

会议论文与期刊论文主题分布的演化关系研究^{*}

郭骅¹ 李庆昊¹ 夏天伊¹ 李盛庆²

(1. 河海大学商学院, 南京 211100; 2. 常熟理工学院, 苏州 215500)

摘要: 为加深对跨文献类型的学术研究主题演化的理解, 设计构建会议论文和期刊论文研究主题演化的多步骤分析框架, 包括数据准备、主题识别、相似度测量、时滞测度、演化模式分析、主路径识别和关键主题识别等。以《中国计算机学会推荐国际学术会议和期刊目录》中的A类出版物为例, 分析和呈现2007—2021年人工智能核心领域会议论文、期刊论文的研究主题演化情况。研究发现: 该领域会议论文的研究主题设置整体上领先期刊1年; 多种主题演化模式交叉作用, 扩展人工智能核心领域的科学知识; 会议论文创新促进研究方向的聚焦和知识框架的稳定; 一些较晚出现的会议主题更强烈地影响领域知识的持续构建。此外, 还分析人工智能术语的分布情况和变化趋势。

关键词: 主题识别; 主题演化; 人工智能; 会议论文; 期刊论文; 关键主题

中图分类号: G255 **DOI:** 10.3772/j.issn.1673-2286.2023.10.003

引文格式: 郭骅, 李庆昊, 夏天伊, 等. 会议论文与期刊论文主题分布的演化关系研究[J]. 数字图书馆论坛, 2023 (10): 18-28.

研究主题由单词集合表达, 旨在反映科学文献的核心内涵。主题演化表明研究主题如何随着时间的推移而变化, 包括它是否发展成熟或逐渐消亡, 是否导入知识或被拆分合并^[1]。会议和期刊关注的研究主题可能交叉影响, 对此类情况的主题演化研究还相对缺乏。研究主题如何在会议论文和期刊论文之间转移和发展? 谁处于整体上领先的位置? 不同学科领域的情况差别非常大^[2-4]。由于时间序列上的主题相互交织, 没有一个关联方案对所有情况都有效^[2], 所以本文使用了一个以创新著称的人工智能研究领域的数据集来分析这一问题。

人工智能的概念在20世纪50年代提出。作为一种通用技术, 人工智能的研究横跨了许多学科, 但是其核心属于计算机科学领域^[5]: 人工智能专注于运行通常需要人类智能的计算系统^[6], 或通过使用高级算法和模型模拟人类意识和思维过程, 赋予机器类似人类的能力^[7]。在计算机科学领域, 领先的学术会议常常吸引高质量的原创新论文, 并由元审稿人来把关评审的专业性, 这促进了

知识共享和创新^[8]。通常认为该领域的会议是创新的摇篮, 而期刊的研究成果构成了知识框架^[9]。

已有文献指出人工智能的会议论文引领了研究趋势^[10], 但是从会议到期刊的文献研究主题演化模式仍不清晰, 并且很少有研究聚焦核心技术。本文试图描绘2007—2021年人工智能核心领域的顶级出版物的研究蓝图, 达成以下目标: ①测度从会议到期刊的学术论文研究主题设置的领先-滞后程度; ②理解从会议到期刊的学术论文研究主题演化特征, 识别演化路径和发现关键主题。

1 文献综述

1.1 主题分析技术

常见的主题分析技术包括引文分析、词汇分析和主题建模分析。引文分析通过统计引用关系识别研究主

收稿日期: 2023-07-23

^{*}本研究得到国家自然科学基金项目“面向代表作评价的科研论文学术价值测度研究”(编号: 21CTQ017)资助。

题,但引文的延迟性可能会扭曲分析结果^[11]。在研究主题设置方面,影响力也并不一定表现为引用^[12]。词汇分析包括词频分析、共词分析和术语聚集。基于单词或术语的方法对词的选择很敏感,并且可能缺乏对主题的足够解释^[12]。概率主题模型将主题定义为单词和文档的分布,基于概率挖掘大规模文本中的语义,生成以单词集合表达的主题,总体上优于之前的技术^[13]。许多模型已经被应用于主题识别和表示,如LDA (Latent Dirichlet Allocation)、CTM (Correlated Topic Model) 和BTM (Biterm Topic Model) 等,尤其LDA模型已经成为语义分析的核心算法^[14]。

1.2 领先-滞后程度

领先-滞后程度计量源自引文统计,如科学论文发表和首次被引的时间差^[15]或参考文献的年龄特征^[16]。有关主题的讨论从关键词、术语,甚至特定的研究设备方面对不同文献之间的主题时滞进行分析^[16]。但是对于大多数的两组语料而言,完全相同的主题实际上并不存在;即便在同一研究领域,不同主题的领先-滞后程度也是分散的^[17],因此有必要衡量整体的时滞值。共词网络分析通过比较两个共词网络的结构相似性来确定领先-滞后关系和程度^[18]。基于LDA的方法则提供了另一条路径,即通过语义相似性确定时滞值^[19]。

1.3 主题演化分析

主题常常具有演化的特征,即一个主题影响或刺激了其后的另一个^[20]。主题演化分析大致分为两类:主题强度演化和主题内容演化。强度演化描述主题的热度变化^[21];内容演化则进一步评估主题之间的语义相似度^[22],包括构建演化图和分析演化路径。主题演化图表征了文本流的主题演化模式,是演化路径的图形化集合^[20]。由于演化路径并非全部显著和可靠,研究人员需要提取关键路径以追踪研究领域的发展轨迹^[23]。主路径的概念来自引文网络分析,在语义网络分析中得到发展^[24]。主路径搜索以遍历权重为基础,通常选择最长或权重和最高的路径^[25]。文本特征差异影响分析技术的选择,对于长跨度的科学研究主题,基于时间序列的相似度计算方法较为适用^[20]。学者最初根据频率和强度识别路径上的关键主题,近年来测量语义网络的节点中心性指标的方法逐渐流行^[26-27]。

