情报学报　2024 年 2 月　第 43 卷　第 2 期
Journal of the China Society for Scientific and Technical Information, Feb. 2024, 43(2): 167-180

DOI: 10.3772/j.issn.1000-0135.2024.02.004

基于三维主题特征测度的新兴主题识别研究

郑德俊，程　为

（南京农业大学信息管理学院，南京　210095）

摘　要　识别领域新兴主题有利于及时跟踪领域发展的最新动态，为科研工作者的选题以及科研管理者的决策提供情报支撑。本文提出一种基于三维主题特征测度的新兴主题识别方法，基于BERTopic对领域语义知识进行主题建模，以文献为基本单位进行主题表示，构建基于时间、引用和关联的三维主题特征指标框架，用于新兴主题识别；并以文本分类领域为例，验证本文方法的可行性与有效性。研究发现，以文献为基本单位表示主题能辅助主题深入挖掘，三维主题特征指标框架具有较好的适应性与扩展性，本文提出的新兴主题识别方法存在泛化应用的参考价值。在理论层面，能为新兴主题识别的相关研究提供一种可参考的方法和思路；在实践层面，可作为一种参考工具应用于科技情报分析、领域发展态势分析等场景。

关键词　新兴主题识别；主题建模；主题特征测度；文本分类

Emerging Topic Recognition Based on Three-Dimensional Topic Feature Measurement

Zheng Dejun and Cheng Wei

(*College of Information Management, Nanjing Agricultural University, Nanjing*　210095)

Abstract: Recognizing emerging topics is conducive to monitoring the latest trends in development over time, providing valuable information support for researchers' topic selection and research managers' policy decisions. In this study, an emerging topic recognition method based on a three-dimensional topic feature measurement is proposed. First, topic modeling is conducted using domain semantic knowledge from BERTopic, representing topics by documents as the basic unit. Next, a three-dimensional topic feature index framework based on time, reference, and correlation is constructed to identify emerging topics. The feasibility and effectiveness of the proposed method are discussed through empirical studies, using domain data on text classification as examples. The findings reveal that using documents as the basic unit enhances the exploration of topic features, the three-dimensional topic feature index framework demonstrates good adaptability and expansibility, and the proposed method can be generalized application in other domains. At the theoretical level, this work provides a reference method for emerging topic recognition research. At the practical level, it can serve as a reference tool for scientific and technological intelligence analysis and domain development trend analysis.

Keywords: emerging topic recognition; topic modeling; topic feature measurement; text classification

0　引　言

新兴主题是一个相对概念，随时间推移而动态变化，是指在观测时间点的未来一段时间内，具有发展潜力与应用价值并处于萌芽期或上升期阶段的研究主题[1-2]。识别领域新兴主题有利于及时跟踪领

收稿日期：2023-04-30；**修回日期**：2023-08-27

作者简介：郑德俊，男，1968年生，教授，博士生导师，主要研究领域为计量分析、知识服务与评价，E-mail：zdejun@njau.edu.cn；程为，男，1998年生，博士研究生，主要研究领域为文本挖掘与科学计量。

域发展的最新动态，厘清领域前沿研究主题的分布概况，辅助认知学科内在的发展机制与轨迹[3-4]，能够为领域未来发展与应用提供可参考的方向，为科研工作者的选题以及科研管理者的决策提供情报支撑。领域新兴主题识别是情报分析领域的一项热点研究内容，相关研究总结了其研究路径，识别方法部分包括主题建模与主题新兴程度测度两个递进阶段[5]。目前，一方面，主题建模基于关键词、摘要或全文抽取特征词进行主题聚类与表示，但该方法强调特征词的共现或语义关联，通常忽视单篇文献更丰富的内外部特征信息，使得后续主题的特征测度与分析维度相对单一[6]；另一方面，主题新兴程度测度以时间、引文网络与相似度作为切入点，选用一个或少数几个定量指标进行计算，当面临领域特点不同或主题类型不同时，较少指标由于揭示的主题信息不够系统、全面，难以适应具体问题下的合适指标选用[7]。

基于此，本文以完整摘要内容为基本单位进行语义向量表示和主题建模，并以摘要代表单篇文献作为线索串联时间、引用等相关特征信息，探索构建综合时间、引用与关联的多维度指标框架，以更全面、细致地挖掘与表示主题特征，实现新兴主题的识别。在理论层面，能够为新兴主题的特征挖掘与测度提供一套可借鉴与扩展的特征指标框架，以期为新兴主题识别的相关研究提供一种可参考的方法和思路；在实践层面，本文提出的新兴主题识别方法可作为一种参考工具应用于科技情报分析、领域发展态势分析等场景，为新兴主题发现提供数据支持。

1 相关研究

新兴主题识别包括主题识别与新兴主题发现两个阶段的任务。其中，主题识别方法分为两类：一是网络社区发现法，构建共被引网络、直接引文网络、文献耦合网络、共词网络、语义网络等识别研究主题[8-10]；二是基于内容挖掘法，实现主题特征词抽取与表示[11]。新兴主题发现则通常采用主题新兴度、主题新颖性、主题成长度等指标，识别并衡量新兴主题的发展潜力[11-13]。本文以文本内容为主题建模的基础，从基于内容挖掘的主题建模与主题新兴程度测度指标两个方面梳理相关研究。

1.1 基于内容挖掘的主题建模

主题建模是一种无监督的聚类算法，挖掘数据集中语义单元的潜在关联性进而划分主题。在建模算法上，LDA（latent Dirichlet allocation）及其改进模型应用较为广泛[14]，比较有影响力的改进算法有LDA2vec[15]、动态LDA算法[16]、融合高斯函数加权的LDA算法[17]等，其面向科技文献的主要应用场景有主题挖掘[18-19]、主题演化分析[20-21]、学术评价[22-23]等。近年来，随着语义向量嵌入模型的广泛应用，topic2vec[24-25]、BERTopic[26-27]等算法在主题建模中取得了较好效果。在特征词抽取上，相关研究基于年份-关键词词频矩阵识别研究热点[10]；结合客户价值细分模型，对高价值关键词进行筛选进而识别热点主题[28]；融合关键词顺序与词频、文献与关键词关联关系等构建关键词综合影响力模型，进而识别领域热点主题[29]等，使得主题建模的结果具有更强的可解释性。上述研究虽然有在关键词的基础上深入文献摘要或全文进行主题挖掘，但仍以主题特征词为基本单位对主题进行建模，导致文献更丰富的内外部特征信息无法得到充分利用。

1.2 主题新兴程度测度指标

主题新兴程度测度指标的特征基础可以分为时间要素、引文网络与语义挖掘三大类，主要包括如下代表性研究。在时间要素中，时序关系下特征词的首次出现时间、平均时间和拐点时间是衡量主题新颖度的重要指标[12,30]，另外，按时间切片的形式衡量主题的发展历程与成长性也具有有效性[13,31]。引文网络是指基于引用关系（引用、共被引和引用耦合）构建复杂网络，在主题聚类的基础上，分析不同阶段主题的知识流动路径与强度[32-33]，进而明确主题在引用网络中的定位与影响力，将其作为新兴主题不确定性和模糊性的测度指标[34]，实现新兴主题的预测。在语义挖掘中，主要通过相似度计算考量不同主题之间的语义距离[35]，从内容层面测度主题之间的差异性，并将这种差异性表达为主题创新度、主题新颖性或主题新兴度[36-37]。上述主题测度指标都在一定程度上反映了主题新兴程度，并得到了实践检验，具有深入研究的价值，是本文的借鉴对象。但是，在科学计量与评价领域强调具体问题具体分析的要求下[38]，需要集成侧重不同主题特征的已有指标与新指标，以强化多指标融合视角下识别结果的可解释性，提升人工判定的准确性与客观性。

