融合专利异质特征的颠覆性技术识别方法研究*
——以人工智能领域为例

吴洁　刘彦良　谢小东　唐健廷　盛永祥
（江苏科技大学经济管理学院，镇江 212100）

摘要：颠覆性技术是推动产业变革和科技进步的重要力量，准确和快速地识别出颠覆性技术，对于各国把握技术先机、持续提升国家竞争力具有关键性意义。首先通过专利计量指标构建量化特征，并基于规则抽取知识单元获得文本特征；随后采用多模态融合方法将异质性特征统一表征，运用XGBoost算法实现分类识别；最后应用BERTopic模型对识别结果进行颠覆性技术主题挖掘。以人工智能领域为例进行实证研究，结果证明提出的融合异质性特征的识别方法取得较高准确率，在该领域具有良好的实际应用效果。

关键词：颠覆性技术；XGBoost；数据融合；技术识别；专利分析；人工智能
中图分类号：G350　DOI：10.3772/j.issn.1673-2286.2025.07.005
引文格式：吴洁，刘彦良，谢小东，等. 融合专利异质特征的颠覆性技术识别方法研究：以人工智能领域为例[J]. 数字图书馆论坛，2025，21（7）：42-53.

当前，我国正处于创新驱动发展的关键转型阶段，同时面临发达国家技术封锁与传统技术范式制约的双重挑战。在此背景下，颠覆性技术凭借其强大的突破性，展现出摆脱技术困境、重塑竞争格局的重要潜力。作为我国建设科技强国的重要抓手，颠覆性技术创新的战略地位已经提升至国家层面，继《中华人民共和国国民经济和社会发展第十三个五年规划纲要》首次提出颠覆性技术创新以来，中国共产党第十九次全国代表大会也强调了发展颠覆性技术的重要性。《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》[1]进一步指出“加强前沿技术多路径探索、交叉融合和颠覆性技术供给”。科学地识别颠覆性技术有助于明确关键研发方向、优化资源配置，确保我国在科技竞争中抢占先机，从而增强我国综合实力。

由于技术无法直接作为识别对象，而专利是技术创新的有效表征，对颠覆性技术的识别可转化为对专利数据的分析。厘清颠覆性技术的核心特征，并从专利数据中准确识别具有颠覆性的技术具有重要意义。近年来相关学者基于专家意见法[2-3]、科学计量法[4-5]、文本挖掘法[6-7]、机器学习法[8–10]等方法识别颠覆性技术，但是存在以下不足：现有研究通常基于单一数据维度的信息计算，尚未充分挖掘异质数据间的互补价值；对于文本特征的细粒度处理关注不足，缺乏针对专利文本语义的精细化抽取。基于此，本文融合计量数据与文本数据异质特征以实现信息互补，同时构建知识单元抽取规则来细化数据粒度，并对融合特征开展相关性分析。多模态融合方法系统地整合了计量指标特征与文本语义特征，实现了结构化与非结构化数据的深度融合。

收稿日期：2025-06-22
*本研究得到江苏省社会科学基金项目“专利转化视角下江苏推动‘双高协同’创新发展研究”（编号：25ZHB010）、江苏省研究生科研与实践创新计划项目“融合专利计量指标和文本语义的颠覆性技术识别研究”（编号：KYCX25_4313）资助。

1　相关研究

颠覆性技术的概念最早由Bower等[11]提出，他们认为颠覆性技术能够突破传统的技术路线，并彻底改变现有市场格局，构建起新的价值体系。针对颠覆性技术的识别，学术界已进行诸多探索，基于计量指标的识别方法是建立科学的评估体系，将技术特征转化为可以计算的指标来识别颠覆性技术[5]，如从知识关联性、技术潜力和市场潜力角度[12]构造技术评价体系。但是这类方法主要依赖结构化特征，难以捕捉技术内涵。为弥补这一不足，部分研究引入文本挖掘方法，从非结构化数据中提取关键技术概念，分析语义的创新内容和内在联系，如Liu等[13]采用Word2Vec对文本数据进行向量化处理，运用K-means聚类分析得到能源领域潜在颠覆性技术。Jia等[14]通过分析文献的关键词共现网络和关键词频率变化来识别电子信息通信行业的颠覆性技术。针对文本挖掘中粒度不足的问题，部分学者[15]提出了知识单元抽取的思路，能够在细粒度层面增强语义表示的结构性，有助于减少噪声干扰并提炼核心技术信息[16]。然而，目前尚缺乏将知识单元抽取方法引入颠覆性技术识别任务的研究。为进一步提升单一信息维度的识别效果，部分研究已尝试引入数据融合[17]，如Yu等[18]开发图信号处理框架，定义3种不同类型的信号并进行融合。近年来，机器学习技术的不断发展为融合处理后的特征提供了有力的分析手段。李欣等[19]基于机器学习方法对融合后的多模态数据实现高价值专利的识别；Deng等[20]提出一种融合文本、专利权人及基本指标等多维特征的高质量专利评估方法，使用机器学习算法实现专利价值评估。

现有研究方法存在着部分局限性：①缺乏对融合前特征的细粒度处理与语义抽取，导致文本信息利用不充分；②不同特征结构的分析方法往往难以兼顾，异质特征间的关联性未被充分检验，可能影响识别模型的准确性。据此，本文为细粒化文本内容，关注技术方法和应用领域，基于规则和句法分析来抽取摘要中的知识单元，去除无关噪声数据；鉴于异质性特征结构内容不同，借鉴多模态特征融合方法，通过将数值指标和文本语义结合，充分考虑专利数据蕴含的特征，在保留特征独特信息的基础上，利用XGBoost模型对融合特征进行颠覆性技术识别。

2　颠覆性技术识别模型

2.1　模型框架

基于融合专利的计量指标与文本内容这两种类别数据，构建颠覆性技术异质性特征识别框架，如图1所示。具体而言：首先，从incoPat数据库中获取专利数据，根据颠覆性技术特征计算数值指标和颠覆性标签，同时对专利摘要文本进行知识单元抽取和语义向量化处理；其次，借鉴多模态融合方法将数值指标和文本语义进行融合，并通过相关性分析确保特征变量的有效性与稳定性；最后，以专利的颠覆性为标签，将融合特征数据输入XGBoost模型进行训练，在识别出颠覆性技术专利后，利用BERTopic模型挖掘颠覆性技术主题。