2 数据与方法

2.1 研究设计

提出一个多步骤框架用于分析跨文献类型的研究主题演化情况,包括数据准备、主题识别、相似度测量、时滞测度、主题演化图合成、主路径识别和关键主题识别。研究步骤如下:①构建语料库,从Web of Science和计算机科学文献库DBLP (Database Systems and Logic Programming) 中收集会议论文和期刊论文,将论文标题和摘要作为研究语料,将文本语料根据出版时间按年份划分为子集;②识别各子集的研究主题,根据困惑度、一致性和JS (Jensen-Shannon) 散度指标联合确定主题数和建立概率主题模型,以解决单一指标的过拟合问题,以及论文和期刊两类子集的规模差异导致的不均衡问题;③测算从会议到期刊的论文研究主题时滞,根据余弦相似度指标衡量不同子集之间的主题相似程度,按年份构建相似度矩阵,使用匈牙利算法寻找不同年份的研究主题在会议和期刊论文之间转移的最优匹配组合,采用最小二乘法对这一组合进行线性拟合,以获得样本期的整体最优时滞值而非单一年份的最优解;④构建主题演化关系,设置演化规则,在排除弱相似性主题后,合成主题演化图;⑤识别主要研究脉络,采用穷举法获取候选路径,根据节点与源之间的累积权重识别主路径,以强调从会议产生并影响期刊论文的创新研究;⑥识别关键的会议研究主题和主题词,构建从会议到期刊的学术论文研究主题转移的分析场景,测算特定场景和业务逻辑下的主题转移概率,运用PageRank算法评价会议研究主题和主题词对期刊论文的长期和广泛的影响。

2.2 数据准备

从第五版《中国计算机学会推荐国际学术会议和期刊目录》^[28]中选取人工智能A类期刊《人工智能》(AI)、《计算机视觉国际期刊》(IJCV)、《机器学习研究杂志》(JMLR)、《IEEE模式分析与机器智能汇刊》(TPAMI),以及A类会议人工智能促进协会(AAAI)会议、国际计算语言学协会(ACL)会议、国际计算机视觉与模式识别会议(CVPR)、国际机器学习大会(ICML)、国际人工智能联合会议(IJCAI)、神经信息处理系统大会(NEURIPS)、计算机视觉国际大

会 (ICCV), 作为论文来源。《中国计算机学会推荐国际学术会议和期刊目录》将出版物/会议分为A、B、C三类, 其中A类表示极少数的顶级出版物/会议, 在历次版本更新中更具稳定性^[29]。

在选择研究材料时考虑以下因素。首先, 人工智能的核心技术属于计算机科学领域^[5, 10]。其次, 由于高度的创新性, 人工智能科学论文的标题、摘要和关键词或许并不包含常用术语^[5], 因此有必要纳入目标出版物/会议的全部论文。最后, 在2007年之后顶级期刊论文的被引频次总是比会议论文多, 表明其知识框架具有稳定性。

要充分评估主题之间的关系, 需要较长的时间跨度^[10], 因此选取2007—2021年作为样本期, 同时还考虑以下因素。首先, 在计算机科学期刊上重新发表基本相同的会议论文的情况曾经相当普遍, 这一传统在2000年之后逐渐改变^[30]。其次, 人工智能科学论文的出版经历了不止一次低谷, 最近一次低谷期是2005—2010年^[31]。最后, 该领域论文数量和平均被引频次在2005年和2006年异常下降, 其后恢复到正常水平^[10]。研究材料包括来自4种期刊的8 047篇论文和7个会议的52 841篇论文。

2.3 研究主题识别

首先提取研究论文的标题和摘要构建文本语料库, 会议和期刊语料按出版时间被分别划进15个子集; 然后进行数据预处理, 去除标点符号、剔除数字、去停用词和还原词形等; 最后利用LDA模型识别各子集的研究主题。LDA模型中文档被表示为潜在主题的随机组合, 主题被表示为单词的概率分布^[32]。单词在文档中的生成概率如式(1)所示。

$$p(w_i) = \sum_{j=1}^K p(w_i | z_i = j) p(z_i = j) \quad (1)$$

式中: $p(w_i | z_i = j)$ 表示主题 j 中单词 w_i 出现的概率, $p(z_i = j)$ 表示文档中主题 j 的比例。

为获得稳定的主题集而不是随机设置主题数, 根据一组有关模型预测能力和可解释性的指标来确定主题数: 首先, 找到困惑度^[32-33]的低值拐点和一致性^[34]的高值拐点, 将二者之间的数值范围设为潜在的取值范围。困惑度是给定模型在测试语料库中的几何平均值似的倒数, 困惑度越低表示模型泛化性能越好^[27]。一致性通过测量主题词的共现来衡量主题质量, 一致性值越高意味着主题内的单词关联性越强^[35]。因此, 该范围内的主题数值具有较好的模型预测能力和可解释性。然后, 在该范围内逐一计算每个数值对应的JS散度^[36],

使用pyLDAvis工具绘制气泡图^[37], 将满足任意两个主题气泡的重叠度小于50%条件的最大值作为主题数目。JS散度衡量两个概率分布之间的相似性, 可以进一步从全局结构评估主题模型的性能。困惑度、一致性和JS散度的计算方式如式(2)~(4)所示。主题数目核算过程见表1。

$$P(W) = \exp \left[- \frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right] \quad (2)$$

式中: P 表示困惑度, W 表示文档中单词的集合, w_d 表示集合中的单词, $p(w_d)$ 表示单词的概率, N_d 表示单词的数量。

$$C = \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{p(w_i, w_j) + 1}{p(w_j)} \quad (3)$$

式中: C 表示一致性, N 表示文档内单词总数, $p(w_i, w_j)$ 是单词 w_i 、 w_j 的共现概率。

$$D_{js}(p \| q) = \frac{1}{2} D_{kl}(p \| m) + \frac{1}{2} D_{kl}(q \| m) \quad (4)$$

式中: D_{js} 表示JS散度, p 表示数据的真实分布, q 表示数据的理论分布, D_{kl} 表示KL (Kullback-Leibler) 散度, m 为 p 和 q 的平均值。

为一些重要主题添加了人工标签: 首先采用点互信息评估单词之间的相关性, 将排名前20的主题词作为表示主题特征的潜在单词; 然后根据文档-主题分布择取每个主题对应的10篇摘要, 阅读这20个单词和10篇摘要, 归纳编写标签, 以全面反映主题的含义^[38]。点互信息的计算方式如式(5)所示。

$$I = \log_2 \frac{p(w_1, w_2)}{p(w_1) p(w_2)} \quad (5)$$

式中: I 表示单词 w_1 和 w_2 之间的点互信息。

2.4 构建语义相似度矩阵

使用余弦相似度衡量不同子集之间的主题相似程度。余弦相似度用 n 维向量空间中两个向量夹角的余弦值来衡量相似性^[39], 其计算方式如式(6)所示。

$$S(T_1, T_2) = \frac{\sum_{k=1}^n p_{1k} \times p_{2k}}{\sqrt{(\sum_{k=1}^n p_{1k}^2) (\sum_{k=1}^n p_{2k}^2)}} \quad (6)$$

