基于AI Agent和文本多特征融合的前沿技术识别方法研究*

李贞贞[1,2]　胡思思[1,2]　钟永恒[1,2]　王辉[1,2]　刘佳[1,2]
（1. 中国科学院武汉文献情报中心，武汉 430071；2. 科技大数据湖北省重点实验室，武汉 430071）

摘要：科技创新是产业高质量发展的强大驱动力，本研究提出一种基于人工智能体（AI Agent）和文本多特征融合的前沿技术识别方法，为把握技术突破方向、探寻技术创新机会提供借鉴。以基金项目、论文和专利数据为研究对象，首先，利用AI Agent进行数据预处理，提取文本关键词；其次，融合科技文献语义特征、分类特征和Biterm主题模型主题特征表征文本向量，增强数据特征，并利用聚类算法划分技术主题；最后，构建基于新颖性、增长性、影响力和创新性特征的前沿技术识别指标体系来识别前沿技术。对微发光二极管领域进行实证研究，融合语义、分类、主题3类特征的聚类方法表现出较好的效果，优于基于单一特征和两项特征的方法，并识别出芯片制造技术、全彩化技术等新兴前沿技术，发光材料与器件、显示屏技术、全息显示、医疗应用、检测与修复、驱动电路等热点前沿技术，以及巨量转移技术等潜在前沿技术，证明了方法的有效性和应用价值。

关键词：前沿技术识别；AI Agent；多特征融合；Biterm主题模型
中图分类号：G255　DOI：10.3772/j.issn.1673-2286.2026.01.004
引文格式：李贞贞，胡思思，钟永恒，等. 基于AI Agent和文本多特征融合的前沿技术识别方法研究[J]. 数字图书馆论坛，2026，22（1）：31-43.

当前，全球新一轮科技革命和产业变革深入发展，技术创新进入前所未有的密集活跃期，前沿技术集中涌现，引发链式变革。世界各国将科技视为决定自身命运的全新竞技场，积极寻求突破性技术，以抢占未来科技发展的制高点。前沿技术通常是指在科技创新过程中最具新颖性、潜在性、前瞻性和引领性的研究方向，具有高创新性、高潜力和高不确定性的特征，它们可能引发行业的变革，甚至重塑整个经济格局[1]。前沿技术识别是指通过系统化方法，识别和追踪某一学科或领域中影响力显著的研究方向、热点问题和新兴趋势，发现当前研究的热点、未来可能的发展方向以及潜在的突破点。及时准确地研判科技发展研究前沿可以有效预测未来研究趋势，掌握科技前沿方向，成为技术预测、创新管理和科技战略决策的关键环节，对于洞察研究趋势、瞄准前沿领域、发展知识服务、促进科技创新具有重要意义。

科技文献作为技术信息的重要载体，为前沿技术识别提供了丰富的数据来源。学术论文反映了科研人员的最新研究成果和学术交流动态，是科技创新的理论基础和知识源泉；专利文献蕴涵了丰富的技术知识前沿，是科学技术创新能力的重要衡量标准；基金项目数据从国家层面反映未来科技创新方向，对科学研

收稿日期：2025-11-14
*本研究得到湖北省技术创新专项软科学研究类重点项目“‘十五五’湖北基础研究重点方向任务及基础条件自主保障能力研究”（编号：2025EDA011）、青海省重点研发与转化计划“盐湖资源绿色高值利用科技成果数字化服务系统”（编号：2023-QY-213）资助。

究未来发展具有前瞻引领作用[2-3]。3类数据在科学研究发展的不同阶段和不同维度都具有独特的价值。现有基于多源数据的前沿技术识别大多为数据集层面的融合，并未深入数据特征层面。本研究使用学术论文、专利和基金项目3类数据源，提出一种基于人工智能体（AI Agent）和多维度数据特征融合的前沿技术识别方法，采用基于大语言模型（Large Language Model，LLM）的AI Agent自动识别科技文献关键词，并融合科技文献文本语义信息、分类信息和主题信息表征科技文献数据特征，通过文本聚类获得技术主题，从多维视角构建科学研究前沿技术识别指标体系，识别出产业的科学研究前沿技术，为我国前瞻布局战略性、引领性基础研究和未来产业提供借鉴。

1 相关研究

1.1 前沿技术识别研究

目前，国内外学者对前沿技术的定义尚未统一。1965年，de Solla Price[4]提出研究前沿为近期发文规模在30~50篇且高被引的文献集合。Chen[5]、Small等[6]认为前沿技术为某一领域正在兴起的且快速增长的技术。Wang等[7]认为研究前沿指引着科技发展的方向和技术创新的路径。刘春文等[8]认为前沿技术指技术系统中较为新颖或发展速度较快，且近期对技术系统影响较大，具有感应、吸收及再创新能力的技术，应同时具有新兴性、确定性以及探索性的特征。武川等[9]认为前沿技术是一定时间内关注度快速提升的某项或某类符合产业发展需求的技术。现有研究中存在诸多与前沿技术相近的概念，包括研究前沿、热点技术、新兴技术、颠覆性技术等，但这些概念之间又有一定的区别。研究前沿是指受到高度关注，具有较大发展潜力、较高影响力和引领性的科学问题，体现了学科发展的最新方向，强调高创新性和高影响力[10]；热点技术是指当前受到高度关注和讨论的研究内容；新兴技术是指处于兴起或发展阶段且存在潜在重大影响，但路径和应用尚不明确的技术，具有新颖性、相对较快增长、连贯性、突出影响和高度不确定性五大特征[11]；颠覆性技术是指通过技术创新，突破并替代传统或主流技术，有潜力颠覆原有产业格局的技术，强调创新的突破性[12]。由此可见，前沿技术与研究前沿内涵较为接近，是在科学问题驱动下产生的新方向，前沿技术关注技术创新，而研究前沿更关注科学研究，研究前沿孕育了前沿技术。前沿技术、新兴技术、颠覆性技术都具有技术创新性特征，新兴技术强调“新”和高潜力，颠覆性技术更强调创新的突破性，而前沿技术已具备战略价值与基础突破潜力，代表世界科技发展的新方向。结合现有学者对前沿技术概念及特征的分析，本研究将前沿技术视为正在兴起的、受关注度增长较快、极具创新能力且开始显现市场潜力的技术，具有新颖性、增长性、影响力、创新性等特征。

