基于特征测度和SciBERT模型的突破性科学创新主题识别研究*
——以农业机器人领域为例

齐世杰　串丽敏　张辉　姚茹　贾倩　赵静娟
（北京市农林科学院数据科学与农业经济研究所，北京 100097）

摘要：突破性科学创新是推动科技进步和社会经济发展的核心动力，准确识别突破性科学创新，对于开辟科技创新的新赛道、实现跨越式发展具有重要意义。首先从突破性科学创新特征入手，综合考虑论文的多样性、均衡性和差异性，利用True Diversity测度指标和学科共现网络进行潜力论文的测度与筛选；然后采用SciBERT+K-means+UMAP方法进行主题建模与可视化，从新颖性、内容变革性、影响力、突变性4个维度构建多特征的突破性科学创新指标体系，融合多种文本挖掘方法识别突破性科学创新主题；最后以农业机器人领域为例，识别出4个突破性科学创新主题：多模态感知与自适应智能决策、机器人仿生驱动与异质能源耦合、群体协同与云边端架构技术、高精度轨迹控制与高动态响应，验证了模型的有效性和可靠性。研究不仅为突破性科学创新主题识别提供了新思路，还为智慧农业领域的原始创新提供了可参考的依据。

关键词：突破性科学创新；主题识别；农业机器人；机器学习；跨学科研究；SciBERT；K-means
中图分类号：G353.1　DOI：10.3772/j.issn.1673-2286.2025.09.003
引文格式：齐世杰，串丽敏，张辉，等. 基于特征测度和SciBERT模型的突破性科学创新主题识别研究：以农业机器人领域为例[J]. 数字图书馆论坛，2025，21（9）：16-27.

科技创新是突破关键核心技术、实现科技自立自强的必由之路。突破性创新是一种不连续、颠覆性的科技创新途径[1]，其创新强度较高，对科技进步、社会发展具有重大的推动作用，对解决我国关键核心技术受限制的困境具有战略意义[2]。突破性科学创新是以知识突破为主的科学范式的变革，是科学发展中取得的重要成果，强调消除科学进步障碍、创造新理论或改进现有理论[3]。2024年，习总书记在全国科技大会上进一步强调：“要切实提升基础研究的原始创新能力，持续产出重大原创性、颠覆性科技成果。”[4]因此，尽早识别突破性科学创新，是决策管理部门有效配置科技资源、合理规划学科发展和前瞻布局产业方向的重要前提，亦是学术界和产业界共同关注的焦点。

农业是保障我国粮食安全的核心基石。随着全球迈入智能化、跨界融合的新形态，农业智能装备技术日新月异。2025年8月，《国务院关于深入实施“人工智能+”行动的意见》将农业机器人列为“农业数智化转型升级”的核心装备，强调其是推动农业现代化、实现“藏粮于技”的关键支撑。因此，基于海量的科技文献，结合深度学习、自然语言处理技术，准确识别农业

收稿日期：2025-07-27
*本研究得到北京市农林科学院科技创新能力建设专项“智库型农业情报研究与服务能力提升”（编号：KJCX20230208）、北京市农林科学院科技创新能力建设专项“面向科研管理的情报研究与服务能力提升”（编号：KJCX20230210）资助。

机器人领域的突破性科学创新方向，对于及时突破领域技术方法的卡点、推动重大原始创新具有十分重要的意义。

本文从突破性科学创新特征入手，综合考虑论文属性和关系网络特性，结合深度学习、自然语言处理等算法，提出基于特征测度和SciBERT模型的突破性科学创新识别方法，用以识别前沿领域的突破性科学创新研究主题，并选取农业机器人领域进行实证分析，为相关领域开展科学决策和科技创新研究提供支撑。

1　相关研究

1.1　突破性科学创新的概念与特征

突破性创新概念的提出源自熊彼特的“创造性的破坏”理论[5]。突破性创新常用来表征科技研究中的重大发现、技术改进、产品性能及市场价值的提升，是一个范围相对较广的概念[6]。国内外突破性创新研究主要从技术、科学、市场3个不同视角展开，涉及基础研究、技术研发和市场应用等多个阶段，其特征较为复杂。申姝婧等[7]认为，突破性创新涉及整个创新链，涵盖从突破性的新思想产生，到新知识、新技术在市场领域应用、更新迭代与推广，以及该技术创新破坏产业或市场格局的过程。其中，突破性科学创新聚焦于基础研究中的重大创新发现，通常会引发科学范式变革，或改变现有领域知识结构，产生重大的影响力[5]。

通过对国内外文献的梳理可知，无论是突破性技术创新还是突破性科学创新，大多包含学科交叉性[8]、新颖性[9]、前瞻性[10]、不连续性[11]、长期性[12]、高风险性[13]、高影响力[14]、随机性[15]特征。在针对突破性科学创新的研究中，马亚雪等[6]认为突破性科学创新的内涵包括3个方面：能够引发科学范式变革、影响范围广、能够改变现有领域知识结构。王娜[16]认为突破性科学创新具有变革性、高度领域颠覆性和革命性，是开创的新研究方向，往往引起创新簇或知识生产的连锁反应，从而为现有的科学体系带来巨大变革。

基于前人研究，笔者认为突破性科学创新应具有一定的内容变革性，通常能够改变领域知识的结构。然而目前研究尚未对该特征进行客观测度与表示，因此，本文尝试将内容变革性这一特性进行指标量化，以此表征知识内容的巨大变化。同时，结合可量化的特征，即新颖性、内容变革性、影响力和突变性，对突破性科学创新进行识别。

1.2　突破性科学创新主题识别方法

突破性科学创新主题识别的方法主要包括定性分析和定量分析两类。定性分析主要依靠专家智慧与经验判断；定量分析通常包括科学计量法、网络分析法和文本挖掘法，是基于文献自身或文献的链接关系，对突破性创新特征进行量化，最终求解研究主题的方法。

（1）科学计量法。通过计量方法对数据特征进行数理统计分析，如发表时间、被引频次等。Arts等[17]利用不同技术类别的专利引用数量来表征学科交叉性；Wang等[18]借助主题包含突破词汇的引文数量来表征主题的影响力。该方法仅基于外部属性，未能表达文本内容及结构等语义信息。

（2）网络分析法。利用引文关系或共现关系构建网络，通过网络属性表征论文特性。王娜[16]根据引用知识，构建了融合跨学科性、一致性、年龄特征、主题突变性的跨学科突破性创新识别模型；张金柱等[19]利用引文词汇和学科差异度来体现突变程度以识别突破性创新。然而，基于引文关系的计算存在一定时滞性。共现网络方面，罗瑞[20]基于主题词共现网络，从网络动态发展、重要主题识别和重要节点识别等维度对科学突破展开识别；刘亚辉等[21]提出可以通过有效捕捉弱信号对突破性创新进行早期识别。网络分析法与文本挖掘相结合可以兼顾文本内容和主题结构，提升主题识别的准确性。

