基于主题模型的领域新兴交叉主题识别研究*
——以作物智能育种为例

齐世杰　串丽敏　赵静娟　张辉　贾倩
（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：准确识别学科交叉前沿主题，有助于了解学科发展脉络，发掘领域重点发展方向，为未来创新性、突破性研究提供参考。提出一种识别新兴交叉主题的方法。首先，提出一种结合学科多样性和学科凝聚性的论文学科交叉性计算方法；其次，利用该方法筛选出具有高学科交叉性的论文，获得潜力论文数据集；再次，采用结合领域词典改进的LDA模型进行研究主题识别；最后，通过构建融合新颖性、突破性和影响力的多维度新兴主题测量模型，识别出新兴交叉主题。选择作物智能育种领域进行实证分析，识别出4个新兴交叉主题，通过资料分析法验证方法的有效性，对基于论文数据识别新兴交叉主题的方法研究与实践具有参考价值。

关键词：学科交叉研究；新兴主题识别；主题建模；作物智能育种
中图分类号：G353.1　DOI：10.3772/j.issn.1673-2286.2024.09.004
引文格式：齐世杰，串丽敏，赵静娟，等. 基于主题模型的领域新兴交叉主题识别研究：以作物智能育种为例[J]. 数字图书馆论坛，2024，20（9）：38-47.

学科交叉融合是解决当今世界复杂问题的重要途径之一，也是创新之源。它不仅能够催生新兴交叉研究方向，更是突破性技术的重要源头。准确探测领域新兴交叉主题，有助于从根源上厘清学科内在的发展机制与轨迹，及时追踪领域交叉创新趋势，进而捕捉新的领域研究增长点，辅助科研管理决策者预先布局突破性创新的方向，为发展新质生产力提供有价值的客观依据。

学者们已就如何从科技文本中探测各类研究主题开展大量研究，其中新兴主题和学科交叉主题是关注的热点，但现有研究多将研究主题的学科交叉性和新兴性分开独立研究，将新兴性与交叉性相结合的主题识别研究尚不多见。在早期，学科交叉是新兴主题形成的内在驱动力之一，更是探测学科生长点的重要手段[1-2]。因此，从学科交叉融合的视角对新兴主题进行探测，能够从根源上增加发现创新性研究的机会，挖掘突破性的创新方向和潜在高价值点。

种业科技创新催生农业新质生产力。全球种业科技已进入“生物技术+人工智能+大数据信息技术”的智能育种时代，通过人工智能决策系统可以设计最佳育种方案，进而定向、高效改良和培育作物新品种。而我国尚处于起步阶段，存在部分前沿和交叉领域基础研究和底盘技术的原始创新能力不足的问题。基于此，本文从学科交叉性角度入手，识别新兴交叉主题，并以智能育种领域为例论证方法的有效性和可行性，研究结果对于我国智能育种领域的发展有一定启发意义。

收稿日期：2024-05-28
*本研究得到北京市农林科学院科技创新能力建设专项“基于学科交叉的农业‘火花技术’早期探测方法与实证研究”（编号：KJCX20240313）、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。

1　相关研究

1.1　新兴主题的概念及特征

新兴主题由Matsumura等[3]在2002年提出，指某研究领域中由多个关键词或词组表示的新主题，代表极具发展潜力的研究方向或趋势。Rotolo等[4]将新兴主题的特征归纳为极强的新颖性、相对快速的增长、一致性、突出的影响、不确定性和模糊性。2018年，Wang[5]认为新兴主题是具有新颖性和一定连贯性、能产生较大科学影响力且发展速度相对较快的主题，主要特征包括新颖性、增长性、一致连贯性和科学影响力。Xu等[6]指出，新兴主题的特点主要体现在时间维度和创新维度上。

1.2　新兴主题识别方法

新兴主题识别主要包括主题探测和新兴特征测度两大重要内容，主题识别方法由基于引用关系向着基于文本内容的趋势发展。新兴主题识别方法主要包括以下3类。①基于引文网络的识别方法，即基于论文的直接引用关系、共被引关系和耦合关系进行主题识别。例如：Small[7]提出利用共被引关系识别新兴主题；Chen[8]将引文与词法分析结合，联合引文分析和爆破检测识别新兴主题。该方法关注论文之间的知识传承，但引用关系往往具有一定的时滞性，导致研究及时性有所欠缺。②基于文本挖掘的识别方法，即基于文本内容挖掘出语义关联的信息，从而发现主题，具体包括主题模型、SAO结构抽取、知识图谱等方法。其中主题模型应用最为广泛，LDA及其改进模型，如LDA2vec、动态LDA算法，取得了良好的效果。此外，一些研究还融合关键词的词频、文献与词关联关系进行主题识别，以提高主题建模的可解释性[9]。但该方法存在主题词太过泛化，与文本特有的学科领域契合度不高的问题。③融合引文网络和文本内容的识别方法。例如：白敬毅等[10]基于引文网络，提出新颖性、增长性、影响力等测度指标，利用LDA模型和多维尺度分析识别新兴主题；Xu等[11]利用动态影响模型提取主题结构及增长性和影响力等指标，使用多任务最小二乘支持向量机识别不同主题。以上方法可以消减时滞性带来的误差，但主题的学科领域契合度不足的问题仍有待解决，现有研究也尚未提及学科交叉性这一识别角度。