2.2　指标体系构建

2.2.1　颠覆性技术特征指标

为系统把握技术的发展脉络，在颠覆性创新理论的基础上，结合生命周期理论，考虑到颠覆性技术专利从诞生到扩散影响的发展路径，构建由技术原创性、市场扩散、战略价值组成的三维指标体系（见表1），对应技术发展的3个阶段：萌芽期、成长期和成熟期。

从技术原创性的角度看，颠覆性技术往往起源于知识基础或技术层面的突破，具有脱离主流技术路径、具备变轨潜力的早期特征；从市场扩散的角度看，处于成长期的颠覆性技术通常开始显现其强劲的扩散能力和日益增强的市场竞争力，随着技术性能的持续改进，其适用范围不断扩大，逐步突破原有的市场壁垒；从战略价值的角度看，成熟阶段的技术能够被后续技术继承与延伸，推动整个技术体系的进一步演进，并在市场中产生持续影响，创造较高的经济与社会价值。

2.2.1.1　技术原创性维度

（1）创新性。技术的创新性取决于它能否用新颖的方式对现有技术进行重新组合，或是能否产生新的技术原理来创造全新的技术路线[21]。目标专利与引用专利的重叠性越小，它的创新性越大。创新性计算方法[22]如式（1）所示，通过后向引用专利的IPC分类号数量与目标专利IPC分类号数量之比进行衡量。

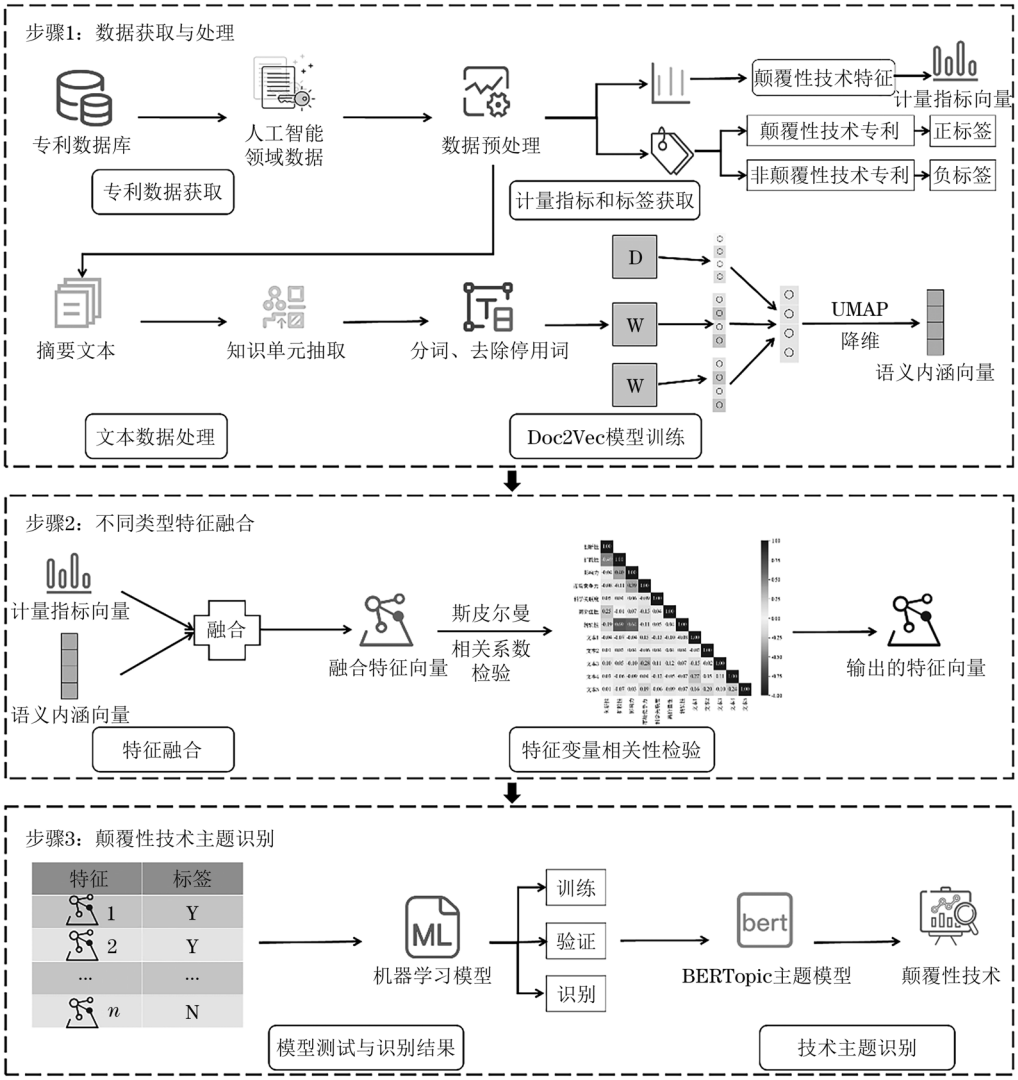

图1　模型框架图

表1　颠覆性技术识别指标

阶段	维度	指标名称	指标含义	参考文献
萌芽期	技术原创性	创新性	技术在原理、方法或应用上的新颖性，是技术颠覆的显著属性	文献[21]、文献[23]
		科学关联度	衡量技术与基础科学理论的关联程度，反映理论深度与研究支撑	文献[7]、文献[24]
		转轨性	改变技术的原有性能标准，推动技术向全新轨道跃迁	文献[22]、文献[25]
成长期	市场扩散	扩散性	打破现有竞争格局，并在进入主流市场过程中呈现出快速扩散趋势	文献[26]、文献[27]
		市场竞争力	反映技术在市场中的竞争优势与突破行业壁垒的能力	文献[4]
成熟期	战略价值	影响力	衡量技术被后续研究或应用采纳的程度，体现持续影响作用	文献[26]、文献[28]
		高价值性	反映技术引领未来发展和推动产业形成的潜力	文献[10]、文献[29]

$$I_{ij}^{\mathrm{AI}}=\frac{\sum_{j=1}^{n}C_j\Big/n}{C_i}(j=1,2,\cdots,n) \quad (1)$$

