

科研合作群体中的信息茧房识别及 破茧主题推荐研究

陈翔¹, 黄璐^{2,3}, 曹晓丽⁴, 任航⁵

(1. 北京理工大学管理学院, 北京 100081; 2. 北京理工大学经济学院, 北京 100081; 3. 数字经济与政策智能工业和信息化部重点实验室, 北京 100081; 4. 中国农业大学图书馆, 北京 100193;
5. 北京理工大学(珠海), 珠海 519088)

摘要 随着科研人员之间的合作关系逐渐固化, 合作群体中的信息茧房问题将抑制学科交叉融合及科研合作创新水平提升。本文提出一套面向科研合作群体的信息茧房识别及破茧主题推荐方法。一方面, 构建时间序列的作者合著-关键词语义相似度双层网络, 利用增量社区发现算法获取各年累积合著网络中的社区结构, 基于双层网络的作者和关键词的对应关系计算每个作者对应的研究主题向量, 综合考虑研究主题同质性和新颖性指标, 对科研合作群体中的信息茧房进行识别; 另一方面, 构建信息传播影响力度量模型测度作者节点的破茧潜力, 生成考虑破茧潜力排序的作者合著-关键词语义相似度双层网络, 并提出基于重启随机游走的作者研究主题推荐算法帮助科研人员破茧。最后, 选取计算机科学领域开展实证研究, 对本文方法进行有效性验证。

关键词 信息茧房; 同质性; 信息传播影响力; 双层网络; 破茧主题推荐

Research on Information Cocoon Identification and Cocoon Breaking Topic Recommendation in Research Collaboration Groups

Chen Xiang¹, Huang Lu^{2,3}, Cao Xiaoli⁴ and Ren Hang⁵

(1. School of Management, Beijing Institute of Technology, Beijing 100081; 2. School of Economics, Beijing Institute of Technology, Beijing 100081; 3. Digital Economy and Policy Intelligentization Key Laboratory of Ministry of Industry and Information Technology, Beijing 100081; 4. China Agricultural University Library, Beijing 100193; 5. Beijing Institute of Technology, Zhuhai 519088)

Abstract: As collaborative relationships among researchers become increasingly entrenched, the emergence of information cocoons within scientific collaboration groups may hinder interdisciplinary integration and limit advancements in collaborative scientific innovation. This paper proposes a method for identifying information cocoons and recommending breakthrough research topics in scientific collaboration groups. First, a time-series, two-layer network comprising co-authorship and keyword semantic similarity was constructed. An incremental community detection algorithm is applied to extract the evolving community structure in the cumulative co-authorship network over time. Each author's research topic

收稿日期: 2025-06-03; 修回日期: 2025-11-02

基金项目: 国家自然科学基金面上项目“新媒体环境下信息茧房的形成演化机理及破解策略研究”(72371026), “新兴产业创新生态系统的演化、预测和评价: 基于动态异质网络分析视角”(72274013)。

作者简介: 陈翔, 1976年生, 博士, 教授, 主要研究方向为信息茧房、推荐系统; 黄璐, 通信作者, 1984年生, 博士, 教授, 主要研究方向为科技创新管理、四链深度融合, E-mail: huanglu628@163.com; 曹晓丽, 1999年生, 博士, 馆员, 主要研究方向为科技创新管理、文献计量分析; 任航, 2001年生, 博士研究生, 主要研究方向为科技创新管理、文献计量分析。

vector was calculated based on the correspondence between the authors and keywords across a two-layer network. Information cocoons are identified by jointly considering topic homogeneity and novelty metrics. Second, an information dissemination influence measurement model is constructed to measure the potential of author nodes to break out of a cocoon. Then, a co-authorship - keyword semantic two-layer network considering the ranking of potential to break out of the cocoon is generated, and an author topic recommendation algorithm based on restarted random walk (ATR_RWR) is proposed to help researchers break out of the cocoon. An empirical analysis was conducted in the field of computer science to validate the effectiveness of the proposed method.

Keywords: information cocoon; homogeneity; information dissemination influence; dual layer network; cocoon breaking topic recommendation

0 引言

科研合作是知识创造与技术创新的重要驱动力,多样化的知识背景有助于激发协同创新潜力、拓展研究领域边界^[1-2]。然而,在现实科研网络中,受学术邻近性和合作偏好影响,研究人员更倾向于与兴趣相近或已有合作基础的学者建立联系,逐渐形成稳定而封闭的学术圈层。同时,学术服务平台的个性化推送机制进一步强化了研究者对特定主题的信息偏好^[3]。两者共同作用,加剧了科研合作群体内部信息和合作关系的同质化,导致学术“信息茧房”现象。信息茧房在科研合作中呈现双重效应:聚焦同质化主题能够促进研究专注推动研究纵深发展;而过度的同质化会固化思维,阻碍跨学科交流与合作^[3-5]。特别是在当前科技快速发展、交叉融合的背景下,破除信息茧房、促进知识跨界流动已成为推动突破性创新的迫切需求。

科研合作领域的信息茧房界定如下:科研合作群体长期聚焦同质化研究主题,导致群体信息窄化、思维固化,进而引发认知趋同、群体极化、创新能力受限的现象^[6-7]。信息茧房与研究专注度虽然在研究主题集中性方面存在一定重合,但二者在形成机制、行为性质及影响结果上存在本质差异。研究专注度是研究者基于学术兴趣和专业知识积累所形成的积极学术聚焦行为,有助于推动研究者深耕特定领域并提升学术产出质量^[8-9];而信息茧房则源于信息窄化、社会选择及网络结构固化等因素所导致的认知封闭与合作依赖,其负面效应表现为抑制跨领域知识流入、降低研究主题新颖性与多样性,进而削弱科研群体的协同创新能力。

当前,有关科研领域信息茧房的研究主要集中于质性分析,探究其形成机制或社会心理影响,缺乏从科研合作群体视角出发的系统性定量分析^[3]。现有信息茧房破茧研究通常停留在理论分析层面,鲜有研究从推荐算法角度提出可行解决方案^[10]。因

此,亟须构建一种可识别科研合作群体信息茧房并为其推荐破茧主题的系统方法。

本文提出一套基于复杂网络分析与重启随机游走算法的科研合作群体信息茧房识别与破茧主题推荐方法。一方面,构建时序作者合著-关键词语义相似度双层网络,利用增量社区发现算法识别动态合作群体,测度合著社区中研究主题同质性和新颖性指标,有效捕捉合作群体结构及研究主题的动态演化,识别信息茧房现象,并分析其成因和强弱演化特征;另一方面,基于合著社区节点的信息接收能力和发布能力构建节点信息传播影响力度量模型,对作者节点的破茧潜力进行排序,综合考虑不同破茧潜力的作者节点及其在双层网络中与关键词间的关联关系,提出基于重启随机游走的破茧主题推荐算法,提升破茧推荐效果。本文以计算机科学领域为例进行实证分析,验证所提方法的有效性,为优化科研合作与知识创新生态提供参考。

1 相关研究

1.1 科研合作群体中信息茧房的概念研究

信息茧房(information cocoons)概念最早由桑斯坦教授于2006年提出,用于描述个体在信息获取过程中因持续偏好自身兴趣而受限于狭窄信息空间^[11]。国外研究主要借助“选择性接触”理论来解释这一现象,强调个体主动过滤异质信息的过程^[12-13]。在学术领域,“信息茧房”常与“回音室”并用,两者均体现信息环境的封闭性与认知趋同^[14],但侧重点有所不同,信息茧房更关注个体基于偏好做出的主动选择^[4,12],而回音室则强调信息在封闭社交网络结构中的重复放大与传播机制^[15-16]。

在科研合作情境下,信息茧房的形成既具有一般信息行为中的同质性选择特点,又呈现独特性。共性体现在个体倾向选择与自身观点、兴趣或知识

结构一致的信息内容与交互对象。差异性主要体现在两个方面。①形成机制。一般信息行为场景中的茧房通常由算法推荐与社交互动驱动^[17]，而科研合作中的茧房依托于制度化的学术交流与合作网络。②茧房特性。一般信息行为场景中的茧房受用户行为反馈和平台机制驱动，呈现持续演化的动态特征^[14,18]；而在科研合作中，研究者的既有研究兴趣、学术背景与合作历史更易形成稳定的合作路径依赖与认知惯性^[19]，导致信息环境更具结构性与长期封闭性。这种特征使得合作网络中的知识流动趋于内向化，跨领域交流的难度进一步增加。基于此，本文聚焦“科研合作群体中的信息茧房”，旨在识别由长期聚焦同质主题导致的群体信息窄化和创新受限问题，并探讨如何通过引入新主题和跨领域合作打破信息壁垒。

1.2 信息茧房的识别与破茧方法研究

在信息茧房研究中，“识别”是揭示其存在与演化规律的前提。学者主要从两个维度开展识别研究。①用户行为视角。通过分析用户内容选择与社交互动的同质化特征来识别茧房^[14,20]。②推荐系统与算法视角。基于个体同质化、群体极化等量化指标构建评估框架，从推荐机制层面识别信息茧房^[21]。

