期刊嵌入揭示科学主题分类方案中隐藏的跨学科模式

吕卓奇[1] 和柯庆[1],*

[1]香港城市大学计算学院数据科学系，中国香港（日期:
2025年12月30日）

arXiv:2512.22524v1 [cs.DL] 27 Dec 2025

摘要

主题分类方案是组织、评估和检索科学知识的基础。Scopus 等由专家编制的系统提供了广泛使用的分类体系，但往往存在粒度粗糙、主观性强以及难以适应新兴跨学科领域等问题。基于引文网络的数据驱动替代方案虽前景可期，却缺乏针对科学文献语义内容的严格外部验证。在此，我们提出一种新颖的定量框架，利用分类任务评估期刊分类方案的有效性。通过使用超过 2300 万篇论文摘要，我们证明，基于 Periodical2Vec (P2V)（一种从论文层级引文中学习得到的期刊嵌入）进行 k-均值聚类所生成的标签，其分类性能显著优于 Scopus 及其他数据驱动基线方法（例如引文、共引和 Node2Vec 变体）。通过比较不同分类方案下的期刊划分，我们在科学地图上发现了两种结构模式：（1）学科边界的重组——拆分过于宽泛的类别（例如，将“医学”细分为“肿瘤学”“心脏病学”等专科），同时合并人为割裂的类别（例如“化学”与“化学工程”）；（2）识别连贯的跨学科簇——例如“生物医学工程”“医学伦理学”和“信息管理学”，它们分散在多个类别中，却在引文空间中彼此关联。这些发现表明，基于引文的期刊嵌入不仅在预测效度方面优于传统分类体系，还能提供一幅动态、精细的科学地图，更好地反映当代研究固有的专业化与跨学科特征。

关键词：科学 o　　科学、期刊主题分类、分类体系、科学结构、科学地图

* q.ke@cityu.edu.hk

一、引言

科学知识通过学科分类系统进行组织，这些系统是学术生态系统的基础设施。在整个研究生命周期中，这些系统发挥着至关重要的作用：作者提交稿件时依靠这些系统标注相关研究领域，读者则借助它们在浩如烟海的出版物中查找所需内容；此外，它们还为科学计量分析提供了必要框架，以描绘科学的整体图景 [1, 2]。这些分类的重要性远不止于简单归类——它们从根本上塑造了科学研究成果的传播与评价方式，并最终影响科学演进的衡量与引导方式。除了学术功能之外，它们还深刻影响信息获取与管理、个人／机构的学术绩效评价 [3],研究资助分配，以及科学政策制定 [4]。

传统分类方案（如 Web of Science (WoS) 和 Scopus 提供的方案）大多由内部专家手动设计，会为论文、期刊、会议论文集和图书指定预设的类别标签。尽管这些分类方案应用广泛且十分重要，但往往存在诸多局限，例如方法论缺乏透明度 [5], 主观性 [6], 粒度不足 [3, 7],覆盖范围有限，以及无法快速适应新兴跨学科领域 [8],因此可能无法充分体现科学研究细致且不断演变的特性，进而导致潜在的错误分类或疏漏 [9], 以及不公平的出版物评价 [10]。

人们曾多次尝试改进或取代 WoS [11–13] 和 Sco- pus [14–16]分类方案。这些数据驱动方法主要挖掘论文之间细微的引文关系，与传统方法相比，具有更好的可扩展性和灵活性。然而，尽管已有这些尝试，仍存在两个关键缺口。首先，缺乏利用引文关系本身之外的外部数据源进行系统验证，以交叉核验所提出分类的合理性 [5], 。文本、关键词或资助类别中蕴含的丰富身份信息，尚未得到充分探索和利用，因而无法用于测试。其次，现有评价主要依赖聚类一致性和集中度等自指性指标 [16], ，或采用作者定义的阈值进行人工检查 [5], ，却未建立普适的定量指标，无法对不同分类方案进行客观比较。这些局限凸显了建立严格的跨模态验证基准和评价指标的必要性，以便评估

有效性　　不同分类方案在大量学术期刊中的表现　　.

为应对这些挑战，本研究旨在将先进的嵌入方法应用于大规模引文网络，从而为期刊生成更具意义的数据驱动主题分类。延续我们以往的研究 [17, 18], ，我们对论文间引文网络进行抽样，生成 107,107,571 条引文路径，并在这些路径上应用 Word2Vec [19] ，以期生成能够反映期刊语义关系的向量表示。随后，可通过对嵌入进行聚类来获得分类标签。接着，我们从基于语义的分类任务和多种主题一致性指标的角度，将这些嵌入与 Scopus 及其他数据驱动分类方案进行比较，评估其有效性。具体而言，我们首先对 2,300 多万篇论文摘要进行文档分类，以评估所提出的科学期刊学科划分的有效性。其次，我们开展主题建模分析，衡量研究论文中的自然主题结构与各种分类方案之间的一致程度。这种受控的方法使我们能够客观评估不同方法在多大程度上符合科学生态系统的实际组织方式和结构，而这些组织方式和结构体现在研究论文的内容之中。此外，我们发现，期刊嵌入不仅能够提高主题分类的准确性，还能为科学出版格局的演变提供洞见。通过识别隐藏的跨学科联系和新兴研究领域，它有助于更细致地理解科学进步及其对学术评价、科研资助和政策制定的影响。

II. 期刊分类系统概述

A. 传统的专家制定分类系统

目前应用最广泛的期刊学科分类系统，是由 Web of Science (WoS) 和 Scopus 等大型文献数据库开发并维护的系统。WoS 目前将期刊划分为 254 个学科类别，而 Scopus 使用 All Science Journal Classification (ASJC) 系统，该系统包含 27 个主要学科领域。这些分类方案主要由内部专家编制，他们依据编辑判断、期刊范围声明，以及在某些情况下的引文模式，为期刊分配学科类别

期刊 [1, 6]。此类专家编制的分类框架在文献计量学和科研评价中发挥着核心作用，支持从计算领域标准化引文指标 [7] 到开展科研评估等各种应用。

然而，基于专家的分类系统存在若干广为人知的局限。首先，分类分配过程往往不透明且带有主观性，导致不同数据库之间以及不同时期的分类结果不一致 [6, 7]。其次，类别的粒度参差不齐：某些生物科学领域被划分为许多专业类别，而医学等学科则划分不足，被归入宽泛的类别 [7, 13]。第三，WoS 和 Scopus 的分类系统几乎只关注期刊，对会议论文集的覆盖有限甚至完全缺失，而会议论文集是计算机科学等某些快速发展的领域的主要出版渠道。这种有限的覆盖范围进一步限制了这些系统反映科学活动全貌及跨学科联系的能力。最后，专家编纂耗费大量人力，阻碍了系统的扩展，也限制了其及时适应科学格局变化的能力。