式中：I_{ij}^{AI}为创新性；C_j为后向引用专利j的IPC分类号数量；n为后向引用专利数量；C_i为目标专利IPC分类号数量。

（2）科学关联度。颠覆性技术的出现源自对基础研究的理解和延伸，结合前沿知识理论进行突破创新。使用领域内专利引用科技文献数量代表科学关联度[7, 24]，在所属领域内引用的科技文献数量越多，与知识前沿关联性越强，如式（2）所示。

$$I_q^{\mathrm{SR}}=\frac{\sum_{p\in q}t_p}{N_q} \quad (2)$$

式中：I_q^{SR}为技术领域的科学关联度；p为技术领域q中的专利；t_p为专利p引用的科技文献数量；N_q为技术领域q的全部专利数量。

（3）转轨性。颠覆性技术拥有非连续跃迁的特性，显著区别于渐进性技术的发展路径。使用专利的IPC分类号研究转轨性，确定所属领域和引用类别的多样性，反映目标专利对先前技术的依赖性。计算方法[25]如式（3）所示，通过衡量目标专利引用结构与其自身差异的大小来评估转轨性的高低。

$$I_{ij}^{TN}=\frac{\sum_{j=1}^{n}C_{ij}^{T}}{C_j^{B}}(j=1,2,\cdots,n) \tag{3}$$

式中：I_{ij}^{TN}为转轨性；C^{T}_{ij}为专利i区别于原始专利j的IPC分类号数量，专利i是专利j的前向引用专利；C_j^{B}为原始专利j的全部后向引用的IPC分类号数量。

2.2.1.2　市场扩散维度

（1）扩散性。专利凭借扩散性打破经济和技术的壁垒，吸引更多产业领域的技术跟进[26]。扩散性的计算方法[22]如式（4）所示，通过前向引用专利与目标专利的IPC分类号数量之比予以衡量。

$$I_{ik}^{DI}=\frac{\sum_{k=1}^{m}C_k\Big/m}{C_i}(k=1,2,\cdots,m) \tag{4}$$

式中：I_{ik}^{DI}为扩散性；C_k为前向引用专利k的IPC分类号数量；m为前向引用专利数量。

（2）市场竞争力。较多的权利要求数量，意味着专利扩大了技术边界和法律保障范围，增强在市场中的排他性和独特性，带来更多的商业开发价值。使用领域内专利权利要求数量代表目标专利的市场竞争力[4]，其占比高能为技术顺利打破垄断、完成颠覆创造条件，如式（5）所示。

$$I_q^{MC}=\frac{\sum_{p\in q}c_p}{N_q} \tag{5}$$

式中：I_q^{MC}为某技术领域q的市场竞争力；c_p为专利p的权利要求数量。

2.2.1.3　战略价值维度

（1）影响力。颠覆性技术具有显著的引领性，其影响力体现在对后续技术发展与变革的推动作用上[5]。引用关系反映了潜在的技术流动与继承[30]，使用领域内被引频次代表目标专利的影响力[4]，其占比高表明技术在后续研究中参考性和借鉴性强，如式（6）所示。

$$I_q^{TI}=\frac{\sum_{p\in q}f_p}{N_q} \tag{6}$$

式中：I_q^{TI}为影响力；f_p为专利p的被引频次。

（2）高价值性。高价值的颠覆性技术在推动新领域发展、突破既有技术瓶颈方面具有较大贡献，其核心价值在于对后续创新的驱动能力。衡量颠覆性技术的高价值性，可通过分析其在专利引用中的IPC分类号占比来实现，比值越高，涉猎领域越多样，如式（7）所示。

$$I_{ik}^{HV}=\frac{\sum_{k=1}^{m}C_{ik}^{V}}{C_k^{B}}(k=1,2,\cdots,m) \tag{7}$$

式中：I_{ik}^{HV}为高价值性；C^{V}_{ik}为专利i与专利k共享的IPC分类号数量；C_k^{B}为专利k的所有后向引用的IPC分类号数量。

2.2.2　细粒度知识单元抽取

专利文本既蕴含着丰富的技术创新价值，又是专利知识单元的载体。句粒度知识单元是文献中表达特定语义的语句，具有序列性和功能性[31]。摘要凝练地囊括了知识应用领域和技术方法原理信息，适合细粒度的深度挖掘[32]。通过从摘要的语义特征中提取知识单元，可将分析的精细度从篇章层面转换至精炼的句子层面，这有助于简化复杂语义，减少噪声干扰，抓住技术核心。

技术的颠覆性往往来源于领域内的突破和方法上的创新，为深入挖掘专利摘要的技术内涵，将摘要中的知识单元划分为技术领域知识单元和技术方法知识单元。基于摘要的句法规则及关键动词列表，利用正则表达式进行匹配，完成对领域和方法的知识单元提取。技术领域知识单元通常位于句子中的宾语位置，其核心动词短语包括“涉及”“适用于”等，由此能够有效识别摘要中与技术领域相关的句子内容。技术方法知识单元通过识别动词列表如“提供”“包括”等[33]后续的短句，匹配技术方案的相关描述。按照这两类线索词分别确定知识单元在摘要中的待提取范围，使用正则表达式完成对要素的提取，并加以合并，

获得一项专利的完整知识单元。对于无法进行规则匹配的专利摘要，引入自然语言处理技术进行句法组块分析，实现全部专利摘要的知识单元覆盖。知识单元抽取示例见表2。

表2　知识单元抽取结果示例

知识单元类型	摘要	抽取结果
技术领域知识单元	本发明提供一种问答处理方法、装置、电子设备及存储介质；涉及人工智能的自然语言处理技术……	涉及人工智能的自然语言处理技术
技术方法知识单元		提供一种问答处理方法、装置、电子设备及存储介质

在完成知识单元的抽取之后，依据特定领域的特征构建专属词典。通过分词处理以及去除停用词的操作形成文本语料库，并将其输入Doc2Vec模型。作为Word2Vec模型的扩展，Doc2Vec克服了忽略语义关联和语境信息的局限性，能够很好地结合语义内容并把文本映射至固定长度的向量[34]。这种高维文本特征也可能导致冗余信息的增加以及模型训练过程的复杂化，引入UMAP降维方法来解决这一问题。UMAP降维在尽可能保留数据的全局结构的同时，可以有效降低计算困难度，优化后续的特征融合。

