基于node2vec模型的多维度科研合作学者推荐研究

袁永旭[1,2]　成韬[1]
（1. 山西医科大学管理学院，太原 030001；2. 山西医科大学图书馆，太原 030001）

摘要：为了更好更全面地发现研究主题契合度高且合作可能性大的学者，提出一种基于图嵌入模型的多维度学者推荐模型。首先，使用CSSCI数据库作为数据源，构建作者间全连通网络，将作者间关键词相似度、合作关系、引用关系及机构属性等维度信息转化为作者间连线的权重。其次，使用node2vec图嵌入模型对该网络进行深度学习，得到每个作者的节点坐标向量。最后，通过计算作者间向量相似度达到合作学者推荐的目的。结果表明，提出的多维图嵌入模型能够有效地完成合作学者的推荐，可以为学者提供有价值的推荐名单，从而促进学术合作和交流。

关键词：科研合作；学者推荐；图嵌入；node2vec；word2vec
中图分类号：G203　DOI：10.3772/j.issn.1673-2286.2024.12.009
引文格式：袁永旭，成韬. 基于node2vec模型的多维度科研合作学者推荐研究[J]. 数字图书馆论坛，2024，20（12）：77-86.

1963年，著名文献计量学家Derek J. de Solla Price在《Little Science，Big Science》一书中指出，进入20世纪以后，随着通信技术的更新，学者间的交流成本不断降低，论文的合著比例不断升高，科学研究呈现由个人研究向团队合作转变的趋势，合著论文也逐步成为主流的科研方式[1]。1978年，Beaver等[2]指出，科研合作能有效地提升科研产出的质量。Adams等[3]也探讨了科研合作的重要性及其对科研产出的促进作用。邱均平等[4]使用相关性分析，论证了作者的合作程度与其科研成果的质量之间存在着显著的正相关关系。Fischbach等[5]指出，合著论文的被引频次远高于非合著论文，并且合著人数越多，其被引频次也越高。

从以上研究不难看出，科研合作对于推动科学技术的发展和提高学者学术影响力具有重要作用。随着大数据时代的到来，以文献信息为主的学术数据也呈现高速增长的态势[6]。不过海量的文献信息在为学者提供丰富科研资源的同时，也给诸多学者造成了“信息迷航”[7]的困扰。随着科学研究复杂性的增加和跨领域研究的需要，利用这些资源快速找到与自己兴趣相合的科研合作伙伴，日益成为提升学者科研效率和创新能力的关键，同时也成为图书情报领域一个越来越值得重点关注的课题。

1　国内外相关研究

目前，国内外关于科研合作学者推荐的研究主要关注基于学者研究内容相似性的推荐和基于学者间网络的推荐，以及融合内容与网络的综合推荐。

基于学者研究内容相似性的推荐主要运用自然语言处理模型对学者的文献信息进行处理，并计算作者之间的研究主题相似度，实现研究主题相似度

收稿日期：2024-09-27

较高的合作学者推荐。Rosen-Zvi等[8]在隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）模型的基础上提出融合作者信息的作者-主题模型（Author-Topic Model，ATM），从而完成合作学者的推荐。刘萍等[9]在对作者及其文献关键词形成的耦合网络进行学术社区划分的基础上，运用LDA模型构建作者的研究主题分布，从而计算作者研究领域相似度，完成合作学者推荐。蒲姗姗[10]同样使用ATM模型，考虑到学者研究主题集中度对推荐结果的影响，提出了一种基于知识互补的合作学者推荐模型。熊回香等[11]先使用词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）算法计算作者所发表文献的关键词的权重，再使用word2vec词向量模型进行向量化表示，通过比较向量的余弦相似度，得出作者之间研究领域的相似度。Lee等[12]使用医学主题词表（Medical Subject Headings，MeSH）提取论文关键词，并结合TF-IDF算法进行作者相似度的计算和推荐排序。李仲等[13]将稀疏分布式表征方法（Sparse Distributed Representations，SDR）用于合作学者推荐，也取得了较为理想的推荐结果。

基于学者间网络的推荐则主要依靠学者间产生的合著网络、共被引网络和共引网络等完成合作学者的推荐。张金柱等[14]在构建学者合著网络的基础上，利用LINE网络表示学习方法，得到每个作者的向量表示，并计算每个作者之间的相似度，从而完成合作学者的推荐。余传明等[15]借助基于深度学习的网络表示学习方法，对科研合作网络进行推荐模型构建。邓少伟等[16]则在合著网络的基础上，考虑到了合著作者的近邻关系作者，提出一种基于共同作者学术关系的学者推荐系统。朱云霞等[17]则在构建学者合著网络和共引网络的基础上，计算每个节点作者的节点度和中心性，并以此为学者推荐的依据。Yang等[18]改进随机游走算法，构建融合机构网络和合著网络的异构网络，经过对此网络的随机游走，得到学者的推荐结果。

融合推荐综合考虑作者研究内容的相似性以及他们在作者网络中的关系，利用node2vec等图嵌入模型进行推荐。Liu等[19]将共同主题、学者间合著关系、引文关系等嵌入异构网络以实现作者推荐。林原等[20]提取文献题录数据中的作者、关键词、机构数据，依据它们之间的关系构建了三层异构共现网络，并使用node2vec模型对此网络进行网络表示学习，将它们投射为低维向量，根据向量间的相似度实现推荐；但该研究将3种不同类型的节点放入一个网络中进行相似度计算，存在一定不合理之处。熊回香等[21]综合考虑文献节点的时间特征和语义特征来构建包含作者和文献两种节点的加权异质网络，从而实现合作推荐，但该研究同样存在在同一网络中计算不同类型节点相似度的问题。

