

基于BERT-LDA的国外LIS领域学科交叉研究 演化分析与前沿主题识别

袁毅 孟盈

(华东师范大学经济与管理学院, 上海 200062)

摘要: 识别学科交叉研究的前沿主题, 并对演化趋势进行分析, 有助于揭示学科交叉融合的方向, 为未来创新性、突破性研究提供参考。首先, 基于引文视角构建测度论文学科交叉性的指标, 识别具有学科交叉性的研究论文; 其次, 通过BERT-LDA模型识别研究主题, 利用余弦相似度计算主题之间的相似度, 构建主题演化路径; 最后, 基于新颖度、增长性、关注度、影响力构建前沿主题识别指标体系, 识别具有前沿性的学科交叉研究主题。以图书情报学 (Library and Information Science, LIS) 为例展开研究, 研究结果显示, 2004—2023年该学科领域的交叉研究主题呈现出逐渐细化和深入的特点, 主要集中在信息挖掘与知识发现、互联网信息行为、医疗信息学3个方面; 现阶段学科交叉研究前沿主题为医疗数据模型、舆情治理与情感分析、机器学习与深度学习; 基于信息技术的研究方法和其在不同领域的应用研究具有良好的应用前景, 有可能成为未来LIS领域的核心研究主题。

关键词: 研究前沿; 主题演化; 学科交叉; BERT-LDA; 主题识别; 图书情报学; 信息科学; 图书馆学

中图分类号: G353.1 **DOI:** 10.3772/j.issn.1673-2286.2024.09.001

引文格式: 袁毅, 孟盈. 基于BERT-LDA的国外LIS领域学科交叉研究演化分析与前沿主题识别[J]. 数字图书馆论坛, 2024, 20(9): 1-15.

学科交叉的概念可以追溯到20世纪初, 目前学界普遍接受的是2004年美国国家科学院提出的定义, 它强调将不同学科领域的知识、方法和理论相结合, 以解决那些超出单一学科范围或研究实践领域的问题, 推动科学进步^[1]。复旦大学校长、中国科学院院士金力^[2]在报告中提到, 百余年来, 诺贝尔奖获奖者有41.02%属于交叉学科, 尤其是21世纪以来, 跨学科成果占半数以上。学科交叉已经成为科学研究范式变革的一个重要基础, 不仅成为领域颠覆性技术的重要源头, 还形成了诸多前沿研究热点。主题识别和演化分析有助于从微观角度识别和挖掘这些研究热点, 帮助研究人员了解各领域研究的发展全貌和演变趋势, 并通过研究前沿识别预测未来研究的发展走向。

图书情报学 (Library and Information Science, LIS) 作为一门学科形成于20世纪70年代, 由传统图书馆学 (Library Science) 和源自文献学 (Documentation) 的信息科学 (Information Science) 融合而成^[3]。LIS与我国的图书情报学科有较程度的重合。2022年9月, “图书情报与档案管理”一级学科正式更名为“信息资源管理”。信息资源管理学科的发展本身就是不同学科交叉融合的结果, 学科的交叉融合为信息资源管理学科的研究带来了更先进的理论、方法与技术支持^[4]。识别国外LIS领域学科交叉研究的前沿主题并分析研究主题的演化趋势, 对于我国信息资源管理学科的跨学科发展具有一定启示意义。因此, 本文提出以下研究问题。

(1) 国外LIS领域的研究论文与哪些学科产生交叉? 交叉融合的发展趋势有什么特征?

(2) 国外LIS领域的学科交叉研究主题有哪些? 呈现出怎样的演化趋势?

(3) 国外LIS领域的学科交叉研究前沿是什么? 如何理解前沿主题的融合模式以提供学科发展的思路 and 方向?

为了解决以上问题, 本研究选取2004—2023年国外LIS领域的学科交叉研究为数据来源, 分析其与各学科交叉程度的变化趋势; 同时基于BERT-LDA方法识别学科交叉研究主题, 提供更加丰富的语义表示, 量化分析研究主题的演化特征, 并构建新颖度、增长性、关注度、影响力4个指标来识别研究前沿, 以期帮助研究人员了解领域内的研究发展趋势和脉络, 促进跨学科的交流合作。

1 相关研究

1.1 LIS领域学科交叉研究

学术界围绕LIS领域的学科交叉开展了大量研究。马岩等^[5]通过对中外LIS领域期刊论文研究方法的研究, 发现研究方法的跨学科性逐渐明显, 为图书馆学和情报学研究提供了更为多元和丰富的研究路径; Chang等^[6]采用直接引用、数目耦合、共同作者分析3种文献计量方法识别1978—2007年LIS领域研究的学科交叉演化情况, 该项研究表明LIS领域跨学科性正在增强; Figuerola等^[7]使用LDA主题建模方法分析1978—2014年LIS领域的研究进展, 揭示出19个重要研究主题, 并将其归纳为过程、技术、图书馆和特定领域的信息应用4个主要范畴, 同时发现LIS领域越来越呈现出多学科特征, 与特定领域(商业、医疗、法律、教育)相结合的研究论文数量呈现显著增长的趋势, 使LIS领域比原先更宽泛也更深入; Huang等^[8]使用直接引用法和合作者分析法比较了1978—2007年信息科学的跨学科演变过程, 发现信息科学学者引用最多的是LIS领域出版物, 其共同作者主要也来自LIS领域; 宋艳辉等^[9]使用关键词分析北美LIS领域和中国LIS领域博士学位论文的学科交叉特征, 研究表明北美LIS领域与信息管理相关的主题频次呈现上升态势, 而与传统图书馆学相关的主题频次增长放缓。

通过梳理文献发现, 国内外学者关于LIS领域学科交叉的研究成果较为丰富, 但仍存在一定的局限性。

在研究方法上, 现有关于研究主题识别的研究多停留在文献外部特征(如关键词等)层面, 较少通过机器学习和深度学习等算法考虑语义关联关系, 进而深入抽取文献潜在的核心内容; 在研究对象上, 由于国内外的LIS领域并不是完全等同的关系, 做对比研究时会存在一定偏差; 在研究内容上, 虽然学科交叉主题识别的研究成果较为丰富, 但少有研究涉及研究主题的演化分析及研究前沿的识别。

1.2 主题识别方法

当前主题识别的方法主要有关键词识别法、引文分析法、文本挖掘法。在基于关键词的研究主题识别中, 研究人员通常通过词频分析^[10]、关键词共现网络^[11]、关键词聚类^[12]等方法, 从关键词之间的关联关系中发现研究主题, 挖掘深层次的知识价值。引文分析法通过分析学术文献中的引用关系来揭示文献之间的交互作用和影响, 根据引用关系类型的不同可分为共被引分析^[13]、文献耦合分析^[14]和直接引用分析^[15]3种。美国科学情报研究所推出的ESI(Essential Science Indicators)数据库使用共被引分析法, 成功将全球科技文献划分为22个专业领域进行前沿分析, 被多国广泛用于跟踪学科发展的动态趋势^[16]。我国公开发布的研究前沿报告, 如中国工程院的《全球工程前沿》、中国科学院的《研究前沿》、中国农业科学院的《全球农业研究热点前沿》等, 同样以共被引分析为核心方法来识别文献主题。在大数据和人工智能技术的加持下, 文本挖掘法作为数据挖掘的一个重要分支, 逐渐在信息检索、自然语言处理等领域取得显著进展。相较于传统的关键词识别法, 文本挖掘法不仅能够捕捉和处理丰富的语义信息, 理解词汇之间的关联、语境及深层含义, 也弥补了引文分析法时间滞后性的不足, 为研究主题的识别提供了新的思路和方法。文本挖掘法中主题概率模型最具代表性, 如: 范云满等^[17]基于LDA主题模型探测文档集中的新兴主题; 冯佳^[18]将LDA主题模型与主题强度、新颖度指标相结合, 为大数据环境下的研究前沿的识别提供了方法借鉴。虽然LDA模型已在主题建模领域被学者们广泛应用, 但也存在一些劣势, 即LDA模型假设文档是由词袋表示的, 忽略了单词之间的顺序和上下文信息, 无法捕捉到单词之间的语义

关联。已有研究证明BERT模型在捕捉上下文语义关联上的能力^[19]，将BERT模型的词向量作为输入，可以为LDA模型提供更丰富的语义表示，提高主题建模的准确性^[20]。

本研究拟通过BERT-LDA方法识别国外LIS领域的学科交叉研究主题，提高主题识别结果的准确性。

2 研究设计

本研究提出一个学科交叉研究前沿主题识别框架（见图1），核心内容包括如下5个部分：学科交叉论文界定、数据预处理、学科交叉研究主题识别、学科交叉研究主题演化分析、学科交叉研究前沿主题识别。