现有信息茧房的破茧方法研究主要分为三类^[4]：①理论分析与政策性建议，从新闻传播、信息技术、情报学等跨学科视角探讨信息茧房的形成机制与干预路径；②模型与仿真分析，通过定量分析探索影响因素，并提出综合性解决思路；③推荐算法优化，针对推荐系统易推送同质内容的问题改进算法，以提升信息多样性^[22]。

然而，现有研究主要聚焦于网络用户和新媒体场景，对科研合作群体的系统性研究仍较为缺乏。在理论体系和方法路径上，从“识别”到“破茧”的系统框架尚未形成，破茧方法在不同情境下的适用性也亟待验证^[23]；尤其是科研合作情境与一般信息行为在破茧目标和技术应用上存在显著差异，使得既有方法难以直接迁移。①破茧目标差异。一般场景旨在提升用户接触异质信息的机会，强调推荐准确性与多样性的平衡；科研合作则以跨学科融合与创新能力提升为核心，更强调新研究方向的引入以及与原有研究的协调^[24]。②技术应用差异。一般场景拥有较成熟的破茧技术体系，如内容过滤优化、社交网络分析和多样性增强算法等；而科研合

作场景的破茧技术仍处探索阶段，主要聚焦于跨学科合作^[25]和学术主题推荐^[26]，尚未形成系统化方法框架。

2 研究设计

本文方法框架如图1所示，主要包括科研合作群体的信息茧房识别和面向科研合作群体破茧的主题推荐两个部分。

2.1 科研合作群体的信息茧房识别

科研合作群体的信息茧房主要体现在论文合作关系与研究内容偏好中。基于此，本文构建时间序列的作者合著-关键词语义相似度双层网络，并使用增量社区发现算法识别动态合作群体；在此基础上，测度研究主题同质性与新颖性，以捕捉群体结构及研究主题演化；最后，从“同质性增加、新颖性降低”的趋势识别信息茧房现象，并分析其产生原因和强弱演化特征。

2.1.1 作者合著-关键词语义相似度双层网络构建

本文构建的时间序列双层网络包括作者合著网络层与关键词语义相似度网络层，用于后续科研合作群体识别、信息茧房测度及破茧主题推荐研究，如图2所示。

(1) 作者合著网络构建

先对所有论文作者进行消歧处理，然后基于作者间的合著关系构建每年的累积合著网络。累积网络包括从初始年份至当年的所有作者节点，全面反映科研人员的合作情况^[27]。

(2) 关键词语义相似度网络构建

采用Word2Vec模型将关键词映射为低维稠密向量，据此构建每年的关键词语义相似度网络。Word2Vec模型结构简单，无须复杂的预处理和微调，适合资源有限的环境，更适用于本文数据条件^[28]。本文从论文标题与摘要中提取关键词，并使用在文献计量研究中表现更优的Skip-Gram模型进行训练^[29]。随后，基于关键词向量间的余弦相似度构建关键词语义相似度网络。

(3) 作者合著-关键词语义相似度双层网络构建

基于每年作者与关键词间的关联关系，构建时间序列的作者合著-关键词语义相似度双层网络。其中，将关键词在作者所有论文中出现的总频次占作者所有关键词出现频次之和的比例作为作者合著网络层与关键词语义网络层间的连边权重。

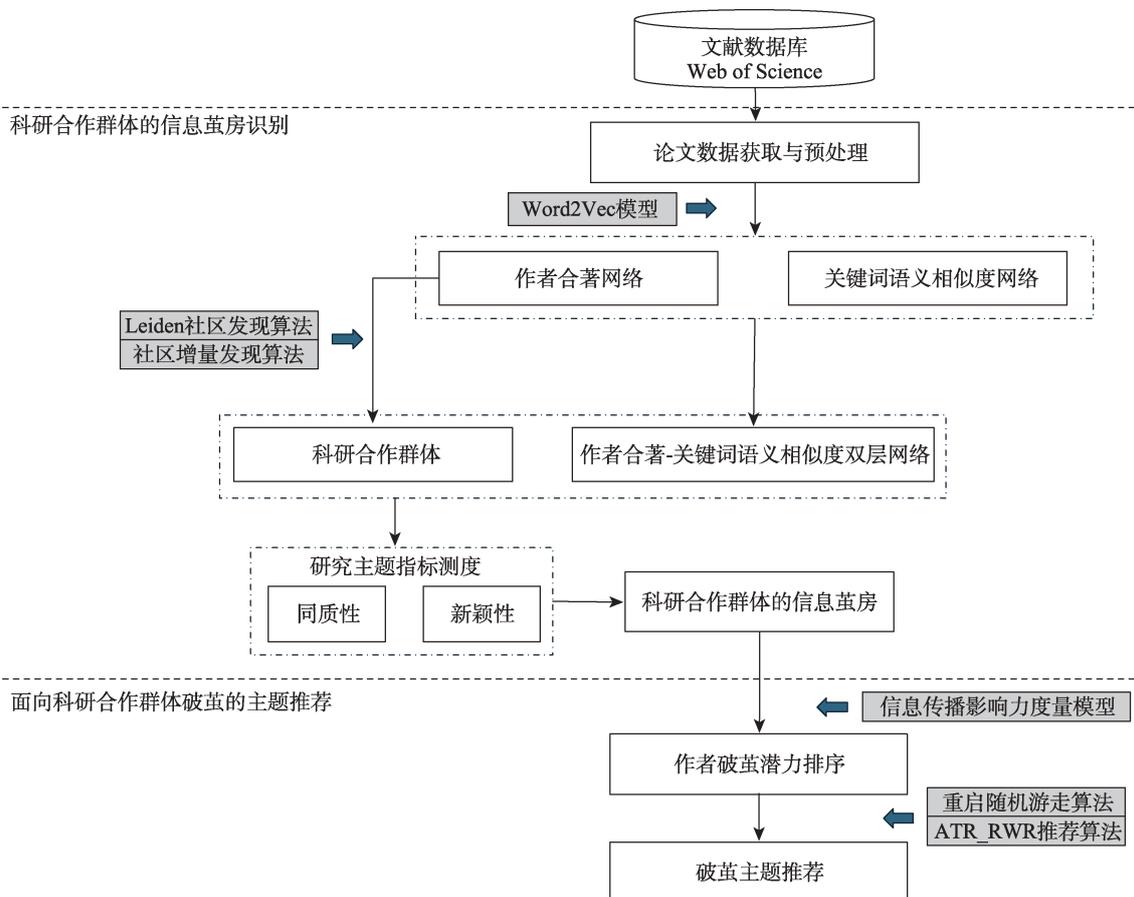


图1 科研合作群体中的信息茧房识别及破茧主题推荐方法框架

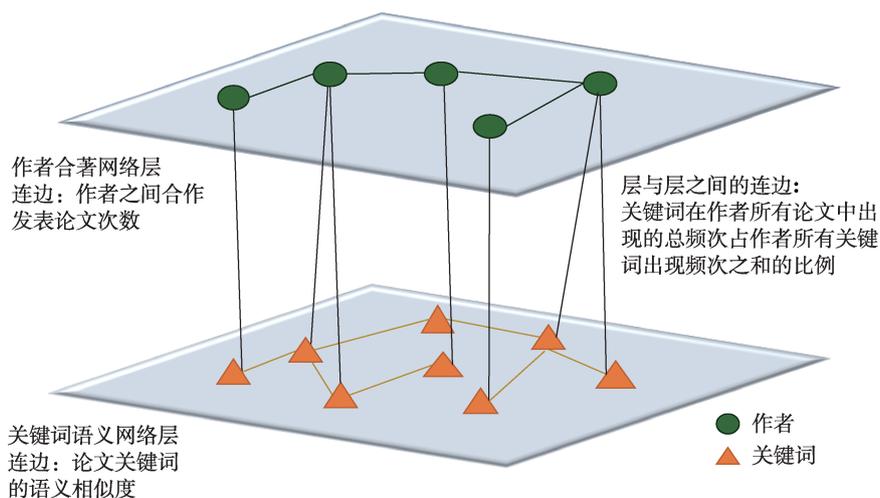


图2 作者合著-关键词语义相似度双层网络

2.1.2 基于增量社区划分的科研合作群体识别

社区划分有助于识别合著网络中研究领域相近的作者群体。首先,利用Leiden算法对初始年度合著网络进行社区划分,以确定基准社区结构。Leiden是一种基于模块度的社区发现算法,适用于

节点和连边分布不均的大规模合著网络,相较于传统算法,具有划分社区结构更紧密、避免非连通社区、计算效率高等优势^[30-31],能够获得稳健的初始科研合作群体。其次,本文采用增量社区发现算法对第二年及之后的累积合著网络进行动态调整,得

到各年累积的合著社区划分结果,可以有效捕捉科研合作群体结构及其动态演化。增量社区发现算法是一种动态网络的社区划分方法,能够根据网络结构随时间的变化更新社区划分,依据模块度增益实现社区结构的实时优化^[32],从而有效捕捉科研合作群体的动态演化。

