

情报理论与实践
Information Studies: Theory & Application
ISSN 1000-7490, CN 11-1762/G3

《情报理论与实践》网络首发论文

题目：“论文—专利”关联视角下的新兴技术识别研究
作者：张凯，吕璐成，韩涛，赵亚娟
网络首发日期：2024-05-13
引用格式：张凯，吕璐成，韩涛，赵亚娟.“论文—专利”关联视角下的新兴技术识别研究[J/OL]. 情报理论与实践.
<https://link.cnki.net/urlid/11.1762.G3.20240511.1616.002>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

●张 凯^{1,2}, 吕璐成^{1,2}, 韩 涛^{1,2}, 赵亚娟^{1,2}

(1.中国科学院文献情报中心, 北京 100190; 2.中国科学院大学经济与管理学院信息资源管理系, 北京 100190)

“论文—专利”关联视角下的新兴技术识别研究*

摘要: [目的/意义]基于“论文—专利”关联视角, 文章通过新兴技术抽取与量化研究的方式识别“从论文到专利的创新链条上对未来技术发展趋势有引领作用”的新兴术语。[方法/过程]创新性地结合了 Termolator 算法和 GPT 提示学习的新术语提取方法。该方法通过对比实验, 探索了 GPT 提示学习在术语抽取中的应用效果, 并且显著提高了术语抽取的准确性和召回率。进一步, 利用 Minibatch Kmeans++ 算法对术语识别结果进行聚类, 形成技术主题, 并通过多维指标量化分析方法对这些新兴技术主题进行识别和分类。[结果/结论]将新兴技术术语划分为热点型、前沿型、应用型和潜在型新兴术语, 实现对技术术语主题的有效识别和分类。研究成果表明, 该方法能够有效揭示大模型研究领域中对未来技术发展趋势有引领作用的新兴技术, 为新兴技术术语识别提供新途径。[局限]技术术语向量化表征和新兴技术主题识别指标阈值确定存在一定局限性, 需要进行进一步研究。

关键词: GPT 提示学习; 大语言模型; 新兴技术识别; 文本挖掘

Research on Emerging Technology Identification from the Perspective of "Paper - Patent" Correlation

Zhang Kai^{1,2}, Lü Lucheng^{1,2}, Han Tao^{1,2}, Zhao Yajuan^{1,2}

(1. National Science Library, Chinese Academy of Sciences, Beijing 100190; 2. Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190)

Abstract: [Purpose/significance] Based on the perspective of "paper-patent" correlation, this paper identifies emerging terms that "lead the development trend of future technology in the innovation chain from paper to patent" through the way of emerging technology extraction and quantitative research. [Method/process] This research innovatively adopts Termolator algorithm and GPT prompt learning to extract new terms. This method explores the application effect of GPT prompt learning in term extraction through comparative experiments, and significantly improves the accuracy and recall rate of term extraction. Furthermore, Minibatch Kmeans++ algorithm is used to cluster the results of term recognition, forming technical topics, and these emerging technical topics are identified and classified by multidimensional index quantitative analysis method. [Result/conclusion] This paper divides the emerging technical terms into hot, cutting-edge, applied and potential emerging terms, and realizes the effective recognition and classification of the topics of technical terms. The results show that this method can effectively reveal the emerging technologies that lead the development trend of future technologies in the field of large model research and provide a new way for the identification of emerging technology terms. [Limitations] This study has certain limitations in the vectorization representation of technical terms and the determination of the threshold of emerging technology topic recognition indicators which deserve further study.

Keywords: GPT prompt learning; large language models; emerging technology identification; text mining

0 引言

新兴技术识别 (Emerging Technology Identification, ETI) 是一个系统性的过程, 用于发现和评估正处于发展初期但有潜力显著影响社会、经济或科技领域的技术。这一过程涉及对技术发展趋势的监测、分析和预测, 旨在早期识别那些可能引起重大变革的技术创新^[1]。

自 2022 年 11 月 ChatGPT 发布以来, 大语言模型 (Large Language Models, LLMs), 简称大模型, 在各种自然语言处理任务上表现出的强大性能, 引起了学术界和产业界的广泛关注。大模型是指基于神经网络的大规模、预训练统计语言模型^[2], 拥有数十亿参数, 其规模可达数百 GB 甚至更大。这种庞大的规模赋予了大模型强大的表达和学习能力^[3], 使其在多项 NLP 任务上的表现趋近于人类水平^[4]。

当前, 大模型领域的研究正迅速成为学术界和工业界的焦点。随着新技术的不断涌现, 各国政府开始制定相关政策, 以促进大模型技术的发展和应用。在此背景下, 大模型领域新兴技术识别显得尤为关

* 本文为国家自然科学基金青年科学基金项目“技术距离视角下的技术融合模式、特征及预测研究”(项目编号: 72304268)和国家社会科学基金项目“支撑 AI4Science 的科技图书馆知识服务内容研究”(项目编号: 22BTQ019)的成果。

键。这不仅有助于理解和预测 AI 科技发展趋势，对于企业和政府来说，能够早期识别并评估潜在技术创新至关重要^[5]。对企业而言，它能揭示最新技术趋势和研究方向，助力创新和维持市场竞争力。对政府而言，它支持科技政策制定，促进科研资金有效分配，推动大模型技术的健康和可持续发展^[6]。

尽管新兴技术识别具有显著的实用价值，但如何准确、高效地识别和评估这些技术依然是一个难题。针对这一问题，本文提出了一种基于文献计量和文本挖掘方法，旨在从专利和学术论文文本中识别大模型领域新兴术语，并采用多维指标量化分析方法进行新兴技术主题识别和分类。

1 文献综述

关于新兴技术的定义，学术界尚未形成统一标准。Burmaoglu 等^[7]认为“新兴技术的涌现是高度创造性科学网络中的一个循环过程。它在特定的时间框架内表现出新颖性、协同性、趋势的不规则性、高功能性和连续性。”美国情报高级研究计划局（IARPA）FUSE 计划提出了新兴技术指标的标准，分别是新颖性、持久性、社区性和增长性。Rotolo 等^[8]在 FUSE 和其他先前研究的基础上提出了新兴技术的 5 大特征：激进的新颖性、相对的快速增长性、连贯性、潜在的重大影响以及不确定性和模糊性。他们认为新兴技术的涌现是一个持续的、动态的过程，其突出的特点在于对未来的影响。