（3）文本挖掘法。利用自然语言处理、机器学习方法对文本内容进行挖掘。陈虹枢等[3]利用LDA-Word2vec抽取主题，从新颖性、突变性、影响力和学科交叉性维度进行测度，开展突破性创新识别研究。随着机器学习、深度学习技术的迅速发展，BERT预训练模型在文本挖掘领域被广泛应用，可敏锐捕捉输入序列的全局依赖关系，具有能够学习文本更深层次语义表示的优势。目前多数研究基于BERT的改进模型，或与LDA、K-means结合的主题聚类方法，能够增强结果的可解释性[22]。SciBERT是在BERT基础上使用生物医学以及计算机科学领域共计114万篇科技论文的全文数据进行预训练的优化模型，能更好地理解和表示科学领域的专业术语和概念，专注于科学文本的理解和表示[23]，目

前多用于命名实体识别[24]、语步识别[25]、文本表示[26]等研究。例如，黄怡等[27]针对专利数据，将PaECTER和SciBERT模型用于技术主题的对比分析研究。然而，SciBERT模型还较少应用于突破性科学创新主题识别，未能验证和发挥其捕捉科学文本语义内容的优势，基于SciBERT模型的主题识别为本文提供了思路。

综上所述，众多学者针对突破性创新展开了大量有价值的研究，但仍存在不足：①突破性科学创新的测度特征和维度较为局限、不够全面；②识别结果的可解释性不佳，缺少能够深度挖掘和表示上下文语义信息的结合多种机器学习算法的主题提取方法，针对大规模科学文本训练的模型应用更是少见。突破性创新依赖多学科的交互作用与知识重组，其中跨界知识重组对突破性创新的影响最大，往往产生更高程度的创新[2,28]，而科技论文是作者表达其完整科学思想和逻辑的最小单元，也是体现知识重组的具有上下文语义的完整单元。因此，通过定量化方法对涉及跨界知识重组的高潜力论文进行遴选，将从数据源头上提高突破性科学创新识别结果的精准度，使识别层次更加细化。

本文以突破性科学创新特征为出发点，提出基于特征测度和SciBERT模型的突破性科学创新识别方法。首先，基于学科交叉指标遴选高潜力论文，并通过SciBERT预训练模型生成词向量，不仅能捕捉专业术语，极大限度保留文献的上下文语义关系，还能实现跨学科词汇的准确分辨，提高了主题识别的准确度。然后，结合K-means聚类方法和UMAP降维算法，解决了“高纬度稀疏、主题隐含”的问题，提高了主题挖掘的效率和稳定性，在保留主题结构基础上实现可视化。最后，构建融合新颖性、内容变革性、突变性和影响力的多维特征突破性科学创新指标体系，弥补当前研究特征维度不够全面的不足。

2 研究设计

研究框架如图1所示。首先，借助Web of Science（WOS）和JournaI Citation Reports（JCR）平台获取论文数据和学科信息，通过学科交叉特征指标筛选具有潜力的论文数据；其次，通过主题模型挖掘语义主题；然后，综合多维特征计算主题的突破性创新潜力；最终，根据得分确定突破性科学创新。

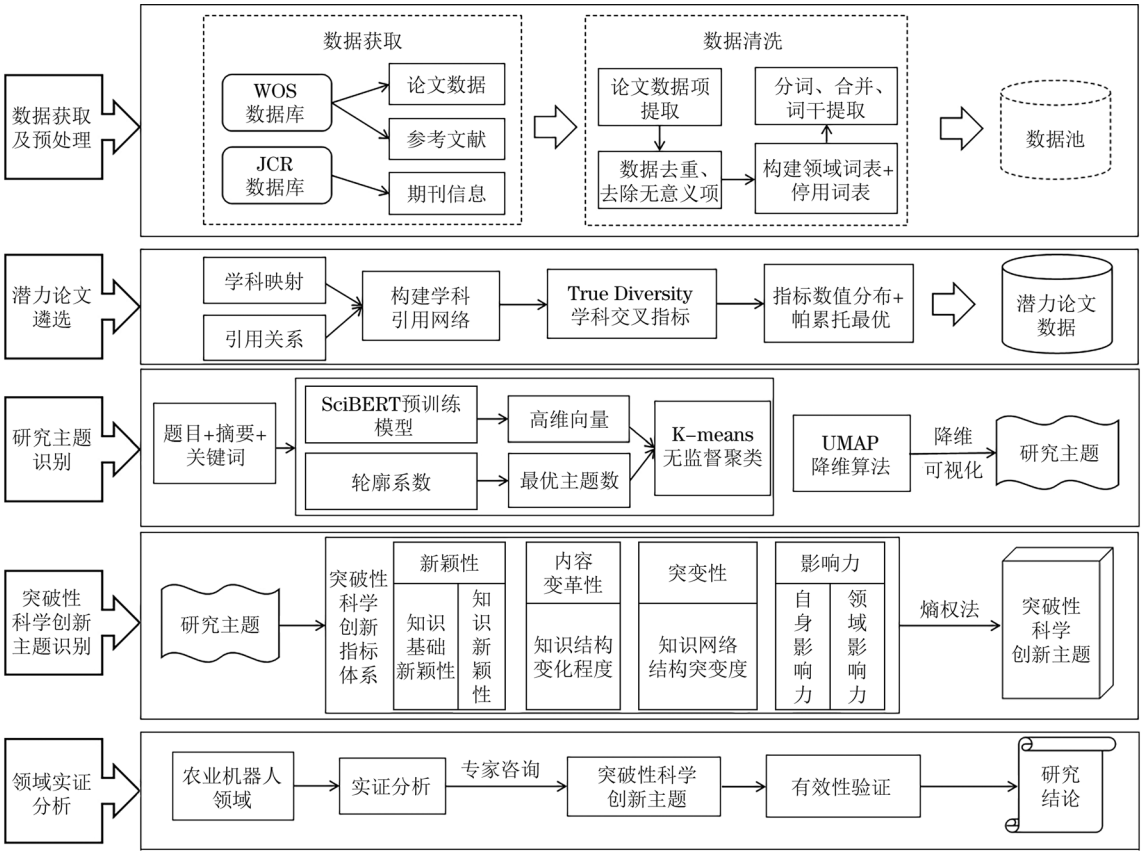

图 1 研究框架

2.1　跨学科潜力论文遴选

论文的跨学科性测度有多种角度，包括学科多样性、学科凝聚性和学科平衡性。Zhang等[29]提出了综合测度指标True Diversity（TD），融合了以上3个测量角度，同时结果具有良好区分度，受到了同行的高度关注与认可。其计算方法如式（1）所示。

$$I_{\mathrm{TD}}=\frac{1}{\sum\left(1-S_{ij}\right)P_xP_y} \tag{1}$$

式中：I_{TD}表示文献的TD值；S_{ij}表示文献i和文献j的学科相似性；P_x和P_y分别表示学科x和学科y的参考文献数量与所有参考文献数量的比。TD值越大，表明目标文献集的学科交叉程度越高，进一步反映文献主题的学科交叉程度高。对主题下文献集的TD值进行加权平均后，得到主题的学科交叉指标。