1.3　学科交叉测度方法

学科交叉测度指标大致可分为学科多样性指标和学科凝聚性指标。学科多样性指标测度角度包括学科丰富性、平衡性和差异性，测度方法包括信息熵、布里渊指数、学科集成化指数、跨领域引用指数、Rao-Stirling指数。学科多样性指标测度主要以论文及引用文献为基础，对所属期刊的学科类别进行分析，以Rao-Stirling指数最为常用，但其对于学科共现网络较为依赖，存在计算繁琐、更新困难、无评判准则等局限。学科凝聚性指标通常侧重于社会网络分析，常用指标包括网络密度和平均路径长度、学科凝聚度、中介中心性等。此外，也有学者将学科多样性和学科凝聚性指标融合，形成综合测算指标。例如：Rafols等[12]融合学科多样性和网络一致性测度了单篇论文的学科交叉性；陈赛君等[13]基于Stirling多样性和致密性构建了Φ指标。

综上可知，在论文学科交叉测度方面，融合学科多样性和学科凝聚性是一种有效方法。在新兴主题识别方面，融入学科交叉性识别新兴领域方向是一种新思路。因此，本文从学科交叉角度入手，先筛选高学科交叉性论文形成潜力论文数据集，再通过多特征指标和主题挖掘实现新兴主题识别，为明晰科技创新方向、加快发展新质生产力提供助力。

2　研究设计

首先，提出融合学科多样性和凝聚性的论文学科交叉综合测度指标TD_C指数，通过测算定量筛选出具有高学科交叉性的潜力论文；其次，基于领域词表优化的LDA模型对潜力论文集进行主题识别；再次，结合时间特征、论文自身特征和网络特征，从新颖性、突发性和影响力3个维度构建新兴主题识别模型；最后，选取智能育种领域进行实证研究，识别领域中的新兴交叉主题。研究框架如图1所示。

2.1　论文学科共现网络构建

论文和参考文献的学科分类是建立学科共现矩阵的前提。本研究中论文数据来自科睿唯安Web of Science数据库，Web of Science中JCR（Journal Citation Reports）分类体系的认可度高且应用广泛，因此以其中的256个

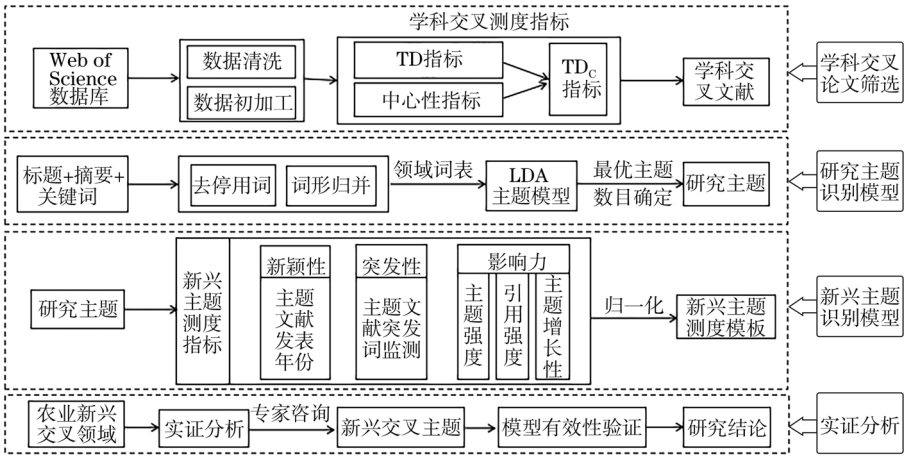

图 1　新兴交叉主题识别研究框架

学科类别为分类依据，将论文映射至各个学科。

首先，通过Web of Science数据库下载论文的字段信息，提取学科、参考文献等字段建立数据表；其次，对JCR中的期刊和其所属学科类别信息，以及期刊缩写与期刊全称信息进行提取，构建期刊-学科类别对照表；再次，基于该对照表，将论文的参考文献根据其所在期刊映射至相应的学科类别；最后，根据参考文献-学科映射结果，建立论文-参考文献学科共现网络。

2.2　学科交叉论文提取

识别高学科交叉性论文是新兴交叉主题识别的前提，当前学科交叉测度内容主要包括学科多样性和凝聚性。Rao-Stirling指数是学科多样性测度中较为成熟的指标。Zhang等[14]在其基础上进一步改进，提出TD（True-Diversity）指标，其学科间区分度较高，受到学界的广泛关注。学科凝聚性方面，中介中心性指标应用较为广泛，在学科交叉测度中具有良好的效果[15]。

综合考虑学科多样性和学科凝聚性，参考TD指标和中介中心性指标，以单篇论文为计算单位，通过两指标乘积的方式计算得到论文的学科交叉综合指数TD_C，据此对论文的学科交叉性进行全面评估。TD、中介中心性、TD_C指标的计算方法如式（1）~式（4）所示。

$$D=\frac{1}{1-\sum_{i,j}p_i p_j d_{ij}} \tag{1}$$

$$C_i=\sum\nolimits^{1}_{s\neq i\neq t}\frac{n_{st}^{(i)}}{g_{st}} \tag{2}$$

$$C_a=\sum\nolimits^{1}_{n}\frac{C_i}{n} \tag{3}$$

$$\delta_{TD_C}=D\cdot C_a=\frac{1}{1-\sum_{i,j}p_i p_j d_{ij}}\cdot\sum\nolimits^{1}_{n}\frac{C_i}{n} \tag{4}$$