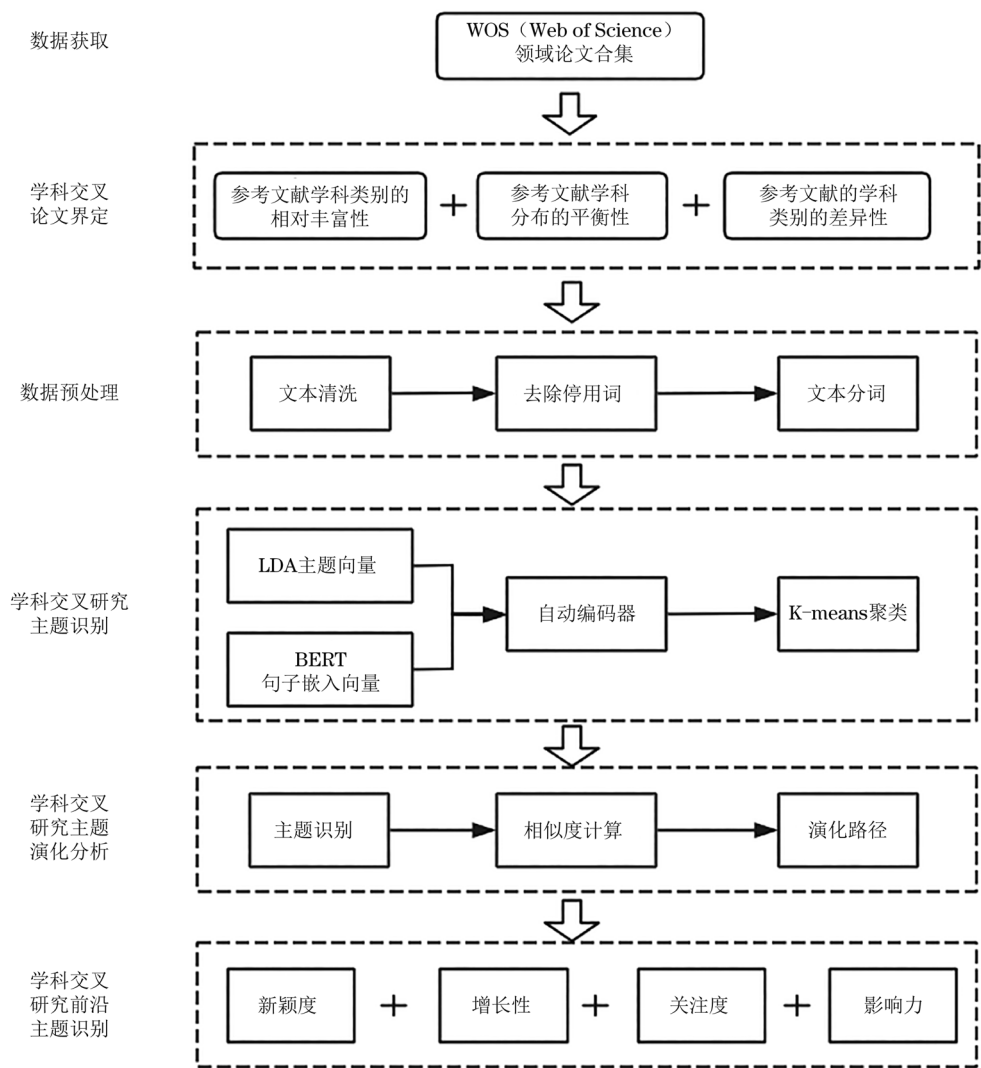


图1 学科交叉研究前沿主题识别框架

2.1 学科交叉论文界定

引用与被引用关系是学科间知识流动最直接的体现，国内外学者常利用引文分析法对不同领域和学科间的知识流动情况、规律或趋势展开研究^[21]。因此，选择从引文视角识别学科交叉论文。Stirling提出学科交叉多样性测量的3个维度，即丰富性（Variety）、平衡性（Balance）和差异性

（Disparity），被认为是比较成熟的测度指标^[22]。其中：丰富性用于衡量论文引文涵盖的学科广度，学科数量越多，论文本身的学科交叉性越强；平衡性用于衡量各学科的贡献均衡度，使用香农均匀度来表示；差异性用来计算论文引文中不同学科之间的差异程度，通常使用学科之间的相似系数来计算，如果学科间的相似系数较小（如商业和农学），那么认为学科之间的差异性较大，学科交叉程度更高。Leydesdorff

等^[23]在此基础上设计了一个新指标DIV来测量学科交叉多样性。相对丰富性、平衡性、差异性和DIV的计算公式如式(1)~式(4)所示。

$$I_{\text{Re_Variety}} = n/N \quad (1)$$

$$I_{\text{Balance}} = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2 \bar{x}} \quad (2)$$

$$I_{\text{Disparity}} = \frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}}{n(n-1)} \quad (3)$$

$$I_{\text{DIV}} = I_{\text{Re_Variety}} \cdot (1 - I_{\text{Balance}}) \cdot I_{\text{Disparity}} \quad (4)$$

式中: n 表示论文参考文献的学科类别总数; N 表示学科类别总数; x_i 表示学科 i 的参考文献数量; $\bar{x}$ 表示参考文献平均数量; d_{ij} 表示学科 i 和学科 j 之间的距离。

2.2 学科交叉研究主题识别

2.2.1 BERT

BERT是一种基于Transformer模型的预训练语言表示方法,相比传统的自然语言处理模型,BERT采用双向Transformer编码器架构,在处理文本时能够更全面地捕捉输入序列的全局依赖关系,并综合考虑上下文信息。在BERT模型中,对于输入的文本,要将其分词后映射到对应的嵌入向量,使用多头自注意力机制对经过Transformer编码器处理的向量进行残差连接和归一化,经过前馈神经网络和残差网络处理后得到BERT语义特征向量,如式(5)所示。

$$d = w_{ij}(\omega + \delta + \rho) \quad (5)$$

式中: d 表示语义特征向量; w_{ij} 表示权重系数; ω 、 δ 、 ρ 分别表示文本的词向量、文本向量、位置向量。

2.2.2 LDA模型

LDA主题模型的基本原理是通过概率分布来估计文档中单词出现的概率,从而发现文档中的主题和关键词。LDA主题模型的联合概率分布定义如式(6)所示。

$$p(\omega_m, z_m, \theta_m, \phi | \alpha, \beta) = \prod_{n=1}^{N_m} p(\omega_{m,n} | \phi_{z_{m,n}}) p(z_{m,n} | \theta_m) \cdot p(\theta_m | \alpha) p(\phi | \beta) \quad (6)$$

式中: ω_m 表示文档 m 中的单词集合; z_m 表示文档 m 中单词对应的主题集合; θ_m 表示文档 m 的主题分布; ϕ 表示所有文档共享的主题-词分布; α 和 β 表示Dirichlet分布的参数,分别用于生成文档的主题分布 θ_m 和主题-词分布 ϕ ; N_m 是文档 m 中的单词数量; $z_{m,n}$ 表示文档 m 中第 n 个单词对应的主题; $\omega_{m,n}$ 表示文档 m 中第 n 个单词。

2.2.3 向量拼接

在文本表示方面,BERT模型用于生成文档的句子嵌入向量,以捕捉文档的语义信息;LDA模型则用于生成文档的概率主题向量,反映每个文档在不同主题上的分布情况。为了更好地综合这两种信息,使用自动编码器连接这两种向量,将高维的拼接向量映射到低维空间,从而进一步提取其潜在语义信息,为后续的学科交叉研究主题识别提供更准确和全面的特征表示,如式(7)所示。

$$d'_m = (\mu; d_m) \quad (7)$$

式中: d'_m 表示融合BERT与LDA特征向量后的文本向量化表示; μ 表示LDA主题模型生成的概率主题向量; d_m 表示BERT模型生成的句子嵌入向量。

2.2.4 聚类

为了进一步识别学科交叉研究主题,借助UMAP(Uniform Manifold Approximation and Projection)算法对自动编码器学习到的潜在空间表示进行降维和可视化,采用K-means算法对潜在空间表示进行聚类分析,以发现文本数据中的潜在结构和模式,识别出具有潜在影响力和重要性的研究主题。主题聚类中通常需要确定主题数量,一般借助一些评价指标来评估模型的拟合效果,如困惑度(Perplexity)、主题一致性(Topic Coherence)等。本研究选择困惑度作为确定最优主题数量的评价指标,具体计算方式如式(8)~式(9)所示。

$$P(M) = \exp \left\{ - \frac{\sum_{m=1}^M \log_2 p(X_m)}{\sum_{m=1}^M N_m} \right\} \quad (8)$$

$$p(X_m) = p(z|m)p(x|z) \quad (9)$$

式中: P 表示困惑度函数; M 表示测试集中文档的数量; X_m 表示文档 m 中的词; $p(X_m)$ 表示文档 m 中单词

序列的概率; $p(z|m)$ 表示在给定文档 m 的条件下, 主题 z 的概率; $p(x|z)$ 表示在给定主题 z 的条件下, 单词 x 的概率。

2.3 学科交叉研究主题演化分析

为了更直观地揭示LIS领域学科交叉研究主题演化趋势, 将研究数据划分为4个时间窗口, 以BERT-LDA模型输出的主题词概率分布为权重, 通过同一主题下的主题词加权和将主题概率向量转化为主题语义向量, 利用相邻时间窗口内的主题语义向量之间的余弦相似度计算主题之间的相关性, 并选择具有强关联关系的主题对绘制研究主题演化路径。余弦相似度计算方法如式(10)所示。

$$\text{sim}(\mathbf{T}_i, \mathbf{T}_j) = \frac{\sum_{k=1}^n x_k(\mathbf{T}_i) \cdot y_k(\mathbf{T}_j)}{\sqrt{(\sum_{k=1}^n x_k^2(\mathbf{T}_i)) \cdot (\sum_{k=1}^n y_k^2(\mathbf{T}_j))}} \quad (10)$$

