

基于引用关系的政策与论文关联图谱 构建与应用研究*

任超¹ 杨孟辉² 许嘉元³

(1. 宁波工程学院人文与艺术学院, 宁波 315211; 2. 中国人民大学信息资源管理学院, 北京 100872;
3. 中国科学院文献情报中心, 北京 100190)

摘要: 当前, 科学研究成果在政策制定中的作用日益凸显, 但二者之间的隐性关联尚不明晰且挖掘难度较大, 亟须构建系统性工具以帮助政策制定者准确有效地选择科学证据。从政策引用的视角出发, 旨在基于大规模数据构建政策与论文关联图谱, 在此基础上为政策制定提供科学论文推荐。在图谱构建阶段, 以Overton和OpenAlex两大数据库为基础数据源, 设计并构建融合多类型、多层级实体的知识图谱模式层(包括9种实体与5种关系), 并采用自顶向下的知识图谱构建策略, 抽取并生成47 327 880个语义三元组, 存储于Neo4j图数据库中, 实现高效查询与可视化支持。在图谱应用阶段, 使用6种知识图谱推理技术, 结合5种评价指标对结果进行评价, 结果表明所提方法能够更为准确和高效地为政策制定推荐科学论文。研究不仅为实现基于知识图谱技术的政府决策提供了可行框架和具体建议, 也为政策智能化支持工具的探索提供了有价值的理论与实践基础。

关键词: 引文分析; 知识图谱; 政策引用论文; 论文推荐; 循证决策

中图分类号: G353.1; D63 **DOI:** 10.3772/j.issn.1673-2286.2025.08.008

引文格式: 任超, 杨孟辉, 许嘉元. 基于引用关系的政策与论文关联图谱构建与应用研究[J]. 数字图书馆论坛, 2025, 21(8): 76-84.

近年来, 以实证证据为基础的政策制定范式逐渐兴起, 强调在政策形成与调整的过程中广泛收集并合理利用相关证据^[1-2], 包括科学式证据、实践性证据、地方性证据3种类型^[3]。其中, 科学式证据指的是通过一些科学方法或符合科学规定的程序或过程而产生的信息, 而政策文件对科学论文的引用行为正是基于这类证据进行决策的重要外在表现。当前, 科学论文凭借其严谨的研究设计、系统的数据分析及严格的同行评议机制, 已成为政策文件中最为重要且首选的权威信息来源^[4-5]。然而, 将科学论文真正有效地引入政策制定实践依然存

在诸多挑战^[6], 即多源、异构、实时更新的海量数据使得政策制定者面临严重的信息过载, 且政策与论文之间的深层次内在联系和知识链条尚未被广泛挖掘。从海量复杂的学术文献中甄别、筛选并提炼与政策主题高度契合的信息, 始终是一项复杂且充满挑战的任务。

鉴于知识图谱的关联挖掘和智能推荐能力有助于显著提升有效信息的发现效率, 为科学决策提供更加丰富和有力的支持^[7], 本文拟使用知识图谱技术推荐学术论文, 为政策制定提供科学证据, 也为实现智能、高效、可追溯的循证决策提供支持。

收稿日期: 2025-06-17

*本研究得到国家社会科学基金项目“基于图书全内容的知识发现与智能服务研究”(编号: 22BTQ068)、宁波工程学院科研启动基金项目“多源数据驱动的科学传播生态系统研究”(编号: 24KQ050)资助。

1 文献综述

1.1 政策引用论文相关研究

政策对学术论文的引用是近年来替代计量学领域一个迅速发展的核心议题^[8]。围绕这一议题, 学界将科学论文在政策文件中的引用频次作为衡量科学论文社会影响力的关键标尺, 相关的实证探索大致可归为3个主要方向。①对政策引用现象的宏观描绘与特征分析, 此类研究旨在揭示科学论文如何被政策所用。例如, 刘晓娟等^[9]从被引文献角度系统分析了具有何种学科背景、发表于何种期刊的成果更易获得政策青睐, 从施引政策角度分析了学术知识具体转化为了何种类型与层级的公共政策, 以及从引用行为本身分析了科学论文在政策文本中扮演的角色(是作为立论依据、背景知识还是具体方法)。②对影响政策引用关键因素的深度探析, 此类工作聚焦于解释政策引用现象背后的驱动机制。例如, Fang等^[10]从传播学视角发现, 科学论文在社会媒体中的传播广度是其被政策引用的重要正向预测因子, 从而揭示了学术影响力向社会影响力转化的路径。余厚强等^[11]的计量分析则进一步验证, 成果本身的替代计量得分、传统被引频次, 乃至标题长度与是否获得基金支持等, 均是影响其被政策引用概率的显著变量。③对政策引用指标本身有效性与应用前景的评估与拓展, 此类研究旨在检验该指标的信度与效度, 并探索其应用边界。这包括对政策引用指标能否真实反映社会影响力的效度检验^[12], 对Altmetric.com等核心数据源中政策引用数据的准确性评估与偏差分析^[13], 以及利用机器学习方法, 探索以社交媒体热度等先行指标预测未来政策引用的可行性, 为科研影响力的早期识别提供创新工具^[14]。