式中：D表示TD指标，i和j表示两个不同的学科，p_i和p_j分别表示i和j学科的概率，d_{ij}表示两个学科在学科网络中的相对距离；C_i表示学科i的中介中心性，s和t表示与i连接的其他学科，$n_{st}^{(i)}$表示连接学科s和学科t且经过学科i的最短路径的数量，g_{st}表示连接学科s和学科t的最短路径的数量；C_a表示论文a的中介中心性，n表示所属的学科数量，论文的中介中心性是学科中介中心性的均值。TD指标值越大，学科之间的相似性越大。中介中心性越大，代表节点在学科网络结构中承担的桥梁作用越大。

基于论文的学科映射，可得到学科的概率p_i和p_j，借助学科共现网络，利用余弦相似度计算两个学科在学科网络中的相对距离d_{ij}。将学科矩阵导入UCINET软件，计算得每篇论文的中介中心性。将每篇论文的中介中心性指标和TD指标值相乘得到综合指数TD_C。

2.3　基于LDA的主题建模

LDA模型由Blei等[16]于2003年提出，是一种无监督机器学习的文本挖掘方法。该模型能够对论文进行内容层面的语义分析，但存在各个主题词之间的语义关联性较小、可解读性不强的问题[17]。通过加入智慧农业领域词表规范主题词作为用户词典，扩充领域词汇，辅助主题解释，提高LDA主题词准确度。

主题数目确定是文本主题抽取的关键步骤，关系到结果的好坏。困惑度（Perplexity）和一致性（Coherence）是目前最优主题数目评估方法中效果较

好的指标[18]。困惑度代表了主题归属的不确定性，值越小效果越好；一致性表征了主题的连贯性，值越大代表结果越好[19]。综合困惑度和一致性确定主题数目，以提高准确性，同时能够避免主题数目过多造成的过拟合现象。困惑度和一致性计算方法如式（5）~式（7）所示。

$$\delta_{\text{Perplexity}} = \exp\left(-\frac{\sum \log p(w)}{\sum_{d-1}^{M} N_d}\right) \tag{5}$$

$$p(w) = p(z|d) \cdot p(w|z) \tag{6}$$

$$\delta_{\text{Coherence}} = \sum_{(v_i, v_j) \in V} \text{score}(v_i, v_j, \varepsilon) \tag{7}$$

式中：M表示文档数量，N表示数据集的词语数量之和，$p(w)$表示数据集中每个单词出现的概率；$p(z|d)$表示文档中每个主题出现的概率，$p(w|z)$表示数据集中每个单词在每个主题下出现的概率；score表示某主题中词语之间的一致性得分，v_i和v_j表示主题中的词语，ε表示平滑系数。

在LDA模型主题建模过程中，通过设定最优主题数目、迭代次数等参数训练模型，会生成主题-文档矩阵，每篇论文对应一个主题概率分布，选择其中概率最高的主题作为每篇论文的主题，以此判定论文的主题归属。

2.4　新兴主题识别

围绕新兴主题“新”和“兴”两个主要特征，以主题的时间属性、引用属性和关键词共现属性为特征要素，构建融合新颖性、突发性和影响力3个维度的测度模型。

（1）新兴主题测度指标。针对新兴主题特征，构建三维新兴主题测度框架体系。①主题新颖性。以论文的发表年份表征时间维度上的“新”，主题下论文的平均发表时间越晚，新颖性越强[20]，计算方法为主题下所有论文发表年份之和除以论文数量。②主题突发性。打破研究主题现有状态的突增现象是判断研究主题新兴与否的直观依据[1]。Kleinberg[21]提出突发词监测算法来揭示新兴主题。该算法能够探测出短时间内频率急剧上升的突发词，由此确定某个领域的新兴趋势和潜在热点，受到广泛应用。选用该突发词监测算法作为主题突发性维度的指标。③主题影响力。从主题强度、引用强度和增长性3个方面综合度量主题影响力。其中：主题强度由某主题下研究论文数量占比表示，能够直观体现主题的研究热度；引用强度由主题下所有论文的被引频次的均值表示，体现某研究主题在同行间的影响力；增长性指标是主题下所有论文数量的平均年增长率，体现了一个主题随时间的推移，受到的关注度的变化情况。新兴主题测度指标及含义见表1。

表1　新兴主题测度指标及含义

维度	指标计算方法	含义
新颖性	某主题下论文的平均发表年份	体现研究主题时间维度的“新”
突发性	主题词中突发词概率值之和	捕捉具有活跃性、潜在热度的研究主题
影响力	主题强度，即某主题下论文数量与所有主题论文数量之比	表征研究主题的热度
	引用强度，即某主题下论文被引频次之和与论文数量之比	表征研究主题在领域中的影响力
	增长性，即主题下所有论文数量的平均年增长率	表征研究主题被同行关注的趋势

