情报学报 2022 年 7 月 第 41 卷 第 7 期
Journal of the China Society for Scientific and Technical Information, Jul. 2022, 41(7): 720-732

DOI: 10.3772/j.issn.1000-0135.2022.07.006

学术文本词汇功能识别——在论文新颖性度量上的应用

罗卓然[1,2]，陆 伟[1,2]，蔡 乐[1,2]，程齐凯[1,2]

（1. 武汉大学信息管理学院，武汉 430072；2. 武汉大学信息检索与知识挖掘研究所，武汉 430072）

摘 要 为进一步挖掘学术论文新颖性的丰富内涵，本文以组合创新理论为基础，开展了基于词汇功能的学术论文新颖性度量研究。以ACM（Association for Computing Machinery）Digital Library收录的论文为数据，提出了面向CS（computer science）领域进一步预训练的词汇新颖性计算方法和基于语义相似度的问题-方法组合新颖度计算流程，分别计算了问题词、方法词、问题-方法组合和论文的语义新颖性，并将本文语义新颖性计算方法与已有的词频共现新颖性计算方法进行了对比。研究结果表明，ACM Digital Library收集的论文在研究方法和研究问题上创新度均较高，相较于已有的论文新颖性计算方法，本文提出的方法能从语义层面捕获更为精细的新颖性差异。

关键词 新颖性度量；词汇功能；问题-方法组合；预训练模型

Application of Lexical Functions in Novelty Measurement of Academic Papers

Luo Zhuoran[1,2], Lu Wei[1,2], Cai Le[1,2] and Cheng Qikai[1,2]

(1. *School of Information Management, Wuhan University, Wuhan* 430072;
2. *Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan* 430072)

Abstract: To further investigate the novelty measure of academic papers, we conducted a study on the novelty measurement of academic papers based on lexical functions. Specifically, we propose a method to calculate the novelty of words for further pre-training in the field of computer science based on the data in papers in Association for Computing Machinery (ACM) Digital Library and present a novelty calculation process for "question-method" combinations. We calculated the novelty of the test set data and performed a comparison between our method and existing methods. Results show that the papers collected from the ACM database are more innovative in terms of both research methods and research questions, and the method proposed in this paper can capture more fine-grained novelty differences at the semantic level.

Key words: novelty measure; lexical function; question-method combination; pre-trained model

1 引 言

从科学的发展来看，科学研究开始于问题发现[1]，美国著名科学哲学家L·劳丹曾在其著作《进步及其问题——一种新的科学增长理论》中强调，科学研究的目的是解决问题；问题和方法是科研工

收稿日期：2021-07-27；**修回日期**：2022-01-07

基金项目：国家社会科学基金重大项目“基于认知计算的学术论文评价理论与方法研究”（17ZDA292）。

作者简介：罗卓然，女，1993年生，博士研究生，主要研究领域为创新评价、数据挖掘，E-mail：zoraluo@whu.edu.cn；陆伟，男，1974年生，教授，博士生导师，主要研究领域为信息检索与可视化、数据智能与创新评价、AI人机协同等；蔡乐，男，1998年生，硕士研究生，主要研究领域为数据挖掘、深度学习；程齐凯，男，1989年生，副教授，主要研究领域为信息检索、科技情报分析。

作的重要组成内容，其中问题和方法的描述是科学话语的重要组成部分[2]，它以特定的形式和程度表现在论文中，固化为论文中的某些词汇或词汇组合[3]。在创新学研究中，组合往往被看作创新产生的一个重要来源。创新理论的鼻祖约瑟夫·熊彼特（Joseph Alois Schumpeter）在其著作《经济发展理论》中提出创新（innovation）是已有生产要素和生产条件的组合[4]，该观点后来得到了国际上许多有影响力学者的支持[5-6]。目前，学术界对于学术文本中的“创新”这一概念还未形成统一定义，常见的指代词如新颖性、创新力、颠覆性、innovation、novelty、creativity、fresh ideas、disruptive innovation等从创新的内容、时间、价值、影响等层面描述了创新的特征。学术研究成果的新颖性（novelty）能够在某种程度上反映其创新性或前沿性[7]，由于成果的价值一般需要较长的时间才能体现出来，在科研评价研究中常用新颖性描述研究成果的创新特质。通过文献调研与分析，本文发现学术论文的新颖性主要源于研究问题、研究方法、研究结论等元素的重组与结合，其中研究问题与研究方法的组合是形成创新的重要方式[2]。

在科学研究领域，研究人员发现影响最大的科学研究成果主要基于以往工作的组合，尤其那些非典型的组合[8-11]，并提出新颖性的主要来源是已有元素的重组或既有元素与新概念的组合[12-13]。此外，组合新颖性的内容和形式也不拘一格，国内外学者从参考文献组合[14]、参考文献的期刊组合[8-9,15]、词汇组合共现[12,16-17]等内容的组合对科学创新进行了研究。上述研究从组合创新的视角研究了科研论文的创新范式，为学术论文新颖性度量和创新性评价提供了理论和方法基础。然而，这种从期刊组合或参考文献组合的角度度量新颖性的方法，在脱离论文内容的情况下测度论文新颖性，对新颖性的解释力度还有所欠缺。值得注意的是，部分研究从论文词汇组合的角度开展了新颖性研究，这类研究的对象更接近创新本体的内容层面，但是仅从词汇组合频率的角度计算新颖性[18-19]，而缺少考虑词汇之间的语义差异，这种情况下可能会忽略新颖性的重要特征。例如，对生物医学词汇之间的组合和生物医学与计算机科学词汇的组合而言，后者是一种跨学科词汇的组合，这种组合能为新颖性来源和创新扩散的研究提供重要线索。挖掘组合词汇的语义内涵，可以揭示不同跨领域研究背后的知识交叉与融合情况[20]，有助于从词汇功能的角度揭示论文新颖性的语义内涵[21]。

学术文本的词汇功能是根据文本所在的语义环境对其承担的语义角色和功能的认知和理解[22]。学术论文作为科研成果载体，其核心问题和核心方法解释了论文待研究的问题和解决途径[23]，是体现论文新颖性和价值的重要功能元素。目前，国内外关于学术论文中的研究问题或研究方法的研究，主要集中在领域研究主题识别[24]、研究方法库构建[25]、跨学科研究问题[26]与研究方法分析[27-28]等方面，而将问题与方法的组合应用在论文新颖性测度上的研究相对较少。

为进一步探索面向文本内容层面的新颖性度量方法，本文以组合新颖性理论为基础，以学术论文细粒度词汇功能语义差异为切入点，利用深度学习预训练模型获取蕴含语义信息的词向量，提出面向CS（computer science）领域进一步预训练的词汇新颖性计算方法，通过模型对比实验证明本文的预训练模型表现效果更好。最后，将提出的语义新颖性计算方法与已有的共现率新颖性计算方法进行比较，结果表明，本文提出的方法能够捕获词汇及词汇组合之间更细粒度的新颖性差异。

2　相关研究

2.1　学术文本词汇功能研究

术语抽取是海量文献内容分析研究的基础，其中不同术语的功能识别是分析术语语义功能的重要环节。伴随着细粒度文本挖掘和实体抽取研究的深入，文本词汇功能识别研究引起了越来越多的关注，学者们从内容元素、概念类型、词汇功能和知识元等角度开展了词汇功能相关研究。Kondo等[29]将标题中的内容元素分为head、method、goal和other四类，并通过构建特定领域的方法/技术演化路径构建了技术趋势图生成系统。Gupta等[30]将学术文献的词汇功能分为话题、技术和领域三类并实现其自动识别。Tsai等[31]将收录于ACL（Association for Computational Linguistics）数据库中的科学文献中的概念分为技术（technique）和应用（application）两个功能类别，并提出了用于识别、归纳和聚类这两类概念的算法，研究结果可为深入了解ACL社区的研究进展、变化和趋势提供有用的见解。Tuomaala等[32]对LIS（library and information science）领域1965—2005年发表的研究论文进行了内容分析，分析了研究论文主题分布与采用的方法