1.2 政策知识图谱相关研究

知识图谱作为一种由节点与边构成的图状数据结构, 能够将非结构化知识显性化与体系化^[15], 目前在数字人文^[16]、知识组织^[17]、医疗健康^[18]等领域得到广泛应用, 也已成为赋能政策研究的有力工具。当前, 主要形成了3个研究方向。①政策文本的深度解构与语义分析, 这类任务的核心在于将无结构的复杂政策文本拆解为结构化的、可计算的知识单元。为实现这一目标, 研究者通常构建一个包含政策主体、客体、工具等核心

要素的本体模型^[19], 进而借助深度学习或关联规则等技术, 完成对文本的细粒度信息抽取^[20]。通过这一过程, 政策得以转化为由关键要素及其关系构成的语义网络, 为精准剖析政策的内部结构提供了可能^[21]。②跨政策的关联挖掘与网络发现, 此类工作致力于揭示政策体系的宏观图景。例如, 韩娜等^[22]利用关联规则推理, 在特定主题下构建政策协同性图谱, 从而能够直观地评估不同政策在目标上的一致性或潜在冲突。这种将孤立的政策点连接成大规模政策网络的方法, 为审视政策体系的内在一致性、发现政策空白或功能重叠提供了全新的计算分析视角^[21]。③面向用户的智能检索与知识服务, 其核心目标在于将上述结构化的知识进一步转化为面向用户的智能检索与知识服务系统, 彻底优化政策信息的组织、管理与获取效率。无论是面向海量的科技政策^[23], 还是聚焦于新冠疫情等特定领域的公共政策^[24], 研究者均可通过构建知识图谱并利用Neo4j等图数据库进行管理, 实现对复杂语义查询的支持, 用户因此能够摆脱传统关键词检索的局限, 极大提升了政策知识发现的效率与广度。

综上, 在现有研究中, 关于政策引用论文的研究多集中于引用频次等定量指标分析, 而对政策背景、决策环境以及社会背景等影响因素的定性探讨仍显不足。尤其是在政策制定过程中, 各主体之间的直接关系尚未得到系统剖析, 对政策文件本身的深入挖掘也相对缺乏, 这为本文研究提供了有力的切入点。同时, 现有知识图谱应用的主要局限在于, 其在动态政策环境下的适应性与演化过程关注较少, 尽管部分研究尝试构建基于知识图谱的政策推荐系统, 但对于政策与其他知识系统间复杂关联的洞察仍须进一步深化。针对上述不足, 本文基于大规模政策和论文数据, 尝试通过知识图谱技术构建政策与科学论文的关联图谱, 并使用知识图谱推理算法为政策制定推荐科学论文, 作为循证决策的科学式证据。

2 研究设计

在复杂多变的公共治理与科技环境中, 政策制定亟需以证据为基础的系统方法来提升科学性、可追溯性与透明度, 政策文件对科学论文的引用正是此过程中的一个关键且可观测的信号。围绕这一核心关系进行建模, 具有不可替代的独特优势: ①引用是明确的证据指向, 天然承载了方向性与时序性, 能够将政策文本

与科学论文有机连接,形成可解释的证据链;②大规模数据集,如Overton^[25]和OpenAlex^[26],提供了规模化、结构化的引用与元数据,便于跨源对齐与持续更新;③引用关系适配图推理与链路预测任务,既能产出可验证的结果,又可通过路径与规则给出可读的推荐理由。因此,构建基于引用关系的政策与论文关联图谱,并在这一基础上开展推理驱动的论文推荐,既回应了循证的核心诉求,也为大规模、可解释的政策智能化提供了坚实的技术抓手。

(1) 政策与论文关联图谱构建。以Overton与OpenAlex为数据底座,围绕2021年发布的政策文件及引用的论文数据开展系统整合,构建政策与论文关联图谱。政策与论文关联图谱以知识图谱技术为基础,以图结构对政策、论文相关的实体和关系进行表示,以便更好地理解和分析政策与论文之间的关系,以及论文对政策的影响。

在模式层中,采用自顶向下的方法设计本体与模式;在数据层中,以模式层为基础进行知识抽取,包括实体抽取和关系抽取;在数据存储与可视化阶段,使用两种方法对抽取的知识图谱数据进行存储:一种是以三元组的形式保存至本地CSV文件,另一种则是导入Neo4j图数据库。

(2) 政策与论文关联图谱应用。面向政策制定的科学论文推荐是政策与论文关联图谱应用的核心目标。具体来说,循证政策制定中的科学论文推荐就是从大量的待选论文中选择出若干合适的论文,这实际上就是知识图谱推理过程,即通过一个已知的实体和关系推理出另一个实体,或者通过两个实体推理它们之间的关系^[27]。

