发展^[32], GNN 开始融合自然语言处理 (natural language processing, NLP) 技术, 利用文本中的丰富语义信息来提高模型性能。如 BERT (bidirectional encoder representations from transformer) 预训练模型已被应用于节点分类, 以提高模型性能^[33]。此外, 文本增强的知识图谱表示学习方法被提出改进链接预测和实体分类^[34]。随着 LLM 的发展, 越来越多的工作将 LLM 与图数据相结合, 一些研究通过

LLM 增强实体链接 (entity linking, EL) 的能力^[35], 一些新的研究也尝试将 LLM 与推荐系统相结合, 提升模型的推荐效果和可解释性^[36]。尽管这些研究进展显著, 但现有模型在挖掘不依赖于特定模型的连边因子、不能充分利用文本信息和节点间的结构关系, 以及在提高模型可解释性方面仍面临挑战。

2 语言模型增强连边因子挖掘框架

2.1 问题描述

在 GNN 中, 链接预测通常通过计算节点的嵌入向量的相似性来实现。本文的任务可以描述为通过语言模型增强连边因子挖掘框架对图连边因子进行挖掘, 对引文网络图进行图表示增强, 进一步进行链接预测。输入是引文网络图 $G = (V, E)$, 其中包含节点集合 V 和边集合 E , 输出是增强后的引文网络图 $G' = (V, E, \mathcal{E})$ 及节点对 (u, v) 之间的连接概率, 其中 $\mathcal{E}$ 表示图增强后的边特征集合。

2.2 预备工作

本节引入图数据以及标准的 GNN 符号, 将引文网络图表示为 $G = (V, E)$, 其中 V 表示节点集合, E 表示边集合, 每个节点 v 包含文本内容 t_v 。GNN 可以表示为一个函数 $f(\cdot)$, 用于聚合节点和边的特征信息, 为每个节点 $u \in V$ 生成表示向量。给定节点 u 的特征 x_u 、与节点 u 相连的连边特征 e_{uv} 和图的拓扑结构, 经过多层网络后, 最终输出 $y_u = f(x_u, \{(x_v, e_{uv}) : (u, v) \in E\})$, 即节点的表示。这里, $(u, v) \in E$ 表示节点 u 和节点 v 之间存在边, 而 e_{uv} 表示节点 u 和节点 v

之间边的特征，最终得到的 y_u 以 $\mathbb{R}^{D \times N}$ 矩阵的形式表示节点嵌入，其中 N' 表示输出特征的维度。

LLM^[37]指一类强大的自然语言处理模型，其通过大规模的文本数据进行训练，以学习单词、短语甚至整个句子的语义和语法特征。在引文网络链接预测任务中，LLM学习的语义信息可以帮助模型更好地理解论文内容，发现文本之间的关联性，从而为链接预测任务提供更加准确和全面的信息支持。

在传统的GNN中，通常只考虑节点特征。为了有效地将边特征 e_{uv} 应用于各种GNN中，笔者对图神经网络的消息传递函数进行了调整，使边特征能够通过改进后的消息传递函数传递到节点中，从而实现语言模型增强连边因子挖掘框架在各类图神经网络中的充分利用。大部分GNN会在消息传递中为模型添加自环以提高模型的性能^[22]，然而，通过语言模型增强连边因子挖掘框架获取的边特征是两个不同节点之间的连边关系，没有涉及自环的特征，因此需要对该情况进行处理，保证连边数量与边特征的数量相匹配。

首先为每个节点添加自循环边，加入边集合中。

$$E = \{(u, v) | (u, v) \in E\} \cup \{(u, u) | u \in V\} \quad (1)$$

将自循环边的特征初始化为节点特征，并与现有的边特征进行拼接。

$$e_{uu} = x_u \quad (2)$$

$$\mathcal{E} = \{e_{uv} | (u, v) \in E\} \cup \{e_{uu} | u \in V\} \quad (3)$$

在消息传递阶段，将节点特征 x_v 与边特征 e_{uv} 拼接到一起，用 h 来表示，作为输入特征。

$$h = \text{concat}(x_v, e_{uv}) \quad (4)$$

对原特征 h 进行线性变换运算。

$$h' = W \cdot h \quad (5)$$

其中， W 的维度为 $(N+L) \times N'$ ，确保拼接后的特征向量可以被处理。

2.3 模型结构

语言模型增强连边因子挖掘框架主要由两个模块构成，如图1所示，分别是粗粒度因子挖掘模块和细粒度因子挖掘模块。算法的具体流程如算法1所示。

算法1：语言模型增强连边因子挖掘图表示增强算法

输入：引文网络图 $G = (V, E)$ ，每篇论文的节点文本信息 t_v ，其中 $v \in V$

输出：增强后的引文网络图 G' ，节点 u 和节点 v 之间的连接概率 y_{uv} ， $u, v \in V$

- 1: 提取每个节点 t_v 的文本信息
- 2: 初始化粗粒度因子集合 F^c ，并将每个粗粒度因子加入 F^c 中
- 3: 设置细粒度因子数量 K
- 4: 使用K-means算法挖掘 K 个细粒度因子
- 5: 将细粒度因子加入细粒度因子集合 F^f
- 6: **for all** $(u, v) \in E$
- 7: $\text{Response}_{u,v}^{\text{coarse}} \leftarrow \text{LLM}(\text{Prompt}(t_u, t_v, F^c))$
- 8: **for all** $(u, v) \in E$
- 9: $\text{Response}_{u,v}^{\text{fine}} \leftarrow \text{LLM}(\text{Prompt}(t_u, t_v, F^f))$
- 10: **for all** $(u, v) \in E$
- 11: $e_{u,v} \leftarrow \lambda \cdot \text{Embedding}(\text{Response}_{u,v}^{\text{coarse}}) + (1 - \lambda) \cdot \text{Embedding}(\text{Response}_{u,v}^{\text{fine}})$
- 12: **for all** $(u, v) \in E$
- 13: 将 e_{uv} 加入 $\mathcal{E}$ ， $\mathcal{E}$ 表示增强后边特征的集合
- 14: **for all** $u \in V$
- 15: 将 e_{uu} 加入 $\mathcal{E}$