和策略，解释了研究问题和研究方法之间的联系。Heffernan 等[2]认为科学研究是问题提出和解决的过程，将科学文献中的词汇功能分为研究问题和解决方法，并训练分类模型对短语是否为问题或方法进行二值判断。近年来，国内学者也对学术文本术语及词汇功能识别展开了一些探索和研究。赵洪等[33]构建了面向理论术语的深度学习模型，研究了该模型中理论术语的特征构造和标注方法，并通过实验对比验证了该模型的有效性。王昊等[34]对情报学理论方法进行研究，利用深度学习模型开展了训练与测试，发现术语实体的长度、训练语料量、实体的类型和数量等因素也与识别结果直接相关。李贺等[35]构建了学术论文的研究问题、理论、方法、结论 4 个知识元本体，提出了基于知识元的学术论文创新性判断方法。章成志等[36]将研究方法分为论文使用研究方法和论文引用研究方法，以《情报学报》10 年的论文全文为数据对象，利用神经网络模型抽取了研究方法实体并分析了其使用情况，发现情报学学科领域中使用频次和引用频次最高的均是与实验相关的研究方法。化柏林[28]通过对文献中研究方法内容描述的分析，将学术论文中的方法知识元总结为方法定义知识元、方法关系知识元、方法特点知识元、方法流程知识元和方法功能知识元 5 种类型。程齐凯等[37]提出了一种基于深度学习和标题生成策略的学术文本词汇功能识别模型，基于 seq2seq 模型和 attention 机制的方式捕获词汇的多层语义信息，实现了学术文本中问题词和方法词的生成。陆伟等[38]构造了一种基于规则标题的数据标注方法对数据进行标注，并利用 BERT（bidirectional encoder representation from transformers）预训练模型对输入的文本进行向量化表征，利用 LSTM（long short-term memory）对关键词进行自动判别以实现论文关键词的问题或方法的识别。

2.2 组合新颖性度量研究

在学术论文新颖性度量与评价研究领域，不少学者试图将基于人工甄别的传统新颖性度量方式转化为自动识别的新型评价方式。作为创新模式研究的重要范式之一，组合目的是对创新发展和创新扩散过程进行理论化与建模[39-40]。从组合内容和方式来看，代表性研究为参考文献的期刊组合。Uzzi 等[8]率先提出了基于重组的论文创新性度量，他们分析了来自 Web of Science 中 1950—2000 年发表的近 1790 万篇文献，发现论文新颖性与先前工作的非常规组合有较大相关性。Boyack 等[15]基于 Uzzi 等[8]的方法，以 Scopus 中收录的期刊为数据对象，利用基于期望标准差的 K50 指标替代了 *Z*-score 指标，结果显示，该方法可以在文献发表后的更早期得出同样的结论。Wang 等[9]将科学研究视为一个组合过程，通过检查已发表的论文是否首次对参考期刊进行组合来衡量科学的新颖性。除了参考文献的期刊组合之外，有研究者直接利用参考文献的组合来度量文献的新颖性。Mukherjee 等[14]基于参考文献的共被引网络建立了“常规性-新颖性”的二维坐标系，将论文划分为 4 个创新类型。Ponomarev 等[41]认为，开创性成果是基于对已有研究的回顾与总结，提出了基于出版物引用动态检测方法，并建立了论文创新性预测模型。Tahamtan 等[10]认为一篇论文中参考文献的不寻常组合可以揭示其新颖性潜质，通过分析论文引文网络中不同类型、不同主题的组合，归纳出了创新性论文常见的主题组合模式。此外，部分学者从与论文直接相关的词汇角度度量了论文的新颖性。Azoulay 等[12]通过检查论文中的 MeSH 主题词对，计算未出现在 PubMed 上所有先前文献中的词对所占的比例，来衡量出版物的重组特征与新颖性，发现论文的重组程度与引文量之间存在负相关关系。Yan 等[40]定义了论文的新组合和新组件，提出了一种利用论文的关键字测度组合新颖性的方法。从问题词和方法词的角度，王艳艳等[18]利用人工的方法抽取科技文献中的问题和方法，将问题、方法作为两个维度构建了新颖性评估方法模型。钱佳佳等[19]根据词频和词组合的频次，提出了一种基于问题-方法组合的科技论文新颖性度量方法。Luo 等[42]考虑了词汇的年龄和语义差异，提出了从词汇生命指数和语义相似度两个角度计算论文新颖性的方法。综上，相关研究从期刊组合、引文组合、主题词组合等角度开展了组合新颖性研究，也有从问题词和方法词的不同功能角度探索了论文新颖性测度，为本文的研究提供了良好借鉴的同时也存在研究数据不足、方法受限等情况。在此现状下，本文发现从语义层面度量论文新颖性仍有进一步探索的空间。

学术论文的研究问题与研究方法是表达学术文本新颖性的主要功能词汇，这种具有特殊语义功能词汇的组合为新颖性研究提供了新思路。因此，本文在前期学术文本词汇功能研究的基础上开展词汇功能在论文新颖性度量上的研究。

3　基于语义相似度的“问题-方法”新颖性度量方法

3.1　数据选择

在程齐凯等[37]、陆伟等[38]前期关于词汇功能的研究基础上，本文利用论文研究问题、研究方法及其组合来测度论文的新颖性。为此，需要在论文中预先提取表征研究问题与研究方法的词汇。由于论文的研究问题或研究方法可能不只一个，本文仅抽取了每篇论文中主要的问题词和方法词，即将论文认为是某一问题与某一方法的组合。本文中的主要问题词是指能够代表论文核心研究问题的词或词组，主要方法词是指用于表征论文为研究解决问题所采用的方法、模型、工具或途径的词或词组。实际中存在部分论文涉及多个研究问题或方法的情况，对于本文研究的组合新颖性而言，测度主要问题和主要方法的组合已能够达到本文的研究目的，而多问题与多方法的自动抽取研究是下一步待解决的问题。

本文将 ACM （Association for Computing Machinery）Digital Library（下称 ACM 数据库）作为数据来源，该数据库收录了计算机领域权威和前瞻性的出版物，提供了解计算机和信息技术领域资源的窗口。陆伟等[38]提出的问题方法识别模型整体准确率、召回率和 F1 值分别达到 0.83、0.87 和 0.85，优于传统模型的效果。本文利用该模型提取了 ACM 数据库中 1968—2018 年的 200182 篇文献的研究问题词和研究方法词，并比较了模型识别效果与人工判断的差异，在随机筛选的 100 条数据中主要问题方法词识别一致性为 82%。然后，抽取了每篇论文的 DOI 号、题目、摘要、关键词、发表时间等题录信息，统计截止到 2021 年 2 月论文在 ACM 数据库中显示的被引量。数据清洗操作中删除了字段为空的数据记录，保留了 200103 条包含题录信息和被引量在内的“问题-方法”记录数据，并将其保存在数据库中，实验数据随时间的数量分布如图 1 所示。统计每组“问题-方法”对出现频数，再按照字母升序的方式为每一个问题词和方法词构建索引。最后，在数据库中对所有的记录数据进行条件查询，并为每条记录的论文设置索引 ID，从实验数据中随机抽取 2018 年的 200 条记录作为分析数据，剩余的 199903 条数据作为历史对照数据。

3.2　技术基础

为从语义层面计算问题词与方法词的新颖性差异，本文采用深度学习预训练模型，在大规模科学文本数据集上训练问题方法词和方法词的词向量模型。词向量是一种将词表示成向量的无监督学习技术，代表性的词向量训练模型有 word2vec[43]、GloVe[44]、BERT[45]等。2018 年，谷歌提出的 BERT 模型刷新了自然语言处理领域的 11 个方向的最佳指标，是继 word2vec 之后深度学习方法在自然语言处理中的又一突破。BERT 模型利用 Transformer[46]构造多层双向编码，该模型训练的词向量可用于文本相似度相关任务中。Su[47]于 2020 年提出的 SimBERT 模型是经过微调的 BERT 模型，在文本相似度任务上效果提升显著，可见 BERT 模型在语义相似度判断上仍具有较好的表现。此外，SciBERT 是 Beltagy 等[48]提出的一种基于 BERT 的预训练语言模型，该模型在 BERT 的基础上进一步在大型多领域的科学出版物语料库上进行了无监督预训练，提高了模型处理下游自然语言处理任务的性能，该模型能用于解决缺乏高质量、大规模标注科学数据的问题。