国内外学者对前沿技术识别任务的研究大致可归纳为定性和定量两类。定性方法主要指专家定性评价法，通常依赖领域专家的知识和经验，主观性较强且耗时耗力。定量方法通过分析客观数据来发现研究前沿，剔除了人为主观因素，且时间成本低，主要包括基于引文的识别方法、基于指标体系的识别方法，以及基于文本挖掘的识别方法。①基于引文的识别方法通过挖掘引用关系形成网络结构，揭示引用关系中蕴含的知识流动和知识组织结构，包括直接引用分析、共被引分析、耦合分析等方法[13]，通过对核心文献、作者和机构的分析，识别特定领域的研究前沿。高楠等[14]选取情报学领域的高关注度文献集进行共被引分析、耦合分析，识别与对比国内和国外的前沿主题；姚茹等[15]基于共被引聚类识别研究前沿，将核心文献数据及施引文献数据作为演化分析中不同文献来源的基础数据。②基于指标体系的识别方法通过构建前沿特征指标体系对数据进行指标测算，筛选前沿技术主题。苗红等[16]充分关注前沿技术的应用场景，建立前瞻性、先导性、探索性、颠覆性4个维度的前沿技术识别指标体系，基于前沿特性指标筛选专利与论文数据集，并使用GloVe模型和K-means聚类算法分析前沿技术主题。③基于文本挖掘的识别方法包含了早期的共词分析方法、词频分析方法、潜在语义分析（Latent Sementic Analysis，LSA）、概率潜在语义分析（Probabilistic Latent Semantic Analysis，PLSA），以及近年来广泛应用的潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）模型、聚类算法等。LDA是常用的前沿主题识别方法之一：刘自强等[17]利用LDA模型进行基金、论文研究主题探测，综合新兴度、关注度指标和战略坐标图进行初始研究前沿判别；林晗等[18]对LDA的拓展模型HLDA（Hierarchical Latent Dirichlet Allocation）展开研究，在方法核心过程nCRP（Nested Chinese Restaurant Processes）中引入了文本相似度因子，提高

了前沿主题识别任务的准确性。聚类算法因能够从大量复杂数据中简单高效地挖掘内在结构和分布规律，在前沿主题识别中发挥了重要作用。武川等[9]、刘盼盼等[19]使用K-means聚类算法对前沿技术专利主题进行聚类，进而识别前沿技术领域。此外，研究表明通过语义分析技术实现多维度特征的关系融合，能够有效揭示文本信息的特征，提升前沿探测结果的完备性。刘小玲等[20]构建专利文本向量、引文向量和分类向量，综合三者得到基于多属性融合的专利向量，在此基础上通过专利聚类获得技术主题；曾文等[21]将科技信息语义关系、分类关系和引用关系融合成一个新的关系解释实体特征，并采用K-means算法对文献进行聚类，结合决策主题余弦相似度和个体经验对多维度科技信息进行主题融合。

1.2　LLM与AI Agent

LLM的快速发展极大推动了自然语言处理任务的进步，其强大的少样本学习能力和生成式任务处理逻辑，解决了深度学习模型训练难、迁移难的痛点。LLM利用深度学习技术，在大规模数据集上进行训练，从而理解并生成新的文本，以高度熟练的方式使用人类语言进行交流、理解和互动，显著提升自然语言处理任务的准确性和泛化能力，在关键词提取、文本生成、自动摘要和分类等自然语言处理任务中展现出卓越的性能[22-24]。AI Agent是一种模仿人类智能行动的智能化系统，其通过独立思考、调用工具来逐步完成给定任务，能够解决幻觉、缺乏特殊知识、知识过时等LLM问题[25]。AI Agent以LLM为核心驱动，具有规划、记忆、工具使用和行动四大核心模块，通过自主理解环境，推理问题、制定解决问题计划，作出适应性决策，并在一组工具的帮助下采取行动，以更智能、更高效的方式自主完成复杂任务[26-27]。目前，AI Agent已在各行业深度融合与加速落地，如数据挖掘、智能推荐、图像生成、程序编写、输出报告、智能驾驶等[28-29]。AI Agent代表了AI从“被动回答问题”到“主动解决问题”的范式转变，利用AI Agent来提高工作任务的执行效率，成为具有重要应用前景的研究方向。

综上所述，现有基于文本挖掘的前沿技术识别方法通常仅利用文本语义信息进行建模，忽略了数据源丰富的分析维度，而对于多特征融合的研究较少。本研究采用文本挖掘方法，以学术论文、专利和基金项目3类数据为研究对象，提出了一种基于AI Agent和多特征融合的前沿技术识别框架，构建AI Agent提取文本关键词，并构建语义向量、分类向量和主题向量进行文本综合向量表征，采用聚类算法划分主题类簇，设计前沿技术识别指标体系，对微发光二极管（Micro-LED）技术领域的前沿技术进行深入探讨，为领域前沿技术识别方法创新研究提供新思路。

2　研究方法

2.1　研究框架

本研究构建基于AI Agent和多特征融合的前沿技术识别方法框架（见图1），主要包括4个环节。①多源数据收集与预处理。制定检索式，获取基金项目、论文和专利数据，构建AI Agent提取文本关键词。②多特征融合的数据向量化。分别基于论文、专利和基金项目的文本语义信息、分类信息和主题信息进行文档向量表征。③技术主题划分。利用聚类算法划分技术主题，抽取聚类簇特征词，采用LLM和人工结合方式对技术主题信息进行凝练。④前沿技术特征识别。基于前沿技术的新颖性、增长性、影响力和创新性特征，构建前沿技术识别指标，形成多源数据关联识别结果。

2.2　基于AI Agent的数据预处理

关键词能够精确、有效地体现文本信息，在主题挖掘中具有较高的分析价值。传统方法采用分词技术对文本数据进行预处理，然而领域专业词语通常具有复合结构的特点，即由两个或两个以上的词语组合而成，可能导致分词处理后词语长度超过2个的未登录词无法有效识别[30]。LLM在文本理解和跨任务生成方面表现出强大的能力，为文本挖掘提供了有效的技术支撑[31]。而AI Agent是由LLM驱动、可独立运行的实体，能够对外部环境主动交付并自主执行复杂任务[32]。为此，本研究构建了一个数据预处理AI Agent，用于提取文本关键词。利用提示工程（Prompt Engineering）精心设计自然语言提示词，通过定义明确的角色、任务、抽取规则和输出格式，引导LLM输出具有代表性的特征词汇，为后续技术识别提供高质量数据。基于AI Agent的关键词提取流程如图2所示。