B. 算法与数据驱动方法

除了基于专家的分类之外，研究人员还开发了多种算法驱动和数据驱动的替代方案。主流研究利用引文网络的结构来描绘科学的组织方式，并生成新的分类方案。例如， [13] 采用期刊交叉引文分析并结合层次聚类，以验证和完善学科分类。研究结果表明，与专家编纂相比，数据驱动方法能够更客观地识别分类错误的期刊，并揭示潜在的学科结构。同样， [20] 以既有期刊类别为参照，系统比较了五种基于引文和共引相似性度量的科学地图，凸显了基于网络的方法在揭示科学结构方面的潜力。后续研究采用了更复杂的算法，通过引文流优化期刊分类 [14–16]。此外，研究还开发了结合引文数据、书目耦合和文本分析的混合方法，以提高分类的稳健性和特异性 [11, 12, 16]。近来，图嵌入技术的进步使表征学习得以应用于科学网络 [17, 18]，展现出构建精细、可扩展且自适应分类系统的潜力，这些系统

能够更好地捕捉不断演变的科学传播中的相互连通性。这些发展反映出一种日益增长的共识：数据驱动的算法框架不仅能够补充基于专家的系统，在某些情况下甚至能够超越它们，更好地描绘科学结构的演变。

尽管方法上有这些创新，但针对海量研究论文，对各种期刊分类方案进行外部验证和基准测试仍然不足。现有评估大多依赖对类别分配的人工检查 [6, 7,16], ，或依赖模块度 [11], 、轮廓系数 [11, 12] 和余弦相似度 [4]等内部聚类指标。使用非引文数据源进行系统的跨模态验证仍然鲜见，例如利用文本内容、作者提供的关键词或资助机构类别 [5,8]。由于缺乏广泛认可的外部基准，不同分类方案之间难以进行定量、客观的比较，我们评估其语义连贯性、实用价值以及适应新兴领域或跨学科领域的能力也因此受到限制。

III. 结果

A. 文档分类任务

为系统、定量地评估不同期刊分类方案的有效性 $\mathcal{C}$，我们使用微软学术图谱（MAG）数据集中的23，322，430篇论文摘要开展大规模文档分类任务。指导我们设计的基本原则是，良好的分类方案应 $\mathcal{C}_i$ 使训练好的分类器能够根据论文的语义内容，有效预测其所属学科领域。如果 $\mathcal{C}_i$ 分类方案确实反映了有意义的学科边界，那么被分配到同一类别的论文就应具有语义相似性，而这种相似性应可通过计算方法检测出来。更具体地说，在训练语料库相同的情况下，我们预期，对期刊进行更有意义的划分，能够使分类器更擅长检索同一类别中的论文，即获得更高的召回率。通过保持输入文本特征和模型结构完全相同，仅改变由不同分类方案得出的标签，我们得以分离出分类方案本身的影响 $\mathcal{C}_i$。

在此，我们比较六种不同的分类方案 $\mathcal{C}$：(1) Scopus 传统学科分类；将 k-均值应用于 (2) 引文矩阵；(3) 引文+Node2Vec；(4)

共引矩阵；(5) 共引+Node2Vec；以及 (6) Periodical2Vec。对于每种 $\mathcal{C}_i$，我们将期刊聚类标签（或 Scopus 的 ASJC 学科领域，见 § V A）用作目标类别，并采用相同的哈希向量化与互补朴素贝叶斯（CNB）分类器组合进行模型训练和评估。分类结果用于评估 $\mathcal{C}_i$ 生成有意义的期刊分组的有效性。得分越高，表示通过某种方案获得的划分与 MAG 数据集论文摘要中蕴含的语义信息越一致。

我们设定了两种预测场景：多分类和一对其余分类。在前一种场景中，为评估每个 $\mathcal{C}_i$，我们使用单个分类器预测论文的类别标签。表 I 汇总了六种方法在精确率、召回率和 F1 分数方面的宏平均和加权平均表现。这里的宏平均值指各类别得分的非加权平均值。虽然宏平均精确率相近，但基于 Periodical2Vec (P2V) 的划分取得了显著更高的召回率，符合我们的预期。Node2Vec 旨在从原始稀疏引文向量中学习稠密向量，但未能带来提升。请注意，$\mathcal{C}^{Scopus}$ 中的期刊在各类别间的分布高度不均衡（表 S1）：由于 Scopus 的粒度较粗（见 § III C），很大一部分期刊属于医学领域。这使 $\mathcal{C}^{Scopus}$ 在加权指标上天然占优，因为规模最大的类别的表现会主导加权平均值。尽管 Scopus 具有这一优势，我们基于引文嵌入的方案（$\mathcal{C}^{P2V}$）在宏平均和加权平均指标上始终优于所有其他方法。具体而言，我们的方法取得了最高的召回率和 F1 分数，表明其在所有类别（包括样本较少的类别）中都具有稳健的表现。这表明，我们的方法显著提升了整体分类的一致性和泛化能力。

对于一对其余分类场景，我们迭代训练 K 个不同的二元分类器，其中 K 是类别总数。对于每个分类器，我们选择一个类别作为正类，其余所有类别均为负类。图 1 展示了各方法的宏平均精确率-召回率（PR）曲线和受试者工作特征（ROC）曲线。$\mathcal{C}^{P2V}$ 的 PR 表现最佳，曲线下面积（AUC）也最高，优于所有其他方法。我们强调，在像我们这样的高度不平衡二分类设置中，基于 PR 曲线的指标比 ROC AUC 更具信息量，也更可靠，适合用于