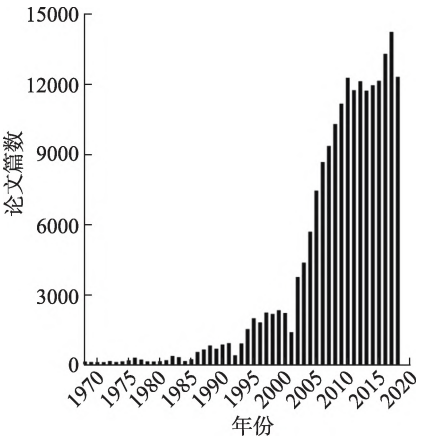

图 1　论文数量分布

鉴于科学语料在词汇功能与内容含义层面具有高度的专业性和领域区分度，直接使用 SciBERT 的问题在于对所有的输入向量都倾向于编码到一个较小的空间区域内，导致大多数的问题方法词对都具有较高的相似度分数，不利于语义新颖性差异化度量。为此，本文参考文本表示领域的常规做法[49-50]，再次引入 ACM 语料做进一步预训练，在获取更好语言模型的同时得到更能表征问题词和方法词真实差异的向量表示。语言模型效果的常用评价指标是困惑度（perplexity），在一个测试集上得到的困惑度越低，说明建模的效果越好[51]。本文选择困惑度作为模型评价指标。

3.3 模型训练与词汇新颖性计算

为从语义层面度量学术论文中研究问题词汇与研究方法词汇的新颖性差异，本文基于BERT模型将词汇表示成词向量的形式，将利用这些词向量表示辅助计算“问题-方法”组合的新颖性。进一步地，本文提出一个面向CS领域进一步预训练（further pretrain）的词汇新颖性计算方法，如图2所示。本文在SciBERT的基础上引入ACM数据库中200182篇论文中的标题及摘要信息，通过无监督训练任务根据句子上下文来预测的概率分布，实现对SciBERT的进一步预训练，通过对模型调参和训练，生成面向ACM论文语料的词向量表征模型SciBERT-further。

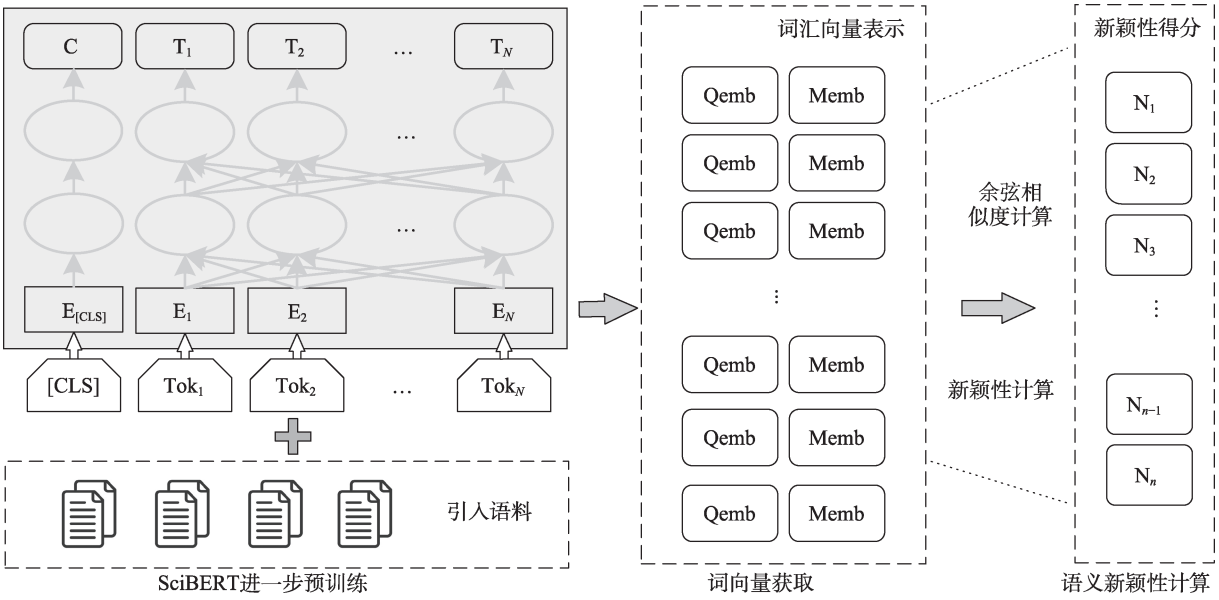

图2 面向CS领域进一步预训练的词汇新颖性计算方法

进一步预训练模型效果验证。首先，对收集到的ACM语料进行分句并统计句子信息，结果表明，25%的句子是短句，在15词以内，75%的句子在27词以内，最大句长76词。为尽可能完全覆盖语料中的句子，再训练时设置模型最大句长为72。在打乱句子顺序后，按照9∶1的方式划分训练集和测试集。然后，针对本文相似的问题-方法在编码后的表示空间中应当相近，不同的问题-方法应相距较远的需求，为获取更好的词汇级词向量表示，对同一样本利用打乱词序、特征裁剪两种方式进行数据增强，同时利用模型的第一层词汇编码和最后一层句子编码实现信息融合。最后，在测试时选择了模型困惑度作为评测指标，对于测试集，将其测试样本全部融合计算，取平均值计算该指标。训练集的模型损失和测试集的困惑度分别如图3a和图3b所示。

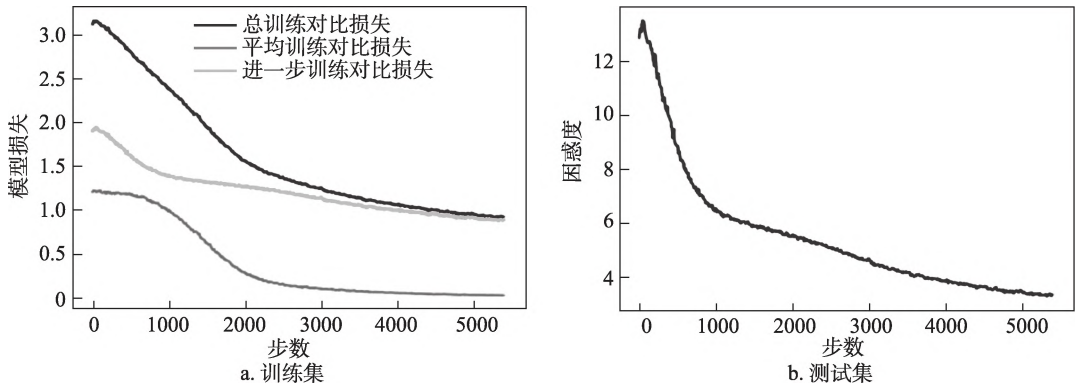

图3 进一步预训练中模型损失和困惑度变化图

此外，本文在文本语义匹配任务（semantic textual similarity，STS）的 STS12、STS13、STS14、STS15、STS16 这 5 个数据集上进行了实验，并对比了 Avg. GloVe、BERT、SciBERT 和 SciBERT-further 模型在无标注的 STS 数据上的训练效果，具体得分如表 1 所示。结果显示，在完全一致的设置下，本文提出的 SciBERT-further 模型相对于 Avg.GloVe 模型平均提升了 3%，相对于 BERT 提升了 10.5%，相对于 SciBERT 平均提升了 17%，表明本文提出的 SciBERT-further 模型能较好地表征词汇真实特征，且比在类似任务上采用 BERT 模型的表现更好[42]。