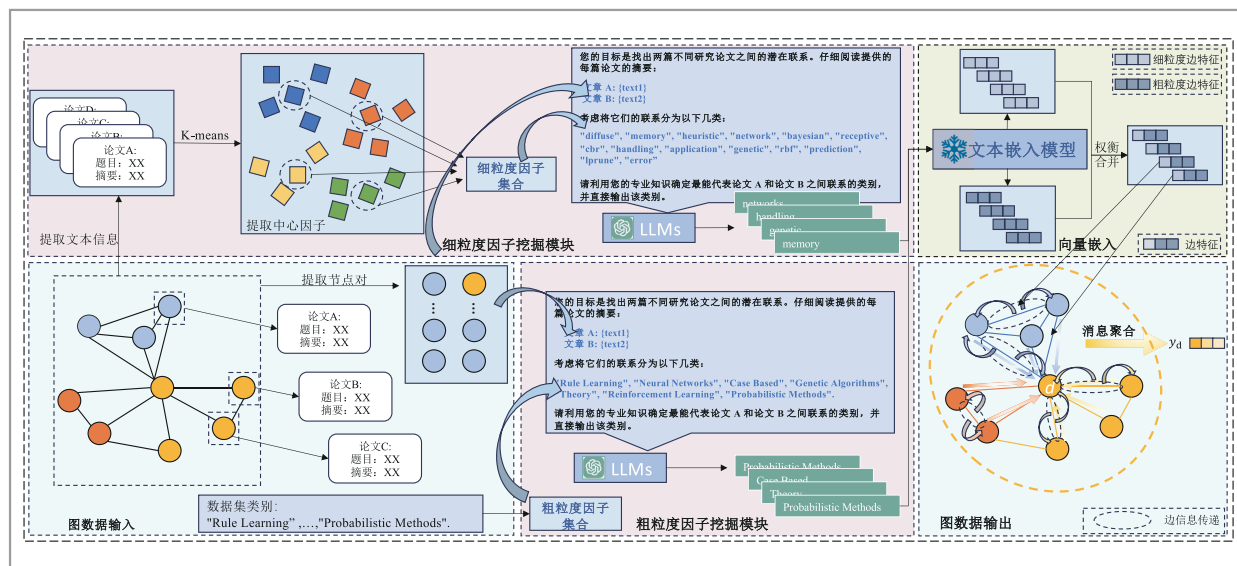


图1 语言模型增强连边因子挖掘框架

- 16: $G' = (G, \mathcal{E})$
- 17: 初始化损失阈值 ϵ 和最大迭代次数 T
- 18: 初始化当前损失 Loss
- 19: 初始化迭代计数 $t \leftarrow 0$
- 20: **while** $t < T$ **and** $\text{Loss} > \epsilon$ **do**
- 21: $y_u = \text{GNN}(x_u, \{(x_v, e_{uv}) : (u, v) \in E\})$
- 22: 计算 $\hat{y}_{uv}$ 与实际连接状态之间的预测得分
- 23: $\hat{y}_{u,v} = \text{Sigmoid}(y_u \cdot y_v)$
- 24: 使用交叉熵损失函数计算损失:
- 25: $\text{Loss} = -\sum_{(u,v) \in E} \mathbb{I}(y_{uv}=1) \cdot \log \hat{y}_{uv} + \mathbb{I}(y_{uv}=0) \cdot \log(1 - \hat{y}_{uv})$
- 26: 迭代计数 $t \leftarrow t + 1$

2.4 粗粒度级别因子挖掘

在对图数据集中的文本信息进行探索和分析时，首先需要从宏观视角入手，重点关注数据集中存在的显式关系，结合精心设计的提示词和 LLM 的强大外部知识，深入挖掘节点间的关系。

粗粒度模块利用 Cora 等数据集中预先

分类的信息，比如“规则学习”“神经网络”“基于案例的推理”等学术领域，将这些信息融入 LLM 的提示文本 t_i 中。这不仅为 LLM 提供了丰富的背景知识，还确保其输出与学术分类紧密相关。这样的设计在实际应用中有助于 LLM 更准确地理解节点间的关系。

在提示文本的构造过程中，需要确保其适用于图结构，以指导 LLM 在图数据中进行更深入的推理分析，从而挖掘节点间的边因子信息。为此，笔者充分利用大语言模型的强大推理能力，通过构建清晰明确的提示文本来引导 LLM 的输出。在提示文本的设计方面，LLM 的目标是提取粗粒度和细粒度的边因子。设计分层提示文本 t_i 策略，针对粗粒度信息，设置主题和类别；对于细粒度信息，则设计更具体的上下文条件。引导式问题在提示文本中起到关键作用，可以帮助 LLM 更好地理解 and 生成相关内容。例如，“请描述论文 A 和论文 B 在某一特定方面的联系”，并提供示例回答，以帮助模型理解预期输出。此外，在

实现过程中，将任务分解为可以并行处理的多个子任务，分别挖掘不同类别的边因子，并通过同步机制确保各个线程之间的数据同步和结果整合，从而大大降低时间开销。在实际操作中，这一过程可以通过式(6)表示。

$$\text{Response}_{ij}^{\text{coarse}} = \left(\text{LLM} \left(\text{Prompt} \left(t_i, t_j, F^c \right) \right) \right) \quad (6)$$