在知识图谱推理中,常见任务是通过一个已知的实体和关系推理出另一个实体,或者通过两个实体推理它们的关系,即 $(h, ?, r)$ 、 $(?, t, r)$ 和 $(h, t, ?)$ 。以三元组(政策A, 论文B, 引用)为例,当论文B未知时,该三元组就变为(政策A, ?, 引用),即尾实体缺失。在模型训练时,将知识图谱中的每个实体都放在尾实体的位置上,并且放入相应的知识图谱嵌入模型的得分函数,计算不同实体作为该三元组的尾实体的得分,也就是该三元组的合理性,得分最高的实体会被视为知识图谱推理的结果。同时,将大量政策与论文知识图谱中的三元组数据输入推理模型后,该模型能够较为全面地学习不同实体和关系间的潜在特征,进而实现对未知三元组的有效推理,因此能够在一定程度上解决

循证政策制定中的科学论文推荐问题。

基于此,使用6种主流的知识图谱推理技术进行训练与推理,使用5种评价指标和6种评价方式对模型效果进行评价,以衡量面向政策的科学论文推荐效果。

3 政策与论文关联图谱的构建

3.1 数据获取

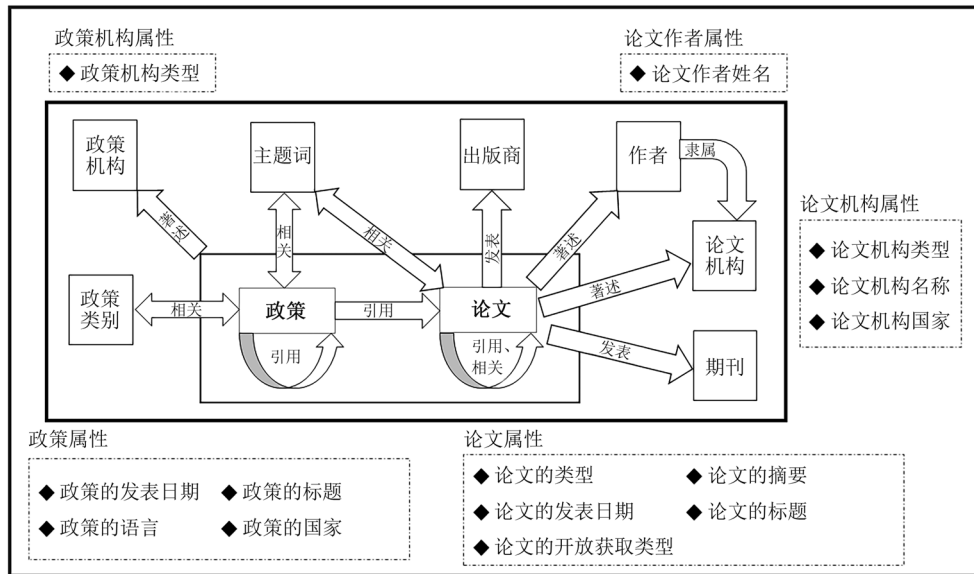
研究数据来源于Overton数据库和OpenAlex数据库。检索并获得了Overton数据库中2021年发布的全部政策数据,共计618 429条数据,在数据预处理后筛选出588 786条有效数据,其中有51 619份政策文件引用了468 194篇科学论文,存在着750 346个政策对论文的引用关系。为获取上述科学论文数据,使用OpenAlex数据库进行检索。使用OpenAlex数据库检索上述468 194篇科学论文的DOI,最终获得了44 875份政策引用的419 936篇论文,共包含655 249个引用关系。

同时,从Overton数据库中收集了2021年发布的政策对其他政策的引用数据,最终获得了82 281份政策文件引用了141 924份政策,存在655 249个政策对政策的引用关系。由于部分政策同时引用了政策和论文,最终获得了95 549份政策文件引用了419 936篇科学论文和141 924份政策。

3.2 知识图谱模式层

根据收集到的数据构建政策与论文关联图谱模式层(见图1),共包括9种实体、5种关系和14种属性。

实体信息包括政策、论文、政策机构、论文机构、主题词、政策类别、出版商、作者、期刊。这些实体属于不同类型、不同层级,它们之间存在一定的逻辑关联和连接关系,共同构建起丰富的知识结构,也体现了知识图谱面向多维度、多层次信息融合的特点。政策和论文属于文档实体,分别指单篇政策文件和单篇科学论文,直接承载了决策过程中的主要证据,其中政策实体包括引用论文的政策、引用政策的政策、被政策引用的政策,同时也存在某些政策在引用了政策或论文的同时也被其他政策所引用,本文在此不作详细区分,仅将政策分为施引政策和被引政策两类。同理,论文实体也被分为施引论文和被引论文两类。政策机构和论文机构属于组织实体,代表对应的机构名称,体现政策制定



与科学研究的社会结构依托,其中:政策机构是指发表该政策的机构名称,其属性为该机构的类型,一般可分为政府机构、智库和国际组织等;论文机构是指论文作者的署名机构,其属性包含机构类型,以高等院校和科研院所为主。主题词和政策类别属于内容层面的实体,用于描述文档的主题和分类,涵盖了内容层面的聚合与分类,其中:主题词指Overton数据库提供的政策主题词和OpenAlex提供的论文主题词;政策类别是指Overton数据库提供的政策所属类别,包括环境、教育、卫生、科学技术等18个大类。出版商、作者、期刊作为其他类别实体,也在知识图谱中承担不同角色,确保了科学论文评价、溯源及影响追踪的可操作性,这些实体信息均从OpenAlex提供的数据库字段提取而来。

