

情报理论与实践
Information Studies: Theory & Application
ISSN 1000-7490, CN 11-1762/G3

《情报理论与实践》网络首发论文

题目： 基于端到端图表示学习的新兴主题预测方法研究
作者： 郭思成，梁镇涛，陈传夫
网络首发日期： 2025-07-10
引用格式： 郭思成，梁镇涛，陈传夫. 基于端到端图表示学习的新兴主题预测方法研究[J/OL]. 情报理论与实践.
<https://link.cnki.net/urlid/11.1762.G3.20250709.1536.002>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

●郭思成，梁镇涛，陈传夫（武汉大学信息资源研究中心，湖北 武汉 430072）

基于端到端图表示学习的新兴主题预测方法研究*

摘要：[目的/意义]在科技创新驱动的背景下，准确预测新兴主题有助于把握未来科技发展趋势，对指引研究方向、优化产业布局具有重要价值。[方法/过程]针对新兴主题的产生与专利分类共现间的逻辑一致性，将新兴主题预测任务迁移至链路预测场景，设计了一种端到端的图表示学习新兴主题预测框架，并以基因工程领域专利数据对预测效果进行了验证。[结果/结论]实验结果表明，该方法能够捕捉未来新兴主题，拥有较好的稳健性和泛化能力。预测结果具备科学性，可以成为辅助新兴主题预测的有效方法路径。

关键词：新兴主题预测；图表示学习；专利分类共现网络；链路预测

Research on Emerging Topic Prediction Methods Based on End-to-End Graph Representation Learning

Guo Sicheng, Liang Zhentao, Chen Chuanfu

(Information Resources Research Center of Wuhan University, Hubei Wuhan 430072)

Abstract: [Purpose/significance] In the context of technological innovation, accurate prediction of emerging topics helps to grasp the future development trend of science and technology, and is of great value in guiding research directions and optimizing industrial layout. [Method/process] Aiming at the logical consistency between the generation of emerging topics and the co-occurrence of patent classification, migrating the emerging topic prediction task to the link prediction scenario, designs an end-to-end graph representation learning emerging topic prediction framework, and verifies the prediction effect with patent data in the field of genetic engineering. [Result/conclusion] Experimental results show that this method can capture future emerging topics, and the model has good robustness and generalization ability. The prediction results are scientific and can become an effectively method path to assist the prediction of emerging topics.

Keywords: emerging topic prediction; graph representation learning; patent classification co-occurrence network; link prediction

0 引言

科学知识的创新是当下的关键议题，在促进技术更新和经济增长方面有不可或缺的作用，及时地发现新的科学或技术主题对于科研机构、政府、企业等具有重要意义。新兴主题被视为特定科技领域下受关注、新颖且具备高度发展前景的一类研究主题^[1]，能够代表科学研究中极具发展潜力的研究方向和趋势^[2]。新兴主题作为知识与技术演进的鲜明载体，对其准确预测，能够揭示学术与产业的创新焦点，有助于及时调整策略、优化资源配置，为相关组织提供决策的依据，从而确保竞争中的有利地位。

主题预测是一项基于阶段性数据或信息，预测未来主题或趋势的探索性研究。新兴主题预测则主要关注尚未完全形成的科学研究或技术领域，更具有风险性和未知性。现有的新兴主题预测研究主要通过测度和建模主题的内部属性或关联特征，以量化的工程性方法揭示新兴主题的演化规律、潜在关联或未来发展趋势^[3-6]。多集中在对主题某类属性趋势的预测，较少前瞻性地预测未来新兴主题的具体内容或类型。得益于机器学习、深度学习相关技术在计算机视觉、自然语言处理等领域取得的突破，图表示学习方法在面向网络化的数据结构上具有天然的优势，通过挖掘节点单元间深层关联和复杂模式，可对节点进行含义丰富且高质量的向量化表达^[7]。鉴于此，考虑到基因工程领域研究的创新价值、应用潜力以及跨学科属性等新兴特征，本文构建基因工程领域专利分类共现的技术主题网络，针对技术主题的潜在融合趋势与新兴主题形成的内在逻辑一致性，将新兴主题预测问题形式化理解为图表示学习支持下的IPC共现链路预测任务，在端到端的训练框架下综合设置推导式和归纳式预测方案，探索新兴主题前瞻性预测的有效方法。

1 相关研究现状

1.1 新兴主题的界定及特征

新兴主题的具体定义可能因学科或领域的不同而有所差异。在情报学领域，自 D. J. D. S. Price 开创性地将新兴主题与研究前沿的概念结合，并强调其新颖性以来，研究者们从

*本文为国家自然科学基金创新群体项目“信息资源管理”（项目编号：71921002）和湖北省自然科学基金青年项目“基于‘科学-技术’网络表示学习和语义融合的新兴技术趋势预测”（项目编号：JCZRQN202500694）的成果。

不同视角、背景和方法论层面出发,对这一概念进行了界定与解释^[8]。大部分研究对新兴主题的新颖性特征达成共识,认为其具备与现有知识体系和技术成果的差异性和独特性,对已有的理论框架或技术瓶颈能形成一定的突破,并为该领域的发展提供新的方向和可能性^[9-10]。也有研究者强调,增长性是新兴主题的重要属性之一。尽管该主题尚未成为主流热点,但具有巨大的成长潜力和发展空间,能够快速增长并被广泛应用^[11-12]。新兴主题的影响力也是学术界普遍认可的特点,是新观念、新方法、新发明的综合体现,常常在科学理论的基础上实现突破,或推动现有产业的转型升级,进而对市场经济结构和社会发展产生巨大的潜在影响^[13]。新兴主题通常与市场需求、应用场景紧密相连,具备转化为实际成果的潜力,但其发展往往依赖于基础研究,并可能受到技术发展路径的制约^[14-15]。因此,也往往呈现出显著的连续性特征,它们从现有知识体系中逐步演变,展现出渐进性和持续性。此外,也有学者指出新兴主题通常表现为由特定的关键词或词组形成的一类知识主题^[9,16-17]。随着跨学科合作的日益增多,学科间交叉和知识的融合变得愈加频繁,往往催生新兴的知识领域,这种交叉性的关键词和词组也是新兴主题的重要信号^[18-19]。在衡量新兴主题的标准时,除普遍提及的新颖性、增长性等特征,也可更多地考虑其交叉性。

1.2 新兴主题预测

新兴主题预测常伴随不确定性和复杂性。早期的预测在很大程度上是基于专家经验进行的定性分析,包括德尔菲法^[20]、同行评议^[21]、层次分析^[22]等。在推断的过程中,其辨识的准确性多依赖专家自身的知识广度和深度,主观性较强。

现阶段研究更多是通过界定新兴主题的特征内涵进行属性测度,构建特征指标,预测其未来发展趋势或变化规律,涉及文献计量、深度学习等。如基于新颖性、增长性、影响力等指标构建新兴指数特征^[3],结合主题模型聚类提取主题术语,以深度神经网络进行时间序列预测或曲线拟合等方法^[5,23-24],从而实现对新兴主题的热度^[25]、强度^[26]、突破性^[27]等发展趋势的预测;也有部分研究从网络视角出发,关注主题彼此或与外部因素间的关系,如共现、引用或主题与作者、文献等外部联系^[28-29],通过挖掘主题关系的拓扑结构和动态变化,以揭示新的关联机会或新兴状态,涉及复杂网络分析、链路预测^[30-31],或图神经网络^[32]等方法。相关研究多是以数据驱动为核心的工程性方法的综合运用,预测过程逐渐采用更先进的技术、构建更复杂的模型,提供了多种有效的方法框架。然而,研究目标多侧重于预测新兴主题的某类趋势,而非直接预测未知新兴主题的出现;少部分研究将带有回溯性质的早期探测或演化分析称之为预测,其本质上是一种面向过去的“未来”。