1.3 相关研究述评

目前，一方面，在主题建模上，基于特征词的

主题表示难以充分挖掘与表示主题的多维特征，有必要利用摘要的完整语义信息进行主题建模，使无监督的主题聚类取得更符合领域知识分布特征的结果，具有更强的可解释性，并且能够充分融合文献发表时间等信息丰富主题特征的观测视角；另一方面，以时间、引用或关联 3 个维度中的某一指标作为新兴主题识别的依据，越来越难以适应主题多元化发展的场景，因此，有研究尝试融合不同指标进行新兴主题识别并比单一指标取得了更好的效果[39-40]。然而，这部分研究仍是选用单维度或多维度的少数几个指标，未形成系统的指标框架，在主题特征的挖掘深度与广度上仍具有一定局限性，有必要在筛选现有指标的基础上，提出新的主题特征指标，构建 3 个维度并列共存、相互补充、相互验证的指标框架，以不同主题特征作为切入点识别新兴主题。

更全面的主题语义知识表示是优化主题建模结果的重要手段，系统指标框架的建立是更客观评估主题新兴程度的工具。因此，以摘要作为主题建模、表示与特征测度的基本单位，符合深度主题知识挖掘的需求；探索基于时间、引用和关联的三维主题特征指标框架，对新兴主题识别具有更广泛的应用价值。

2　新兴主题识别方法

2.1　识别方法概述

现有研究中基于特征词的主题表示，存在难以充分揭示主题特征信息、单个或少数指标难以全面且深入地挖掘主题特征信息的问题。本文尝试解决这两个方面的问题，提出新兴主题识别方法的实现框架，如图 1 所示。主要操作如下：第一，获取数据包括领域文献题录及引用数据，通过数据预处理构建摘要语料；第二，以文献为基本单位，基于语义词向量嵌入摘要语料实现领域主题建模及评估，并通过主题置信概率保证主题建模的效果；第三，构建三维主题特征指标框架，分别基于 3 个维度的指标计算识别新兴主题；第四，融合各个维度下的识别结果相互补充来辅助人工判定，汇总形成领域新兴主题识别结果；第五，通过基于 LDA+word2vec+similarity 的方法对比分析、指标相关性计算分析、资料分析法来综合评估本文方法的有效性。其中，充分挖掘文献的语义内涵并以文献摘要为基本单位进行主题建模、构建融合多因素的细粒度主题特征指标框架是本文的创新所在。

2.2　主题建模及评估

2.2.1　BERTopic 主题建模

传统主题建模算法，如 LSA（latent semantic analysis）、PLSA（probabilistic latent semantic analysis）、LDA 等，通过词袋表示进行建模，忽略了词间的语义关系，不能解释文档语料中词的上下文，难以准确表示文档。BERT（bi-directional encoder representations from transformers）及其改进模型能够生成融合文档语料上下文语义信息词向量与句子向量，在该方式下，相似文本在向量空间中更接

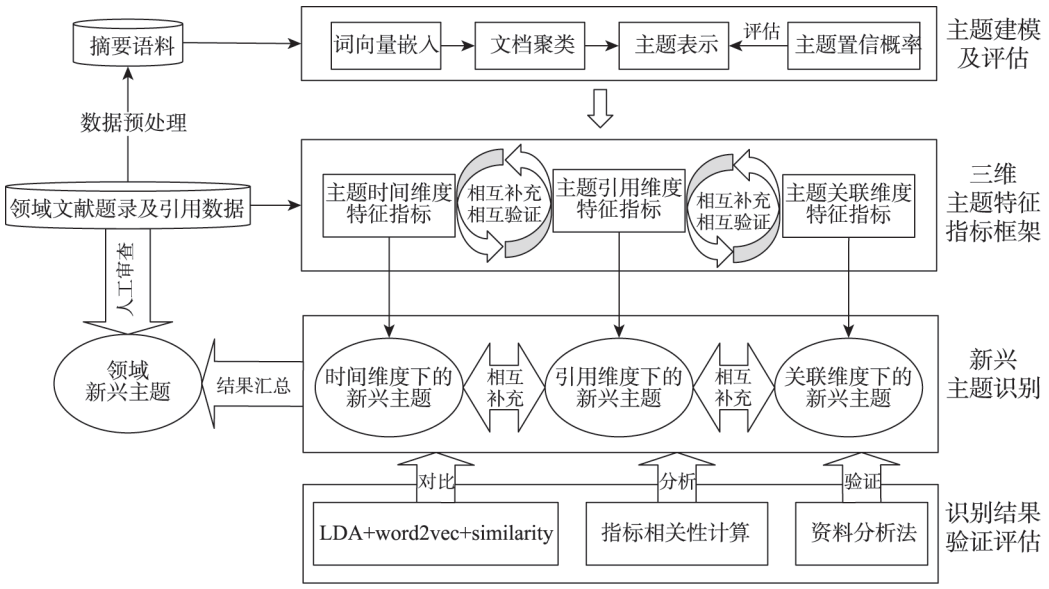

图 1　新兴主题识别方法的实现框架

近[41]。BERTopic融合深度语义向量与传统聚类方法对主题进行建模，本文舍弃抽取特征词进行主题表示的过程，核心思路分为两个阶段：①通过词嵌入的预训练模型得到文档语料的深度语义向量；②通过HDBSCAN（hierarchical density-based spatial clustering of applications with noise）、*k*-means等聚类算法进行聚簇处理，以文档为基本单位实现领域研究主题建模。以摘要文本为语料，基于BERTopic挖掘更深层次的语义信息进行向量表示，进而以摘要为基本单位进行主题聚类与表示，相较于主题特征词，能够融合更丰富的信息如参考文献与施引文献的语义信息作为后续主题特征测度的数据基础。

2.2.2 评估方法

当存在N个样本与K个主题时，BERTopic模型会计算任意一个样本分别聚类至K个主题的概率，并最终将其聚类至最大概率对应的主题；将N个样本被聚类至最大概率对应的主题的平均概率称为主题置信概率，用于评估主题建模的结果。主题置信概率最小取值为$1/K$，此时建模效果最差，N个样本聚类至各主题完全随机；理想状态下，主题置信概率取值为1，即样本属于某一主题的概率为100%且属于其他主题的概率为0%，此时建模达到理想的最佳效果，主题内部样本高度集中，且与主题外部的样本高度分离，聚类结果不存在任何误差可能。在区间$[1/K, 1]$内，主题置信概率取值越高，主题建模效果越好。

2.3 三维主题特征指标框架

主题的时间、引用和语义特征是评估主题新兴程度的主要参考线索，吸纳现有指标并补充新指标，本文构建了包含14个指标的三维主题特征指标框架。

2.3.1 时间维度

时间维度下，主题新兴度计算思路有主题中最新样本的时间属性、主题中最早样本与观测时间的时间间隔、主题内样本的时间属性的平均值等。虽然以上计算方法可以直观反映主题的重要时间点，但是容易受到极端单样本的影响。因此，基于上述指标，着重关注观测时间的最近一段时间的主题样本分布情况，考虑到观测时间点不一定能够以完整的年份为间隔划分样本，将主题中样本的局部分布与领域中样本的整体分布进行对比，按不均等划分时间段的方式衡量主题在观测时间点的发展潜力。时间维度的各主题特征指标如表1所示。