式中: $\text{sim}(\mathbf{T}_i, \mathbf{T}_j)$ 表示向量 $\mathbf{T}_i$ 和 $\mathbf{T}_j$ 之间的余弦相似度; $x_k(\mathbf{T}_i)$ 和 $y_k(\mathbf{T}_j)$ 分别表示向量 $\mathbf{T}_i$ 和 $\mathbf{T}_j$ 的第 k 个元素。

2.4 研究前沿指标体系构建

普莱斯(D. J. Price)于1965年首次提出了研究前沿这一概念, 认为引文网络中最近的、被广泛引用的文献集合就是活跃的研究前沿^[24-25]。前沿主题识别的研究中产生了诸多测量的指标, 如新颖度、增长性、关注度、影响力等, 研究者对这些指标制定了具体的测度方法, 如新颖度是通过评估某主题的参考文献及其引用的知识基础的发表时间来界定的^[26]。本研究综合考虑学术界的普遍认知, 将研究前沿定义为那些蕴含巨大发展潜力、正处于兴起阶段且未来可能吸引广泛注意力的研究主题或领域, 通常具有前瞻性、影响力、成长性^[27]。为了更精准地衡量研究主题的前沿性, 结合本研究数据源的独特性质, 从以下4个维度对研究主题的前沿性进行测度。

(1) 新颖度: 可用论文的平均发表年份来表示。论文发表时间越晚, 说明该主题中近期研究论文越多, 新颖度越高, 越可能成为研究前沿主题。新颖度计算方法如式(11)所示。

$$\delta_{N,z} = \sum_{m=1}^{M_t(z)} \frac{T_m(z)}{M_t(z)} \quad (11)$$

式中: $M_t(z)$ 表示时间段 t 内主题 z 相关的论文总数; $T_m(z)$ 表示时间段 t 内主题 z 相关论文 m 的发表时间。

(2) 增长性: 可以通过相关主题论文产出量的年增长率来衡量。年增长率反映了该主题在研究领域中的发展速度和动态变化, 如果一个主题的相关论文产出量年增长率较高, 则表明该主题的研究在近年来得到了更多的关注和投入, 具有较高的研究价值。增长性计算方法如式(12)所示。

$$\delta_{I,z} = \frac{1}{4} \sum_{t=1}^4 \frac{M_{t+1}(z)}{M_t(z)} \quad (12)$$

(3) 关注度: 可以通过相关主题论文数量占总论文数量的比例来衡量。如果一个主题的相关论文数量占总论文数量的比例较高, 则说明该主题在研究领域中具有较高的关注度, 研究者们对该主题的兴趣更高。关注度计算方法如式(13)所示。

$$\delta_{A,z} = \frac{M_t(z)}{\text{sum}(M_t)} \quad (13)$$

(4) 影响力: 被引频次反映了研究论文在学术界和行业中的影响力和重要性, 如果一个主题相关论文的被引频次较高, 则说明该主题在学术界和行业中具有较大的影响力, 对相关领域的发展产生了积极的影响, 促进了后续研究和研发工作的开展。影响力计算方法如式(14)所示。

$$\delta_{E,z} = \sum_{m=1}^{M_t(z)} \frac{C_m(z)}{M_t(z)} \quad (14)$$

式中: $C_m(z)$ 表示时间段 t 内主题 z 相关论文 m 的被引频次。

3 数据采集与结果分析

3.1 数据获取

根据2023年版JCR (Journal Citation Reports) 的Information Science & Library Science分区, 选择影响因子排名前21的外文期刊为数据来源, 期刊信息如表1所示。获取WOS数据库中2004—2023年发表在这21种期刊上的所有论文, 经过对数据的处理和筛选, 去除无摘要、重复、空白、不符合要求的数据后, 共获得24 206篇有效论文。

为了得到论文参考文献所属学科信息, 从WOS数据库下载文献的题录数据, 包含标题、作者、摘要、参

表1 研究数据来源期刊列表

序号	期刊名称	影响因子	分区
1	International Journal of Information Management	21.0	Q1
2	Information & Management	9.9	Q1
3	Information Processing & Management	8.6	Q1
4	Telematics and Informatics	8.5	Q1
5	Government Information Quarterly	7.8	Q1
6	Journal of Management Information Systems	7.7	Q1
7	MIS Quarterly	7.3	Q1
8	Journal of Computer-Mediated Communication	7.2	Q1
9	Journal of Knowledge Management	7.0	Q1
10	Journal of Enterprise Information Management	6.5	Q1
11	Journal of Organizational and End User Computing	6.5	Q1
12	Journal of the American Medical Informatics Association	6.4	Q1
13	Information and Organization	6.3	Q1
14	International Journal of Geographical Information Science	5.7	Q1
15	Journal of Information Technology	5.6	Q1
16	Information Technology for Development	4.8	Q2
17	Journal of Global Information Management	4.7	Q2
18	Profesional de la Informacion	4.2	Q2
19	Social Science Computer Review	4.1	Q2
20	Scientometrics	3.9	Q2
21	Journal of Informetrics	3.7	Q2

考文献等,使用Leydesdroff^[28]开发的WOS文献题录信息处理程序isi.exe对每篇论文的参考文献信息进行提取,以得到参考文献所在期刊信息。由于isi.exe提取的参考文献信息是期刊名称的缩写,需要借助WOS的ESI数据库进行全称与缩写的对照。

得到论文参考文献所在期刊后,还需要对每篇参考文献的所属学科进行识别。JCR数据库提供了收

录的期刊信息,涵盖期刊名称、ISSN、语言、所属学科等,使用自编程序关联信息后,最终得到论文、参考文献、期刊全称、学科类别对照表,示例如表2所示。

3.2 LIS领域学科交叉论文确定

为了进一步识别研究主题及演化趋势,将所收集到

表2 论文、参考文献、期刊全称、学科类别对照表(示例)

论文编号	参考文献	作者	年份	期刊全称	DOI	学科
n1	Adam MTP, 2017, BUS INFORM SYST ENG+, V59, P277	Adam M T P	2017	Business & Information Systems Engineering	10.1007/s12599-016-0451-3	Computer Science, Information Systems
n1	Anusha AS, 2020, IEEE J BIOMED HEALTH, V24, P92	Anusha A S	2020	IEEE Journal of Biomedical and Health Informatics	10.1109/JBHI.2019.2893222	Computer Science, Information Systems; Computer Science, Interdisciplinary Applications; Mathematical & Computational Biology; Medical Informatics
n24206	Watson R. T., 1997, Journal of Management Information Systems, V13, P91	Watson R T	1997	Journal of Management Information Systems	10.1016/0378-7206(91)90056-8	Computer Science, Information Systems; Information Science & Library Science; Management
n24206	WHITE KB, 1984, INFORM MANAGE, V7, P331	White K B	1984	International Journal of Information Management	10.1016/0378-7206(84)90028-4	Information Science & Library Science

的数据划分为4个时间窗口: 2004—2008年、2009—2013年、2014—2018年、2019—2023年。基于式(1)~式(4)编写Python程序, 分别计算每篇论文的相对丰富性、平衡性、差异性、DIV指标, 其中学科间的距离由Chavarro等^[29]构建的学科相似度矩阵得到。LIS领域的论文学科交叉测度结果如表3所示。根据学科交叉论文遴选原则, 将综合指标数值排名数据集前50%的论文作为学科交叉论文。

表3 学科交叉测度结果 (示例)

论文编号	相对丰富性	平衡性	差异性	DIV
n1	0.322	0.562	0.379	0.053
n2	0.129	0.580	0.285	0.016
n3	0.266	0.505	0.356	0.047
n4	0.260	0.488	0.240	0.032
n5	0.149	0.377	0.275	0.026
n6	0.323	0.638	0.323	0.038
n7	0.212	0.599	0.278	0.024