1.3 图表示学习及相关应用

图表示学习,也被称为图嵌入、图神经网络或网络表示学习。其关注的是描述非欧几里得空间中节点关系的网络图,被广泛用于各种数据产生、交换、存储的场景,如通信网络、交通网络、药物分子网络等,其均可以通过图进行表示^[33]。与其他类型的知识表示学习目的类似,图表示学习在保留原网络中复杂的拓扑结构信息的同时将高维特征嵌入低维的欧几里得空间,以低维稠密向量形式对网络进行多维度的信息描述,为节点分类、聚类、链路预测、社区发现等下游任务提供有效的支撑^[34]。

在科技情报分析场景中,图表示学习方法也逐渐得到更多的应用。例如,以 GAE 识别技术融合轨迹并研究相应技术子领域在融合过程中的动态作用^[35];基于知识图谱训练的 TransE 和 SentenceBERT 文本表示功能,设计初始专利集检索的有效方法^[36];基于 Transformer 和图嵌入的专利态势分析自动分类模型,辅助研发期间的信息搜索^[37];基于技术主题时序网络,以 Node2Vec 结合链路预测模型识别未来技术网络^[38];基于科技文献关键词共现加权网络的匿名随机游走序列向量,与公认颠覆性技术网络的关键词向量间的 DTW 距离衡量相似程度,发现潜在颠覆性技术^[39];结合 GraphSAGE 和 PatentBERT 进行图与语义表示学习,优化专利引文网络的链路预测方法^[40];以 CPC 层级信息结合图神经网络进行多维度特征综合表示,在细粒度层预测技术汇聚^[41]。

综上,本文在已有研究基础上界定新兴主题,认为其是一种具备新颖性、增长性、高影响力、交叉性,以主题词或词组进行呈现的领域主题,常伴随学科交叉或知识的融合出现。从网络视角出发,引入图表示学习相关方法结合端到端训练框架,以专利分类潜在融合趋势为依据,预测可能产生的新兴主题,符合其交叉性特点,能够具备一定的

前瞻性。

2 研究设计

2.1 任务定义

新兴主题预测任务的关键之一在于发现领域内尚未显现的潜在趋势和知识演化路径。主题通常是由多个相似性个体知识单元的社区性聚集所形成，在专利分析中，一个 IPC 分类号本身就代表一批有类似性质的知识单元，IPC 间的共现关系可以解释为原有技术主题融合或交互产生新主题的潜在信号^[42-44]。

考虑到新兴主题具备的新颖性、增长性、交叉性等特征，当两个 IPC 分类首次在同一专利中共现时，这种“0”到“1”的转变本身便是一种新颖性的直接体现，从无到有的技术融合过程可以视为新兴技术的种子或萌芽^[45]，且两者在分类体系中的层级距离越远，可被视为不同技术领域间更大跨度的交叉。例如，当传感器技术（G01N）与人工智能（G06N）的初次共现可能预示着物联网感知智能化的新技术方向，当纳米材料技术（B82Y）和生物医学配品（A61K）的共现预测表明跨领域纳米药物递送技术的兴起。选择特定高速发展领域的专利分类为切入点，基于对这种专利分类间潜在融合趋势的分析，可以实现对新兴、跨领域的技术发展趋势、机会以及技术创新路径的探索^[46]，可能会成为预测新兴主题的有效途径。

IPC 共现网络体现了不同技术领域主题下的关联与交叉，可以视为一种技术主题的融合网络，图表示学习技术能够学习到揭示网络节点特征属性的优化表示。因此，本文以 IPC 共现网络为载体，基于技术主题间潜在融合趋势和新兴主题产生间的逻辑一致性定义任务目标，将新兴主题预测迁移到以图表示学习支撑的 IPC 共现链路预测场景中，构建端到端联合训练框架，综合设计全局窗口下的推导式与局部窗口下的归纳式预测方案，研究科学动态的预测方法以实现对新主题的前瞻性预测。

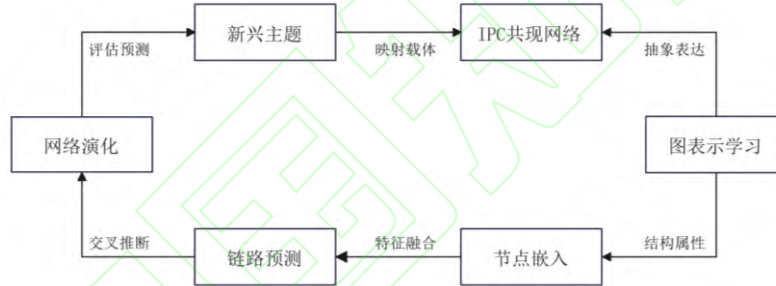


图 1 新兴主题预测任务逻辑

Fig.1 Logic of emerging topic prediction

2.2 模型构建

2.2.1 端到端图表示学习新兴主题预测框架构建 如 2.1 节中所述，在 IPC 共现网络中，新兴主题的产生与技术的融合可能伴随两个 IPC 节点单元间新关系的建立。针对这种逻辑间一致性，本文将新兴主题预测任务形式化为基于 IPC 共现网络的链路预测问题。为了实现该目标，设计了一种端到端的图表示学习预测框架，在关注节点表示生成的同时注重高效利用这些表示进行预测。

该框架包含一个图表示学习模型作为编码器模块以及一个多层感知机作为解码器模块。编码器模块负责将图信息嵌入每个节点的表示中，实现对节点自身及拓扑信息的特征捕捉；解码器模块通接收拼接或点积等形式的节点嵌入组合，将其解码为链路存在的概率，以此实现对新兴主题的预测。该过程中无须预先获得节点的表示，其神经网络从原始的图输入最终预测结果的输出都是一个连续的可微分过程，预测模型会在每一个训练批次内同时优化两个模块中参数，如图 2 所示。

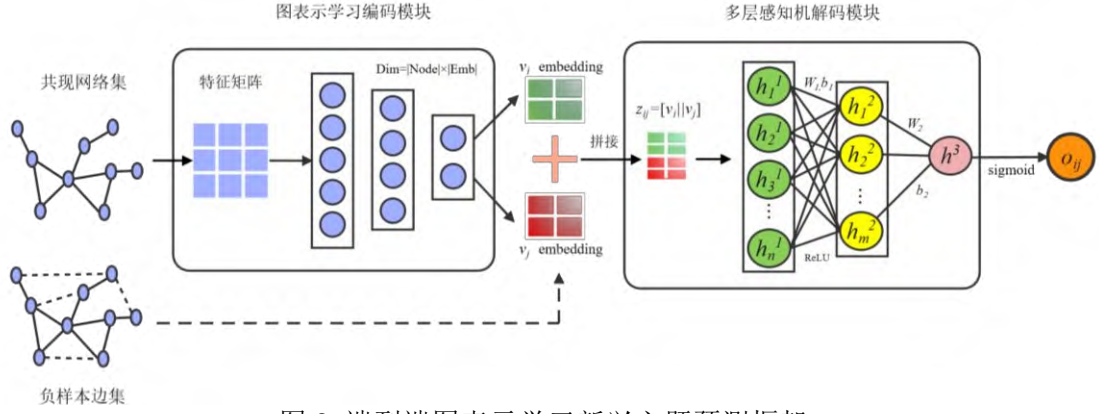


图 2 端到端图表示学习新兴主题预测框架

Fig.2 End-to-end graph representation learning framework for emerging topic prediction

2.2.2 训练过程 本文构建的 IPC 共现网络 $G = (V, E, X)$ 中, 节点 $v_i \in V$ 表示特定的 IPC 号, 其中每条边 $(v_i, v_j) \in E$ 表示节点 v_i 和 v_j 两项 IPC 在某项或多项专利中的共现关系, X 代表相关权重属性, 以节点度作为初始特征。其表示学习任务目标是学习一个函数 $f: v_i \rightarrow v_i \in R^d$, 将原图中的 IPC 节点 V 映射到低维空间 $d \ll |V|$ 中, 得到向量表示 $f(v)$, 实现对其在图中的相对位置、邻近关系或者属性等方面的捕捉。