2.1.3 基于研究主题同质性和新颖性的信息茧房识别

研究主题同质性增加是科研群体信息茧房的显著特征之一^[11],能反映科研群体信息窄化及群体极化的问题。大量研究指出,科研人员通常基于自身研究兴趣和领域选择合作伙伴及研究方向^[19,33-35],形成相对封闭的知识结构,使群体内部研究主题高度同质^[36],进一步加剧了“群内同质”和“群间异质”的分化^[6]。研究主题新颖性降低是科研群体信息茧房的另一显著特征,反映出科研群体内部创新能力不足的问题。信息茧房会限制研究者接触新颖信息^[14],导致合作网络的新颖度降低^[3],最终削弱科研群体的创新能力。

因此,本文构建研究主题同质性与新颖性指标,分别刻画科研合作群体的信息窄化与创新不足的特征,以识别信息茧房并进一步分析茧房产生原因和强弱演化特征。鉴于学界尚无统一的科研合作群体信息茧房量化标准,本文参考 Uzzi 等^[37]的研究,基于合著社区的研究主题同质性与新颖性的长期动态变化趋势进行判定:若同质性持续上升且新颖性下降,则表明信息茧房逐渐形成;相反,若同质性下降且新颖性增加,则说明信息茧房现象可能减弱或被打破。因此,本文不采用固定阈值判定标准,而是基于研究主题同质性和新颖性指标的长期动态变化趋势来识别科研合作群体中的信息茧房。

首先,基于作者合著-关键词语义相似度网络和合著社区划分结果,对作者关联的关键词向量加权平均,得到不同作者各年份的研究主题向量^[38]。第 t 年所有作者的研究主题向量集合 V_t 公式为

$$V_t = \{v_t(a_1), \dots, v_t(a_i), \dots, v_t(a_M)\} \quad (1)$$

其中, $v_t(a_i)$ 表示第 t 年第 i 个作者的研究主题向量; T 为总年数, $1 \leq t \leq T$; M 表示第 t 年作者总数, $1 \leq i \leq M$ 。

其次,参考 Sacharidis^[39]的研究,计算各社区内研究主题向量的两两余弦相似度并取平均值,将其作为群体研究主题同质性指标。第 x 个合著社区 C_x 的研究主题同质性 TH_x 的计算公式为

$$TH_x = \frac{1}{M(M-1)} \sum_{i \neq j} \text{sim}(a_i, a_j) \quad (2)$$

其中, $\text{sim}(a_i, a_j)$ 表示第 x 个合著社区 C_x 中作者 a_i 和作者 a_j 的研究主题向量余弦相似度。

最后,对比相邻两年合著社区对应的累积关键词,获得新一年的新关键词集合。参考 Sacharidis^[39]的研究,对社区内每个新关键词,计算其与上一年社区全部关键词的相似度,取相似度平均值的倒数作为该新关键词的新颖度。合著社区中新研究主题的新颖性指标则为所有新关键词的新颖度平均值。第 x 个合著社区 C_x 的新研究主题新颖性 TN_x 计算公式为

$$TN_x = \frac{1}{N_{\text{new}}} \left[\sum_{w_{\text{new}} \in C_x} \frac{N_{\text{old}}}{\sum_{w_{\text{old}} \in C_x} \text{sim}(w_{\text{new}}, w_{\text{old}})} \right] \quad (3)$$

其中, $\text{sim}(w_{\text{new}}, w_{\text{old}})$ 表示第 x 个合著社区 C_x 中新关键词 w_{new} 和旧关键词 w_{old} 的向量余弦相似度; N_{new} 和 N_{old} 分别表示第 x 个合著社区 C_x 中新关键词总数和旧关键词总数。

2.2 面向科研合作群体破茧的主题推荐

科研合作群体破茧可从提升研究主题新颖性或降低研究主题同质性入手。一方面,通过引入新主题增强群体的信息多样性和新颖性;另一方面,通过促进跨领域或跨社区合作打破知识封闭、降低主题同质化程度。本文从“引入新主题”视角构建破茧机制,通过作者破茧潜力排序促进信息流入,并结合双层网络结构提出破茧主题推荐算法,为科研人员推荐有助于突破信息茧房的研究主题。

2.2.1 基于信息传播影响力的作者破茧潜力排序

科研合作群体的破茧关键在于识别具备较强外部信息吸收与内部扩散能力的核心作者节点。作者在合著网络中既是信息接收者又是信息发布者,其破茧潜力取决于两种能力的综合表现^[40]。基于此,本文从两个维度构建信息传播影响力度量模型^[41-42]。

①信息接收能力。衡量作者与社区外节点的连接强度,反映其吸收外部知识的广度与深度,接收能力越强,其所能引入的外部信息越丰富,越有助于降低同质性。

②信息发布能力。衡量作者在社区内部的扩散效率,反映其在群体中的影响力及推动知识共享的能力。基于上述两种能力,本文构建的信息传播影响力度量模型如下文所述。

首先,借鉴中介中心性的概念衡量作者节点的

信息接收能力。基于合著网络, 计算归一化后的节点与社区外节点连边数 S_i^{out} , 以及连接社区外节点的中介能力 B_i^{out} , 并基于熵权法对连边数和中介能力进行权重分配, 度量节点 a_i 的信息接收能力 RC_i 。具体计算公式为

$$B_i^{\text{out}} = \frac{1}{(M^{\text{in}} - 1) \times M^{\text{out}}} \sum_{\substack{f \neq i \neq t \\ f \in C^i, t \notin C^i}} \frac{n_{ft}^i}{g_{ft}} \quad (4)$$

$$\text{RC}_i = W_s^{\text{out}} \times S_i^{\text{out}} + W_B^{\text{out}} \times B_i^{\text{out}} \quad (5)$$

其中, $f \in C^i$ 表示节点 a_f 属于节点 a_i 所在社区; $t \notin C^i$ 表示节点 a_t 不属于节点 a_i 所在社区; n_{ft}^i 表示经过作者节点 a_i 且为最短路径的路径数量; g_{ft} 表示连接作者节点 a_f 和作者节点 a_t 的最短路径的数量; M^{in} 表示节点 a_i 所在社区节点总数; M^{out} 表示节点 a_i 所在社区外的节点总数; W_s^{out} 和 W_B^{out} 分别为使用熵权法计算得到的连边数 S_i^{out} 和中介能力 B_i^{out} 的权重。

其次, 采用类似的方法测度作者节点的信息发布能力。本文计算归一化后的节点与社区内节点连边数 S_i^{in} , 以及连接社区内节点的中介能力 B_i^{in} , 然后基于熵权法对连边数和中介能力进行权重分配, 度量节点 a_i 的信息发布能力 PC_i , 计算公式为

$$B_i^{\text{in}} = \frac{2}{(M^{\text{in}} - 1)(M^{\text{in}} - 2)} \sum_{\substack{f \neq i \neq t \\ f, t \in C^i}} \frac{n_{ft}^i}{g_{ft}} \quad (6)$$

$$\text{PC}_i = W_s^{\text{in}} \times S_i^{\text{in}} + W_B^{\text{in}} \times B_i^{\text{in}} \quad (7)$$

其中, W_s^{in} 和 W_B^{in} 分别为使用熵权法计算得到的连边数 S_i^{in} 和中介能力 B_i^{in} 的权重。

当节点仅与社区内或社区外节点存在连边时, 本文认为该节点很难传播具有新颖性的研究主题, 对信息茧房破茧可能没有作用。因此, 将节点信息接收能力值 RC_i 和发布能力值 PC_i 相乘, 得到作者节点 a_i 的信息传播影响力 CI_i , 计算公式为

$$\text{CI}_i = \text{RC}_i \times \text{PC}_i \quad (8)$$

最后, 基于测度的信息传播影响力, 对各社区节点分别进行破茧潜力排序; 依据该排序更新双层网络中的合著网络, 助力作者获取新研究主题。在社区内, 破茧潜力大的节点优先接收外界信息, 并向其他节点传递。将社区内的无向边转化为破茧潜力较小节点指向较大节点的有向边, 同时根据合作频次调整权重。社区间的无向边仍然保持, 但权重调整为原有权重与所有社区节点连边权重平均值的比值。得到更新后的第 $T-1$ 年 (T 为总年数) 的作者合著-关键词语义相似度双层网络, 记为 $G = G_{aa} \cup G_{aw} \cup G_{ww}$ 。其中, G_{aa} 表示有向合著网络,

G_{aw} 表示作者-关键词双层网络, G_{ww} 表示关键词语义相似度网络。双层网络对应的邻接矩阵 R 为

$$R = \begin{bmatrix} R_{aa} & R_{aw} \\ R_{wa} & R_{ww} \end{bmatrix} \quad (9)$$

其中, R_{aa} 表示合著网络对应的邻接矩阵; R_{aw} 表示双层网络对应的邻接矩阵; R_{wa} 表示 R_{aw} 的转置矩阵; R_{ww} 表示关键词语义相似度网络对应的邻接矩阵。