关系信息包括引用关系、相关关系、隶属关系、发表关系、著述关系。引用关系包括政策引用政策、政策引用论文、论文引用论文,引用关系作为知识传递和证据支撑的核心纽带,系统刻画了政策与政策、政策与论文、论文与论文之间的直接或间接影响链条,为决策溯源、证据追踪和知识流动分析奠定了基础;相关关系包括政策与主题词的相关关系、论文与主题词的相关关系、政策与类别的相关关系、论文与论文的相关关系,相关关系进一步强化了内容语义层面的实体关联,使政策与主题词、政策与类别、论文与主题词、论文与论文之间的主题关联与影响路径得以清晰表达;隶属关系包括论文作者与机构的隶属关系,隶属关系细化个体作者与组织机构之间的结构性联系,为机构绩效评

估以及科研合作网络分析提供了关键支持;发表关系包括论文与出版商的发表关系、论文与期刊的发表关系,著述关系包括政策与机构的著述关系、论文与作者的著述关系、论文与机构的著述关系,发表关系和著述关系分别将论文的学术发表与研究产出的社会分布展现出来,有助于快速识别权威出版资源和领域领军人物。

属性信息包括政策的发表日期、政策的语言、政策的标题、政策的国家、论文的类型、论文的开获取类型、论文的发表日期、论文的摘要、论文的标题、论文机构类型、论文机构名称、论文机构国家、论文作者姓名、政策机构类型。这些属性既有助于实现高效的信息筛选、检索与排序,也为后续的量化评估与智能推荐模型提供了丰富的特征输入,在提升政策决策过程的效率与证据精度方面发挥着关键作用。

3.3 知识图谱数据层

知识图谱数据层的主要任务是以模式层为基础进行知识抽取,包括实体抽取和关系抽取。抽取的实体数据量和字段信息如表1所示。论文实体数据共6 173 890条,包含了被政策直接引用的419 936篇核心论文,以及这些核心论文所引用的5 753 954条参考文献;政策实体数据共371 710条,包括95 549条施引政策数据、141 924条被其引用的一级被引政策数据,以及134 237条更深层次的二级被引政策数据;主题词实体数据共

表1 实体数据量和字段信息

实体	数据量/条	占比/%	字段信息
政策	371 710	4.76	唯一标识符、名称、发表日期、语言、标题、国家、标签
论文	6 173 890	79.12	唯一标识符、名称、开放获取类型、发表日期、摘要、标题、标签
政策机构	937	0.01	唯一标识符、名称、政策机构类型标签
论文机构	22 096	0.28	唯一标识符、名称、论文机构类型、论文机构名称、论文机构国家、标签
主题词	149 423	1.92	唯一标识符、名称、标签
政策类别	18	0.01	唯一标识符、名称、标签
出版商	7 117	0.09	唯一标识符、名称、标签
作者	1 049 871	13.45	唯一标识符、名称、论文作者姓名、标签
期刊	27 914	0.36	唯一标识符、名称、标签

有149 423条,包括128 204条政策主题词数据和39 014条论文主题词数据,其中有17 795条重合。

实体的字段信息主要包括唯一标识符、名称、标签和属性信息。①唯一标识符是全局唯一的字符串,用于精确区分图谱中的每个实体。政策、论文、论文机构、作者4种实体的唯一标识符分别使用Overton提供的政策ID和OpenAlex提供的论文ID、论文机构ID、作者ID,其他5种实体的唯一标识符是由本研究生成的、可唯一识别的字符串。②名称是指实体的自然语言描述或标题,在知识图谱中用于全文检索和用户交互。③标签是指实体的分类标记或类型,通常为一个或多个类别标签,定义了实体的本体类,可支持推理和模式匹配。④属性信息是描述该实体内在特征和外在联系的一系列“键-值”对,除了ID、名称和标签之外,其他所有描述性的信息都属于属性。

同时,也从上述数据中抽取出了5种关系,关系的类型、名称和数量信息如表2所示。引用关系总体占比超过了44%,尤其是论文引用论文关系占比超过36%;论文与论文的相关关系占比超过了18%,政策和论文与主题词、类别之间的相关关系占比超过了25%。

3.4 知识图谱存储与可视化

使用两种方法对抽取到的知识图谱数据进行存储:一种是以三元组的形式保存至本地CSV文件,另一种则是导入Neo4j图数据库。

在第一种存储方式中,三元组是指(头实体,关

表2 关系的类型、名称和数量信息

关系类型	关系名称	数据量/条	占比/%
引用关系	政策引用政策	998 386	2.33
	政策引用论文	2 427 773	5.67
	论文引用论文	15 524 161	36.26
相关关系	政策与主题词的相关关系	6 911 339	16.14
	论文与主题词的相关关系	3 512 940	8.21
	政策与类别的相关关系	624 059	1.46
	论文与论文的相关关系	7 777 518	18.17
发表关系	论文与出版商的发表关系	412 270	0.96
	论文与期刊的发表关系	405 014	0.95
著述关系	政策与机构的著述关系	223 324	0.52
	论文与作者的著述关系	1 904 848	4.45
	论文与机构的著述关系	738 678	1.73
	论文与第一作者的著述关系	419 936	0.98
隶属关系	论文作者与机构的隶属关系	927 954	2.17