2.2.3　异质性特征向量融合

数据的异质性源于其属性（如格式、表示或类型）的差异[35]，可划分为句法异质性、语义异质性和统计异质性。异质性体现在不同模态信息的质量、结构和表示形式上[36]，文本特征与结构化指标在数据类型上存在显著差异，本文将二者称为异质性特征。来自同一数据源内部的结构化量化特征和非结构化语义特征在表征形式和数据结构上有着系统性差异，这种不同致使两类特征无法直接进行运算，需要通过预处理与同一表达方式实现特征体系的融合。

异质性特征融合的过程在于综合利用不同类型数据所承载的互补信息，从多维度进行深入分析，以创建更加全面的特征表示，从而提高模型的性能。可计量的指标特征反映了颠覆性技术的属性，文本特征则深挖语义内容，蕴含着技术核心和创新点，二者虽形式不同，但具有互补性。若只依靠一种类型的特征进行技术识别，则会忽略掉一些关键要素。为充分利用专利数据中的多维信息，借鉴多模态联合架构的思想来有效融合指标特征和文本特征，采用拼接方式[19]，确保不同形式的信息得以保留。在转换来源异质的语义内涵和计量指标特征后映射到一个统一的向量空间，将其组合形成融合向量，子空间的每个内容都对应各模态的信息。

在机器学习建模中，若直接将未经处理的融合特征输入模型，可能会忽视因素之间的自相关性，导致模型出现过拟合或多重共线性等问题，使得模型的识别效果和稳定性欠佳。斯皮尔曼相关系数的分析结果能为模型筛选主要因素提供参考。因此，在将融合特征输入模型之前，需要进行斯皮尔曼分析来筛选变量、优化特征组合，以便提升模型的性能和稳定性。

2.3　识别模型构建

在知识单元抽取与特征融合的基础上，进一步构建基于机器学习的识别模型，将融合后的异质特征作为输入变量，以颠覆性指数为标签，通过对历史专利数据的分析，识别出具有颠覆性潜力的技术创新。

2.3.1　颠覆性指数标签

CD指数由Funk等[37]首次提出，随后逐渐获得学术界的广泛认可，并被应用于颠覆性方面的研究[38]。CD指数具有量化颠覆性创新和技术演化的潜力，当一项具有颠覆性的专利出现后，它能够中断既有技术族群的演进路径，即后续专利只会引用它本身，而不会继续遵循该专利之前的技术路线[39]，从而促成技术的迭代更新。CD指数计算如式（8）所示。

$$I_t^{CD}=\frac{1}{N_t}\sum_{i=1}^{n}\frac{-2a_{it}b_{it}+a_{it}}{w_{it}} \tag{8}$$

式中：N_t为在时间t的专利数量；w_{it}为专利i在时间t内的权重，参考文献[37]将w_{it}赋值为1；若后续专利引用了目标专利，则a_{it}为1，反之则为0；若后续专利引用了目标专利的后向引证专利，则b_{it}为1，否则为0。CD指数的取值范围为[−1, 1]。在考虑到CD指数的局限性和本文应用范围的基础上，将CD指数等于1且被引频次超过5[38, 40]的目标专利认定为颠覆性技术，确保技术具备基本影响力。CD指数计算图例如图2所示，展现了CD指数为−1、1和1/6的专利的引证关系。

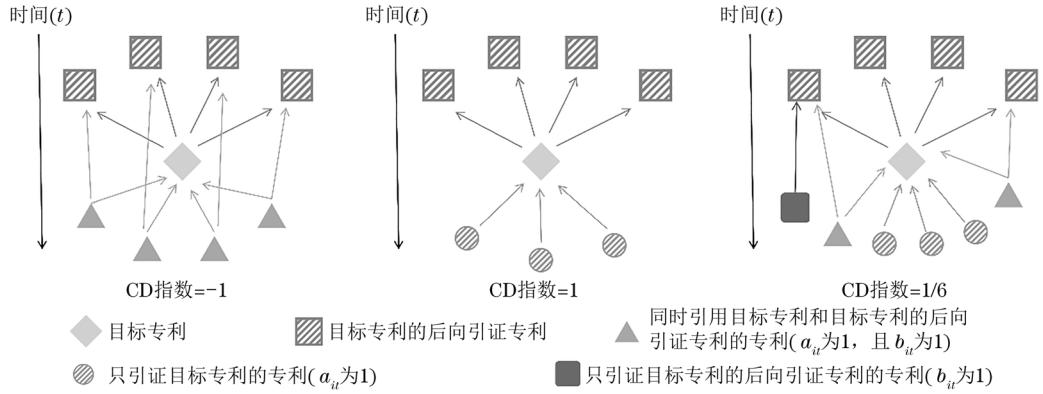

图2　CD指数计算图例

2.3.2　机器学习模型

机器学习方法在处理海量数据时表现出色，并且易于实现。在模型的选择上，考虑了多种常用于分类任务的机器学习算法，如XGBoost、随机森林、决策树、K最近邻（KNN）等模型。XGBoost与随机森林在处理高维数据和特征间的复杂交互方面具有显著优势；决策树模型具有较好的可解释性，便于分析特征的重要性；KNN模型能够有效捕捉样本间的局部相似性。

3　实证研究——以人工智能领域为例

人工智能是数字时代推动国家发展和企业进步的重要数字技术，并且在促进创新过程中扮演着关键的角色。人工智能技术多次被纳入政府工作报告，这一举措充分表征了人工智能在国家创新体系中的重要地位，同时也吸引了科研与产业领域的高度关注。因此，以人工智能为研究对象，旨在从这一前沿领域识别颠覆性技术。

3.1　数据检索与获取

选取incoPat数据库作为研究数据的来源，参考我国颁布的《关键数字技术专利分类体系》和《人工智能标准化白皮书（2021版）》设定检索策略，具体的检索式为：TIAB=（自然语言处理 OR 机器学习 OR 计算机视觉 OR 增强现实 OR 目标追踪 OR 语音处理 OR 情绪分析 OR 场景理解 OR 生物特征识别）。2010年以前中国人工智能发展尚处于萌芽阶段，自2011年起进入涌现期，2024年专利数据尚未完全公开，因此将数据时间范围确定为2011年1月1日—2023年12月31日。经数据库初步检索得到24 657项中国有效发明专利，获取摘要、引证次数、被引证次数、权利要求数量等著录项字段。在数据预处理、噪声清除后，获得23 120条数据。