2.2　基于SciBERT模型的主题识别

提出融合SciBERT、K-means和UMAP的主题识别方法，整合了语义嵌入技术、分布式语义表示及降维聚类算法，能够在高维语义空间中获得更清晰的科学文本的分群，解决传统主题识别方法语义关联弱、分群边界模糊等问题，提高科学主题可解释性。

（1）文本向量生成。利用SciBERT进行文本向量化，生成高维向量。SciBERT通过双向Transformer编码器捕捉摘要中词语、短语和句子的上下文依赖关系，生成能反映科学语义的嵌入。文本向量通过提升主题节点间语义相似度，提高主题识别的准确性。

（2）最优主题判定。轮廓系数可以衡量聚类的紧密度和分离度。借助轮廓系数判断最优主题数量，避免无监督聚类中主观设定主题数量导致的聚类效果偏差[22]。调用Python程序scikit-learn包对形成的文本向量进行操作，利用式（2）计算轮廓系数。

$$S(\boldsymbol{i})=\frac{b(\boldsymbol{i})-a(\boldsymbol{i})}{\max\left\{a(\boldsymbol{i}),b(\boldsymbol{i})\right\}} \tag{2}$$

式中：$a(\boldsymbol{i})$表示主题内不相似度，即向量$\boldsymbol{i}$与同主题内其他点不相似程度的平均值，体现紧密度；$b(\boldsymbol{i})$表示主题间不相似度，即向量$\boldsymbol{i}$与其他主题的平均不相似程度的最小值，体现分离度；$S(\boldsymbol{i})$表示轮廓系数。所有样本的$S(\boldsymbol{i})$均值为聚类结果的轮廓系数，取值范围为[−1，1]，值越大表明同一聚类内部越紧密，类簇之间越远，效果越好。

（3）主题聚类与提取。K-means是流行的无监督机器学习算法，通过迭代优化快速地将数据集中的对象划分为K个主题。利用K-means对向量进行无监督聚类，迭代优化后生成主题，并采用TF-IDF算法完成主题词提取。联合SciBERT模型使用，能够从高维语义空间中发现隐含主题结构，避免了传统关键词表面相似的匹配误区，同时提高计算效率。

（4）降维与可视化。UMAP是目前流行的使高维数据降维的方法，可以在降维过程中同时保留局部网络结构和全局网络结构，具有显著优势。借助Python调用umap-learn和matplotlib库，进行降维操作和可视化展示，可以在保留局部主题信息基础上，直观展示聚类子团体和全局团体关系。

2.3　突破性科学创新主题识别

2.3.1　指标体系构建

判别一个科学主题是否具有突破性创新潜力，关键在于该科学主题或主题的来源数据是否体现了突破性创新的基本特征。基于前人研究，综合考虑论文及引文属性、网络结构等多个方面因素，构建了融合新颖性、内容变革性、突变性和影响力4个维度和6个指标的突破性科学创新主题测度指标体系（见表1）。

表1　突破性科学创新主题测度指标体系

维度	指标	含义
新颖性	知识基础新颖性	体现研究主题知识基础（参考文献）的新旧程度
	知识新颖性	时间维度表征研究主题中知识的新旧程度
内容变革性	知识结构变化程度	体现研究主题文献结构的变革性
突变性	知识网络结构突变度	捕捉研究主题的突变点和科学技术的大幅度跳跃点
影响力	自身影响力	表征主题论文对领域其他主题或论文产生的影响
	领域影响力	表征研究主题在研究领域网络结构中的桥梁、连通、中心位置的复合影响力

（1）新颖性。新颖性是突破性科学创新的基础，文中主要体现为时间维度的“新”，包括知识基础新颖性和知识新颖性。

参考文献是科技论文的知识基础，论文的参考文献较晚发表，证明论文比较关注最新研究进展[22]，故选

取主题中近5年参考文献的占比来衡量该主题的知识基础新颖性，计算方法如式（3）所示。主题的知识新颖性以该主题中所有论文的平均出版年份衡量，计算方法如式（4）所示。

$$Y_j = \frac{R_{\text{near5}}}{R_{\text{total}}} \tag{3}$$

$$E_j = \sum_{n=1}^{N} \frac{T_n}{N} \tag{4}$$

式中：Y_j表示第j个主题的知识基础新颖性；R表示主题中的参考文献数量；E_j表示第j个主题的知识新颖性；T_n表示第n篇文献的出版年份；N表示主题中的文档数量。

（2）内容变革性。内容变革性主要表征知识结构的变化程度，是基于引文网络的测度指标。Bornmann等[30]指出颠覆性系列指标能够识别出被广泛认为具有变革性的文献。颠覆性指数最早由Funk等[31]提出，后经Wu等[32]整合拓展，可以通过前向引用和后续引用的组合衡量论文对既有知识的发展或颠覆程度。许佳琪等[33]将颠覆性指数作为技术替代性维度指标，识别脑科学领域的颠覆性主题。Wu等[34]利用修剪焦点文献引用级联方法改进了颠覆性指数，避免了庞大繁冗引文网络的计算。因此，借鉴Wu等[34]提出的公式，以主题论文集的平均颠覆性指数衡量内容的知识结构变化程度，颠覆性指数计算方法如式（5）所示。

$$D = \frac{f_S - f_D}{f_S + f_D} \tag{5}$$

式中：f_S为仅引用焦点论文而不引用其参考文献的次数，f_S>1；f_D为同时引用焦点论文及其参考文献的次数。

（3）突变性。在科学发展过程中，网络结构中跳跃式的断层处是主题突变的关键节点，是测度突变性的重要位置[35]。关于主题的突变检测有诸多算法，Kleinberg算法是突变词检测的经典算法，通过动态规划方法来计算词汇在不同时段的最优状态序列，以识别突变词[36]。Small等[37]利用该算法和关键词演化来追踪领域发展趋势。本文选用相对成熟、应用广泛的Kleinberg突变词检测算法计算突变性维度的指标，计算方法如式（6）~式（8）所示。

$$F(w) = \frac{F_w^k}{\overline{F_w} + 2\sigma_w} \tag{6}$$

$$\overline{F_w} = \frac{1}{p}\sum_{k=1}^{p} F_w^k \tag{7}$$

$$\sigma_w = \sqrt{\frac{1}{p}\sum_{k=1}^{p}\left(F_w^k - \overline{F_w}\right)^2} \tag{8}$$