上述相关研究在推动科研合作学者推荐研究进展的同时拓宽了研究的思路，但同时也存在一些不足，因此本文提出构建作者间全连通网络，将作者间的合作、引用、关键词相似度等信息转化为作者间连线的权重，使用node2vec图嵌入模型对该网络进行深度学习，得到每个作者的节点坐标向量，并通过计算作者间向量相似度达到合作学者推荐的目的。这样做可以避免使用作者合著关系网络进行模型训练时出现的部分核心作者或核心作者群与其他核心作者之间没有发生合著行为而导致的网络割裂现象，保证了作者网络的全连通性，也解决了以往推荐参考维度单一的问题。

2　模型构建

从作者信息的4个维度（关键词、合著、引用、机构）提取数据，构建词相似度矩阵、作者合著矩阵、作者引用矩阵、作者机构矩阵4个作者链接关系矩阵，并对该矩阵形成的作者网络进行node2vec图嵌入模型学习，生成作者节点向量并完成作者的相似度计算，最终实现合作学者的推荐。研究流程如图1所示。

2.1　多维度模型构建

（1）关键词维度。学者的研究成果主要体现在其学术论文当中，而论文的关键词则是对论文研究内容和方法的高度凝练。对于单个作者来说，其发表论文的关键词在很大程度上可以代表他的研究方向，而其中词频越高的关键词在作者研究方向分析中所占的权重也越大。利用TF-IDF算法计算作者所有发表论文的每个关键词的权重，并按照该权重对每位作者的全部关键词进行word2vec词向量计算，得到各作者的关键词相似度。

TF-IDF算法是一种常见的文本加权技术，用于评估词汇在文档中的重要性，将其应用于文献关键词权值计算的方法如式（1）~式（3）所示。

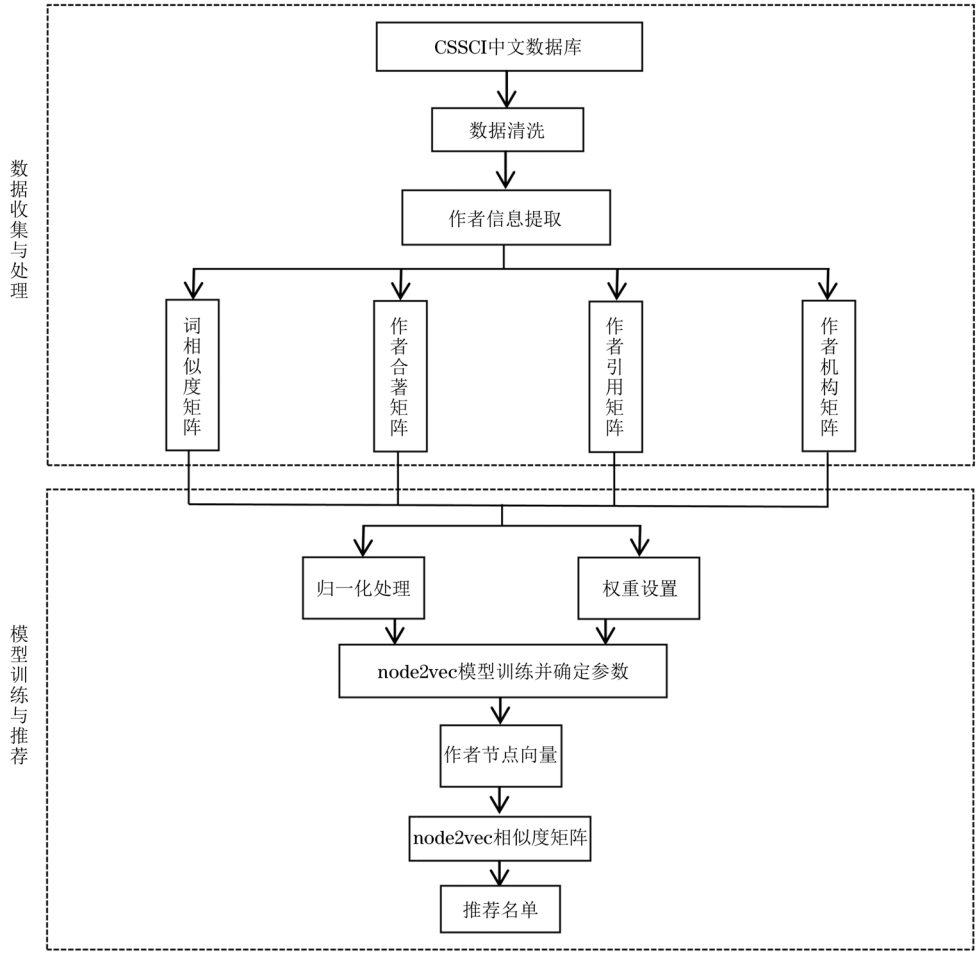

图1　研究流程

$$\mathrm{TF}(t,d)=\frac{f(t,d)}{\mathrm{TotalTerms}(d)} \tag{1}$$

$$\mathrm{IDF}(t,D)=\log\frac{N(D)}{\mathrm{DF}(t,D)} \tag{2}$$

$$w_{\text{TF-IDF}}=\mathrm{TF}(t,d)\times\mathrm{IDF}(t,D) \tag{3}$$

式中：$f(t,d)$是关键词t在关键词集合d中出现的次数；$\mathrm{TotalTerms}(d)$是关键词总数；$N(D)$是作者发表的论文总数；$\mathrm{DF}(t,D)$是作者发表论文中包含关键词t的论文数。TF-IDF值$w_{\text{TF-IDF}}$为每个关键词在作者研究方向中所占的权重。