表 1　SciBERT-further 与其他方法在无监督情况下的性能比较

模型	数据集					均值
	STS12	STS13	STS14	STS15	STS16	
Avg.GloVe	55.14	70.66	59.73	68.25	63.66	63.49
BERT	39.39	59.34	49.66	66.02	66.19	56.18
SciBERT	35.82	50.46	41.66	60.58	57.35	49.17
SciBERT-further	59.96	68.57	62.27	71.29	71.32	66.68

问题词和方法词新颖性计算。提取学术论文“问题-方法”数据集中的问题词和方法词，在 SciBERT-further 模型中计算并获取上述词的词向量。然后计算当前问题词和方法词与已有词汇空间中所有词汇的余弦相似度，取最大值，计算词汇的新颖性，问题词和方法词的新颖性计算方式分别为

$$\text{quesNov} = 1 - \max\left(\frac{V_q \cdot V_{q_i}}{\|V_q\| \cdot \|V_{q_i}\|}\right) \tag{1}$$

$$\text{methodNov} = 1 - \max\left(\frac{V_m \cdot V_{m_i}}{\|V_m\| \cdot \|V_{m_i}\|}\right) \tag{2}$$

其中，quesNov 表示问题词新颖性，V_q 表示当前问题词的词向量，V_{q_i} 表示问题词域的第 i 个问题词的向量表示，计算 V_q 和 V_{q_i} 的余弦相似度，用 1 减去最大的向量余弦相似度，得到 quesNov 的值，若 V_q 与 V_{q_i} 越相似，则表示 V_q 的新颖性越小；methodNov 表示方法词新颖性，V_m 表示当前方法词的词向量表示，V_{m_i} 表示方法词域中第 i 个方法词的向量表示，用 1 减去最大的向量余弦相似度，得到 methodNov 的值。

3.4　“问题-方法”组合新颖性计算

对于论文中的“问题-方法”组合，在学术论文“问题-方法”数据集中查找当前问题词或当前方法词是否存在。若存在，则表明是旧的研究问题或研究方法；若不存在，则表示当前词在已有的问题词域或方法词域中不存在，属于新的研究问题或研究方法。组合新颖性计算的是相对新颖性，即当前组合词相对于组合对象的所有历史组合词的新颖性。这里对问题方法词是否存在进行了精确查找，只要之前在数据集中未出现过即为新词。语义相似度用在计算组合对象的新颖性上，即对旧的问题词或方法词，计算它的当前组合词与历史组合词序列之间的相似度。在钱佳佳等[19]对“问题-方法“组合划分的基础上，本文从词汇组合方式上将“问题-方法”组合进一步分为 5 种类型：“新问题+新方法”组合、“新问题+旧方法”组合、“旧问题+新方法”组合、“旧方法+旧问题”旧组合和“旧方法+旧问题”新组合。

对于“旧问题+新方法”和“新问题+旧方法”的组合而言，在已有的问题空间中分别提取与其组合过的词，形成旧问题的方法序列和旧方法的问题序列。由于本文主要从词汇功能组合的角度研究“问题-方法”组合，因此计算的是当前组合词与已有组合序列的相似度。因此，对于“旧问题+新方法”组合，“新方法”不是相对于全部方法词域来说的，而是相对于旧问题的方法序列而言，即只要当前方法词没有与当前问题的方法词序列组合过，对于当前的组合来说该方法即为新方法。然后，计算当前方法词的组合新颖性，分别计算当前方法词与旧问题的组合序列中各个方法词的相似度。最后，将当前组合词的新颖性得分赋值给“问题-方法”组合，得出最终组合新颖性。基于语义相似度的“问题-方法”组合新颖性计算流程如图 4 所示。

对于旧问题或旧方法的组合而言，本文将“旧问题”和“旧方法”称作当前词，与其组合的对象称作组合词。对于“问题-方法”组合中的当前词 t，要测度其组合的新颖性，则需要判断其组合词 p 的相对新颖性。例如，对于现有研究中已存在的旧问题 t，首先枚举与该问题组合过的所有方法，形成 t 的历史组合序列 $P(p_1, p_2, \cdots, p_n)$。利用 SciBERT-further 模型计算当前组合词 p 的向量表征 V_p 与 P 中各个历史组合词的词向量的余弦相似度，计算方式为

$$\text{combSim}_i = \frac{V_p \cdot V_{P_i}}{\|V_p\| \cdot \|V_{P_i}\|} \tag{3}$$

其中，V_{P_i} 表示序列 P 中的第 i 个元素的词向量表征；combSim_i 表示 V_p 与 V_{P_i} 的余弦相似度。

“问题-方法”组合的相似度取当前组合词 p 与

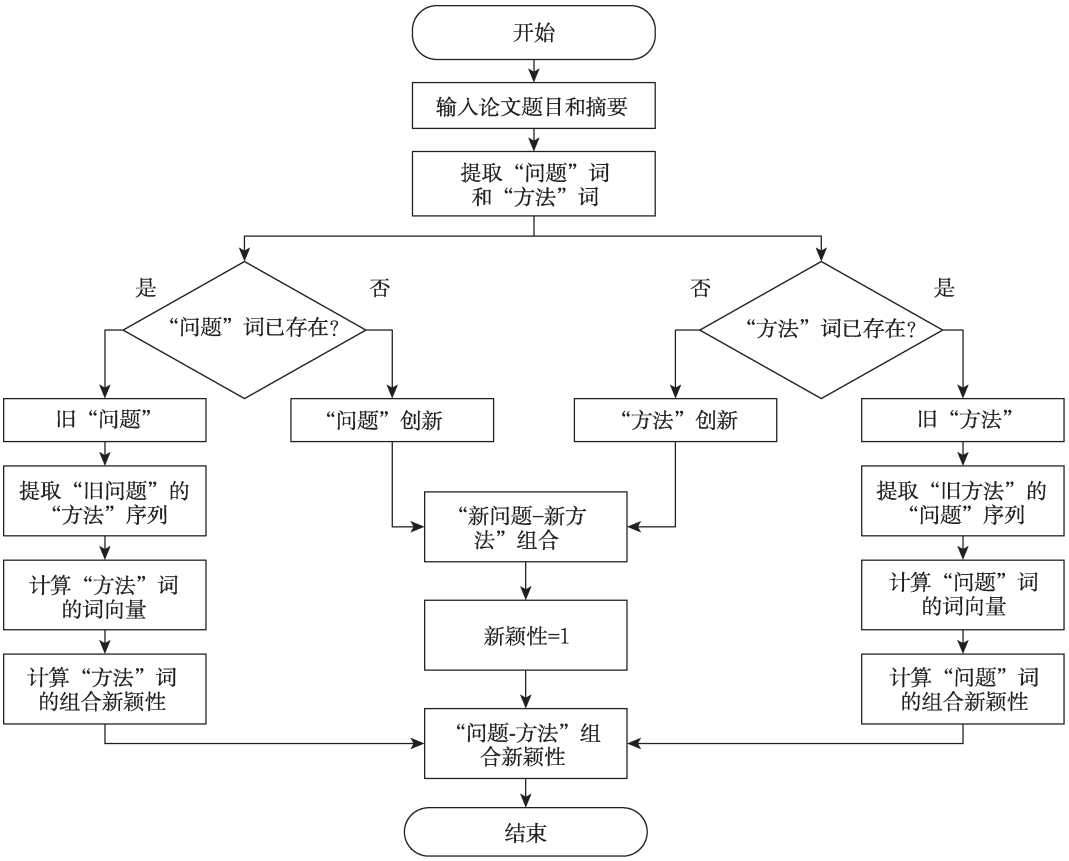

图4 基于语义相似度的"问题-方法"组合新颖度计算流程

当前词t的历史组合序列P中各个元素的最大相似度值，“问题-方法”组合的相似性越高，表示该组合的新颖性越低，将“问题-方法”的新颖性得分定义为$\mathrm{combNov}(t,p)$，计算方法为

$$\mathrm{combNov}(t,p)=1-\max(\mathrm{combSim}_1,\mathrm{combSim}_2,\mathrm{combSim}_3,\cdots,\mathrm{combSim}_n) \quad (4)$$