（2）新兴主题测度模型。根据指标的量纲和取值范围，选择不同数据处理方式进行指标归一化。新颖性指标计算中将发表年份（1980—2030年）分散到0~1范围内处理，突发性、影响力及二级指标均采用最大-最小值法进行归一化。采用客观赋权法和线性叠加的方法，构建主题测度模型。

3　实证分析

3.1　数据采集与处理

通过Web of Science和InCites平台进行论文数据采集。InCites平台是基于Web of Science核心合集建立的科研分析平台，提供多个国家的多种学科分类体系，有助于科研人员的多维度分析与数据筛选。

借助专家咨询，对智能育种进行主题拆解与词汇整合，组建检索式如下：TS=（（smart agriculture OR digital agriculture OR digitization OR intelligence OR big data OR artificial intelligence OR machine learning OR deep learning OR neural network OR Internet of Things OR cloud compute OR information technology OR genome information OR image analysis OR AI）and breed）。利用InCites平台，选择Crop Science学科类别，文献类型限定为Article和Review，时间范围是2018年1月1日—2023年12月31日，检索时间是2024年1月15日。共得到文献检索记录6 752条，经过人工筛选，最终确定6 315篇论文为数据源，下载年份、标题、摘要、关键词等字段，同时下载其参考文献用于学科交叉性计算。

3.2 智能育种领域学科交叉论文提取

（1）指标计算。提取6 315篇论文与参考文献的期刊全称，按照JCR形成参考文献-学科映射对和论文-参考文献学科映射对，利用德温特数据分析工具（Derwent Data Analyzer，DDA）生成学科共现矩阵。结合Leydesdorff等[22]提出的全局学科距离矩阵，编程计算学科余弦相似度、学科出现的概率以及TD指标值，并将学科共现矩阵导入UCINET，计算论文的中介中心性，最终得到每篇论文的TD_C指数。

（2）筛选高学科交叉性论文。计算6 315篇论文的TD_C指数（见表2），统计并绘制分值分布直方图（见图2）。结合直方图判读原文可知，大多数论文的TD_C指数集中于20~60，大于60的论文学科交叉性较高。因此，提取数值大于60的808篇论文作为新兴主题识别的潜力论文数据集。

表2 智能育种领域论文学科交叉性计算结果（部分）

序号	入藏号	论文信息	标题	TD_C指数
1	WOS：000448876600006	Wang YQ，2020，AGRONOMY-BASEL，V10，P	Analysis on efficiency and influencing factors of new soybean producing farms	204.36
2	WOS：000754557400024	Vrahatis AG，2021，ADV EXP MED BIOL，V1338，P199	Mating type distribution，genetic diversity and population structure of *Ascochyta rabiei*，the cause of *Ascochyta* blight of chickpea in western Iran	146.51
3	WOS：000754488200008	Krokidis MG，2021，ADV EXP MED BIOL，V1339，P	PhenoFlex—an integrated model to predict spring phenology in temperate fruit trees	143.16
4	WOS：000754557400003	Vergis S，2021，ADV EXP MED BIOL，V1338，P13	Large-scale RNAseq analysis reveals new insights into the key genes and regulatory networks of anthocyanin biosynthesis during development and stress in cassava	139.47
5	WOS：000454396400029	Spanoghe MC，2020，GENET RESOUR CROP EV，V67，P947	Genetic patterns recognition in crop species using self-organizing map：the example of the highly heterozygous autotetraploid potato（*Solanum tuberosum* L.）	134.66

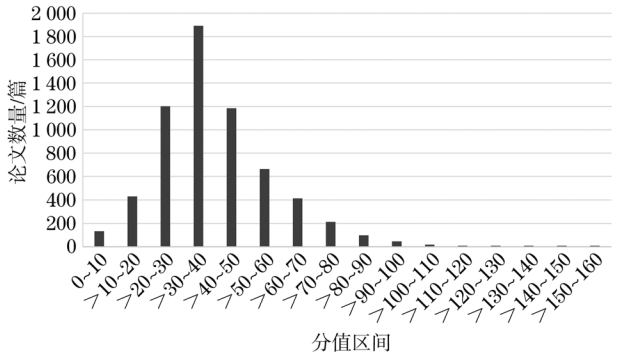

图2 论文学科交叉性分值分布直方图

3.3 智能育种领域基于LDA模型的主题识别

（1）数据预处理。针对潜力论文数据集，抽取数据集中每篇论文的标题、摘要和关键词字段作为语料库，将专家共同构建的智能育种领域的关键词表定义为用户词典，同时加入停用词表，利用Python jieba软件包对语料进行分词，进行去除停用词、词形归并和词干提取等数据预处理，通过扩充用户词典与停用词表优化分词效果，提高主题词提取的准确率。

（2）主题识别。借助Python的gensim库，对语料进行LDA模型训练，设置迭代20次，获取主题-文档矩阵、主题词-文档矩阵，并基于pyLDAvis库开发工具对主题进行可视化展示。