最早的词向量表示方法是独热编码，其原理是将词表示为一个词汇表长度的二进制编码序列，每一个词有其独特的序号，在此编码逻辑下，计算机可以识别文本中的每个词。但独热编码存在两个问题：其一是无法体现词语之间的语义联系，其二是词汇表过长导致的维度灾难。2013年Mikolov等[22-23]提出的word2vec自然语言处理模型则有效地解决了这些问题。目前常用的word2vec模型主要基于跳字（Skip-Gram）模型和连续词袋（Continuous Bag of Words，CBOW）模型，这两种模型都使用了浅层的神经网络模型，通过最大化词汇之间的相似度来训练词向量，其中Skip-Gram通过目标词来预测上下文词，CBOW则刚好相反。已有许多研究表明word2vec模型能有效地识别词之间的语义关系，并高效地将词投影为低维向量[24-26]。本研究使用Skip-Gram模型对所得关键词进行word2vec训练，由此可以得到每个作者基于关键词向量的研究方向的向量表示，通过计算每两个作者之间词向量的余弦相似度，将作者间研究方向的相似度量化为数据，从而进行后续的研究。

（2）合著维度。合著论文是最常见、也是最容易统计的科研合作行为，Li等[27]指出，合著论文是科研合作的一种“显性产物”。在合作撰写论文时，每一位作者都将自己的知识贡献给了这篇文章，同时也获取了其他合著者的知识，在一定程度上共享了对方的研究领域。

在学者合著论文过程中所产生的信任关系有助于合作双方产生新的合作关系，同时，信任传递理论[28]指出，有共同合作者的学者间也更容易产生合作。因此，将合著次数作为作者间链接关系的指标之一纳入研究维度。本研究将提高合作过的作者之间的链接权重，从而提高模型在拥有共同合著者作者之间的游走概率，使与目标学者合作距离更短的作者更容易被推荐。

（3）引文维度。引用是指作者在文章中使用他人的研究成果或数据作为依据，以支持自己观点的行为，施引者与被引者之间存在较为显著的研究领域相似性。此外，虽然引用本身并不属于科研合作行为，但有研究指出引用与合著之间存在一定的促进关系，作者往往会与引用次数较多的学者存在较好的社交关系，从而促进合作[29]。因此，本研究将引用次数和被引次数作为作者间链接关系指标纳入研究维度。

（4）机构维度。机构是指作者所属的研究机构，一般为大学、研究院等。Evans等[30]在对英国商业与管理领域的合著网络进行社区检测后指出，学者在进行合作者挑选的时候，更倾向于选择同一机构的学者，合著论文中同机构作者合著现象也最为普遍。因此，本研究将作者机构信息纳入研究维度。

在获取上述4个维度的作者信息之后，将它们融入作者链接关系矩阵，该矩阵纳入全体目标学者，矩阵中的数值即为每两个作者间形成的4个维度的数据经过归一化处理后加权相加的值，该值的大小体现了作者间链接关系的密切程度。

2.2 基于node2vec模型的学者推荐

得到作者链接关系矩阵之后，使用node2vec模型训练该矩阵中的作者节点，反复训练并调整参数，最终生成每个作者的向量表示，并根据向量间的余弦相似度完成作者间相似度的计算。

（1）node2vec模型介绍。node2vec是一种图嵌入模型，最早由Grover等[31]在2016年提出。这种方法可以通过对模型内部参数的调整，让图嵌入结果在偏向同质性和偏向同构性当中进行权衡。同时，该模型不仅可以对节点进行设置，还可以对节点间边的权重进行设置，从而对带有丰富语义信息的网络进行学习训练。node2vec的基本任务是将网络图中每个节点映射到低维空间，形成多维（维度数一般为2的指数）空间向量表达。

图2为图嵌入模型的经典示意图，图中的节点间存在同构性和同质性。同质性是指节点高度互联且属于相似网络集群或社区，其嵌入结果应更为相似，即距离更近的两个节点的低维特征更相似，如图2中节点U和节点S_1；同构性是指网络中节点的局部结构相似，其嵌入结果应更为相似，如图2中的节点U和节点S_6。同质性强的两个节点具有相似的特征或属性，更可能产生链接，在社交网络中，体现为拥有相似兴趣、背景或观点的人更容易建立合作关系；同构性强的两个节点具有相似的网络结构位置，在社交网络中，体现为在不同团体中处于相似位置的两个人更容易产生联系。

不同的图嵌入模型在遍历图中节点时，采用的探索策略是不同的，常见的两种探索策略分别为广度优先探索（Breadth-First Search，BFS）和深度优先探索（Depth-First Search，DFS），如图2箭头所示。BFS更倾向于得到网络节点的同构性，因为BFS会更多地在当前节点的邻域中遍历，即对节点周围的网络结构进行一次扫描，从而判断当前节点是该局部网络的中心节点还是边缘节点抑或是连接性节点；而DFS更倾向于得到网络节点的同质性，因为尽管DFS可能通过多次跳转走到距离较远的节点，但从概率的角度来说，其更有可能在一个社区内部游走，从而使得生成的随机游走序列更容易表达社区内部节点的相似性。