式中：$F(w)$表示词w的突变性值；F_w^k表示词w在时间段k的出现频次；$\overline{F_w}$表示词w在整个研究区间内出现频次的平均值；σ_w表示词w在整个研究区间内出现频次的标准差。

（4）影响力。突破性科学创新的影响力可以分为自身影响力和领域影响力。自身影响力用主题下论文集的被引频次来度量，表示论文所构成主题形成的绝对影响力，计算方法如式（9）所示。

$$I_j = \sum_{n=1}^{N} f_n \tag{9}$$

式中：I_j表示主题自身影响力；f_n表示该主题下第n篇文献的被引频次。

领域影响力用来测度主题在所形成网络中的地位。在网络分析中，网络中心性常被用于度量节点在网络中的影响力，相关指标包括接近中心性、中介中心性以及度中心性等[38]，以度中心性、中介中心性和接近中心性的等权叠加计算得到领域影响力。

度中心性为社交网络中与一个节点直接连接的节点数量的比例，能够反映节点的直接影响力和活跃度，计算方法如式（10）所示。

$$C_D^v = \frac{N_{\text{degree}}^v}{N-1} \tag{10}$$

式中：C_D^v是节点v的度中心性；N_{degree}^v是节点v的度，即与之相连的节点数；N是网络中的总节点数。

接近中心性为一个节点到其他所有节点的最短路径长度之和的倒数，能够反映节点在网络中的位置，以及信息从该节点传播到其他节点的速度，标准化计算方法如式（11）所示。

$$C_C^v = \frac{N-1}{\sum_{u=1}^{N-1} d(v,u)} \tag{11}$$

式中：C_C^v是节点v的接近中心性；$d(v,u)$是节点v到节点u的最短路径长度。

中介中心性为网络中经过某个节点的最短路径数占最短路径总数的比例，该数值高，代表该节点相较于节点两端的非相连节点拥有着更强的控制信息和利用资源的能力，计算方法如式（12）所示。

$$C_B^v = \sum_{s \neq v \neq t} \frac{g_{st}^v}{g_{st}} \tag{12}$$

式中：C_B^v是节点v的中介中心性；s和t是与节点v连接的其他节点；g_{st}^v是连接s和t，且经过v的最短路径的数量；g_{st}是连接s和t的最短路径的数量。

2.3.2　熵权法综合评价

动态权重分配方法可以避免主观性，还可以增强指标体系的科学性与适应性。熵权法是目前流行的方法之一，通过衡量每个指标所包含的信息量设置权重，信息量越大，权重越高。计算公式参考文献[39]。

3　实证分析

3.1　数据收集与预处理

以WOS平台和InCites平台为数据源获取数据，在InCites平台选择农业科学和工程学，联合领域专家对“农业机器人”进行主题解构，形成检索式：TS=（Agricultural robot* or Farming robot* or Planting robot* or Picking robot* or Weeding robot* or Harvesting robot* or Fertilization robot* or Sorting robot* or Milking robot* or Agricultural drones* or Agricultural UAV* or Unmanned crop protection machine* or Plant protection drone* or Plant protection UAV*）。文献类型设置为Article和Review，时间范围为2010—2024年，共得到检索记录3 750条，经人工筛选，得到3 517篇论文。下载论文的题录信息与参考文献、施引文献记录。

3.2　潜力论文筛选

（1）学科交叉指标计算。首先进行学科映射，提取文献、参考文献的学科类别和期刊字段，根据JCR提供的期刊名称对照表，形成“文献-学科”“参考文献-学科”映射对，利用德温特数据分析工具生成学科共现矩阵。根据式（1），编程计算学科余弦相似度、学科出现的概率以及TD指标值。最终得到论文学科交叉指标测度得分，Top10如表2所示。

（2）潜力论文筛选。统计TD指标数值分布并绘制直方图，得分区间统计见图2。该图符合帕累托分布的长尾特性，约82%的数集中在1.0~<11.5范围内。因此，以分值大于等于11.5的397篇论文为突破性科学创新的潜力论文数据集。

3.3　主题识别

以397篇论文为对象，提取论文的题目、摘要和关键词作为语料库，进行停用词去除、同义词合并等文本清

表2　论文学科交叉指标测度得分Top10

WOS号	标题	题录信息	TD值
WOS：001075215700001	Perceptual soft end-effectors for future unmanned agriculture	Ye WK，2023，SENSORS-BASEL，V23，P	21.77
WOS：000896621500001	Challenges for future robotic sorters of mixed industrial waste：a survey	Kiyokawa T，2024，IEEE T AUTOM SCI ENG，V21，P1023	19.06
WOS：000883359400001	A soft gripper design for apple harvesting with force feedback and fruit slip detection	Chen KW，2022，AGRICULTURE-BASEL，V12，P	18.54
WOS：001090685300001	Developments and challenges of miniature piezoelectric robots：a review	Li J，2023，ADV SCI，V10，P	18.24
WOS：001127798300001	Fluid-driven and smart material-driven research for soft body robots	Xu SM，2023，PROG NAT SCI-MATER，V33，P371	17.81
WOS：001062309000001	Classification-design-optimization integrated picking robots：a review	Xiang JY，2024，J INTELL MANUF，V35，P2979	17.57
WOS：000842447800002	A review of the design of load-carrying exoskeletons	Liang JJY，2022，SCI CHINA TECHNOL SC，V65，P2051	17.14
WOS：001154097400001	Accessorizing quadrupedal robots with wearable electronics	Kim MS，2024，ADV INTELL SYST-GER，V6，P	16.96
WOS：001061173700001	Tactile-sensing technologies：trends，challenges and outlook in agri-food manipulation	Mandil W，2023，SENSORS-BASEL，V23，P	16.90
WOS：001241298700001	Collaborative orchard pesticide spraying routing problem with multi-vehicles supported multi-UAVs	Xu JH，2024，J CLEAN PROD，V458，P	16.68