新兴技术识别的一种方法以技术主题为研究对象，采用基于主题模型的方法识别新兴技术。这种方法能够从大量科学文献中检测到快速成长的技术主题，主要利用文本数据中丰富的语义信息，通过“术语向量表征”识别新兴术语，将识别出的新兴术语聚类成主题后，对新兴主题特征进行量化来识别新兴技术。孙蒙鸽等^[9]借助 Node2Vec 网络表征方法量化新兴技术“过去、现在及未来”三大时间维度特征——“融合性、新颖性及潜在的科学影响力”，用特征值筛选技术主题是否具有新兴性。曹锟等^[10]在共词网络结构特征和语义表示的基础上，构建模型进行新兴术语的遴选和新兴分数的量化，并运用 Node2Vec 图表示学习算法对新兴术语的向量进行编码。针对新兴主题特征量化，Wang^[11]提出新兴技术主题识别框架：新颖性、增长性、一致性和科学影响力，并在图书情报学（LIS）文献上进行实验。Xu 等^[12]综合衡量新兴主题 5 个特征：新颖性、增长性、持久和连贯性、高影响力、不确定性和模糊性，并进一步将新兴技术主题划分为 7 种模式，针对不同新兴主题提供了不同研发布局策略。这类方法的不足之处在于向量化表征准确率不高，由此导致新兴技术特征与计算得到的计量指标之间存在可解释性不强的问题。

另一部分学者以细粒度的技术术语为研究对象，对新兴术语的特征进行量化，进而识别出新兴术语。针对新兴术语特征量化，Carley 等^[13]提出了技术涌现指标（Technology Emergence Indicator, TEI），即 TEI 方法。TEI 方法以特定的科学技术领域为重点，挖掘某一时期的论文或专利，计算技术出现的 4 个属性：新颖性、持久性、社区性和成长性，识别新兴技术。Liu 等^[14]在识别出新兴术语的基础上通过一个三维评估框架系统地评估了新兴术语，主要考虑术语的三个属性：持久性、社区性和成长性，并且通过灵敏度分析表明这些指标具有良好的健壮性。Jiang 等^[15]基于技术知识流（Technological Knowledge Flow）的视角，通过对知识吸收、增长和扩散等方面的全面剖析，构建了术语的前向和后向被引网络并阐释了反映新兴术语属性的多维指标。这类研究基于快速增长的技术创新过程，过于偏重主观增长性指标，而忽略了术语语义特征，且论文和专利从提交申请到发表存在时间上的滞后性^[16]。

针对目前研究中存在的问题，本文结合文献计量和文本挖掘，提出一种新兴技术的抽取和识别方法。本文改进传统技术涌现指标方法，避免过于偏重主观增长性问题，同时加强指标可解释性，对识别出的新兴技术进行了系统性评价，从而提升了新兴技术识别的准确性。针对论文和专利从提交申请到发表存在时间上的滞后性，本文综合利用大模型领域内的高价值论文和专利数据，充分考虑从学术研究到专利申请转化过程中的时间滞后现象，结合多维指标量化分析方法，提出全新的新兴术语识别方法。

2 研究方法

本文基于技术涌现指标，并结合文献计量和文本挖掘方法，提出了一种新兴技术的识别和评价框架，如图1所示。第一，数据获取与预处理阶段。本文爬取2017—2023年人工智能领域A类会议论文，并针对大模型领域制定检索策略获取2017—2023年专利数据，在完成数据去重与清理后，提取标题和摘要字段，而后对论文和专利文本进行筛选。第二，新兴术语抽取与聚类阶段。采用了一种结合 Termolator 算法和 GPT 提示学习的方法抽取技术术语。进一步，利用 MiniBatch Kmeans++ 算法对术语进行聚类，形成技术主题。第三，新兴技术识别与分类阶段。考虑技术术语的三个维度特征——新颖性、热点性、不确定性，通过 CRITIC 权重计算方法计算论文新兴分数和专利新兴分数，结合多维指标量化分析方法，将新兴技术术语划分为热点型、前沿型、应用型和潜在型新兴术语。最后对识别出的新兴术语进行解读，验证本文方法的科学性和准确性。

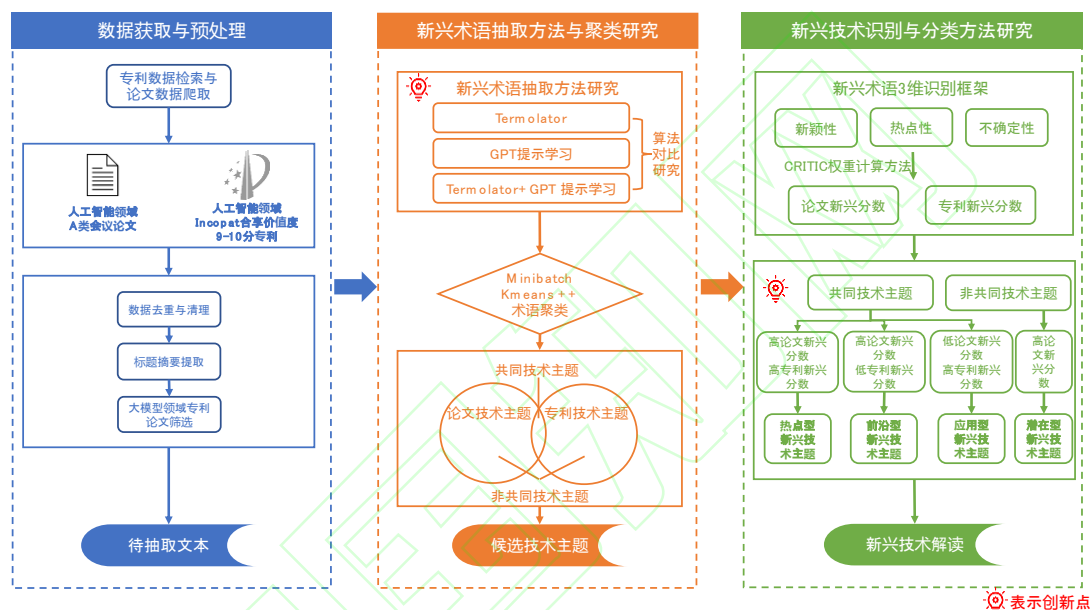


图1 新兴术语抽取与识别研究框架

Fig.1 Research framework of emerging term extraction and identification

2.1 新兴术语抽取方法研究

2.1.1 Termolator 术语抽取算法 本文采用 Termolator 算法提取候选技术术语。Termolator^[17]是由美国纽约大学开发的开源高性能术语提取系统，结合了多种不同的方法来提升术语覆盖范围和精度。Termolator 的运行流程如图2所示，分为三个阶段：①术语分块和缩写，识别文本中的潜在术语；②分布式排序，