其中， $\text{Response}_{ij}^{\text{coarse}}$ 表示 LLM 在接收到与节点 i 和 j 相关的文本信息 t_i 和 t_j ，以及粗粒度因子集合 F^c 后生成的输出。具体的文本提示构造如图 2 所示，首先，利用图节点的文本信息引导 LLM，之后将“Paper A”和“Paper B”分别将两个节点的文本信息 t_v 和 t_u 输入模型中。同时，输入候选粗粒度因子集合 F^c （如“Rule Learning”等）来限定 LLM 的输出范围，减少模型输出偏离主题的情况。最后，充分利用 LLM 对文本的理解能力，生成与挖掘相关的因子类别。通过清晰的提示词，LLM 能够更好地理解节点间的关系，并进行准确的预测。

2.5 细粒度级别因子挖掘

为了更深入地理解并提升模型的效果，本节将重心转向更细致的分析层面。除了已知的分类信息，节点中还包含丰富的文本信息，这些信息不仅描述了节点的特性，还暗示了节点之间潜在的连接方式，对全面探索节点间边的构成至关重要。仔细分析和解析节点内的文本信息，能够揭示网络结构中更微妙和复杂的关系模式。

细粒度模块用 K-means^[14] 聚类算法深入探索节点内的文本数据，以识别文本数据的分布特征，并进一步理解节点之间的连接性。这种方式不仅可以识别不同的文本数据簇，基于文本相似性揭示节点间的

潜在联系，还可以更深入地理解网络内的信息流动以及连接节点的边之间的复杂关系。

K-means 聚类算法通过最小化每个数据点到其聚类中心的距离之和来形成聚类簇，从而在每个聚类簇内保持较高的相似性，同时在不同聚类簇之间保持较低的相似性。具体而言，该算法旨在最小化以下成本函数。

$$J = \sum_{i=1}^D \sum_{k=1}^K w_{ik} \cdot \|d_i - \mu_k\|^2 \quad (7)$$

其中， J 是成本函数， D 是节点数量， K 是因子数量。笔者将在实验部分探讨 K 的数目对实验效果的影响，每个因子是聚类中的中心词。 w_{ik} 是一个指示变量，当文本数据点属于聚类 k 时，为 1，否则为 0。 d_i 为文本数据点， μ_k 是聚类 k 的中心， $\|d_i - \mu_k\|^2$ 是文本数据点 i 到其聚类中心 μ_k 的欧几里得距离的平方。

通过将 K-means 聚类算法应用于网络节点内的文本数据，细粒度模块可以识别出不同的文本主题语义聚类。这些聚类反

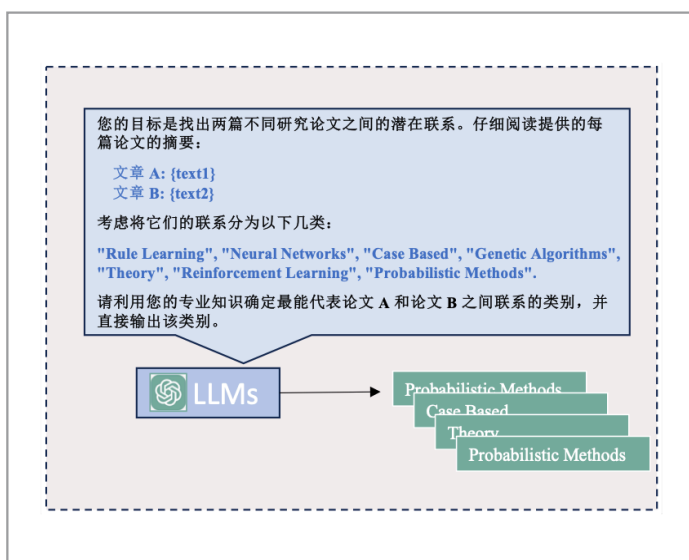


图2 提示文本模版

映了根据主题或风格对文本数据的聚合，从而揭示了基于文本相似性的节点之间的潜在连接。例如，主题相似的文本可能指向具有相似功能或角色的节点，而这种相似性可能难以通过传统的网络分析方法捕获。将通过 K-means 聚类算法获得的不同文本数据聚类的中心文本视为细粒度因子集合 F^f ，这些中心文本代表了每个聚类内文本数据的核心主题或特征。将它们引入 LLM 中作为新的提示文本，利用 LLM 对文本的深度理解和生成能力，探索和识别节点之间更详细和深刻的关系与互动，如主题相关性、概念连接或功能相似性，此过程可以表达为以下内容。

$$\text{Response}_{i,j}^{\text{fine}} = \text{LLM}(\text{Prompt}(t_i, t_j, F^f)) \quad (8)$$

其中， $\text{Response}_{i,j}^{\text{fine}}$ 表示 LLM 在接收到与节点 i 和节点 j 相关的文本信息 t_i 和 t_j ，以及细粒度因子集合 F^f 后生成的输出。

2.6 特征融合策略

融合粗粒度特征和细粒度特征后的连边特征可以全面反映节点之间的关系。前文中获得的粗粒度级别的答案 $\text{Response}_{i,j}^{\text{coarse}}$ 和细粒度级别的答案 $\text{Response}_{i,j}^{\text{fine}}$ 分别通过嵌入模型后获得高维的向量表示，再将两个维度的向量进行聚合，获得最终的连边特征 $e_{i,j}$ 。嵌入函数 $\text{Embedding}(\cdot)$ 可以将 LLM 的文本回答转换为固定长度的特征向量，利用调节参数 λ 来混合粗粒度特征和细粒度特征，具体如式 (9) 所示。

$$e_{i,j} = \lambda \cdot \text{Embedding}(\text{Response}_{i,j}^{\text{coarse}}) + (1 - \lambda) \cdot \text{Embedding}(\text{Response}_{i,j}^{\text{fine}}) \quad (9)$$