系,尾实体)的形式,其中的关系包括抽取到的5种关系和14种属性信息,尾实体中也有一部分是属性信息。最终获得了47 327 880个三元组,以CSV的形式保存至本地,以便后续使用。

在第二种存储方式中,在服务器中搭建了Neo4j数据库,使用Neo4j的批量导入功能将上述9种实体、5种关系和14种属性全部导入Neo4j数据库。在大规模知识图谱项目中,数据导入丢失率通常为5%~15%。最终成功导入了7 802 976条实体数据、38 590 380条关系数据,数据丢失率为9.84%,属于正常范围。

所构建的政策与论文关联图谱为科学证据的组织与利用提供了核心支撑。通过系统整合政策文件、学术论文、机构、主题词等多源异构数据,图谱将分散的信息以实体和关系的形式结构化表达,从而为科学证据的全面梳理和高效利用创造条件。与传统的信息检索方式相比,图谱不仅实现了证据的有机融合和语义关联,还揭示了政策与科学论文间的复杂联系,如引用、相关、著述等多维路径,极大提升了证据溯源和链条分析的能力。此外,图谱中的属性信息能够支持基于主题、时序等多维度的证据智能推荐与检索,便于政策制定者快速获取权威且有针对性的科研支撑。因此,构建的政策与论文关联图谱作为科学证据的结构化载体和智能推理平台,为循证决策提供了坚实的数据基础和方法保障,有助于提升决策的科学性、系统性与透明度。

4 政策与论文关联图谱的应用

构建政策与论文关联图谱的最终目的是服务于循证决策的现实需求, 打通科学知识与实践之间的通道。鉴于“为政策制定者精准匹配科学证据”是连接二者最直接、最关键的应用场景, 本章将聚焦于图谱最核心的应用功能——面向政策制定的科学论文推荐。该应用在技术上可转化为知识图谱领域经典的链接预测任务, 即预测政策实体与论文实体之间可能存在的引用关系。为此, 遵循实证研究的逻辑: 首先, 阐明实现该应用所需的技术方法与效果评价体系; 其次, 详述包括数据准备与参数设置在内的模型构建全过程; 最后, 通过展示和分析实验结果, 系统地验证该推荐应用的实际效能与稳健性。这一结构旨在完整地呈现本研究成果从理论构建到实践应用的闭环。

4.1 科学论文推荐方法与效果评价

使用6种主流的知识图谱推理技术, 即3种基于平移的模型TransE^[28]、TransH^[29]和TransD^[30], 以及3种基于张量分解的模型Simple^[31]、DistMult^[32]和ComplEX^[33]。

由于科学论文推荐方法是常用的知识图谱推理技术, 推荐效果评价指标使用知识图谱推理中常用的5种指标, 即MR (Mean Rank)、MRR (Mean Reciprocal Ranking)、HITS@1、HITS@3、HITS@10^[34]。MR用于衡量知识图谱嵌入模型在预测头尾实体关系时的准确性, MR值越小, 模型的推理能力越强。MRR通过计算逆排名的平均值来度量模型的性能, 该指标越大越好。HITS@ n 是指排名小于 n 的三元组的平均占比, 一般地, 取 n 等于1、3或者10, 该指标越大越好。

此外, 在知识图谱推理模型评估中, 通常采用6种方式: 原始数据左实体效果评估、原始数据右实体效果评估、原始数据平均效果评估、过滤数据左实体效果评估、过滤数据右实体效果评估, 以及过滤数据平均效果评估。这些评估依赖于虚拟三元组, 即通过替换原始三元组中的元素生成虚假三元组, 用于模拟模型对未知事实的推断和泛化能力。其中, 左实体指头实体, 右实体指尾实体。原始数据左实体效果评估预测左实体性能, 考虑所有虚拟三元组, 包括训练、验证、测试集中已存在的正确三元组, 衡量模型对已知与未知事实的处理能力; 右实体效果评估类似, 针对右实体; 平均效果评估

为二者均值, 提供综合度量。过滤数据效果评估则排除已存在的正确三元组: 左实体评估强调模型区分已知与未知左实体的能力; 右实体评估类似, 针对右实体; 平均评估为二者均值, 突出对未知事实的泛化能力。

4.2 科学论文推荐模型构建

(1) 数据准备。由于数据量过于庞大, 无法将全部数据一次性输入知识图谱推理模型进行训练和测试。尝试以政策为核心、月份为单位, 随机选取某一月数据进行模型的训练和测试。

首先, 从全部数据中选取了2021年2月发布的政策数据, 共计45 770条, 其中有7 713条引用了论文或其他政策; 其次, 从全部数据中抽取出这7 713条政策数据相关的所有信息, 以三元组的形式保存; 最后, 考虑到知识图谱推荐任务的特性, 论文的摘要、论文的标题和政策的标题信息在三元组数据中并无较大的作用, 其中包含的丰富的文本特征也难以在知识图谱推理过程中展现出来, 且会大量增加运行复杂度, 将这3种属性删除, 最终获得了25种三元组。