根据术语在前景和背景语料库中的出现频率对其进行排序；③重排序，根据格式良好性度量和相关性度量对第二阶段的前 N 个术语重新排序，得到最终的候选术语。

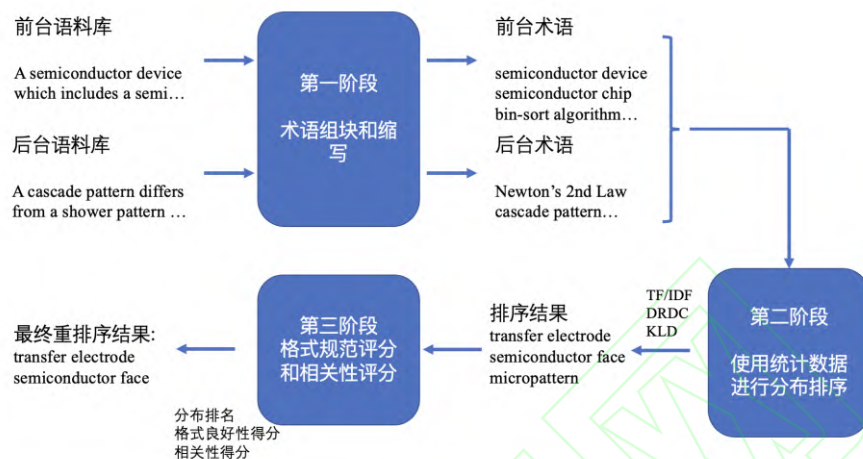


图 2 Termolator 系统框架
Fig.2 Termolator system framework

2.1.2 GPT 提示学习 GPT 提示学习（Prompt Learning）方法利用预训练的语言模型，通过设计特定的提示或问题形式，指导模型针对特定的任务生成所需输出。提示本质上充当了模型输入和期望输出之间的桥梁，使得无须或仅需少量微调即可适应新任务，显著降低了传统机器学习方法中所需的大量标注数据和计算资源。

本文通过 GPT 提示学习的方法进行术语抽取，分成任务描述、提出要求标准、指定输出格式、提供少量示例（Few-shot Prompting）4 部分。提示词设计结构如图 3 所示。



图 3 提示词设计框架
Fig.3 Prompt word design framework

2.2 多指标量化分析

本文采用多维度量化分析方法，通过对新兴技术主题的新颖性、热点性以及不确定性等指标进行综合考量，识别新兴技术。具体而言，笔者考虑主题出现的时间点来评估新颖性，考虑主题在文献中出现

的频率来衡量热点性，考虑主题在活跃期与基准期的信息熵之差表示主题的不确定性。通过这种方法能够平衡新颖性、热点性和不确定性，更全面地理解新兴技术特性。

1) 新颖性。新颖性考虑术语出现的时间维度，即一个技术术语出现的时间越晚，则认为其新颖程度越高^[18]。本文利用某主题所包含论文的平均发表年或专利的平均申请年来表示每个主题的新颖性，具体计算见公式（1）：

$$\text{Novelty}_k = \frac{\sum_{i=1}^n y_i}{n} \quad (1)$$

式中， Novelty_k 表示主题 k 的新颖性； n 表示主题 k 所包含的专利或论文数量； y_i 表示论文的发表年或专利的申请年。

2) 热点性。热点性考虑术语的出现频率，即术语在活跃期出现的频率越高，则认为术语的热点性越高。本文利用某主题所包含的文档数量与所有主题所包含平均文档数量的比值来表示主题的热点性，具体计算见公式（2）：

$$\text{Hot}_k = \frac{Kn_{kt}}{\sum_1^K n_{kt}} \quad (2)$$

式中， Hot_k 表示主题 k 的热点性； n_{kt} 表示时间段 t 内主题 k 所包含的文献数； K 表示时间段 t 内的主题数量。

3) 不确定性。不确定性考虑术语研究方向的模糊性，Rotolo 等^[8]认为新兴技术的突出特点在于对未来的影响，因此在出现阶段具有不确定性和模糊性。本文更倾向于识别目前尚处于早期阶段，不确定性还在增加的新兴技术，根据主题在活跃期与基准期的信息熵之差表示主题的不确定性，具体计算见公式（3）、公式（4）：

$$\text{Uncertainty}_k = H_{\text{活跃期}} - H_{\text{基准期}} \quad (3)$$

$$H_k = -\sum_{i=1}^n p_i \log_2 p_i \quad (4)$$

$$p_i = \frac{m_{it}}{M_t}$$

式中， H_k 表示主题 k 的信息熵； p_i 表示主题 k 中术语在文献中出现的概率； n 表示主题 k 中术语的数量； m_{it} 表示时间段 t 内术语 i 出现的次数； M_t 表示时间段 t 内主题 k 中所有术语出现的总次数。

本文对以上三个指标进行归一化处理后，利用 CRITIC 法计算各测度指标的权重和新兴分数。CRITIC 法是一种客观赋权法^[19]，通过结合标准间的相关性评估，综合考虑了数据对比强度和冲突性这两项指标，消除了相关性较强的指标对最终结果的影响，并减少指标之间的信息重叠^[20]，更有利于得到可信的评价结果。

2.3 新兴技术识别

通过对学术论文和专利数据中识别出的技术主题进行比较分析，将其分为共同存在主题和非共同存在主题，再结合指标计算得出的论文新兴分数和专利新兴分数，进一步将识别出的技术主题进行分类。