表 I. 使用不同标签训练的分类器之间的测试集多分类性能比较。结果取10折交叉验证的平均值；数值表示为

均值 ± 标准差。

	宏平均			加权平均			Rank	
	精确率	召回率	F1	精确率	召回率	F1	平均精确率	Loss
Periodical2Vec	60.62 ± 0.24	**50.62** ± 0.04	**51.69** ± 0.04	61.21 ± 0.10	60.38 ± 0.03	**58.81** ± 0.03	**74.47** ± 0.02	**0.0550** ± 0.0001
引文	60.70 ± 0.59	41.67 ± 0.03	44.26 ± 0.04	58.54 ± 0.03	57.87 ± 0.03	56.91 ± 0.03	72.17 ± 0.02	0.0853 ± 0.0001
引文+Node2Vec	46.31 ± 0.27	33.22 ± 0.03	31.17 ± 0.04	47.13 ± 0.08	47.78 ± 0.03	41.43 ± 0.03	64.59 ± .02	0.0828 ± 0.0001
共被引	**62.93** ± 0.17	47.50 ± 0.04	49.14 ± 0.05	61.14 ± 0.03	60.40 ± 0.02	58.76 ± 0.02	74.45 ± 0.02	0.0598 ± 0.0001
共被引+Node2Vec	38.57 ± 2.09	12.51 ± 0.03	12.75 ± 0.05	59.52 ± 0.32	**60.86** ± 0.03	57.29 ± .03	74.42 ± 0.02	0.1294 ± 0.0002
Scopus	60.93 ± 0.38	34.41 ± 0.03	36.28 ± 0.03	**61.35** ± .07	60.20 ± 0.04	55.90 ± 0.04	74.07 ± .03	0.0574 ± 0.0001

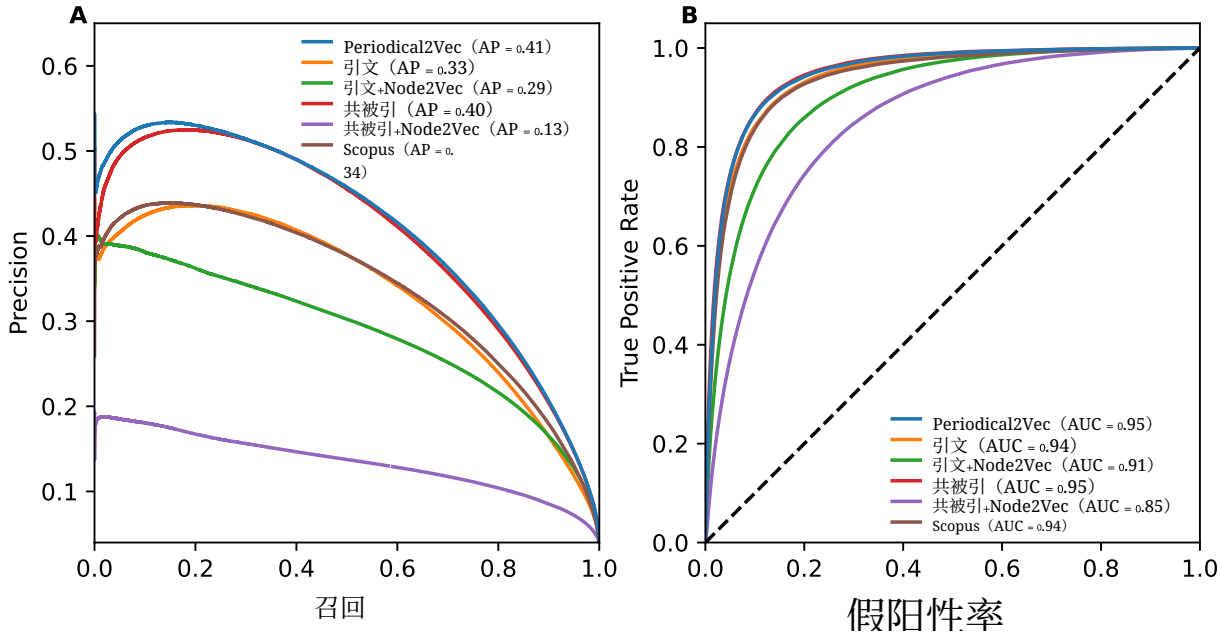


图 1. 使用不同分类标签训练的分类器之间的一对其余分类性能比较。曲线绘制为所有单个类别结果的宏平均。（**A**）精确率-召回率曲线。括号中的数字表示平均精确率。（**B**）受试者工作特征曲线。括号中的数字表示曲线下面积。

衡量分类器在现实世界中的有效性 [21]。因此，我们方法更高的宏平均 AP 不仅表明其整体区分能力强，也尤其体现出其在从代表性不足的学科领域中检索相关条目方面表现更佳。

综合来看，这些结果证实了基于期刊嵌入得出的标签具有更强的判别能力和类别可分性，无论是在一般的多类别设置中，还是在区分单个类别与其他类别时，均是如此。

B. 主题标签一致性

为进一步评估不同分类方案的语义连贯性和有效性，我们对从论文摘要中检测出的主题与基于分类得出的标签进行一致性分析。此分析旨在检验这些标签是否与学术文献中的自然主题分布相符。具体而言，我们将潜在狄利克雷分配（LDA） [22] 应用于整个摘要语料库，从而推断出一组能够体现科学文献潜在主题结构的潜在主题。每篇文档都被分配到概率最高的主题，随后将这些主题分配与各分类方法生成的期刊标签进行比较。我们使用三种广泛采用的聚类一致性指标，对LDA推断的主题与分类标签之间的一致性进行定量评估，分别是标准化互信息（NMI）、调整兰德指数（ARI）和Fowlkes-Mallows指数（FMI）。

表 II 展示了结果， $\mathcal{C}^{P2V}$ 在全部三项指标中得分最高（NMI $= 0.3195$、ARI $= 0.1693$、FMI $= 0.2169$），其次是 $\mathcal{C}^{Co-citation}$。与 Scopus 相比， $\mathcal{C}^{N2V}$和 $\mathcal{C}^{Citation}$ 与 LDA 推断出的主题结构的一致性较低。这些结果表明，基于引文的期刊嵌入所得到的簇与科学文献的潜在主题结构最为吻合，优于两种直接从原始引文中得出的替代性数据驱动方法，也优于既有的专家整理的 Scopus 分类。

表 II. LDA 指定的论文主题与不同分类方案指定的标签之间的标准化互信息、调整兰德指数和 Fowlkes–Mallows 指数。

	Periodical2Vec	引文	引文+Node2Vec	共被引	共被引+Node2Vec	Scopus
NMI	**0.3195**	0.2662	0.2748	0.3168	0.1672	0.3010
ARI	**0.1693**	0.1127	0.1306	0.1633	0.0324	0.1371
FMI	0.2169	0.1804	0.1816	**0.2203**	0.1631	0.2106

总体而言，这些发现提供了有力的定量证据，表明基于引文嵌入的分类方法不仅提升了文档分类等下游任务的表现，还能生成在语义上更符合科学研究内在主题结构的标签分配。

C. 科学的结构

前述分析通过定量评估，系统验证了我们基于引文的分类系统的有效性。在此基础上，我们现在将重点转向更具探索性的研究，考察期刊在不同分类方案中的归类差异，并揭示反映科学研究结构演变的隐藏模式。