根据智能育种领域的数据规模，设定候选主题数目K范围为2~10，步长为1，迭代计算主题的一致性和困惑度，绘制曲线图（见图3~图4）。可见，主题数目为5个时，一致性达到第一个峰值；当主题数目为9个时，一致性数值达最高点，但困惑度降低。结合主题分析结果，K=9时，出现过拟合现象。通过对一致性和困惑度的综合判断，发现主题数目为5个时，主题识别的效果较好，因此，确定最佳主题数目为5个。

依据构建的主题-文档矩阵、主题词-文档矩阵，通过Python程序调取5个主题及所对应的论文、主题词，选取每个主题下概率较高的主题词作为主题的核心特征词，统计整理各主题的TOP 15核心特征词，如表3所示。

3.4 智能育种领域新兴交叉主题识别

3.4.1 新兴主题识别

根据提出的多维度新兴主题测度指标，计算智能育种领域每个交叉研究主题的新颖性、突发性和影响

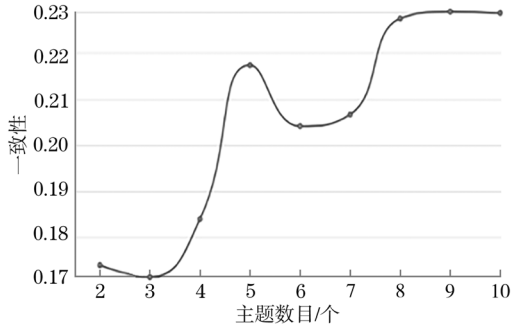

图3　一致性分布曲线

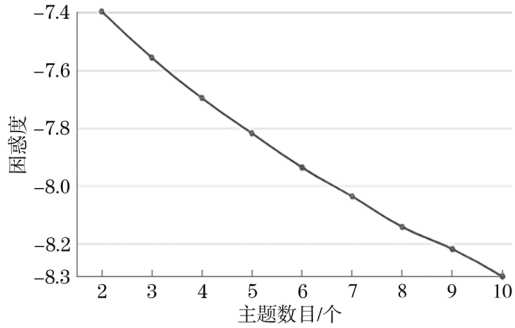

图4　困惑度分布曲线

表3　智能育种领域研究主题及核心特征词

主题	核心特征词
下一代测序技术/分子标记技术	分类、全基因组、大豆、鉴定、单核苷酸多态性、等位基因、有机、平均、长度、算法、特征、关联分析、生长、耐旱性、标记
作物生长/生产预测	网络分析、温度、氮、番茄、回归、生长模型、大豆、产量预测、序列发散、作物表面模型、全基因组关联、遗传资源、多重性状、产量变异性、代谢抗性
农情智能监测	多层感知器、调节亏缺灌溉、氮素、等位基因、聚类、杂交、LSTM神经网络、基因组、多态性、qtl分析、直接耦合分析、主坐标分析、作物模拟模型、遗传多样性分析、土壤化学计量
图像处理技术及应用	计算机辅助图像分析、qtl分析、三维图像分析、统计耦合分析、弹性网络模型、连续投影算法、注意力机制、养分管理、生境管理、耐旱性、根系多样性、表型分析、花期监测、种子分级、密度估计
作物智能生产管理	夜间温度、浓度、重量、肥料发芽曲线、空间数据分析、图像分析、纹理轮廓分析、无人机、作物表面模型、分子动力学模拟、长度密度模型、废物管理、害虫管理、甲烷排放、自动化育种

力的指标值，为了能够突出新涌现的新兴主题，减弱影响力的长期积累效果，将新颖性、突发性和影响力的权重分别设置为0.4、0.4和0.2，按照最大-最小值法归一化，计算得到5个主题的总分数，结果见表4。

表4　智能育种领域交叉主题新兴性测算结果

主题名称	新颖性	突发性	影响力	总分数
下一代测序技术/分子标记技术	2 019.93	0.07	11.49	0.932
图像处理技术及应用	2 019.94	0.37	9.49	0.929
农情智能监测	2 020.30	0.12	9.64	0.885
作物生长/生产预测	2 020.13	0.19	9.44	0.826
作物智能生产管理	2 020.14	0.25	8.93	0.599

从新颖性来看，农情智能监测、作物智能生产管理、作物生长/生产预测的论文发表年份较近，平均年份为2020年；从突发性来看，图像处理技术及应用、作物智能生产管理、作物生长/生产预测主题的热度增长较快，是迅速引起关注的热点主题；从影响力来看，下一代测序技术/分子标记技术、农情智能监测、图像处理技术及应用的研究论文数量较多，是智能育种领域持续受关注的研究主题。

综合各项指标得分来看，5个主题中作物智能生产管理分数最低，得分为0.599，且与其他4个主题的分数差距较大，故选取研究主题新兴性综合得分在0.8以上的主题，即下一代测序技术/分子标记技术、图像处理技术及应用、农情智能监测和作物生长/生产预测为新兴主题。