式中: $S(T_1, T_2)$ 是主题 T_1 和 T_2 之间的余弦相似度, p_{1k} 和 p_{2k} 是主题在第 k 个时间片的概率, n 为样本期。余弦值越接近1, 相似度越高。最终构建了一个228×228的主题相似度矩阵。

表1 主题数目的核算过程

年 份	期刊论文主题数			会议论文主题数		
	困惑度低值拐点	一致性高值拐点	最终设定值	困惑度低值拐点	一致性高值拐点	最终设定值
2007	3	8	6	4	7	6
2008	4	5	4	4	10	6
2009	4	4	4	4	11	6
2010	3	6	6	4	9	7
2011	3	9	6	3	8	4
2012	3	7	7	4	4	4
2013	3	11	8	5	12	9
2014	3	8	5	5	12	8
2015	3	6	5	8	10	10
2016	3	9	7	8	11	10
2017	3	9	9	8	9	11
2018	3	8	6	8	12	9
2019	3	8	5	18	10	16
2020	3	9	6	18	12	16
2021	3	10	6	18	12	16

2.5 领先-滞后程度测度

定义年度主题相似度为1个会议子集和1个期刊子集的所有主题的相似度均值,以反映不同子集在整体上的相似程度。年度主题相似度矩阵的元素值为0.13~0.73。使用匈牙利算法^[40]求解所有子集的研究主题从会议向期刊转移的一对一时滞匹配组合。匈牙利算法是一种组合优化算法,其输出为最小和值的矩阵元素组合。为确保得到相似度之和最大的时滞匹配组合,对原始矩阵进行定值取反的预处理。最终得到一组研究主题时滞组合,该组合包含的元素不一定是其所在行或列的最大值,却涵盖了所有行和列的最优匹配组合。

在剔除异常值后,采用最小二乘法^[41]对最优匹配组合进行线性拟合。最小二乘法通过最小化误差的平方和寻找数据的最佳匹配函数。考虑到时间切片为整数年份,年度主题相似度矩阵为方形对称矩阵^[19],使用线性方程 $y=x+c$ 拟合,时滞值 c 为整数,具体公式如式(7)所示。

$$L_{\min}^2 = \sum_1^n \frac{(y-x-c)^2}{2} \quad (7)$$

式中: L 表示观测值与函数值的距离, y 和 x 分别为期刊和会议论文的出版年份。

2.6 构建主题演化图

主题演化图是一个加权有向图,节点表示主题,连

边表示演化关系,连边的权重表示主题相似度^[20]。当相似度大于一定阈值时认为存在演化关系。测试了不同的阈值,包括相同和不同出版场所之间的主题相似度的均值,以及均值的倍数(0.5倍、 $\sqrt{2}$ 倍、 $\sqrt{3}$ 倍和2倍)。结果显示,将阈值设置为 $\sqrt{3}$ 倍均值时可以比较清楚地显示主题演化关系。类似的,从会议到期刊的相似度阈值被设置为均值的 $\sqrt{3}$ 倍。

将演化图划分为两个区间,会议主题分布在上半区,期刊主题在下半区。会议和期刊节点标签由数字和字母组合而成,其中数字代表年份,字母C代表会议,字母J代表期刊。例如:2007C2表示根据2007年的会议论文数据识别出的第2个主题;同理,2007J2表示根据2007年的期刊论文数据识别出的第2个主题。排除了未能向期刊主题演化的会议主题和无法向会议主题溯源的期刊主题,以突出会议主题的演化路径。图中连边的宽度表示主题之间相似性,节点的高度表示主题的出度和/或入度。

2.7 演化主路径识别

为候选路径设定两个约束条件以聚焦会议论文的研究脉络:①所有候选路径的源位于上半区、汇位于下半区;②删除会议-会议的连边,使每一个会议主题节点都成为源。对所有源进行分治,以确定由会议主题引发的最重要的期刊主题演化路径。对于每个源,采用穷

举搜索的方法生成候选路径。根据候选路径上的节点与源之间的权重和来识别主路径。识别出的主路径具有以下特征：①路径上的期刊主题节点在整体上与源最相似；②除源以外，其他节点之间的关系不一定是最强的；③主路径不一定是最长路径。路径权重 U_{path} 的计算方式如式(8)所示。

$$U_{path} = \sum S(T_{source}, T_i) \quad (8)$$

式中： $S(T_{source}, T_i)$ 表示路径上的源主题 T_{source} 和路径上其他主题 T_i 之间的相似度。

2.8 关键主题识别

与带主题偏好的页面游走不同^[42]，采用带相似性偏好的主题游走的思路，使用PageRank算法^[43]识别关键主题。将探索研究主题视为一个不断循环的两步骤过程，即“方案设计+信息检索”：方案设计指研究人员分析文献并推断如何设置检索方案；信息检索指根据检索方案获取新的文献。由于信息寻求行为由认知状态介导并修改^[44]，方案设计主导了主题探索。

在网页排名中，阻尼系数 d 表示用户通过当前网页跳转到链接网页的概率，通常设置为0.85^[45]， $1-d$ 则表示通过输入网址跳转到新页面的概率。在引文网络排名中， d 被设置为0.5^[46]。在主题探索的场景中，研究人员既可能独立探索新的研究主题（不依托已有文献是不现实的，可以认为所有已知文献对新主题不产生特定影响，类似于白噪声），也可能通过寻找之前的研究脉络来设置新的主题。这一过程可以模拟为通过主题节点的连边反向查找与该主题相似的前置主题。因此将 d 定义为一个主题连接之前主题的概率，即它们之间的相似度。计算网络中所有节点对的相似度均值得到 d 为0.61，表征了样本数据的主题相似度偏好。节点 i 的排名值的表达式如式(9)所示。

$$R_i = \frac{1-d}{N} + d \sum_{j=1}^N (R_j \times S_{ij}) \quad (9)$$

式中： R_j 表示链接到节点 i 的节点 j 的排名值， S_{ij} 表示 i 与 j 之间的相似度， N 表示节点数。

3 分析结果

3.1 会议-期刊的整体时滞程度

会议-期刊主题的时滞值线性拟合图（见图1）显

示，样本期中的期刊论文主题设置在整体上比会议论文滞后1年。 $R^2=0.76$ ， $P=0.000\ 02$ ，模型得到充分解释且统计学意义显著。年度主题相似度矩阵及最优匹配组合见表2，用粗体显示年度主题相似度的最优匹配组合，每个元素所在的行和列分别表示对应的会议年份和期刊出版年份。