在知识图谱推理实验中, 将训练集、测试集和验证集以8:1:1的比例进行划分, 其中共包含25种关系、1 537 478个实体或属性, 训练集有2 698 051个三元组, 测试集和验证集均有337 256个三元组。实验中的数据均使用三元组的形式输入, 具体的数据处理方法和6种知识图谱推理技术的参数设置参考文献[34]。

(2) 模型参数设置。以上述数据为输入, TransE、TransH、TransD、Simple、DistMult和ComplEX模型的参数设置如下。

在数据加载阶段, 6种模型的批次数为100, 线程数为8, 采样模式为正常; 使用伯努利分布, 使用过滤; 负样本中实体数为25, 关系数为0。在模型定义时, 所有模型的嵌入维度数均为200, TransE、TransH、TransD模型的范数设置为1, 使用范数归一化。TransE、TransH、TransD模型的损失函数为Margin Loss, Margin分别设置为5.0、4.0、4.0; Simple、DistMult和ComplEX模型的损失函数为Softplus Loss。在模型训练阶段, TransE和TransD模型的训练轮数为1 000, 学习率为1.0; TransH模型的训练轮数为1 000, 学习率为0.5; Simple、DistMult和ComplEX模型的训练轮数均为2 000, 学习率为0.5, 模型优化使用Adagrad算法。所有模型均使用GPU运行。

4.3 科学论文推荐的结果分析

使用TransE、TransH、TransD、Simple、DistMult和Complex这6种知识图谱推理算法对2021年2月的政策与论文关联图谱进行推理,以MRR、MR、HITS@10、HITS@3和HITS@1这5种指标为评价标准,得到推理实验结果。实际上,上述模型在本质上执行知识图谱推理中的实体预测任务,能够实现对所有头实体和尾实体的预测,因此所获得的结果是在对多种三元组关系推理的基础上获得的平均效果,但这一结果并不

能直观地展示模型对政策引用论文这一三元组的实体预测效果。

鉴于此,从上述6种模型中选出效果相对较好的TransE和TransH模型,单独针对政策引用论文这一三元组进行推理。具体来说,同样使用2021年2月的数据,在保证训练集不变的情况下,将测试集和验证集中除政策引用论文之外的其他24种三元组全部删除,其他参数和操作不变,最终修改后的测试集样本数据量为4 184条,训练集样本数据量为4 394条。TransE和TransH模型获得的推理结果如表3所示。

表3 TransE和TransH模型对政策引用论文的推理结果

推理算法	评估方式	MRR	MR	HITS@10	HITS@3	HITS@1
TransE	过滤数据左实体效果	0.478 548	536.642 212	0.667 543	0.522 228	0.382 170
	过滤数据右实体效果	0.132 490	2 399.305 420	0.230 402	0.145 076	0.079 111
	过滤数据平均效果	0.305 519	1 467.973 877	0.448 972	0.333 652	0.230 641
TransH	过滤数据左实体效果	0.451 051	449.993 317	0.638 862	0.495 937	0.353 250
	过滤数据右实体效果	0.115 965	3 300.206 543	0.215 583	0.122 610	0.067 161
	过滤数据平均效果	0.283 508	1 875.099 976	0.427 223	0.309 273	0.210 206

由于在过滤数据中的效果评估是更接近真实的评估方式,表3仅展示了过滤数据左实体效果评估、过滤数据右实体效果评估和过滤数据平均效果评估3种评估方式,评价指标使用MRR、MR、HITS@10、HITS@3和HITS@1这5种评价指标。

与最初得到的结果对比,表3中所有的评估方式和评价指标都是更优的:所获得的MRR更高、MR更低,说明整体排名质量更高;而HITS@10、HITS@3和HITS@1的明显提升则表示推荐列表中正确三元组的数量明显增加。这是由于最初结果为对25种三元组进行推理的平均效果,仅对政策引用论文进行推理的效果明显优于对25种三元组进行推理的效果,或许是其余24种三元组的推理结果拉低了整体平均值,这也表明所构建的基于知识图谱推理的科学论文推荐模型更适合处理对政策引用论文的推理。

此外,在对政策引用论文进行推理时,左实体表示施引政策,右实体表示被引论文。从表3可以看出,TransE和TransH模型在过滤数据左实体效果评估中获得了更好的效果,这说明模型在对左实体进行推理时会获得更高的准确率。同时,TransE模型获得了最优效果,尤其是在过滤数据左实体效果评估和过滤数据右实体效果评估中的HITS@10分别为0.667 543和0.230 402。这意味着在针对政策引用论文的推理

任务中,在给定右实体(被引论文)的基础上,模型从候选政策集合中选出10份施引政策时,命中的概率为66.754 3%;在给定左实体(施引政策)的基础上,模型从候选论文集合中选出10篇被引论文时,命中的概率为23.040 2%。