该方法有效地整合了粗粒度和细粒度两个层次的文本信息，融合的连边特征可以直接被应用于图神经网络中进行消息聚

合，用于各种图分析任务，如链接预测、图分类或节点分类等，提高模型处理复杂图数据的性能和准确性。

3 实验与分析

3.1 数据集

本文在 3 个公开学术数据集——DBLP、M10^[38]和 Cora^[39]上进行实验，与其他图数据集相比，这些数据集提供了分类标签和原始文本，便于开展明确且可解释的案例分析。在预处理阶段，笔者首先移除了无意义的连词和介词，例如“of”“or”“and”等。为减少低频词可能造成的噪声，只保留在数据集中出现超过 10 次的词汇。此外，将引文数据集中频繁且无意义的词汇进行剔除，例如“paper”“propose”等。利用 bge-large-en-v1.5 模型^[40]对文本进行嵌入。处理后的数据集的统计信息列在表 1 中。

3.2 基线模型

为了充分测试语言模型增强连边因子挖掘框架增强各种 GNN 的有效性，选取了多个具有代表性的图表示学习的基线模型，所有这些模型都可以考虑图神经网络中的连边特征。GCN^[22]简化了图上的卷积操作，从而构建了一个可扩展且高效的通用图模型。GAT^[23]在图神经网络中引入了一种基于注意力的节点聚合机制。DeeperGCN (deeper graph convolutional network)^[41]通过 SoftMax 和 PowerMean 进行更细腻和可调节的聚合策略。GATv2^[42]解决了其前身原始 GAT 层的静态注意力限制的问题，通过允许每个节点考虑任何其他节点实现动态注意力。UniMP^[43]模型基于 Transformer 架构，利

表1 数据集统计信息表

符号	种类	节点数/个	边数/条	类别	特征维度/维
DBLP	引文网络	2 708	5 429	7	1 024
M10	引文网络	60 744	105 781	4	1 024
Cora	引文网络	10 310	5 923	10	1 024

用多头点积注意力动态权衡邻近节点的贡献，允许聚合上下文相关的特征。

3.3 实现细节

实验过程使用 PyTorch 来训练语言模型增强连边因子挖掘框架以及选定的基线模型，实验在一台配备有 RTX 3090 GPU 的 Linux 服务器上进行，模型的优化采用了 Adam 优化器^[44]优化模型参数，并使用 GPT-3.5^[12]作为框架中推理的 LLM。通过实验效果反馈，细粒度选项数量设定为 14，这表示模型在数据集中挖掘 14 种不同的连接因素。训练模型的过程中使用了 bge-large-en-v1.5 模型^[40]，向量嵌入的维度设定为 1 024。为了控制模型的训练过程，学习率设置为 10^{-4} 。在特征融合策略中，参数 λ 设置为 0.3。为了确保实验的公平性和一致性，本文对于实验的关键参数和配置进行了仔细的设置，对于不同的模型在不同数据集上的表现，根据训练效果确定模型在 200~1 500 个周期内进行训练，以确保每个模型能够充分学习数据中的模式和特征。在实验部分，给出的结果呈现的是不同模型在进行 5 次重复实验后得到的均值效果。

3.4 性能比较结果

链接预测作为一种被广泛应用于评估

图学习模型的任务，通过基于两个节点的嵌入相似性来预测节点间的链接。由于其与连边语义挖掘的强相关性，笔者主要在该任务上进行相关模型效果验证。为了评估模型的性能，首先为每个节点随机选择与本身存在的边等量的负边（即无关联节点）。实验设置了不同的训练集比例以提高实验的全面性，包括 10%、20%、25%，同时保持验证集的比例为 5%。

训练集的比例由前两者确定。模型的预测性能使用接收者操作特征曲线（下面积（area under the curve, AUC）^[45]来衡量。除了 GCN 在 Cora 上进行的 10% 以及 DBLP 上的部分测试，语言模型增强连边因子挖掘框架一致地增强了基线模型的性能。总体而言，语言模型增强连边因子挖掘框架可以显著增强每个模型对节点之间关系的理解，从而获得更准确的链接预测，具体结果见表 2。实验结果显示，引入语言模型无关的语言模型增强连边因子挖掘框架在几乎所有情况下均能显著提升基线模型在链接预测任务上的性能。此外，语言模型增强连边因子挖掘框架增强效果对于不同的模型，在不同的训练数据下效果有所差异，例如，GATv2 模型在 Cora 数据集上的效果在训练比例较大的情况下更明显，训练比例为 25% 时，语言模型增强连边因子挖掘框架加强后的 GATv2 效果提升了 0.0386，而训练比例为 10% 时，语言模

表2 链接预测结果

数据集	训练集	GAT	+语言 模型增强 连边因子 挖掘	GATv2	+语言 模型增强 连边因子 挖掘	DGCN	+语言 模型增强 连边因子 挖掘	UniMP	+语言 模型增强 连边因子 挖掘	GCN	+语言 模型增强 连边因子 挖掘
Cora	25%	0.8504	0.8718	0.8267	0.8653	0.7566	0.7590	0.8511	0.8648	0.6195	0.6514
	15%	0.8222	0.8374	0.8198	0.8205	0.7298	0.7450	0.8319	0.8442	0.6125	0.6197
	10%	0.8225	0.8246	0.8273	0.8374	0.7044	0.7141	0.7594	0.8289	0.6294	0.5976
DBLP	25%	0.9390	0.9409	0.9441	0.9463	0.9548	0.9550	0.9604	0.9600	0.9051	0.9046
	15%	0.9159	0.9131	0.9259	0.9278	0.9365	0.9394	0.9406	0.9438	0.9075	0.9176
	10%	0.8941	0.8955	0.9090	0.9121	0.9198	0.9251	0.9275	0.9288	0.9001	0.9102
M10	25%	0.9185	0.9277	0.9529	0.9599	0.9424	0.9460	0.9560	0.9561	0.8415	0.8797
	15%	0.9147	0.9258	0.9492	0.9540	0.9342	0.9359	0.9496	0.9515	0.8452	0.8636
	10%	0.9132	0.9183	0.9482	0.9509	0.9252	0.9353	0.9429	0.9443	0.8300	0.8699