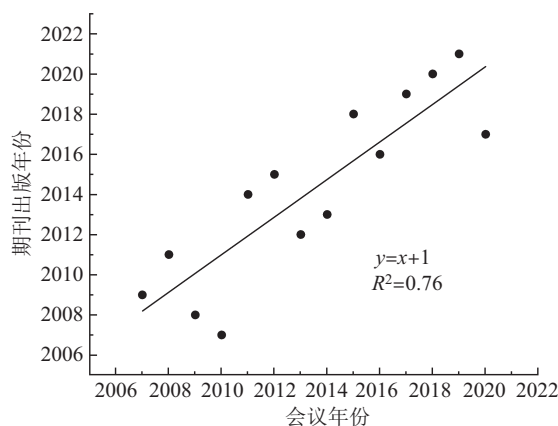


图1 会议-期刊主题的时滞值线性拟合图

3.2 会议-期刊的主题演化模式

从会议到期刊的主题演化路径共有103 803条。从图2可见，不同年份的学术会议对其后学术期刊研究主题设置产生的直接影响差别很大，如2014年有7个会议主题直接影响学术期刊研究主题，而2013年没有会议主题产生影响。图2中，会议主题节点的高度表示出度，出度越大，由会议主题引发的期刊主题越多；期刊主题节点的高度表示入度，入度越大，导向期刊主题的会议主题就越丰富。一些研究主题在演化中扮演了重要角色，如会议主题2008C2 (cluster analysis) 和2010C4 (image recognition)，以及期刊主题2017J2 (linear optimization) 和2021J1 (kernel method)；而另一些主题仅产生了微不足道的作用，如2010C3和2016J7。

有两类演化模式在网络中特别突出。一是对期刊主题产生长期影响的会议主题的演化模式（见图3）。会议主题2010C4不断分化出新的研究方向，对后续8个时间片的期刊主题产生影响。二是受到多个时间片会议主题影响的期刊主题的溯源模式（见图4）。一组不同时间片的期刊主题，即2017J2、2019J5和2021J1，交叉溯源至一组不同时间片的会议主题。虽然图示的连边宽度差别很大，但实际上在数据尺度缩放之前，所有主题相似度都在阈值以上。

表2 年度主题相似度矩阵及最优匹配组合

类 别	2007C	2008C	2009C	2010C	2011C	2012C	2013C	2014C	2015C	2016C	2017C	2018C	2019C	2020C	2021C
2007J	0.50	0.60	0.59	0.57	0.52	0.60	0.14	0.58	0.32	0.35	0.37	0.31	0.23	0.24	0.21
2008J	0.54	0.64	0.63	0.61	0.57	0.65	0.16	0.62	0.35	0.38	0.39	0.34	0.25	0.26	0.23
2009J	0.61	0.68	0.67	0.66	0.62	0.70	0.20	0.68	0.41	0.43	0.44	0.39	0.28	0.31	0.26
2010J	0.51	0.61	0.59	0.58	0.54	0.62	0.17	0.60	0.35	0.37	0.39	0.35	0.25	0.27	0.24
2011J	0.53	0.65	0.63	0.59	0.56	0.64	0.16	0.62	0.35	0.37	0.39	0.34	0.25	0.26	0.23
2012J	0.52	0.62	0.62	0.59	0.56	0.64	0.18	0.62	0.37	0.39	0.40	0.35	0.25	0.28	0.24
2013J	0.52	0.60	0.60	0.58	0.55	0.64	0.17	0.62	0.37	0.38	0.39	0.34	0.25	0.28	0.23
2014J	0.59	0.67	0.66	0.65	0.63	0.71	0.21	0.69	0.43	0.44	0.45	0.41	0.30	0.32	0.28
2015J	0.58	0.65	0.63	0.63	0.61	0.70	0.23	0.67	0.43	0.43	0.45	0.41	0.29	0.33	0.28
2016J	0.57	0.64	0.62	0.62	0.60	0.69	0.23	0.66	0.43	0.44	0.45	0.43	0.31	0.34	0.29
2017J	0.54	0.60	0.60	0.59	0.56	0.65	0.20	0.63	0.40	0.42	0.43	0.40	0.29	0.32	0.27
2018J	0.55	0.62	0.60	0.61	0.59	0.68	0.24	0.66	0.45	0.46	0.48	0.45	0.33	0.36	0.30
2019J	0.57	0.65	0.65	0.64	0.61	0.73	0.24	0.70	0.47	0.49	0.56	0.51	0.37	0.40	0.33
2020J	0.53	0.58	0.55	0.58	0.55	0.65	0.23	0.62	0.45	0.47	0.53	0.52	0.38	0.40	0.33
2021J	0.56	0.62	0.61	0.62	0.60	0.70	0.24	0.68	0.47	0.49	0.56	0.53	0.40	0.42	0.35