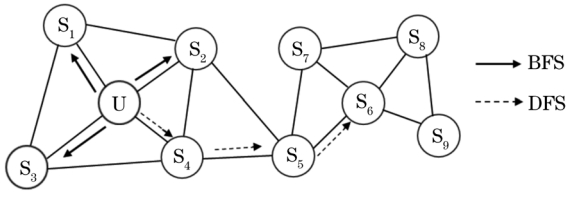

图2　图嵌入模型示意图

目前使用node2vec进行科研合作推荐的研究均将作者或作者和文献、机构等作为节点，若节点间存在联系则建立一条连线。这样的操作虽然较易实现，但可能导致节点凝簇成团，其原因是学者在选择合作对象和引用对象时，更倾向于选择同机构或合作过的学者，而节点团之间的联系非常稀疏，甚至存在断联现象，这违背了科研合作推荐的初衷。因此，本研究将作者节点进行全连接，即每两个作者之间均存在连线，并按照作者间词向量相似度及合著、引用、机构关系的强弱为这些连线赋予不同的权重。使用node2vec模型对作者链接矩阵进行训练，经过对不同参数下结果的分析，确定模型的参数，并生成节点的向量值。

（2）作者节点相似度计算和推荐。在得到作者节点的n维向量之后，计算每两个作者之间的向量相似度，以此为推荐指标。本研究采用余弦相似度进行作者节点向量的相似度计算，如式（4）所示。

$$\cos(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{i=1}^{n}(\boldsymbol{x}_i\times\boldsymbol{y}_i)}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{x}_i)^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{y}_i)^2}} \quad (4)$$

式中：$\boldsymbol{x}$、$\boldsymbol{y}$分别为计算相似度的两个作者的多维向量表示；$\boldsymbol{x}_i$、$\boldsymbol{y}_i$分别为向量$\boldsymbol{x}$、$\boldsymbol{y}$在第i维度上的向量；n为node2vec所设置的维度数量。由此计算出的$\cos(\boldsymbol{x},\boldsymbol{y})$的结果处于（–1，1）区间。该结果越接近1，代表两个作者相似度越高；反之则代表两个作者相似度越小。在完成相似度计算之后，针对每个作者，给出网络中与其相似度排名前10的作者，形成科研合作学者推荐名单，并分析结果的有效性。

3　实证研究

3.1　数据源的选择和收集

研究需要用到论文的篇名信息、作者机构信息、关键词信息和参考文献信息，因此，选择CSSCI数据库中分类为“图书馆、情报与文献学”且影响因子大于2的19种期刊的2019—2023年的文献题录信息作为源数据，尽可能保证使用的数据涵盖图书情报领域的活跃作者和研究热点。使用CSSCI数据库官方下载渠道，得到初始数据15 837条。初始数据示例如表1所示。

表1　提取整理得到的初始数据

序号	篇名	作者及机构	关键词	参考文献
1	新时代图书馆发展的……	[陈传夫]武汉大学/[陈一]武汉大学	新时代/图书馆发展/中国实践	略
2	中国特色公共图书馆事业……	[苗美娟]对外经济贸易大学/[陆晓曦]北京语言大学/[张皓珏]北京工商大学	公共图书馆/中国特色/公共文化服务体系/……	略
3	公共图书馆在中国减贫事业中……	[肖鹏]中山大学/[赵庆香]武汉大学	公共图书馆/减贫/可持续发展目标/公共图书馆宣言	略
4	古籍传承性保护再认识	[姚伯岳]天津师范大学/[周余姣]天津师范大学/[王鹜嘉]天津师范大学	古籍保护/传承性保护/非物质文化遗产/学科建设	略
15 837	突发事件网络舆情传播……	[苏妍嫄]燕山大学/[张金纯]燕山大学/[张亚明]燕山大学	突发事件/网络舆情/观点演化/舆情传播/……	略

3.2　数据清洗和核心作者群的确定

对这15 837条题录数据中的无效数据进行识别删除，对作者及机构词条当中的作者进行分割，并依据作者分割结果得到单一作者词条的题录数据39 895条。此外，为了区分不同署名次序作者的贡献，在每条题录中加入一个新的数据，即“作者署名次序”，按出现顺序依次标记为1、2、3等。

绑定作者和机构以实现同名作者的消歧。在计算作者发文量时，使用邱均平等[4]研究中的相对发文量作为合著作者权重的计算依据，统计所有作者相对发文量，各合著作者权重的计算标准如表2所示。

实际数据中存在极少数论文的作者数量超过6名，考虑到其权重过小（小于0.05），对第六作者之后

表2　合著作者权重

作者数量/名	权重					
	第一作者	第二作者	第三作者	第四作者	第五作者	第六作者
1	1.00					
2	0.60	0.40				
3	0.60	0.25	0.15			
4	0.60	0.20	0.10	0.10		
5	0.60	0.10	0.10	0.10	0.10	
6	0.60	0.10	0.10	0.10	0.05	0.05