在图表示学习编码模块, 需要考虑端到端框架下的模型可微分性, 本文拟引入一批业内认可度较高的图神经网络模型, 通过对比筛选出最适用于预测框架的有效方法。涉及的模型包括: GCN^[47]; GAT^[48]; GraphSAGE^[49]; GraphMAE^[50]等。其各自在计算效率、局部抽象、泛化能力、任务适配等方面可能具备独特优势。

基于图表示学习编码后的嵌入节点特征, 对于每对待预测的节点对 (v_i, v_j) , 通过拼接两者表示向量 $z_{ij} = [f(v_i) || f(v_j)]$, 继续传入多层感知机部分进行链路预测, 经过两次权重矩阵和偏置的线性变换, 其间引入 ReLU 以及 sigmoid 激活函数增加非线性, 强化模型的表达能力并确保其输出值 o_{ij} 分布于 $(0,1)$:

$$o_{ij} = \text{sigmoid}[W_2 * \text{ReLU}(W_1 * z_{ij} + b_1) + b_2] \quad (1)$$

预测框架以端到端的训练设置对图表示学习编码模块到多层感知机解码模块整体联动优化。针对观测到的正样本边集合 E^+ 与负样本边集合 E^- , 共同使用二元交叉熵损失函数 L 对模型进行约束, 以衡量模型预测结果与实际链接标签之间的差异:

$$L = -\sum_{(i,j) \in E^+} \log(o_{ij}) - \sum_{(i,j) \in E^-} \log(1 - o_{ij}) \quad (2)$$

实验利用 PyTorch 深度学习框架以及结合传统神经网络和图计算框架特点的 DGL 库相关功能, 构建端到端预测模型, 实现对图数据的定义、训练和推断过程。设置最大迭代次数 $\text{num_epcoh} = 300$, 采用早停策略 $\text{patience} = 10$, 以 Adam 优化器进行参数更新, 损失函数使用 BCE 交叉熵损失, 使用 Huggingface 的 transformers 库的学习率调度方法 $\text{get_linear_schedule_with_warmup}$ 创建动态变换的学习率, 动态学习率最大上限设置为 $\text{lr} = 0.001$, 在该方法下学习率在 $\text{num_warmup_steps} = 15$ 个步中线性增加至最大上限, 之后经历线性衰减为 0, 持续到 $\text{num_training_steps} = \text{num_epochs}$ 。端到端预测模型中图表示学习部分隐藏层和输出层嵌入维度设置为 $d = 768$, 多层感知机维度设置为 $(768 * 2, 128 * 1, 1)$, 图的游走或路径长度以及聚合阶数等长度类参数均设置为 2 (根据 3.1.2 节网络结构测度进行设定)。整体实验流程中涉及随机数种子的环节重复进行 5 次, 以多次的平均结果作为模型的最终性能。

2.2.3 检验方法 实验根据模型输出概率与设定阈值判断未来新兴主题产生的可能性。由于 IPC 共现网络中边数据集的不均衡特点, 对所有节点对而言正样本数量可能远小于负样本, 因此本文选取不过于依赖数据分布特性的 4 项指标查准率 (Precision), 查全率 (Recall), F1 分数, AUC-ROC 进行性能的评估, 如表 1 所示。4 项指标分别衡量了预测为正的样本中

实际是正样本的比例；真正正样本中被预测出来为正的的比例；查准和查全两项指标的调和平均值；ROC 曲线下面积，面积越大模型分类性能越好。

表 1 检验指标及计算方式

Tab.1 Evaluation metrics & calculation

Precision	$\text{Precision} = \frac{TP}{TP + FP}$
Recall	$\text{Recall} = \frac{TP}{TP + FN}$
F1	$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$
AUC-ROC	$\text{AUC} = \sum \left(\frac{\text{FPR}_{i+1} - \text{FPR}_i}{2} \times (\text{TPR}_{i+1} + \text{TPR}_i) \right) \quad \text{TPR} = \text{Recall},$ $\text{FPR} = \frac{FP}{FP + TN}$

3 实验过程及检验

实验流程涉及数据的来源与处理、网络测度、实验设置及结果等内容。详细思路如图 3 所示。

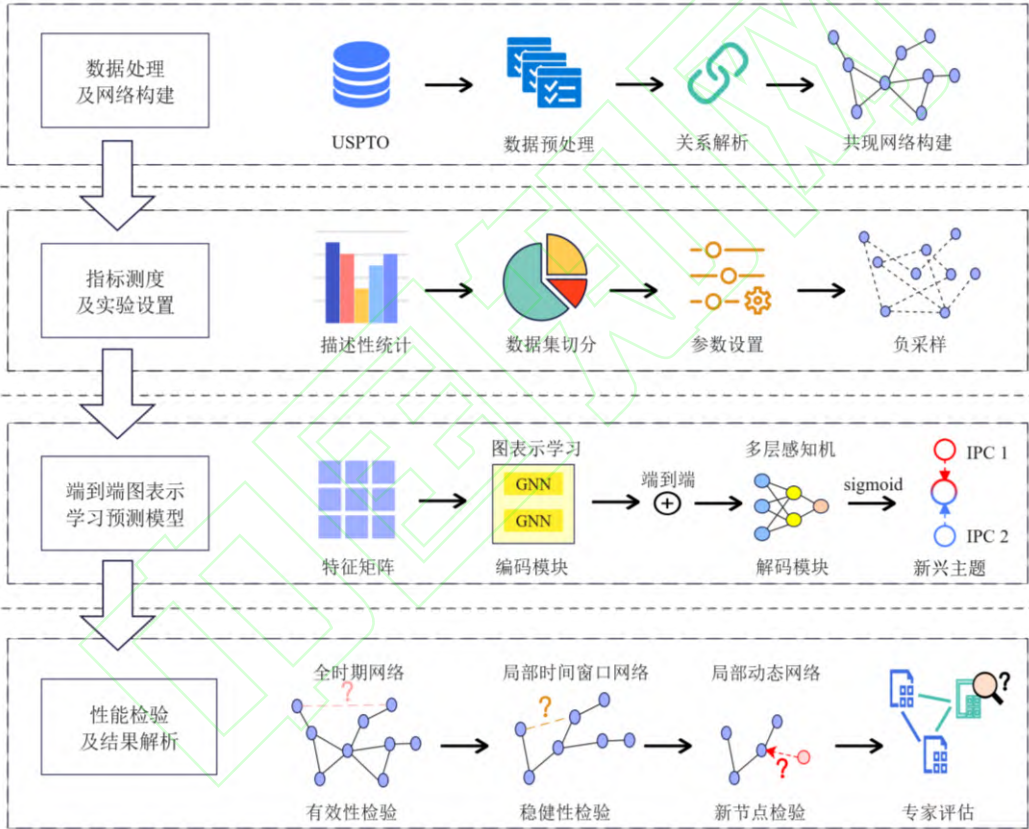


图 3 实验过程及思路

Fig.3 Experimental process & ideas

3.1 数据及统计

3.1.1 领域选取及数据来源 在当下技术革命浪潮中，基因工程领域不断涌现新兴技术和前沿主题，相关技术往往具备高度的创新价值、激增的专利申请数量、广泛的领域覆盖力，逐渐颠覆传统产业链，推动行业格局重塑，成为全球关注焦点。对应了新兴主题的新颖性、增长性、高影响力特征。例如，CRISPR-Cas9 基因编辑技术凭借其其对基因组定点精确修改的可能性，革新了基因操作，相关研究和应用呈现指数级增长，在疾病治疗、农业育种等多个领域产生了深远影响^[51]。此外，基因工程也具有深刻且固有的跨学科性，如与重组 DNA 相关的 CAR-T 细胞疗法同时涉及 IPC 分类 C12N 15/09 和 A61K 35/17 等^[52]。其中重大突破往往源于不同知识基础和方法论的协同组合，正是新兴主题“交叉融合”的直接体现。这些新兴