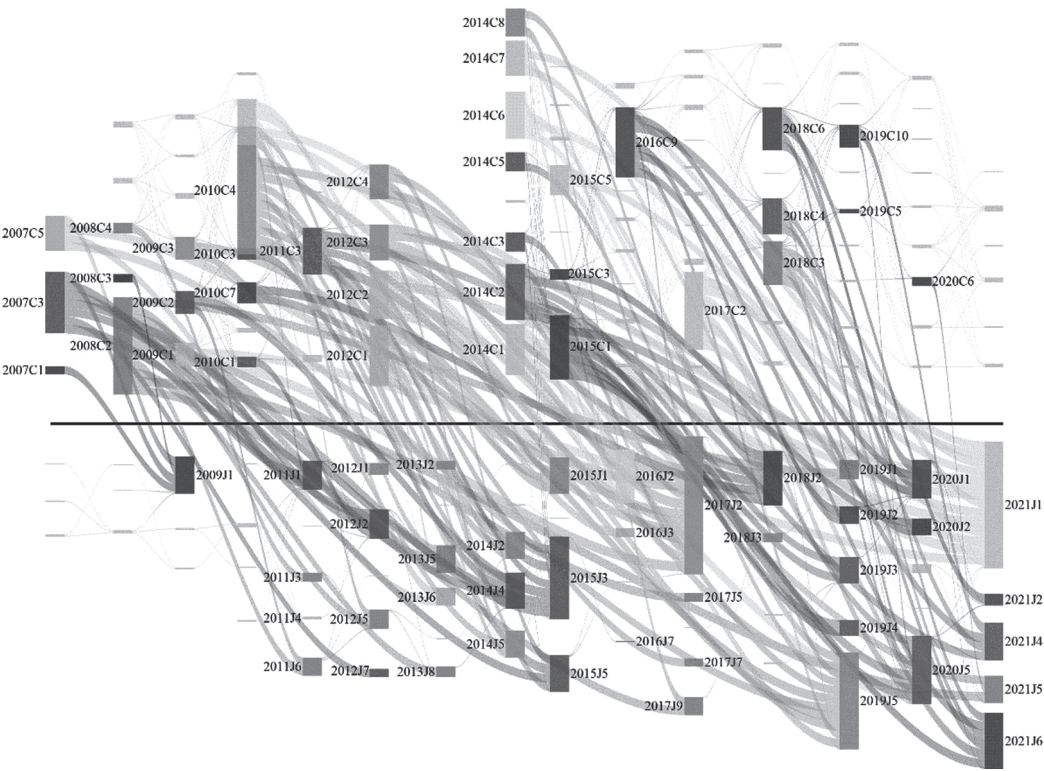


图2 人工智能核心领域从会议到期刊的研究主题演化

3.3 会议-期刊的演化主路径

共识别出35条主路径，它们都从会议主题出发到达期刊主题。主路径的形成过程中，产生主题继承、分

裂和融合的现象。例如：会议主题2008C2、2008C3分别由期刊主题2011J1和2009J1继承；期刊主题2015J5分裂为2016J2、2016J3和2016J4；会议主题2011C3、2014C3和2014C5融合为期刊主题2015J3。

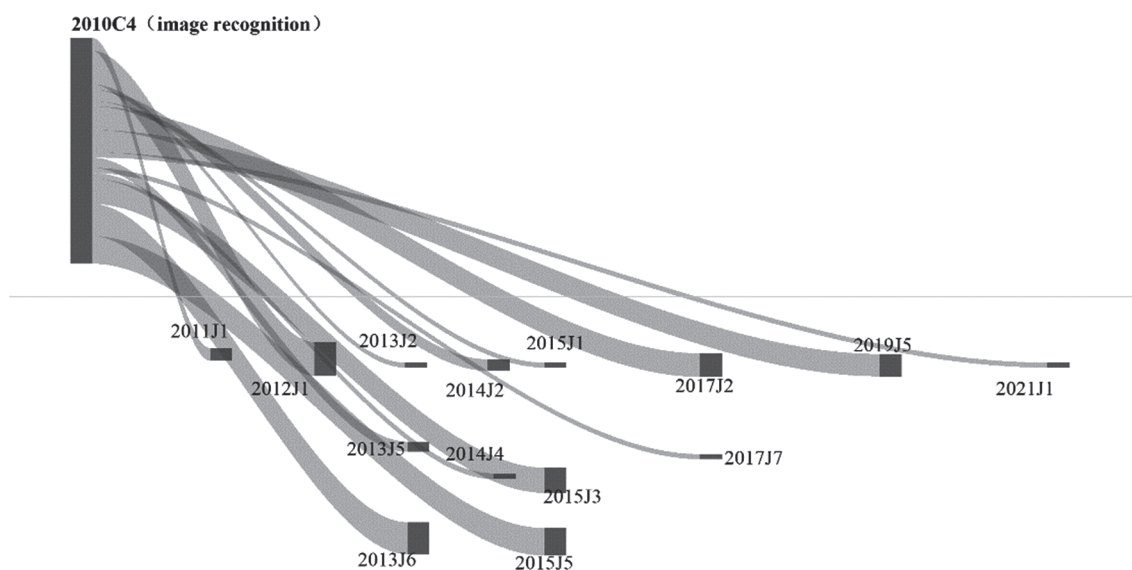


图3 对期刊主题产生长期影响的会议主题的演化模式

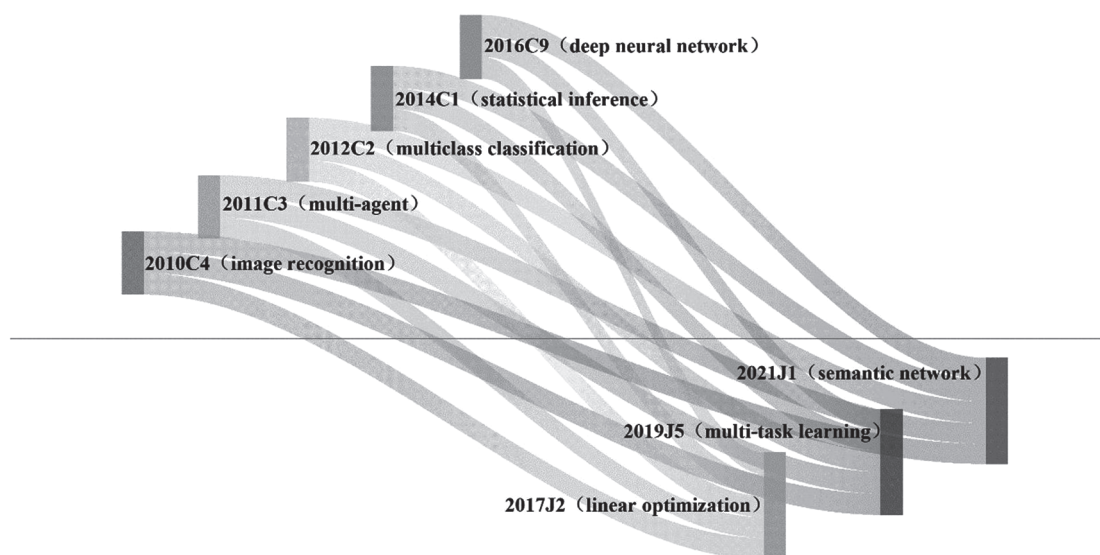


图4 受到多个时间片会议主题影响的期刊主题的溯源模式

排除连边数小于4的主路径,其余所有主路径几乎都最终汇聚到两条链路上,并且那些未能汇聚的较短路径所含的节点多数处于这两条链路上。链路1: 2017J2 (linear optimization) → 2018J2 (face recognition) → 2019J5 (multi-task learning) → 2020J1 (semantic network) → 2021J1 (kernel method); 链路2: 2017J9 (3D recognition) → 2018J3 (first-order method optimization) → 2019J1 (image segmentation) → 2020J3 (graph model) → 2021J5 (feature learning)。两条链路均处于2017—2021年,链路内部的节点具有强关联性,围绕相似的高频主题词如learn