的作者不予计算。基于同名消歧的结果以及合著作者权重表，对所有作者的相对发文量进行统计，发现发文量最高的作者相对发文量为40.10篇。使用普莱斯定律[1]来确定核心作者群，普莱斯定律公式如式（5）所示。

$$M = 0.749\sqrt{m_{\max}} \quad (5)$$

式中：$m_{\max}$为最高的作者发文量。经计算$M \approx 4.743$，取其最近整数，即发文量大于等于5篇的作者为核心作者，得到核心作者390名，并以此为后续任务的基础。

3.3 作者关键词筛选与word2vec相似度计算

对390名核心作者全部文献中的关键词按照频次进行统计，并使用式（1）~式（3）计算每个作者名下关键词的权重，依据这个权重对每位作者的关键词进行word2vec词向量计算。word2vec模型参数设置如表3所示。

表3 word2vec参数设置

参数名	设置值	解释
Vector_size	100	向量维度
Windows	5	取词窗口大小
Workers	4	同时运算线程数

在得到每位作者的词向量后，使用式（4）计算每两个作者之间的关键词相似度，并生成矩阵。由于本研究综合考虑4个维度的信息，在计算得到关键词相似度后，对其进行归一化处理，从而方便4个维度的融合。归一化的实现借助Python中的常规归一化函数df_normalized=（df−min_val）/（max_val−min_val），其中min_val为最小值，max_val为最大值。经此函数转换后，最大值归一化为1，最小值归一化为0。归一化后的关键词相似度矩阵如表4所示。

表4 归一化后的相似度矩阵（部分）

作者	陈传夫	陈一	肖鹏	姚伯岳	周余姣	张靖	余厚强
陈传夫	0	0.388 74	0.390 96	0.383 29	0.286 93	0.277 42	0.196 21
陈一	0.388 74	0	0.260 09	0.205 40	0.230 47	0.254 22	0.286 25
肖鹏	0.390 96	0.260 09	0	0.287 97	0.150 68	0.440 15	0.423 63
姚伯岳	0.383 29	0.205 40	0.287 97	0	0.360 56	0.181 73	0.276 22
周余姣	0.286 93	0.230 47	0.150 68	0.360 56	0	0.281 29	0.143 66
张靖	0.277 42	0.254 22	0.440 15	0.181 73	0.281 29	0	0.309 34
余厚强	0.196 21	0.296 25	0.423 63	0.276 22	0.143 66	0.309 34	0

3.4 合著矩阵、引用矩阵及机构矩阵的构建

构建一个390×390的作者合著矩阵，矩阵中每个元素代表对应两位作者的合著次数，并将此矩阵存入Excel表格，如表5所示。

表5 作者合著矩阵（部分）

作者	陈传夫	陈一	肖鹏	姚伯岳	周余姣	张靖	余厚强
陈传夫	0	3	0	0	0	0	0
陈一	3	0	0	0	0	0	0
肖鹏	0	0	0	0	0	3	0
姚伯岳	0	0	0	0	2	0	0
周余姣	0	0	0	2	0	0	0
张靖	0	0	3	0	0	0	1
余厚强	0	0	0	0	0	1	0

引用矩阵反映的是作者间的引用关系。要创建引用矩阵，首先需要对基础数据中的参考文献字段进行数据处理，删掉不属于390位核心作者的参考文献，然后将参考文献进行分割，得到将每位作者及其引用作者一一对应的列表。依照此引用对应表，构建一个390×390的引用矩阵，矩阵中每个元素代表两者互相引用的次数，将该矩阵存入Excel，如表6所示。

同样利用基础数据生成作者机构矩阵，矩阵中1代表两人同属一个机构，0代表不属于一个机构，其构建原理与上类似，不予赘述。另外，在处理数据时发现有几位作者在研究覆盖的5年间存在机构变换情况，因此，将此类作者其他维度的信息进行融合，但保留了不同的机构属性，并在机构矩阵中体现为某一作者与现机构和原机构作者的机构相似度均为1。

表6　作者引用矩阵（部分）

作者	陈传夫	陈一	肖鹏	姚伯岳	周余姣	张靖	余厚强
陈传夫	0	12	1	0	0	2	0
陈一	12	0	1	0	0	0	0
肖鹏	1	1	0	0	0	4	0
姚伯岳	0	0	0	0	3	6	0
周余姣	0	0	0	3	0	4	0
张靖	2	0	4	6	4	0	0
余厚强	0	0	0	0	0	0	0

对这3个矩阵中的值也进行归一化处理，使矩阵中最大值为1，最小值为0。

3.5　node2vec模型训练与参数设置

在开始node2vec模型训练之前，需要对以上4个矩阵进行权重设置，即设置每个维度对模型的影响力大小。董文慧等[32]在经过层次分析和专家调查法研究后，得到专家在挑选合作学者时不同影响因素的权重值。结合其研究成果与本研究预设目标，最终选定关键词维度、合著维度、引用维度及机构维度的权重分别为0.5、0.2、0.2和0.1。这组权重设置可以根据使用者的不同需求进行调整，本研究采用此权重设置是为了更加凸显作者间研究领域相似度对最终推荐名单的影响，以期在兼顾既往合作关系、引用关系和机构属性的基础上，为学者推荐更多尚未产生合作却有着相似研究领域的学者。