表1 时间维度的主题特征测度指标

指标名称	基本内涵	计算说明	参考指标
最早发文点	主题中发表时间最早的文献，主题发展早，新兴程度一定程度上较低	取值为主题中最早发表文献的发表年份	**首次出现的年份**[28]
最新发文点	主题中发表时间最新的文献，主题最新研究越近，越反映其新兴程度	取值为主题中最新发表文献的发表年份	**最后出现的年份**[28]
平均发文点	主题中各文献的平均发表时间，在一定程度上反映主题的新兴程度	取值为主题中各文献发表年份的平均值	主题热度[42]
相对增长度	相对于领域最近两年的发文增长速度，主题相比于最近两年的发文增长速度的偏差程度，反映主题发展相比于领域发展的增长态势	取值为主题最近两年（观测时间所处年份与上一年份）发文增长速度减去领域最近两年发文增长速度	平均增长率[30]

注：“参考指标”列加粗文本表示直接使用的已有指标，未加粗文本表示对已有指标的借鉴与优化。

2.3.2 引用维度

文献间的引用关系通常用于探测知识的跨主题流动，可以揭示主题间的关联程度、测度主题在领域中所处的位置，相关测度指标主要基于主题的被引频次、施引频次等基本计量指标衡量主题在领域中的核心度。但是，被引文献与施引文献客观存在的时间先后关系在一定程度上反映了知识更新的周期，而上述指标对引用关系潜在的时间关系关注较少。因此，融合主题内部和跨主题引用数据的时间属性，采用表2中的指标来综合测度主题引用维度的特征。

2.3.3 关联维度

相似度计算是衡量主题新兴度的重要手段，当前新兴主题发现研究通常以词共现关系及其频次、词向量等作为相似度计算基础。但是，词在主题中不是孤立存在或以简单的共现关系存在的。因此，本文以标题为基本单位，充分挖掘标题的语义信息构建向量，进而测度主题内外部的语义关联程度。

表 2　引用维度的主题特征测度指标

指标名称	基本内涵	计算说明	参考指标
主题内引度	主题内部的知识流动强度，值越高一定程度上反映主题发展越成熟，属于热门主题	主题内部各文献相互引用次数之和除以主题内文献总数	**主题知识流动强度**[43]
内引时差	主题内部知识的更新周期，取值小时反映主题正处于快速发展阶段	主题内所有引用记录的时间差之和除以引用记录总数	主题扩散度[44]
跨主题入度	知识跨主题流入的强度，值越高反映主题吸收跨主题的知识能力越强，是新兴主题	主题内部各文献的参考文献来源外部主题的记录之和除以主题内文献总数	**主题知识流动强度**[43]
入度时差	主题吸收领域其他主题知识的平均周期，取值小时反映其是及时探测知识融合的新兴主题	主题内所有跨主题施引记录的时间差之和除以跨主题施引记录总数	主题扩散度[44]
跨主题出度	知识跨主题流出的强度，值越高反映主题跨主题的知识扩散能力越强，是经典热门主题	主题内部各文献的施引文献来源外部主题的记录之和除以主题内文献总数	**主题知识流动强度**[43]
出度时差	其他领域主题吸收主题知识的平均周期，取值小时表示其为具有跨主题影响潜力的新兴主题	主题内所有跨主题被引记录的时间差之和除以跨主题被引记录总数	主题扩散度[44]

注：“参考指标”列加粗文本表示直接使用的已有指标，未加粗文本表示对已有指标的借鉴与优化。

同时，仅考虑主题内部或主题间的语义关联会忽略非领域内的相关数据，在当前多学科领域知识交叉融合的背景下，跨领域知识流动程度能够揭示研究的潜在价值，有必要获取文献完整引用数据用于主题特征测度。SimCSE（simple contrastive sentence embedding）基于对比学习的思想，利用自监督学习来提升句子的表示能力，能够充分学习文本的语义知识[45]。因此，本文以文献标题作为输入，选择 sup-simcse-bert-base-uncased 预训练模型，输出表示文献的 768 维语义向量。以主题中各文献语义向量的平均向量作为主题的语义向量。关联维度的各指标详情如表 3 所示。

表 3　关联维度的主题特征测度指标

指标名称	基本内涵	计算说明	参考指标
主题内聚度	主题的内部联结强度表征，取值越大表明主题发展越成熟	主题内任意两条文献语义向量的余弦相似度的平均值	主体密度[46]
主题交叉度	主题间联结强度的表征，取值越大表明主题处于核心地位，受到广泛关注，是领域当前热点主题	主题语义向量与其他任意一个主题语义向量的余弦相似度的平均值	主题相似度[36]
施引丰富度	参考文献的整体差异性程度，值越高代表知识吸纳越丰富，是新兴主题	计算文献任意两条参考文献语义向量的余弦相似度的平均值，由 1 减去该值表示文献的施引丰富度，计算主题内所有文献的施引丰富度的平均值（低于两条参考文献的文献不参与计算）	学科交叉度[47]
被引丰富度	施引文献的整体差异性程度，值越高代表知识扩散越丰富，是具有潜在影响力的新兴主题	计算文献任意两条施引文献语义向量的余弦相似度的平均值，由 1 减去该值表示文献的被引丰富度，计算主题内所有文献的被引丰富度的平均值（低于两条施引文献的文献不参与计算）	学科交叉度[47]

注：“参考指标”列均是对已有指标的借鉴与优化。

3　实证研究

3.1　数据来源与预处理

以“文本分类”领域为例进行实证，限定 Web of Science 核心合集，为提升检索结果与领域的相关性，不额外限制“text classification”这一通用概念，并限制其同义概念必须以词组形式出现。因此，构建检索式“TS=((text classification) OR ("document classification") OR ("document categorization") OR ("text categorization") OR ("text tagging") OR ("document tagging"))”进行检索，检索时间为 2022 年 10 月 2 日，出版日期截至 2022 年 9 月 30 日，得到检索结果 28095 条，经过人工判断初步剔除不相关或弱相关记录，得到 25714 条记录。为获取更规范、完整的数据，在 2022 年 10 月 7 日至 2022 年 10 月 11 日，遍历检索结果中每一文献的 DOI（digital object identifier），通过开源学术搜索引擎 Semantic Scholar 提供的 API（application programming interface）获取文献的题录信息、参考文献与施引文献数据。由于检索结果中部分文献没有 DOI 或文献未被 Semantic

Scholar 收录，最终通过 API 获得 23096 条文献的 JSON（JavaScript object notation）数据，包括 770559 条参考文献记录和 685406 条施引文献记录。本文实验所用数据虽然无法涵盖领域所有文献，但数据已具有一定规模，能够较全面地反映领域的主要研究内容。

基于 NLTK（natural language toolkit）库，对 23096 条文献的摘要进行大写转小写、分词、词形还原和去停用词的预处理，形成摘要语料。由于领域文献的主题均与“文本分类”相关，为防止词频过高的词集中于某一主题进而导致过多文献被分类至该主题，在基本去停用词表的基础上，选取词频超过 10000 的词并基于人工筛选补充停用词表，基本统计信息如表 4 所示。

表 4 补充停用词表的统计信息

词	频次	词	频次
use	37604	study	14748
text	33226	paper	13459
classification	31077	analysis	13350
model	26714	document	13079
method	26280	system	12291
data	23729	show	11370
feature	21679	algorithm	11154
propose	20740	word	11125
result	20130	network	10838
information	15910	performance	10816
base	15163	task	10217
approach	15038	different	10099