和image展开。

3.4 关键会议主题和主题词

PageRank算法考虑了主题的持续演化,因此排名靠前的会议主题对领域知识构建产生了最大的累积效应,见表3。从经验判断,很容易误认为那些早期的会议主题较多地影响领域知识构建,但测量结果却说明一些相对晚的会议主题更强烈地影响了其后的期刊研究,例如2016C9 (deep neural network) 和2015C1 (multi-label learning)。

单词在主题中出现的概率和主题对网络的影响共同反映了单词的重要性,因此定义主题词的影响力是它在主题中出现的概率与主题的PageRank值之乘积的累加。比较影响力指标与单词在主题中的累积出现频次,可以发现表3中用“*”标记的单词的重要性

显著上升。人工智能领域研究的早期重点是形式推理和符号推理,其后陆续产生了神经网络、专家系统、智能代理等研究热潮^[47-48],主题词learn的影响力远超其他单词,印证了样本期内机器学习及其相关技术的成功应用。

表3 关键的会议主题和主题词

会议主题	PageRank	主题词(影响力)
2016C9 (deep neural network)	0.019 69	learn* (0.001 47)、neutral (0.000 36)
2008C2 (cluster analysis)	0.017 05	network (0.000 68)、training (0.000 32)
2007C3 (feature selection)	0.016 00	task (0.000 50)、distribution* (0.000 25)
2010C4 (image recognition)	0.015 54	image* (0.000 50)、adversarial* (0.000 69)
2015C1 (multi-label learning)	0.015 46	feature (0.000 38)、gradient* (0.000 64)

3.5 人工智能术语使用分析

由于基于概率的单词分布的可解释性仍然有局限性,进一步分析领域术语的使用情况。许多组织和个人提供了人工智能术语集:有的术语集规模十分有限,如联合国教科文组织的《人工智能术语表》^[49];有的术语集收录范围过于宽泛,如机器之心的人工智能词汇集项目^[50]。考虑到术语集的专业性和规模,选取维基百科的《人工智能术语表》^[51]进行分析。该术语表包含345个与人工智能、其子学科和相关领域研究有关的术语,并配套有387篇参考文献。

在语料库中搜索到269个人工智能领域术语,在会议语料集里先出现的术语有104个,在期刊中先出现的有26个。在会议中先出现的术语出现的时间平均比期刊早4.36年。聚焦其中在期刊中出现频次较高和较新的术语,可以发现convolutional neural network(卷积神经网络)和deep learning(深度学习)特别深刻地影响了样本期内的学术交流,而有关generative adversarial network(生成式对抗网络)的研究在2015年首次出现在顶级会议中,直到2019年才开始被顶级期刊出版。

使用一个术语在语料中的出现频次占所有术语出现频次的比例来表征术语受到研究者关注的程度,图5展示了典型术语受关注程度的变化趋势:一些术语的受关注程度在样本期内保持提升,如computer vision(计算视觉);一些术语在出现频次下降后,重新得到研究者的重视,如feature extraction(特征提取);还

有一些术语的受关注程度先提升再下降,如Boltzmann machine(玻尔兹曼机)在会议论文中的出现频次占比在2013年达到峰值,接着在2015年和2016年的期刊论文中也达到峰值。

4 结语

本文采用基于概率主题模型的混合方法,构建跨文献类型的研究主题演化的多步骤分析框架,分析2007—2021年人工智能核心领域的顶级学术出版物,试图加深对从会议到期刊的研究主题演化模式的理解,形成如下结论:①样本期内顶级会议的研究主题在整体上领先顶级期刊1年;②两类模式在从会议向期刊的主题演化过程中产生突出作用,即对期刊主题产生长期影响的会议主题和受到多个时间片会议主题混合影响的期刊主题,不同演化模式联合作用,融合和分裂出许多新主题;③主题演化主路径在2017—2021年汇聚成两条,表明人工智能核心领域的研究形成了稳定的方向和知识框架;④一些相对较晚的会议主题比许多早期主题更强烈地影响了人工智能领域知识的持续构建,并且主题词learn影响力最高,显示样本期的人工智能领域研究处于“机器学习→深度学习→强化学习”的蓬勃发展阶段;⑤根据人工智能术语的分布情况和变化趋势发现,术语出现的时间差更长,达到4.36年。卷积神经网络、深度学习和生成式对抗网络是样本期内会议-期刊学术交流的最重要术语。