3.4.2　新兴主题解读

结合专家咨询与文献判读，对新兴主题内容进行深入解读，以准确提炼主题方向，也为方法的验证环节提供可靠性的事实依据。

（1）下一代测序技术/分子标记技术。生物育种与现代信息技术加速融合，大数据、人工智能开始应用于基因型检测、分子标记、表型处理等方面，正在成为辅助育种、提高育种效率的重要手段。①下一代测序技术。下一代测序技术的发展和应用推动了全基因组范围内的基因测序和表观遗传修饰位点鉴定与功能机制研究，是作物智能育种的基础。一方面，人工智能有助于对基因组数据展开深度分析，如土壤菌落的基因组鉴定[23]、小麦品种的转录组鉴定与对比[24]、鹰嘴豆重组自交系的基因分型[25]等。将基因型的鉴定与机器学习相结合来预测表型是新兴方向，如Monreal等[26]利用16SrDNA下一代测序技术，分析土壤细菌群落的组

成和相对丰度，建立了一个生态功能概念模型。Liu等[27]通过比较高光合效率小麦品种BN207与其亲本的光合生理和转录组，发现了影响光合效率的关键基因。②分子标记技术。标记辅助选择思想在全基因组范围内的扩展是智能育种的重要组成部分。分子标记技术与人工智能技术的结合主要表现为基因型到表型的预测，如抗病性、产量和品质等，相关研究通过机器学习算法分析基因型和表型数据、融入大数据的精准育种模式、智能育种平台的构建以及育种策略的优化。Shin等[28]利用基因分型测序方法对油棕种间杂交和回交后代进行分析，通过单标记-性状关联分析鉴定了与性状相关的分子标记及影响脂肪酸合成的关键候选基因，并利用两种机器学习算法评估分子标记对表型值的预测能力。还有研究者基于智能育种平台自动化获取与解析基因型和表型数据，以及羽衣甘蓝、香菇、木薯等的QTL（Quantitative Trait Locus）快速定位性状[29-31]。

（2）图像处理技术及应用。图像处理技术是推动农业信息化发展与智能化赋能的关键技术之一，研究主题主要涉及光谱、热红外、遥感卫星等成像技术在作物分类/种子识别、病虫害识别、杂草识别、作物生长/产量预测等方面的应用。①作物分类/种子识别。通过结合深度学习算法、图像处理与遥感技术对作物进行识别。Liu等[32]建立了一种基于高光谱特征融合的油菜品种精准识别模型，识别率高于93.71%。②病虫害识别。利用光谱遥感技术获取虫害作物的光谱特征，结合机器学习建立估测模型，识别多种病虫害。目前，小麦条锈病与叶锈病耦合识别[33]、油菜茎上黄斑细球绦虫子实体密度识别[34]、小麦赤霉病识别[35]等方面都已有相关研究。③土壤特征及土地适宜性评价。Ismaili等[36]基于多种机器学习算法验证了物候参数对土壤适宜性预测影响最大。

（3）农情智能监测。农情智能监测是智慧农业的重要组成部分，基于无线传感器网络、远程遥感技术、智能农业平台等，实现对农田环境、作物生长、气象气候和农业管理等方面的实时监测和评估，以提高农作物产量和质量，降低生产成本和风险。①农田环境监测。采用土壤墒情仪[37]、土壤分析仪[38]、土壤温度传感器[39]等工具，对土壤环境和养分含量进行监测。Kelly等[40]研究了农田土壤水分监测对灌溉用水效率的影响。Zhang等[41]构建了土壤质地和总碳预测模型，分析了美国土壤质地和总碳的短期变化特征。②农作物生长监测。通过植被指数测量[42]、作物生理参数测量[43]、遥感技术[44]等，对作物的生长状态、叶绿素含量等进行监测。Qi等[45]基于无人机多光谱图像特征与植被指数构建了花生叶绿素预测模型，为最优作物类型选取、肥效评价及种植密度管理提供决策支持。③气象气候监测。利用气象传感器、卫星遥感等，对温度、湿度、降水量、风速等进行监测。Cheng等[46]基于新一代全球降水测量卫星、植被指数和地表温度提出了一种新型的综合遥感干旱指数，用于监测气象干旱和农业生态干旱。④农业管理措施监测，涉及灌溉量、施肥量、病虫害等信息监测。Qian等[47]提出了一种基于遥感的长短期记忆模型，用于亚像素尺度农田秋冬灌溉程度的实时监测，有助于精准管控灌溉时间与灌溉用水量。

（4）作物生长/生产预测。作物生长/生产预测是农业大数据分析、图像识别技术、遥感技术和深度学习等多种技术综合应用的结果。①基于大数据和人工智能算法，开发作物生长及预测模型。García-Martínez等[48]基于RGB图像特征构建了玉米种植密度预测模型。Kumar等[49]基于图像处理与随机森林算法，构建了冬小麦耐旱性预测模型。②作物物质含量预测。Cho等[50]构建了甜瓜固溶体浓度和水分含量的预测模型。此外，基于神经网络的作物叶片钼、铜等重金属含量的预测模型可用于作物施肥与灌溉的精准监控[51]。③作物产量预测。相关学者基于图像识别技术分别构建了水稻[52]、小麦[53]等的产量预测模型，为精准调控作物产量影响因素、优化作物生产模式与管理等提供支撑。