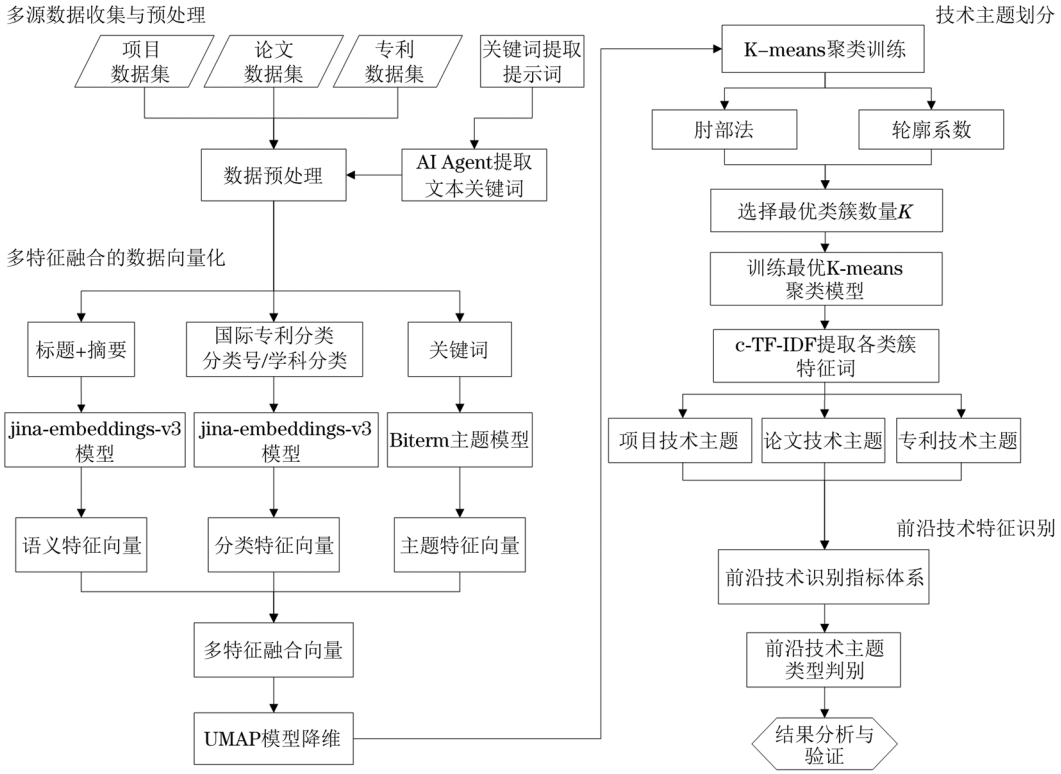


图1　研究框架

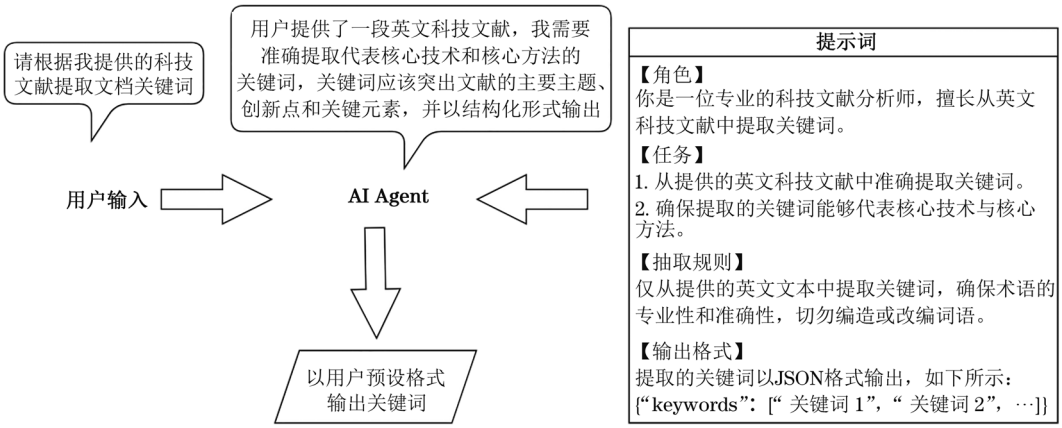


图2　基于AI Agent的关键词提取流程

2.3　多特征融合的数据向量表征

本研究融合科技文献语义特征、分类特征和主题特征揭示文本信息的综合特征。其中，用文本内容表示语义特征，用分类信息表示分类特征，用主题概率分布表示主题特征。

（1）语义特征向量。拼接文本标题和摘要得到文本语料，采用jina-embeddings-v3模型训练文本语料得到文本向量，用来表示文本在语义上的相似程度。jina-embeddings-v3是Jina AI公司于2024年9月18日发布的一个前沿的多语言文本嵌入模型，具有特定任务的低秩适应适配器，能够为查询、检索、聚类、分类和文本匹配等任务生成高质量嵌入，在多语言数据和长文本检索任务中具有先进的性能[33]。

（2）分类特征向量。为有效获取分类信息在文本中的特征，不同于构建共同分类强度表示的方法，本研究将分类序列视为一个句子进行建模，获取文本分类的语义表示，最大限度保留其位置信息和语义信息[34]。以专利文献为例，本研究将一个分类号视为一个词语，多个分类号组成的序列视为一个句子，通过学习分类

号在序列中的上下文语境，得到每个专利序列的向量表示。本研究采用国际专利分类（International Patent Classification，IPC）分类号大组（前6位）表示专利分类信息，拥有相同分类号数量越多的专利，在技术功能和应用领域上越相似。同样地，将科技论文的学科分类序列视为一个句子进行文本向量建模，使用jina-embeddings-v3模型训练论文和专利的分类信息，得到分类向量。

（3）主题特征向量。主题模型是一种无监督的学习方法，其通过文本挖掘技术发现科技文献中具有高度概括性的主题信息，能够辅助研究人员快速把握核心技术[35]。Biterm主题模型（Biterm Topic Model，BTM）[36]通过提取文档中词共现的方式，提高主题模型对文本的分类效果，常应用于评论挖掘[37]、话题检测[38]等任务。BTM采用词对（Biterm）作为主题建模的单元，有效增加了传统主题建模的词语数量。词对中两个词语共现次数越多，越有可能由同一个主题产生，这种显式的词语共现模式增强了主题模型学习能效，解决了短文本特征稀疏的问题[39]。考虑到提取关键词后的文本更接近短文本，本研究采用BTM对文档进行主题建模。将AI Agent提取的文本关键词作为BTM的输入，结合Rényi熵和主题一致性来确定最佳主题数量[40]，采用最佳主题数量训练BTM，获得文本主题概率向量。

以上3类特征向量分别处于不同的特征空间，为避免加权求和、注意力融合等处理改变这些特征的原始结构，导致部分特征丢失、不同特征之间边界模糊等问题，本研究保留所有原始特征信息，将语义特征向量、分类特征向量和主题特征向量这3类向量通过首尾相连的方式进行拼接，得到文本的多特征融合向量。

2.4　技术主题划分

将得到的科技文献向量表征作为模型输入，通过文本聚类和主题定义得到技术主题类簇划分结果。

（1）文本聚类。本研究首先采用UMAP（Uniform Manifold Approximation and Projection）模型对文本的多特征融合向量进行降维处理，将高维向量映射到低维向量中，保留更多的全局特征，同时提升后续聚类算法的效果。其次使用K-means算法分别对论文、专利和基金项目进行聚类，将降维后的多特征融合向量作为模型的输入，结合肘部法和轮廓系数来确定最佳主题数。肘部法计算不同K值下的聚类误差平方和（Sum of Squared Error）；轮廓系数结合了簇内的紧密度和簇间的分离度两个因素，值越高表示数据点在簇内越紧密，同时与最近簇的距离越远。

（2）主题表示与描述。通过K-means算法可聚类得到文本向量类簇，使用c-TF-IDF（Class-Based Term Frequency-Inverse Document Frequency）算法提取每个类簇的特征词，对类簇进行主题表示。c-TF-IDF算法是一种用于文本分类和主题建模的技术，它通过将传统TF-IDF方法应用于类别或主题级别而非文档级别，发现每个类别中的代表性词语。对于每个类簇的代表性特征词，利用LLM基于输入的一组特征词凝练出对应的技术主题描述，并结合专家判读，提高技术主题的描述质量。