图 2 使用桑基流可视化，考察期刊在 Scopus 分类与我们的数据驱动分类中如何被划分到不同类别。首先，图中显示，基于 P2V 得到的数据驱动聚类比分布于 Scopus 分类方案中的期刊更为均衡：在 $\mathcal{C}^{Scopus}$ 中，23% 和 16% 的期刊分别被归入“医学”和“社会科学”，而在 $\mathcal{C}^{P2V}$ 中，最大的簇仅包含 6% 的期刊。两种分类方案在规模分布上的差异，既体现了彼此的契合，也反映出分歧。具体而言，数据驱动聚类揭示了 Scopus 宽泛类别内部更精细的子结构。例如，在 Scopus 中被归为“医学”的期刊分布于多个簇中，如侧重心脏病学的簇 7、聚焦肿瘤学的簇 4，以及其他若干簇；这反映出 Scopus 将丰富的内部结构压缩为单一类别。类似地，社会科学期刊被划分为多个专业簇，如簇 19“语言学”（例如，语言与文学）和 #25“教育学”（例如，教学与教师教育），与 Scopus 单一化的归类方式有所不同。这些模式表明，传统分类往往将引文交流方式迥异的领域混为一谈。

与此同时，一些 $\mathcal{C}^{Scopus}$ 类别（如数学）在 $\mathcal{C}^{P2V}$ 中有一一对应关系，而另一些类别则被合并到同一簇中。例如，大多数化学工程和化学期刊被归入簇 11；同样，大多数标注为“经济学、计量经济学与金融学”和“商业、管理与会计”的期刊被归入簇 24。这两种情况都反映了 Scopus 分类可能存在的过度划分缺陷：这些领域通过引文模式紧密交织于研究实践之中，因此自然归入同一簇，却被 Scopus 人为地分开。除合并之外，一些较小的类别，包括“牙医学”“兽医学”“心理学”和“地球与行星科学”，主要对应簇 3、16、9 和 12，并因纳入来自其他

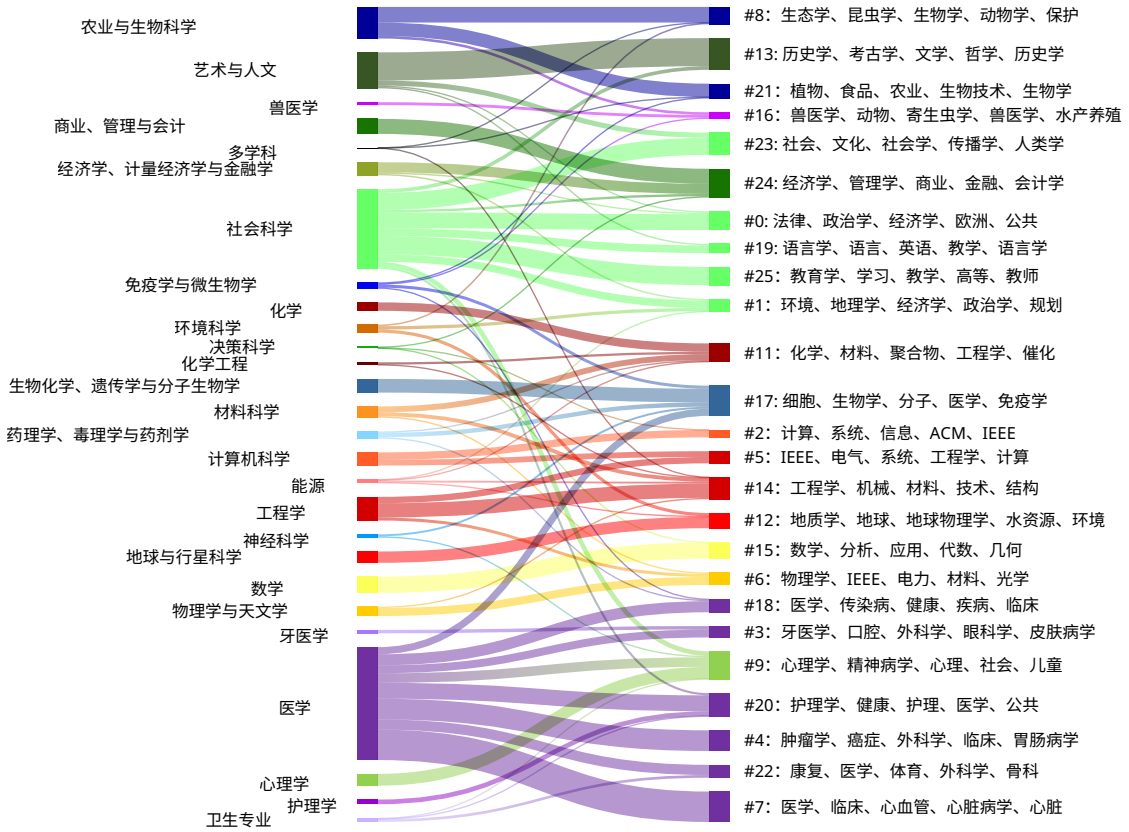


图 2. 桑基图展示了期刊在 Scopus 类别（左侧节点）与数据驱动的聚类（右侧节点）之间的不同归类方式。右侧每个节点均列出期刊名称中出现频率最高的 5 个词，以显示各簇的主题。流线的粗细表示在两种分类方案之间重新分配的期刊比例。为便于观察，我们过滤掉期刊数少于 $\min(10\% \times N_{source}, 50)$ 的流线，其中 N_{source} 为源节点（左侧）的期刊数。可通过 https://lyuzhuoqi.github.io/ periodical-clustering/sankey/snakey_kmeans_filtered.html 查看该桑基图的交互式在线版本。

Scopus 类别的流量而扩展，这凸显了基于引文的嵌入如何捕捉跨学科联系和子领域粒度。通过整合分散的 Scopus 类别并拆分规模过大的类别，可以实现自然的规模均衡，从而反映实际的研究活动规模。

为了进一步描绘科学的细微结构，图 3A–B 展示了基于期刊 P2V 嵌入的二维投影，并分别根据 $\mathcal{C}^{Scopus}$ 和 $\mathcal{C}^{\mathrm{P2V}}$ 中分配的标签为其着色，使我们能够观察期刊的分布与相互作用，

此外，还可以细致考察各簇的边界。图 3A 显示，许多位于簇边缘附近的期刊获得了相互冲突的 Scopus 标签（例如，位于“兽医学”簇边缘的期刊被标记为“农业与生物科学”）。相比之下， k-均值簇的边界更为清晰（图 3B）。