3.2 主题建模

基于 23096 条文献的摘要语料进行主题建模实验，主要分为 4 个步骤：①选择 BERTopic 中处理英文文本的默认嵌入模型 all-MiniLM-L6-v2，将每一摘要文本的语义信息映射到一个 384 维的稠密向量空间；②基于默认的 UMAP（uniform manifold approximation and projection）降维算法对摘要向量进行降维，为平衡计算开销与信息量大小，以区间[2,10]内的整数作为候选空间维数；③选择 HDBSCAN 算法进行聚类，以区间[2,100]内的整数作为候选最小聚类样本数；④基于 sklearn 库的文本特征抽取实现主题序列化，训练 BERTopic 模型时，nr_topics 设为“auto”，由模型自动迭代生成最佳主题数。当降维空间维数为 5 时，模型聚类结果相对稳定，重复实验，可以获得相似的聚类结果；当最小聚类样本数分别为 65、66、68 时，模型取得较好效果，主题置信概率均在 90% 左右。对实验结果进行人工审查后，最终确定最小聚类样本数为 66，此时有 4319 条样本属于离群文档或无法划分主题归属的文档，剩余 18777 条样本被模型分别聚类至 42 个主题，主题置信概率为 90.12%，从定量评估的角度可以认为主题建模结果较为合理。

分别在 42 个主题中随机选取部分摘要样本人工研读，总结各主题的基本内涵，42 个主题的基本信息如表 5 所示。将 18777 条摘要样本由高维空间映射至二维语义空间，其在 42 个主题的分布情况如图 2 所示，图中各主题的示例特征词由 BERTopic 模型给出以便区分与可视化，并非主题的实际表示方式。

在定性评估方面，图 2 中主题内部的样本分布相对集中，不同主题间的界限明显，达到了较好的聚类效果。其中，部分主题如“功能性磁共振成像”的样本相对游离孤立，考虑到该部分主题也具有分析的需要，不对聚类结果做进一步人工处理。综合来看，主题建模结果具有较强的可解释性，未出现违反客观事实的重大误判，建模结果可信。

3.3 新兴主题识别

3.3.1 指标计算

基于三维主题特征指标框架，计算各主题特征指标值。在时间维度，有 4 个主题的最早发文点为 1991 年，取值最小；3 个主题最早发文点为 2012 年，取值最大；除极少数主题外，其他主题的最新发文点均为 2022 年。以上两个指标的区分度较差，难以用于各主题的对比分析。计算各主题平均发文点与相对增长度，如图 3a 所示。在引用维度，基于参考文献数据集，构建领域 18777 篇文献的引用网络，共包含 45658 条引用关系，计算指标值如图 3b 和图 3c 所示。图 3b 和图 3c 中的节点数字代表主题编号，节点大小分别代表内引时差和主题内引度。在关联维度，对于施引丰富度与被引丰富度两个指标，构建每一文献的参考文献标题集与施引文献标题集，基于 SimCSE，以标题文本输入 sup-simcse-bert-base-uncased 预训练模型构建语义向量，采用余弦相似度分别计算参考文献标题集与施引文献标题集的相似度矩阵，进而计算指标值，如图 3d 所示。

3.3.2 识别结果

由图 3a 可以发现，“虚假信息监测”的平均发

表 5　42 个主题基本信息

主题编号	主题名称	样本数	主题编号	主题名称	样本数
0	医学信息挖掘	5644	21	创新技术发现	167
1	图像识别	2768	22	文本聚类	163
2	社交平台用户情感分析	1940	23	灾害识别与预警	151
3	基于统计的文本分类技术	934	24	领域自适应迁移学习	131
4	语音识别	707	25	融合多模态的情感分类	131
5	基于神经网络的文本分类	483	26	实体关系抽取	126
6	语义网	462	27	虚假信息监测	125
7	教学实践中的知识分类	452	28	法律文本智能挖掘	118
8	多标签文本分类	353	29	功能性磁共振成像	117
9	作者风格与文本体裁分类	308	30	程序错误识别与分类	114
10	主题挖掘技术	295	31	自动问答	109
11	句子级情感分析	281	32	命名实体识别	96
12	方面级情感分析	270	33	非英语语言文本分类	96
13	网络安全系统	247	34	自动摘要	93
14	引文分类与推荐	212	35	音乐流派与情感分类	81
15	项目合同文本挖掘	209	36	网络讽刺信息识别	78
16	网络平台负面言论检测	203	37	术语语义化与消歧	75
17	对抗式生成网络模型	202	38	图神经网络技术	74
18	金融信息分类与挖掘	202	39	空间知识标注与计算	74
19	文本过滤	182	40	关键词抽取	68
20	事故识别	168	41	用户消费满意度情感分析	68

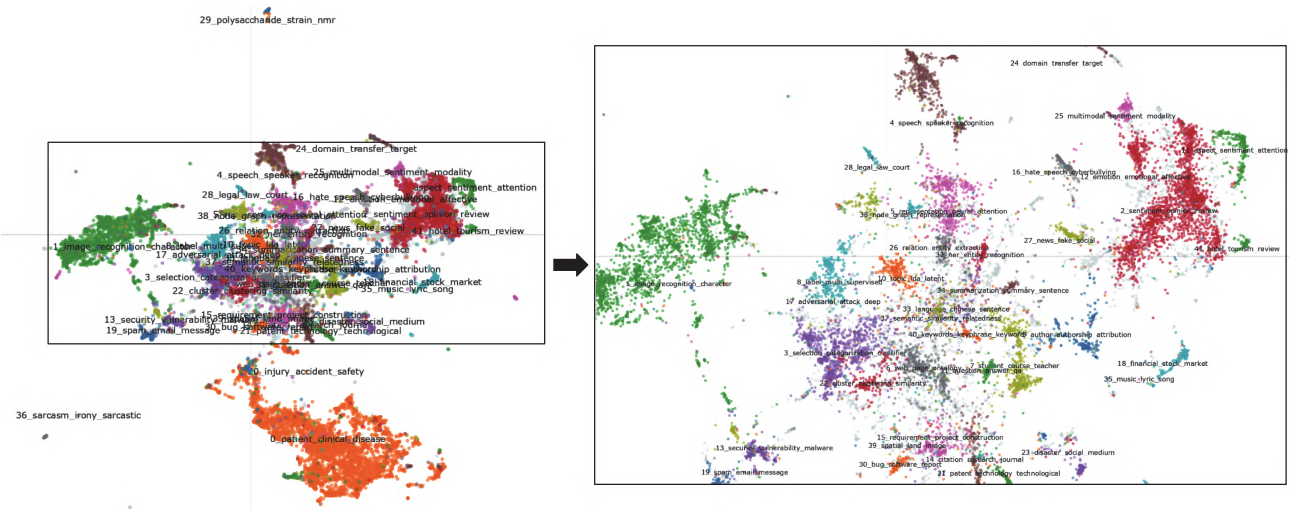

图 2　摘要样本在二维语义空间的可视化分布

文点（2020.008）距观测点最近，表明在 2020 年前后有关虚假信息监测的研究引起领域学者的广泛关注；另外，“网络平台负面言论检测”“图神经网络技术”“对抗式生成网络模型”等主题的平均发文点都在 2019 年以后，表明它们具有较新的研究场景与内容。“音乐流派与情感分类”和“创新技术发现”两个主题的平均发文点都在 2016 年左右，但二者的相对增长度遥遥领先其他主题，均超过 0.58，表明与领域整体发展相比，这两个主题的研究在当下具有贴合外部需求、融合新兴技术并且更与时俱进的研究内容，迎来了新的小高峰。

由图 3b 可以发现，主题的出度时差与入度时差基本都在两年以上，而“融合多模态的情感分类”“网络平台负面言论检测”“图神经网络技术”等主题内引时差、出度时差与入度时差均较小，知识更新的速度较快，更具有产生新知识的潜力。