2.5　前沿技术识别指标体系构建

本研究基于前沿技术的内涵和特点，综合考虑技术的新颖性、增长性、影响力和创新性等典型特征，构建前沿技术识别指标体系。

（1）新颖性。新颖的技术主题通常预示着新的研究方向，可通过计算论文平均出版时间、专利公开时间及基金项目立项时间测度研究主题的新颖性。论文出版平均时间、专利公开时间及基金项目立项时间可以体现研究主题是否为新近出现，越是新近出现的研究主题越容易包含新的研究内容。参考部分学者做法[1,19,21]，采用某一主题下全部支持文档的平均发表时间表示主题新颖性。假定技术主题k的支持文档总数为m篇，该技术主题新颖性计算公式如式（1）所示。

$$N_k = \sum_{i=1}^{m} \frac{y_i}{m} \tag{1}$$

式中：y_i为第i篇文档的发表年份。通过Min-Max标准化将计算结果归一化为（0，1]，值越接近1表示技术主题越新颖。

（2）活跃性。技术增长性和影响力特征可由主题的活跃性来综合表征[19]。增长性是描述技术主题发展趋势的特征，反映了主题的活跃程度，可通过计算主题支持文档数量的年增长率来衡量主题的增长性特征。影响力是描述主题吸引力的特征，反映了研究主题的学术关注程度，通过论文被引频次、专利施引频次测度技术影响力。主题的增长性越高，影响力越广泛，表明该研究主题的增长越快，其研究价值受到的关注度也越高，产生作用的程度越大，越有可能成为前沿主

题。假定技术主题k的支持文档总数为m篇，其在n年中每一年的支持文档数分别为d_1、d_2、…、d_n，参考张家惠等[41]定义的计算公式，技术主题活跃性计算公式如式（2）所示。

$$P_k=\frac{1}{n}\sum_{j=1}^{n}\frac{d_j-d_{j-1}}{d_{j-1}}\times\ln(1+\sum_{i=1}^{m}\frac{c_i}{m}) \quad (2)$$

式中：d_j、d_{j-1}分别为主题在第j和$j-1$年的支持文档数量，通过比值计算主题的年平均增长率，用于衡量主题的增长性特征；c_i为主题k中第i篇文档的被引频次，通过计算主题k支持文档的篇均被引频次，衡量主题的影响力特征。将篇均被引频次对数化，以使影响力特征和增长性特征更加平衡，因为篇均被引频次远大于年平均增长率。

（3）创新性。通过重组创新性指标测算技术的创新程度，以此衡量创新性。以专利数据为例，将专利IPC大组两两组合，并与申请年以前的IPC大组组合进行对比，指标值越高，则专利的创新程度越高。同理，对论文和基金项目数据的关键词进行成对组合，如果某组合在之前的文献中未出现，说明产生了知识的新组合，具有创新性。借鉴已有研究做法[42-44]，创新性指标的计算方法如式（3）所示。

$$I_k=\frac{1}{m}\sum_{i=1}^{m}\frac{p_i}{g_i},p_i\leqslant g_i \quad (3)$$

式中：p_i为在第i个专利申请年之前从未出现的IPC大组组合数目；g_i为第i个专利的IPC大组组合数目。

根据构建的前沿技术识别指标，以各个指标均值为分界点，指标计算结果大于平均值的确定为“高”，小于平均值的为“低”，将技术主题划分为热点前沿技术主题、新兴前沿技术主题、潜在前沿技术主题、非前沿技术主题4种类型（见图3）。①潜在前沿技术主题具有高新颖性特征，但活跃性和创新性均较弱，表明该类主题尚处于早期萌芽阶段，随着时间推移逐步改进，未来有较高潜力成为前沿技术方向；②新兴前沿技术主题的新颖性和创新性均高于平均值，但其活跃性低，表明该类主题是近期出现且具有较强创新能力的技术，正处于新兴阶段，其增长性和影响力尚未显现；③热点前沿技术主题是前沿技术成熟后演变而成的热门研究内容，该类主题活跃性高于平均值，表明其增长性和影响力均较强，虽已出现一段时间，但仍受到广泛的关注；④非前沿技术主题为已出现一段时间，但活跃性或创新性较低的主题，该类主题的受关注度和影响力较弱，创新能力偏低。

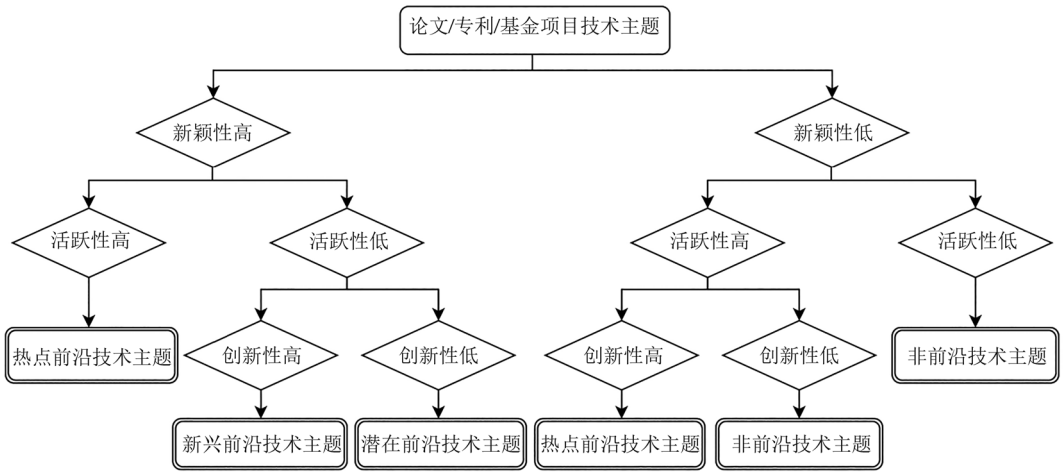


图3　前沿技术主题类型划分

3　实证分析

3.1　实验数据收集

Micro-LED是一种将微米级半导体发光二极管以矩阵形式高密度集成在一个芯片上的显示技术，具有高度小型化、薄膜化、集成化的特点，可广泛应用于消费电子、虚拟现实/增强现实设备、车载显示、医疗显示等多个领域，被称为次世代显示技术的发展趋势[45-46]。本研究选取Micro-LED领域为科技前沿识别的研究对象，以该领域专利文献、科技论文和基金项目3类数据为信息源。其中，专利数据来源于Derwent Innovations Index数据库，论文数据来源于Web of Science Core Collection数据库，2类数据设

定检索式均为：TS=（“Micro led” OR “Micro-led”） or TS=（Display AND（chip soldering or massive transfer or bonding drive or full-color or（testing and repair））），基金项目数据来源于万方数据科慧平台，设定检索式为：TS=（“Micro led” OR “Micro-led”）。检索日期为2025年1月6日，删除无关数据后，共获得专利数据25 748条、论文数据7 618条、基金项目数据1 168条。

3.2　数据预处理

本研究基于阿里云研发的百炼平台，以DeepSeek-V3为底座，构建科技文献数据预处理AI Agent来提取文本关键词。通过调用AI Agent接口实现对文本的数据预处理，传统数据预处理方法和基于LLM的数据预处理方法示例对比结果如表1所示。由表1可知，LLM在识别复合术语词上具有显著的成效，在文本数据预处理方面展现出较好的效果。