根据学科交叉论文识别结果, 参考Biglan^[30]提出的学科分类框架, 将LIS领域论文的参考文献划分至19个具体领域(不包含LIS), 分别为数学、物理学、生物学、计算机科学、工程学、医学、环境科学、艺术学、历史学、语言学、哲学、人文地理、人类学、心理学、社会学、管理学、经济学、法学、教育学。

在学科交叉论文遴选的基础上, 进一步对参考文献所属学科进行引用重要程度统计, 计算不同学科类别的参考文献数量在每篇论文参考文献中的占比, 以更

清楚地了解LIS领域研究的学科交叉情况。结果显示, 计算机科学、管理学、经济学、医学、心理学、社会学、工程学、人文地理、环境科学是LIS领域的核心交叉学科。为了深入了解学科交叉随时间变化的趋势, 选取上述9个学科进行演化分析, 计算该学科类别所有参考文献占比的平均值作为该年度的学科引用占比, 分析结果如图2所示。可以看出, 计算机科学始终是LIS领域的重要研究主题, 为LIS领域研究提供技术支撑, 促进了“问题域-方法论”交叉模式的形成^[31]。同时, 在信息规模不断扩大的趋势下, 各个领域的信息内容越来越容易获取, LIS领域的研究范式和理论方法可以更广泛地用于解决其他应用领域的研究问题, LIS与心理学、社会学、工程学、环境科学的交叉研究越来越受到重视^[32]。随着研究问题的不断深入和拓展, 结合实践发展的需求, “方法论-问题域”交叉模式将有望转变成“问题域-问题域”, 促进新兴交叉学科的形成^[33]。

3.3 LIS领域学科交叉研究主题识别

在识别学科交叉研究主题方面, 将论文的标题和摘要作为分析内容, 经过一系列数据预处理步骤后使用BERT-LDA模型进行主题识别及抽取, 通过困惑度计算最佳主题数量, 进行K-means聚类提取核心研究主题。对于每一个识别出的研究主题, 选择概率排名前10的词作为该主题的主题词, 最终在4个时间窗口分别识别出了80、80、120、140个主题词, 并依据这些主题

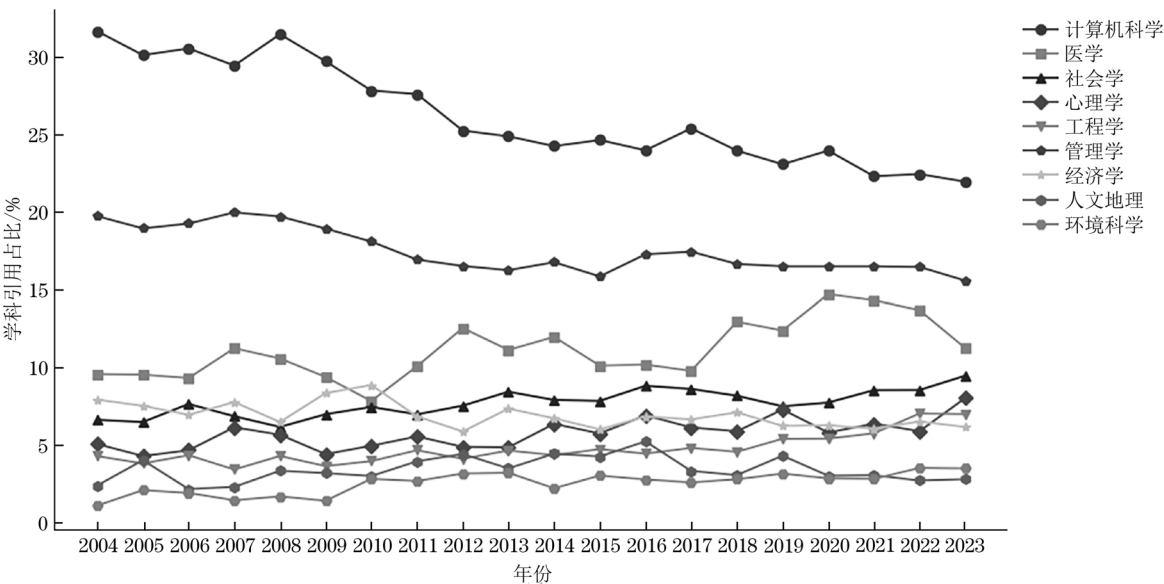


图2 核心交叉学科演化趋势

词的含义总结归纳得到8、8、12、14个研究主题。为了更直观地展示和理解这些主题，进一步深入分析主题词

含义和该主题下的代表性论文，结合每个主题词在相应主题中的权重，对每一个主题进行命名，如表4所示。

表4 研究主题分布

时间窗口	年份	主题数量/个	主题名称
T1	2004—2008	8	健康信息服务、科学计量、科学出版与影响力评估、信息技术与信息服务、信息检索与用户行为、页面设计与用户体验、搜索引擎、电子商务
T2	2009—2013	8	医疗信息系统、健康信息分析、科学计量、政策与技术发展、知识抽取、算法优化、信息检索与用户行为、搜索引擎与排名推荐
T3	2014—2018	12	健康信息分析、医疗信息系统、科学计量、知识抽取、技术识别与预测、算法优化、空间人文与地理数据分析、政策与技术发展、网络舆情、信息推荐、社交媒体、情感分析
T4	2019—2023	14	医疗信息系统、健康信息分析、医疗数据模型、机器学习与深度学习、科学计量、知识抽取、时空大数据分析、商业与市场决策、技术识别与预测、企业情报分析、气候信息学、政策发布与政府事件、社交媒体、舆情治理与情感分析

为了验证本研究所使用方法的科学性，进一步利用主题一致性对BERT-LDA模型效果做出检验，同时对单独使用LDA、BERT进行学科交叉研究主题识别的效果，3种建模方法的识别效果对比如表5所示。可以看出BERT-LDA方法的一致性值为0.552，优于其他两种方法，意味着本研究所使用的建模方法在学科交叉研究主题识别方面有较好的效果。

表5 3种建模方法的识别效果对比

方法	一致性
LDA	0.414
BERT	0.467
BERT-LDA	0.552

3.4 LIS领域学科交叉的演化分析

为了深入探讨LIS领域学科交叉研究主题的演化路径，通过Python程序计算相邻时间窗口主题之间的相似度，将所有主题间相似度的平均值作为筛选的阈值，选择具有高相似度的主题进行可视化分析，演化路径如图3所示。

基于主题识别得到的研究主题分布和演化路径可以看出，LIS领域4个时间窗口内的学科交叉研究主题之间存在着紧密的联系。健康信息服务、健康信息分析、医疗数据模型等主题聚焦于医疗领域的信息应用，且技术手段随着时间的推移不断升级，推动了医疗服务的信息化、数字化、智能化发展。从信息处理的角度出发，科学计量、知识抽取、技术识别与预测、机器学习与深度学习研究主题的演变代表了信息挖掘与知识发现的核心研究方向，研究目标聚焦于利用技术手段从

海量数据中提取有价值的信息和知识，助力LIS领域发展。此外，从T1的搜索引擎、信息检索与用户行为到T4的社交媒体、舆情治理与情感分析，可以看出互联网的信息行为也始终是研究的核心主题，学者们通过研究特定话题下的信息行为、情感表达和社会关系揭示互联网信息传播的规律和事件观点，进而提出营造网络信息环境和治理网络舆情的相关建议。因此，2004—2023年LIS领域学科交叉研究主题的核心研究方向可归纳为信息挖掘与知识发现、互联网信息行为和医疗信息学。其中：信息挖掘与知识发现是计算机科学、统计学等领域的研究方法在信息处理方面的研究应用，方法体系的不断完善使其可以对各个领域的信息进行处理；由于网络信息涉及的领域范围较广，目前关于信息传播与分析的研究已扩展至政治、医学、金融、环境等多个领域，即通过信息行为的研究范式为不同背景下的信息传播提供应对策略；医疗信息学则是学科交叉背景下产生的新兴交叉学科，研究人员通过分析和讨论医学领域的信息数据为其提供解决方案，目前已形成较为成熟的研究体系。

(1) 信息挖掘与知识发现。此类论文主要是LIS与计算机技术、统计学、应用数学的交叉，研究对象包括科研论文、专利数据、网络信息等。各阶段的研究主题演化图如图4所示。以科学和技术层面的信息挖掘为例，在T1阶段，科学计量和科学出版与影响力评估是研究的重点，研究主要集中在对科学论文的计量、分析以及科研成果的评估上。知识图谱作为一种基于图的数据结构，被广泛应用于科学知识的展示和分析^[34]。进入T2阶段，研究重心逐渐转向科学计量与算法的结合，算法的发展为科学计量提供了更强大的分析工具，使得