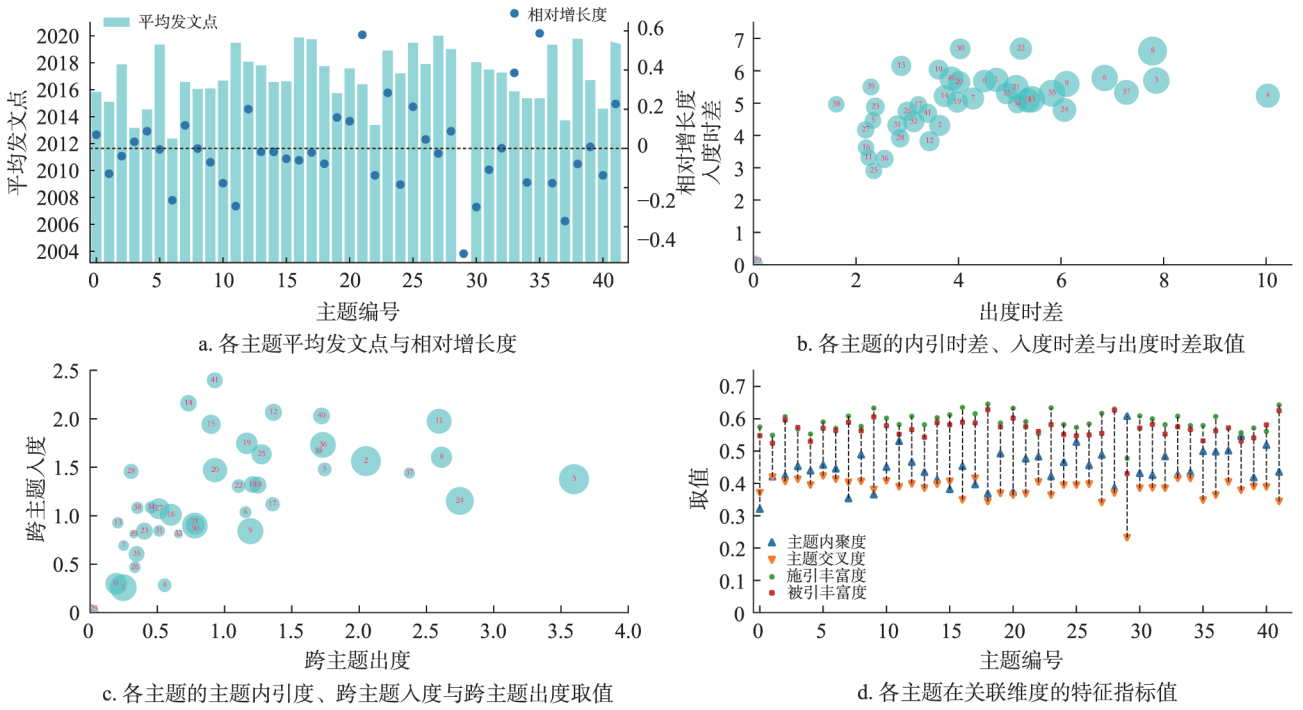

图 3 各主题三维主题特征指标值对比图

由图 3c 可以发现，“用户消费满意度情感分析”“引文分类与推荐”等跨主题入度较高且跨主题出度较小，即它们广泛吸收了同领域其他主题的知识，但还未在其他主题大范围知识扩散，在将来具有更广阔的发展前景。“空间知识标注与计算”“自动问答”“自动摘要”等主题内引度较小且跨主题出度较小，表明其仍处于发展上升期，具有发展潜力。

由图 3d 可以发现，“医学信息挖掘”由于样本数最多，主题内聚度最低；而“功能性磁共振成像”是一个比较孤立的主题，主题内聚度高而交叉度低，在前面各项指标中它也是一个具有明显特征差异的主题，与文本分类领域本身关联性较弱。“作者风格与文本体裁分类”“金融信息分类与挖掘”等主题内聚度与主题交叉度均较低，表明主题具有相对新颖、独到的研究内容，仍处于发展上升期。“金融信息分类与挖掘”“法律文本智能挖掘”“用户消费满意度情感分析”“作者风格与文本体裁分类”等被引丰富度与施引丰富度均较高，表明这些主题在未来具有跨领域知识融合与发现的价值。

在 2022 年 10 月这个观测点上，综合图 3 的各指标结果，判定文本分类领域的新兴主题，如表 6 所示。结合表 6，基于对部分领域文献的内容分析，归纳文本分类领域未来的新兴研究主题如下：①文本分类前沿技术的改进，如图神经网络、对抗式生成网络等模型的优化；②文本分类方法在通用场景下的下游任务创新，如网络信息的智能挖掘与质量评估、情感分析在多模态数据与细粒度分类标准上的扩展等；③文本分类方法结合领域知识特征在垂直领域的深入应用，如金融、法律等领域的知识结构化建模与价值发现；④文本分类与其他相关技术的协同优化，如增强自动问答与自动摘要的自然语言可理解性、提升多模态数据细粒度分类的准确性等。

3.3.3 结果与分析

在实验中，最早发文点和最新发文点由于区分度较差未作为新兴主题识别的依据，因此，对任意

表 6 三维特征指标测度下的文本分类领域新兴主题识别结果

主题特征维度	新兴主题
时间维度	虚假信息监测、网络平台负面言论检测、图神经网络技术、对抗式生成网络模型、音乐流派与情感分类、创新技术发现
引用维度	用户消费满意度情感分析、引文分类与推荐、空间知识标注与计算、自动问答、自动摘要、融合多模态的情感分类、网络平台负面言论检测、图神经网络技术
关联维度	作者风格与文本体裁分类、金融信息分类与挖掘、法律文本智能挖掘、用户消费满意度情感分析

一个主题，由3个维度共12个指标值定量描述。例如，“虚假信息监测”可表示为[({平均发文点: 2020.008}, {相对增长度: −0.0259}); ({主题内引度: 0.9280}, {内引时差:2.4138}, {跨主题出度: 0.5120}, {出度时差: 2.1875}, {跨主题入度: 1.0720}, {入度时差: 4.1716}); ({被引丰富度: 0.5548}, {施引丰富度: 0.6169}, {主题内聚度: 0.4882}, {主题交叉度: 0.3409})]。对任意一个主题，分别取12个指标值在42个主题中的升序排名值，对于平均发文点等与新兴程度正相关的指标，单指标的新兴程度量化结果为排名值，对于主题内聚度等与新兴程度负相关的指标，单指标的新兴程度量化结果为43减去排名值；3个维度量化结果取下属单指标量化结果的平均值；整体量化结果取3个维度量化结果的平均值。例如，“虚假信息监测”在时间、引用与关联维度的新兴程度量化结果分别为24、28.6667、25.25，整体量化结果为25.9722。汇总表6中的新兴主题，基于上述处理过程绘制图4，主题标签大小取决于整体新兴程度。

在图4中，“作者风格与文本体裁分类”“自动摘要”等主题仅在单一维度新兴程度较高，“引文分类与推荐”“融合多模态的情感分类”等主题同时在2个维度新兴程度较高，而仅有“用户消费满意度情感分析”“法律文本智能挖掘”等少数主题在3个维度新兴程度均较高。各主题在三维空间中分布较为离散，且极少有主题在3个维度的新兴程度均较高，这说明从不同维度分类讨论新兴主题是有价值且符合客观结果的，结合具体主题的实际特征，综合考量各维度的指标值能够有效识别不同类型的新兴主题。