为了确保研究的时效性和动态性，按照时间将专利数据划分为两个数据集：数据集1涵盖2011年1月1日—2020年12月31日的专利数据，共计14 472项；数据集2包含2021年1月1日—2023年12月31日的专利数据，共计8 648项。数据集1用于模型训练，数据集2则基于数据集1所训练的模型识别颠覆性专利。

3.2　指标文本数据处理

3.2.1　异质性特征向量融合

专利分析的量化指标可以通过数值反映颠覆性技术的特征，按照2.2.1节中提出的计算方法，分别提取数据集中专利样本的各技术特征指标，依据2.2.2节的知识单元规则，从专利摘要中抽取技术领域知识单元和技术方法知识单元，对于未被提取到的知识单元，借助自然语言处理的句法分析进行处理。为了保证分词的准确性和技术术语识别的可靠性，结合人工智能领域专家知识和搜集到的人工智能词汇，构建2 427项中英文术语作为自定义词典的内容。采用Doc2Vec模型将预处理文本转换为高维语义向量，利用UMAP算法进行降维处理。以融合数据训练机器学习模型的评估结果为依据，尝试不同的Doc2Vec模型参数和UMAP算法。

在Doc2Vec的输出向量维度数为100时，对经UMAP降维的不同维度进行了准确率对比，结果如图3所示。当使用UMAP算法把维度数降为5时，文本表示的效果最佳。降维后的5个知识单元向量分别用文本1~5表示，将它们和指标特征进行多模态融合，形成一个综合特征向量，并且对其进行Z-score归一化处理，避免

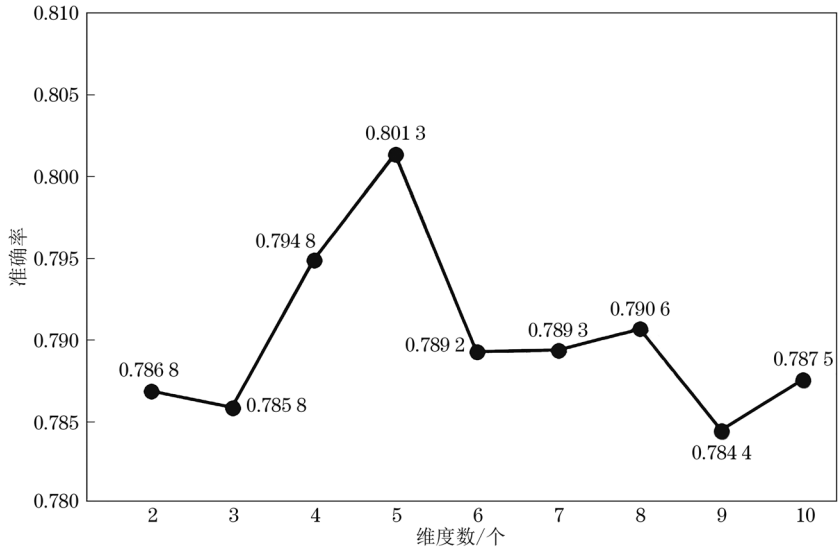

图3　不同维度准确率对比

不同类型特征因量纲差异在模型训练中产生偏差。

3.2.2　输入特征相关性分析

为衡量融合指标之间的依赖关系，选用斯皮尔曼方法分析变量之间的相关程度，结果如图4所示，相关系数的绝对值越大、方框颜色越深代表了相关性越强。若变量间的相关系数绝对值大于0.80则意味着可能引发多重共线性问题，研究中相关系数最大为0.69，进一步考察多重共线性的程度。利用方差膨胀因子（VIF）进行多重共线性检验，当VIF≥10时，认为变量之间存在明显多重共线性问题，指标间依赖性很大。检验结果如表3所示，各变量间VIF＜4，表明所选取的指标不存在多重共线性，后续可以构建机器学习模型。

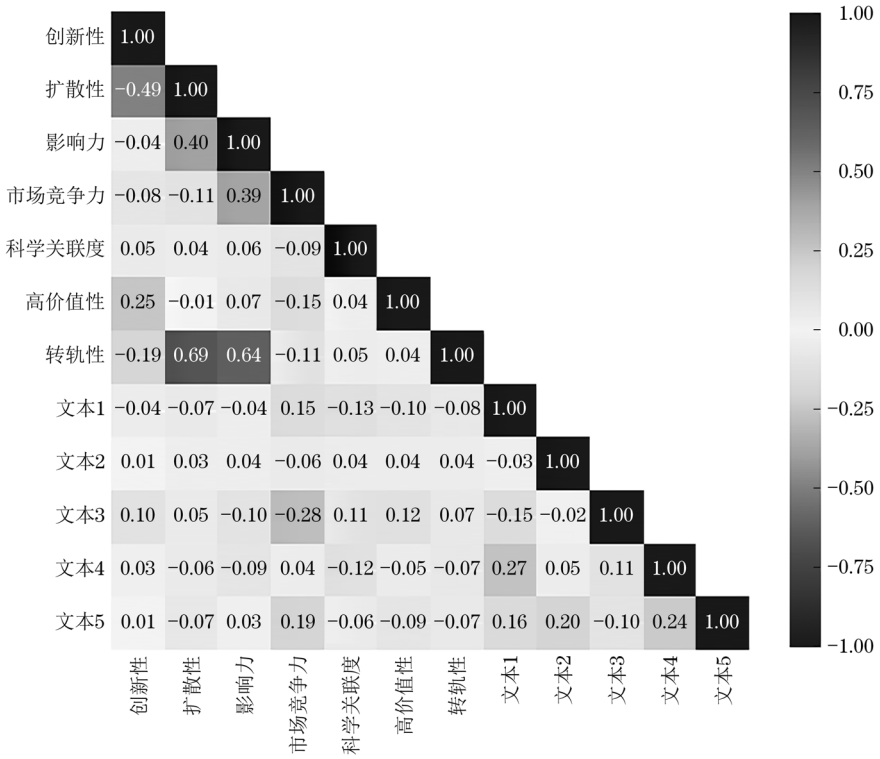

图4　斯皮尔曼相关系数热力图

表3　多重共线性检验结果

指标	VIF	指标	VIF	指标	VIF
创新性	1.16	科学关联度	2.54	文本2	1.06
扩散性	1.26	高价值性	1.07	文本3	1.11
影响力	2.04	转轨性	1.17	文本4	1.19
市场竞争力	3.54	文本1	1.14	文本5	1.12