3.5 结果有效性验证

新兴主题识别是一种预估性工作，没有可衡量识别结果准确性的通用定量标准。本研究参考郝雯柯等[54]的验证方法，采用资料分析法验证文本分类领域新兴主题识别结果的科学性。应继锋等[55]在《第5代（5G）作物育种技术体系》中提到：基于基因型大数据、表型大数据、环境大数据构建的基因型-表型-环境模型，以及人工智能技术，特别是图像识别技术、数字化图像处理技术等现代技术的快速发展将作物育种引向了新的阶段。《2023全球农业研究热点前沿》提到了机器视觉在农业生产中的应用、多源遥感技术在作物产量估测中的应用、无人机遥感在农业监测中的应用、智慧农业决策支持系统、基于深度学习的作物病害自动识别等热点前沿方向[56]。汪海等[57]关注作物表型-环境大数据获取解析（新一代传感器、作物表型高通量获取设施装备、物联网和表型智能解析）、多重组学大数据分析、多维大数据驱动的智能

育种预测模型构建、育种大数据存储管理与应用等。

综合来看，本研究识别出的智能育种领域新兴交叉主题基本贴合了上述相关表述，证明了本文方法的有效性与准确性。

4　结语

本研究基于智能育种领域的科技论文数据，通过学科交叉性测度、LDA主题抽取和新兴性测度，共识别出4个研究主题。论文的创新性主要在于：①提出了融合学科多样性与凝聚性的TD_C指数，用于学科交叉性测度；②利用农业词表对LDA模型进行修正，提高领域的适用性；③在学科交叉性测度基础上，构建了包含3个维度的新兴主题测度模型，完成新兴交叉主题识别。

需要指出的是，文中所提方法可以拓展至其他学科领域，须构建对应学科和领域的专业词表进行模型训练与优化，进行多领域实验结果的交叉验证。同时还可以丰富数据源，进一步引入专著、专利等科技文本，使数据源多样化。随着科学研究的交叉程度日趋加深，未来可以重点关注如何从更多维度或机制层面客观地识别新兴方向，提高主题表达的准确度，深入开展对学科交叉新兴领域的分析与研究，助力面向重大需求的交叉创新方向的科学布局，培育发展各领域的新质生产力。

参考文献

[1] 张雪，张志强，朱冬亮. 基于时间序列分析的潜在学科交叉前沿主题识别研究[J]. 情报理论与实践，2024，47（4）：152-162.

[2] 杨金庆，张力. 学科交叉视角下新兴主题识别特征分析：以医学信息学为例[J]. 情报工程，2021，7（4）：3-12.

[3] MATSUMURA N，MATSUO Y，OHSAWA Y，et al. Discovering emerging topics from WWW[J]. Journal of Contingencies and Crisis Management，2002，10（2）：73-81.

[4] ROTOLO D，HICKS D，MARTIN B R. What is an emerging technology? [J]. Research Policy，2015，44（10）：1827-1843.

[5] WANG Q. A bibliometric model for identifying emerging research topics[J]. Journal of the Association for Information Science and Technology，2018，69（2）：290-304.

[6] XU H Y，WINNINK J，YUE Z H，et al. Multidimensional scientometric indicators for the detection of emerging research topics[J]. Technological Forecasting and Social Change，2021，163：120490.

[7] SMALL H. Co-citation in the scientific literature：a new measure of the relationship between two documents[J]. Journal of the American Society for Information Science，1973，24（4）：265-269.

[8] 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.

[9] XU M，LI G J，WANG X D. Detecting emerging topics by exploiting probability burst and association rule mining：a case study of library and information science[J]. Malaysian Journal of Library & Information Science，2020，25（1）：47-66.

[10] 白敬毅，颜端武，陈琼. 基于主题模型和曲线拟合的新兴主题趋势预测研究[J]. 情报理论与实践，2020，43（7）：130-136，193.

[11] XU S，HAO L Y，AN X，et al. Emerging research topics detection with multiple machine learning models[J]. Journal of Informetrics，2019，13（4）：100983.

[12] RAFOLS I，MEYER M. Diversity and network coherence as indicators of interdisciplinarity：case studies in bionanoscience[J]. Scientometrics，2010，82（2）：263-287.

[13] 陈赛君，陈智高. 学科领域交叉性及对其测度的Φ指标：以我国科学学研究领域为例[J]. 科学学与科学技术管理，2014，35（5）：3-12.

[14] 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.

[15] LEYDESDORFF L. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals[J]. Journal of the American Society for Information Science and Technology，2007，58（9）：1303-1319.

[16] BLEI D，NG A，JORDAN M. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research，2003，3：993-1022.

[17] 张彪，吴红，高道斌，等. 基于潜在高被引论文与高价值专利的创新前沿识别研究[J]. 图书情报工作，2022，66（18）：72-83.

[18] 闫盛枫. 融合词向量语义增强和DTM模型的公共政策文本时序建模与演化分析：以“大数据领域”为例[J]. 情报科学，2021，39（9）：146-154.

[19] 冯艳铭，郝志梅，董春栎. 基于LDA模型的老年人生活满意度主题挖掘与文本实证分析[J]. 华北理工大学学报（社会科学版），2024，24（2）：19-25.

[20] SCHMIEDEL T, MÜLLER O, VOM BROCKE J. Topic modeling as a strategy of inquiry in organizational research: a tutorial with an application example on organizational culture[J]. Organizational Research Methods, 2019, 22(4): 941-968.