本文将论文的新颖性$\mathrm{Novelty}(D)$定义为问题词新颖性、方法词新颖性以及问题-方法组合新颖性三项的算数平均值，即

$$\mathrm{Novelty}(D)=(\mathrm{quesNov}+\mathrm{methodNov}+\mathrm{combNov}(t,p))/3 \quad (5)$$

若一篇论文存在多个问题与方法，则逐个计算问题词、方法词以及所有的问题-方法组合的新颖性，对这些新颖性得分取算数平均值就得到论文新颖性。

4 “问题-方法”新颖性测度实验

4.1 基于语义相似度的“问题-方法”新颖性计算

采用训练得到的词向量模型SciBERT-further计算得到所选问题词和方法词的词向量，并根据公式(1)~公式(4)计算词和组合的新颖性。由于计算出的新颖性得分均较小，不能显著体现不同组合之间的差异性，为便于数据可视化分析，本文对数值小于1的新颖性得分进行了分值归一化处理，计算方式为

$$\mathrm{noveltyNormal}=\frac{\mathrm{noveltyScore}-\mathrm{noveltyScore}_{\min}}{\mathrm{noveltyScore}_{\max}-\mathrm{noveltyScore}_{\min}+t} \quad (6)$$

其中，noveltyNormal表示归一化后的新颖性得分，取值范围为[0,1)；noveltyScore表示计算出的词和组合的新颖性得分，$\mathrm{noveltyScore}_{\min}$表示测试集数据中新颖性得分的最小值，$\mathrm{noveltyScore}_{\max}$表示测试集数据中新颖性得分最大值；为避免分母为0，在分母中加上常数t，这里取t=0.001。

通过上文的模型训练与新颖性计算，得到了测试集中200篇论文的“问题-方法”新颖性得分，其中“问题”词、“方法”词和“问题-方法”组合的新颖性得分取值范围均为[0,1]，具体分布如图5所示。图中绿色的圆点表示“问题-方法”组合新颖性得分，圆点左边蓝色和右边黄色的柱状线分别表示论文研究问题和研究方法的新颖性得分。由统计数据和图6可知，2018年发表的200篇论文中，“旧

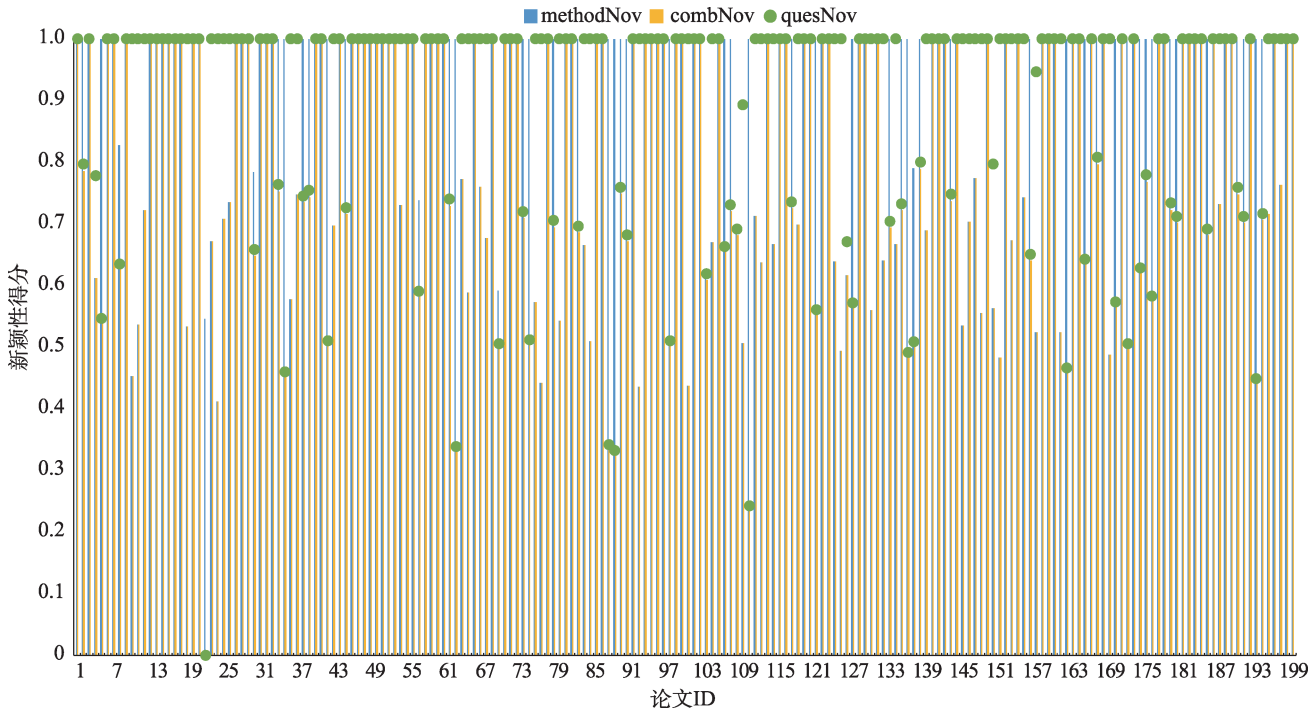

图 5　基于语义相似度的“问题-方法”新颖性得分（彩图请见 https://qbxb.istic.ac.cn/CN/volumn/home.shtml）

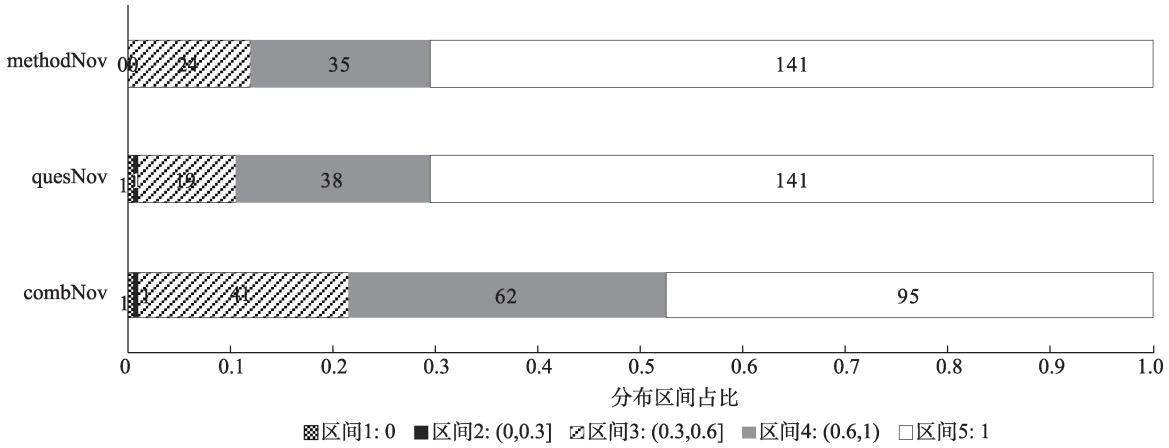

图 6　“问题-方法”新颖性取值分布区间

问题+旧方法”的论文有 1 篇，占所有测试论文的 0.5%，说明对于 ACM 数据库中收录的计算机领域的论文而言，同一个研究问题采用与已有研究完全相同的方法进行研究的论文占极少数，而多数研究属于“新问题+旧方法”或者“旧问题+新方法”的组合。此外，“新问题+新方法”的论文有 95 篇，占所有测试论文的 47.5%，由此可见，近半数的研究具有问题和方法两个层面的创新。

此外，本文对三类新颖性得分进行了区间分布统计，按区间将新颖性取值分为 5 个部分：区间 1，新颖性得分为 0；区间 2，新颖性得分取值范围(0, 0.3]；区间 3，新颖性得分取值范围(0.3,0.6]；区间 4，新颖性得分取值范围(0.6,1)；区间 5，新颖性得分取值为 1。本文将词汇新颖性的阈值设置为同类型所有词新颖性得分的中位数，统计结果表明，本实验中问题词和方法词新颖性阈值均为 1。