型增强连边因子挖掘框架加强效果仅提升了0.0101。而对于UniMP模型而言，在Cora数据集上的表现在训练比例较小时更明显，例如，训练比例为15%时，语言模型增强连边因子挖掘框架对加强后的UniMp效果提升了0.0123，而在训练比例为10%时，效果提升了0.0695。因此，需要针对不同应用场景选择合适的模型进行增强。

3.5 消融实验

为了展示笔者提出的每个模块的必要性，本节利用前面介绍的几个模型，在不同的训练集比例下精心设计了消融实验。具体来说，笔者使用的模型包括完整的语言模型增强连边因子挖掘框架和两个子模型：原始模型和粗粒度模块。原始模型指的是未经增强的初始模型，而粗粒度模块表示仅通过粗粒度层次增强的模型。消融实验结果显示如图3所示，所有实验都是在M10数据集上进行的。值得注意的是，所有模型的参数都

经过了精细调整以达到最佳性能，实验设置与前文一致。如预期，完整的语言模型增强连边因子挖掘框架优于所有子模型。原始模型的性能通常低于粗粒度模块，表明粗粒度模块能够挖掘节点间因素信息，从而提高模型的性能。然而，只有当同时挖掘粗粒度和细粒度因素时，才能达到最佳性能。

3.6 超参数实验

笔者研究对语言模型增强连边因子挖掘框架最关键的超参数——细粒度因子数量的影响，通过调整该参数，保证其他超参数在最优设置的情况下，在M10数据集进行了链接预测任务，如图4所示。

细粒度因子数量（表示为 K ）对语言模型增强连边因子挖掘框架的性能影响也在笔者的研究中进行了仔细检查。灵敏度分析结果揭示了在不同模型和训练集比例下，最优因子数量的变化。对于GCN，当训练集比例为10%时，20个因子是最佳选择。而在15%的训练比例下，15个因子表

现最佳。在DGCN中，无论训练集多大，增加因子数量总能显著提升性能。相对而言，GATv2的性能对因子数量的敏感度较低，表明其性能在不同因子数量下变化不大。这些发现为不同需求下模型参数的优化提供了指导。

3.7 可视化分析

在研究中，模型可解释性是评估其实际应用价值的重要因素之一。通过设计具体的案例研究来可视化语言模型增强连边因子挖掘框架挖掘的连边因子，本节旨在深入探讨模型在实际操作中如何揭示学术论文之间的隐藏关系。本文选择Cora数据集作为实验对象，该数据集涵盖了从规则学习到神经网络的7个不同研究领域。

本节选取不同的案例进行粗粒度和细粒度的可视化分析，如图5所示，所有子图均通过DFS（depth first search）策略识别出起始节点及其连接的节点后构建出。图中节点的标号从0至6依次对应案例中粗粒度分析的图例，如标号为3的节点代表属性为Genetic Algorithm的论文。粗粒度视角下，图5（a）表明，同一类别的大部分节点之间的粗粒度连边因子相同，验证了学术论文在引用过程中倾向于引用同领域的其他论文的假设。细粒度视角提供了对节点链接深层次因素的洞见，细粒度连边因子是在数据集中不显式存在的、对数据集中的文本信息进行深度挖掘后产生的隐式连边因子。图5（b）展示了即使在同一领域内，论文之间的联系也可能由不同的潜在因子驱动。例如，即使两篇论文都属于遗传算法领域内容，它们之间的链接也可能由特定的研究方法或实验技术这类更具体的因素所决定，即文本中存在的细粒度因子。图5（c）和图5（d）探