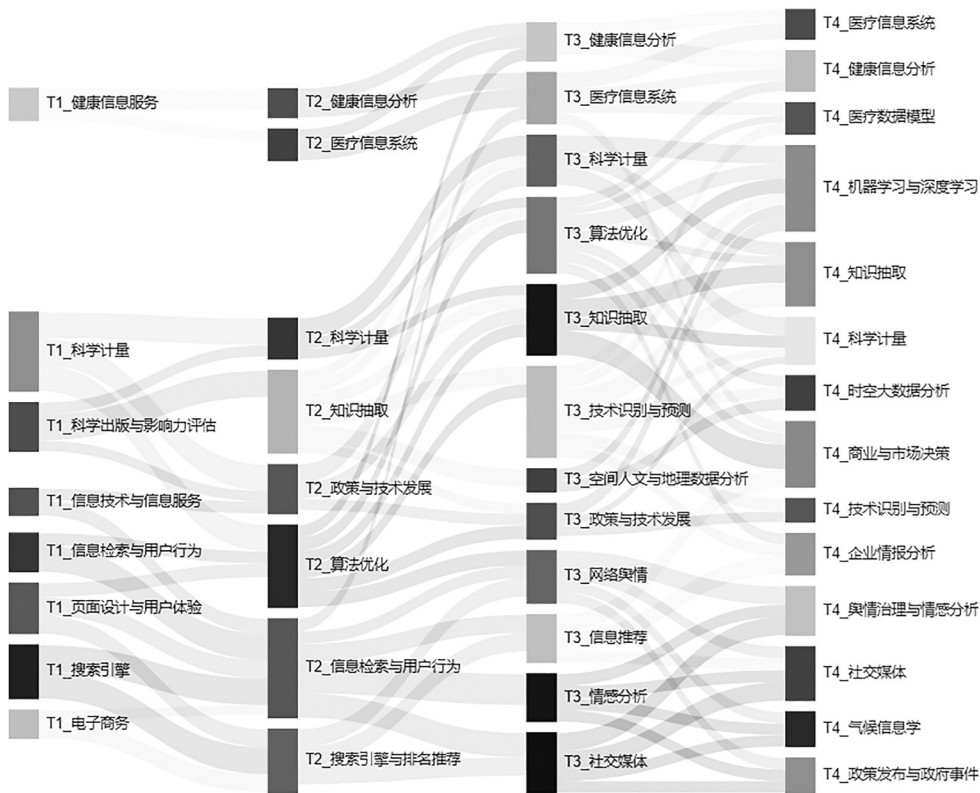


图3 研究主题演化路径

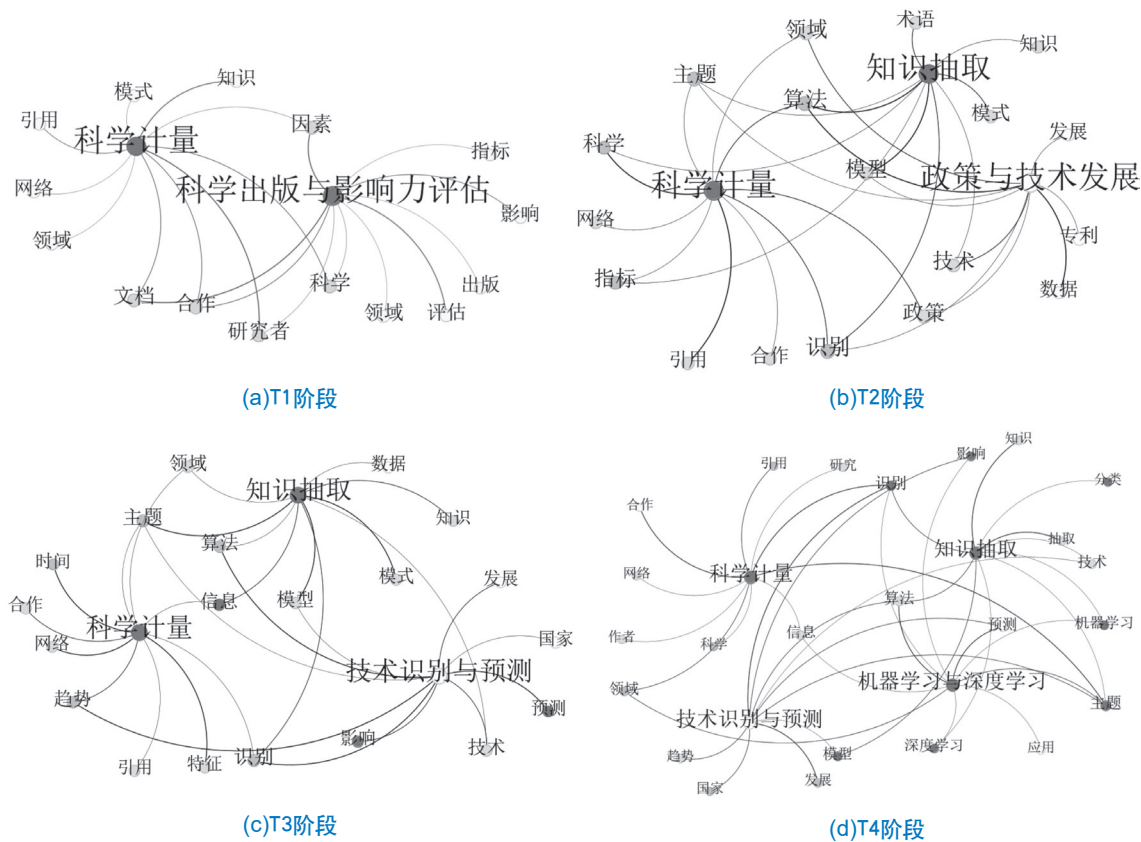


图4 信息挖掘与知识发现研究主题演化图

研究人员能够更精确地分析科学数据,揭示科学发展的内在规律^[35]。同时,算法的应用也开始在政策制定和科技发展领域发挥重要作用^[36]。T3阶段,在信息技术的加持下,研究者们开始利用数据挖掘技术,从海量的数据中提取有价值的信息和模式,以揭示用户行为、知识传播、技术创新等蕴含的潜在特征和演变趋势^[37-38]。而在T4阶段,越来越多的学者尝试将机器学习、深度学习等应用于产学研项目(论文、专利、基金)的研究,以识别科研发展趋势,提高模型准确率;技术预测也进一步拓展至金融^[39]、市场^[40]、地理^[41]、医疗^[42]等领域,信

息挖掘的应用领域愈发多元化,如LSTM与ST-ResNet算法的结合被用来预测时空流量,为城市发展和交通管理提供了更加精准和有效的解决方案^[43]。总体而言,随着计算机技术的不断发展,计算机科学领域的相关方法为LIS领域的研究提供了愈发高效和精准的技术工具,推进LIS领域研究范式变革与知识升级。

(2) 互联网信息行为。此核心主题的研究路径可概括为“信息系统用户行为—行为动机与信息策略—社交媒体信息行为—网络舆情传播与治理”,各阶段的研究主题演化图如图5所示。在T1阶段,信息检索与