3.3　机器学习模型构建与评估

通过3.2节所述的特征向量融合和标签计算方法，获得机器学习模型所需的数据输入。数据集1的每个专利样本均包含综合特征向量和对应的二分类标签，按照8∶2的比例划分为训练集和测试集，用于模型的构建与性能评估。

3.3.1　对比实验

为探究多种机器学习方法在融合特征数据上的分类效果，选择XGBoost、随机森林、决策树和KNN模型，采用五折交叉验证与网格搜索方法优化超参数。XGBoost模型的最优参数配置如下：max_depth=6，learning_rate=0.05，n_estimators=100，gamma=0.1。在模型训练完成后，以准确率、召回率、F1分数和ROC曲线下面积（AUC）为评估指标。不同模型效果对比如表4所示。由实验结果可知，XGBoost在多种算法中表现最优，其准确率、召回率、F1分数和AUC值均高于其他模型，对融合特征的处理能力更为出众。

表4　不同模型效果对比

模型	准确率	召回率	F1分数	AUC
XGBoost	0.80	0.82	0.67	0.88
随机森林	0.79	0.79	0.64	0.87
决策树	0.76	0.50	0.51	0.67
KNN	0.62	0.50	0.39	0.60

3.3.2　消融实验

为进一步评估异质性特征选取的有效性，设计消融实验进行测试分析，通过移除特征组件来解析各模块对识别任务的贡献程度，消融实验效果对比如表5所示。

由表5可知，移除特征组件后模型的评估指标均低于原模型。去除文本特征模型的性能有所下降，但整体来看保留单一指标特征的效果下降程度最小，模型能保持相对稳定的表现。去除指标特征模型的准确率下降至0.62，召回率、F1分数和AUC均大幅度降低，模型效果下降十分明显。抽取知识单元的异质性特征融合模型通过过滤文本冗余信息，有效捕捉正类样本语义的关键信息。相较于未抽取知识单元的模型，所提模型的召回率和AUC值显著提升，意味着原始文本中存在噪声，抽取知识单元能够减少信息干扰，强化模型对细粒度技术内容的关注。

表5　消融实验效果对比

模型	准确率	召回率	F1分数	AUC
所提模型	0.80	0.82	0.67	0.88
去除文本特征	0.70	0.77	0.56	0.76
去除指标特征	0.62	0.30	0.28	0.52
未抽取知识单元	0.74	0.75	0.59	0.79

3.4　颠覆性技术主题识别与分析

经过机器学习模型对比筛选后，将数据集2的融合特征输入训练好的XGBoost模型，完成颠覆性技术识别。模型在2021—2023年的8 648项样本中筛选得到了656项具有颠覆性的专利，识别结果如表6所示。

为进一步对识别所得颠覆性专利进行主题挖掘，采用BERTopic建模方法生成具有较强解释性的主题分类结果，选择paraphrase-multilingual-MiniLM-L12-v2模型生成语义嵌入，将UMAP降维参数设为n_neighbors=15、n_components=5，HDBSCAN聚类参数设为min_cluster_size=10，共得到5个有效技术主题，如图5所示。

依据图5呈现出的主题词内容，为确保主题名称的语义准确，利用GPT-5对5个技术主题进行客观的技术描述，并结合领域专家意见，将这5个技术主题分别对应为自主可信计算、智能信息抽取、三维重建、智慧环境监测和智能机器人。自主可信计算旨在通过可信硬件与算法机制确保智能系统的安全与可控性，体现出较高的创新性与科学关联度；智能信息抽取

表6　识别结果（部分）

公开（公告）号	CD指数	公开（公告）号	CD指数	公开（公告）号	CD指数
CN113765898A	1	CN112863665A	1	CN112881993A	1
CN113782125A	1	CN114152825A	1	CN114219212A	1
CN114666564A	1	CN112666464A	1	CN113311799A	1
CN116531089A	1	CN114723701A	1	CN113588692A	1

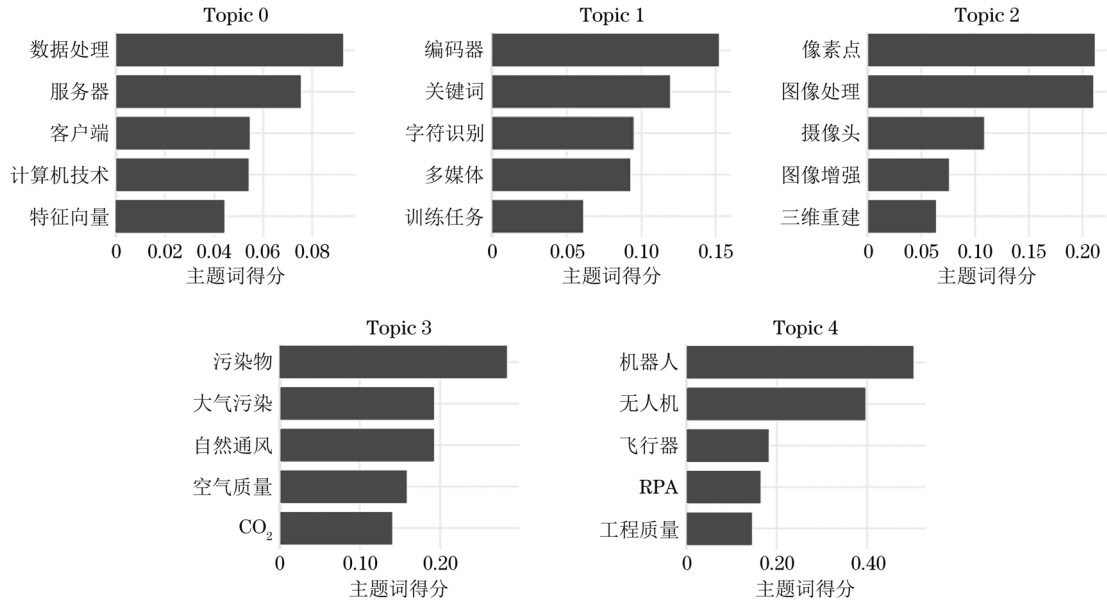

图5　主题聚类可视化结果

通过跨模态知识融合与语义理解算法，应用于司法、教育等领域，展现出较强的扩散性与影响力；三维重建技术突出转轨性特征，通过深度学习与虚拟现实技术的融合，实现从二维图像识别到沉浸式场景构建的技术跃迁；智慧环境监测具备较强的影响力，利用人工智能算法对生态和环境进行实时监测与预测，助力绿色可持续发展；智能机器人作为人工智能的集成应用形态，不仅体现了技术集成的高价值性，也推动了制造业与服务业的深度变革，有着较强的市场竞争力。