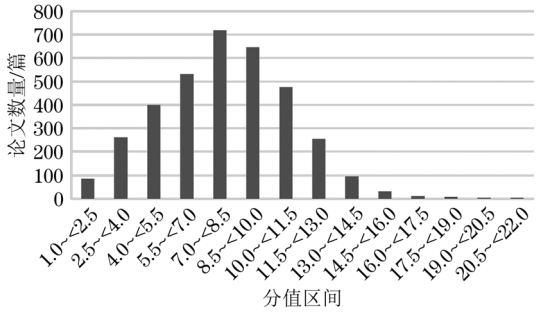

图2　论文学科交叉指标得分区间统计

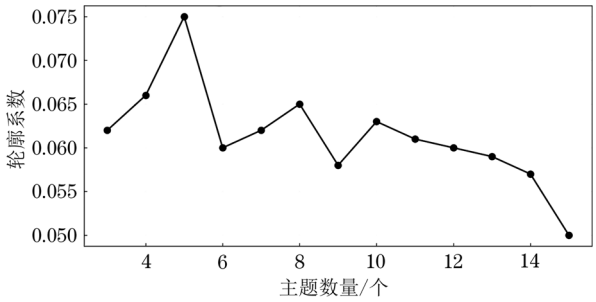

图3　主题轮廓系数计算结果

洗操作；利用SciBERT模型（allenai/scibert_scivocab_uncased）进行文本向量化，为每篇论文生成向量表示，得到一个397×768维的矩阵，每行代表一篇论文的向量；通过调用scikit-learn库计算轮廓系数（见图3），通过最高分值对应主题数是5，判断最佳主题数量为5个；设置主题数量K=5，利用K-means算法进行聚类，结合UMAP算法进行降维操作与聚类可视化，以便观察聚类效果（见图4），每个节点是文献语义向量在低维空间的投影，节点数量即文献的数量；借助TF-IDF算法，对每个主题的语料进行特征词提取以表征主题，主题及核心特征词如表3所示。

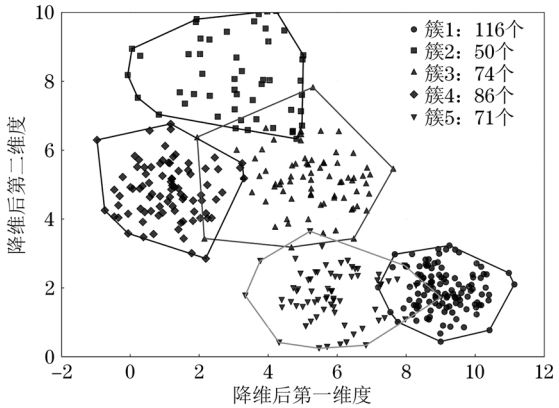

图4　最优主题聚类结果可视化

表3　主题及核心特征词

主题名称	特征词
多模态感知与自适应智能决策（簇1）	SLAM、Recognition and localization、Trajectory tracking、3D point cloud、Path planning、Mapping、Computer vision、Object detection、Image processing、CNN
机器人仿生驱动与异质能源耦合（簇2）	Energy efficiency、Shape memory alloy、Robotic fish、Adaptive、Solar energy、Regenerative braking technology、Soft materials、Soft actuators、Grippers、End-effectors
群体协同与云边端架构技术（簇3）	Human-robot interaction、Pneumatic systems、Inspection robot、Multimodal、YOLOv5、Collaborative homework、Weeding robot、Spray UAV、Dual arm collaborative
高精度轨迹控制与高动态响应（簇4）	Objective optimization、Sensors、Control、Robotic perception system、Transformer、YOLOv5、Retainer、Pick、Mechanical arm、Self monitoring
机器人自主学习与训练（簇5）	Deep learning、Strategy、UAV、Autonomous learning、Learning strategy pool、Precision farming、Neural network、Supervised learning、Particle swarm algorithm、State recognition、Particle swarm optimization algorithm、Spectral diagnostic model

3.4　突破性科学创新主题识别

（1）构建共现网络。共现分析是对文献特征信息共现现象的定量研究方法，用来揭示文献信息中内容关联的隐含知识[40]。利用Python提取K-means算法识别的5个主题结果及主题下的论文信息，构建局部引文网络、共词网络、主题共现网络。步骤如下：①根据论文及其参考文献和施引文献，调用networkx库构建主题间引文网络，计算文章颠覆性指数，将论文集的平均颠覆性指数作为主题的颠覆性指数值；②获取论文的时间信息、“论文-主题”矩阵、主题词分布，利用Python计算突变性；③分别针对5个主题子网络，调用Python的networkx库计算每个主题的中心性指标。

（2）突破性科学创新主题识别。基于共现网络，调用Python程序计算农业机器人领域每个研究主题的新颖性、内容变革性、突变性和影响力4个维度的6个指标值，利用熵权法进行逐级赋权，按照极大极小值数据标准化，计算得到5个主题的权重与总分数（见表4）。结

表4　突破性科学创新主题得分计算结果

维度	权重	多模态感知与自适应智能决策	机器人仿生驱动与异质能源耦合	群体协同与云边端架构技术	高精度轨迹控制与高动态响应	机器人自主学习与训练
新颖性	0.246	0.307	0.434	0.419	0.385	0.382
内容变革性	0.202	0.620	0.476	0.433	0.513	0.343
突变性	0.334	0.345	0.531	0.950	0.959	0.236
影响力	0.218	1.851	1.356	1.916	3.264	0.919
总和	1.000	0.817	0.758	1.073	1.491	0.605

果中，机器人自主学习与训练的总分最低，且在突破性和影响力两个维度中，与组内最高值差距过大，分别是最高值的24.6%和28.2%，均小于30%，因此选取前4个主题为突破性科学创新主题。

3.5　突破性科学创新主题分析

将每个主题的论文信息进行整理，邀请智慧农业领域的专家共同解读，结果如下。

（1）多模态感知与自适应智能决策。传统农业机器人依赖单一传感器（如仅激光雷达/视觉）实现环境感知，存在信息维度缺失、目标识别误差大的缺陷。该主题通过多模态传感器融合（多光谱+激光雷达+3D视觉）与深度学习算法耦合，突破“环境感知-作物认知-作业决策”的单向线性逻辑，构建“多维度环境信息-作物生理状态-精准作业指令”的闭环认知体系。例如：Guo等[41]融合触觉与视觉数据提出轻量级MDETR网络，姿势检测准确率达96%，果蔬抓取的速度提高到29.6 ms级别；Lee等[42]将亚硝酸盐传感器嵌入玉米秆实现自主监测，使农田环境数据的获取从人工抽样转向实时连续感知，稻草检测率达94%。