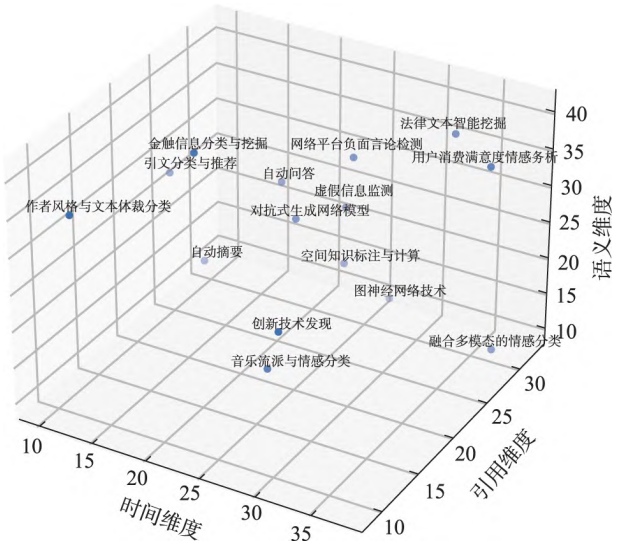

图4　主题新兴程度在三维空间的量化分布

3.4　方法评估

3.4.1　方法对比分析

递进式组合使用LDA、word2vec与similarity是目前较为常用的新兴主题识别方法，为验证本文识别方法（以下简称“A方法”）的有效性，增加LDA+word2vec+similarity方法（以下简称“B方法”）作为实验对照组。首先，对于预处理后的摘要语料，基于LDA识别领域主题及主题特征词；其次，利用word2vec训练主题特征词的词向量；最后，基于特征词向量，计算主题与其他主题余弦相似度的平均值，用1减去该值来表示主题的新兴度。基于gensim库训练LDA主题模型，训练时通过语料库的次数为10，文档-主题分布的先验alpha与主题-词分布的先验均设置为“auto”，以区间[2,50]内的整数作为候选主题数，主题数为27时主题困惑度最低，达到最优建模效果；主题由分布概率较高的N个特征词表示，设置N为30，设置词向量维度为100。经计算，得到对照组新兴主题识别结果，如表7所示。

通过表7可以发现，与A方法识别结果相比，B方法难以有效表示主题的语义信息，导致难以识别更细致的主题差异，同时，因为仅有新兴度一个指标，难以综合考虑主题的各特征因素，导致结果具有一定的局限性。例如，排名第1、3、6位的主题在内涵上隶属对照实验中的“医学信息挖掘”，而实验数据集更多是文本分类技术与应用的相关文

表7　对照组实验的新兴主题识别结果

排名	主题	部分特征词	新兴度得分
1	精神类医学信息挖掘	disorder; child; mental; depression; symptom; diagnostic; personality; cognitive; criterion; dsm	0.7779
2	中文法律文本智能挖掘	chinese; legal; sign; law; case; gesture; judgment; tibetan; license; article; court	0.7727
3	生物信息挖掘	specie; cancer; cell; food; plant; soil; breast; genetic; group; water; disease	0.7546
4	情感识别	emotion; event; emotional; expression; detection; temporal; job; traffic; affective; sarcasm	0.7401
5	语音识别应用	speech; speaker; recognition; language; voice; error; audio; discourse; acoustic; automatic	0.7275
6	患者风险评估	patient; risk; disease; high; treatment; low; group; rate; year; clinical; case; conclusion; injury	0.7257

献，大量医学相关文献的缺失导致对照实验高估了其新兴度，导致误判；排名第4位的主题“情感识别”属于新兴主题，但其主题范畴过于宽泛，A方法能识别出更细分的新兴主题“用户消费满意度情感分析”“融合多模态的情感分类”“网络平台负面言论检测”等；单个指标的局限性也导致A方法识别出的新兴主题在对照实验中被大量漏识。

通过与B方法的对比分析，可以认为本文方法具有新兴主题识别粒度更细、更准确、更全面的优势。

3.4.2 指标相关性分析

对42个主题的各指标值进行相关性分析，样本数小于50条，因此，选择夏皮洛-威尔克（Shapiro-Wilk，S-W）法进行检验，变量检验结果均不符合正态分布；根据变量的数据类型和分布形态，选取斯皮尔曼（Spearman）相关系数法对变量进行相关性检验，结果如表8所示。从表8可以发现，一方面，14个指标两两之间大部分不存在显著的相关关系，这表明各指标具有差异性，形式上相互补充，能够从不同角度更加全面地揭示主题特征，这在3.3节中的实证得以验证。另一方面，部分指标之间存在正向或负向的显著性相关关系，正向相关性反映了相关指标之间内在的协同性，如内引时差、入度时差与出度时差3个指标彼此存在正向相关性，从图3b中也可发现这一规律，“融合多模态的情感分类”等主题在3个指标上的取值均较小，在图中的分布相对集中且与其他主题存在明显特征差异，能够更准确地发现新兴主题；负向相关性反映了指标之间在极端值内涵上的互斥性，如最早发文点通常较早，反映了一个主题的发展时间与成熟

表8 各指标相关性分析

	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a. 最早发文点	1													
b. 最新发文点	0.245	1												
c. 平均发文点	**0.787****	0.264	1											
d. 相对增长度	−0.086	0.264	0.134	1										
e. 主题内引度	0.170	−0.045	0.071	−0.092	1									
f. 内引时差	**−0.587****	0.110	**−0.726****	0.057	0.179	1								
g. 跨主题出度	0.271	0.264	0.033	−0.179	**0.335***	0.177	1							
h. 出度时差	**−0.454****	0.264	**−0.690****	0.018	0.108	**0.895****	0.295	1						
i. 跨主题入度	**0.315***	0.264	0.234	0.022	0.185	−0.074	**0.684****	−0.016	1					
j. 入度时差	**−0.533****	0.264	**−0.576****	−0.063	−0.085	**0.642****	−0.009	**0.631****	−0.159	1				
k. 被引丰富度	0.131	0.264	0.198	0.130	0.109	0.090	0.269	0.106	**0.415****	−0.036	1			
l. 施引丰富度	**0.429****	0.264	**0.593****	0.255	0.147	−0.187	0.082	−0.172	0.269	−0.285	**0.766****	1		
m. 主题内聚度	0.258	−0.264	−0.005	−0.196	0.040	−0.187	0.272	−0.233	0.178	−0.254	**−0.517****	**−0.489****	1	
n. 主题交叉度	−0.252	0.264	−0.282	−0.153	−0.204	0.177	**0.348***	**0.359***	0.159	0.304	−0.062	**−0.309***	−0.147	1

注：*表示 $p<0.05$，**表示 $p<0.01$，***表示 $p<0.001$；粗体表示存在显著相关关系的两个变量的相关性系数。

度，与新兴程度相互对立，因此，它与内引时差等多个指标存在负向相关性，从侧面反映了这些指标作为新兴主题识别依据的科学性与合理性。

3.4.3　识别结果验证

由于新兴主题识别是一种预测性任务，没有一种通用的定量标准衡量识别结果的准确性[5,30]，因此，采用资料分析法验证文本分类领域新兴主题识别结果的科学性。在 2022 年 1 月 1 日至 2023 年 3 月 1 日，收集 Web of Science 核心合集、中文社会科学引文索引（Chinese Social Sciences Citation Index，CSSCI）与中国科学引文数据库（Chinese Science Citation Database，CSCD）里文本分类相关的中英文综述，对其内容进行深入分析。相关文献谈到如何优化深度学习模型、改进前沿技术[48]；文本分类将重点关注算法改进、信息拓展以及二者的相互融合，并探索特定领域应用[49]；应当加强情感分析与知识问答对自然语言的理解能力[50]等。综合来看，本文结合识别结果总结的新兴研究主题基本贴合了上述相关表述，证明了本文方法的有效性与准确性。