为量化这种结构差异，我们计算逐元素聚类相似度 [24]——这是一种样本层面的度量，用于比较不同分类系统间标签的一致性；随后采用反距离加权 [25] (IDW)进行空间插值，生成图 3C。该可视化突显了人工分类与实际引文模式之间的一致与分歧。例如，我们在图 3C 中用斜体标注一致性较高的区域，包括“数学”“统计学”“艺术、拉丁美洲研究”等。对于分歧较大的区域（相似度较低），我们观察到两种不同的模式。第一种对应于宽泛学科类别中的复杂子结构，包括“耳鼻喉科学”“眼科学”和“皮肤病学”。这些领域呈现出尤其碎片化的微观结构模式（在图 3C 中以常规字体标注），与我们在图 2 中观察到的情况一致，表明 Scopus 将专业领域人为地混为一谈，尽管这些领域的引文模式与医学主流领域有所不同。第二种分歧较大的情况对应于主题凝聚、且具有高度跨学科性的期刊簇（例如“信息管理学”“生物医学工程”和“医学伦理学”，在图 3C 中以粗体标注）；在 $\mathcal{C}^{Scopus}$和 $\mathcal{C}^{\mathrm{P2V}}$中的一个或两者作用下，这些期刊簇会形成不稳定的异质簇，揭示了其跨学科特性。表 S2–S3 列出了位于这些流动边界上的期刊详细名单。

总体而言，某些领域中持续存在的一致性证实，引文网络能够捕捉科学实践的关键特征；与此同时，观察到的差异则凸显了当代科学研究的动态性、跨学科性和精细化特征。

IV. 讨论

本研究的主要贡献是引入利用预测任务对不同期刊分类进行定量评估的方法。通过基于论文摘要的语义交叉验证，我们证明，与其他方法相比，我们的基于嵌入的方法不仅与论文层面的文本信息更加吻合，而且在

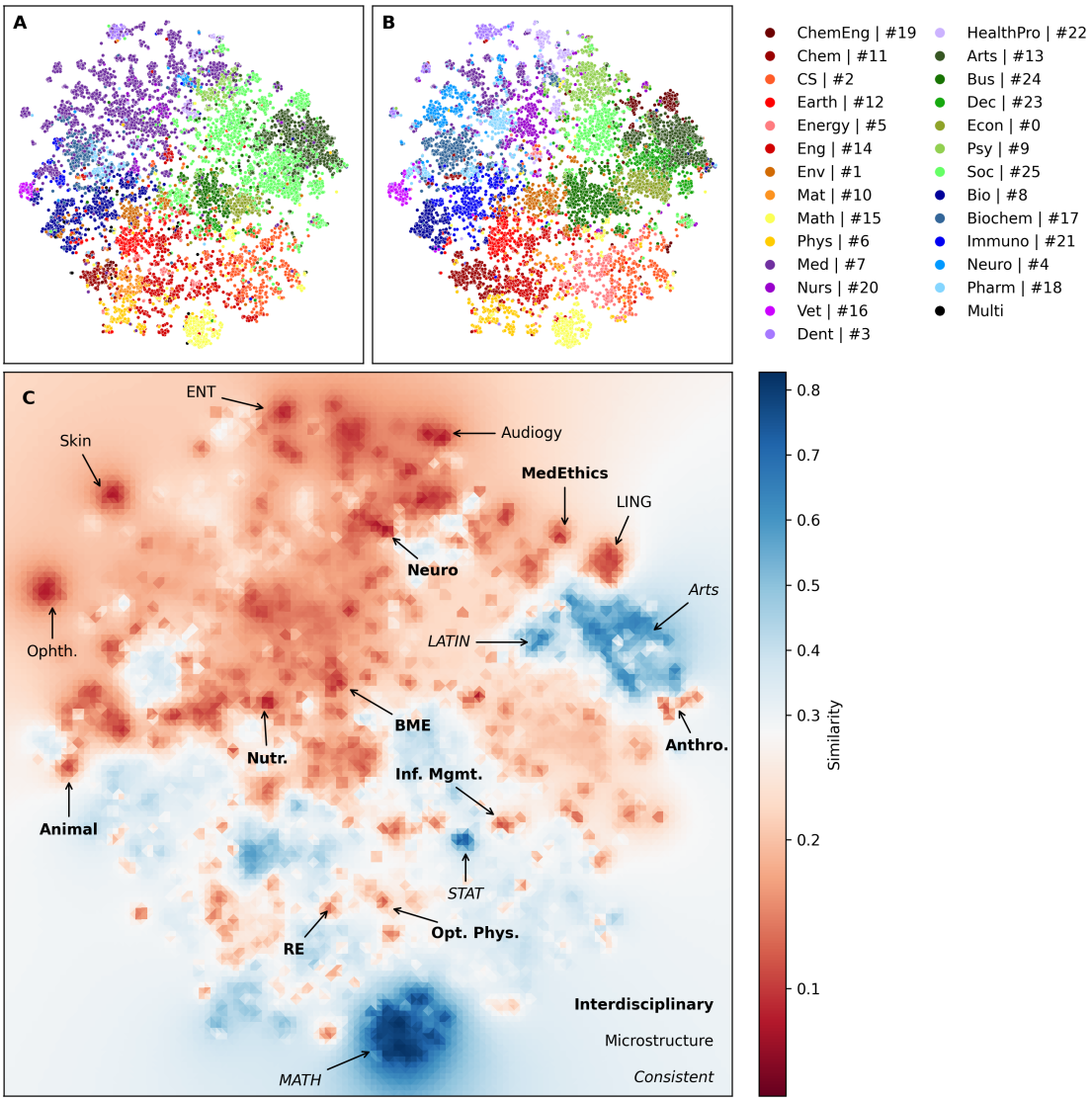


图 3。比较 $\mathcal{C}^{Scopus}$ 和 $\mathcal{C}^{P2V}$。（A–B）使用 t-SNE [23] 对期刊 P2V 嵌入进行二维投影。（A）和（B）共用图例，以便直观追踪簇。（**A**）按 $\mathcal{C}^{Scopus}$ 划分。颜色表示标签。（**B**）按 $\mathcal{C}^{P2V}$ 划分。颜色表示簇编号。（**C**）Scopus 与聚类生成标签之间逐元素聚类相似度的插值热图，并标注了具有显著低/高相似度的代表性主题内聚簇（详细簇成员见表 S2–S5）。箭头的线型表示两个分类方案之间产生异常相似度的原因。插值基于反距离加权（IDW），使用幂参数 $=2$。