3.3　技术主题识别结果

3.3.1　特征向量

下载jina-embeddings-v3模型到本地服务器，采用SentenceTransformer调用模型获取每篇文档的语义向量和分类向量，任务选取“Classification”。为验证将分类信息视为文本序列进行语义向量表征的有效性，分别对专利分类号和论文学科分类，采用语义向量和共同分类强度表示方法开展研究对比，结果如表2所示。根据分类向量对比结果，专利分类号“H01L-027；H01L-033”和“H01L-033”分别与“H01L-025；H01L-033”计算分类相似度，采用共同分类强度计算结果均为0.50；而对比发现，相较于“H01L-033”，“H01L-027；H01L-033”两个分类均以H01L开头，与“H01L-025；H01L-033”两个分类应更为相近。科技论文学科分类“Engineering，Electrical & Electronic；Optics；Physics，Applied”与“Optics；Physics，Multidisciplinary；Physics，Condensed Matter”虽只有“Optics”一个分类相同，但“Physics，Applied”和“Physics，Multidisciplinary；Physics，Condensed Matter”的一级学科分类相同，有着极高的相似度，分类语义相似度计算结果0.91比共同分类强度0.20更为合理。

3.3.2　主题识别模型对比

为验证本研究所提方法的有效性，将基于多特征融合的向量与基于单一特征、两项特征的向量技术主题识别模型进行比较，利用轮廓系数验证聚类效果，体现簇内紧密度和簇间分离度。BTM使用Python中的bitermplus包来构建，模型参数统一设定为α=50/K、β=0.01，迭代次数为1 000次。将语义特征、分类特征、主题特征分别简写为D1、D2、D3。设定主题数量的变化范围为[5，100]，主题数变化步长为5，使用UMAP模型将向量降至500维，采用K-means算法进行聚类，使用不同特征向量作为模型输入的对比实验结果如图4所示。基金项目数据因没有分类字段，故无D2相关实验结果。

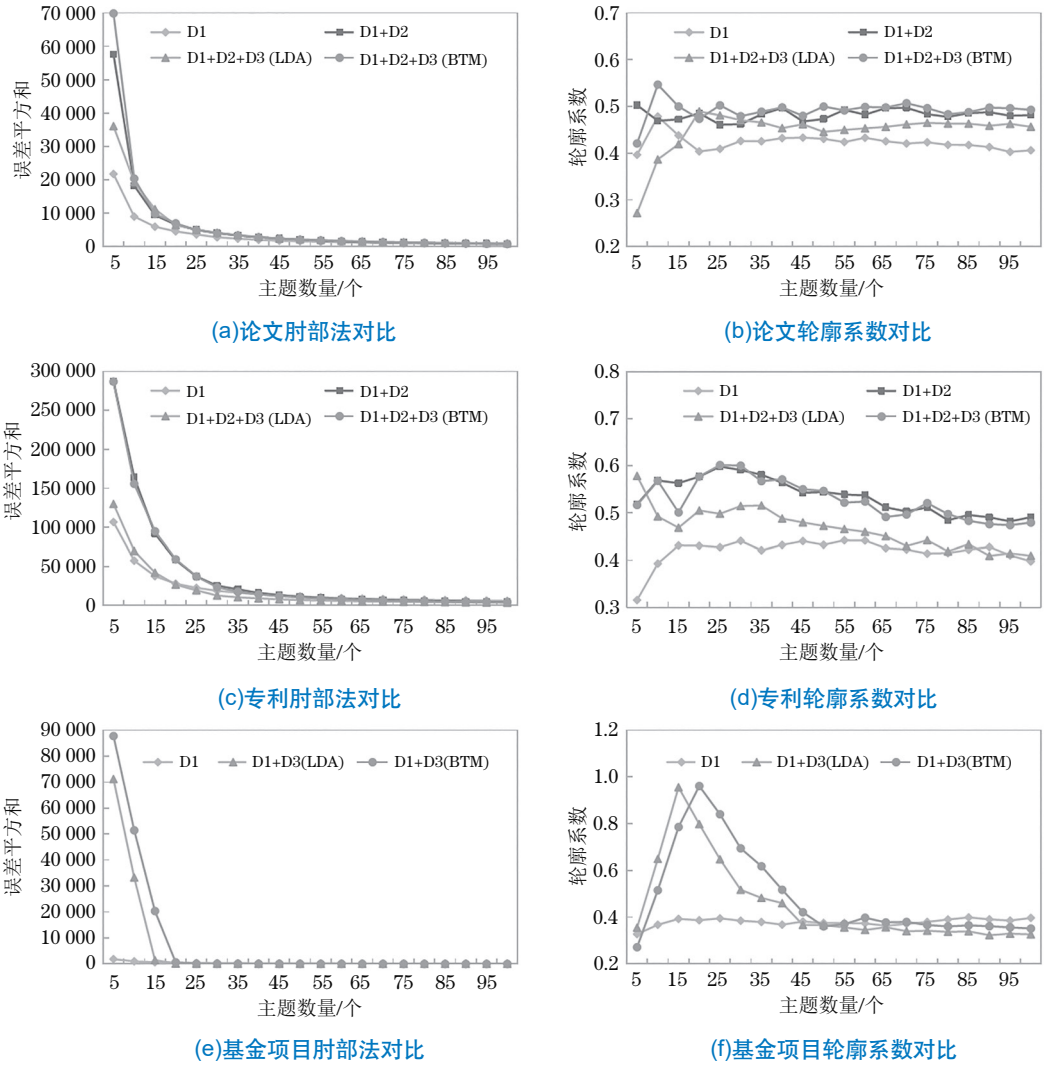
由图4可知，对于论文和专利数据，D1+D2的轮廓系数显著高于D1，表明分类向量的加入有效提升了单一语义向量的聚类效果，对文本数据的区分有较强的影响。当增加D3（LDA）时，轮廓系数相较于D1+D2有较大幅度的降低，说明LDA主题模型的主题概率向量未能有效区分文本间的关系；而当增加D3（BTM）时，轮廓系数有一定的提升，尤其是论文和基金项目数据的轮廓系数，专利数据中增加D3（BTM）后的轮廓系数最大值也优于D1+D2，表明使用BTM来表征数据的主题特征，可提升文本聚类的效果。本研究提出的融合语义、分类、主题3类特征的聚类方法表现出了较好的效果，优于单一特征和两项特征的聚类结果，表明通过增

表1　提取关键词示例

专利分类号	标题	分词+去停用词结果	AI Agent提取关键词
CN114975699-A；CN114975699-B；WO2024021158-A1	Single-chip integrated preparation method of full-color nitride semiconductor micro-LED array，involves using bottom surface electrode bonding technology to bond full-color micro-LED array flat sheet and driving circuit board	single-chip、integrated、preparation、method、full-color、nitride、semiconductor、micro-led array、involve、use、bottom、surface、electrode、bonding technology、bond、full-color、micro-led array、flat、sheet、drive、circuit、board	single-chip integrated preparation method（单芯片集成制备方法）、full-color nitride semiconductor micro-LED array（全彩氮化物半导体微LED阵列）、bottom surface electrode bonding technology（底面电极键合技术）