更直观地说,给定一个示例三元组事实(政策A,论文B,引用关系),在头实体缺失时,上述模型推荐的结果列表中,政策A位于列表前10的概率为66.754 3%,位于列表前3的概率为52.222 8%,位于列表前1的概率为38.217 0%;在尾实体缺失时,上述模型推荐的结果列表中,论文B位于列表前10的概率为23.040 2%,位于列表前3的概率为14.507 6%,位于列表前1的概率为7.911 1%。因此,模型在为政策制定提供科学论文推荐时,会为每一份政策提供一个候选论文列表,而真正对这份政策有价值的论文在候选论文列表前10位的概率为23.040 2%,在前3位的概率为14.507 6%,在第1位的概率为7.911 1%,这表明所提方法已实现相当优异的推荐效果。

5 结语

本研究以政策文件对科学论文的引用为基础,成功构建了一个大规模的政策与论文关联图谱。基于

Overton与OpenAlex两大权威数据源, 设计了涵盖9种实体、5种关系的本体模型, 并完成了包含7 802 976条实体数据、38 590 380条关系数据的大型知识图谱构建。在应用层面, 将面向政策的科学论文推荐问题转化为知识图谱的链接预测任务, 实验结果验证了所提方法的可行性与有效性: 在为政策推荐科学论文作为决策依据时, 模型能够实现高达23%的前10名命中率, 显著提升了证据检索的效率。这一成果不仅为多源异构情报数据的融合与应用提供了可行的技术方案, 也为实现智能、高效、可追溯的循证决策提供了新的方法论支持。

本研究主要有以下局限: ①数据仅限于2021年的静态快照, 难以反映政策与学术研究的动态演进, 可能导致时间偏置与主题漂移风险; ②本体设计采用自顶向下方法, 虽有助于明晰结构, 但可能遗漏新兴概念或隐含关系, 进而影响证据链条的完整性与推荐的可解释性。为降低上述偏差, 后续研究将引入多年度与实时更新的数据源, 构建具备时间戳与有效期的动态知识图谱, 并引入自底向上的本体演化机制, 结合多种方式持续扩充与修正本体结构, 以提升知识图谱的鲁棒性与可扩展性。

参考文献

- [1] 周志忍, 李乐. 循证决策: 国际实践、理论渊源与学术定位[J]. 中国行政管理, 2013 (12): 23-27, 43.
- [2] 魏夏楠, 张春阳. “循证决策”30年: 发展脉络、研究现状和前沿挈领: 基于国内外代表性文献的研究综述[J]. 现代管理科学, 2021 (4): 26-36.
- [3] 张继亮. 循证政策: 政策证据的类型、整合与嵌入[J]. 社会科学, 2019 (11): 39-47.
- [4] 方志超, 郑尔特. 科学论文政策影响力计量的数据库选择: 基于Altmetric与Overton的比较[J]. 图书情报知识, 2025, 42 (1): 18-28.
- [5] 曹喆, 张琳, 尚媛媛. 如何提升中国学术论文的国际政策影响力: 基于多维计量特征分析的策略研究[J]. 图书情报知识, 2025 (1): 29-43.
- [6] 任超, 杨孟辉, 赵群. 循证政策中的科学证据特征分析: 以新冠疫情防控政策为例[J]. 情报理论与实践, 2023, 46 (7): 98-106, 124.
- [7] 周红磊, 张海涛, 刘伟利, 等. 面向重大突发事件应急管理的事件知识图谱构建及场景应用[J]. 情报学报, 2024, 43 (12): 1453-1466.
- [8] 余厚强, 肖婷婷, 王曰芬, 等. 政策文件替代计量指标分布特征研究[J]. 中国图书馆学报, 2017, 43 (5): 57-69.
- [9] 刘晓娟, 周若卿, 肖云彤, 等. 政策提及指标在学术成果社会影响力评价中的应用价值分析[J]. 情报理论与实践, 2024, 47 (11): 47-55.
- [10] FANG Z C, COSTAS R, TIAN W C, et al. An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics[J]. Scientometrics, 2020, 124 (3): 2519-2549.
- [11] 余厚强, 李龙飞. 政策文件替代计量指标影响因素研究[J]. 情报理论与实践, 2021, 44 (7): 28-36.
- [12] HAUNSCHILD R, BORNMANN L. How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data[J]. Scientometrics, 2017, 110 (3): 1209-1216.
- [13] YU H Q, CAO X T, XIAO T T, et al. How accurate are policy document mentions? A first look at the role of altmetrics database[J]. Scientometrics, 2020, 125 (2): 1517-1540.
- [14] KALE B, KALE B, SIRAVURI H V, et al. Predicting research that will be cited in policy documents[C]//Proceedings of the 2017 ACM on Web Science Conference. New York: ACM Press, 2017: 389-390.
- [15] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53 (3): 582-600.
- [16] 朱兰兰, 霍婕, 高玉婷. 馆藏家谱文献知识化开发: 价值、主体与过程[J]. 数字图书馆论坛, 2023, 19 (11): 38-45.
- [17] 徐晨飞, 唐佳林. 方志书目提要语义化知识组织与知识发现研究[J]. 数字图书馆论坛, 2024, 20 (11): 30-42.
- [18] 孟秋晴, 郑铭瑞, 田玥璐, 等. 面向在线健康社区UGC的医疗健康知识图谱构建研究: 以小儿腹泻病为例[J]. 数字图书馆论坛, 2024 (8): 9-18.
- [19] 赵雅洁, 冯凌子, 袁军鹏, 等. 多维细粒度政策知识图谱构建方法[J/OL]. 数据分析与知识发现: 1-21[2025-04-09]. <https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFD&filename=XDTQ20250408001>.
- [20] 马续补, 张阁静, 李波, 等. 基于多要素语义关联的政策知识图谱构建[J]. 情报理论与实践, 2025, 48 (8): 93-104.
- [21] 张维冲, 王芳, 赵洪. 基于全要素网络构建的大规模政策知识关联聚合研究[J]. 情报学报, 2023, 42 (3): 289-303.
- [22] 韩娜, 马海群, 刘兴丽. 基于知识图谱的政策文本协同性推理研究[J]. 情报科学, 2021, 39 (11): 180-186.