在方法上,本文扩展了PageRank算法在主题分析

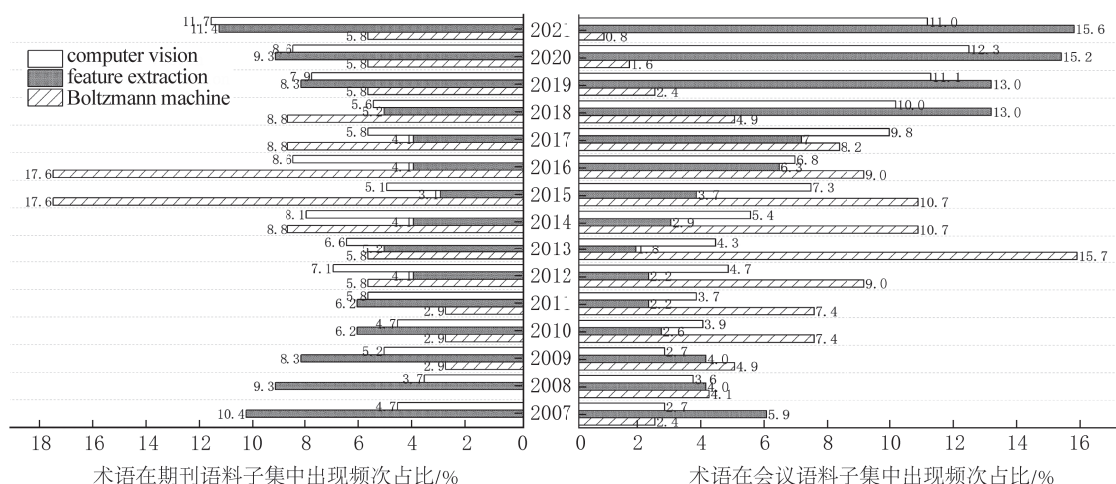


图5 典型术语在各语料集中的受关注程度

领域的应用。与其他的网络实践相比,存在两个有意的区别:①引文网络链接可以被精确跟踪,引文链的平均路径长度大约为2^[46];而主题演化网络通过相似性溯源,平均路径远超前者。②互联网链接将用户从正在浏览的网页导向另一个网页,而设定新的研究主题的行为是研究人员向已知的文献探索未知的可能的过程,两者在场景设定和参数设置方案上存在差异。

本研究存在局限和不足。首先,不同领域的主题演化模式可能存在差异,不同层次和不同阶段的研究论文也可能有不同的演化趋势,本文尚未细分领域、层次和阶段,研究结果还不能代表人工智能研究的总体趋势。其次,如果未来能够将领域术语和概率主题模型有机结合,可更加有效地识别和解释主题模型。有关跨文献类型的学术研究主题演化模式研究还有许多值得进一步探索之处。

参考文献

- [1] WANG X G, HE J, HUANG H, et al. MatrixSim: a new method for detecting the evolution paths of research topics[J]. Journal of Informetrics, 2022, 16 (4): 101343.
- [2] TU Y N, SENG J L. Research intelligence involving information retrieval-an example of conferences and journals[J]. Expert Systems with Applications, 2009, 36 (10): 12151-12166.
- [3] MONTESI M, OWEN J M. From conference to journal publication: how conference papers in software engineering are extended for publication in journals[J]. Journal of the American Society for Information Science and Technology, 2008, 59 (5): 816-829.
- [4] DA SILVA ALMENDRA V, ENĂCHESCU D, ENĂCHESCU C. Ranking computer science conferences using self-organizing maps with dynamic node splitting[J]. Scientometrics, 2015, 102 (1): 267-283.
- [5] WANG P. On defining artificial intelligence[J]. Journal of Artificial General Intelligence, 2019, 10 (2): 1-37.
- [6] CHARTRAND G, CHENG P M, VORONTSOV E, et al. Deep learning: a primer for radiologists[J]. Radiographics, 2017, 37 (7): 2113-2131.
- [7] JAKHAR D, KAUR I. Artificial intelligence, machine learning and deep learning: definitions and differences[J]. Clinical and Experimental Dermatology, 2020, 45 (1): 131-132.
- [8] YANG Z B, ZHANG S H, SHEN W J, et al. Artificial intelligence related publication analysis based on citation counting[J]. IEEE Access, 2018, 6: 52205-52217.
- [9] FIALA D, TUTOKY G. Computer science papers in web of science: a bibliometric analysis[J]. Publications, 2017, 5 (4): 23.
- [10] HO Y S, Wang M H. A bibliometric analysis of artificial intelligence publications from 1991 to 2018[J]. COLLNET Journal of Scientometrics and Information Management, 2020, 14 (2): 369-392.
- [11] JARIĆ I, KNEŽEVIĆ-JARIĆ J, LENHARDT M. Relative age of references as a tool to identify emerging research fields with an application to the field of ecology and environmental sciences[J]. Scientometrics, 2014, 100 (2): 519-529.
- [12] WANG X F, ZHANG S, LIU Y Q. ITGInsight-discovering and visualizing research fronts in the scientific literature[J]. Scientometrics, 2022, 127 (11): 6509-6531.