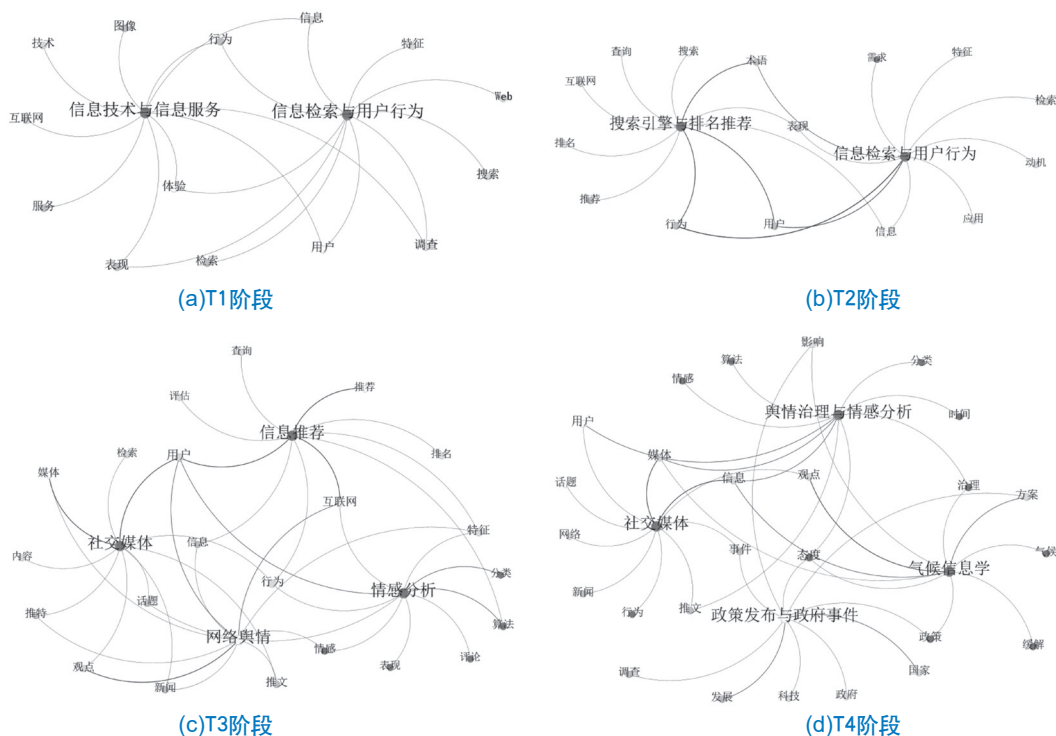


图5 互联网信息行为研究主题演化图

用户行为、信息技术与信息服务是研究的重点,研究者普遍关注如何更有效地从海量信息中检索到所需的内容,探索更有效的信息组织和可视化方法,以提升用户体验和信息展示的效果,便于获取信息。到T2阶段,研究重点扩展至搜索引擎与排名推荐,研究者开始关注各种情景下用户的心理和行为特征,试图基于信息行为领域的研究范式分析用户意图、心理机制和行为动机,以更准确地满足用户的个性化需求。T3阶段,信息传播的重要渠道从搜索引擎转变成社交媒体,以推特为代表的社交媒体平台上的内容开始受到广泛关注。用户行为的相关研究逐渐转向社交媒体平台上的舆情传播、情感分析等研究问题,分析社交媒体上的信息流

动和用户行为能够揭示公众的观点和情感倾向,便于舆情治理^[44]。在T4阶段,舆情仍然是研究的热点,研究人员也愈发关注舆情传播过程中与传播后的治理策略,期望改善信息传播环境^[45]。随着研究范围的不断扩展、学科间的深入交叉融合,相关研究主题也逐渐扩展至政治选举新闻、科技政策发布、气候变化、金融发展等领域,交叉模式有望从“方法论-问题域”转向“问题域-问题域”,促进新兴交叉学科的产生^[46]。研究人员通过情感分析对博文、推文、新闻等社交媒体内容进行深入洞察,以了解公众对于气候变化、政治政策变化等热点话题的情感态度和看法^[47-48],为政策制定者、企业等提供更加精准和有价值的参考。总体而言,在互联网信

息行为这一主题的核心演化路径上, 学科交叉研究呈现出研究领域和研究问题逐渐多元化的特点, 研究人员发现了LIS领域的方法论在其他领域应用的可能性, 且这种研究范式的应用有望促成新的交叉学科领域产生。

(3) 医疗信息学。随着公众对健康问题的日益重视和健康需求的不断增长, LIS领域的研究范畴逐渐扩展到健康信息领域, 研究问题聚焦于健康医疗信息系统的开发、健康大数据的深入分析以及健康信息行为模式的探索。从图6可以看出, T1阶段的健康信息服务主题在T2阶段转向具体的应用领域, 如医疗信息系统、健康信息分析。医疗信息系统是提升医疗服务和决策效率的信息化工具^[49], 电子健康记录作为其重要的组成部分, 实现了患者信息的电子化存储、传输和共享, 使得医疗信息的存储、检索和使用变得更为便捷和高效, 为医疗服务提供了更加全面和准确的数据支持。到了T3

阶段, 健康信息分析的研究和应用进一步深化, 同时在线医疗社区逐渐兴起, 平台上的信息内容成为健康信息行为研究的重要依据。以技术为媒介的在线医疗社区为患者和医生提供了更加便捷的交流和互动平台, 使患者能够更加方便地获取健康知识和医疗建议, 进一步缩小了城乡健康差距^[50]。进入T4阶段, 在机器学习等算法模型的支持下, 研究者开始广泛利用电子健康记录或医疗健康系统中的数据进行建模, 以此评估患者风险和识别疾病特征, 预测患者可能出现的健康风险, 形成系统的医疗解决方案, 为医生提供决策支持^[51-53]。区块链技术的应用则为医疗信息资源的组织、管理、传播和利用带来了更高的安全性和可靠性, 通过去中心化、分布式存储和加密机制, 确保了数据的完整性和隐私性。如Zou等^[54]提出了一种基于区块链技术的电子医疗系统, 在推动跨机构数据共享的同时加强了隐私保护, 提高诊断和治疗的准确性和效率。

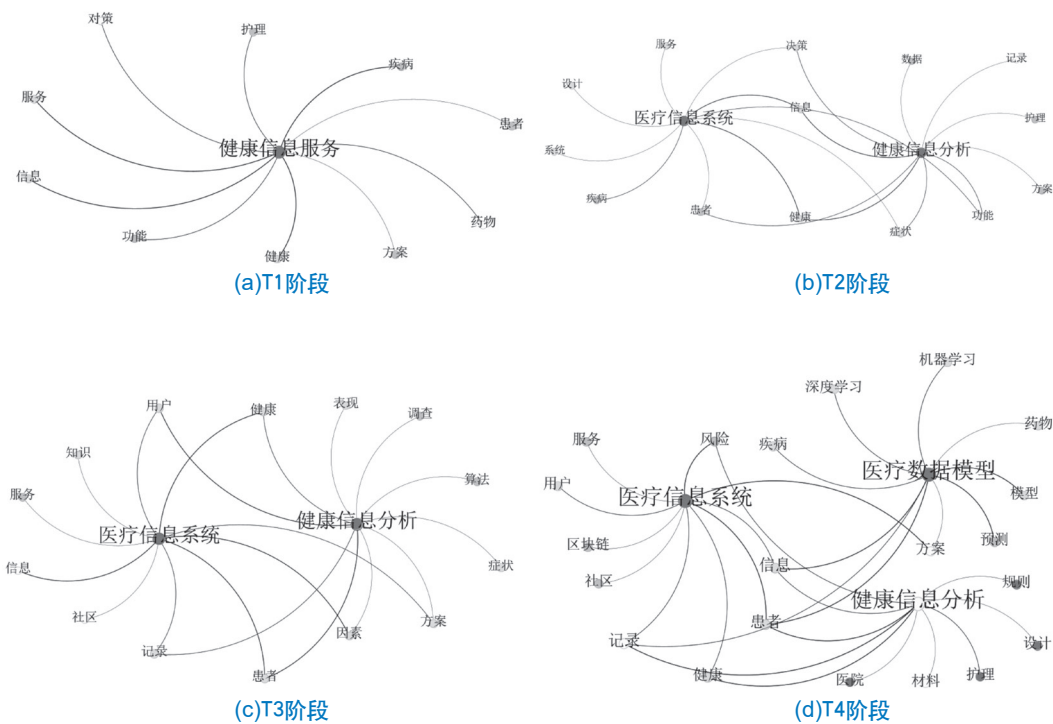


图6 医疗信息学研究主题演化图

3.5 LIS领域学科交叉的前沿主题识别

为了进一步识别LIS领域内具有前沿性的研究主题, 依据前文2.4节所构建的研究前沿指标体系, 根据每个主题下的论文计算主题的新颖度、增长性、关注度、影响力, 选择各项指标得分均高于平均值的主题为

研究前沿主题, 最终识别出4个阶段的研究前沿主题, 如表6所示。

根据识别结果, T1阶段的前沿主题主要聚焦于信息技术和信息服务领域, 主要交叉学科为计算机科学、心理学, 研究者们普遍关注信息技术如何赋能信息服务领域。T2阶段的前沿主题进一步聚焦于科学计量、

表6 研究前沿主题识别结果

时间窗口	主题名称	新颖度	增长性	关注度	影响力	核心交叉学科	交叉模式
T1	信息技术与信息服务	2 006.220	1.140	0.272	55.727	计算机科学、心理学	问题域-方法论
T2	科学计量	2 011.085	1.102	0.175	43.077	计算机科学、管理学、数学	问题域-方法论
T2	算法优化	2 011.191	1.091	0.153	41.015	计算机科学	问题域-方法论
T2	政策与技术发展	2 011.021	1.021	0.180	42.711	经济学、社会学	方法论-问题域
T3	社交媒体	2 016.102	1.052	0.163	29.229	新闻传播学、心理学、社会学	方法论-问题域
T3	健康信息分析	2 016.041	1.021	0.103	31.376	医学、计算机科学	问题域-问题域
T3	科学计量	2 016.090	1.050	0.121	31.037	计算机科学、管理学、数学	问题域-方法论
T4	舆情治理与情感分析	2 021.283	1.173	0.159	10.722	管理学、新闻传播学、社会学	方法论-问题域
T4	医疗数据模型	2 021.260	1.175	0.117	10.727	医学、计算机科学	问题域-问题域
T4	机器学习与深度学习	2 021.219	1.159	0.135	11.122	计算机科学、统计学	问题域-方法论