特征的典型缩影，也意味着相关主题背后的技术原理、专利布局能够成为研究新兴主题的代表性样本，保证了选择该领域数据进行预测方法验证的合理性和普适性。因此，本文以基因工程领域的 IPC 分类号为基础开展研究。基于 USPTO 专利数据库进行数据获取，通过选取 4 位国际专利分类号 IPC 为 C12N 15 下的所有相关专利并以美国专利分类为 435/440 的专利进行补充，考虑到结果的可验证性，时间范围确定在 1980—2020 年，经清洗共得到专利数据 49416 项。结合官方文档对分类体系规则的定义，上述分类号涵盖了对 DNA 或 RNA 等遗传物质进行基因结构修饰的过程。这类过程在自然界中通常不会自然发生，涉及如基因工程、突变、细胞融合及基因编辑等新兴技术。

3.1.2 网络相关指标测度 通过对每项专利对应的 IPC 号进行梳理，抽取其中 IPC 共现关系，以 Networkx 库构建 IPC 共现网络并对相关指标进行测度，包括图的规模、密度、度分布、平均路径长度等、这些指标将具体影响模型训练的相关过程。其中密度、聚类系数等可反映网络呈现的最基本的“小世界”特征、平局路径长度影响到生成游走序列或聚合范围的窗口大小，如表 2 所示。

表 2 IPC 共现图结构指标测度
Tab.2 IPC co-occurrence graph structural index measurement

节点数	3925
边数	108659
密度	0.014
聚类系数	0.7719
平均路径长度	2.401

由上可知，所构建的 IPC 共现网络具备 3925 个节点和 108659 条边，是一个相对大型的网络。其密度为 0.014，表明各 IPC 节点之间的连接相对稀疏，然而聚类系数 0.7719 却是一个相对高的聚类系数，这意味着网络中的节点间是具有彼此紧密相连的倾向，这说明网络可能由许多不同的高连通性的小型社区构成，而彼此之间的连接较少；平均路径长度为 2.401 表示网络中的连通分量内节点到任意的另一个节点的平均距离约为 2.4，对该规模的网络来说，呈现出的是一个相对很短的路径。总体上可认为在该网络中，不同的 IPC 节点在多个局部间形成较密集的汇聚，不同的局部间也存在相对较少的联系，具备“小世界”性质。此外，通过测度网络节点度分布情况对其无标度属性进行验证，结果如图 4 所示，其展现出一个具有长尾特征的幂律分布。综上可以推断所构建的 IPC 共现网络具有无标度特性，少数中心节点拥有较高的节点度，网络具备一定的自组织性和韧性，这种结构的普遍性也表明该网络与许多其他科学知识网络或真实网络在生长机制上有着高度的相似性，随机删除或添加节点不会对网络整体结构造成过大波动，这也是后续预测实验展开的重要保证。

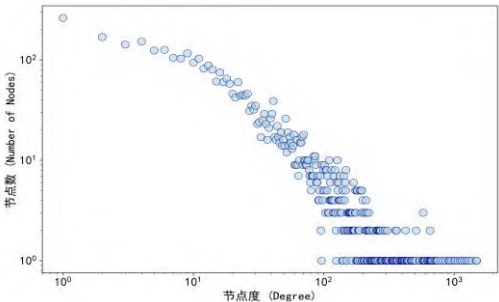


图 4 IPC 共现网络节点度分布
Fig.4 IPC co-occurrence network nodes degree distribution

3.2 实验设置

3.2.1 预测方案及抽样原则 参考无监督网络表示学习的设置思路^[53]，首先以全时段数据集构建 IPC 共现网络检验预测方法有效性，数据划分上采用设置随机数种子的方式，以不重复随机抽样将正样本按 8:2 的比例划分为训练集与测试集，同时在训练集中抽取 10%的边作为验证集正样本，保证彼此间的独立与互斥性。同时，划分 4 个连续的 10 年局部时间窗口检验模型的稳健性，排除因为特定网络结构偶然带来的良好性能。此外，

进一步在全年度窗口下分别设置 3 个部分重叠的 20 年的时间窗口，以每个窗口中前 15 年数据作为训练基础预测后 5 年中可能出现的新主题，检验模型在“冷启动”环境中新节点加入情况下的泛化能力。综合使用推导式和归纳式的预测方案切分数据集对本文所提出的新兴主题预测方法进行动态验证，具体如图 5 所示。

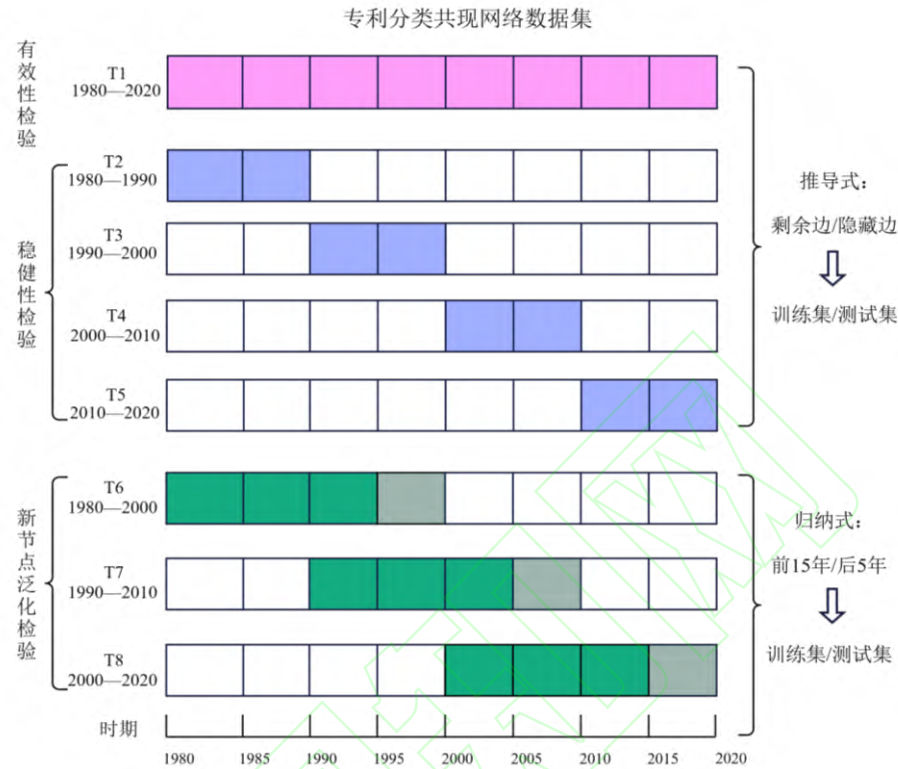


图 5 预测方案及数据集切分策略

Fig.5 Prediction scheme and data set segmentation strategy

由于 IPC 共现网络呈现的无标度与“小世界”特性，新增共现边所产生的新兴主题对于网络的全部节点对而言只能占据很小的比例。具体到链路预测的任务场景中，这种不平衡表现在已有的真实链接和未知可能存在链路的失衡，因此在对负样本的抽取策略上也应还原网络演化的自然规律。为探知真实的链接新增机制，本文以 2005 年为时间节点，对 2006 年新增的所有连边进行分析，通过计算新增边节点对在 2005 年之前原网络上的最短路径长度 d ，对新增连边的特征规律进行总结，结果如图 6 所示。由图 6 可知，2006 年中有 76.8% 的新增连边的节点对在之前的网络中最短路径长度为 2，剩余的少量节点对分布在最短路径长度为 3~5 的区间中，还有约 20% 的新增连边是伴随着新节点的加入。因此，负样本采样过程中确保在训练集、验证集、测试集生成不少于正样本数量负边的前提下，有 80% 的负样本是来自原网络中最短路径长度为 2 的节点对，以此模拟真实环境。避免新增连边差异过大，导致负样本易于区分。