由图 6 可见，测试集中的问题词和方法词的新颖性值的数量分布在 5 个区间的呈现一致性，即位于区间 1 的新颖性为 0 的最少，而新颖性为 1 的最多，说明在 ACM 收录的论文中无论是研究问题还是研究方法，与已有的主题完全重合的占比非常小，只占到所有分析数据的 0.5%，而 70.5% 的问题

词和方法词的新颖性为1，即在已有的主题词空间中均未出现过。从“问题-方法”组合的角度看，组合新颖性值要整体小于单个问题词或单个方法词的新颖性值的分布，新颖性为1的组合占所有测试数据的47.5%，组合新颖性值位于区间3和区间4的数据占所有数据的51.5%，表明“问题-方法”组合中有一半是具有中度新颖性的。整体而言，通过词向量语义相似度计算的不同新颖性区间的数值差异明显，问题词和方法词在不同新颖性区间的数量分布呈现相同的分布特征，亦表明不同功能的词汇在语义相似度上具有一致性，说明本文提出的基于词向量语义距离计算的“问题-方法”组合新颖性能够测度不同词汇之间的新颖性差异。

4.2 结果分析与讨论

采用以上方式计算出论文的“问题-方法”组合新颖性的得分后，为进一步解释该方法的度量效果，本文分别从高新颖性的高被引和高频词两个角度对结果进行实例分析。

从高新颖性和高被引角度来看，本文结合论文的被引量指标，从高新颖性得分（问题、方法、组合新颖性得分均为1）的论文中，列举了排名前五的论文，如表2所示。由表2可知，新颖的研究主题包括用户和项目关系学习、Ad-Hoc搜索、上下文感知计算系统、网络型数据挖掘、个性化检索等，与主题相对应的新颖的研究方法包括潜在关系度量学习、语法软匹配、将语境利用在递归推荐系统中、基于相似度的多功能图嵌入和随机点击模型。由此可见，对计算机领域近些年的研究而言，若以论文的被引量代表论文的影响力，从问题和方法组合新颖性的角度来看，ACM数据库中收录的新颖性和影响较强的论文研究主题与信息检索、用户信息行为、推荐系统密切相关，问题的解决方法则采用深度学习、人机协同、图网络等衍生方法，与用户行为、情境感知、决策匹配等情景的相关性更高。

表2 ACM数据库2018年高新颖性论文示例

论文标题	问题词	方法词	被引量
Latent Relational Metric Learning via Memory-based Attention for Collaborative Ranking	collaborative ranking	latent relational metric learning	137
Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search	ad-hoc search	soft-matching n-gram	158
Latent Cross：Making Use of Context in Recurrent Recommender Systems	contextual recommendation	latent cross	149
VERSE：Versatile Graph Embeddings from Similarity Measures	web-scale data mine	VERtex similarity embeddings	137
Position Bias Estimation for Unbiased Learning to Rank in Personal Search	person search	random click model	103

从词频角度来看，词的出现次数能够反映该话题的热度和关注度。本文统计了测试集中问题词和方法词的频次，并分别选取了2个高频问题词和2个高频方法词，获取与其相关的论文信息，如表3所示。高频问题词“人机交互（human-robot interaction）”和“无线网络（wireless network）”是计算机领域经典的研究问题。示例论文Q1-1和Q1-2围绕经典研究问题“人机交互”开展了研究，Q1-1讨论了如何进一步探索不同的反馈方法，并研究它们对信任、控制分配和工作负载的影响，属于采用新方法解决旧问题的研究。论文Q1-2开发了一个基于任务对话和聊天机器人的人机交互多通道系统，并证明了该系统中应用强化学习是有益的，是旧问题+旧方法新组合类的研究。这两篇论文研究了同样的旧研究问题，Q1-2采用了热门的深度学习模型强化学习（reinforce learning），在发表后获得了比Q1-1更高的被引量，表明用旧方法+旧问题组合在新颖性上可能比新方法+旧问题弱一点，但是影响力不一定比新方法低，因为旧方法可能在某阶段引起了大量的研究兴趣，例如，Q1-2中的“强化学习”一词虽然在1998年就已出现，但随着近些年智能计算和深度学习的发展，强化学习再度受到了较多的关注。示例论文Q2-1和Q2-2研究了计算机工程领域无线网络（wireless network）的问题。Q2-1提出了一个处理器支持的超低延迟调度实现PULS（propellant utilization loading system），用于测试无限网络下行调度协议的超低延迟需求。Q2-2提出了无线网络拓扑选择和组件规模调整的设计空间探索方法，其新颖性类别为旧问题+旧方法的新组合，研究方法是旧方法且受到的关注较少，发表后获得的被引量较低。

高频方法词社交媒体（social media）和机器学习（machine learning）是近年来人工智能方向的热点词，示例论文M1-1和M1-2研究了“社交媒体”

表 3　高频问题词和方法词组合论文示例

论文编号	论文标题	问题词/频次	方法词/频次	新颖性得分(问题,方法,组合,总分)	被引量	新颖性类型
Q1-1	Feedback Methods in HRI: Studying their effect on Real-Time Trust and Operator Workload	human-robot interact/174	real-time trust, workload and operator workland/1	(0.51444,1, 0.50677,0.67374)	5	旧问题+新方法组合
Q1-2	Combining Chat and Task-Based Multimodal Dialogue for More Engaging HRI: A Scalable Method Using Reinforcement Learning	human-robot interact/174	reinforce learning/94	(0.56547,0.49925, 0.49925,0.52132)	23	旧问题+旧方法新组合
Q2-1	PULS: Processor-Supported Ultra-Low Latency Scheduling	wireless network/120	ultra-low latency schedule/1	(0.54873,1, 0.54055,0.69643)	8	旧问题+新方法组合
Q2-2	Optimized selection of wireless network topologies and components via efficient pruning of feasible paths	wireless network/120	efficient pruning/3	(0.77714,0.52083, 0.52083,0.60627)	4	旧问题+旧方法新组合
M1-1	Snitches, Trolls, and Social Norms: Unpacking Perceptions of Social Media Use for Crime Prevention	crime prevent/6	social media/214	(0.4173,0.42347, 0.41108,0.41728)	9	旧问题+旧方法新组合
M1-2	Falling for Fake News: Investigating the Consumption of News via Social Media	fake news/1	social media/214	(1,0.25121, 0.25121,0.50081)	70	新问题+旧方法组合
M2-1	Machine Learning-based Prediction of ICU Patient Mortality at Time of Admission	patient mortality prediction/1	machine learning/380	(1,0.59433, 0.59433,0.72955)	10	新问题+旧方法组合
M2-2	Machine learning for software engineering: models, methods, and applications	software engineering/160	machine learning/380	(0.30884,0.58775, 0.30423,0.40027)	10	旧问题+旧方法新组合

作为研究方法时的应用。M1-1 研究了人们如何看待社交媒体在其社区中支持预防犯罪的使用，属于常规旧问题+旧方法的新组合，新颖性较低且发表后获得的引文量较少。M1-2 研究了人们对社交媒体新闻的态度，研究结果突出了打击假新闻传播的困难，该研究是将旧的研究方法应用在新的热门研究问题“虚假新闻检测”上的案例，问题的新颖性使论文获得了较大的关注。示例论文 M2-1 和 M2-2 是将机器学习作为研究方法的应用案例。M2-1 开展了将机器学习技术应用于预测医院重症监护室病人死亡率的研究，是用旧方法解决新问题的案例。M2-2 围绕机器学习在软件工程中所面临的挑战，以及机器学习如何从软件工程方法中受益开展了研究，是旧问题与旧方法的新组合的案例。这两篇论文是机器学习技术应用于不同领域的案例，均获得了 10 次引用，表明机器学习技术具有较强的推广应用性。整体而言，无论是对于高频问题词还是方法词而言，新颖性仅是从词的新旧层面测量新颖性，而论文发表后的被引量不仅取决于研究问题或研究方法的新颖程度，还受到研究问题本身的适用性的影响。