3.5　结果验证

国家政策为颠覆性技术主题识别结果的可靠性验证提供了有力支撑[41]。以政策验证法为外部参考，通过检索2022—2024年国家层面发布的相关政策文件，发现所识别的5个人工智能技术主题均在政策中获得呼应（见表7），从而在一定程度上验证了识别结果的合理性与现实相关性。所识别出的颠覆性技术主题在国家政策文件中有所体现，这一结果从政策角度印证了模

表7　政策文本解读与对应的技术主题

技术主题	政策文件名称	年份	发布主体
自主可信计算	中共中央办公厅 国务院办公厅关于完善市场准入制度的意见	2024	中共中央办公厅、国务院办公厅
智能信息抽取	关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见	2022	工业和信息化部等部门
三维重建	虚拟现实与行业应用融合发展行动计划（2022—2026年）	2022	工业和信息化部等部门
智慧环境监测	关于加快建立现代化生态环境监测体系的实施意见	2024	生态环境部
智能机器人	国家能源局关于加快推进能源数字化智能化发展的若干意见	2023	国家能源局

型识别的有效性，也增强了结论的可靠性与说服力。政策层面的关注说明识别出的技术方向具备良好的发展环境与实践潜力。

4　结语

新质生产力来源于颠覆性技术创新的突破，准确识别颠覆性技术有助于掌握未来经济发展的主动权，抢占全球科技竞争制高点。在现有指标方法与文本方法的基础上，本文提出改进路径，即在计量指标的框架中引入语义特征，并通过细粒度文本处理实现两类特征的异质融合，从而在兼顾全面性的同时提高识别精度，以此构建融合异质性特征的颠覆性技术识别框架。首先，通过分析颠覆性技术特征，计算得到各特征指标测度值；其次，在更加细粒化挖掘语义内涵后，利用Doc2Vec算法将文本中的知识单元转换为向量表示；再次，融合技术特征和知识单元向量将其构成识别指标体系，作为识别颠覆性专利的基础；最后，依托XGBoost模型拟合融合指标和颠覆性标签的非线性关系，对识别得到的颠覆性专利进行BERTopic主题建模，识别出颠覆性技术主题。以人工智能领域为例，验证了方法的有效性。研究结果有助于该领域决策制定者拟定技术研发战略，推动科技创新发展并形成新质生产力，为企业筛选和培育颠覆性技术提供有力支持，减少试错成本。

本文的创新之处在于：①构建颠覆性技术识别的新框架，通过多模态融合方法系统整合计量指标特征与文本语义特征，实现结构化与非结构化数据的深度融合；②设计基于人工智能领域的知识单元抽取方法，突破传统文本分析的粗粒度局限，更细粒化地测度文本内涵。本文依然存在一定的不足，只利用了专利数据进行分析，未来的研究可以结合更多的数据源。此外，还可进一步扩展至其他领域进行更加泛化的研究和分析。

参考文献

[1] 中华人民共和国国家发展和改革委员会. 中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要[EB/OL]. [2025-06-14]. https://www.ndrc.gov.cn/xxgk/zcfb/ghwb/202103/t20210323_1270124.html.

[2] RATHORE B，GUPTA R，BISWAS B，et al. Identification and analysis of adoption barriers of disruptive technologies in the logistics industry[J]. The International Journal of Logistics Management，2022，33（5）：136-169.

[3] SOMMARBERG M，MÄKINEN S J. A method for anticipating the disruptive nature of digitalization in the machine-building industry[J]. Technological Forecasting and Social Change，2019，146：808-819.

[4] 陈育新，卢俊，韩毅，等. 基于专利文献的颠覆性技术识别研究：以人工智能为例[J]. 情报学报，2022，41（11）：1124-1133.

[5] 李乾瑞，郭俊芳，黄颖，等. 基于专利计量的颠覆性技术识别方法研究[J]. 科学学研究，2021，39（7）：1166-1175.

[6] 金可怡，周立军，杨静. 基于SAO结构的颠覆性技术关联机会发现路径研究[J]. 情报杂志，2024，43（9）：84-91，111.

[7] 刘俊婉，庞博，徐硕. 基于弱信号的颠覆性技术早期识别研究[J]. 情报学报，2023，42（12）：1395-1411.

[8] 黄颖，张慧，叶冬梅，等. 基于Gartner曲线的颠覆性技术生命周期模型构建研究：以新一代人工智能技术为例[J]. 现代情报，2025，45（8）：70-84.

[9] 徐硕，李静鸿，安欣，等. 基于专利术语的颠覆性技术识别及实证研究[J]. 图书情报工作，2024，68（2）：62-72.

[10] 王莉晓，陈伟，邱含琪. 基于机器学习的颠覆性技术弱信号识别模型研究[J]. 数据分析与知识发现，2024，8（S1）：63-75.

[11] BOWER J L，CHRISTENSEN C M. Disruptive technologies：catching the wave[J]. Harvard Business Review，1995，73（1）：43-53.

[12] 罗素平，寇翠翠，金金，等. 基于离群专利的颠覆性技术预测：以中药专利为例[J]. 情报理论与实践，2019，42（7）：165-170.

[13] LIU X W，WANG X Z，LYU L C，et al. Identifying disruptive technologies by integrating multi-source data[J]. Scientometrics，2022，127（9）：5325-5351.

[14] JIA W F，XIE Y P，ZHAO Y N，et al. Research on disruptive technology recognition of China's electronic information and communication industry based on patent influence[J]. Journal of Global Information Management，2021，29（2）：148-165.

[15] 冯儒佳，王忠义，王艳凤，等. 科技论文的多粒度知识组织框架研究[J]. 情报科学，2016，34（12）：46-50，54.

[16] 吴海婷，唐晓波，董克. 引入语义引用信息的专利创新性测度研究[J]. 现代情报，2025，45（9）：140-149.