下游任务中表现也更出色；与其他基于引文的方法以及 Scopus 等由专家整理的分类体系相比，皆是如此。这一能力使我们的方法成为描绘

并细致描绘科学不断演变的格局。通过揭示隐藏的跨学科联系、整合分散的领域，并展现宽泛类别内部的子结构，这些数据驱动的嵌入为科学知识提供了动态表征，使其更贴合实际研究实践。

我们承认本研究存在两个主要局限。首先，用于比较的期刊分类系统并不全面。例如，本文未涵盖 Web of Science学科类别等其他广泛使用的方案，但未来可以对此进行考察。其次，单向量表示可能无法充分捕捉多学科期刊的语境多样性，例如$PNAS$ 这类期刊，其范围横跨多个不同的研究群体。第三，尽管我们基于引文的方法优于其他方法，但我们也认识到，将引文模式与文本语义相结合仍有尚未发掘的潜力。多模态学习的最新进展[26]表明，将引文嵌入与基于文本的表示 [27, 28] 融合，或许能够更好地捕捉科学实体拓扑。尽管存在这些限制，我们的框架仍为若干有前景的研究方向奠定了基础。未来研究可以探索：（1）将引文嵌入与最先进的自然语言处理表示（例如基于 Transformer 的嵌入）相结合的混合模型，超越基础元数据，实现语境化的语言理解；（2）此类多模态系统如何提升新兴领域检测能力，尤其是在快速发展的领域中，文本所承载的创新可能早于引文激增出现。

V. 数据与方法

A. 数据

在本研究中，我们使用了 12 月 2021 [29], 获取的一个版本的微软学术图谱（MAG），其中包含 96,072, 424 篇论文，这些论文发表于 49, 006 种期刊和 4,551 场会议，时间跨度为 1800 年至 2021 年。鉴于科学领域格局随时间快速演变 [18], ，我们将分析范围限定为 2010 年及之后发表的期刊和会议论文，以聚焦最新快照。为避免偏向历史学，我们仅保留相应十年内发表的论文，以及这些论文之间的引文关系（即，单条引文记录中的来源出版物和目标出版物都必须发表于同一十年）。最终，我们的数据集包含 41,820,928 篇论文和

来自 46,074 种期刊的 359,442,336 条引文。

我们使用 Scopus 学科领域类别为期刊添加标签，作为我们自标注标签的手动标注对应方案，用于基准测试。Scopus 是一个数据库，涵盖生命科学、社会科学、物理科学和健康科学等顶级学科领域中的 40,878 种同行评审期刊 [30]。该数据库采用 ASJC（All Science Journal Classification）方案，将期刊划分为 27 个主要学科领域（包括“多学科”），且期刊可以同时属于多个领域。我们将某一期刊的前 50 个相似同行中大多数期刊共有的学科领域（取决于期刊 P2V 嵌入之间的余弦相似度）分配给该期刊，作为其单标签。

最后，我们根据期刊名称，在我们的嵌入与 Scopus 之间匹配了 20,038 种期刊。值得注意的是，Scopus 不提供会议论文集的标签，而这些会议也被纳入嵌入的构建过程中。

B. 分类方案

我们采用五种算法方法获取期刊分类，这些方法都涉及对期刊向量应用 k-均值。第一种方法是 P2V，该方法此前已为 37,594 种期刊准备完毕 [18]。

接下来，对于另外四种备选向量表示（引文、引文+Node2Vec、共被引和共被引+Node2Vec），我们构建一个 43982×43982 期刊—期刊引文矩阵，其中每个条目 E_{ij} 记录从期刊 i 到期刊 j的引文数量。具体而言，对于每篇引用了发表于期刊 j的论文的论文，都会统计引文从施引期刊 i 指向被引期刊 j 的次数。该矩阵刻画了期刊之间的引文流，其中数值 E_{ij} 对应于期刊 i 中的期刊对发表于期刊 j的论文所作引文的总数。我们按行对矩阵进行归一化；因此，第 i行中的每个条目 i表示该期刊引用其他期刊的可能性。

第二种向量表示（引文）是直接将该矩阵的各行用作每种期刊的高维向量，矩阵中编码了期刊之间原始的引文信息。

第三种方法（引文+Node2Vec）基于未归一化的引文矩阵构建期刊–期刊引文网络。在该网络中，节点代表期刊，边

表示期刊之间的引文流。随后，我们在该引文网络上应用 Node2Vec [31], 这一常用的图嵌入技术，以获得期刊的向量表示。我们采用 SNAP 实现 [32] 并遵循默认超参数：嵌入维度为 $d = 128$、游走长度为 $l = 80$、每个源节点的游走次数为 $r = 10$、返回超参数为 $p = 1$，以及出入超参数为 $q = 1$。该方法与我们的方法的关键区别在于：我们在论文级引文网络（其中每篇论文都是一个节点）上进行随机游走，而该方法则在期刊级引文网络上运行，因此我们的方法可能保留更细粒度的信息。

第四种（共被引）方法构建一个 43728×43728 期刊-期刊共被引矩阵，其中每个条目 E_{ij} 记录期刊 i和期刊 j 被同一篇论文共同引用的次数。具体而言，对于每篇同时引用期刊 i和期刊 j的论文，这两种期刊之间的共被引次数加一。因此，该矩阵根据期刊在文献中被共同引用的频率，反映期刊之间的相关程度。与引文矩阵类似，我们按行对共被引矩阵进行归一化，使每个条目反映某一期刊与其他期刊共同被引的相对倾向。随后，直接将该矩阵归一化后的各行用作每种期刊的高维向量表示。

第四种（共被引+Node2Vec）方法根据未归一化的共被引矩阵构建期刊-期刊共被引网络，其中节点代表期刊；如果两种期刊曾被同一篇论文共同引用，则在它们之间建立一条边，边权重对应其共被引次数。然后，我们在该共被引网络上应用Node2Vec，为每种期刊学习低维向量嵌入。

获得期刊嵌入后，我们对这些期刊嵌入应用 k-均值，以生成期刊标签。簇的数量设为 $K = 26$，与 Scopus 分类中的 26 个学科领域相对应（不包括多学科类别）。这一选择便于在评估聚类性能时直接与 Scopus 标签进行比较。

C. 文档分类任务

分类任务旨在评估不同方案所生成标签的一致性和可解释性。为此，我们使用来自 MAG 的 23,322, 430 篇论文摘要执行了文档分类任务。表 S1 展示了本研究所用论文的领域分布。所采用的分类器由用于特征提取的 Hashing Vectoriza- tion [33] 和用于分类的互补朴素贝叶斯 (CNB) [34]组成。HashingVectorizer 通过哈希函数将文本特征映射为固定维度的向量，实现特征哈希降维。对于一个包含词元 ($t_1, t_2, ..., t_n$) 的文档 d，哈希变换定义如下：

$$\mathbf{x}_d = \left[\sum_{j:h(t_j)=1} f(t_j), \sum_{j:h(t_j)=2} f(t_j), ..., \sum_{j:h(t_j)=m} f(t_j) \right] \tag{1}$$