由于专利申请往往基于先前的学术成果^[21]，从学术研究到专利申请的转化过程表现出显著的时间滞后现象^[22]。因此相对于论文中呈现的技术，专利中的技术表现出了时间上的滞后性。对于在论文和专利中共同存在的技术主题，可以进一步划分为热点型新兴技术、前沿型新兴技术和应用型新兴术语。仅在学术论文中出现的技术主题，根据它们的论文新兴分数划分为潜在型新兴技术。以下是分类的具体特征：

1) 热点型新兴技术主题：指新近出现且对其所属领域产生显著影响的技术主题。它们受到该领域研究者的高度关注，通常在论文和专利技术主题中都具有较高的新兴分数，反映了它们在学术和实际应用中的活跃度及其潜在的影响力。

2) 前沿型新兴技术主题：具备显著创新性和快速发展特征的技术主题，这些技术主题的潜在影响力和发展深度，有望超越现有的科学理解和技术预期。前沿型新兴技术主题通常在论文技术主题中具有较高的新兴分数。而在专利中的新兴分数较低。

3) 应用型新兴技术主题：指那些在论文中已经发展成熟，但在专利中新近出现的技术主题，它们再近期内学术研究的活跃度有所下降，但对特定领域的实践应用仍具有深远意义。应用型新兴技术属于共同存在主题，并且在论文技术主题新兴分数中得分较低，在专利技术主题新兴分数中得分较高。

4) 潜在型新兴技术主题：涵盖那些极具创新性且当前尚未成为研究热点的技术主题。这些技术因其高度的新颖性和潜在的应用前景，被认为有可能在未来演变为热点型新兴技术。结合专利技术的滞后性，高楠等^[23]研究认为潜在型新兴术语仅出现在一种数据源中，因此本研究选择仅出现在论文数据中出

现并具有较高论文新兴分数的技术主题为潜在型新兴技术主题。该类技术主题的特点是在论文技术主题新兴分数中得分较高。

3 基于大模型领域的实证研究

3.1 大模型领域技术术语抽取

本文综合考虑会议论文和高质量专利中的技术术语，爬取 2017—2023 年中国计算机学会（CCF）推荐的人工智能领域 A 类会议论文，如表 1 所示，完成去重、缺失值处理后得到论文 52627 篇。在 incopat 数据库中，构建专利检索词进行专利检索，得到了 2017—2023 年合享价值度 9~10 分专利的大模型领域专利 18002 篇。获取到每条文献数据的标题、摘要、作者、所属机构等字段，并通过关键词筛选进一步筛选其中大模型领域的文献。

表 1 人工智能领域 A 类会议论文数据集
Tab. 1 Data set of Class A conference papers in the field of artificial intelligence

会议	年份
AAAI	2017—2023
NeurIPS	2017—2022
ACL	2017—2023
CVPR	2017—2023
ICCV	2017, 2019, 2021
ICML	2017—2023
IJCAI	2017—2023

本文对从论文和专利文本数据中抽取出技术术语方法展开了相关研究，人工抽取 200 篇论文标题数中 374 个术语，采用多种术语抽取算法进行对比实验，实验结果如表 2 所示。

表 2 术语抽取算法对比实验
Tab. 2 Comparison experiment of terminology extraction algorithms

	完全匹配			部分匹配		
	准确率（%）	召回率（%）	F1 值（%）	准确率（%）	召回率（%）	F1 值（%）
算法 1	67.25	31.76	43.15	78.95	37.29	50.66
算法 2	95.54	69.65	80.54	96.28	71.55	82.09
算法 3	83.25	90.61	86.77	88.83	96.69	92.59

注：算法 1 表示 Termolator 术语抽取算法；算法 2 表示 gpt 提示学习抽取方法；算法 3 表示“Termolator + gpt”术语抽取方法。

从术语抽取实验结果可以发现，单纯使用 Termolator 算法的实验结果较差，损失了大部分候选术语。本文创新性地采用了 GPT 提示学习方法，通过合理构建提示词，大幅提升了术语抽取准确率。通过 Termolator 和 GPT 提示学习结合方法，分别抽取 13806 个论文术语和 4389 个专利术语。

3.2 术语向量化表征与主题识别

本文采用 OpenAI 词嵌入模型对抽取术语进行向量化表征。OpenAI 词嵌入模型将文本转换为数值形式的向量，使计算机能够处理和理解自然语言。它的主要特点包括文本深层含义的理解，高维数据的压缩，并可用于后续聚类等机器学习算法的实现。

本文选择的词向量模型为 OpenAI 最新推出的 TEXT-EMBEDDING-3-SMALL，其术语表征的结果更为准确。本文设置向量维度为 128 维，方便后续对术语进行聚类。

表 3 OpenAI 模型参数设置
Tab.3 OpenAI model parameter Settings

参数	设定值	参数说明
model	TEXT-EMBEDDING-3-SMALL	词向量模型
dimension	128	向量维度

得到术语的向量表征后，采用 MiniBatch-KMeans ++算法对术语进行技术主题聚类。该算法通过小批量（mini-batch）处理来提高大规模数据集上的聚类效率，并改进了初始聚类中心的选择方法，以提高算法的收敛速度和聚类质量。采用轮廓系数法确定聚类结果质量，当聚类数为 28 时，轮廓系数得分最高，因此选择将术语聚成 28 个类簇，专利聚成 25 个类簇，如图 4 和图 5 所示。