[17] 冯立杰，秦浩，王金凤，等. 融合专利数据与社交媒体数据的潜在颠覆性技术识别：基于深度学习模型[J]. 情报学报，2024，43（2）：181-197.

[18] YU H Q, LIANG Y A. A framework for predicting scientific disruption based on graph signal processing[J]. Information Processing & Management, 2024, 61(6): 103863.

[19] 李欣, 李悦, 冯野, 等. 基于多模态数据融合与机器学习的高价值专利早期识别方法[J]. 情报杂志, 2024, 43(6): 134-144.

[20] DENG N, ZHANG J A. HMFM: a method for identifying high-value patents by fusing multiple features[J]. Computers, Materials & Continua, 2025, 82(2): 2235-2254.

[21] 张金柱, 王秋月, 仇蒙蒙. 颠覆性技术识别研究进展综述[J]. 数据分析与知识发现, 2022, 6(7): 12-31.

[22] 王海军, 于佳文. 基于专利发展路径的颠覆性技术识别: 以智能语音领域为例[J]. 科技管理研究, 2022, 42(6): 170-181.

[23] 马永红, 孔令凯, 林超然, 等. 基于异构数据的颠覆性技术识别研究: 以智能制造装备领域为例[J]. 现代情报, 2022, 42(7): 92-104.

[24] 赵一鸣, 孙运涛, 孙晓蕾. 颠覆性技术早期识别方法及其在氢能源领域的应用[J]. 图书情报工作, 2024, 68(17): 95-108.

[25] 王海军, 于佳文. 基于专利和微博的颠覆性技术主题识别研究: 以人工智能领域为例[J]. 中国科技论坛, 2024(7): 83-94, 109.

[26] 武建龙, 刘禹彤, 陈劲, 等. 基于专利挖掘和Gompertz模型的颠覆性技术识别方法研究[J]. 科研管理, 2024, 45(4): 62-72.

[27] 李文清, 齐晓曼, 庄子艾, 等. 基于专利文献的颠覆性技术识别框架研究: 以绿氢行业为例[J]. 世界科技研究与发展, 2024, 46(2): 226-238.

[28] 许佳琪, 汪雪锋, 雷鸣, 等. 从突破性创新到颠覆性创新: 内涵、特征与演化[J]. 科研管理, 2023, 44(2): 1-13.

[29] 陈伟, 王莉晓. 颠覆性技术弱信号识别漏斗模型研究[J]. 图书情报工作, 2024, 68(8): 97-111.

[30] 唐健廷, 吴洁, 谢小东. 基于动态价值与多层网络的关键核心技术识别[J]. 情报杂志, 2025, 44(8): 136-145.

[31] 李广建, 袁钺. 基于深度学习的科技文献知识单元抽取研究综述[J]. 数据分析与知识发现, 2023, 7(7): 1-17.

[32] JEE J, SHIN H, KIM C, et al. Six different approaches to defining and identifying promising technology through patent analysis[J]. Technology Analysis & Strategic Management, 2022, 34(8): 961-973.

[33] 唐晓波, 吴海婷, 吴佳琳. 基于特征知识元的专利语义引用识别方法研究: 以量子计算领域为例[J]. 情报理论与实践, 2023, 46(10): 86-95.

[34] 谢小东, 吴洁, 盛永祥, 等. 基于图卷积网络的发明人跨领域合作伙伴识别方法[J]. 情报杂志, 2024, 43(4): 175-183, 167.

[35] KAMM S, VEEKATI S S, MÜLLER T, et al. A survey on machine learning based analysis of heterogeneous data in industrial automation[J]. Computers in Industry, 2023, 149: 103930.

[36] LIANG P P, ZADEH A, MORENCY L P. Foundations & trends in multimodal machine learning: principles, challenges, and open questions[J]. ACM Computing Surveys, 2024, 56(10): 1-42.

[37] FUNK R J, OWEN-SMITH J. A dynamic network measure of technological change[J]. Management Science, 2016, 63(3): 791-817.

[38] PARK M, LEAHEY E, FUNK R J. Papers and patents are becoming less disruptive over time[J]. Nature, 2023, 613(7942): 138-144.

[39] 何郁冰, 徐美娟, 荆晶, 等. 基于创新采用框架的颠覆性技术识别研究[J]. 科学学研究, 2025, 43(1): 66-78.

[40] BORNMANN L, DEVARAKONDA S, TEKLES A, et al. Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers[J]. Quantitative Science Studies, 2020, 1(3): 1242-1259.

[41] 吕鲲, 张未旭, 靖继鹏. 基于CLIP-LDAGV多模态信息融合的颠覆性技术主题识别研究: 以新能源领域为例[J]. 情报学报, 2025, 44(3): 353-368.

作者简介

吴洁, 女, 博士, 教授, 博士生导师, 研究方向: 知识管理。
刘彦良, 女, 硕士研究生, 通信作者, 研究方向: 知识管理, E-mail: 185377325@qq.com。
谢小东, 男, 博士研究生, 研究方向: 数据挖掘。
唐健廷, 男, 博士研究生, 研究方向: 知识管理。
盛永祥, 男, 博士, 教授, 硕士生导师, 研究方向: 复杂网络。

Disruptive Technology Identification Method Integrating Patent Heterogeneous Features: A Case Study of Artificial Intelligence Field

WU Jie　LIU YanLiang　XIE XiaoDong　TANG JianTing　SHENG　YongXiang
(School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212100, P. R. China)

Abstract: Disruptive technologies serve as a crucial driving force behind industrial transformation and technological advancement. Accurately and efficiently identifying such technologies is essential for countries to seize technological opportunities and continuously strengthen national competitiveness. First, we construct quantitative features using patent bibliometric indicators and extract textual features by applying rule-based knowledge unit extraction. Then, we employ a multimodal fusion approach to unify heterogeneous features and use the XGBoost algorithm to achieve classification and identification. Finally, we apply the BERTopic model to extract disruptive technology themes from the identification results. An empirical study conducted in the field of artificial intelligence demonstrates that the proposed recognition method integrating heterogeneous features achieves high accuracy and exhibits excellent practical application effectiveness in this domain.

Keywords: Disruptive Technology; XGBoost; Data Fusion; Technology Identification; Patent Analysis; Artificial Intelligence

（责任编辑：刘伟）