构建以作者为节点的全连通网络，节点间连线的权重即为相似度矩阵中的值。使用node2vec模型对该网络进行表示学习，得到作者向量以及余弦相似度。在node2vec模型当中，dimensions是指节点投射的向量维度数，更高的维度数可以更细致地反映节点信息，但计算压力会相应增大，在实际训练中一般会选择32、64、128等2的指数作为维度数，本研究折中选择64维。walk_length是每次随机游走的步长，经过多次比对模型结果发现：当游走步长小于400时，由于没有遍历到多数节点，推荐结果随机性过强；而当游走步长大于1 600时，由于遍历到过多节点，节点间同质性过强，最终的推荐名单同样存在随机性强的问题，因此折中将游走步长定为1 000。p值是返回参数，q值是进入参数，二者数值一般互为倒数，此处将p值设置为2，q值设置为0.5，目的是使模型的游走更倾向于DFS。其余参数均使用node2vec默认值，其具体数值设置如表7所示。

表7　node2vec模型参数设置

参数名	设置值	解释
dimensions	64	嵌入向量的维度
walk_length	1 000	每次随机游走的步长
num_walks	30	每个节点启动的随机游走数
p	2	返回参数
q	0.5	进入参数
workers	4	并行的线程数
windows	10	窗口大小
batch_words	4	每个批次处理的节点数

3.6　推荐结果分析与模型评价

在上述权重和参数设置下，生成了node2vec相似度矩阵，选取与每位作者相似度排名前10的作者形成最终推荐名单。为了评估推荐名单的价值，建立综合研究领域相似度、合著维度、引用维度和机构维度的评分标准：①研究领域相似度维度，将推荐名单中包含的word2vec相似度排名前10的作者的人数作为评价指标，即每位作者最终推荐名单中每存在一个word2vec相似度排名前10的作者，就为其加10分；②合著维度，依据作者间合作关系形成的合著网络，进行最短合作距离的计算，合作距离越短代表两个作者之间产生合作的概率越高，对每位目标学者的推荐名单的合作距离取倒数，分值为倒数与10的乘积；③引用维度，每位作者的推荐名单中每存在一次引用关系，就为其加1分；④机构维度，每位作者的推荐名单中每存在一名机构位于同省域的作者，就为其加2分。

经过对最终推荐名单的计算得到评分结果，如表8所示。分值越高表示推荐名单的价值越高，也代表目标学者与其推荐名单中的学者在未来产生合作的可能性越高。

随机选择学者A和学者B的推荐名单为例进行模型的分析与评价，如表9~表10所示，其中作者关注领域源自中国知网。出于隐私考虑，隐去作者真实姓名。

为了佐证推荐名单的价值，将目标学者与每位作者在中国知网中的关注领域词条进行对比，并进行具体分析。

在表9所示的10名作者中，有两名作者与目标学者

表8　最终推荐名单评分（部分）

目标学者	word2vec相似度排名前10作者数/名	合作距离倒数和	互引次数	机构位于同省域作者数/名	分值
陈传夫	2	1.29	21	8	69.9
陈一	1	2.35	16	8	65.5
肖鹏	4	0.79	4	2	55.9
姚伯岳	2	0	0	1	22.0
陆阳	2	0	0	1	24.0

表9　学者A的推荐名单

作者	所属机构	word2vec相似度	互引次数	node2vec相似度	作者关注领域
Author1	黑龙江大学	0.605 88	46	0.739 12	数据安全、知识图谱、数据治理、政策文本分析等
Author2	中山大学	0.637 06	0	0.737 81	知识服务能力评价、城市治理、标准情报服务等
Author3	河南师范大学	0.451 22	1	0.702 43	高校图书馆、知识产权、专利分析、图书馆学等
Author4	黑龙江大学	0.647 75	23	0.702 37	公共数据、总体国家安全观、政府开放数据等
Author5	中南财经政法大学	0.432 60	12	0.700 21	开放数据、政府数据、隐私保护、美国、高等教育等
Author6	南开大学	0.402 76	0	0.700 12	公共图书馆、情报学、知识管理、学科建设等
Author7	江苏大学	0.488 35	1	0.698 57	高校图书馆、战略规划、数字中国、战略规划等
Author8	武汉大学	0.479 44	5	0.694 40	电子政务、政府数据治理、开放数据、数据治理等
Author9	中国人民警察大学	0.476 72	0	0.691 66	知识图谱、网络舆情、文献计量、情报感知等
Author10	湖北工业大学	0.436 37	0	0.677 55	知识图谱、文献计量、知识管理、数字图书馆等

表10　学者B的推荐名单

作者	所属机构	word2vec相似度	互引次数	node2vec相似度	作者关注领域
Author11	南京财经大学	0.456 19	0	0.760 87	个性化、学科发展、用户模式、学术虚拟社区等
Author12	南京理工大学	0.534 77	0	0.751 07	数字人文、精神障碍、知识本体、共生模式等
Author13	中山大学	0.446 51	2	0.748 88	影响因素、用户需求、大众创新、用户体验等
Author14	吉林大学	0.579 31	8	0.748 62	影响因素、扎根理论、社交网络、用户、信息传播等
Author15	南京大学	0.452 85	1	0.719 50	在线健康社区、影响因素、用户、健康信息态度等
Author16	武汉科技大学	0.403 27	1	0.715 90	扎根理论、影响因素、在线知识共享网络等
Author17	南开大学	0.549 70	20	0.714 77	健康信息搜索、信息需求、信息行为研究等
Author18	南昌大学	0.423 89	2	0.713 46	扎根理论、用户需求、用户关系、服务模式等
Author19	华中师范大学	0.589 43	0	0.711 21	影响因素、众包、知识共享、社交网络、机制等
Author20	中国科学院文献情报中心	0.417 38	1	0.705 45	开放获取、数据共享、用户行为、数据管理服务等