其中， $h: \mathcal{T} \to \{1, 2, ..., m\}$ 表示将词元映射到特征索引的哈希函数，$f(t_j)$表示词元频率， $m = 2^{20}$ 指定特征空间维度。该方法在保持稀疏表示的同时，无需存储词汇表 [33]。互补朴素贝叶斯是一种适用于类别不平衡数据集的概率分类器。与传统朴素贝叶斯算法不同，后者使用类别 y 本身的数据来估计似然 $P(x_i|y)$；互补朴素贝叶斯则基于各类别的补集来估计参数，也就是说，使用所有不属于类别 y的数据。这一改动有助于减少对多数类的偏倚，并提升性能，尤其是在类别分布偏斜的文本分类任务中。对于文档$\mathbf{x} = (x_1, x_2, \ldots, x_n)$，互补朴素贝叶斯通过最大化以下得分来预测类别 y:

$$\hat{y} = \arg\max_y \left[\log P(y) - \sum_{i=1}^{n} x_i \log P(x_i \mid \bar{y}) \right]$$

其中， $P(y)$是类别 y的先验概率， $P(x_i \mid \bar{y})$是特征 x_i在类别 y的补集条件下的似然。

我们考虑两种评估分类性能的设置：多类别分类和一对其余分类。在多类别设置中，我们使用所有类别的精确率、召回率和 F1 分数的宏平均值与加权平均值作为指标来评估分类性能。此外，我们还采用两种基于排序的指标来评估分类器性能：排序平均精度（RAP）和排序损失。RAP 衡量每个

的表现如何　样本中，真实标签的排名高于错误标签，其定义为　　s

$$\mathrm{RAP} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{Y}_i|}\sum_{j\in\mathcal{Y}_i}\frac{|\{k : f_i(j) \geq f_i(k),\ k \in \mathcal{Y}_i\}|}{r_{ij}} \tag{2}$$

其中，　N 是样本数量，　$\mathcal{Y}_i$ 是样本 i的真实标签集合，　$f_i(j)$是标签 j的预测得分，　r_{ij} 是样本 i中标签 j 的排名。排序损失用于量化排序错误的标签对（错误标签排在真实标签之前）的平均占比，计算公式为

$$\text{Ranking Loss} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{Y}_i||\overline{\mathcal{Y}_i}|}\left|\{(j,k) : f_i(j) \leq f_i(k),\ j \in \mathcal{Y}_i,\ k \notin \mathcal{Y}_i\}\right| \tag{3}$$

其中，$\overline{\mathcal{Y}}_i$ 是样本 i的错误标签集合。这些指标有助于了解聚类派生标签与数据集语义结构的吻合程度。

在一对多设置中，我们训练 K 个分类器，并针对每个分类器，在不同阈值下绘制精确率-召回率（PR）曲线和受试者工作特征（ROC）曲线。为提供一个综合衡量各分类器性能的指标，我们计算PR图的宏平均精度（AP）：

$$\text{Macro-AP} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{AP}_k,$$

其中，AP_k 是类别 k的AP。对于ROC曲线，我们计算曲线下面积（AUC），定义为

$$\text{Macro-AUC} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{AUC}_k,$$

其中，AUC_k 表示 ROC 曲线下面积。两个宏平均指标确保所有类别的权重相等，从而均衡评估分类器在各类别上的表现。

D. 主题-标签一致性分析

为评估由 $\mathcal{C}$ 分配的标签与文档底层主题结构之间的语义连贯性，我们对包含 23,322,430 篇论文摘要的语料库进行潜在狄利克雷分配（LDA） [22] 。首先，我们通过比较主题数量在 10 至 200 之间时的主题一致性得分，确定最佳主题数

步长为 10，并使用包含 1,166,122 篇（5%）摘要的抽样子集。一致性评分（C_v） [35]衡量每个主题中高分词之间的语义相似度，从而对主题可解释性进行定量评估。我们选择 $T = 30$ 个主题，此时一致性评分达到最大值（图 S8）。

对于语料库中的每篇文档 d，LDA 都会给出其属于各个主题的概率分布 T 主题$\theta_d =$ (θ_{d_1}，……，θ_{dK})，其中 θ_{dk} 表示文档 d 属于主题k的比例。我们将每篇文档归入其主导主题 t_d：$t_d = argmax_k\theta_{dk}$。为了量化主题分配与各种分类方案之间的一致性，我们计算了标准化互信息（NMI）、调整兰德指数（ARI）和 Fowlkes-Mallows 指数（FMI）。这些指标用于衡量以两种不同方式划分同一数据集时的一致程度；数值越高，表明发现的主题结构与分类方案越吻合，分类也具有更强的语义意义。NMI 的计算公式为：