算法优化、政策与技术发展。算法是数据处理和分析的核心工具，其研究和发展对于推动领域进步具有关键作用，在分类、聚类、引文分析、影响力评估等方面帮助科研人员更精确地分析科学数据^[55]。除了研究数据外，政策支持下的技术发展同样是研究者们重点关注的课题，技术发展往往与经济发展密切相关，如何挖掘科技文本中的关键信息以推动科技创新与社会进步是相关学者广泛关注的问题，相关研究拓宽了LIS领域研究方法论的应用视角^[56]。

T3阶段的前沿主题更加多元化，包括社交媒体、健康信息分析、科学计量。研究信息行为的情境逐渐从传统转向各种社交媒体，社交媒体情境下的研究涉及用户行为、信息传播机制等多个方面，研究话题分布于政治、政策、环境、健康等多个领域，结合心理学相关理论和知识收集并分析用户发表的网络信息有助于了解公众信息感知和信息行为的成因和影响因素，从而方便相关主体采取干预措施^[57-58]。电子健康记录记载着患者个人信息、病历信息等，为疾病诊断和预测提供了丰富的研究数据，研究人员通常需要使用信息技术对电子健康记录中的信息进行深入挖掘和处理分析，以进一步完成疾病检测、临床预测、药物剂量检测等研究工作，为医学领域的信息服务质量提升提供支持，促进交叉学科研究的进一步深入^[59]。

T4阶段的前沿主题聚焦于医疗数据模型、舆情治理与情感分析、机器学习与深度学习。相比前一阶段的研究前沿，规模逐渐增长的海量医学信息为医疗数据模型的建立提供了数据基础，促进其在医疗数据的挖掘与分析、健康管理系统的构建、疾病预测与诊断等方面的应用，使医疗数据模型逐渐成为推动健康信息学领域决策与创新的强大引擎^[42]。舆情分析仍是此阶段

的研究前沿，相关研究大多基于信息行为的研究范式，利用新兴算法模型对公众意见和情感进行监测和分析，如何及时发现并监测特定领域和特定话题下的网络舆情、降低舆情给公众的负面影响是研究者正在关注和重视的问题，交叉学科更加多元化^[45]。机器学习和深度学习作为人工智能领域的重要分支，在信息检索、文本挖掘、趋势预测、信息抽取等方面展现出了巨大的应用潜力^[43]，为LIS领域的信息计量、行为分析、情报分析提供了新的方法。

综合以上识别结果，可以看到数据挖掘、机器学习和深度学习等研究正在不断突破技术大关，研究粒度不断细化，研究方法逐渐深入，为LIS领域的数据处理、信息分析和知识发现提供了新的思路和方法。同时，以网络舆情研究为代表的信息分析推动了LIS与计算机科学、医学、经济学、地理学等更多学科的深度融合，方法论所应用的研究领域正在不断丰富。此外，医疗健康与信息分析的结合也正在促使健康信息学这一交叉学科的形成，推动LIS领域向纵深拓展，帮助解决人民健康的实际问题。

4 结语

本研究以WOS数据库中LIS领域内2004—2023年的学术论文为数据源，借助学科交叉测度指标筛选出具有跨学科特征的论文数据，采用BERT-LDA主题模型识别LIS领域的研究主题，同时借助主题相似度计算各阶段主题的相似性，绘制演化路径，并在此基础上基于新颖度、增长性、关注度、影响力4个指标计算研究主题的前沿性，得出如下研究结论：①从学科交

叉情况来看, LIS与其他学科的交叉研究主要集中于计算机科学、管理学、经济学、医学、心理学、社会学、工程学、人文地理、环境科学等学科, 计算机科学始终是LIS领域研究的技术支撑, 且LIS与其他学科领域的交叉研究正在不断拓展和延伸; ②2004—2023年LIS领域学科交叉研究有3条主要的演化路径, 分别是信息挖掘与知识发现、互联网信息行为、医疗信息学, 且研究呈现出“整合—分裂”的模式, 学科方法论的应用领域正在逐渐细化; ③现阶段的学科交叉研究的前沿主题主要聚焦于3个方面, 即医疗数据模型、舆情治理与情感分析、机器学习与深度学习。总的来说, 当前LIS领域的学科交叉研究主题正在不断突破技术大关, 基于信息技术的研究方法和其在不同领域的应用将是未来一段时间内研究者首要关注的方向。

本研究选择国外LIS领域的重要期刊论文作为研究对象, 识别核心交叉学科, 分析学科交叉研究前沿主题和演化路径, 可以为国内信息资源管理领域的研究发展提供创新思路和应用指导。

参考文献

- [1] National Academy of Sciences. Facilitating interdisciplinary research[M]. Washington, DC: National Academies Press, 2005.
- [2] 金力. 金力: 学科交叉已成为高水平科研与顶尖人才的重要特征[EB/OL]. [2024-03-10]. <https://fddi.fudan.edu.cn/c9/8c/c18965a444812/page.htm>.
- [3] 于良芝, 樊振佳. 信息处理进阶: LIS学科独特性再探究[J]. 情报资料工作, 2020, 41(5): 5-13.
- [4] 马捷, 何美慧, 冯文, 等. 交叉与边界: 论信息资源管理跨学科发展策略[J]. 数字图书馆论坛, 2024(1): 1-12.
- [5] 马岩, 郑建明, 刘佳静. 2023年LIS期刊论文研究方法应用态势[J]. 图书馆论坛, 2024, 44(8): 8-13.
- [6] CHANG Y W, HUANG M H. A study of the evolution of interdisciplinarity in library and information science: using three bibliometric methods[J]. Journal of the American Society for Information Science and Technology, 2012, 63(1): 22-33.
- [7] FIGUEROLA C G, GARCÍA MARCO F J, PINTO M. Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA[J]. Scientometrics, 2017, 112(3): 1507-1535.
- [8] HUANG M H, CHANG Y W. A study of interdisciplinarity in information science: using direct citation and co-authorship analysis[J]. Journal of Information Science, 2011, 37(4): 369-378.
- [9] 宋艳辉, 朱李, 舒非, 等. 中美图情学博士学位论文主题跨学科发展比较研究[J]. 情报学报, 2022, 41(5): 536-548.
- [10] CHEN C M. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.
- [11] CALLON M, COURTIAL J P, TURNER W A, et al. From translations to problematic networks: an introduction to co-word analysis[J]. Social Science Information, 1983, 22(2): 191-235.
- [12] 郝晋清, 王珺. 主题-时序视域下中国图书馆学会年会主题分布及演化分析[J]. 图书情报工作, 2019, 63(2): 107-119.
- [13] SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents[J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.
- [14] KESSLER M M. Bibliographic coupling between scientific papers[J]. American Documentation, 1963, 14(1): 10-25.
- [15] GARFIELD E. Citation indexes in sociological and historical research[J]. American Documentation, 1963, 14(4): 289-291.
- [16] 王小梅, 邓启平, 李国鹏, 等. ESI研究前沿的科学图谱及在纳米领域的应用[J]. 图书情报工作, 2017, 61(12): 106-112.
- [17] 范云满, 马建霞. 基于LDA与新兴主题特征分析的新兴主题探测研究[J]. 情报学报, 2014, 33(7): 698-711.
- [18] 冯佳. 研究前沿识别与分析方法研究[D]. 长春: 吉林大学, 2017.
- [19] 王秀红, 高敏. 基于BERT-LDA的关键技术识别方法及其实证研究: 以农业机器人为例[J]. 图书情报工作, 2021, 65(22): 114-125.
- [20] 赵月华, 朱思成, 苏新宁. 面向网络虚假医疗信息的识别模型构建研究: 一种基于预训练的BERT模型[J]. 情报科学, 2021, 39(12): 165-173.
- [21] 范晴晴, 李长玲, 荣国阳, 等. 跨学科输入知识对学科发展的影响力分析: 以图书情报学科为例[J]. 情报科学, 2023, 41(2): 79-85.
- [22] 韩正琪, 刘小平, 徐涵. 基于Rao-Stirling指数的学科交叉文献发现: 以纳米科学与纳米技术为例[J]. 图书情报工作, 2018, 62(1): 125-131.
- [23] LEYDESDORFF L, WAGNER C S, BORNEMANN L. Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient[J]. Journal of Informetrics, 2019, 13(1): 255-269.