2.2.2 基于重启随机游走的破茧主题推荐

在破茧主题推荐中, 本文综合考虑作者破茧潜力及其在双层网络中与关键词间的关联关系, 为作者推荐跨社区主题和基于目标作者的外层关联主题, 以促进多样化知识流入, 打破信息茧房。已有研究表明, 加强不同社区间的连接可有效减弱信息茧房^[13]。同时, 目标作者外层关联节点所涉及的主题也能为其提供新的研究方向, 推动突破认知边界, 实现主题破茧。

为实现上述目标, 本文在重启随机游走 (random walk with restart, RWR) 基础上引入破茧概率, 提出基于重启随机游走的作者研究主题推荐算法 (author topic recommendation_random walk with restart, ATR_RWR), 为科研人员推荐跨社区主题和基于目标作者的外层关联主题。RWR 在游走过程中设置“重启”机制, 可根据节点偏好灵活调整游走方向, 既能捕捉网络的全局结构, 又能保持与目标节点的关联性^[43-44]。其优势体现在: ①跨社区跳转能力强, 多次重启可跨越既有合作网络的结构边界, 帮助作者获取社区外关键词; ②能捕捉外层关联关系, RWR 在游走中可沿着“从破茧潜力小到大”的路径前进, 识别外层作者所关联的新主题^[45], 拓展目标作者的研究领域, 抑制研究趋同。因此, ATR_RWR 能同时吸收跨社区信息与外层作者知识, 为科研人员提供多样性和跨社区的研究主题, 进而帮助合作群体突破信息茧房。

在随机游走过程中, 当目标作者节点到达任一作者节点后, 下一步继续向作者节点移动会比向关键词节点移动更快跳出现有群体, 更有可能实现茧房突破, 据此, 引入破茧概率 $\alpha \in (0, 1)$ 。从任一作者节点出发, 可以沿 G_{aa} 或 G_{aw} 游走——以概率 α 在作者有向合著网络 G_{aa} 上进行破茧游走, 或以概率 $1 - \alpha$ 在双层网络 G_{aw} 上游走。基于此, 可得到 ATR_RWR 算法的破茧状态转移概率矩阵, 即

$$\mathbf{R}' = \begin{bmatrix} \alpha \mathbf{R}_{aa} & (1 - \alpha) \mathbf{R}_{aw} \\ (1 - \alpha) \mathbf{R}_{wa} & \mathbf{R}_{ww} \end{bmatrix} \quad (10)$$

ATR_RWR 算法对群体信息茧房中的每个目标作者节点分别进行游走迭代, 其过程可以表示为

$$\mathbf{P}^{t+1} = (1 - \lambda) \mathbf{R}' \times \mathbf{P}^t + \lambda \mathbf{P}^0 \quad (11)$$

其中, $\lambda \in (0,1)$ 是重启概率, 表示在游走中每一步重新回到目标作者节点的概率值; 列向量 \mathbf{P}^0 表示重启动向量; 列向量 \mathbf{P}^t 、 \mathbf{P}^{t+1} 分别表示迭代第 t 次和第 $t+1$ 次时所有节点与目标作者节点之间亲和度的概率分布。当 $|\mathbf{P}^{t+1} - \mathbf{P}^t| < 10^{-6}$ 时, 认为该迭代过程达到稳定状态, 得到稳态概率分布 \mathbf{P} 。基于该分布中关键词与目标作者的偏好关联度, 本文对所有关键词从大到小进行排序, 为目标作者推荐其所在社区外排序靠前的关键词。

3 实证研究

计算机科学涵盖计算机科学理论、计算机体系结构、网络与信息安全、人工智能等多个子领域, 领域内作者合作关系多样^[46], 有助于识别相对稳定的合作群体, 进而分析科研合作中的信息茧房现象。同时, 该领域知识迭代迅速、论文产出规模大, 可以有效捕捉学者研究主题演化^[47]。因此, 本文选取计算机科学领域的论文数据进行实证研究。

3.1 计算机科学领域科研合作群体的信息茧房识别

①参考 Yu 等^[48]和 Zhang 等^[49]的研究, 选取中国计算机学会推荐的 A 类计算机科学领域国际期刊和学术会议来构建论文检索式, 从 Web of Science 核心合集数据库下载 2014—2023 年的英文论文数据, 包括作者、标题、摘要、作者关键词和年份等记录, 共计 114573 条。②使用 ItgInsight^①文本挖掘软件对数据进行预处理, 即完成作者消歧与合作关系提取, 构建各年累积合著网络; 对论文标题、摘要及作者关键词分词后, 提取每年词频及长度至少为 2、术语度最大的 3000 个关键词作为核心关键词集合。③使用训练好的 Word2Vec 模型生成词向量, 其中词向量维度设置为 100、窗口大小为 5。④基于主题向量间的余弦相似度以及作者与核心关键词间的关联关系, 构建 2014—2023 年的作者合著-关键词语义相似度双层网络, 包含 6305 个作者和 7423 个

核心关键词。

接下来, 采用 Leiden 算法对 2014 年合著网络进行社区划分, 得到作者个数均大于 30 的 30 个社区, 并按作者数量从大到小依次编号为 G1~G30。为确保社区初始划分结果的可靠性, 本文从参数设定与随机稳健性两个方面进行验证。

(1) 分辨率参数选择。分辨率取值会直接影响社区划分的颗粒度。本文在 2013—2014 年的初始合著网络上测试不同分辨率取值, 如图 3 所示, 当分辨率为 1.0 时, 模块度最高, 划分质量最佳。因此, 将分辨率设置为 1.0。

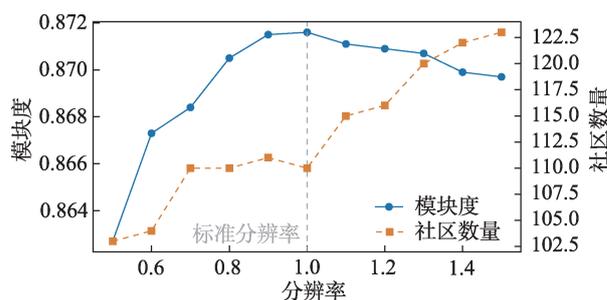


图3 Leiden算法分辨率参数选择

(2) 随机稳健性验证。为检验 Leiden 算法在随机初始化条件下的鲁棒性, 本文在分辨率为 1.0 的条件下, 设置不同随机种子多次运行 Leiden 算法, 并计算每次划分结果与基准划分 (seed=0) 之间的归一化互信息 (normalized mutual information, NMI)。NMI 能在标签排列不唯一的情况下衡量不同划分结果间的一致性^[50], 是检验社区发现鲁棒性的常用指标^[51]。如图 4 所示, 所有 NMI 值均高于 0.90, 平均值为 0.94, 表明算法划分结果对随机初始化不敏感, 具有良好的稳健性。

基于上述验证结果, 本文将 2014 年的划分结果

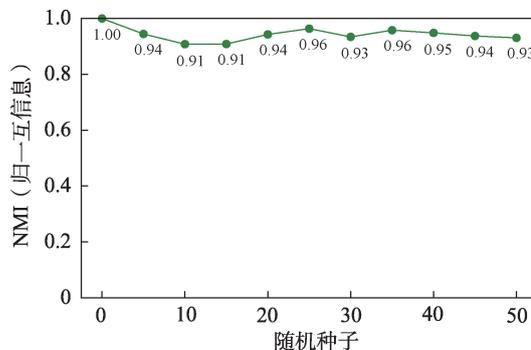


图4 Leiden算法稳健性分析

① ItgInsight 是一款科技文本挖掘和可视化分析工具, 详情见官方网站 itginsight.com

作为初始社区，并据此对2015—2023年的累积合著网络开展增量社区发现，将所有节点动态加入既有的30个社区。一般来说，模块度在0.3~0.7表明社区划分效果较好。在本文中，2015—2023年各年的模块度均大于0.3，说明增量社区划分效果较好。合著社区各年节点数分布情况如表1所示。

由表1可见，各社区的最大、最小、平均节点数均随年份增长而不断上升，表明网络规模持续扩大。同时，社区节点数标准差逐年增大，说明社区规模差异逐渐扩大，即部分社区节点数快速增长，而部分社区保持相对稳定或增长缓慢。

在构建作者合著-关键词语义相似度双层网络

后，本文为每位作者生成对应的研究主题向量，并基于余弦相似度计算各个社区的研究主题同质性和新颖性。其中，新颖性因需对比前一年的关键词，故从2015年开始计算。为精准刻画两类指标的动态变化，本文同时采用平均值与“中位数+置信区间”两种方式分析数据分布。其中，平均值反映整体趋势，中位数能降低极端值影响，而置信区间能揭示指标波动的统计稳定性。2014—2023年合著社区研究主题同质性和新颖性的平均值和中位数（95%置信区间）变化趋势如图5和图6所示。其中，图5b和图6b中的实线表示指标的中位数数值，阴影部分表示中位数的95%置信区间。