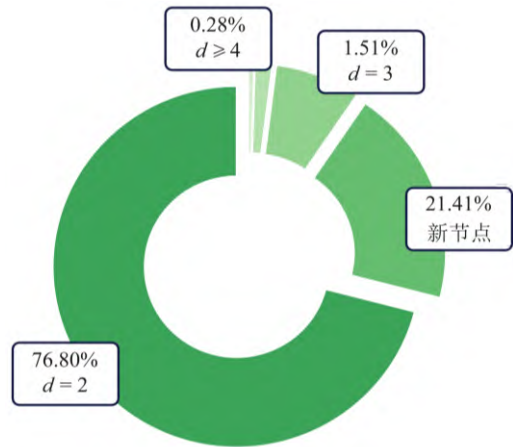


图6 节点间新增链接最短距离分布

Fig.6 Distribution of the shortest distance of new links

3.2.2 对比方法 根据 2.2.2 节中涉及的 5 种不同图表示学习方法分别构建端到端预测模型。此外，除可微分的端到端预测框架外，加入相对较传统的基于矩阵分解的表示方法和基于拓扑结构相似性的特征构建方法，以及基于随机游走的图表示学习方法作为特征输入，下游同样接入多层感知机从而设计对比模型，进一步比较各模型的性能表现，验证最有效的模型构建方式。

1) 基于矩阵分解的方法。通过对图的邻接矩阵或相关的矩阵进行分解，得到节点低维嵌入并以此作为特征。实验选择 SVD 奇异值分解和 NMF 非负矩阵分解两种方法。

2) 基于拓扑结构相似性的方法。基于拓扑结构指标构建特征，根据节点对在网络中的结构相似性程度衡量其间存在链接的可能性。实验选择局部结构相似性指标 CN、AA、RA、JA 以及全局信息相似性指标 PageRank 和 Katz 共同进行，在标准化后为节点对生成一个[CN,AA,RA...]拼接特征向量。

3) 基于随机游走图表示学习的方法。本质上是将基于图的随机游走序列视为类似文本中词的上下文环境，以词嵌入工作中经典的 Skip-gram 模型进行训练。实验选择了 Deepwalk 和 Node2Vec 两种方法。

3.3 实验结果

3.3.1 有效性检验 以全时期数据为基础，验证了使用上述模型进行新兴主题预测的有效性，以 4 类不同的 9 种模型进行总结，各项性能指标如表 3 所示。

表 3 多方法对比的新兴主题预测结果评价

Tab.3 Evaluation of emerging topic prediction results

Method	Model	Precision	Recall	F1	ROC-AUC
矩阵分解	奇异值分解 SVD	0.5297	0.6174	0.5702	0.6551
	非负矩阵分解 NMF	0.5882	0.5495	0.5682	0.6778
相似度计算	拓扑结构指标相似度	0.6925	0.6227	0.6557	0.7408
随机游走图表示学习	DeepWalk	0.6528	0.7021	0.6766	0.7420
	Node2Vec	0.7173	0.7295	0.7233	0.7948
端到端图表示学习	GCN	0.8090	0.8169	0.8129	0.8923
	Gat	0.8125	0.8299	0.8211	0.8998
	GraphMAE	0.8304	0.8316	0.8310	0.9079
	GraphSAGE	0.8233	0.8469	0.8349	0.9150

从实验结果可以看出：

作为图表示学习的对比方法中，矩阵分解和相似度计算在性能上表现相对较弱，模型对捕获到的超过一半的实际正向链接中有 40%以上都返回了错误的结果。可能是由于模型在一定程度上欠缺对图中深层结构或节点间动态交互信息的感知能力。

基于随机游走的图表示学习方法中，Node2Vec 在本任务下的表现各方面略优于 DeepWalk，但两者 4 项指标均实现了对前面模型的超越。这类方法能够在一定程度上提升预测性能，这可能是因为它们能更好地保留图的结构信息，并更敏锐地捕捉到节点间的非线性复杂关系。

基于端到端图表示学习的 4 类模型在全部指标上均呈现出中更优秀的表现，GCN 和 GAT 通过图卷积操作和注意力机制，能够灵活地处理图中节点之间的深层关联，而 GraphSAGE 和 GraphMAE 则通过多种集成策略进一步提升了模型的性能，具体到每个模型的指标项都在 0.8 以上。这表明端到端的方法通过在网络参数上的直接优化，可能更能捕获到图中的高阶邻居关系、聚集结构等复杂模式。其中，GraphMAE 的查准率在所有方法中达到最高的 0.8304，GraphSAGE 尽管稍低，但 4 项指标的综合表现达到了所有方法中的最高峰值，在这种设置下进行新兴主题预测工作可能更具备科学性。

3.3.2 稳健性检验 为了确保模型在不同数据结构和分布下的鲁棒性，本文采取了基于年限的数据集再分割策略，引入多个连续的局部时间窗口来对上节中最优模型性能进行更深入的验证。检验模型是否仅在某一特定数据集结构下表现出色，或是有效地排除因某些特定的网络结构偶然产生的优越性能，从而确保模型的预测能力在各种不同的场景中都是稳健的。

通过对 4 个不同子时间窗口下，不同规模不同分布结构的 IPC 共现网络分别以端到端的 GraphSAGE 模型进行新兴主题预测实验，结果如表 4 所示。

表 4 稳健性检验数据集分布及结果评价

Tab.4 Robustness test dataset distribution and evaluation

时间	网络规模(节点数/边数)	Precision	Recall	F1	ROC-AUC
T2(1980—1990)	320 / 1386	0.6968	0.7572	0.7258	0.8108
T3(1991—2000)	778 / 4900	0.7767	0.8538	0.8134	0.8926
T4(2001—2010)	963 / 18278	0.8113	0.8725	0.8408	0.9018
T5(2011—2020)	3664 / 95578	0.8200	0.8575	0.8383	0.9118

根据结果可知，4 个时间窗口内的模型基本呈现出了良好的性能，其中 1980—990 年由于网络规模较小，训练数据可能有所不足，依然保持了 0.7258 的 *F1* 值和 0.8108 的 ROC-AUC 分数，在后续 3 个时间段内随着数据量增加模型性能也呈逐步上升趋势，综合来看模型整体表现良好，各项性能指标与有效性校验中基本一致。基于此，通过排除时间因素和数据结构带来的随机性，验证了以端到端的 GraphSAGE 模型进行新兴主题预测的稳定效果。

3.3.3 新节点泛化检验 在新兴主题的产生和演化的初期阶段，往往与现有知识体系间发生互动融合、建立关联，该过程通常伴随全新的“冷启动”节点的加入。考虑到 GraphSAGE 能够通过聚合邻居信息更新节点表示的特点，如果测试集中的“冷启动”节点能与训练过程中的“旧”图建立关联，那么模型对这些新节点的预测能力也有望得到验证。因此，本文以归纳式的任务设置检验模型对于新节点加入情况下的泛化能力。这实际上是对模型的鲁棒性以及泛化能力提出了更高的考验，需要模型在面对更多未知的不确定环境下，仍能提供有意义和准确的预测效果。

本文分别在 3 个时期内对模型面对新节点加入后的泛化能力进行验证，为了避免数据泄露和保持预测过程的规范性，测试集只考虑每个时期内的后 5 年网络中的新增边，以反映新兴主题的出现。同时，考虑到不同类型新增边存在预测难度的差异，两端全由“冷启动”的未知节点组成的“纯新边”可能更难被准确预测。因此，除了统计后 5 年中未在训练集中出现的新加入节点外，也对这种“纯新边”在测试集中所占的比例进行了计算。表 5 显示了各时期数据集的网络规模及对应评估结果。