由上述分析可知，论文研究问题或方法的新颖性与发表后一定时期内能获得的被引量有一定联系，但计算组合新颖性得分与被引量之间的相关性发现，其未达到显著程度，将其可能的原因总结为两点。其一，对于某些研究问题，方法的创新可能获得更大的影响，这是由于有的经典问题本身就带着“光环效应”，它可能是一个还未攻克的难题或瓶颈，也可能本就属于热点问题。其二，论文发表后的被引量或许可以反映一定的新颖性，但却不能完全揭示新颖性或创新性的特征内涵。一方面，对于经典的理论或方法，新颖性的研究会面临一些来自外部的阻力，包括来自现有科学范式的抵制[52]；另一方面，由于受限于研究问题范围的影响，也许在该问题上某方法的新颖性较高，但是这个问题还没有受到相应的关注，或许需要更长的时间才能发现其新颖性并将其纳入后续的研究中。

4.3　新颖性计算方法对比分析

本文提出的基于语义相似度的“问题-方法”组合新颖性计算方法是深度学习模型在词汇新颖性度量上的应用。为进一步比较本文提出的方法与已有方法的差异，利用钱佳佳等[19]提出的基于问题-方法组合共现率的科技论文新颖性计算公式，计算了 200 条分析数据的共现率新颖性，将该方法计算的问题新颖性、方法新颖性、组合新颖性和论文新颖性的结果与本文提出的语义新颖性计算结果进行了比较，如图 7a~图 7d 所示。其中 quesNov、methodNov、combNov 和 paperVov 分别表示问题词、方法词、组合和论文的语义新颖性计算结果，nov_Q、nov_M、nov_Q2M 和 nov_D 分别表示问题词、方法词、组合和论文的词频共现率新颖性计算结果。图 7 中三角形表示本文语义新颖性计算结果，圆点表示共现率新颖性的计算结果。对于单个词的新颖

性，由图 7a 和图 7b 可知，共现率新颖性的计算结果呈现明显的两极分化，集中在新颖性为 1 和新颖性小于 0.6。相较而言，语义新颖性的分布更为均匀，表明基于词汇语义方法捕捉到的新颖性更为精准，这一现象在图 7c 中得到了更为显著的验证。由图 7c 可知，共现率新颖性的计算结果几乎全部集中在新颖性为 1 的区域，表明用该方法计算的组合新颖得分几乎全部是 1，象征着问题-方法组合都是一样的新颖性，然而实际情况中的组合并不都是新颖的，受限于基于词频共现率的新颖性计算的局限性，该方法不能区分更为细微的新颖性差异；而基于语义的新颖性计算方法弥补了该方法的这一局限，能够捕获细微的差异。例如，语义新颖性计算方法计算的 augment reality 和 augment reality game 之间的差异就比 augment reality 和 blockchain 之间的差异要小，前两者在向量空间中更为接近，相似度更高且相对新颖性不如后两者；而基于词频共现率的新颖性计算认为这两组词的相对新颖性是一样的，这将会在较大程度上损失新颖性测度精度。共现率新颖性计算方法中的实验将论文新颖性计算公式中的问题、方法和问题-方法对的权重分别设为 0.25、0.25 和 0.5，即给问题-方法组合更大的权重，该做法在组合新颖性的理论层面是有意义的，然而受限于基于词频共现的新颖性计算方法，论文新颖性结果的整体分布更为紧密（聚集在 0.8 附近），导致新颖性结果的差异更小，如图 7d 所示。总的来说，对比实验的结果表明，基于语义相似度的问题-方法组合新颖性计算方法要优于基于词频共现的新颖性计算方法，前者利用词向量的空间语义捕捉优势能计算出更为精细的新颖性。

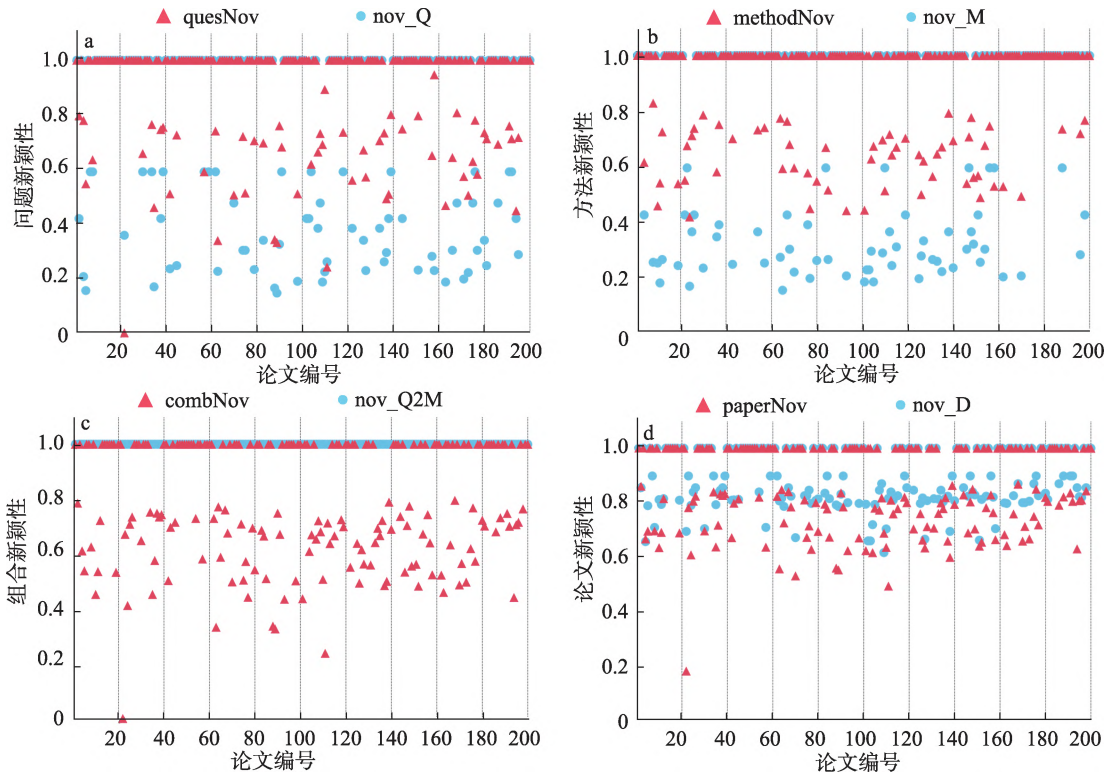

图 7 两种新颖性计算方法对比

5 总结与展望

科学问题作为科学研究的逻辑起点，其解决方法是促进科学研究深入与发展的助推器。科学研究问题和研究方法的识别对科技前沿追踪和创新研究发现具有重要研究意义。近年来随着内容分析研究的流行，从学术文本内容视角对学术论文进行细粒度挖掘，是图书情报学领域的一个新视角，其中学术论文词汇语义功能的识别能够帮助学者快速了解学术论文的核心内容，有助于厘清研究问题、研究方法的演化过程和发展模式，辅助于论文创新识别和新颖性度量研究。

本文以组合创新理论为基础，以具备词汇语义功能的学术论文问题词和方法词为数据，从问题与

方法组合的语义层面研究了论文新颖性度量方法。与已有新颖性计算方法进行比较，发现本文提出的方法能捕获问题词、方法词和问题-方法组合之间更为精细的新颖性差异。本文的不足之处是问题词和方法词的识别效果在某种程度上会影响论文新颖性计算结果。本文提出的计算方法更类似于计算机领域新颖性追踪（novelty track）的方法，该方法是独立于问题词和方法词本身的，但结果的解释却依赖于词汇识别结果，更为准确的词汇识别结果将会使本文的研究结果更具有可解释性和延伸价值，如用于新颖性和影响力之间的关系分析、创新扩散的规律分析等研究上。此外，问题新颖性、方法新颖性及组合新颖性与论文影响力之间的联系也是值得进一步探索的方向。

参 考 文 献

[1] K. R. 波珀. 科学发现的逻辑[M]. 查汝强, 邱仁宗, 译. 北京: 科学出版社, 1986.