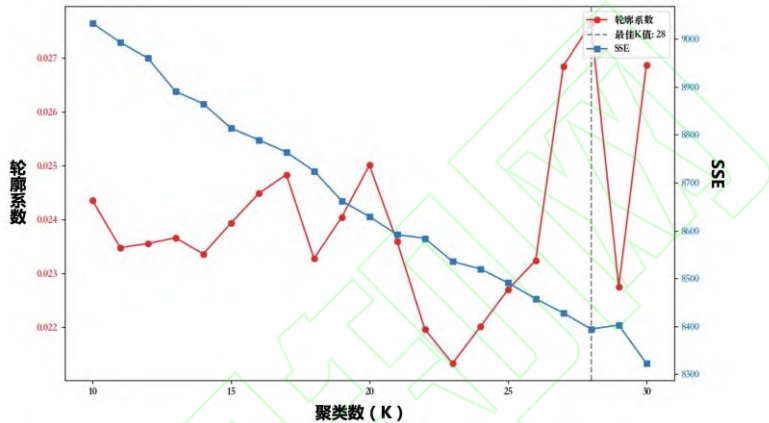


图 4 论文术语聚类表现

Fig.4 Clustering performance of paper terms

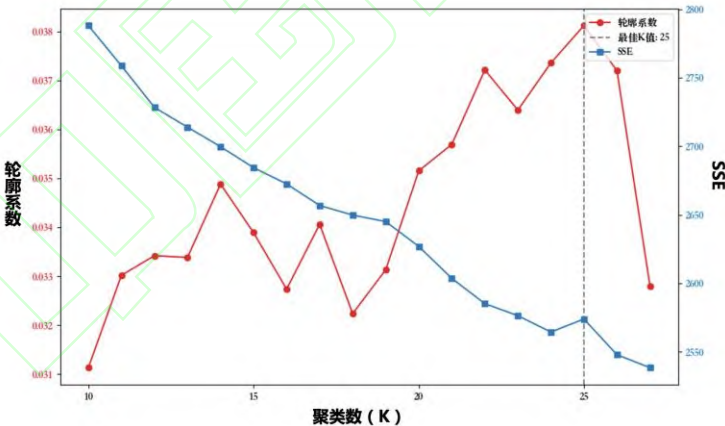


图 5 专利术语聚类表现

Fig.5 Clustering performance of patent terms

由于篇幅限制只展示前三技术主题内容，如表 4 和表 5 所示，论文技术主题更偏向于技术研究，由于专利的商业化程度更高，专利技术主题更偏向领域应用。

表 4 论文技术主题识别结果

Tab.4 Result of paper technical topic identification

主题编号	技术主题名称	部分主题词
Paper_1	知识蒸馏和领域知识注入	Injecting Background Knowledge, Few Sample Knowledge Distillation, Long-Tail Knowledge, Model Interpretability Knowledge Distillation, Knowledge Graph-Augmented Abstractive Summarization, Decision-Based Knowledge Distillation

2	Paper_ 多模态大模型预 训练	Multilingual Model, Agnostic Multilingual Information Retrieval, Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension, Anisotropic Cross-Lingual Model, Multilingual Transformer, Dataset Multilingual Simile Dialogue
3	Paper_ 人工智能中的伦 理考虑和社会影响: 公平、偏见和质量控 制	Personality Profiling Models, Privacy Policies, Detoxifying Language, Extracted Healthcare Terms, Quality Assessment, Statistical Bias, Hyena Hierarchy, Inductive Biases', Extrinsic Fairness Evaluation

表 5 专利技术主题识别结果
Tab.5 Results of patent technology subject identification

主题编号	技术主题名称	部分主题词
Patent_1	计算机视觉与模式识别	Character Recognition Improved Training, Lightweight Video Classification Method, Full-Supervision Video Pedestrian Re-Identification Method, Character Recognition Network Model Training, Forgery Detection of Human Face Image, Face Recognition
Patent_2	文本分类与多标签分类	Paper Classification Method, Picture Classification, Fusion-Based Text Classification Method, Hierarchical Label Text Classification, Text-Level Text Coherence Classification Method, AI Training-Oriented Multi-Modal Data Set Labeling Method, Fine-Grained Paper Classification Method
Patent_3	大模型微调与应用	False Comment Identification Model Based On a Structural Attention Enhancement Mechanism, Training Entity Type Recognition Model, Abstractive Summarization Language Model, Explanation Analysis Model, Unsupervised Video Abstraction Model

3.3 多维指标量化分析

本文分别对论文和专利中识别出的技术主题进行多维指标量化分析，并通过各特征的筛选条件（Novelty > 2021, Hot > 0, Uncertainty > 0）初步筛选得到了 20 个论文技术主题和 15 个专利技术主题，相关的结果如表 6 和表 7 所示。

表 6 论文技术主题多维指标计算结果
Tab.6 Calculation results of multidimensional indicators of technical topics of the paper

主题名称	新颖性	热点性	不确 定性	新兴分数
少样本学习与零样本学习	2021.992	0.313	0.948	0.668
预训练语言模型的整合和领域专业化	2021.758	4.290	0.170	0.626
持续学习和基础模型迁移	2021.636	0.403	0.834	0.481
提示学习和预训练模型的应用	2021.666	0.445	0.680	0.444
扩散模型	2021.657	0.876	0.512	0.417
文本与图片生成	2021.697	1.080	0.374	0.403
知识蒸馏和领域知识注入	2021.694	0.693	0.465	0.401

人工智能中的伦理考虑和社会影响:公平、偏见和质量控制	2021.612	0.827	0.530	0.400
大模型领域多模态内容合成、检索和解释	2021.351	1.599	0.642	0.390
推理与问答系统	2021.770	0.340	0.409	0.385
大模型对话系统评估与分析	2021.452	2.383	0.311	0.383
针对不同应用的 Transformer 结构创新和增强	2021.774	0.644	0.139	0.318
自监督学习、弱监督学习与半监督学习	2021.469	1.395	0.302	0.305
多模态大模型预训练	2021.615	0.801	0.166	0.271
持续学习与迁移学习	2021.331	1.902	0.174	0.242
多模态学习与上下文表示学习	2021.201	0.367	0.666	0.230
语义解析与句法分析	2021.453	0.797	0.245	0.228
对抗学习和对比学习	2021.485	0.481	0.203	0.201
大模型的微调方法	2021.453	0.461	0.022	0.122
机器翻译与命名实体识别	2021.316	0.491	0.139	0.106

表 7 专利技术主题多维指标计算结果
Tab. 7 Calculation results of multi-dimensional indicators of patent technology topics

主题名称	新颖性	热点性	不确定性	新兴分数
多模态内容合成与模型解释	2021.272	5.631	0.155	0.594
基于知识图谱的特征表征与预训练	2021.351	0.311	0.588	0.485
计算机视觉与模式识别	2021.002	0.418	0.696	0.424
注意力机制在神经网络中的应用	2021.181	1.509	0.248	0.335
目标状态监测与控制	2021.274	2.377	0.034	0.317
多模态学习与跨模态融合	2021.657	0.129	0.017	0.293
医疗信息处理与智能医疗系统	2021.256	0.041	0.307	0.293
目标检测与识别	2021.501	0.467	0.071	0.289
网络安全性检测与风险评估	2021.360	0.770	0.123	0.285
知识表示与融合、文本摘要生成	2021.492	0.219	0.075	0.271
大模型微调与应用	2021.466	0.160	0.066	0.253
实体识别与智能检测	2021.308	0.319	0.139	0.245