表2　分类向量对比结果

数据类型	类别1	类别2	分类语义相似度	共同分类强度
专利文献	H01L-027; H01L-033	H01L-027; H01L-033	1.00	1.00
		H01L-033	0.86	0.50
		H01L-025; H01L-033	0.93	0.33
	H01L-033	H01L-027; H01L-033	0.86	0.50
		H01L-033	1.00	1.00
		H01L-025; H01L-033	0.87	0.50
科技论文	Optics	Engineering, Electrical & Electronic; Optics; Physics, Applied	0.85	0.33
		Optics; Physics, Multidisciplinary; Physics, Condensed Matter	0.86	0.33
		Physics, Applied	0.74	0.00
	Engineering, Electrical & Electronic; Optics; Physics, Applied	Optics; Physics, Multidisciplinary; Physics, Condensed Matter	0.91	0.20
		Physics, Applied	0.87	0.33
		Optics	0.85	0.33



(a)论文肘部法对比　(b)论文轮廓系数对比

(c)专利肘部法对比　(d)专利轮廓系数对比

(e)基金项目肘部法对比　(f)基金项目轮廓系数对比

图4　模型对比实验结果

强文本多维度特征的方法，可以进一步提高文本的表征能力，获得较好的主题聚类效果。

3.3.3　主题聚类结果

根据以上分析，本研究综合考虑语义特征、分类特征、主题特征来进行数据的向量表征，设置最佳主题数分别对论文、专利和基金项目数据进行聚类，得到技术主题识别结果。在构建BTM时，利用Rényi熵计算得到论文、专利和基金项目的最佳主题数分别为12、15、17。采用K-means聚类算法分别对论文、专利和基金项目进行聚类，最佳K值利用肘部法和轮廓系数进行判定，分别为9、25、19。利用c-TF-IDF算法计算得到各技术主题的核心特征词，结合LLM和专家经验解读聚类结果，将判定的相同主题方向进行合并，确定主题名称，得到各技术主题内容，结果示例如表3所示。

表3　技术主题特征词示例

数据类型	主题名称	主题特征词
专利	显示屏技术	chip on board display screen（板载芯片显示屏）、curved surface display device（曲面显示器）、color light emitted diode display device（彩色发光二极管显示屏）、electrophoresis display device（电泳显示器）、full color reflective display（全彩反射式显示器）
	巨量转移技术	giant transfer device（巨量转移装置）、transfer mold（转移模具）、transfer seal（转移密封）、transferred chip（已转移芯片）、chip transfer device（芯片转移装置）
论文	芯片制造技术	chip solder bumps（芯片焊料凸块）、flip chip assembly（倒装芯片组件）、microchips（微芯片）、solder flip chip assembly（焊接倒装芯片组件）、chip on glass（晶玻接装）
	全息显示	volume holographic optical elements（体积全息光学元件）、generated hologram（生成全息图）、color hologram（彩色全息图）、metaholograms（元全息图）、full color holographic optical element（全色全息光学元件）
基金项目	背板技术	layer composite patterned lead frame substrate（层复合图案化引线框架基板）、ic backplane substrate（集成电路背板基板）、patterned sapphire substrate（衬底）、metal substrates（金属基材）、driving circuit backplane（驱动电路背板）
	全彩化技术	full color volume visualization（全色体积可视化）、color organic thin film（彩色有机薄膜）、three primary color wavelength bands（三原色波段）、differential display coloring（差分显示着色）、color control utility（颜色控制实用程序）

3.4　前沿技术识别结果

根据前沿技术识别指标体系，分别计算基金项目、论文和专利的技术主题新颖性、活跃性和创新性，得到3类数据的前沿技术主题指标测算得分。本研究将3类数据视为同等重要，将多数据类型中相同技术主题的指标值均值作为主题最终得分，遴选出的前沿技术主题如表4所示。

3.5　结果分析与验证

根据表4所示的Micro-LED领域的新兴前沿技术主题、热点前沿技术主题和潜在前沿技术主题，结合情报专家经验，分别从各类型技术主题中选取代表性主题，通过文献溯源法对技术主题进行定性调研与分析解读，验证识别结果是否符合实际情况。

（1）新兴前沿技术主题。该技术主题涉及Micro-LED领域近期兴起并得到广泛关注的技术，包括全彩化技术、芯片制造技术等主题。其中，全彩化技术是在基金项目和论文中共同出现的主题。Micro-LED全彩化方案决定了Micro-LED实际实现色彩的丰富程度，全彩发光是Micro-LED微显示屏实际应用需要应对的挑战之一[47]。基于量子点颜色转换器的全彩化Micro-LED是较具代表性的技术方案，正受到越来越多的关注[48]。此外，芯片制造技术属于“高、尖”和新兴的重大复杂性技术，是世界各国争夺的技术战略高地，成为科技领域的新兴研究前沿[49]。

（2）热点前沿技术主题。该技术主题是过去一段时间内Micro-LED领域出现并被广泛关注的具有重要影响力的技术，包含发光材料与器件、显示屏技术、全息显示、医疗应用、检测与修复、驱动电路等主题。在基金项目、论文和专利中均出现的技术主题为发光材料与器件、显示屏技术。Micro-LED发光材料与器件由以氮化镓（GaN）为主的无机材料构成，提供了更高的亮度和对比度。Micro-LED材料快速发展，不断涌现出新的研究成果和前沿热点[50-51]。Micro-LED显示屏技术以其高亮度、高对比度和高分辨率的特点，成为下一代显示技术的代表，具有广泛的应用前景，其研发和产业化

表4　Micro-LED领域前沿技术主题划分结果（部分）

数据类型	主题名称	新颖性	活跃性	创新性	技术主题类型
基金项目+专利+论文共同主题	芯片制造技术	0.666	0.606	0.318	新兴前沿技术主题
	发光材料与器件	0.659	0.922	0.330	热点前沿技术主题
	显示屏技术	0.614	0.691	0.400	热点前沿技术主题
基金项目+论文共同主题	全彩化技术	1.000	0.530	0.546	新兴前沿技术主题
基金项目+专利共同主题	背板制造技术	0.548	0.605	0.367	非前沿技术主题
论文主题	全息显示	0.633	0.747	0.591	热点前沿技术主题
专利主题	背板材料	0.814	0.583	0.195	潜在前沿技术主题
	医疗应用	0.708	0.738	0.376	热点前沿技术主题
	封装技术	0.001	0.691	0.076	非前沿技术主题
	光学元器件	0.526	0.722	0.122	非前沿技术主题
	通信应用	0.409	0.886	0.241	非前沿技术主题
	巨量转移技术	0.849	0.559	0.043	潜在前沿技术主题
	检测与修复	0.731	0.664	0.269	热点前沿技术主题
	驱动电路	0.709	0.725	0.106	热点前沿技术主题

进程不断加快，是大家竞相研究的热点前沿技术[52]。

（3）潜在前沿技术主题。该技术主题是Micro-LED领域近期出现并受到广泛关注，但研究数量不足、具有较大研究潜力的技术，包含巨量转移技术。巨量转移是将大量微米级芯片从生长基板上分离出来，并快速且精准地转移到目标基板的过程，是Micro-LED显示屏量产制造的核心[53]，被中国机械工程学会列入2021—2022年度“前沿科学问题和工程技术难题”，已成为该领域的国际竞争制高点[54]。