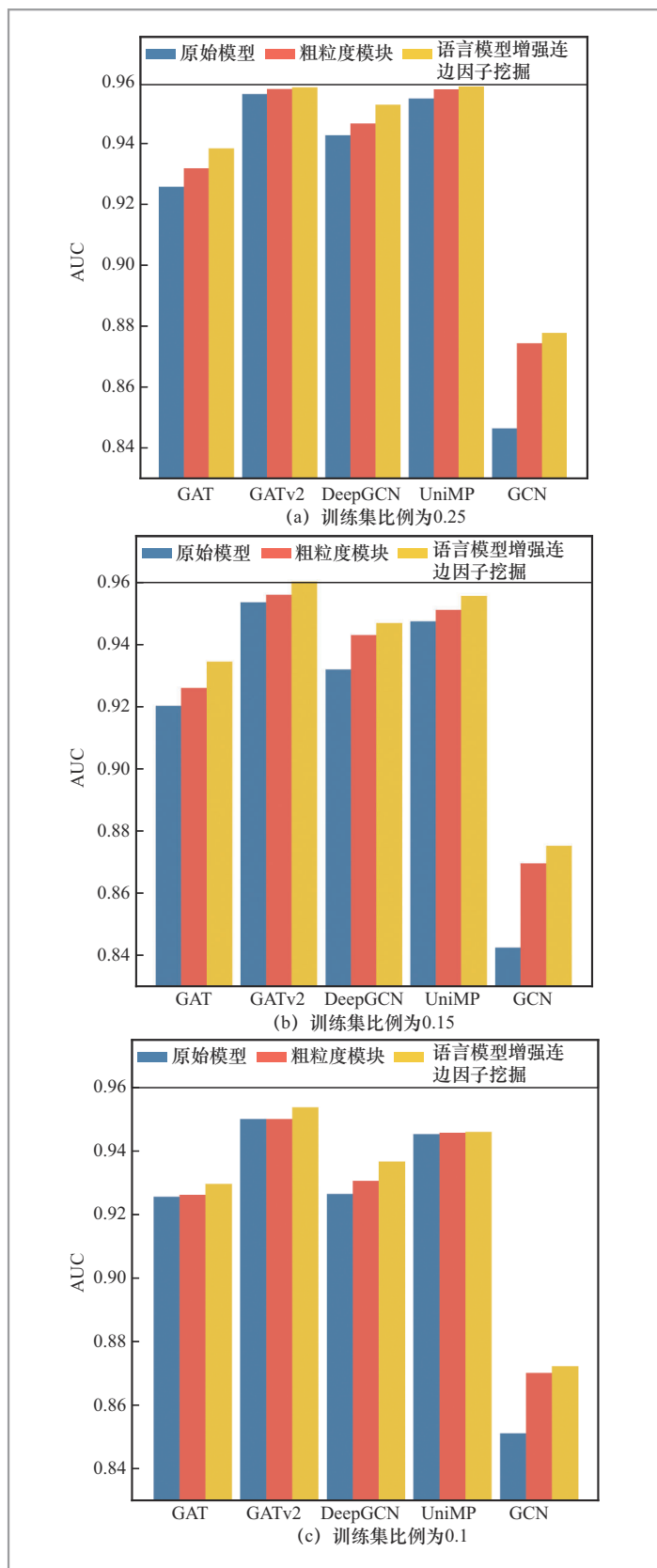


图3 消融实验结果

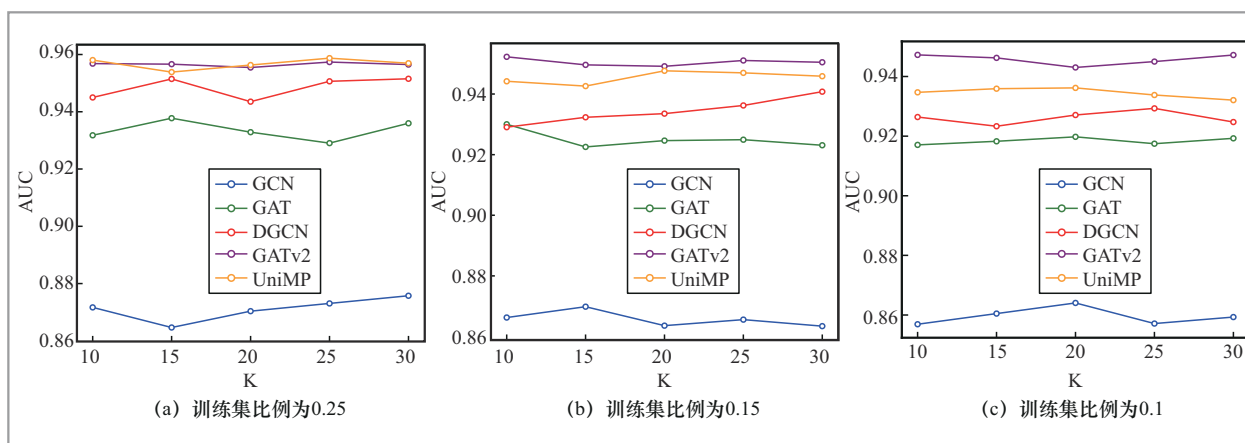


图4 超参数实验

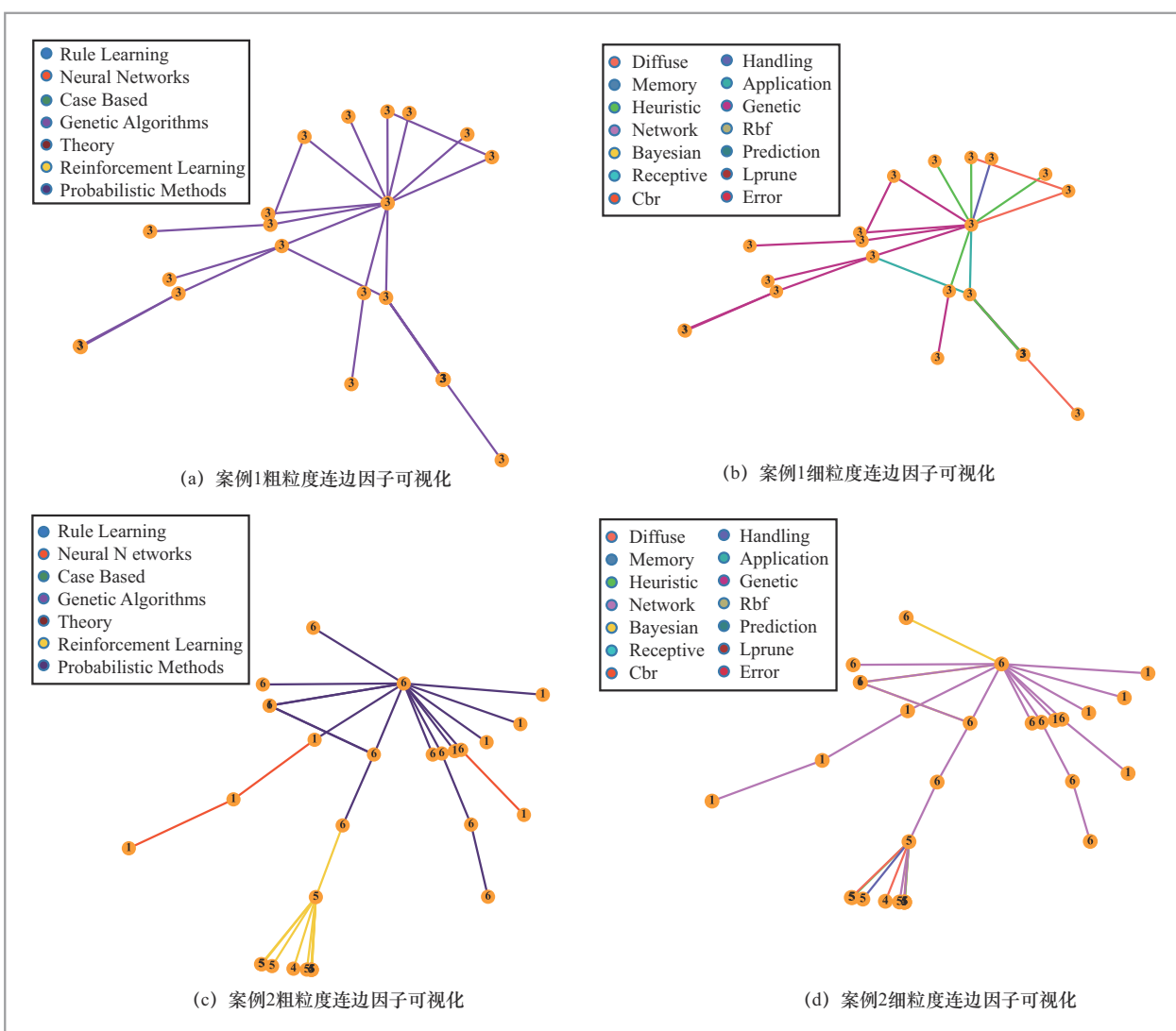


图5 案例可视化分析

讨了跨领域的引用和连接。尽管论文跨越不同的研究领域，但是语言模型增强连边因子挖掘框架能够揭示出促成这些跨领域引用的共同因子。例如，属于增强学习和概率方法的论文可能因为共同关注“网络”而被链接在一起。这种跨领域的链接不仅丰富了模型对学术引用网络结构的理解，也展示了语言模型增强连边因子挖掘框架在挖掘潜在连接因子方面的强大能力。

4 结束语

语言模型增强连边因子挖掘框架利用 LLM 整合两种语义粒度因子的方式来挖掘连边关系的驱动因素：数据集本身的因子类别，以及通过文本聚类算法从潜在文本语义空间提取细粒度因子。充分的实验证明，在与连边挖掘最相关的链接预测任务上，通过连边特征增强后的图表示模型在 3 个真实的引文网络图数据集上均达到了最优的性能。与已有的相关方法相比，本文提出的框架不仅在深入挖掘节点之间关系和连接动机方面表现卓越，还在理解数据背后的形成机制和模式上提供了一定的可解释功能。此外，本文所提框架的可拓展性较强，可以应用到一些类似的研究领域。例如，该模型可以用于构建更丰富准确的知识图谱，帮助推荐系统理解用户需求，以及在社交网络中推荐具有共同兴趣的用户。

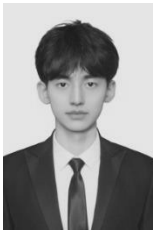
参考文献：

- [1] 李国杰. 大数据与计算模型[J]. 大数据, 2024, 10(1): 9-16.
LI G J. Big data and computing models[J]. Big Data Research, 2024, 10(1): 9-16.
- [2] 林晶晶, 冶忠林, 赵海兴, 等. 超图神经网络综述[J]. 计算机研究与发展, 2024, 61(2): 362-384.
LIN J J, YE Z L, ZHAO H X, et al. Survey on hypergraph neural networks[J]. Journal of Computer Research and Development, 2024, 61(2): 362-384.