文本分类与多标签分类	2021.330	0.170	0.133	0.240
自动驾驶与模式识别	2021.310	0.147	0.033	0.181
文本生成与交互式对话系统	2021.131	0.105	0.134	0.167

3.4 新兴术语识别与分类

本文在识别共同存在主题时，计算了专利文献主题与论文文献主题间的余弦相似度，将主题对数按照相似性大小的分布情况绘制了如图 6 所示的趋势图。

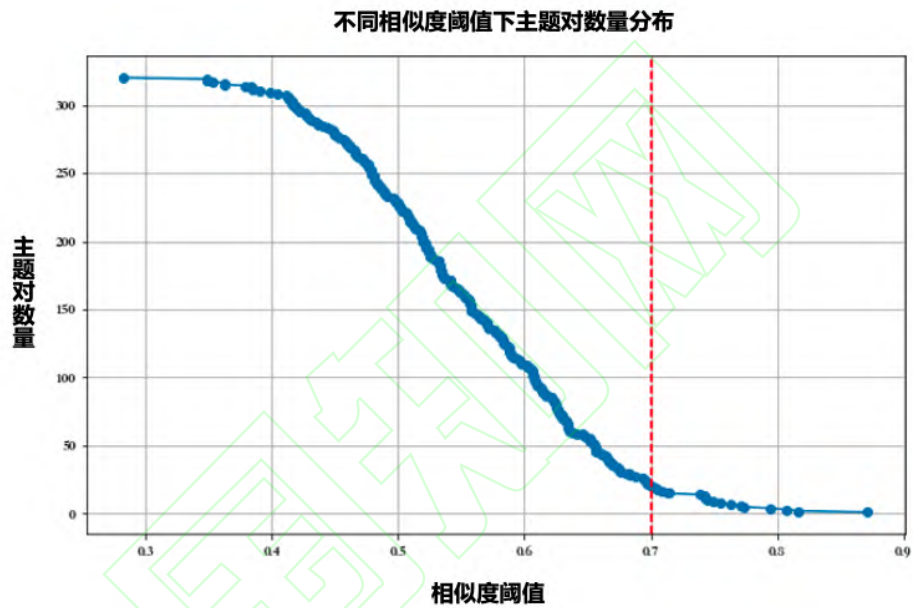


图 6 不同相似度阈值下主题对数量分布
Fig. 6 Number distribution of topic pairs under different similarity thresholds

本文假设大多数主题对的内容是彼此无关的，即多数主题对为非共同存在主题，这些无关主题对之间的相似度大于 0.7 后，满足条件的主题对数量下降趋势迅速变缓，这说明相似度大于 0.7 的各个主题对之间的相似度趋于接近，因此可将该阈值作为区分共同存在主题与非共同存在主题的标准。

基于相关研究中设定指标阈值的经验，通过对各个主题内容进行分析，将论文主题新兴分数大于 0.4 识别为“高论文新兴分数”，将专利主题新兴分数大于 0.4 识别为“高专利新兴分数”。根据本文定义的新兴技术类别标准，对各技术主题进行了归类整理，结果如图 7 所示。

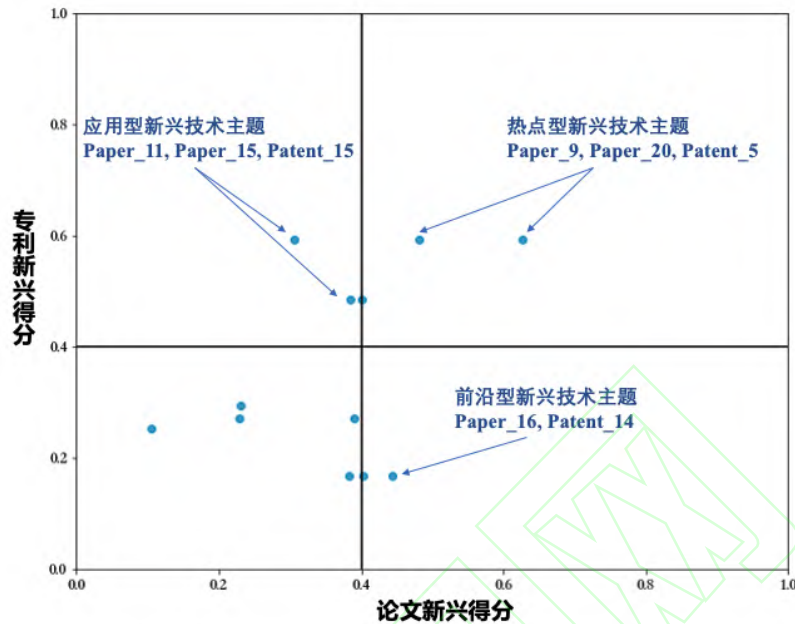


图 7 新兴技术主题识别分类结果

Fig.7 Emerging technology topic identification classification results

1) 热点型新兴技术主题。热点型新兴技术主题的划分要求为主题在共同技术主题中，并且论文新兴得分大于 0.4，专利新兴得分大于 0.4。在共同技术主题中，满足要求的热点型新兴技术主题包括 Paper_9（预训练语言模型的整合和领域专业化）、Paper_20（持续学习和基础模型迁移）和 Patent_5（多模态内容合成与模型解释）。

以 Paper_9 为例，预训练语言模型的整合和领域专业化是自然语言处理领域的重要研究方向，旨在提高模型对特定领域的理解能力和任务执行性能。在学术界，如何高效地从多个领域或知识源中整合知识到预训练语言模型中，提升模型在特定领域的理解能力和性能是目前主流的研究热点^[24]。我国《新一代人工智能发展规划》^[25]和美国国家科学技术委员会（NSTC）发布的《国家人工智能研究和发展战略计划》^[1]中都明确强调了推进人工智能，特别是预训练语言模型的整合和领域专业化技术的重要性。