- [24] 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.
- [25] 张雪, 张志强, 曹玲静, 等. 学科领域研究前沿识别方法研究进展[J]. *图书情报工作*, 2022, 66 (12): 139-151.
- [26] GUO H N, WEINGART S, BÖRNER K. Mixed-indicators model for identifying emerging research areas[J]. *Scientometrics*, 2011, 89 (1): 421-435.
- [27] 陈稳, 陈伟. 科学与技术对比视角下的前沿主题识别与演化分析[J]. *情报杂志*, 2022, 41 (1): 67-73, 163.
- [28] LEYDESDROFF L. Words and co-words as indicators of intellectual organization[J]. *Research Policy*, 1989, 18 (4): 209-223.
- [29] CHAVARRO D, TANG P, RAFOLS I. Interdisciplinarity and research on local issues: evidence from a developing country[J]. *Research Evaluation*, 2014, 23 (3): 195-209.
- [30] BIGLAN A. The characteristics of subject matter in different academic areas[J]. *Journal of Applied Psychology*, 1973, 57 (3): 195-203.
- [31] 马费成, 张帅. 我国图书情报领域新兴交叉学科发展探析[J]. *中国图书馆学报*, 2023, 49 (2): 4-14.
- [32] STACHOFSKY J, SCHAUPP L C, CROSSLER R E. Measuring the effect of political alignment, platforms, and fake news consumption on voter concern for election processes[J]. *Government Information Quarterly*, 2023, 40 (3): 101810.
- [33] DWIVEDI Y K, HUGHES L, KAR A K, et al. Climate change and COP26: are digital technologies and information management part of the problem or the solution? an editorial reflection and call to action[J]. *International Journal of Information Management*, 2022, 63: 102456.
- [34] ORTEGA J L, AGUILLO I F. Visualization of the Nordic academic web: link analysis using social network tools[J]. *Information Processing & Management*, 2008, 44 (4): 1624-1633.
- [35] BRANDTZÆG P B. Social networking sites: their users and social implications—a longitudinal study[J]. *Journal of Computer-Mediated Communication*, 2012, 17 (4): 467-488.
- [36] CHEN Y L, CHIU Y T. An IPC-based vector space model for patent retrieval[J]. *Information Processing & Management*, 2011, 47 (3): 309-322.
- [37] VALVERDE-REBAZA J C, ROCHE M, PONCELET P, et al. The role of location and social strength for friendship prediction in location-based social networks[J]. *Information Processing & Management*, 2018, 54 (4): 475-489.
- [38] REKIK R, KALLEL I, CASILLAS J, et al. Assessing web sites quality: a systematic literature review by text and association rules mining[J]. *International Journal of Information Management*, 2018, 38 (1): 201-216.
- [39] YANG Y, QIN Y, FAN Y Y, et al. Unlocking the power of voice for financial risk prediction: a theory-driven deep learning design approach[J]. *MIS Quarterly*, 2023, 47 (1): 63-96.
- [40] PEREIRA M M, FRAZZON E M. A data-driven approach to adaptive synchronization of demand and supply in omnichannel retail supply chains[J]. *International Journal of Information Management*, 2021, 57: 102165.
- [41] SAEEDIMOGHADDAM M, STEPINSKI T F. Automatic extraction of road intersection points from USGS historical map series using deep convolutional neural networks[J]. *International Journal of Geographical Information Science*, 2020, 34 (5): 947-968.
- [42] ZHANG Z Q, YAN C, ZHANG X M, et al. Forecasting the future clinical events of a patient through contrastive learning[J]. *Journal of the American Medical Informatics Association*, 2022, 29 (9): 1584-1592.
- [43] REN Y B, CHEN H F, HAN Y, et al. A hybrid integrated deep learning model for the prediction of citywide spatio-temporal flow volumes[J]. *International Journal of Geographical Information Science*, 2020, 34 (4): 802-823.
- [44] TURCOTTE J, YORK C, IRVING J, et al. News recommendations from social media opinion leaders: effects on media trust and information seeking[J]. *Journal of Computer-Mediated Communication*, 2015, 20 (5): 520-535.
- [45] LIU J, LIU L Y, TU Y, et al. Multi-stage Internet public opinion risk grading analysis of public health emergencies: an empirical study on microblog in COVID-19[J]. *Information Processing & Management*, 2022, 59 (1): 102796.
- [46] PAPAGIANNIDIS S, MARIKYAN D. Environmental sustainability: a technology acceptance perspective[J]. *International Journal of Information Management*, 2022, 63: 102445.
- [47] UPADHYAYA A, FISICHELLA M, NEJDL W. Towards sentiment and temporal aided stance detection of climate change tweets[J]. *Information Processing & Management*, 2023, 60 (4): 103325.
- [48] WILSON C. The socialization of civic participation norms

- in government?: assessing the effect of the open government partnership on countries' e-participation[J]. Government Information Quarterly, 2020, 37 (4): 101476.
- [49] MENON N M, YAYLACICEGI U, CEZAR A. Differential effects of the two types of information systems: a hospital-based study[J]. Journal of Management Information Systems, 2009, 26 (1): 297-316.
- [50] GOH J M, GAO G G, AGARWAL R. The creation of social value: can an online health community reduce rural-urban health disparities? [J]. MIS Quarterly, 2016, 40 (1): 247-263.
- [51] LIU X K, CHUBAK J, HUBBARD R A, et al. SAT: a surrogate-assisted two-wave case boosting sampling method, with application to EHR-based association studies[J]. Journal of the American Medical Informatics Association, 2022, 29 (5): 918-927.
- [52] WEIR C R, TABER P, TAFT T, et al. Feeling and thinking: can theories of human motivation explain how EHR design impacts clinician burnout? [J]. Journal of the American Medical Informatics Association, 2021, 28 (5): 1042-1046.
- [53] LIU S R, SCHLESINGER J J, MCCOY A B, et al. New onset delirium prediction using machine learning and long short-term memory (LSTM) in electronic health record[J]. Journal of the American Medical Informatics Association, 2022, 30 (1): 120-131.
- [54] ZOU R P, LV X X, ZHAO J S. SPChain: blockchain-based medical data sharing and privacy-preserving eHealth system[J]. Information Processing & Management, 2021, 58 (4): 102604.
- [55] YANG P, GAO W, TAN Q, et al. A link-bridged topic model for cross-domain document classification[J]. Information Processing & Management, 2013, 49 (6): 1181-1193.
- [56] SOURBATI M. The digital switchover as an information society initiative: the role of public policy in promoting access to digital ICTs[J]. Telematics and Informatics, 2011, 28 (4): 295-302.
- [57] YANG J, ROJAS H, WOJCIESZAK M, et al. Why are "others" so polarized? perceived political polarization and media use in 10 countries[J]. Journal of Computer-Mediated Communication, 2016, 21 (5): 349-367.
- [58] YAQUB U, CHUN S A, ATLURI V, et al. Analysis of political discourse on twitter in the context of the 2016 US presidential elections[J]. Government Information Quarterly, 2017, 34 (4): 613-626.
- [59] XIAO C, CHOI E, SUN J M. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review[J]. Journal of the American Medical Informatics Association, 2018, 25 (10): 1419-1428.

作者简介

袁毅, 女, 博士, 教授, 博士生导师, 研究方向: 信息计量、互联网信息资源管理。

孟盈, 女, 硕士研究生, 通信作者, 研究方向: 信息组织, E-mail: 1123517193@qq.com。

Evolutionary Analysis and Frontier Identification of Interdisciplinary Research in Foreign LIS Field Based on BERT-LDA

YUAN Yi MENG Ying

(School of Economics and Management, East China Normal University, Shanghai 200062, P. R. China)

Abstract: Identifying interdisciplinary research frontier topics and analyzing evolutionary trends can help reveal the direction of interdisciplinary integration, and provide references for future innovative and breakthrough research. Firstly, based on the perspective of citation, we construct indicators to measure the interdisciplinary nature of research papers, and identify research papers with interdisciplinary nature. Secondly, we use the BERT-LDA model to identify research topics, calculate the similarity between topics using cosine similarity, and construct topic evolution paths. Finally, we construct a research frontier topic identification index system that includes topic novelty, topic growth, topic attention, and topic influence to identify interdisciplinary research topics with cutting-edge characteristics. This paper takes the Library and Information Science (LIS) discipline as an example to conduct research. The results show that interdisciplinary research topics in this field from 2004 to 2023 have gradually become more refined and in-depth, mainly focusing on information mining and knowledge discovery, Internet information behavior, and medical informatics. At present, the interdisciplinary topics of medical data model, public opinion governance and sentiment analysis, and machine learning and deep learning have cutting-edge characteristics. Research methods based on information technology and their applications in different fields have good application prospects and may become the core research topics in the future field of LIS.

Keywords: Research Frontier; Topic Evolution; Interdisciplinarity; BERT-LDA; Topic Identification; LIS; Information Science; Library Science

(责任编辑: 王玮)