- [3] HENRIQUE B M, SOBREIRO V A, KIMURA H. Building direct citation networks[J]. Scientometrics, 2018, 115(2): 817-832.
- [4] 徐立祥, 许巍, 陈恩红, 等. KENN: 线性结构熵的图核神经网络[J]. 软件学报, 2024, 35(5): 2430-2445.
XU L X, XU W, CHEN E H, et al. KENN: graph kernel neural network based on linear structural entropy[J]. Journal of Software, 2024, 35(5): 2430-2445.
- [5] 赵港, 王千阁, 姚烽, 等. 大规模图神经网络系统综述[J]. 软件学报, 2022, 33(1): 150-170.
ZHAO G, WANG Q G, YAO F, et al. Survey on large-scale graph neural network systems[J]. Journal of Software, 2022, 33(1): 150-170.
- [6] ZHAO H K, ZHAO C, ZHANG X, et al. An ensemble learning approach with gradient resampling for class-imbalance problems[J]. Inform Journal on Computing, 2023, 35(4): 747-763.
- [7] 朱健甫, 郝中楷, 刘淇, 等. 通向大分子的探索空间: 一种高效的分子生成方法[J]. 计算机科学技术学报, 2022, 37(6): 1464-1477.
ZHU J F, HAO Z K, LIU Q, et al. Towards exploring large molecular space: an efficient chemical genetic algorithm[J]. Journal of Computer Science and Technology, 2022, 37(6): 1464-1477.
- [8] MENG J Y, YU K, WU X F. Enhancing graph neural networks with edge features through sequential representation[C]// Proceedings of the 2020 IEEE 6th Inter-