$$\mathrm{NMI}(T, C) = \frac{MI(T, C)}{\sqrt{H(T)H(C)}} \tag{4}$$

其中，$H(T)$ 和 $H(C)$ 分别表示主题分配和类别标签的熵：

$$H(T) = -\sum_{t \in T} p(t) \log p(t) \tag{5}$$

NMI 的取值范围为 0 到 1，其中 1 表示主题与标签完全相关，0 表示二者完全独立。ARI 的定义如下：

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}, \tag{6}$$

其中，RI 是兰德指数，用于衡量两种数据聚类之间的相似度；$E[RI]$ 是随机聚类的兰德指数期望值。最后，FMI 的计算方式如下：

$$FMI = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}, \tag{7}$$

其中，TP是真正例的数量，FP是假正例的数量，FN是假负例的数量。

致谢

本研究获得国家自然科学基金委员会（72204206）、香港城市大学（项目编号：9610552、7005968）和香港研究院的资助

面向数据科学。

---

[1] W. Gl¨anzel 和 A. Schubert, 《为科学计量评价目的设计的科学领域与子领域新分类方案》, Scientometrics **56**, 357 (2003).[2] L. Waltman 和 N. J. Van Eck, 《构建科学出版物级分类系统的新方法》, 美国信息科学与技术学会会刊 **63**, 2378 (2012).[3] J. Zhang、S. Gruber 和 R. Frietsch, 《分类粒度对研究机构和组织跨学科绩效评估的影响》, 《数据与信息科学杂志》 **10**, 61 (2025).[4] I. Rafols 和 L. Leydesdorff, 《基于内容的期刊分类与算法分类：科学传播动态及索引者效应的视角》, 美国信息科学与技术学会会刊 **60**, 1823 (2009).[5] Q. Wang 和 L. Waltman, 《Web of Science 和 Scopus 期刊分类系统准确性的大规模分析》, 《信息计量学杂志》**10**, 347 (2016).[6] A. I. Pudovkin 和 E. Garfield, 《寻找语义相关期刊的算法程序》, 美国信息科学与技术学会会刊 **53**, 1113 (2002).[7] L. Leydesdorff 和 L. Bornmann, 《在评价型文献计量学中将“领域”操作化为 WoS 学科类别（WCS）：以“图书馆与信息科学”和“科学与技术研究”为例》, 信息科学与技术协会会刊 **67**, 707 (2016).[8] M. Thelwall 和 S. Pinfield, 《Scopus 期刊领域分类的准确性》, Scientometrics **129**, 1097 (2024).[9] Z. Shen、J. Zhang 和 A. Zeng, 《物理化学期刊代表性不足》, ChemRxiv (2024).[10] Y. Liao、L. Li 和 Z. Shen, 《期刊引文指标目前是否适用于艺术与人文学科期刊？》, arXiv 预印本 arXiv:2503.23400 (2025).[11] F. Janssens、L. Zhang、B. De Moor 和 W. Gl¨anzel, 《用于验证和改进主题分类体系的混合聚类》, 《信息处理与管理》**45**, 683 (2009).

[12] B. Thijs、L. Zhang 和 W. Gl¨anzel, 《用于验证和改进主题分类体系的书目耦合与层次聚类》, Scientometrics **105**, 1453 (2015).[13] L. Zhang、F. Janssens、L. Liang 和 W. Gl¨anzel, 《用于验证和改进文献计量学研究中基于期刊的主题分类的期刊交叉引文分析》, Scientometrics**82**, 687 (2010).[14] A. J. G´omez-N´u˜nez、B. Vargas-Quesada、F. de Moya-Aneg´on 和 W. Gl¨anzel, 《通过参考文献分析改进 SCImago 期刊与国家排名（SJR）的主题分类》, Scien- tometrics**89**, 741 (2011).[15] A. J. G´omez-N´u˜nez、V. Batagelj、B. Vargas-Quesada、F. Moya-Anegon 和 Z. Chinchilla- Rodr´ıguez, 《通过社区检测优化 SCImago 期刊与国家排名分类》, 《信息计量学杂志》**8**, 369 (2014).[16] A. J. G´omez-N´u˜nez、B. Vargas-Quesada 和 F. de Moya-Aneg´on, 《更新 SCImago 期刊与国家排名分类：一种采用 Ward 聚类和引文指标替代组合的新方法》, 信息科学与技术协会会刊**67**, 178 (2016).[17] H. Peng、Q. Ke、C. Budak、D. M. Romero 和 Y.-Y. Ahn, 《学术期刊的神经嵌入揭示复杂的学科组织结构》, Science Advances **7**, eabb9004 (2021).[18] Z. Lyu 和 Q. Ke, 《通过历时期刊嵌入描绘科学结构的变化》, Chaos, Solitons & Fractals **201**, 117295 (2025).[19] T. Mikolov、I. Sutskever、K. Chen、G. Corrado 和 J. Dean, 《词语和短语的分布式表示及其组合性》, 载于神经信息处理系统第*26*届国际会议论文集——第*2*卷, NIPS'13（Curran Associates Inc., Red Hook, NY, USA, 2013）, 第3111–3119页。[20] K. W. Boyack、R. Klavans 和 K. B¨orner, 《描绘科学主干》, Scientometrics**64**, 351 (2005).[21] J. Davis 和 M. Goadrich, 《精确率-召回率曲线与 ROC 曲线之间的关系》, 载于第*23*届国际机器学习会议论文集（2006）, 第233–240页。[22] D. M. Blei、A. Y. Ng 和 M. I. Jordan, 《潜在狄利克雷分配》, Journal of Machine Learning Research**3**, 993 (2003).[23] L. v. d. Maaten 和 G. Hinton, 《使用 t-SNE 可视化数据》, Journal of Machine Learning

$\mathrm{R}_{\mathrm{earch}}$ h **9**, 2579（2008）

.[24] A. J. Gates、I. B. Wood、W. P. Hetrick 和 Y.-Y. Ahn，元素中心聚类比较统一了重叠与层次结构，《Scientific Reports》**9**，8574（2019）。[25] D. Shepard，针对非规则间隔数据的二维插值函数，载于*1968* 年第 *23* 届 *ACM* 全国会议论文集 （1968），第 517–524 页。[26] P. Xu、X. Zhu 和 D. A. Clifton，基于 Transformer 的多模态学习：综述，《IEEE Transactions on Pattern Analysis and Machine Intelligence》**45**，12113（2023）。[27] I. Beltagy、K. Lo 和 A. Cohan，SciBERT：面向科学文本的预训练语言模型，arXiv 预印本 arXiv:1903.10676 10.18653/v1/D19-1371（2019）。[28] J. Dong、Z. Lyu 和 Q. Ke，借助语言模型系列理解科学的演进，arXiv 预印本 arXiv:2409.09636（2024）。[29] A. Sinha、Z. Shen、Y. Song、H. Ma、D. Eide、B.-j. P. Hsu 和 K. Wang，Microsoft Academic 服务（MAS）及其应用概述，载于第 *24* 届万维网国际会议论文集 （2015），第 243–246 页。[30] Elsevier，内容 - Scopus 的运作方式 - Scopus - — Elsevier 解决方案。[31] A. Grover 和 J. Leskovec，Node2Vec：面向网络的可扩展特征学习，载于第 *22* 届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集（2016），第 855–864 页。[32] J. Leskovec 和 R. Sosič，SNAP：通用网络分析与图挖掘库，《ACM Transactions on Intelligent Systems and Technology》**8**，1:1（2016）。[33] K. Weinberger、A. Dasgupta、J. Langford、A. Smola 和 J. Attenberg，大规模多任务学习的特征哈希，载于*ICML* （2009），第 1113–1120 页。[34] J. D. Rennie、L. Shih、J. Teevan 和 D. R. Karger，解决朴素贝叶斯文本分类器的不合理假设，载于第 *20* 届国际机器学习会议论文集（*ICML-03*） （2003），第 616–623 页。[35] M. R¨oder、A. Both 和 A. Hinneburg，探索主题连贯性度量的空间，载于第八届 *ACM Web* 搜索与数据挖掘国际会议论文集（2015），第 399–408 页。