从以上结果可以看出，利用AI Agent对文本进行预处理，使得聚类得到的类簇特征词为更能表达语义信息的复合词语，而通过考虑科技文献多维度的特征，能够有效提升主题分类的准确性。前沿技术主题识别结果较为准确，表明本研究提出的方法具有一定的可行性，能够有效识别领域前沿技术。

4　结语

坚持科技创新的战略地位，是我国在新兴产业和未来产业发展中赢得主动权的有力保障。准确、有效的前沿技术识别方法有助于前瞻性分析各研究领域的研究前沿和技术前沿，为我国宏观战略决策、科技创新主体把握科技前沿方向提供有力支撑，从而在全球科技竞争中抢占先机。本研究以论文、专利和基金项目3类科技数据为基础，构建AI Agent提取文本关键词，综合语义信息、分类信息和主题信息多维度特征融合表征数据特征，综合运用新颖性、活跃性和创新性等指标识别前沿技术主题类型，并在Micro-LED领域开展前沿技术识别实证研究。分析发现，Micro-LED领域包括芯片制造技术、全彩化技术等新兴前沿技术主题，发光材料与器件、显示屏技术、全息显示、医疗应用、检测与修复、驱动电路等热点前沿技术主题，以及巨量转移技术等潜在前沿技术主题。本研究表明，基于AI Agent和多维度数据特征融合的方法，能够从科技数据中有效提取技术主题。结合专家智慧和定量指标进行前沿识别，有助于提高前沿识别的准确性和前沿探测的前瞻性，为知识服务和技术创新提供支撑。

本研究仍存在一些局限：①技术主题识别结果不够细致，淹没了科技文献数量较少的技术主题，未来将考虑利用深度学习算法改进关键技术主题识别方法，进一步提升技术主题识别准确性和完备性；②仅选用3类科技数据，且仅对一个领域进行实践，数据范围与规模有待扩展，未来将通过增加数据的多样性和规模来增强前沿识别的全面性和可靠性；③仅利用了单一AI Agent执行特定任务，未来将进一步探索多AI Agent与专家协同的前沿技术识别策略。

参考文献

[1] CHUAN L M, ZHAO J J, QI S J, et al. Research frontiers in the field of agricultural resources and the environment[J]. Applied Sciences, 2024, 14(12): 4996.

[2] YAMAN A, SARTONO B, INDRAWATI A, et al. Automated multi label classification on fertilizer themed patent documents

in Indonesia[J]. DESIDOC Journal of Library & Information Technology，2022，42（4）：218-226.

[3] 曹树金. 探寻未来科学研究前沿 服务国家科技创新：评《未来新兴科学研究前沿识别研究》[J]. 现代情报，2025，45（5）：3-4.

[4] DE SOLLA PRICE D J. Networks of scientific papers[J]. Science，1965，149（3683）：510-515.

[5] CHEN C M. CiteSpace II：detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology，2006，57（3）：359-377.

[6] SMALL H，BOYACK K W，KLAVANS R. Identifying emerging topics in science and technology[J]. Research Policy，2014，43（8）：1450-1467.

[7] WANG X F，ZHANG S，LIU Y Q. ITGInsight−discovering and visualizing research fronts in the scientific literature[J]. Scientometrics，2022，127（11）：6509-6531.

[8] 刘春文，黄鲁成，苗红，等. 基于ECT-Dim的前沿技术识别方法：以养老环境辅助生活技术为例[J]. 情报杂志，2023，42（8）：77-82，76.

[9] 武川，王宏起，王珊珊. 前沿技术识别与预测方法研究：基于专利主题相似网络与技术进化法则[J]. 中国科技论坛，2023（4）：34-42.

[10] 许海云，龚兵营，杨俊浩，等. 新兴研究主题识别方法研究进展与前瞻[J]. 图书情报工作，2025，69（3）：135-150.

[11] 张甜，陈进东，周晓纪，等. 基于潜在影响力预测和多源信息融合的新兴技术识别方法[J]. 情报杂志，2025，44（9）：134-142，133.

[12] 叶选挺，余子明，梁文庭，等. 颠覆性技术创新项目资助模式：国外经验与发展策略[J/OL]. 科学学与科学技术管理：1-22[2025-12-12]. https://doi.org/10.20201/j.cnki.ssstm.20251202.002.

[13] BOYACK K W，KLAVANS R. Co-citation analysis，bibliographic coupling，and direct citation：which citation approach represents the research front most accurately? [J]. Journal of the American Society for Information Science and Technology，2010，61（12）：2389-2404.

[14] 高楠，周庆山. 基于共被引方法的情报学前沿领域识别与演进趋势分析[J]. 现代情报，2024，44（5）：3-19.

[15] 姚茹，乌吉斯古楞，张学福. 基于知识流的研究前沿主题演化分析方法研究：以“基因组编辑技术及其在农作物中的应用”研究前沿为例[J]. 情报理论与实践，2023，46（8）：165-174.

[16] 苗红，王浩桐，李伟伟，等. 面向应用场景的前沿技术识别方法[J]. 情报杂志，2025，44（1）：95-103.

[17] 刘自强，岳丽欣，方曙. 基于主题扩散演化滞后的研究前沿趋势预测方法研究[J]. 情报理论与实践，2023，46（6）：145-154.

[18] 林晗，汤珊红，高强，等. 基于改进HLDA的前沿主题挖掘方法研究[J]. 情报理论与实践，2022，45（11）：188-194.

[19] 刘盼盼，王丽. 战略性新兴产业前沿技术探测研究：以量子计算领域为例[J]. 中国科技论坛，2024（6）：46-57.

[20] 刘小玲，谭宗颖. 基于专利多属性融合的技术主题划分方法研究[J]. 数据分析与知识发现，2022，6（S1）：45-54.

[21] 曾文，王海燕，陈峰，等. 科技前沿探测中的科技信息融合方法研究[J]. 情报学报，2023，42（8）：906-914.

[22] 朱丹浩，黄肖宇，李尧霖，等. 基于大语言模型的法律文本的自动摘要方法[J]. 数据分析与知识发现，2025，9（6）：35-46.

[23] 罗鹏程，王继民，聂磊. 基于生成式大语言模型的文献资源自动分类研究[J]. 情报理论与实践，2024，47（12）：174-182.

[24] CHEN Z Y，XU L C，ZHENG H T，et al. Evolution and prospects of foundation models：from large language models to large multimodal models[J]. Computers，Materials & Continua，2024，80（2）：1753-1808.

[25] KALYUZHNAYA A，MITYAGIN S，LUTSENKO E，et al. LLM agents for smart city management：enhancing decision support through multi-agent AI systems[J]. Smart Cities，2025，8（1）：19.

[26] WANG Y X，WU Z Y，YAO J F，et al. TDAG：a multi-agent framework based on dynamic task decomposition and agent generation[J]. Neural Networks，2025，185：107200.