- [13] CURISKIS S A, DRAKE B, OSBORN T R, et al. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit[J]. *Information Processing & Management*, 2020, 57 (2) : 102034.
- [14] BLEI D M. Probabilistic topic models[J]. *Communications of the ACM*, 2012, 55 (4) : 77-84.
- [15] SENGUPTA I N, HENZLER R G. Citedness and uncitedness of cancer articles[J]. *Scientometrics*, 1991, 22 (2) : 283-296.
- [16] HICKS D, MELKERS J, ISETT K R. A characterization of professional media and its links to research[J]. *Scientometrics*, 2019, 119 (2) : 827-843.
- [17] HU B B, DONG X L, ZHANG C W, et al. A lead-lag analysis of the topic evolution patterns for preprints and publications[J]. *Journal of the Association for Information Science and Technology*, 2015, 66 (12) : 2643-2656.
- [18] CHEN X W, LI J P, SUN X L, et al. Early identification of intellectual structure based on co-word analysis from research grants[J]. *Scientometrics*, 2019, 121 (1) : 349-369.
- [19] 杨金庆, 陆伟, 吴乐艳. 面向学科新兴主题探测的多源科技文献时滞计算及启示: 以农业学科领域为例[J]. *情报学报*, 2021, 40 (1) : 21-29.
- [20] MEI Q Z, ZHAI C X. Discovering evolutionary theme patterns from text: an exploration of temporal text mining[C]//Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. 2005: 198-207.
- [21] ZHU H M, QIAN L, QIN W, et al. Evolution analysis of online topics based on ‘word-topic’ coupling network[J]. *Scientometrics*, 2022, 127 (7) : 3767-3792.
- [22] MA J, WANG L, ZHANG Y R, et al. An integrated latent Dirichlet allocation and Word2Vec method for generating the topic evolution of mental models from global to local[J]. *Expert Systems with Applications*, 2023, 212: 118695.
- [23] XU H Y, WINNINK J, YUE Z H, et al. Topic-linked innovation paths in science and technology[J]. *Journal of Informetrics*, 2020, 14 (2) : 101014.
- [24] HUMMON N P, DEREIAN P. Connectivity in a citation network: the development of DNA theory[J]. *Social Networks*, 1989, 11 (1) : 39-63.
- [25] CHEN L, XU S, ZHU L J, et al. A semantic main path analysis method to identify multiple developmental trajectories[J]. *Journal of Informetrics*, 2022, 16 (2) : 101281.
- [26] 胡吉明, 田沛霖. 文本智能计算研究的主题挖掘与演化分析[J]. *情报杂志*, 2021, 40 (4) : 139-146.
- [27] DE BATTISTI F, FERRARA A, SALINI S. A decade of research in statistics: a topic model approach[J]. *Scientometrics*, 2015, 103 (2) : 413-433.
- [28] 中国计算机学会学术工作委员会. 中国计算机学会推荐国际学术会议和期刊目录(第五版)[EB/OL]. [2023-09-01]. <https://www.ccf.org.cn/c/2019-04-25/663625.shtml>.
- [29] 中国计算机学会学术工作委员会. CCF公布推荐的国际学术会议和期刊目录[EB/OL]. [2023-09-01]. <https://www.ccf.org.cn/wqwxw/ccflwz/2011-07-22/649128.shtml>.
- [30] ZHANG Y H, JIA X Y. A survey on the use of CrossCheck for detecting plagiarism in journal articles[J]. *Learned Publishing*, 2012, 25 (4) : 292-307.
- [31] LIU N, SHAPIRA P, YUE X X. Tracking developments in artificial intelligence research: constructing and applying a new search strategy[J]. *Scientometrics*, 2021, 126 (4) : 3153-3192.
- [32] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. *Journal of Machine Learning Research*, 2003 (3) : 993-1022.
- [33] GRÜN B, HORNIK K. Topicmodels: an R package for fitting topic models[J]. *Journal of Statistical Software*, 2011, 40 (13) : 1-30.
- [34] MIMNO D, WALLACH H, TALLEY E, et al. Optimizing semantic coherence in topic models[C]//The 2011 Conference on Empirical Methods in Natural Language Processing. 2011: 262-272.
- [35] NEWMAN D, LAU J H, GRIESER K, et al. Automatic evaluation of topic coherence[C]//HLT 10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 100-108.
- [36] SIEVERT C, SHIRLEY K. LDavis: a method for visualizing and interpreting topics[C]//Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 2014: 63-70.
- [37] GitHub. API documentation: topic models (e.g. LDA) visualization using D3[EB/OL]. [2023-09-01]. <https://pyldavis.readthedocs.io/en/latest/modules/API.html>.
- [38] HAGEN L. Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models?[J]. *Information Processing & Management*, 2018, 54 (6) : 1292-1307.
- [39] JOACHIMS T. A probabilistic analysis of the Rocchio algo-

- rithm with TFIDF for text categorization[R]. Dortmund: Carnegie-mellon Univ Pittsburgh Pa Dept of Computer Science, 1996.
- [40] KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics, 2005, 52 (1): 7-21.
- [41] EDMONDS J. Maximum matching and a polyhedron with 0, 1-vertices[J]. Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics, 1965, 69B (1/2): 125-130.
- [42] DING Y. Topic-based PageRank on author cocitation networks[J]. Journal of the American Society for Information Science and Technology, 2011, 62 (3): 449-466.
- [43] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web[R]. Stanford: The Stanford University InfoLab, 1999.
- [44] XU Y J. The dynamics of interactive information retrieval behavior, part I: an activity theory perspective[J]. Journal of the American Society for Information Science and Technology, 2007, 58 (7): 958-970.
- [45] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30 (1/2/3/4/5/6/7): 107-117.
- [46] AMODIO P, BRUGNANO L, SCARSELLI F. Implementation of the PaperRank and AuthorRank indices in the Scopus database[J]. Journal of Informetrics, 2021, 15 (4): 101206.
- [47] ADAMI C. A brief history of artificial intelligence research[J]. Artificial Life, 2021, 27 (2): 131-137.
- [48] LUGER G F. A brief history and foundations for modern artificial intelligence[J]. International Journal of Semantic Computing, 2023, 17 (1): 143-170.
- [49] UNESCO. 人工智能术语表[EB/OL]. [2023-02-05]. <https://www.unesco.org/zh/articles/rengongzhinengshuyubiao-0>.
- [50] 机器之心. 人工智能术语数据库[EB/OL]. [2023-03-05]. <https://github.com/jiqizhixin/Artificial-Intelligence-Terminology-Da-tabase>.
- [51] Wikipedia. Glossary of artificial intelligence[EB/OL]. [2023-04-02]. https://en.wikipedia.org/wiki/Glossary_of_artificial_intelligence#References.

作者简介

郭骅, 男, 博士, 副教授, 硕士生导师, 研究方向: 信息智能处理。

李庆昊, 男, 硕士研究生, 研究方向: 信息智能处理。

夏天伊, 女, 硕士研究生, 研究方向: 文献计量。

李盛庆, 男, 博士, 讲师, 通信作者, 研究方向: 信息计量, E-mail: chql11@cslg.edu.cn。

Evolutionary Relationship Between the Topic Distribution of Conference Literature and Journal Literature

GUO Hua¹ LI QingHao¹ XIA TianYi¹ LI ShengQing²

(1. Business School of Hohai University, Nanjing 211100, P. R. China; 2. Changshu Institute of Technology, Suzhou 215500, P. R. China)

Abstract: The study aims to deepen the understanding of the evolution of scholarly research topics across literature types. This paper proposes a multi-step analytical framework for research topic evolution in conference and journal articles, including data preparation, topic identification, similarity measurement, time lag measurement, evolutionary pattern analysis, main path identification, and critical topic identification. Taking the Class A publications in *Directory of Recommended International Academic Conferences and Journals of the China Computer Federation* as examples, the research topic evolution of conference articles and journal articles in the core field of AI from 2007 to 2021 is analyzed and presented. The results show that the research topic setting of conference articles is overall one year ahead of journals; multiple topic evolutionary patterns intersect to expand the scientific knowledge in the field; innovations brought about by conference articles have contributed to the focusing of the research direction and the stabilization of the knowledge framework; and some of the later-emerging conference topics have more strongly impacted the continued construction of the knowledge in the field. In addition, this paper analyzes the distribution and trends of AI terminology.

Keywords: Topic Identification; Topic Evolution; Artificial Intelligence; Conference Literature; Journal Literature; Critical Topic

(责任编辑: 王玮)