4　结论与展望

4.1　研究结论

（1）以文献为基本单位表示主题能辅助主题深入挖掘。传统以特征词进行主题表示的方法通常依靠词间共现或关联关系进行主题特征测度，在一定程度上限制了对主题的深入挖掘。本文以文献为基本单位进行主题表示。其一，主题表示由一系列词的集合替换为包含更丰富语义信息的摘要文本集合，使得主题能够涵盖更细致的语义内容，为主题内涵的凝练与分析提供更易理解的数据支撑，更精准地溯源主题的某一文献进行深入分析；其二，以文献为基本单位表示主题能够扩展主题的内外部关联类型，如特征词之间难以准确表示的引用关系可以通过文献精准表示，为主题的关联分析提供不同的切入点；其三，以文献为基本单位能够融合更多样的相关数据辅助主题特征的测度，如领域文献与非领域文献之间基于引用的相关关系，可以探测知识跨领域流向特定主题的方向及强度，丰富主题观测的思路。

（2）三维主题特征指标框架具有较好的适应性与扩展性。本文从时间、引用与关联 3 个维度构建了包含 14 个指标的主题特征指标框架，从更全面的角度深入考察新兴主题区别于一般主题的特征因素，并通过各指标的整体协同分析定量评估主题，以得到更客观的识别结果。一方面，3 个维度的各指标有不同的侧重因素，以定量指标形式科学化、精细化地展示与描述主题，有利于辅助人工更客观、有效地判定新兴主题。在宏观上，立足领域考量其知识生产及扩散的特殊性选取适用指标对具体问题具体分析，如引用数据较少的领域应适当降低相关指标的重要性；在微观上，针对不同主题类型可以综合各项指标进行差异化解读与评估，如交叉主题、迎来新发展机遇的经典主题、新诞生的主题等。另一方面，指标框架具有较强的扩展性，在面向特定领域或特定需求时，可灵活新增具有测度价值的定量或定性指标，建立更完善适用的指标框架，更好地服务于新兴主题发现。

（3）本文提出的新兴主题识别方法存在泛化应用的参考价值。首先，实验所用数据的时间跨度为 1991—2022 年，领域经历了较长的发展历程，共有 18777 篇文献聚类至 42 个主题，数据具有一定规模，实验数据选取具有合理性。其次，文本分类本身是一个随技术进步不断迭代发展的领域，同时与医学、金融、法律等领域形成具有交叉领域特色的研究内容。除此之外，与图像分类、语音识别等相关领域存在诸多共通性与差异性，实证领域特点鲜明，具有一定代表性。最后，实验证明了本文方法的可操作性，通过与 LDA+word2vec+similarity 方法的对比分析、指标相关性分析及资料分析法讨论了实验结果的科学性与有效性。因此，在结合其他泛化领域自身特征的基础上，参考本文方法，选取适用指标组合使用能够帮助人工动态识别领域新兴主题。

4.2　未来展望

（1）以文献为线索充分挖掘主题特征。以文献为线索，可以串联文献题录各字段信息、全文本内容、引文信息等各类型数据，以更全面的数据作为深入观测主题的切入点。本文初步探讨了从不同视角评估主题新兴度的可行性，未来可从两个方面做进一步的探索。一是在数据范围广度上，融合更多元的数据拓展可能影响新兴主题预测的因素和维度，完善现有指标框架，使得新兴主题的识别结果更加准确、丰富。例如，参考文献与施引文献的发表时间及其被引量等、评估主题跨领域的知识扩散

强度及可赋予更高权重的高影响力知识扩散强度等，还可关联专利、政策文件等多源数据观测新兴主题的其他相关特征因素。二是在指标内涵深度上，基于更深层次的语义知识挖掘，定量描述主题的细粒度特征以更微观地观测主题。例如，针对引用维度的相关指标，可以通过更细致的分类，综合考量引用位置、引用情感与引用动机等语义信息，使得指标在简单计数的基础上向更复杂的语义计量进行深化。

（2）智能评估的需求下减少人工干预。当前，新兴主题的自动识别无法完全消除人工干预，主要体现为两个方面。一是在主题建模阶段，由于不同主题独特的内涵与语境，算法无法准确理解并概括主题，仍依赖人工对主题内涵进行总结；二是在新兴主题识别结果分析阶段，指标虽然能客观量化主题，但是无法自动分析新兴主题的内容，依赖人工进一步评估与解读。针对上述问题，可以明确各指标的权重，提出综合多因素的唯一指标，以主题排序的形式自动生成新兴主题，但仍无法解决新兴主题的智能解读问题。因此，还可以借助现有生成式人工智能技术，面向特定场景下的需求，基于领域知识输入对预训练模型进行继续训练，在人工提示的基础上，根据定量指标计算结果，由机器智能化概括主题内涵并生成新兴主题的定性评估报告，在减少人工干预的同时，通过人机合作达到相互参照、相互验证的效果，以实现更高效的新兴主题发现。

5 结 语

本文提出一种基于三维主题特征测度的领域新兴主题识别方法，具体包括两个方面的优势：①基于 BERTopic 模型对领域知识进行主题建模，相较于特征词，采用包含更丰富语义信息的摘要文本进行主题表示，能够挖掘到更深层次的主题特征；②构建融合时间、引用与关联因素的三维主题特征指标框架，对 14 个指标进行计算与观测，能够基于更广泛、更深入的主题特征挖掘实现更有效的新兴主题发现。

然后，利用文本分类领域相关数据进行了实证研究，验证了本文方法的可行性，识别出虚假信息监测、网络平台负面言论检测、自动问答、自动摘要、图神经网络技术、作者风格与文本体裁分类等新兴主题，并将识别结果归纳为文本分类前沿技术的改进、文本分类方法在通用场景下的下游任务创新、文本分类方法结合领域知识特征在垂直领域的深入应用、文本分类与其他相关技术的协同优化四个方面。通过方法对比分析、指标相关性分析和资料分析法验证了本文方法的有效性，说明该方法具有泛化应用至其他领域的价值。

此外，本文存在一定的局限性：①仅获取领域文献的参考文献与施引文献标题，没有利用相关的更丰富的信息做进一步的挖掘与分析；②仅初步讨论了各指标的有效性，需要进一步明确各指标的权重，进而提出综合性指标。未来研究将做进一步的改进。

参 考 文 献

[1] 卢超, 侯海燕, Ding Ying, 等. 国外新兴研究话题发现研究综述[J]. 情报学报, 2019, 38(1): 97-110.

[2] Liang Z T, Mao J, Lu K, et al. Combining deep neural network and bibliometric indicator for emerging research topic prediction [J]. Information Processing & Management, 2021, 58(5): 102611.

[3] 段庆锋, 闫绪娴, 陈红, 等. 基于媒介比较的学科新兴主题动态识别——altmetrics 与引文数据的融合方法[J]. 情报学报, 2022, 41(9): 930-944.

[4] 钱旦敏, 楼筱湾, 王华麟, 等. 我国信息资源管理学科及其邻近学科视角下的新兴主题识别[J]. 图书馆论坛, 2023, 43(9): 54-64.

[5] 郝雯柯, 杨建林. 基于语义表示和动态主题模型的社科领域新兴主题预测研究[J]. 情报理论与实践, 2023, 46(2): 184-193.

[6] 刘春江, 刘自强, 方曙. 基于 SAO 的技术主题创新演化路径识别及其可视化研究[J]. 情报学报, 2023, 42(2): 164-175.

[7] 贺德方, 潘云涛. 科技评价的内涵、分类与方法辨析及完善策略[J]. 情报学报, 2023, 42(1): 1-9.

[8] Yang S L, Han R Z, Wolfram D, et al. Visualizing the intellectual structure of information science (2006-2015): introducing author keyword coupling analysis[J]. Journal of Informetrics, 2016, 10 (1): 132-150.