[27] 赵浜，曹树金. 生成式AI大模型结合知识库与AI Agent开展知识挖掘的探析[J]. 图书情报知识，2025，42（4）：88-101.

[28] CAO C，WANG F Y，LINDLEY L，et al. Managing Linux servers with LLM-based AI agents：an empirical evaluation with GPT4[J]. Machine Learning with Applications，2024，17：100570.

[29] GARCIA K R，CHEN J，XIAO Y R，et al. Human perception of AI capabilities at classifying perturbed roadway signs[J]. IEEE Transactions on Human-Machine Systems，2025，55（4）：499-507.

[30] 邢玲，程兵. 基于结巴分词的领域自适应分词方法研究[J]. 计算机仿真，2023，40（4）：310-316，503.

[31] HUANG D W，YAN C，LI Q，et al. From large language models to large multimodal models：a literature review[J]. Applied Sciences，2024，14（12）：5068.

[32] 韩雪雯，车尚锟，杨梦晴，等. 多模态数据驱动的AI智能体模式设计[J]. 图书情报工作，2024，68（24）：27-37.

[33] STURUA S，MOHR I，AKRAM M K，et al. Jina-embeddings-v3：multilingual embeddings with task LoRA[EB/OL]. [2025-11-12]. https://arxiv.org/abs/2409.10173.

[34] 张金柱，李溢峰. 专利分类序列和文本语义表示视角下的技术融合预测研究[J]. 情报学报，2022，41（6）：609-624.

[35] 田鹏伟，张娴. 基于异构信息网络融合的专利技术主题识别研究[J]. 情报杂志，2021，40（8）：45-52.

[36] LI X M，ZHANG A，LI C C，et al. Relational biterm topic model：short-text topic modeling using word embeddings[J]. The Computer Journal，2019，62（3）：359-372.

[37] 王曦，陈铎. 基于BTM模型的教育舆情热点主题演化研究：以研究生招生考试为例[J]. 情报科学，2022，40（7）：55-60，77.

[38] 徐菲菲，陈赛红，田宇. 基于BTM模型和改进聚类算法的热点话题检测[J]. 计算机应用与软件，2022，39（5）：283-290.

[39] 吴迪，马文莉，杨利君. 遗忘曲线和BTM词频双层加权微博用户画像[J]. 计算机工程与设计，2023，44（12）：3800-3808.

[40] KOLTCOV S. Application of Rényi and Tsallis entropies to topic modeling optimization[J]. Physica A：Statistical Mechanics and its Applications，2018，512：1192-1204.

[41] 张家惠，丁敬达. 基于BERTopic和LSTM模型的新兴主题预测研究[J]. 情报科学，2025，43（1）：98-105，126.

[42] VERHOEVEN D，BAKKER J，VEUGELERS R. Measuring technological novelty with patent-based indicators[J]. Research Policy，2016，45（3）：707-723.

[43] 黄鲁成，蒋林杉，吴菲菲. 萌芽期颠覆性技术识别研究[J]. 科技进步与对策，2019，36（1）：10-17.

[44] 苗红，连佳欣，李伟伟，等. 基于数据挖掘的前沿技术识别方法与实证研究[J]. 系统工程与电子技术，2024，46（9）：3082-3092.

[45] 程旭东，陈祖康，张针霖，等. 基于量子点颜色转换层的Micro LED研究进展[J]. 新能源进展，2025，13（1）：107-120.

[46] 汤晖，廖智燊，魏玉章，等. Mini/Micro LED巨量转移关键技术与装备研究现状[J]. 振动·测试与诊断，2023，43（5）：839-849，1033.

[47] QI L H，LI P A，ZHANG X，et al. Monolithic full-color active-matrix micro-LED micro-display using InGaN/AlGaInP heterogeneous integration[J]. Light：Science & Applications，2023，12：258.

[48] 周佳，闫金健，刘志强，等. Micro LED当前面临的瓶颈及技术进展[J]. 光电子技术，2023，43（2）：91-113.

[49] 冯昭奎. 中日高科技发展的比较与思考：以半导体芯片制造技术为案例[J]. 日本研究，2022（3）：1-15.

[50] AN H J，KIM M S，MYOUNG J M. Strategy for the fabrication of perovskite-based green micro LED for ultra high-resolution displays by micro-molding process and surface passivation[J]. Chemical Engineering Journal，2023，453：139927.

[51] 张立群，李祥高. 全球关键电子材料应用进展与我国未来发展方向[J]. 中国工程科学，2025，27（2）：249-268.

[52] 卢文胜. LED显示芯片行业现状与发展研究[J]. 中国集成电路，2023，32（12）：30-35.

[53] 严子雯，严群，李典伦，等. 高度集成的μLED显示技术研究进展[J]. 发光学报，2020，41（10）：1309-1317.

[54] 王伟，赵甜甜，刘强，等. Mini/Micro LED巨量转移技术研究与发展现状[J]. 光学 精密工程，2023，31（2）：183-199.

作者简介

李贞贞，女，硕士，副研究馆员，研究方向：科技和产业大数据分析、自然语言处理，E-mail：lizz@mail.whlib.ac.cn。
胡思思，女，硕士，馆员，研究方向：光电产业情报。
钟永恒，男，博士，研究员，研究方向：产业竞争情报与产业技术分析、产业智库与大数据建设。
王辉，男，硕士，副研究馆员，研究方向：产业技术分析、大数据分析。
刘佳，女，博士，副研究馆员，研究方向：大数据情报分析、知识产权分析与评价。

Cutting-Edge Technology Identification Method Based on AI Agent and Multi-Feature Text Fusion

LI ZhenZhen[1,2] HU SiSi[1,2] ZHONG YongHeng[1,2] WANG Hui[1,2] LIU Jia[1,2]
(1. National Science Library (WuHan), Chinese Academy of Sciences, Wuhan 430071, P. R. China;
2. Hubei Key Laboratory of Big Data in Science and Technology, Wuhan 430071, P. R. China)

Abstract: Technological innovation serves as a powerful driving force for the high-quality development of industries. This study proposes a method for identifying cutting-edge technologies based on AI agent and multi-feature text fusion, offering insights into grasping directions for technological breakthroughs and exploring opportunities for technological innovation. Using funding projects, academic papers, and patent data as research objects, this approach first applies AI agent for data preprocessing, extracting textual keywords. Then, the semantic features, classification features, and Biterm topic features of scientific and technological documents are fused to represent the text vector and enhance the data features. The clustering algorithm is used to divide the technical topics. Finally, a cutting-edge technology measurement index system based on novelty, growth, influence, and innovation is constructed to identify cutting-edge technologies. An empirical study is conducted in the field of micro-LED. Clustering methods that integrate semantic, classification, and topic features demonstrate superior performance compared to using only single or dual features. The analysis identifies emerging frontier technologies such as chip manufacturing technology and full-color display technology; highlights popular frontier technologies including light-emitting materials and devices, display technology, holographic display, medical applications, detection and repair, and driving circuits; and also uncovers potential frontier technologies like mass transfer technology. Thus, the effectiveness and application value of the proposed method are proved.

Keywords: Cutting-Edge Technology Identification; AI Agent; Multi-Feature Fusion; Biterm Topic Model

（责任编辑：常春）