表1 合著社区各年节点数分布情况

	2014年	2015年	2016年	2017年	2018年	2019年	2020年	2021年	2022年	2023年
最大节点数	112	217	254	299	291	297	296	304	285	272
最小节点数	17	31	41	55	69	84	93	101	116	121
平均节点数	44.23	82.9	106.67	134.77	150.8	174.13	189.4	199.53	203.83	203.87
节点数标准差	20.08	41.32	47.94	55.54	57.05	58.6	60.63	60.34	53.03	45.65

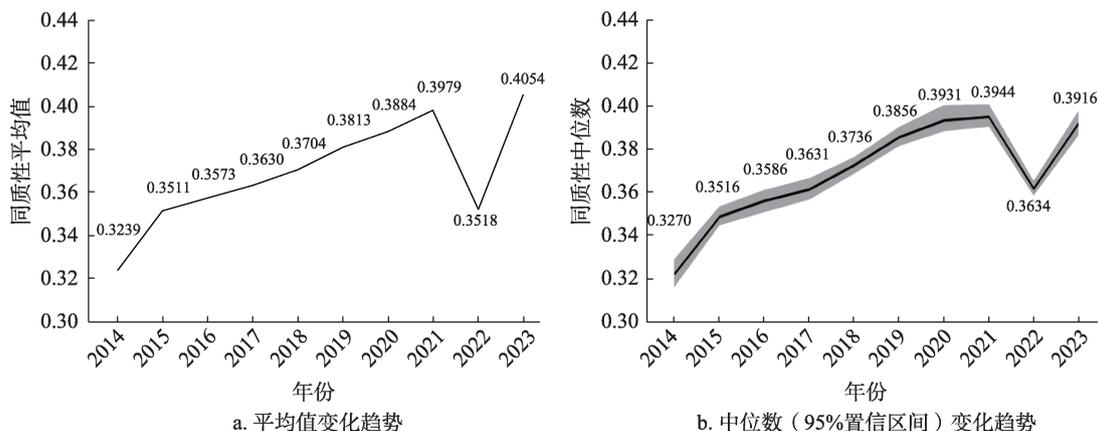


图5 2014—2023年合著社区研究主题同质性平均值和中位数（95%置信区间）变化

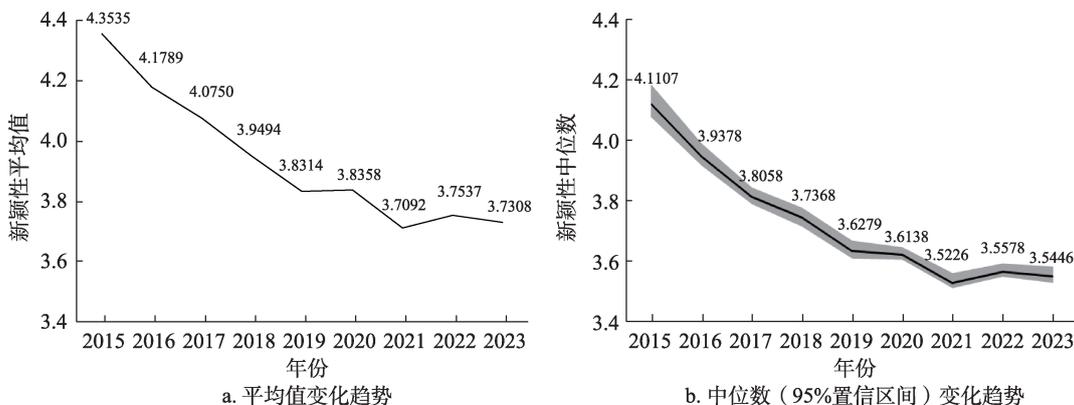


图6 2015—2023年合著社区研究主题新颖性平均值和中位数（95%置信区间）变化

图 5 和图 6 显示, 平均值分析与“中位数+置信区间”分析结论一致。2014—2021 年, 研究主题同质性持续上升, 而新颖性持续下降且置信区间收窄, 表明合作群体的研究内容逐渐趋同, 创新能力减弱, 信息茧房现象逐步显现。2021—2022 年, 同质性平均值与中位数分别下降 11.59% 和 7.86%, 但置信区间未显著扩大, 说明部分社区引入了新研究主题, 使同质性暂时下降; 同期新颖性略有提升, 表明信息茧房现象减弱。2022—2023 年, 同质性再次回升, 平均值较 2022 年提升 15.24%, 中位数提

升 7.76%; 新颖性略有下降 (平均值和中位数分别降低 0.61% 和 0.37%), 表明群体又重新聚焦于同质主题, 信息茧房现象再次加剧。

进一步地, 本文结合各合著社区的关键词社区数量及主题随时间的演化情况, 分析科研合作群体信息茧房的形成原因及强弱变化趋势。具体而言, 基于各年同质性平均值计算相邻年份变化比例, 并分别计算作者数量居前的 G1~G3 合著社区同质性和新颖性的逐年变化比例 (表 2 和表 3), 以揭示不同社区信息茧房变化的差异性。

表 2 2014—2023 年的同质性变化比例 (部分社区)

单位: %

	2014— 2015 年	2015— 2016 年	2016— 2017 年	2017— 2018 年	2018— 2019 年	2019— 2020 年	2020— 2021 年	2021— 2022 年	2022— 2023 年
整体	8.4	1.8	1.6	2.0	2.9	1.9	2.4	-11.6	15.2
合著社区 G1	7.8	0.2	1.3	2.3	1.3	1.6	2.2	-13.1	16.7
合著社区 G2	7.4	1.7	2.7	2.8	1.0	3.9	1.9	-12.4	16.1
合著社区 G3	9.5	6.1	1.7	2.9	3.1	2.7	2.3	-12.7	17.2

表 3 2015—2023 年的新颖性变化比例 (部分社区)

单位: %

	2015— 2016 年	2016— 2017 年	2017— 2018 年	2018— 2019 年	2019— 2020 年	2020— 2021 年	2021— 2022 年	2022— 2023 年
整体	-4.0	-2.5	-3.1	-3.0	0.1	-3.3	1.2	-0.6
合著社区 G1	4.0	-1.6	0.0	-5.8	0.3	-5.3	1.6	-2.5
合著社区 G2	-1.8	-3.6	-1.6	-5.9	0.9	-4.0	0.0	-4.4
合著社区 G3	-1.5	-2.7	-3.1	-2.6	-3.2	-2.8	1.1	-0.3

由表 2 和表 3 可知, 新颖性变化趋势与同质性变化趋势基本相反, 表明使用这两个指标来解释茧房强弱的变化具有可信度。2015—2021 年, 同质性整体保持在 2% 左右的小幅波动, 而新颖性在 2016—2021 年显著下降; 2021—2022 年和 2022—2023 年的同质性出现剧烈波动, 而新颖性则在 2021—2022 年回升, 反映出同质性和新颖性之间的反向关系。整体变化趋势与各合著社区基本保持一致, 因此, 可进一步从单个社区层面对信息茧房产生和强弱演化原因进行分析。

以合著社区 G1 为例, 主要存在以下 4 类变化机制。

(1) 2014—2015 年, 研究主题同质性大幅增长, 作者累积关键词充分反映研究方向。2015 年合著社区 G1 研究主题同质性增大比例仅低于 2023 年, 为 7.8%; 研究主题数量从 15 个增加至 23 个, 新增主题包括目标检测、目标识别、优化理论与算法等。由于初始 2014 年作者关键词较少, 未能全面反映研究兴趣, 群体内作者研究主题差异较大, 同质

性较小; 而 2015 年累积关键词能更全面地表征作者真实的研究方向, 群体内作者研究主题差异为正常水平, 同质性较大。

(2) 2015—2021 年, 信息茧房现象逐步加剧, 合作结构趋于固化, 缺乏新主题加入。合著社区 G1 在此期间的研究主题同质性逐年上升, 增幅在 0.2%~2.3%; 新颖性持续下降, 降幅在 0.3%~5.8%, 与合作关系稳定化相一致。虽然合著社区 G1 在 2015—2021 年分别有 23、23、26、30、30、35、33 个研究主题, 部分年份主题个数有所增加, 但新增主题由既有关键词重新组合形成, 如 2018 年新增主题“度量学习”中的大部分关键词在 2017 年已经存在, 并未带来实质性新内容。同时, 部分年份主题数量反而减少, 作者之间研究方向重叠度提升, 研究主题同质性增大, 导致信息茧房现象逐渐突出。

(3) 2022 年, 信息茧房减弱, 新研究主题进入社区。2022 年合著社区 G1 的研究主题同质性下降 13.1%, 新颖性增加 1.6%。结合语义网络社区划分结果可知, 同年研究主题数量从 2021 年的 33 个增加

至36个,新增“知识表示与推理”“可解释性人工智能”“并行与分布式计算”研究主题,为社区引入了人工智能快速发展背景下的新研究方向,从而增加了研究多样性,减弱了信息茧房。可以认为,新研究主题的出现是科研合作群体突破信息茧房的关键。