传统路径规划与避障算法多基于静态地图/预设规则，在复杂农田环境（如作物倒伏、临时障碍物）中易出现决策滞后导致作业失效问题。该主题通过轻量化模型部署和边缘推理技术，实现针对环境实时变化的决策即时调整。例如：Ji等[43]改进超扭曲滑模与干扰观测器，使无人农业拖拉机在田间动态障碍物场景下的路径跟踪误差从米级降至厘米级，破解规模化农业生产中环境不确定性与决策滞后性的核心矛盾；Zhang等[44]基于自适应神经网络的联合收割机路径规划，让农机在坡地、垄作等复杂地形下的作业效率提升25%，直接推动农业生产降本增效。

（2）机器人仿生驱动与异质能源耦合。融合生物运动/结构原理，通过新材料（如柔性材料、形状记忆合金），结合生物启发控制策略，实现驱动系统从刚性执行到类生命体自适应的跃迁，避免了传统农业机器人依赖刚性电机/液压驱动而存在的能耗高、环境适应性弱、易损伤柔性作业对象的弊端。例如：Duan等[45]受捕蝇草启发，利用二维金属有机框架与石墨开发能耗低的快速响应仿生软体机器人，极大缩短抓取动作的响应时间，为轻量化、低功耗农业机器人提供全新驱动范式；Manduca等[46]采用仿生控制策略，模仿鱼类运动机制，使机器鱼在浑浊水域的机动性与适应性得到提升，拓宽农业机器人在池塘监测、水产养殖等场景的应用边界。此外，基于清洁能源如氧化石墨烯和低密度聚乙烯的高灵敏度仿生软体执行器、基于磁性液态金属液滴的软体可重构循环器更是影响着环保与可持续发展理念。

传统机器人依赖单一电池供电，存在续航时间短、补给依赖外部输入的卡点。该主题通过异质能源采集、耦合、智能分配，突破能源供给的外源性约束。例如：Katiyar等[47]设计的自适应形态软体机器人利用运动时的机械能自给自足，使无外部电源作业时长从2 h延长至8 h；Liu等[48]的可穿戴Terfenol-D磁刻度振动收割机将机械振动能转化为电能，为农田微型监测节点（如土壤墒情传感器）持续供电，从根本上改变农业机器人的能源供给逻辑。仿生驱动与能源耦合技术的交叉融合，催生出低功耗、长续航、环境友好的农业机器人新产品（如柔性果蔬采摘机器人、“光伏-储能”一体化巡检机器人）。

（3）群体协同与云边端架构技术。机器人群体协同作业是未来农场的必然趋势，群体协同与云边端架构共同构建了农业机器人的智能协同神经，“终端感知（终端）-实时决策（边缘）-全局优化（云端）”的云边端协同架构突破了设备孤岛、算力割裂、调度低效的困境。例如：Rysz等[49]开发的随机优化模型降低了果园多机协作场景的人机协作风险，同时作业效率大幅提升；García等[50]基于A*与协同进化算法的多机器人路径

规划，证明该架构能有效平衡局部响应速度与全局资源最优，为未来农场的机器人群体自主作业提供底层支撑。

群体协同可以融合蜂群算法、数字孪生、协同进化算法，实现任务、资源、环境的动态匹配，突破了传统多机协作依赖固定任务分配算法，难以应对农田动态环境（如作物生长差异、突发天气）的问题。例如：Wang等[51]的多目标教学学习优化算法，将“除草+喷雾”无人机的协作任务分配耗时从小时级压缩至分钟级，提高了能源利用率；Wang等[52]提出数字孪生系统，建立协作机器人运动学控制模型，通过虚拟环境预演优化多机协作参数，为复杂农业场景下的多机协同提供理论依据。云边端架构与群体协同技术的结合，推动农业机器人从分散作业向集群服务转型。

（4）高精度轨迹控制与高动态响应。高精度轨迹控制与高动态响应是机器人运动控制的两大核心能力，通过可重构模块化执行器与新型控制算法，可实现微米级操作、毫秒级响应。例如：Kodama等[53]设计的藤蔓状动力软夹持器基于欧拉定理实现仿生抓取，实现负载能力随着缠绕角度增加的效果，可夹持1 660 N的物体；Chellattoan等[54]的相变低压驱动大振幅软体执行器可以在低于50 V的电压下，在7 s内完成复杂动作，实现高动态响应，是提升机器人工作效率的关键技术之一。

该主题融合遗忘反馈控制、级联比例积分控制、多轴协同优化，实现“在线学习-动态调整-全局优化”的闭环控制，突破了传统轨迹规划算法依赖离线建模与预计算，难以应对实时干扰（如风力、作物倒伏）的问题。例如：Shrivastava等[55]的最优关节轨迹跟踪控制算法使空间机械臂的轨迹跟踪误差从传统的15°降至3°，且耗时减少；Blanck-Kahan等[56]级联比例积分控制器，结合深度神经网络，将五杆并联机器人的误差最小化，适应不同轨迹、不同环境任务场景。高精度轨迹控制与高动态响应技术的产业化应用，推动农业机器人从“可用”向“好用”进阶。

3.6 方法验证

参考郝雯柯等[57]提出的资料分析法验证主题识别结果的科学性。赵春江院士[58]在2024世界机器人大会上提出下一步技术应解决的问题：①复杂环境下感知不精准（对应多模态感知）；②动态环境下手、眼、脚、脑协调难（对应多模态感知与高动态响应）；③末端执行器操作不灵巧（对应高精度轨迹控制）。同时，提出多机协同、“边-端-云”协同将会是未来5~20年的发展趋势，将加入更多的智能农业机器人技术创新，如深度学习、人机共融、新材料（对应机器人仿生驱动与异质能源耦合）、触觉反馈（对应多模态感知与自适应智能决策）等[58]。郭胜君[59]指出智能机器人在农业领域的技术挑战与创新：复杂环境下的智能导航和避障（对应多模态感知与自适应智能决策、高精度轨迹控制）。中国农业大学信息与电气工程学院李道亮教授[60]指出，农业机器人智能感知（对应多模态感知）、自主作业路径规划与导航（对应高精度轨迹控制）是农业机器人发展的趋势。

可见，所得突破性科学创新主题与研究报告和科技论文中的结论具有较好的一致性，验证了所提的突破性科学创新主题识别方法具有可信度和有效性，可为及时发现领域创新研究方向提供依据。

4 结语

突破性科学创新是驱动人类认知边界持续拓展的核心动力。精准识别突破性科学创新，不仅能够促进学科领域知识融合、加速科学范式的迭代更新，更能为破解复杂科学领域难题开辟新路径。传统研究对于突破性科学创新特征的表征并不全面，且主题挖掘的语义表示不足，导致可解释性不高。本研究提出基于特征测度与SciBERT的突破性科学创新主题识别方法，创新之处有二：①通过融合深度学习与自然语言处理技术，提出SciBERT+K-means+UMAP的主题识别框架，发挥各算法优势，通过多模型协同分工，最终实现主题识别性能的系统性提升，包括显著增强了模型的上下文语义理解能力，提升主题识别准确率与效率，同时实现了主题结构的深度挖掘与可视化表征；②所构建的多特征指标体系避免了单一指标评价的片面性，从理论上增加了研究的科学性、系统性和全面性，丰富了突破性科学创新的理论研究，同时兼具可拓展性与实际应用价值。实证分析结果可为服务农业强国建设目标、支撑科技创新战略实施提供参考依据，在原始创新能力提升路径与关键领域突破方向上具有实践指导意义。

参考文献

[1] TUSHMAN M L，ANDERSON P. Technological discontinuities and organizational environments[J]. Administrative Science Quarterly，1986，31（3）：439-465.