[2] Heffernan K, Teufel S. Identifying problems and solutions in scientific text[J]. Scientometrics, 2018, 116(2): 1367-1382.

[3] 索传军, 赖海媚. 学术论文问题知识元的类型与描述规则[J]. 中国图书馆学报, 2021, 47(2): 95-109.

[4] 约瑟夫·熊彼特. 经济发展理论[M]. 郭武军, 吕阳, 译. 北京: 华夏出版社, 2015.

[5] Kogut B, Zander U. Knowledge of the firm, combinative capabilities, and the replication of technology[J]. Organization Science, 1992, 3(3): 383-397.

[6] Arthur W B. The structure of invention[J]. Research Policy, 2007, 36(2): 274-287.

[7] 逯万辉, 苏金燕, 余倩. 学术成果主题新颖性与学术引用的相关关系研究[J]. 情报资料工作, 2018(6): 68-73.

[8] Uzzi B, Mukherjee S, Stringer M, et al. Atypical combinations and scientific impact[J]. Science, 2013, 342(6157): 468-472.

[9] Wang J, Veugelers R, Stephan P. Bias against novelty in science: a cautionary tale for users of bibliometric indicators[J]. Research Policy, 2017, 46(8): 1416-1436.

[10] Tahamtan I, Bornmann L. Creativity in science and the link to cited references: is the creative potential of papers reflected in their cited references?[J]. Journal of Informetrics, 2018, 12(3): 906-930.

[11] Fortunato S, Bergstrom C T, Börner K, et al. Science of science [J]. Science, 2018, 359(6379): eaao0185.

[12] Azoulay P, Graff Zivin J S, Manso G. Incentives and creativity: evidence from the academic life sciences[J]. The RAND Journal of Economics, 2011, 42(3): 527-554.

[13] Lee F. Recombinant uncertainty in technological search[J]. Management Science, 2001, 47(1): 117-132.

[14] Mukherjee S, Uzzi B, Jones B, et al. A new method for identifying recombinations of existing knowledge associated with high-impact innovation[J]. Journal of Product Innovation Management, 2016, 33(2): 224-236.

[15] Boyack K W, Klavans R. Atypical combinations are confounded by disciplinary effects[C]// Proceedings of the 19th International Conference on Science and Technology Indicators, Leiden, The Netherlands, 2014.

[16] Hofstra B, Kulkarni V V, Galvez S M N, et al. The diversity-innovation paradox in science[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(17): 9284-9291.

[17] 任海英, 王德营, 王菲菲. 主题词组合新颖性与论文学术影响力的关系研究[J]. 图书情报工作, 2017, 61(9): 87-93.

[18] 王艳艳, 张均胜, 乔晓东, 等. 基于问题—方法矩阵的文献新颖性评估方法[J]. 情报理论与实践, 2021, 44(2): 90-95.

[19] 钱佳佳, 罗卓然, 陆伟. 基于问题-方法组合的科技论文新颖性度量与创新类型识别[J]. 图书情报工作, 2021, 65(14): 82-89.

[20] 徐庶睿, 卢超, 章成志. 术语引用视角下的学科交叉测度——以PLOS ONE上六个学科为例[J]. 情报学报, 2017, 36(8): 809-820.

[21] 陆伟, 孟睿, 刘兴帮. 面向引用关系的引文内容标注框架研究[J]. 中国图书馆学报, 2014, 40(6): 93-104.

[22] 程齐凯. 学术文本的词汇功能识别[D]. 武汉: 武汉大学, 2015.

[23] 程齐凯, 李信, 陆伟. 领域无关学术文献词汇功能标准化数据集构建及分析[J]. 情报科学, 2019, 37(7): 41-47.

[24] Jarvelin K, Vakkari P. Content analysis of research articles in library and information science[J]. Library and Information Science Research, 1990, 12(4): 395-421.

[25] 王芳, 史海燕, 纪雪梅. 我国情报学研究中理论的应用: 基于《情报学报》的内容分析[J]. 情报学报, 2015, 34(6): 581-591.

[26] 王芳, 王向女. 我国情报学研究方法的计量分析: 以1999~2008年《情报学报》为例[J]. 情报学报, 2010(4): 652-662.

[27] Ferran-Ferrer N, Guallar J, Abadal E, et al. Research methods and techniques in Spanish library and information science journals (2012-2014)[J]. Information Research, 2017, 22(1): paper 741.

[28] 化柏林. 针对中文学术文献的情报方法术语抽取[J]. 现代图书情报技术, 2013(6): 68-75.

[29] Kondo T, Nanba H, Takezawa T, et al. Technical trend analysis by analyzing research papers' titles[C]// Proceedings of the 4th Conference on Human Language Technology: Challenges for Computer Science and Linguistics. Heidelberg: Springer, 2011: 512-521.

[30] Gupta S, Manning C D. Analyzing the dynamics of research by extracting key aspects of scientific papers[C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, 2011: 1-9.

[31] Tsai C T, Kundu G, Roth D. Concept-based analysis of scientific literature[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York:

ACM Press, 2013: 1733-1738.

[32] Tuomaala O, Järvelin K, Vakkari P. Evolution of library and information science, 1965-2005: content analysis of journal articles[J]. Journal of the Association for Information Science and Technology, 2014, 65(7): 1446-1462.

[33] 赵洪, 王芳. 理论术语抽取的深度学习模型及自训练算法研究[J]. 情报学报, 2018, 37(9): 923-938.

[34] 王昊, 邓三鸿, 苏新宁, 等. 基于深度学习的情报学理论及方法术语识别研究[J]. 情报学报, 2020, 39(8): 817-828.

[35] 李贺, 杜杏叶. 基于知识元的学术论文内容创新性智能化评价研究[J]. 图书情报工作, 2020, 64(1): 93-104.

[36] 章成志, 张颖怡. 基于学术论文全文的研究方法实体自动识别研究[J]. 情报学报, 2020, 39(6): 589-600.

[37] 程齐凯, 李鹏程, 张国标, 等. 学术文本词汇功能识别——基于标题生成策略和注意力机制的问题方法抽取[J]. 情报学报, 2021, 40(1): 43-52.

[38] 陆伟, 李鹏程, 张国标, 等. 学术文本词汇功能识别——基于BERT向量化表示的关键词自动分类研究[J]. 情报学报, 2020, 39(12): 1320-1329.

[39] Kaplan S, Vakili K. The double-edged sword of recombination in breakthrough innovation[J]. Strategic Management Journal, 2015, 36(10): 1435-1457.

[40] Yan Y, Tian S W, Zhang J J. The impact of a paper's new combinations and new components on its citation[J]. Scientometrics, 2020, 122(2): 895-913.

[41] Ponomarev I V, Williams D E, Hackett C J, et al. Predicting highly cited papers: a method for early detection of candidate breakthroughs[J]. Technological Forecasting and Social Change, 2014, 81: 49-55.

[42] Luo Z R, Lu W, He J G, et al. Combination of research questions and methods: a new measurement of scientific novelty[J]. Journal of Informetrics, 2022, 16(2): 101282.

[43] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[OL]. (2013-09-07). https://arxiv.org/pdf/1301.3781.pdf.

[44] Pennington J, Socher R, Manning C D. GloVe: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.

[45] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[46] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010.

[47] Su J L. SimBERT: integrating retrieval and generation into BERT [EB/OL]. (2020-07-28). https://github.com/ZhuiyiTechnology/simbert.

[48] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2019: 3615-3620.

[49] Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 8342-8360.

[50] Su J L, Cao J R, Liu W J, et al. Whitening sentence representations for better semantics and faster retrieval[OL]. (2021-03-29). https://arxiv.org/pdf/2103.15316.pdf.

[51] 刘少鹏, 印鉴, 欧阳佳, 等. 基于MB-HDP模型的微博主题挖掘[J]. 计算机学报, 2015, 38(7): 1408-1419.

[52] Kuhn T S. The structure of scientific revolutions[M]. 4th ed. Chicago: The University of Chicago Press, 2012.

（责任编辑 潘 尧）