(4) 2022—2023年,信息茧房再度加剧,主题融合导致研究方向再度集中。2023年合著社区G1的研究主题同质性增幅最大,为16.7%,新颖性下降2.5%。这一年的研究主题由36个减少至35个,新关键词与原来属于不同主题社区的关键词均存在相似性,导致部分主题被合并,社区研究方向重新收缩,加剧信息茧房。

3.2 面向计算机科学领域合作群体破茧的主题推荐

为进行科研合作群体破茧策略的实施和验证,本文基于2022年数据对2023年作者进行主题推荐,并利用2023年作者实际科研合作数据进行验证。首先,利用信息传播影响力度量模型对节点破茧潜力进行排序,并以作者在2023年新研究主题的数量作为其“携带外部信息能力”的代理指标,以评估模型有效性。信息传播影响力越大,作者越可能在下一年度引入更多新主题,从而推动合作群体突破信息茧房。本文统计了31个合著社区中信息传播影响力排名前10位的作者在2023年的新研究主题数量,以及占有新主题的比例,详见附表1。进一步地,为便于深入分析,选取作者数量最多的3个合著社区G1~G3,列出其节点总数、信息传播影响力排序前10位的作者,以及2023年新主题个数和占比,如表4所示。

当作者兼具较强的信息接收能力与发布能力时,其信息传播影响力更大;若其中一项能力偏弱,整体影响力明显下降,说明模型能够有效平衡两类能力、识别真实具有破茧潜力的关键节点。此外,由表4可知,在合著社区G1~G3中,信息传播影响力排名前10位的作者贡献了社区约20%的新研究主题,表明高信息传播影响力节点能够为社区带来更多外部新颖信息,具有破茧潜力。

随后,本文计算各社区信息传播影响力前10位作者的新主题数量及其占比的平均值。结果显示,2023年社区平均节点数为203.87,各社区前10位作者平均新主题数为11.91,占全部新主题的21.70%,由此进一步验证,信息传播影响力较大的节点具有显著的破茧潜力。

表4 节点破茧潜力排序及信息传播影响力度量模型有效性分析(合著社区G1~G3)

合著社区	节点总数	信息传播影响力前10位作者	2023年新主题个数	占有新主题比例/%
合著社区G1	262	Wang, Meng	9	12.71
		Zhou, Jie	15	
		Yan, Shuicheng	0	
		Feng, Jiashi	0	
		Gao, Wen	1	
		Wu, Feng	0	
		Ji, Rongrong	2	
		Qin, Jie	0	
		Zhang, Tong	10	
		Lin, Zhouchen	1	
合著社区G2	285	Li, Qi	11	19.68
		Zhang, Wei	10	
		Wu, Fei	5	
		Chen, Guihai	35	
		Zaidman, Andy	3	
		Yang, Yang	10	
		Li, Cheng	0	
		Ren, Xiang	7	
		Cong, Gao	4	
		Li, Xuandong	2	
合著社区G3	263	Zheng, Kai	12	23.89
		Chen, Wei	11	
		Lo, David	37	
		Chen, Enhong	6	
		He, Yuan	2	
		Wang, Xiaofeng	4	
		Xia, Xin	23	
		Tian, Yuan	1	
		Grundy, John	20	
		Zhou, Xiaofang	13	

在此基础上,本文更新2022年具有破茧方向的作者合著-关键词语义相似度双层网络,并采用ATR_RWR算法对2023年作者进行破茧主题推荐。本文设置推荐列表长度RL为10,重启概率为0.5,破茧概率为0.8。为评估推荐效果,依据2023年作者实际采用关键词与算法推荐关键词之间的余弦相似度进行匹配,参考熊回香等^[26]的研究,当相似度超过0.6时,认定为作者偏好关键词。对每位作者而言,其偏好关键词占全部推荐关键词的比例即为推荐准确率。通过计算得到主题推荐的平均准确率为27.76%,在推荐任务中表现良好,证明了本文方法对广大作者群体具有普适性和有效性。

当作者节点所属社区发生变化时,意味着其合作范围扩展到原社区之外,可视为突破信息茧房。为此,本文重点分析各社区中信息传播影响力排名前 10 位且在 2023 年发生社区变动的作者,分析其实际研究主题与推荐主题之间的匹配情况。统计结果(附表 2)显示,这类作者的平均推荐准确率达到 46.83%,显著高于所有作者的整体水平(27.76%),表明推荐主题具有一定的破茧作用。

表 5 破茧作者实际研究主题和被推荐主题的相似度分析(合著社区 G1~G3)

作者	2022 年社区	2023 年社区	准确率/%	2023 年实际使用关键词	2023 年推荐关键词
Zhou, Jie	合著社区 G1	合著社区 G30	70	neural machine translation	cross lingual word embedding
				deep network	gradient vanishing problem
				semantic feature	entity mention
Li, Qi	合著社区 G2	合著社区 G11	75	mobile application	android application
				data leakage	software vulnerability
				facial expression	human face
Tian, Yuan	合著社区 G3	合著社区 G18	65	contrastive learning	collaborative learning
				mobile system	heterogeneous wireless network
				GNN model	text classification task

以作者“Zhou, Jie”为例,其社区由 2022 年的 G1 转为 2023 年的 G30,本文算法为其推荐的主题包括“cross lingual word embedding”“gradient vanishing problem”“entity mention”等,与其 2023 年实际采用的主题“neural machine translation”“deep network”“semantic feature”等高度相关,均属于自然语言处理领域,推荐结果较为准确。结合表 5,可归纳得到如下结论。

(1) 本文方法推荐主题与作者实际研究主题的相关性较高,即推荐准确率较高。例如,本文为作者“Zhou, Jie”推荐的关键词“cross lingual word embedding”作为跨语言语义表示方法,与实际使用关键词“neural machine translation”在模型语义空间构建方面紧密关联。虽然作者“Zhou, Jie”未直接研究跨语言词嵌入,但其在 2023 年提出基于图的多模态融合编码器^[52],说明其已围绕推荐主题展开相关研究,实现了茧房突破。

(2) 本文方法推荐主题具有多样性,为作者破茧提供了多样化的选择。例如,针对作者“Li, Qi”的推荐主题涵盖移动应用、系统安全及计算机视觉等多个技术方向,有助于作者在不同研究领域开展探索,实现信息茧房突破。

(3) 本文方法推荐主题能够有效帮助作者突破信息茧房。例如,作者“Zhou, Jie”长期从事遥感

此外,为增强破茧主题分析的代表性,本文选取合著社区 G1~G3 中发生变化且信息传播影响力处于不同位次的 3 位作者(“Zhou, Jie”“Li, Qi”“Tian, Yuan”)进行验证。表 5 展示了其社区变化情况、推荐准确率,以及推荐关键词与实际关键词的对应关系。结果显示,3 位作者的平均推荐准确率达到 70%,显著高于全部作者平均水平(27.76%),说明本文方法能够有效促进作者突破信息茧房。

成像与图像分析研究,自 2023 年起涉足自然语言处理领域,并与社区外作者开展合作,导致其所属社区发生变化。自然语言处理技术的引入为其遥感研究提供了跨模态语义理解与智能标注能力,拓宽了研究范式,增强了创新力,从而验证了推荐主题的破茧促进作用。

3.3 方法验证

(1) 推荐算法多样性指标对比

为验证本文提出的基于重启随机游走的作者研究主题推荐算法(ATR_RWR)的有效性,本节将其与重启随机游走算法(RWR)、基于用户的协同过滤算法(user-based collaborative filtering, UserCF)、基于项目的协同过滤算法(item-based collaborative filtering, ItemCF)以及图表示学习模型 Node2Vec 进行对比。RWR 是本文的基准算法;UserCF 和 ItemCF 为经典推荐方法,具有较高准确率;Node2Vec 在提升推荐多样性方面具有优势。为兼顾推荐准确性与多样性,将 ATR_RWR 的破茧概率设为 0.8,重启概率设为 0.5。

ATR_RWR 基于破茧主题进行推荐,而作者的个人破茧意愿可能导致其实际关键词与推荐关键词存在差异,进而影响准确性评估。因此,本文侧重于评估推荐的多样性。选用覆盖率与海明距离作为

评价指标。其中,覆盖率指的是所有作者推荐的关键词去重后的个数占总关键词数的比例,反映系统生成不同关键词的广度;海明距离衡量不同作者推荐列表之间的差异度,值越大,说明推荐多样性越好。考虑到推荐列表长度对指标的影响,本文设置RL=10、20、30进行对比实验,结果如表6所示。

表6 本文方法验证

推荐列表长度RL	算法	覆盖率/%	海明距离
10	ATR_RWR	82.04	0.9971
	RWR	46.84	0.9637
	User-CF_20	29.42	0.9141
	User-CF_100	8.52	0.7731
	Item-CF	11.95	0.8946
	Node2Vec	28.26	0.9595
20	ATR_RWR	89.65	0.9950
	RWR	64.98	0.9504
	User-CF_20	44.43	0.9027
	User-CF_100	14.92	0.7690
	Item-CF	7.86	0.8196
	Node2Vec	59.32	0.9441
30	ATR_RWR	91.95	0.9927
	RWR	73.93	0.9399
	User-CF_20	54.89	0.8974
	User-CF_100	20.04	0.7671
	Item-CF	11.30	0.8193
	Node2Vec	87.43	0.9534