2) 前沿型新兴技术主题。前沿型新兴技术主题的划分要求为主题在共同技术主题中，并且论文新兴得分大于 0.4，专利新兴得分小于 0.4。在共同存在的技术主题中，满足高论文新兴得分和低专利新兴得分的前沿型新兴技术主题包括 Paper_16（提示学习和预训练模型的应用）、Patent_14（文本生成与交互式对话系统）。

以主题 Paper_16（提示学习和预训练模型的应用）为例，提示学习是一种新兴的机器学习范式，与传统的预训练语言模型如 BERT^[27-28]、GPT^[29]等使用方法不同。提示学习的核心思想是利用“提示”来引导或激活预训练模型以解决特定任务，这种方法在处理少量数据的情况下尤为有效。相关研究在学术界引起热烈讨论，但在专利中尚未得到大量应用，由此体现了学术界对技术的研究和探索相较于产业界更快。

3) 应用型新兴技术主题。应用型新兴技术主题的划分要求为主题在共同技术主题中，并且论文新兴得分小于 0.4，专利新兴得分大于 0.4。在共同存在的技术主题中，满足低论文新兴得分和高专利新兴得分的应用型新兴技术主题包括 Paper_11（自监督学习、弱监督学习与半监督学习）、Paper_15（推理与问答系统）和 Patent_15（基于知识图谱的特征表征与预训练）。

以主题 Patent_15（基于知识图谱的特征表征与预训练）为例，这个领域的研究旨在如何利用知识图谱（Knowledge Graphs）来增强文本数据的理解和表征能力，以及如何预训练模型以有效利用这种结构化知识。最近的研究趋势表明，这些方法在学术出版物中的出现频率有所下降。这表明研究焦点发生了变化，该领域研究相对已经成熟，其中基本概念已经确立^[30-31]。这些研究已在现实世界场景中得到了广泛应用，被纳入专利和商业技术中。

4) 潜在型新兴技术主题。潜在型新兴技术主题的划分, 要求主题在非共同技术主题中, 并且论文新兴得分大于 0.4。在非共同存在的技术主题中, 只在论文技术主题中存在且具有高论文新兴分数的潜在型新兴技术主题包括 Paper_1 (知识蒸馏和领域知识注入)、Paper_7 (少样本学习与零样本学习)、Paper_8 (文本与图片生成)、Paper_17 (扩散模型)。

以主题 Paper_17 (扩散模型) 为例, 扩散模型 (Diffusion Model) 是一种用于生成模型的先进技术, 通过逐步引入并再逐步去除噪声来生成或修改数据^[32], 尤其是在图像和文本生成中表现出色, 相关研究在 2023 年 CVPR 获得了最佳论文奖。尽管扩散模型展现出了巨大的潜力和灵活性, 相关研究仍处于相对早期阶段, 并面临着提高效率、降低计算成本以及理解模型动态等挑战。特别是在改善模型性能和扩大应用范围方面, 仍需进一步发展, 以便将其以专利发明的形式应用于产业界并成为新的热点型新兴技术主题。

4 结束语

本文旨在通过综合分析大模型领域内的高价值论文和专利数据, 围绕新兴技术识别目标, 采用了一种结合 Termolator 算法和 GPT 提示学习的方法抽取技术术语。进一步, 利用 MiniBatch-Kmeans ++ 算法对术语进行聚类, 形成技术主题并将技术主题划分为共同技术主题和非共同技术主题, 通过多维指标量化分析方法对新兴技术主题进行识别和分类。研究成果表明, 该方法能够有效地揭示大模型研究领域中的热点、前沿、应用和潜在型新兴技术主题, 为新兴技术术语的识别提供了新的途径。具体进展包括:

1) 提出并验证了一种基于 Termolator 算法和 GPT 提示学习的新术语提取方法。该方法通过对比实验, 探索了 GPT 提示学习在术语抽取中的应用效果, 并且显著提高了术语抽取的准确性和召回率。

2) 通过 OpenAI Embedding 模型实现技术术语的向量化表征, 深化了对术语上下文语义信息的理解, 并解决了技术术语主题特征的量化表征问题。

3) 综合利用大模型领域内的高价值论文和专利数据, 并结合多维指标量化分析方法, 利用了技术从学术研究到专利申请转化过程中的时间滞后现象, 综合新兴技术主题的论文新兴分数和专利新兴分数, 实现了对新兴技术主题的有效识别和分类。

尽管本研究在技术术语的提取、向量化表征和主题识别方面取得了一些进展, 但仍存在待改进之处。首先, 使用 OpenAI Embedding 模型进行术语向量表征时, 模型的可解释性不足, 未能充分利用术语间的共现关系、文献间的引用关系和语义相似性等数据特性进行图表示学习。其次, 在确定主题特征指标的阈值设置和主题分类结果上, 本研究借鉴了现有研究的经验。未来研究将探索如何基于数据更加客观地确定阈值, 优化新兴技术的识别与分类方法。□

参考文献

- [1] HALAWEH M. Emerging technology: what is it[J]. Journal of Technology Management & Innovation, 2013, 8: 108-115.
- [2] MINAEE S, MIKOLOV T, NIKZAD N, et al. Large language models: a survey[J/OL]. arXiv, 2024.[2024-02-20]. <http://arxiv.org/abs/2402.06196>.
- [3] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large language models[J/OL]. arXiv, 2023.[2024-02-20]. <https://arxiv.org/abs/2307.06435>
- [4] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems[J/OL]. arXiv, 2020.[2024-02-20]. <http://arxiv.org/abs/1905.00537>.
- [5] XU Shuo, HAO Liyuan, YANG Guancan, et al. A topic models based framework for detecting and forecasting emerging technologies[J]. Technological Forecasting and Social Change, 2021, 162: 120366.
- [6] GAO Xudong. Approaching the technological innovation frontier: evidence from Chinese SOEs[J]. Industry and Innovation, 2019, 26(1): 100-120.
- [7] BURMAOGLU S, SARTENAER O, PORTER A. Conceptual definition of technology emergence: a long journey from philosophy of science to science policy[J]. Technology in Society, 2019, 59: 101126.
- [8] ROTOLO D, HICKS D, MARTIN B R. What is an emerging technology?[J]. Research Policy, 2015, 44(10): 1827-1843.
- [9] 孙蒙鸽, 王燕鹏, 韩涛, 等. 新兴技术的多指标量化识别研究——基于向量表征方法的探索[J]. 图书情报工作, 2022, 66(3): 130-139. (SUN Mengge, WANG Yanpeng, HAN Tao, et al. Research on multi-index quantitative recognition of emerging technologies: an exploration based on vector representation methods [J]. Library and Information Service, 2022, 66(3): 130-139.)