[9] Hou J H, Yang X C, Chen C M. Emerging trends and new developments in information science: a document co-citation analysis (2009-2016)[J]. Scientometrics, 2018, 115(2): 869-892.

[10] 马铭, 王超, 周勇, 等. 基于语义信息的核心技术主题识别与演化趋势分析方法研究[J]. 情报理论与实践, 2021, 44(9): 106-113.

[11] 荣国阳, 李长玲, 范晴晴, 等. 主题热度加速度指数——学科研究热点识别新方法[J]. 图书情报工作, 2021, 65(20): 59-67.

[12] 段庆锋, 陈红, 刘东霞, 等. 基于 LSTM 模型与加权链路预测的学科新兴主题成长性识别研究[J]. 现代情报, 2022, 42(9): 37-

48, 142.

[13] 叶光辉, 王灿灿, 李松烨. 基于SciTS会议文本的跨学科科研协作新兴主题识别及预测[J]. 情报科学, 2022, 40(7): 126-135.

[14] 张东鑫, 张敏. 图情领域LDA主题模型应用研究进展述评[J]. 图书情报知识, 2022, 39(6): 143-157.

[15] Moody C E. Mixing Dirichlet topic models and word embeddings to make lda2vec[OL]. (2016-05-06). https://arxiv. org/pdf/1605.02019.pdf.

[16] 胡吉明, 陈果. 基于动态LDA主题模型的内容主题挖掘与演化[J]. 图书情报工作, 2014, 58(2): 138-142.

[17] 张小平, 周雪忠, 黄厚宽, 等. 一种改进的LDA主题模型[J]. 北京交通大学学报, 2010, 34(2): 111-114.

[18] Qin Y W, Qin X Z, Chen H H, et al. Measuring cognitive proximity using semantic analysis: a case study of China's ICT industry [J]. Scientometrics, 2021, 126(7): 6059-6084.

[19] 沈思, 李沁宇, 叶媛, 等. 基于TWE模型的医学科技报告主题挖掘及演化分析研究[J]. 数据分析与知识发现, 2021, 5(3): 35-44.

[20] Liu H L, Chen Z W, Tang J, et al. Mapping the technology evolution path: a novel model for dynamic topic detection and tracking [J]. Scientometrics, 2020, 125(3): 2043-2090.

[21] Wu H, Yi H F, Li C. An integrated approach for detecting and quantifying the topic evolutions of patent technology: a case study on graphene field[J]. Scientometrics, 2021, 126(8): 6301-6321.

[22] Zhang Y J, Ma J L, Wang Z J, et al. Collective topical PageRank: a model to evaluate the topic-dependent academic impact of scientific papers[J]. Scientometrics, 2018, 114(3): 1345-1372.

[23] 赵蓉英, 戴祎璠, 王旭. 基于LDA模型与ATM模型的学者影响力评价研究——以我国核物理学科为例[J]. 情报科学, 2019, 37(6): 3-9.

[24] 王婷婷, 韩满, 王宇. LDA模型的优化及其主题数量选择研究——以科技文献为例[J]. 数据分析与知识发现, 2018, 2(1): 29-40.

[25] 徐月梅, 吕思凝, 蔡连侨, 等. 结合卷积神经网络和Topic2Vec的新闻主题演变分析[J]. 数据分析与知识发现, 2018, 2(9): 31-41.

[26] Abuzayed A, Al-Khalifa H. BERT for Arabic topic modeling: an experimental study on BERTopic technique[J]. Procedia Computer Science, 2021, 189: 191-194.

[27] 张敏, 沈嘉裕. 突发公共卫生事件中政务短视频主题与用户行为的关联演化研究[J]. 情报杂志, 2023, 42(3): 181-189.

[28] 孙佳佳, 李雅静. 基于关键词价值细分的高价值热点主题识别方法研究[J]. 情报学报, 2022, 41(2): 118-129.

[29] 李慧, 王若婷. 基于文献—关键词双模网络的热点识别方法研究——以数字人文领域为例[J]. 情报理论与实践, 2022, 45(11): 107-114.

[30] 许海云, 张慧玲, 武华维, 等. 新兴研究主题在演化路径上的关键时间点研究[J]. 图书情报工作, 2021, 65(8): 51-64.

[31] Liu X Y, Porter A L. A 3-dimensional analysis for evaluating technologyemergenceindicators[J].Scientometrics,2020,124(1):27-55.

[32] Zhang S T, Han F. Identifying emerging topics in a technological domain[J]. Journal of Intelligent & Fuzzy Systems, 2016, 31(4): 2147-2157.

[33] Li M N, Wang W S, Zhou K Y. Exploring the technology emergence related to artificial intelligence: a perspective of coupling analyses[J]. Technological Forecasting and Social Change, 2021, 172: 121064.

[34] Xu H Y, Winnink J, Yue Z H, et al. Multidimensional scientometric indicators for the detection of emerging research topics[J]. Technological Forecasting and Social Change, 2021, 163: 120490.

[35] 陈虹枢, 宋亚慧, 金茜茜, 等. 动态主题网络视角下的突破性创新主题识别: 以区块链领域为例[J]. 图书情报工作, 2022, 66(10): 45-58.

[36] Kim E H J, Jeong Y K, Kim Y H, et al. Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction[J]. Journal of Informetrics, 2022, 16(1): 101242.

[37] 孙晓玲, 陈娜, 丁堃. 基于组合概率的技术主题新颖性研究[J]. 情报学报, 2022, 41(10): 1015-1023.

[38] 杨瑞仙, 高鑫宁, 董克. 我国学术代表作评价研究进展[J]. 图书情报工作, 2022, 66(17): 129-140.

[39] Xu S, Hao L Y, An X, et al. Emerging research topics detection with multiple machine learning models[J]. Journal of Informetrics, 2019, 13(4): 100983.

[40] 高楠, 高嘉骐, 陈洪璞. 新兴技术识别与演化路径分析方法研究——以集成电路领域为例[J]. 情报科学, 2023, 41(3): 127-135, 172.

[41] Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure[OL]. (2022-03-11). https://arxiv.org/pdf/2203.05794.pdf.

[42] Gao Q, Huang X, Dong K, et al. Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec[J]. Scientometrics, 2022, 127(3): 1543-1563.

[43] 王伟, 梁继文, 杨建林. 基于引文网络的领域主题层次结构识别方法研究[J]. 图书情报工作, 2022, 66(17): 81-92.

[44] Kim M, Baek I, Song M. Topic diffusion analysis of a weighted citation network in biomedical literature[J]. Journal of the Association for Information Science and Technology, 2018, 69(2): 329-342.

[45] Gao T Y, Yao X C, Chen D Q. SimCSE: simple contrastive learning of sentence embeddings[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6894-6910.

[46] 王康, 高继平, 潘云涛, 等. 多位态研究主题识别及其演化路径方法研究[J]. 图书情报工作, 2021, 65(11): 113-122.

[47] 黄菡, 王晓光, 王依蒙. 复杂网络视角下的研究主题学科交叉测度研究[J]. 图书情报工作, 2022, 66(19): 99-109.

[48] 刘航冶, 富铁楠, 杨勇. 互联网开源文本情报智能分析技术综述[J]. 情报杂志, 2023, 42(2): 12-16.

[49] 淦亚婷, 安建业, 徐雪. 基于深度学习的短文本分类方法研究综述[J]. 计算机工程与应用, 2023, 59(4): 43-53.

[50] Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning—based text classification: a comprehensive review[J]. ACM Computing Surveys, 2022, 54(3): Article No.62.

（责任编辑 魏瑞斌）