注:粗体表示最优值。

表6显示,在各列表长度下,ATR_RWR在覆盖率与海明距离两项指标上均优于对比算法,证明本文方法在提升推荐多样性方面具有显著优势。

(2) 基于研究主题同质性和新颖性的推荐效果对比

本文从研究主题的同质性与新颖性出发,分析推荐主题对2023年合作群体信息茧房的影响。具体而言,一方面,将ATR_RWR推荐结果与2023年作者实际研究主题进行比较,分别计算推荐主题与实际主题的同质性和新颖性,以判断推荐是否有助于削弱群体内的信息茧房;另一方面,将ATR_RWR与推荐领域表现突出的图表示学习模型Node2Vec进行对比,采用相同流程生成推荐主题,并计算其同质性和新颖性,结果如表7所示。

由表7可知,本文提出的ATR_RWR算法在打破信息茧房方面具有显著优势。使用该算法后,合著社区的研究主题同质性显著低于2023年实际水

表7 推荐前后合作群体研究主题同质性和新颖性对比(部分合著社区)

		合著社区		
		G1	G2	G3
研究主题同质性	2023年实际	0.4198	0.3967	0.3982
	Node2Vec 推荐后	0.4670	0.4682	0.4693
	本文方法推荐后	0.3821	0.3680	0.3534
研究主题新颖性	2023年实际	3.4758	3.5298	3.5088
	Node2Vec 推荐后	3.5932	3.7092	3.6969
	本文方法推荐后	3.6738	3.7397	3.7908

注:粗体表示最优值。

平,而新颖性则显著提高,说明ATR_RWR能有效减少合作群体中的信息茧房现象。相比之下,Node2Vec推荐结果的同质性反而较实际情况增加,新颖性虽有所提升但幅度较小,整体效果不及ATR_RWR。上述结果表明,ATR_RWR在提升研究主题多样性和新颖性方面更具优势,更能促进科研合作群体突破信息茧房。

4 结论与不足

本文提出了一套面向科研合作群体的信息茧房识别与破茧主题推荐方法,能够有效捕捉合作群体结构及研究主题的动态演化特征,识别群体的信息茧房现象并分析茧房产生原因和强弱演化特征;同时,综合考虑了作者节点的破茧潜力及其在双层网络中与关键词的关联关系,实现了有助于信息茧房突破的研究主题推荐。

基于计算机科学领域论文数据的实证结果表明:①科研合作群体中存在信息茧房现象,主要表现为合著社区研究主题同质性上升、新颖性下降,这主要源于科研人员在合作中形成的研究兴趣、学术背景与合作网络的路径依赖,使跨领域交流受限,创新活力不足;②从节点的信息接收和发布能力两个维度构建信息传播影响力度量模型,综合考虑节点在网络中的结构位置和信息扩散能力,有效识别破茧潜力高的节点,实证结果显示,传播影响力排名前10位的作者贡献了约21.7%的新主题,表明其具有较强的破茧能力;③本文提出的ATR_RWR算法通过引入新主题促进外部信息流入,能够同时兼顾推荐准确性与多样性,有效降低了研究主题同质性、提升新颖性,其中,对所有作者的推荐平均准确率达到27.76%,验证了其在破茧推荐中的有效性。

本文的研究贡献主要包括:①拓展了信息茧房的研究场景,从传统社交领域延伸到科研合作群

体, 利用研究主题同质性与新颖性指标揭示了科研群体的信息窄化、群体极化与创新不足问题; ②构建信息传播影响力度量模型, 从量化信息接收与发布能力的视角评估了科研人员的破茧潜力, 为理解科研合作网络的信息传播与扩散机制提供支持; ③提出 ATR_RWR 算法, 将节点破茧潜力与双层网络关联关系融入推荐框架, 为科研团队配置优化与提升协同创新能力提供了可行策略。

不过, 本文仍存在不足。首先, 受时间与算力限制, 仅基于 2014—2023 年计算机科学领域论文数据展开研究, 而科研合作可能跨学科、跨领域, 不同学科的信息茧房现象也可能存在差异; 其次, 科研合作中的创新与茧房问题除研究主题外, 还可能受到合作者特征、合作模式等因素影响。未来研究可进一步丰富数据来源, 拓展至更多学科场景, 并进一步探讨其他因素对信息茧房的影响, 为破茧推荐引入更多类型的辅助信息。

参 考 文 献

- [1] D' Este P, Llopis O, Rentocchini F, et al. The relationship between interdisciplinarity and distinct modes of university-industry interaction[J]. *Research Policy*, 2019, 48(9): 103799.
- [2] 柳美君, 步一, 杨斯杰. 科研团队成员国别差异性的测度、演变及其与团队产出影响力的关系[J]. *情报学报*, 2024, 43(7): 818-838.
- [3] 袁顺波. 网络环境下科研人员信息茧房的成因要素研究[J]. *情报理论与实践*, 2022, 45(2): 43-48.
- [4] Wang J. Knowledge creation in collaboration networks: effects of tie configuration[J]. *Research Policy*, 2016, 45(1): 68-80.
- [5] 杨芳芳, 宋雪雁, 张伟民. 国内信息茧房研究热点与演进趋势: 兼论静态和动态双重视角[J]. *情报科学*, 2024, 42(5): 169-176, 185.
- [6] Yuan X F, Wang C Y. Research on the formation mechanism of information cocoon and individual differences among researchers based on information ecology theory[J]. *Frontiers in Psychology*, 2022, 13: 1055798.
- [7] 卢国强, 马海群. 信息茧房的形态辨析: 抽象概念与科学量化的交融探索[J]. *情报资料工作*, 2025, 46(6): 15-24.
- [8] 郭亚军, 李天祥, 冯思倩, 等. 算法推荐、信息茧房与“附近的消失”[J]. *图书情报知识*, 2025, 42(2): 156-166.
- [9] 丁敬达, 陈一帆. 专注—持续—延伸: 基于诺贝尔奖获得者研究模式探析的主题——以物理学领域为例[J]. *图书馆杂志*, 2023, 42(8): 100-109.
- [10] 朱红涛, 李姝熹. 信息茧房研究综述[J]. *图书情报工作*, 2021, 65(18): 141-149.
- [11] Sunstein C R. *Infotopia: how many minds produce knowledge* [M]. Oxford: Oxford University Press, 2006.
- [12] 姜婷婷, 许艳闰. 窄化的信息世界: 国外信息茧房、选择性接触与回音室研究进展[J]. *图书情报知识*, 2021, 38(5): 134-144.
- [13] Gu M, Zhao T F, Yang L, et al. Modeling information cocoons in networked populations: insights from backgrounds and preferences[J]. *IEEE Transactions on Computational Social Systems*, 2024, 11(3): 4497-4510.
- [14] 焦媛媛, 高雪, 杜军, 等. 信息茧房: 概念内涵、理论框架及在工商管理领域研究拓展[J]. *南开管理评论*, 2025, 28(8): 196-208.
- [15] 高艺畅, 孙英苹, 刘凤鸣, 等. 社交网络舆情传播的回音室效应研究[J]. *情报理论与实践*, 2022, 45(4): 58-66.
- [16] 岳芳, 樊茂瑞, 高子雅, 等. 开放式知识协同平台中的“回音室效应”研究——以 bilibili“新能源汽车”视频评论为例[J]. *情报理论与实践*, 2024, 47(2): 124-131.
- [17] Liu J Y, Schwarz A, Risius M, et al. Conceptualizing echo chambers and information cocoons: a literature review and synthesis of current knowledge and future directions[J]. *Journal of Strategic Information Systems*, 2025, 34(2): 101904.
- [18] Pariser E. *The filter bubble: what the internet is hiding from you* [M]. New York: Penguin, 2011.
- [19] 钟元生, 高成珍, 朱文强. 融合动态研究偏好和社交信任的潜在科研合作者推荐研究[J]. *情报学报*, 2023, 42(11): 1335-1346.
- [20] Li N, Gao C, Piao J H, et al. An exploratory study of information cocoon on short-form video platform[C]// *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. New York: ACM Press, 2022: 4178-4182.
- [21] Wang X, Huang X W, Sang J T. Understanding the information cocoon: a multidimensional assessment and analysis of news recommendation systems[OL]. (2025-09-14). <https://arxiv.org/pdf/2509.11139>.
- [22] 姜婷婷, 吕妍, 傅诗婷. 应对过滤气泡: 算法策展对用户信息消费行为选择性和态度极端化的影响[J]. *现代情报*, 2024, 44(7): 22-33.
- [23] O'Hara K. In worship of an echo[J]. *IEEE Internet Computing*, 2014, 18(4): 79-83.
- [24] Dong L, Wang Z Q, Tao Z Y, et al. Evaluating interdisciplinary contributions to synchrotron radiation studies: scientometric insights into novelty and disruption[C]// *Proceedings of the Advanced Applications on Qualitative and Quantitative Methods in Information Science*. Cham: Springer, 2025: 108-130.
- [25] Huang L, Chen X, Zhang Y, et al. Dynamic network analytics for recommending scientific collaborators[J]. *Scientometrics*, 2021, 126(11): 8789-8814.
- [26] 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.
- [27] 刘鹏, 夏昊翔. 跨学科研究领域的合著网络演化分析——以“复杂网络”研究领域为例[J]. *情报杂志*, 2015, 34(9): 85-91.
- [28] Mikolov T, Sutskever I, Chen K, et al. Distributed representations