表 5 泛化检验数据集分布及评价

Tab.5 Generalization test dataset distribution and evaluation

时期	前 15 年 (节点/边)	后 5 年 (节点/ 边)	新加入 节点 数	纯新边/ 新增边	纯新 边比 例 (%)	Precisi on	Reca ll	F1	ROC -AU C
T6(1980— 2000)	687/5011	906/7220	505	194/5510	3.5	0.7017	0.7606	0.7300	0.8031
T7(1990— 2010)	1288/18571	998/12013	408	236/7997	2.9	0.7888	0.7159	0.7506	0.8173
T8(2000— 2020)	2446/49477	3452/85756	1847	6203/65287	9.5	0.7014	0.6533	0.6765	0.7463

结果表明，3 个时期下的模型性能相比于有效性检验环节中以全时期数据进行的推导式任务 (*F1*: 0.8349; ROC-AUC: 0.9150) 都出现了下滑。其中 T6 和 T7 的模型 *F1* 都处于 0.75 左右，而 T8 时期下降的程度最高达到了 15.84%。T6、T7、T8 中纯新边的比例依次为 3.5%、2.9%和 9.5%，相比训练数据规模的增大，“纯新边”比例的增长与模型性能的下降呈现出更多的一致性，这可能是由于 GraphSAGE 主要通过新节点的邻居聚合对未知节点进行表示，“纯新边”意味着这样的节点在“旧”图中的邻居更少。同时，结合图模型的 ROC-AUC 分数可以得知，尽管随着越来越多的新节点加入，模型的预测准确性会出现下降，但还是能够保证一个相对较高的基本性能。其中，T7 的 AUC 分数最高为 0.82，其次是 T6 的 0.80，T8 的 AUC 分数为 0.75。验证了模型面对未知新节点具备一定的泛化能力，表明该方法面对网络动态演化中缺少历史数据的挑战时，能够

在一定程度上处理或减轻“冷启动”造成的干扰。

4 预测结果解析

基于训练好的模型，将网络中所有尚未产生链接的节点对（即不存在的边）输入模型进行预测。根据模型输出的预测得分，选择得分高于 0.75 的预测结果，继而筛选出其中节点由 C12N 15 基因工程领域和其他的交叉领域构成的 IPC 节点对，这样的节点对可能更具备成为新兴主题或新兴技术趋势的潜力。表 6 展示了部分得分较高的节点对。

表 6 高得分预测节点对

Tab.6 High-scoring predicted node pairs

序号	节点对		序号	节点对	
	节点 1	节点 2		节点 1	节点 2
1	C12N 15/82	C07D 215/22	12	C12N 15/60	A61K 47/28
2	C12N 15/09	C08F 210/18	13	C12N 15/55	A61P 37/06
3	C12N 15/82	A61K 38/38	14	C12N 15/31	A61K 9/08
4	A01N 43/78	C12N 15/54	15	C12N 15/13	C07C 271/16
5	C12N 15/88	C12P 17/18	16	C12N 15/861	A61Q 5/12
6	C12N 15/09	C08L 5/04	17	C12N 15/12	A61K 31/221
7	C12N 15/89	G01N 33/564	18	C12N 15/52	C07D 333/38
8	C12N 15/74	C12P 21/02	19	A61K 9/107	C12N 15/64
9	C12N 15/53	B01D 57/02	20	G01N 33/554	C12N 15/19
10	C07K 14/54	C12N 15/863	21	C12N 15/87	A61K 35/14
11	C12N 15/86	A61K 8/65	22	C12N 15/82	A61P 37/06

本文邀请了两位基因工程领域专家对预测节点对进行评分，以验证所预测新兴主题的合理性。专家评分从科学性、可解释性和实用性 3 个维度进行评估，每个维度的评分范围为 1 至 5 分。评分矩阵如图 7 所示。同时，根据专家评分结果，以 Spearman 相关系数对两位专家评分矩阵进行一致性检验，以进一步验证预测结果。结果显示相关系数为 0.41， $P\text{-value}=0.0073<0.05$ 。可以认为专家评分间确实存在统计学显著的一致性，表明预测的节点对在科学性、可解释性和实用性方面得到了基本的认可，预测方法具有一定的有效性和科学性。其中两位专家对第 3、13、21 组的节点对给出了较高得分。

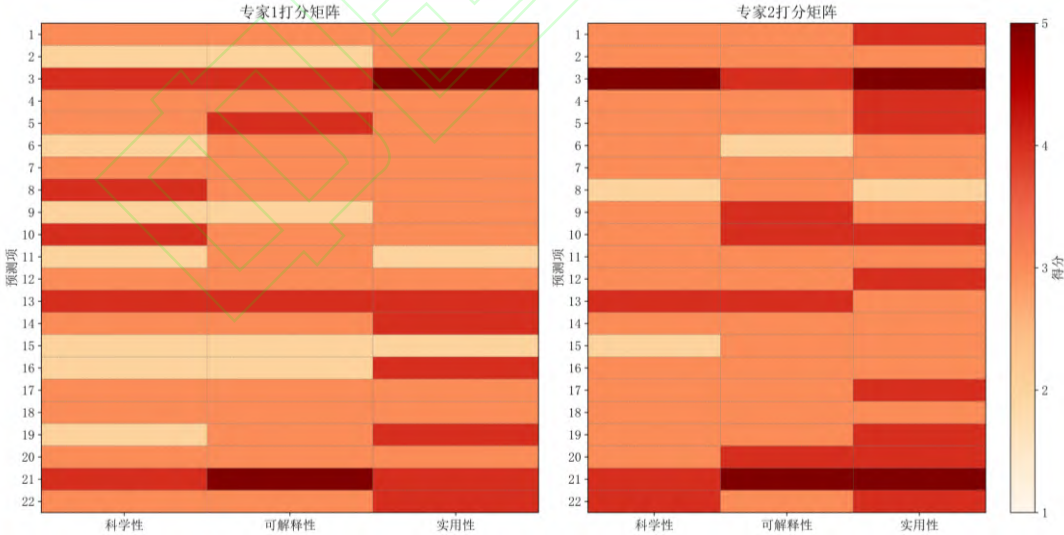


图 7 专家评分热力矩阵

Fig.7 Expert rating heat matrix

针对评分矩阵中共同得分较高的几项节点对展开进一步分析，在解释其合理性的同时更深入地挖掘主题背后的技术支撑、形成因素等重要内容，如：

第三组节点对 C12N 15/82 与 A61K 38/38 中，C12N 15/82 关注植物细胞的基因编辑，A61K 38/38 涉及三类黄连木科植物的肽类药物制剂。该节点对间可能会构成一种基于植物细胞基因编辑技术提取特定肽类药物制剂的新兴技术主题。传统上，肽类药物的生产

通常涉及细菌或真菌等微生物的发酵过程，或者通过化学合成的方法。然而，某种植物可能含有一种具有治疗潜力的肽，但其在自然状态下的浓度很低，但伴随着这项新兴技术，基于植物细胞基因编辑的方法可能会提供一种新的途径，使其更容易被提取并用于相关药物的生产。

第13组节点对C12N 15/55与A61P 37/06中，C12N 15/55代表了编码水解酶的基因技术，A61P 37/06涉及用于降低或抑制人体免疫系统活性的药物。二者间的高概率连边可能预示着会产生一种基于水解酶基因编辑与免疫抑制药物优化的结合的新兴技术主题。以往免疫抑制药物的剂量和效果通常难以精确控制，容易导致免疫系统过度抑制或药物的短时间内消失等副作用。通过基因工程技术定向编辑水解酶的表达，能够调控免疫抑制药物的代谢途径，控制药物在体内的降解速度或降解程度，从而优化药物的作用时间和强度，可能会提高免疫系统相关疾病治疗的精确性和安全性。