[2] 冯南平，刘鹏飞. 面向关键核心技术攻关的突破性创新潜在主体识别研究[J/OL]. 现代情报，1-21[2025-05-17]. https://link.cnki.net/urlid/22.1182.G3.20250515.1129.002.

[3] 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.

[4] 新华社. 习近平：在全国科技大会、国家科学技术奖励大会、两院院士大会上的讲话[EB/OL]. [2025-06-24]. https://www.gov.cn/yaowen/liebiao/202406/content_6959120.htm.

[5] MCCRAW T K. Prophet of innovation[M]. Cambridge：Belknap Press of Harvard University Press，2010：254-256.

[6] 马亚雪，王秀，熊宇宸，等. 突破性科学创新研究：内涵、进展与展望[J/OL]. 现代情报：1-16[2025-02-06]. https://kns.cnki.net/kcms/detail/22.1182.G3.20250205.1643.008.html.

[7] 申姝婧，梁继文，李剑飞，等. 创新全过程视角下突破性创新方向早期识别理论拓展与实践框架研究[J]. 情报理论与实践，2025，48（9）：130-143.

[8] 张金柱，张晓林. 基于专利科学引文的突破性创新识别研究述评[J]. 情报学报，2016，35（9）：955-962.

[9] ZHOU K Z，YIM C K，TSE D K. The effects of strategic orientations on technology- and market-based breakthrough innovations[J]. Journal of Marketing，2005，69（2）：42-60.

[10] DAHLIN K B，BEHRENS D M. When is an invention really radical? [J]. Research Policy，2005，34（5）：717-737.

[11] VERYZER R W. Discontinuous innovation and the new product development process[J]. Journal of Product Innovation Management，1998，15（4）：304-321.

[12] BRONDONI S M. Innovation and imitation：corporate strategies for global competition[J]. Symphonya Emerging Issues in Management，2012：2.

[13] SEMADENI M，ANDERSON B S. The follower's dilemma：innovation and imitation in the professional services industry[J]. Academy of Management Journal，2010，53（5）：1175-1193.

[14] BABA Y，WALSH J P. Embeddedness，social epistemology and breakthrough innovation：the case of the development of statins[J]. Research Policy，2010，39（4）：511-522.

[15] KOSHLAND D E，JR. The Cha-Cha-Cha theory of scientific discovery[J]. Science，2007，317（5839）：761-762.

[16] 王娜. 基于知识流动理论的跨学科突破性创新识别模型构建[D]. 武汉：华中师范大学，2023.

[17] ARTS S，VEUGELERS R. The technological origins and novelty of breakthrough inventions[J]. SSRN Electronic Journal，2013：2230366.

[18] WANG X，YANG X M，DU J，et al. A deep learning approach for identifying biomedical breakthrough discoveries using context analysis[J]. Scientometrics，2021，126（7）：5531-5549.

[19] 张金柱，张晓林. 基于被引科学知识主题突变的突破性创新识别[J]. 现代图书情报技术，2016（S1）：42-50.

[20] 罗瑞. 基于熵值的领域科学突破主题的识别与预测[D]. 北京：中国科学院文献情报中心，2020.

[21] 刘亚辉，许海云. 突破性创新早期识别与弱信号分析综述[J]. 图书情报工作，2021，65（4）：89-101.

[22] 崔文波，张涛，许鑫. 情报学领域典型文本主题聚类算法比较研究[J]. 情报理论与实践，2025，48（7）：199-207.

[23] BELTAGY I，LO K，COHAN A. SciBERT：a pretrained language model for scientific text[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing（EMNLP-IJCNLP）. Stroudsburg：ACL，2019：3613-3618.

[24] 王玉龙，秦春秀，马续补，等. 融合语义增强和知识蒸馏的学术文献知识实体抽取方法研究[J]. 情报学报，2025，44（4）：438-451.

[25] 许钦亚，薛秋红，钱力，等. 融合ChatGPT数据增强的学术论文语步识别方法研究[J]. 图书情报工作，2024，68（17）：84-94.

[26] 冯小东，黄雨杭. 基于SciBERT的科研合作知识交叉测度及其对科研主体持续科研产出的因果效应[J]. 情报学报，2025，44（7）：869-891.

[27] 黄怡，隗玲，张凯. 基于PaECTER-BERTopic与大模型的专利技术主题识别及演化分析：以生成式人工智能领域为例[J]. 数字图书馆论坛，2025，21（2）：1-11.

[28] 杨俊浩，许海云，王超，等. 科学突破主题的学科交叉演化特征分析[J]. 信息资源管理学报，2024，14（4）：70-85.

[29] ZHANG L，ROUSSEAU R，GLÄNZEL W. Diversity of references as an indicator of the interdisciplinarity of journals：taking similarity between subject fields into account[J]. Journal of the Association for Information Science and Technology，2016，67（5）：1257-1265.

[30] 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.

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

[32] WU L F，WANG D S，EVANS J A. Large teams develop and small teams disrupt science and technology[J]. Nature，2019，566（7744）：378-382.

[33] 许佳琪，汪雪锋，陈虹枢，等. 跨领域颠覆性技术主题识别研究：以脑科学技术为例[J]. 图书情报工作，2024，68（15）：44-57.

[34] WU Q，YAN Z Y. Solo citations，duet citations，and prelude citations：new measures of the disruption of academic papers[EB/OL]. [2025-06-08]. https://arxiv.org/abs/1905.03461.

[35] DOSI G. Technological paradigms and technological trajectories：a suggested interpretation of the determinants and directions of technical change[J]. Research Policy，1982，11（3）：147-162.

[36] 李丽娜. 跨学科视角下突破性创新的测度研究[D]. 武汉：华中师范大学，2023.