属于同一机构，分别是Author1与Author4，他们与目标学者存在多次合作关系，而在整个390人名单中，还有多位与目标学者存在合作关系且属于同一机构的作者，但由于合作次数较少，未能在推荐名单中体现出来；另外8名作者虽未与目标学者产生合作关系，但其中4名与目标学者存在引用关系。目标学者A在中国知网中的关注领域词条内容为数据安全、开放数据、数字图书馆、信息公开、知识产权、知识管理、图书馆、公共数据、开放政府数据、信息政策、美国、高校、文献计量、情报学、数字信息资源、知识图谱、总体国家安全观、政策协同、著作权。经比对，发现推荐名单中的10名作者都在不同程度上与目标学者存在词条的耦合现象，其中耦合频率较高的内容为政府数据、政策、知识产权、知识图谱、文献计量、公共数据、高校图书馆以及知识管理。

在表10所示的10名作者中，没有与目标学者产生过合作的作者，因此进一步研究这份名单，发现其中6名作者与目标学者的合作距离较近。其中Author11、Author12、Author14、Author15、Author16、Author20与目标学者的合作距离分别为2、5、2、5、7、4。信任传递原理指出，合作距离短的学者达成合作的可能性更高。表10所示的推荐结果在word2vec相似度上的表现也很好，均远高于平均值0.298 00。目标学者B在中国知网中的关注领域词条为数字人文、公众科学、社会化媒体、影响因素、扎根理论、健康信息实践、用户生成内容、用户体验、知识付费、科研众筹、健康信

息学、感知示能性、移动视觉搜索、众包、社会化搜索、活动理论、在线健康信息搜索、社交网站、动因研究、关联数据。经比对，推荐作者的关注领域与目标学者也存在较大程度的耦合，其中耦合频率较高的内容为用户、扎根理论、健康信息搜索、健康信息、数字人文、影响因素、社交网络、众包。

经过对随机抽取的两名学者的推荐名单分析，认为构建的模型可以有效地发现领域内研究方向相似度高、活跃度较高且较易达成合作的学者。此外，本文只展示了模型中4个维度一组权重设置下的推荐结果，而模型提供权重设置窗口，使用者可以通过对权重的调整来实现其他需求下的合作学者推荐。

4　结论与不足

本研究从科研合作学者推荐这一课题出发，结合当下热度较高的图嵌入模型，为学者推荐与图嵌入模型的结合提供了全新思路。根据关键词相似度、合著次数、引用次数以及所属机构4个维度构建了基于node2vec图嵌入模型的合作学者推荐模型，该模型创新性在于将所有作者节点进行全连通，以作者4个维度相似度的加权之和为节点间连线权重。经实证研究分析，证明了该模型可以高效并准确地完成科研合作学者的推荐任务，推荐的结果在保证较高关键词耦合度的同时，兼顾了合作距离与互引指标，可以为每一位寻找合作学者的用户推荐具有实际利用价值的推荐名单。值得一提的是，本研究虽以图书情报领域数据为例进行实证分析，但该模型并不拘泥于图书情报领域，可以推广至其他领域。另外，本模型可以进行权重调整，从而满足实际使用者对合作学者发现的不同需求。

本研究也存在一些不足之处。①纳入的数据量不足，对作者间合作关系的体现不够完善，在引用关系上统一计算引用与被引次数，而两作者间只存在单向引用关系的情况普遍存在。②在机构维度，未考虑地理因素、机构合作历史等。③缺少同一数据下所提模型与传统推荐模型的推荐结果对比。同时，受限于篇幅，只提供了一组维度权重，只体现了更关注研究领域相似度的推荐结果。在模型效果评价时，由于本研究选取了2019—2023年的文章作为原始数据，在体现领域内最新研究方向的同时，无法使用之后一段时间内的新文章数据来验证模型预测有效性，也无法基于未来作者间实际合作情况评估该模型的有效性。相信在未来的研究中，随着其他学者研究的深入，可以突破这些桎梏，实现更为精准的合作学者推荐。

参考文献

[1] DE SOLLA PRICE D J. Little science, big science[M]. New York: Columbia University Press, 1963.

[2] BEAVER D D, ROSEN R. Studies in scientific collaboration: part I. the professional origins of scientific coauthorship[J]. Scientometrics, 1978, 1(1): 65-84.

[3] ADAMS J D, BLACK G C, CLEMMONS J R, et al. Scientific teams and institutional collaborations: evidence from U.S. universities, 1981–1999[J]. Research Policy, 2005, 34(3): 259-285.

[4] 邱均平，温芳芳. 作者合作程度与科研产出的相关性分析：基于“图书情报档案学”高产作者的计量分析[J]. 科技进步与对策，2011，28(5)：1-5.

[5] FISCHBACH K, PUTZKE J, SCHODER D. Co-authorship networks in electronic markets research[J]. Electronic Markets, 2011, 21(1): 19-40.

[6] XIA F, WANG W, BEKELE T M. Big scholarly data: a survey[J]. IEEE Transactions on Big Data, 2017, 3(1): 18-35.

[7] 肖丽平，娄策群. 互联网发展环境下“信息超限”问题研究[J]. 图书馆学研究，2018(10)：16-21，29.