- [23] 张雨, 吴俊. 科技政策知识图谱构建研究[J]. 数字图书馆论坛, 2021 (8) : 31-38.
- [24] 霍朝光, 钱毅, 祁天娇. 基于开放公文的新肺炎政策知识图谱构建与分析[J]. 档案学通讯, 2021 (2) : 53-62.
- [25] SZOMSZOR M, ADIE E. Overton: a bibliometric database of policy document citations[J]. Quantitative Science Studies, 2022, 3 (3) : 624-650.
- [26] PRIEM J, PIWOWAR H, ORR R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts [EB/OL]. [2025-05-12]. <https://arxiv.org/abs/2205.01833>.
- [27] 任超, 杨孟辉, 杨冠灿, 等. 基于知识图谱的循证政策中科学证据推荐研究: 以新冠肺炎疫情防控政策为例[J]. 图书情报工作, 2023, 67 (2) : 108-118.
- [28] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data [EB/OL]. [2025-05-12]. <https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html>.
- [29] WANG Z, ZHANG J W, FENG J L, et al. Knowledge graph embedding by translating on hyperplanes[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2014, 28 (1) : 8870.
- [30] JI G L, HE S Z, XU L H, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2015: 687-696.
- [31] KAZEMI S M, POOLE D. Simple embedding for link prediction in knowledge graphs[EB/OL]. [2025-05-12]. <https://proceedings.neurips.cc/paper/2018/hash/b2ab001909a8a6f04b51920306046ce5-Abstract.html>.
- [32] YANG B S, YIH W T, HE X D, et al. Embedding entities and relations for learning and inference in knowledge bases[EB/OL]. [2025-05-12]. <https://arxiv.org/abs/1412.6575>.
- [33] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[EB/OL]. [2025-05-12]. <http://proceedings.mlr.press/v48/trouillon16.html?ref=https://githubhelp.com>.
- [34] HAN X, CAO S L, LV X, et al. OpenKE: an open toolkit for knowledge embedding[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg: ACL, 2018: 139-144.

作者简介

任超, 男, 副研究员, 研究方向: 科学计量与政策计量。

杨孟辉, 男, 教授, 研究方向: 信息计量。

许嘉元, 男, 博士, 通信作者, 研究方向: 信息分析, E-mail: xujiayuan@mail.las.ac.cn。

Construction and Application of Policy-Paper Association Graph Based on Citation Relationships

REN Chao¹ YANG MengHui² XU JiaYuan³

(1. School of Humanities and Arts, Ningbo University of Technology, Ningbo 315211, P. R. China; 2. School of Information Resource Management, Renmin University of China, Beijing 100872, P. R. China; 3. National Science Library, Chinese Academy of Sciences, Beijing 100190, P. R. China)

Abstract: While the role of scientific research in policymaking is increasingly significant, the latent associations between them remain obscure and are difficult to mine. This creates an urgent need for systematic tools that can aid policymakers in selecting scientific evidence accurately and effectively. Adopting the perspective of policy citations, this paper aims to construct a knowledge graph of policy-paper linkages from large-scale data to provide scientific paper recommendations for policy formulation. In the construction phase, we utilize the Overton and OpenAlex databases to design a schema layer for the knowledge graph, integrating multi-type and multi-level entities (nine entity types and five relation types). A top-down construction strategy is then employed to extract and generate 47 327 880 semantic triples, which are subsequently stored in Neo4j graph database to enable efficient querying and visualization. In the application phase, we apply six distinct knowledge graph reasoning techniques and evaluate their performance using five evaluation metrics. The results demonstrate that the proposed method can recommend scientific papers for policymaking with high accuracy and efficiency. This study not only provides a feasible framework and concrete suggestions for knowledge graph-driven government decision-making, but also establishes a valuable theoretical and practical foundation for the development of intelligent policy support tools.

Keywords: Citation Analysis; Knowledge Graph; Policy-Citing Paper; Paper Recommendation; Evidence-Based Policy Making

(责任编辑: 王玮)