第21组节点对C12N 15/87与A61K 35/14中，C12N 15/87涵盖了使用如共转化这类非传统的基因引入方法导入外源遗传材料的技术，A61K 35/14包含了血液及人造血液有关的药物制剂。结合这两个IPC，可能会诞生一种以共转化技术实现人造血液的遗传优化的新兴技术主题。人造血液是指通过体外培养和工程改造的细胞系统，具备类似于自然血液的功能和特性。而这项新兴技术以共转化引入或调控与血液功能相关的基因，将外源遗传材料导入与血液相关载体从而改善人造血液的特性和性能，可能会使其在医疗过程中更为安全、高效。

5 结束语

深入理解并准确预测新兴主题，对于揭示相关行业特定期内的创新焦点，为未来的科技与知识发展趋势提供具体方向具有关键意义。图表示学习在处理泛在的网络化科技知识信息上有显著的优势，能够捕获数据的复杂结构和隐藏模式，充分利用其中深层的关联性和结构性信息，为新兴主题的预测提供了创新的视角和技术支持。本文构建了基因工程领域专利分类共现网络作为新兴主题映射延伸的基本载体，针对新兴主题的产生与专利分类共现间的逻辑一致性，以图表示学习技术展开了对新兴主题预测的方法研究，提出了一种将新兴主题预测问题转化为以图表示学习实现链路预测的研究思路。研究设计了端到端的图表示学习新兴主题预测框架，确保在获得最佳预测效果的前提下对模型整体进行高效的共同训练。通过多种方法的有效性对比，确定了该预测框架下的最优模型组合，并以多种检验确保了模型在不同时期数据结构上的稳健表现，以及面对“冷启动”问题中部分新节点加入时的泛化能力，同时对预测结果进行了进一步解读。本文研究的设计在一定程度上保证了预测的前瞻性，能够动态的感知未来会产生新兴主题类型。但其中存在两个方面的局限性：①本文的预测方法更适用于前期主题交叉关联的合并模式下产生的新兴主题的预测场景。虽然模型在一定程度上具备面对新加入专利分类号节点的动态表示能力，但尚且缺失对于完全脱离现有网络的新主题的预测能力。②目前主要聚焦在单一来源的专利信息，多源异构的关联信息加入可能会辅助提升预测的效果，如结合期刊等出版物以科学—技术异质性节点间的紧密关联补充对同一实体单元的细节描述。后续研究可以设计异质网络下的新兴主题预测方法框架，提升模型面对多源异构场景的预测能力；同时，大语言模型由于其优秀的推理和生成表现已逐渐被应用于情报分析中，可考虑有效的策略进行结合，起到进一步的优化效果。

参考文献

- [1] XU Haiyun, WINNINK J, YUE Zenghui, et al. Multidimensional Scientometric indicators for the detection of emerging research topics[J]. Technological Forecasting and Social Change, 2021, 163: 120490.
- [2] MATSUMURA N, MATSUO Y, OHSAWA Y, et al. Discovering emerging topics from WWW[J]. Journal of Contingencies and Crisis Management, 2002, 10(2): 73-81.
- [3] XU Shishuo, LIU Jinbo, LI Songnian, et al. Exploring and visualizing research progress and emerging trends of event prediction: a survey[J]. Applied Sciences, 2023, 13(24): 13346.
- [4] LIU Han, LIU Ying, WANG Yonglian, et al. Hot topics and emerging trends in tourism forecasting research: A scientometric review[J]. Tourism Economics, 2019, 25(3): 448-468.
- [5] LIANG Zhentao, MAO Jin, LU Kun, et al. Combining deep neural network and bibliometric indicator for emerging research topic prediction[J]. Information Processing & Management, 2021,

58(5): 102611.

- [6] CHI Yankin, WONG Raymond, SHEPHERD J. Popularity forecasting for emerging research topics at its early stage of evolution[C]//International Conference on Advanced Data Mining and Applications. Cham: Springer Nature Switzerland, 2022: 290-303.
- [7] XU Mengjia. Understanding graph embedding methods and their applications[J]. SIAM Review, 2021, 63(4): 825-853.
- [8] PRICE D J D S. Networks of scientific papers: the pattern of bibliographic references indicates the nature of the scientific research front[J]. Science, 1965, 149(3683): 510-515.
- [9] SMALL H, BOYACK K W, KLAVANS R. Identifying emerging topics in science and technology[J]. Research Policy, 2014, 43(8): 1450-1467.
- [10] XU Shuo, HAO Liyuan, AN Xin, et al. Emerging research topics detection with multiple machine learning models[J]. Journal of Informetrics, 2019, 13(4): 100983.
- [11] COZZENS S, GATCHAIR S, KANG J, et al. Emerging technologies: quantitative identification and measurement[J]. Technology Analysis & Strategic Management, 2010, 22(3): 361-376.
- [12] TU Yining, SENG Jialang. Indices of novelty for emerging topic detection[J]. Information Processing & Management, 2012, 48(2): 303-325.
- [13] PORTER A L, GARNER J, CARLEY S F, et al. Emergence scoring to identify frontier R&D topics and key players[J]. Technological Forecasting and Social Change, 2019, 146: 628-643.
- [14] ROTOLO D, HICKS D, MARTIN B R. What is an emerging technology?[J]. Research Policy, 2015, 44(10): 1827-1843.
- [15] 许海云, 王超, 龚兵营, 等. 科学-技术关联视角下的新兴研究主题产学研合作适用模式研究——以干细胞研究领域为例[J]. 图书情报工作, 2022, 66(15): 3-1.(XU Haiyun, WANG Chao, GONG Bingying, et al. Research on the applicable mode of industry-university-research cooperation of emerging research topics based on science-technology linkage: a case study of stem cell research[J]. Library and Information Service, 2022, 66(15):3-1.)
- [16] GLÄNZEL W, THIJS B. Using 'core documents' for detecting and labelling new emerging topics[J]. Scientometrics, 2012, 91(2): 399-416.
- [17] 黄鲁成, 唐月强, 吴菲菲, 等. 基于文献多属性测度的新兴主题识别方法研究[J]. 科学与科学技术管理, 2015, 36(2): 34-43. (HUANG Lucheng, TANG Yueqiang, WU Feifei, et al. Research on identification of emerging topics based on multi-attribute measurement of literature[J]. Science of Science and Management of S.&T. 2015, 36(2):34-43.)
- [18] 冷伏海, 王立学. 科技情报研究发展趋势与应用环境分析[J]. 图书情报工作, 2010, 54(4): 9-12, 60.(LENG Fuhai, WANG Lixue. S&T information study: tendencies and applied environments analysis[J]. Library and Information Service, 2010, 54(4): 9-12, 60.)
- [19] PORTER A L, RAFOLS I. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time[J]. Scientometrics, 2009, 81(3): 719-745.
- [20] TSENG Fangmei, CHENG Aichia, PENG Yinung. Assessing market penetration combining scenario analysis, Delphi, and the technological substitution model: the case of the OLED TV market[J]. Technological Forecasting and Social Change, 2009, 76(7): 897-909.
- [21] CHOUDHURY A K, SHANKAR R, TIWARI M K. Consensus-based intelligent group decision-making model for the selection of advanced technology[J]. Decision Support Systems, 2006, 42(3): 1776-1799.
- [22] LEE S, KIM W, KIM Y M, et al. The prioritization and verification of IT emerging technologies using an analytic hierarchy process and cluster analysis[J]. Technological Forecasting and Social Change, 2014, 87: 292-304.
- [23] XU Shuo, HAO Liyuan, YANG Guancan, et al. A topic models based framework for detecting and forecasting emerging technologies[J]. Technological Forecasting and Social Change, 2021, 162: 120366.
- [24] YANG Zhenyu, ZHANG Wenyu, WANG Zhiming, et al. A deep learning-based method for predicting the emerging degree of research topics using emerging index[J]. Scientometrics, 2024, 129(7): 4021-4042.
- [25] 陈稳, 陈伟. 基于计量指标多变量 LSTM 模型的新兴主题热度预测研究[J]. 数据分析与知识发现, 2022, 6(10): 35-45. (Chen Wen, CHEN Wei. Predicting popularity of emerging topics with multivariable LSTM and bibliometric indicators[J]. Data Analysis and Knowledge Discovery, 2022, 6(10): 35-45.)