[8] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. New York: ACM Press, 2004: 487-494.

[9] 刘萍，郑凯伦，邹德安. 基于LDA模型的科研合作推荐研究[J]. 情报理论与实践，2015，38(9)：79-85.

[10] 蒲姗姗. 基于知识互补的科研合作专家推荐模型研究[J]. 情报理论与实践，2018，41(8)：96-101.

[11] 熊回香，李晓敏，杜瑾. 基于学术关键词与共被引的学者推荐研究[J]. 情报学报，2021，40(7)：725-733.

[12] LEE D H, BRUSILOVSKY P, SCHLEYER T. Recommending collaborators using social features and MeSH terms[J]. Proceedings of the American Society for Information Science and Technology, 2011, 48(1): 1-10.

[13] 李仲，韩红旗，吴广印，等. 基于文本稀疏分布式表征的学术合作推荐[J]. 情报科学，2019，37(6)：113-118.

[14] 张金柱，于文倩，刘菁婕，等. 基于网络表示学习的科研合作预

测研究[J]. 情报学报，2018，37（2）：132-139.

[15] 余传明，林奥琛，钟韵辞，等. 基于网络表示学习的科研合作推荐研究[J]. 情报学报，2019，38（5）：500-511.

[16] 邓少伟，罗泽，李树仁，等. 基于论文共同作者学术关系的学者推荐系统[J]. 计算机工程，2013，39（2）：12-17.

[17] 朱云霞，魏建香. 基于共引网络的学者间潜在合作机会发现[J]. 西南民族大学学报（人文社会科学版），2016，37（9）：235-240.

[18] YANG C，LIU T T，CHEN X H，et al. HNRWalker：recommending academic collaborators with dynamic transition probabilities in heterogeneous networks[J]. Scientometrics，2020，123（1）：429-449.

[19] LIU X Y，WU K，LIU B，et al. HNERec：scientific collaborator recommendation model based on heterogeneous network embedding[J]. Information Processing & Management，2023，60（2）：103253.

[20] 林原，王凯巧，刘海峰，等. 网络表示学习在学者科研合作预测中的应用研究[J]. 情报学报，2020，39（4）：367-373.

[21] 熊回香，唐明月，叶佳鑫，等. 融合加权异质网络与网络表示学习的学术信息推荐研究[J]. 现代情报，2023，43（5）：23-34.

[22] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[EB/OL]. [2024-06-16]. https://arxiv.org/abs/1301.3781.

[23] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2024-06-16]. https://arxiv.org/abs/1310.4546.

[24] CHURCH K W. Emerging trends Word2Vec[J]. Natural Language Engineering，2017，23（1）：155-162.

[25] WANG Y C，ZHU L G. Research on improved text classification method based on combined weighted model[J]. Concurrency and Computation：Practice and Experience，2020，32（6）：e5140.

[26] 周练. Word2vec的工作原理及应用探究[J]. 科技情报开发与经济，2015，25（2）：145-148.

[27] LI E Y，LIAO C H，YEN H R. Co-authorship networks and research impact：a social capital perspective[J]. Research Policy，2013，42（9）：1515-1530.

[28] BELANCHE D，CASALÓ L V，FLAVIÁN C，et al. Trust transfer in the continued usage of public e-services[J]. Information & Management，2014，51（6）：627-640.

[29] 邱均平，马凤. 微观层次的引用与合作之间关系的研究：以三位中国图书情报学核心作者为例[J]. 情报理论与实践，2011，34（10）：43-47.

[30] EVANS T S，LAMBIOTTE R，PANZARASA P. Community structure and patterns of scientific collaboration in business and management[J]. Scientometrics，2011，89（1）：381-396.

[31] GROVER A，LESKOVEC J. node2vec：scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2016：855-864.

[32] 董文慧，熊回香，杜瑾，等. 基于学者画像的科研合作者推荐研究[J]. 数据分析与知识发现，2022，6（10）：20-34.

作者简介

袁永旭，男，硕士，研究馆员，研究方向：信息计量与科学评价、图书馆管理与服务，E-mail：yuanzi-4696@163.com。
成韬，男，硕士研究生，研究方向：信息计量与科学评价。

Multi-Dimensional Scholars Recommendation in Scientific Research Cooperation Based on node2vec Model

YUAN YongXu[1,2] CHENG Tao[1]
(1. School of Management, Shanxi Medical University, Taiyuan 030001, P. R. China;
2. Library of Shanxi Medical University, Taiyuan 030001, P. R. China)

Abstract: This article proposes a multi-dimensional scholars recommendation method based on graph embedding model to recognize scholars with a high degree of relevance to the research topic and high potential for collaboration. First, the CSSCI database is used as the data source to construct a fully connected network among authors. Keyword similarity, cooperative relationship, citation relationship, and institutional attributes among authors are integrated into weights for inter author connections. Second, the node2vec graph embedding model is used to perform deep learning on the network and obtain the the node coordinate vectors of each author. Finally, comprehensive vector similarity between authors is obtained, so as to complete the recommendation. The results show that the multi-dimensional graph embedding model proposed in this paper can effectively complete the recommendation of cooperative scholars, and can provide a valuable recommendation list for scholars, so as to promote academic cooperation and communication.

Keywords: Science Collaboration; Scholars Recommendation; Graph Embedding; node2vec; word2vec

（责任编辑：王玮）