- of words and phrases and their compositionality[C]// Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2013: 3111-3119.
- [29] Huang L, Chen X, Zhang Y, et al. Identification of topic evolution: network analytics with piecewise linear representation and word embedding[J]. *Scientometrics*, 2022, 127(9): 5353-5383.
- [30] Traag V A, Waltman L, van Eck N J. From Louvain to leiden: guaranteeing well-connected communities[J]. *Scientific Reports*, 2019, 9: Article No.5233.
- [31] 黄璐, 任航, 曹晓丽, 等. 面向有组织产学研协同创新的合作主题挖掘[J]. *科学学研究*, 2025, 43(3): 548-559.
- [32] 陈晶, 刘志君, 杨新宇, 等. 基于多核心节点的增量式动态社区发现算法[J]. *通信学报*, 2024, 45(3): 66-80.
- [33] 郭顺利, 张雪宁, 苏新宁. 科研人员生成式人工智能应用信息茧房成因及实证研究[J]. *情报理论与实践*, 2025, 48(2): 45-55.
- [34] 赵展一, 钟永恒, 李贞贞, 等. 基于技术匹配与分型优化的企业研发合作伙伴推荐方法研究[J]. *情报学报*, 2025, 44(1): 48-60.
- [35] 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.
- [36] Santos J M, Horta H, Feng S H. Homophily and its effects on collaborations and repeated collaborations: a study across scientific fields[J]. *Scientometrics*, 2024, 129(3): 1801-1823.
- [37] Uzzi B, Mukherjee S, Stringer M, et al. Atypical combinations and scientific impact[J]. *Science*, 2013, 342(6157): 468-472.
- [38] Shen D H, Wang G Y, Wang W L, et al. Baseline needs more love: on simple word-embedding-based models and associated pooling mechanisms[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 440-450.
- [39] Sacharidis D. Diversity and novelty in social-based collaborative filtering[C]// Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization. New York: ACM Press, 2019: 139-143.
- [40] 索琪, 王力媛. 官方—民间舆论场协同驱动的异质层次网络个体传播影响力评价[J]. *情报探索*, 2022(1): 10-18.
- [41] 魏开洋, 邱均平, 刘亚飞. 科研合作中明星作者对学术论文的影响机理研究——基于合著网络的视角[J]. *情报科学*, 2024, 42(2): 174-181.
- [42] 岳婷, 杨立英, 沈哲思. 科研活动中的倒U型现象研究初探[J]. *科学学研究*, 2025, 43(5): 996-1003.
- [43] Jin W, Jung J, Kang U. Supervised and extended restart in random walks for ranking and link prediction in networks[J]. *PLoS One*, 2019, 14(3): e0213857.
- [44] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(3): 355-369.
- [45] 时宇岑, 印莹, 赵宇海, 等. 基于多开发者社区的用户推荐算法[J]. *软件学报*, 2019, 30(5): 1561-1574.
- [46] 赵毅, 章成志, 习海旭. 影响不同于领域国际合作的距离因素相同吗?——来自计算机科学学科的证据[J]. *情报学报*, 2023, 42(12): 1458-1476.
- [47] 柳美君, 石静, 杨斯杰, 等. 科学家研究主题演化速度与科研绩效的关系研究: 基于计算机科学领域的分析[J]. *图书情报工作*, 2024, 68(6): 72-82.
- [48] Yu D J, Xiang B. Discovering topics and trends in the field of artificial intelligence: using LDA topic modeling[J]. *Expert Systems with Applications*, 2023, 225: 120114.
- [49] Zhang G Y, Song J J, Feng Z F, et al. Artificial intelligence applied in gastric cancer: a bibliometric and visual analysis via CiteSpace[J]. *Frontiers in Oncology*, 2023, 12: 1075974.
- [50] Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis[J]. *Physical Review E*, 2009, 80: 056117.
- [51] Tian M Y, Moriano P. Robustness of community structure under edge addition[J]. *Physical Review E*, 2023, 108: 054302.
- [52] Yin Y J, Zeng J L, Su J S, et al. Multi-modal graph contrastive encoding for neural machine translation[J]. *Artificial Intelligence*, 2023, 323: 103986.

(责任编辑 王海燕)

附 录

附表 1 节点破茧潜力排序及信息传播影响力度量模型有效性分析(合著社区 G4~G10)

合著社区	节点总数	信息传播影响力位于不同排序位置的 10 名作者	2023 年新主题个数	占所有新主题比例/%	合著社区	节点总数	信息传播影响力位于不同排序位置的 10 名作者	2023 年新主题个数	占所有新主题比例/%	
合著社区 G4	117	Zheng, Yuanqing	4	32.37	合著社区 G8	198	Liu, Jin	8	24.97	
		Lyu, Siwei	2				Zhang, Jiang	15		
		Xu, Changsheng	4				Gu, Dawu	8		
		Li, Wenbo	1				Zhang, Min	81		
		Chen, Dapeng	3				Gonzalez, Joseph E.	18		
		Yuan, Zejian	2				Schuurmans, Dale	3		
		Li, Hongdong	6				Wu, Jian	48		
		Li, Houqiang	6				Li, Yuan	3		
		Li, Ang	28				Louri, Ahmed	0		
		Geng, Tong	11				Karanth, Avinash	0		
合著社区 G5	247	Yan, Junchi	3	14.02	合著社区 G9	259	Xu, Yi	0	13.69	
		Xu, Lei	12				Gao, Xing	4		
		Wang, Tianyi	0				Wei, Xin	3		
		Wu, Qi	0				Yang, Zhenglu	11		
		Mahoney, Michael W.	2				Liu, Ye	20		
		Li, Jun	37				Li, Yi	8		
		Zhang, Xu	1				Lin, Shang-Wei	3		
		Wang, Lin	6				Zhu, Chenguang	8		
		Eugster, Patrick	0				Wang, Shuohang	12		
		Richtarik, Peter	7				Zeng, Michael	3		
合著社区 G6	279	Li, Guoliang	7	12.50	合著社区 G10	217	Wang, Bo	27	24.30	
		Chai, Chengliang	2				Zhang, Xinyu	1		
		Luo, Yuyu	2				Sun, Xing	3		
		Tang, Nan	1				Xiao, Yanghua	9		
		Wang, Yanhao	5				Shi, Wei	0		
		Chen, Xi	34				Sui, Yulei	15		
		Wong, Weng-Fai	4				Wan, Yao	18		
		Hu, Songlin	2				Li, Ying	10		
		Feng, Jianhua	1				Yu, Philip S.	13		
		Stuerzlinger, Wolfgang	4				Yang, De-Nian	0		
合著社区 G7	265	Li, Keqiu	0	16.53						
		Xiao, Bin	4							
		Wang, Xiaoliang	2							
		Wu, Hao	33							
		Li, Li	11							
		Harman, Mark	3							
		Cordy, Maxime	3							
		Papadakis, Mike	4							
		Le Traon, Yves	6							
		Xiong, Yingfei	16							

附表2 破茧作者实际研究主题和被推荐主题的相似度分析(合著社区G4~G10)

作者	2022年社区	2023年社区	准确率/%	2023年实际使用关键词	2023年推荐关键词
Geng, Tong	合著社区G4	合著社区G26	65	general purpose processor	hardware architecture
				machine learning	data science
				tensor core	graph analytic
Zhao, Han	合著社区G5	合著社区G2	30	deep neural network	dnn accelerator
				feature learning	salient region detection
				feature learning	few shot classification
Zhang, Yang	合著社区G6	合著社区G13	50	face recognition	facial expression
				attack method	poisoning attack
				differential privacy	local differential privacy
Min, Geyong	合著社区G7	合著社区G6	35	reinforcement learning	off policy evaluation
				private data	private key
				baseline algorithm	link prediction task
Wang, Liang	合著社区G8	合著社区G13	80	performance overhead	hardware overhead
				computer vision	visual object tracking
				energy consumption	power dissipation
Wei, Xin	合著社区G9	合著社区G7	35	maximum size	reed solomon code
				upper bound	optimal code
				maximum size	linear code
Zhang, Yue	合著社区G10	合著社区G14	45	data flow	dependency graph
				code snippet	sql query
				vector representation	sentence embedding