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

[38] 姜鑫，王德庄，马海群. 社会网络分析方法在图书情报学科的应用研究[M]. 北京：知识产权出版社，2019：24-26.

[39] 韦景竹，张桐瑞，马芷晴. 基于科技报告的关键共性技术识别[J]. 情报杂志，2025，44（4）：153-159，199.

[40] 图书馆·情报与文献学名词审定委员会. 图书馆·情报与文献学名词[M]. 北京：科学出版社，2019.

[41] GUO C M，ZHU C H，LIU Y C，et al. End-to-end lightweight transformer-based neural network for grasp detection towards fruit robotic handling[J]. Computers and Electronics in Agriculture，2024，221：109014.

[42] LEE M，BERGER A，GURI D，et al. Towards autonomous crop monitoring：inserting sensors in cluttered environments[J]. IEEE Robotics and Automation Letters，2024，9（6）：5150-5157.

[43] JI X，DING S H，WEI X H，et al. Path tracking control of unmanned agricultural tractors via modified supertwisting sliding mode and disturbance observer[J]. IEEE/ASME Transactions on Mechatronics，2024，29（6）：4051-4062.

[44] ZHANG Y，WANG L，LIU Y. Adaptive neural network-based path tracking control for autonomous combine harvester with input saturation[J]. Industrial Robot，2021，48（4）：510-522.

[45] DUAN S S，WEI X，WENG M C，et al. Venus flytrap-inspired data-center-free fast-responsive soft robots enabled by 2D $Ni_3(HITP)_2$ MOF and graphite[J]. Advanced Materials，2024，36（28）：2313089.

[46] MANDUCA G，SANTAERA G，MIRAGLIA M，et al. A bioinspired control strategy ensures maneuverability and adaptability for dynamic environments in an underactuated robotic fish[J]. Journal of Intelligent & Robotic Systems，2024，110（2）：69.

[47] KATIYAR S A，LEE L Y，IIDA F，et al. Energy harvesting for robots with adaptive morphology[J]. Soft Robotics，2023，10（2）：365-379.

[48] LIU H F，ZHAO X X，LIU H K，et al. Magnetostrictive biomechanical energy harvester with a hybrid force amplifier[J]. International Journal of Mechanical Sciences，2022，233：107652.

[49] RYSZ M W，MEHTA S S. A risk-averse optimization approach to human-robot collaboration in robotic fruit harvesting[J]. Computers and Electronics in Agriculture，2021，182：106018.

[50] GARCÍA E，VILLAR J R，TAN Q，et al. An efficient multi-robot path planning solution using A* and coevolutionary algorithms[J]. Integrated Computer-Aided Engineering，2022，30（1）：41-52.

[51] WANG C H，PAN Q K，LI X P，et al. A multi-objective teaching-learning-based optimizer for a cooperative task allocation problem of weeding robots and spraying drones[J]. Swarm and Evolutionary Computation，2024，87：101565.

[52] WANG L H，SHU L，ZHOU H. Multi-robot collaborative flexible manufacturing and digital twin system design of circuit breakers[J]. Applied Sciences，2023，13（4）：2721.

[53] KODAMA H，IDE T，FENG Y H，et al. Vine-like，power soft gripper based on Euler's belt theory[J]. IEEE Robotics and Automation Letters，2024，9（4）：3108-3115.

[54] CHELLATTOAN R，YUDHANTO A，LUBINEAU G. Low-voltage-driven large-amplitude soft actuators based on phase transition[J]. Soft Robotics，2020，7（6）：688-699.

[55] SHRIVASTAVA A，DALLA V K，ABHISHEK K. Optimum joint and trajectory tracking control of multi-axes space manipulator based on amnesia feedback controller[J]. Proceedings of the Institution of Mechanical Engineers，Part C：Journal of Mechanical Engineering Science，2023，237（19）：4546-4563.

[56] BLANCK-KAHAN D，ORTIZ-CERVANTES G，MARTÍNEZ-GAMA V，et al. Neural-optimal tuning of a controller for a parallel robot[J]. Expert Systems with Applications，2024，236：121184.

[57] 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.

[58] 杨光. 中国工程院院士赵春江：农业机器人发展面临五大技术难点[N]. 中国信息化周报，2024-09-02.

[59] 郭胜君. 智能机器人在农业领域的应用与发展[J]. 河北农机，2024（10）：42-44.

[60] 李道亮. 我国农业机器人技术与发展趋势[J]. 人民论坛，2024（24）：70-74.

作者简介

齐世杰，女，硕士，助理研究员，研究方向：前沿技术识别、文本挖掘。
串丽敏，女，博士，副研究员，研究方向：智能知识服务技术。
张辉，男，硕士，助理研究员，研究方向：自然语言处理、技术识别。
姚茹，女，博士，助理研究员，研究方向：知识发现。
贾倩，女，硕士，助理研究员，研究方向：科技情报。
赵静娟，女，硕士，副研究员，通信作者，研究方向：科技情报，E-mail：zhaojj@agri.ac.cn。

Breakthrough Scientific Innovation Topic Identification Based on Feature Measure and SciBERT Model: A Case Study of Agricultural Robots Field

QI ShiJie　CHUAN LiMin　ZHANG Hui　YAO Ru　JIA Qian　ZHAO JingJuan
(Institute of Data Science and Agricultural Economics, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, P. R. China)

Abstract: Breakthrough scientific innovation is the core driving force behind technological progress and socio-economic development. Accurately identifying breakthrough scientific innovations is of great significance for opening up new tracks of scientific and technological innovation and achieving leapfrog development. Starting from the characteristics of breakthrough scientific innovations in scientific paper data, we first comprehensively consider the diversity, balance, and difference of papers, using the True Diversity measurement index and disciplinary co-occurrence network to measure and screen potential papers. Then, we employ the SciBERT+K-means+UMAP method for topic modeling and visualization. From the four dimensions of novelty, content transformation, influence, and abruptness, we construct a multi-feature breakthrough scientific innovation indicator system, integrating various text mining methods to identify breakthrough scientific innovation topics. Taking the field of agricultural robots as an example, four breakthrough scientific innovation topics are ultimately identified: multimodal perception and adaptive intelligent decision-making, bio-inspired actuation and heterogeneous energy coupling for robots, swarm collaboration and cloud-edge-end architectural technology, and high-precision trajectory control and high-dynamic response. The effectiveness and reliability of the model are verified. The research not only provides new ideas for identifying breakthrough scientific innovation topics, but also provides a reference basis for original innovation in the field of smart agriculture.

Keywords: Breakthrough Scientific Innovation; Topic Identification; Agricultural Robot; Machine Learning; Interdisciplinary Research; SciBERT; K-means

（责任编辑：王玮）