- [10] 曹琨, 吴新年, 靳军宝, 等. 基于共词和 Node2Vec 表示学习的新兴技术识别方法[J]. 数据分析与知识发现, 2023, 7(9): 89-99. (CAO Kun, WU Xinnian, JIN Junbao, et al. New technology recognition method based on co-word and Node2Vec representation learning [J]. Data Analysis and Knowledge Discovery, 2023, 7(9): 89-99.)
- [11] WANG Qi. A bibliometric model for identifying emerging research topics[J]. Journal of the Association for Information Science and Technology, 2018, 69(2): 290-304.
- [12] XU Haiyun, WINNINK J, YUE Z, et al. Multidimensional scientometric indicators for the detection of emerging research topics[J]. Technological Forecasting and Social Change, 2021, 163: 120490.
- [13] CARLEY S F, NEWMAN N C, PORTER A L, et al. An indicator of technical emergence[J]. Scientometrics, 2018, 115(1): 35-49.
- [14] LIU Xiaoyu, PORTER A L. A 3-dimensional analysis for evaluating technology emergence indicators[J]. Scientometrics, 2020, 124(1): 27-55.
- [15] JIANG Man, YANG Siluo, GAO Qiang. Multidimensional indicators to identify emerging technologies: perspective of technological knowledge flow[J]. Journal of Informetrics, 2024, 18(1): 101483.
- [16] PORTER A L, GARNER J, CARLEY S F, et al. Emergence scoring to identify frontier R&D topics and key players[J]. Technological Forecasting and Social Change, 2019, 146: 628-643.
- [17] MEYERS A L, HE Yifan, GLASS Z, et al. The termolator: terminology recognition based on chunking, statistical and search-based scores[J]. Frontiers in Research Metrics and Analytics, 2018, 3: 19.
- [18] Emerging technology identification and selection based on data-driven: taking the unmanned systems as an example | 2020 IEEE international conference on Systems, Man, and Cybernetics (SMC)[EB/OL]. [2024-04-24]. <https://dl.acm.org/doi/abs/10.1109/SMC42975.2020.9283323>.
- [19] KRISHNAN A, KASIM M M, HAMID R, et al. A modified CRITIC method to estimate the objective weights of decision criteria[J/OL]. Symmetry, 2021, 13.[2024-03-06]. <https://consensus.app/papers/modified-critic-method-estimate-objective-weights-krishnan/f1ea030d3e0b5d7cbd314ccc568e8459/>.
- [20] SERBIA C C, ŽIŽOVIĆ M, MILJKOVIĆ B, et al. Objective methods for determining criteria weight coefficients: a modification of the CRITIC method[J]. Decision Making: Applications in Management and Engineering, 2020, 3(2): 149-161.
- [21] CHANG Y, YANG P., TSAI-LIN T F. The impacts of academic patenting on paper publication: a quantity-quality examination[J]. PICMET 2010 Technology Management for Global Economic Growth, 2010: 1-10.
- [22] Patent publication and the market for ideas | management science[EB/OL]. [2024-03-04]. <https://pubsonline.informs.org/doi/10.1287/mnsc.2016.2622>.
- [23] 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. (GAO Nan, GAO Jiaqi, CHEN Hongpu. Research on emerging technology identification and evolutionary path analysis: a case study of integrated circuits [J]. Information Science, 2023, 41(3): 127-135,172.)
- [24] LU Qiuhaohao, DOU Dejing, NGUYEN T H. Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models[J]. 2021: 3855-3865.
- [25] 国务院关于印发新一代人工智能发展规划的通知_科技_中国政府网[EB/OL]. [2024-04-24]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. (State Council on the notice issued by a new generation of artificial intelligence development planning of science and technology of the Chinese government network [EB/OL]. [2024-04-24]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.)
- [26] National artificial intelligence research and development strategic plan 2023 update - the Networking and Information Technology Research and Development (NITRD) program.[EB/OL].[2024-04-24]. <https://www.nitrd.gov/national-artificial-intelligence-research-and-development-strategic-plan-2023-update/>
- [27] LAN Zhenzhong, CHEN Mingda, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J/OL]. arXiv, 2020.[2024-03-06]. <http://arxiv.org/abs/1909.11942>.
- [28] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J/OL]. arXiv, 2019.[2024-03-06]. <http://arxiv.org/abs/1810.04805>.
- [29] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[OL]. [2024-04-07]. <https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035>.
- [30] TRIGUERO I, GARCÍA S, HERRERA F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study[J]. Knowledge and Information Systems, 2015, 42: 245-284.

- [31] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J/OL]. arXiv, 2021.[2024-03-06]. <http://arxiv.org/abs/2010.11929>.
- [32] LI Cheng, QI Yali, ZENG Qingtao, et al. Comparison of image generation methods based on diffusion models[C]//2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL). Zhuhai, China, 2023: 1-4.

作者介绍: 张凯, 男, 2000 年生, 硕士生。研究方向: AI4Science 视角下的新兴术语识别, 推荐算法研究等。吕璐成, 男, 博士, 副研究员。研究方向: 知识产权情报研究, 技术挖掘。韩涛, 男, 博士, 研究员。研究方向: AI4Science 视角下的知识服务, AI4Science 前沿态势研究。赵亚娟, 女, 博士, 研究员。研究方向: 知识产权情报研究, 技术挖掘。

作者贡献声明: 张凯, 论文撰写, 数据采集与加工, 算法设计与实现, 算法结果解读。吕璐成, 论文框架设计, 论文修改。韩涛, 分析思路整理, 研究框架优化。赵亚娟, 分析思路整理, 提出修改意见。

录用日期: 2024-04-12