- [26] 叶光辉,王灿灿,李松烨. 基于 SciTS 会议文本的跨学科科研协作新兴主题识别及预测[J]. 情报科学, 2022, 40(7): 126-135. (YE Guanghui, WANG Cancan, LI Songye. Recognition and prediction of emerging topics in interdisciplinary scientific research collaboration based on SciTS conference text[J]. Information Science, 2022, 40(7): 126-135.)
- [27] 曹艺文,许海云,武华维,等. 基于引文曲线拟合的新兴技术主题的突破性预测——以干细胞领域为例[J]. 图书情报工作, 2020, 64(5): 100-113.(CAO Yiwen, XU Haiyun, WU Huawei, et al. Study on radical innovation prediction to emerging technology topics based on citation curve fitting: taking the field of stem cells as an example[J]. Library and Information Service, 2020, 64(5): 100-113.)
- [28] SHIBATA N, KAJIKAWA Y, TAKEDA Y, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications[J]. Technovation, 2008, 28(11): 758-775.
- [29] HUANG Lu, CHEN Xiang, NI Xingxing, et al. Tracking the dynamics of co-word networks for emerging topic identification[J]. Technological Forecasting and Social Change, 2021, 170: 120944.
- [30] XU Haiyun, LUO Rui, WINNINK J, et al. A methodology for identifying breakthrough topics using structural entropy[J]. Information Processing & Management, 2022, 59(2): 102862.
- [31] 刘俊婉,龙志昕,王菲菲.基于 LDA 主题模型与链路预测的新兴主题关联机会发现研究[J]. 数据分析与知识发现,2019,3(1):104-117.(LIU Junwan, LIU Zhixin, WANG feifei. Finding collaboration opportunities from emerging issues with LDA topic model and link prediction[J]. Data Analysis and Knowledge Discovery, 2019,3(1):104-117.)
- [32] 段庆锋,陈红,闫绪娴,等.基于图神经网络异构数据融合的学科新兴主题探测研究[J].情报杂志,2023,42(12):127-133. (DUAN Qingfeng, CHEN Hong, YAN Xuxian, et al. Detecting scientific emergency topic based on heterogeneous data fusion using GCN[J]. Journal of Intelligence, 2023,42(12):127-133.)
- [33] GOYAL P, FERRARA E. Graph embedding techniques, applications, and performance: a survey[J]. Knowledge-Based Systems, 2018, 151: 78-94.
- [34] CAI Hongyun, ZHENG V W, CHANG K C C. A comprehensive survey of graph embedding: problems, techniques, and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616-1637.
- [35] KONG Dejing, YANG Jianzhong, LI Lingfeng. Early identification of technological convergence in numerical control machine tool: a deep learning approach[J]. Scientometrics, 2020, 125(3): 1983-2009.
- [36] SIDDHARTH L, LI Guangtong, LUO Jianxi. Enhancing patent retrieval using text and knowledge graph embeddings: a technical note[J]. Journal of Engineering Design, 2022, 33(8-9): 670-683
- [37] CHOI S, LEE H, PARK E, et al. Deep learning for patent landscaping using transformer and graph embedding[J]. Technological Forecasting and Social Change, 2022, 175: 121413.
- [38] WANG Liang, LI Munan. An exploration method for technology forecasting that combines link prediction with graph embedding: a case study on blockchain[J]. Technological Forecasting and Social Change, 2024, 208: 123736.
- [39] 窦永香,开庆,王佳敏.一种基于图表示学习的潜在颠覆性技术识别方法[J].情报学报,2023,42(6):637-648. (DOU Yongxiang, KAI Qing, WANG Jiamin. Potential disruptive technology identification method based on graph representation learning[J]. Journal of the China Society for Scientific and Technical Information, 2023,42(6):637-648.)
- [40] 胡威,李姝影,张鑫,等.基于图与语义表示学习的专利引文网络链路预测研究[J].数据分析与知识发现,2024,8(10):28-43. (HU Wei, LI Shuying, ZHANG Xin, et al. Link Prediction in patent citation networks based on graph and semantic representation learning[J]. Data Analysis and Knowledge Discovery, 2024,8(10):28-43.)
- [41] 杨冠灿,行佳鑫,鲁国轩,等.基于图神经网络的细粒度技术会聚预测方法研究[J].信息资源管理学报,2023,13(2):95-107. (YANG Guancan, XING Jiaxin, LU Guoxuan, et al. A fine-grained technology convergence prediction method based on graph neural networks[J]. Journal of Information Resources Management, 2023,13(2):95-107.)
- [42] SONG C H, ELVERS D, LEKER J. Anticipation of converging technology areas—a refined approach for the identification of attractive fields of innovation[J]. Technological Forecasting and Social Change, 2017, 116: 98-115.

- [43] TANG Ying, LOU Xuming, CHEN Zifeng, et al. A study on dynamic patterns of technology convergence with IPC co-occurrence-based analysis: the case of 3D printing[J]. Sustainability, 2020, 12(7): 2655.
- [44] 慎金花, 闫倩倩, 孙乔宣, 等. 基于专利数据挖掘的技术融合识别与技术机会预测研究——以电动汽车产业为例[J]. 图书馆杂志, 2019, 38(10): 95-106.(SHEN Jinhua, YAN Qianqian, SUN Qiaoxun, et al. Research on technology fusion recognition and technology opportunity prediction based on patent data mining[J]. Library Journal, 2019, 38(10): 95-106.)
- [45] 吕璐成, 赵亚娟. 基于机器学习的“从 0 到 1”型技术融合预测方法研究[J/OL]. 情报理论与实践, 1-13[2025-05-02]. <http://kns.cnki.net/kcms/detail/11.1762.G3.20250429.1444.005.html>. (LÜ Lucheng, ZHAO Yajuan. Research on the prediction method of the "0 to 1" technology convergence based on machine learning[J/OL]. Information studies: Theory & Application, 1-13[2025-05-02]. <http://kns.cnki.net/kcms/detail/11.1762.G3.20250429.1444.005.html>.)
- [46] SUZUKI J, KODAMA F. Technological diversity of persistent innovators in Japan: two case studies of large Japanese firms[J]. Research Policy, 2004, 33(3): 531-549.
- [47] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.
- [48] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
- [49] HAMILTON W, YING Zhitao, LESKOVEC J. Inductive representation learning on large graphs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. California: Curran Associates Inc, 2017: 1025-1035.
- [50] HOU Zhenyu, LIU Xiao, CEN Yukuo, et al. Graphmae: self-supervised masked graph autoencoders[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington DC, USA, 2022: 594-604.
- [51] FAJRIAL A K, HE Qingqing, WIRUSANTI N I, et al. A review of emerging physical transfection methods for CRISPR/Cas9-mediated gene editing[J]. Theranostics, 2020, 10(12): 5532.
- [52] BRANDT L J B, BARNKOB M B, MICHAELS Y S, et al. Emerging approaches for regulation and control of CAR T cells: a mini review[J]. Frontiers in Immunology, 2020, 11: 326.
- [53] KIPF T N, WELLING M. Variational graph auto-encoders[J]. arXiv preprint, 2016:57.

作者简介: 郭思成 (通信作者, Email: guosccc@whu.edu.cn), 男, 1992 年生, 博士后。研究方向: 科技情报分析, 主题挖掘。**梁镇涛**, 1996 年生, 博士生。研究方向: 科技情报组织与分析。**陈传夫**, 1962 年生, 博士, 教授。研究方向: 信息资源知识产权。

作者贡献声明: 郭思成, 研究设计, 实验, 论文撰写。**梁镇涛**, 参与实验数据处理。**陈传夫**, 提供论文修改建议。

录用日期: 2025-07-02