- national Conference on Computer and Communications (ICCC). Piscataway: IEEE, 2020: 2110–2114.
- [9] GONG L Y, CHENG Q. Exploiting edge features for graph neural networks[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2019: 9203–9211.
- [10] WU L K, ZHAO H K, LI Z, et al. Learning the explainable semantic relations via unified graph topic-disentangled neural networks[J]. ACM Transactions on Knowledge Discovery from Data, 2023, 17(8): 1–23.
- [11] CHANG Y, WANG X, WANG J, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1–45.
- [12] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB]. arXiv preprint, 2023, arXiv: 2303.08774.
- [13] TIAN Y J, SONG H, WANG Z C, et al. Graph neural prompting with large language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.: s.n.], 2024, 38(17): 19080–19088.
- [14] LLOYD S. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28(2): 129–137.
- [15] NEWMAN M E. Clustering and preferential attachment in growing networks[J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2001, 64(2): 025102.
- [16] ADAMIC L A, ADAR E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3): 211–230.
- [17] KATZ L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39–43.
- [18] BRIN S, PAGE L. The anatomy of a largescale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1–7): 107–117.
- [19] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61–80.
- [20] BRONSTEIN M M, BRUNA J, LECUN Y, et al. Geometric Deep Learning: going beyond Euclidean data[J]. IEEE Signal Processing Magazine, 2017, 34(4): 18–42.
- [21] ZHOU J, CUI G, ZHANG Z, et al. Graph neural networks: a review of methods and applications[J]. AI Open, 2020, 1: 57–81.
- [22] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB]. arXiv preprint, 2016, arXiv: 1609.02907.
- [23] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB]. arXiv preprint, 2017, arXiv: 1710.10903.
- [24] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[EB]. arXiv preprint, 2017, arXiv: 1706.02216.
- [25] KIPF T N, WELLING M. Variational graph auto-encoders[EB]. arXiv preprint, 2016, arXiv: 1611.07308.
- [26] ZHANG M H, CHEN Y X. Link prediction based on graph neural networks[EB]. arXiv preprint, 2018, arXiv: 1802.09691.
- [27] TAN Q Y, ZHANG X, HUANG X, et al. Collaborative graph neural networks for attributed network embedding[EB]. arXiv preprint, 2023, arXiv: 2307.11981.
- [28] WU L K, LI Z, ZHAO H K, et al. Sup-

- porting your idea reasonably: a knowledge-aware topic reasoning strategy for citation recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(8): 4275–4289.
- [29] LIU Z, CHEN C, YANG X, et al. Heterogeneous graph neural networks for malicious account detection[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM, 2018: 2077–2085.
- [30] GUO Y, LING Y, CHEN H. A time-aware graph neural network for session-based recommendation[J]. IEEE Access, 2020, 8: 167371–167382.
- [31] YUN S, JEONG M, KIM R, et al. Graph transformer networks[J]. Advances in neural information processing systems, 2019, 32: 11960–11970.
- [32] VASWANI A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.
- [33] DEVLIN J. Bert: Pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2019, arXiv: 1810.04805.
- [34] AN B, CHEN B, HAN X, et al. Accurate text-enhanced knowledge graph representation learning[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). [S.l.: s.n.], 2018: 745–755.
- [35] XIAO Z, GONG M, WU J, et al. Instructed language models with retrievers are powerful entity linkers[EB]. arXiv preprint, 2023, arXiv: 2311.03250.
- [36] ZHAO H K, ZHENG S M, WU L K, et al. LANE: logic alignment of non-tuning large language models and online recommendation systems for explainable reason generation[EB]. arXiv preprint, 2024, arXiv: 2407.02833.
- [37] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large language models[EB]. arXiv preprint, 2023, arXiv: 2307.06435.
- [38] PAN S, WU J, ZHU X, et al. Tri-party deep network representation[C]//Proceedings of the International Joint Conference on Artificial Intelligence 2016. Montreal: AAAI, 2016: 1895–1901.
- [39] BOJCHEVSKI A, GÜNNEMANN S. Deep gaussian embedding of graphs: unsupervised inductive learning via ranking[J]. arXiv preprint, 2017, arXiv: 1707.03815.
- [40] XIAO S, LIU Z, ZHANG P, et al. C-pack: packaged resources to advance general chinese embedding[EB]. arXiv preprint arXiv: 2309.07597, 2023.
- [41] LI G, XIONG C, THABET A, et al. Deepergcn: all you need to train deeper gcns[EB]. arXiv preprint, 2020, arXiv: 2006.07739.
- [42] BRODY S, ALON U, YAHAV E. How attentive are graph attention networks?[J]. arXiv preprint, 2021, arXiv: 2105.14491.
- [43] SHI Y, HUANG Z, FENG S, et al. Masked label prediction: unified message passing model for semi-supervised classification[EB]. arXiv preprint, 2020, arXiv: 2009.03509.
- [44] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB]. arXiv preprint, 2014, arXiv: 1412.6980.
- [45] HANLEY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (roc) curve[J]. Radiology, 1982, 143(1): 29–36.

作者简介



王懋懋（2003–），男，天津大学经济与管理学部本科生，主要研究方向为数据挖掘、大语言模型。



赵洪科（1988–），男，博士，天津大学管理与经济学部副教授，主要研究方向为数据挖掘、机器学习、算法管理。



吴李康（1996–），男，博士，天津大学经济与管理学部讲师，主要研究方向为数据挖掘、大模型决策。



焦之贤（2001–），男，北京大学光华管理学院硕士生，主要研究方向为量化投资、数据挖掘。



黄振亚（1992–），博士，中国科学技术大学副教授，主要研究方向为数据挖掘与人工智能、知识计算与认知推理、智能教育技术与应用。

收稿日期: 2024-07-16

通信作者: 吴李康, wulk@tju.edu.cn

基金项目: 国家自然科学基金项目(No.72101176)

Foundation Item: The National Natural Science Foundation of China (No.72101176)