[21] KLEINBERG J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373-397.

[22] LEYDESDORFF L, CARLEY S, RAFOLS I. Global maps of science based on the new Web-of-Science categories[J]. Scientometrics, 2013, 94(2): 589-593.

[23] MA Y, QIU C W, FAN Y, et al. Genome-wide association and transcriptome analysis reveals candidate genes for potassium transport under salinity stress in wheat[J]. Environmental and Experimental Botany, 2022, 202: 105034.

[24] XU P D, XIE S Q, LIU W B, et al. Comparative genomics analysis provides new strategies for bacteriostatic ability of Bacillus velezensis HAB-2[J]. Frontiers in Microbiology, 2020, 11: 594079.

[25] AMALRAJ A, TAYLOR J, BITHELL S, et al. Mapping resistance to Phytophthora root rot identifies independent loci from cultivated (Cicer arietinum L.) and wild (Cicer echinospermum P.H. Davis) chickpea[J]. TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik, 2019, 132(4): 1017-1033.

[26] MONREAL C M, ZHANG J. An ecological function conceptual model for bacterial communities with high relative abundance in an unplanted and canola (Brassica napus) planted Podzol[J]. Rhizosphere, 2018, 5: 26-31.

[27] LIU H J, ZHU Q D, PEI X X, et al. Comparative analysis of the photosynthetic physiology and transcriptome of a high-yielding wheat variety and its parents[J]. The Crop Journal, 2020, 8(6): 1037-1048.

[28] SHIN M G, ITHNIN M, VU W T, et al. Association mapping analysis of oil palm interspecific hybrid populations and predicting phenotypic values via machine learning algorithms[J]. Plant Breeding, 2021, 140(6): 1150-1165.

[29] REN J, LIU Z Y, DU J T, et al. Fine-mapping of a gene for the lobed leaf, BoLl, in ornamental kale (Brassica oleracea L. var. acephala) [J]. Molecular Breeding, 2019, 39(3): 40-45.

[30] LEE H Y, MOON S, RO H S, et al. Analysis of genetic diversity and population structure of wild strains and cultivars using genomic SSR markers in Lentinula edodes[J]. Mycobiology, 2020, 48(2): 115-121.

[31] ALMEIDA COSTA N, DA SILVA AZÊVEDO H S F, DA SILVA L M, et al. Molecular characterization and core collection evaluation of Manihot esculenta Crantz[J]. Bioscience Journal, 2020, 36: 22-35.

[32] LIU F, WANG F, WANG X Q, et al. Rapeseed variety recognition based on hyperspectral feature fusion[J]. Agronomy, 2022, 12(10): 2350.

[33] WANG H L, JIANG Q, SUN Z Y, et al. Identification of stripe rust and leaf rust on different wheat varieties based on image processing technology[J]. Agronomy, 2023, 13(1): 260-265.

[34] BOUSSET L, PALERME M, LECLERC M, et al. Automated image processing framework for analysis of the density of fruiting bodies of Leptosphaeria maculans on oilseed rape stems[J]. Plant Pathology, 2019, 68(9): 1749-1760.

[35] MAO R, WANG Z C, LI F L, et al. GSEYOLOX-s: an improved lightweight network for identifying the severity of wheat fusarium head blight[J]. Agronomy, 2023, 13(1): 242-247.

[36] ISMAILI M, KRIMISSA S, NAMOUS M, et al. Assessment of soil suitability using machine learning in arid and semi-arid regions[J]. Agronomy, 2023, 13(1): 165-168.

[37] CHAKRABORTY M, MALKANI A, BISWAS K. Hand-held soil moisture meter using polymer coated sensor[J]. IEEE Instrumentation & Measurement Magazine, 2019, 22(5): 24-29.

[38] HAO J, LI F S, JIANG X Y, et al. Improvement approach for determination of cadmium at trace levels in soils by handheld X-ray fluorescence analyzers[J]. Spectrochimica Acta Part B: Atomic Spectroscopy, 2023, 206: 106711.

[39] LIU M. Modified monitoring system of soil temperature based on ARM[J]. Environmental Technology & Innovation, 2021, 21: 101346.

[40] KELLY T D, FOSTER T, SCHULTZ D M, et al. The effect of soil-moisture uncertainty on irrigation water use and farm profits[J]. Advances in Water Resources, 2021, 154: 103982.

[41] ZHANG Y K, HARTEMINK A E. Quantifying short-range variation of soil texture and total carbon of a 330-ha farm[J]. CATENA, 2021, 201: 105200.

[42] KRISHNAN S, INDU J. Assessing the potential of temperature/vegetation index space to infer soil moisture over Ganga Basin[J]. Journal of Hydrology, 2023, 621: 129611.

[43] QIU Q, ZHENG C F, WANG W P, et al. A new strategy in observer modeling for greenhouse cucumber seedling growth[J]. Frontiers in Plant Science, 2017, 8: 1297.

[44] REBOUH N Y，MOHAMED E S，POLITYKO P M，et al. Towards improving the precision agriculture management of the wheat crop using remote sensing：a case study in central non-black earth region of Russia[J]. The Egyptian Journal of Remote Sensing and Space Sciences，2023，26（3）：505-517.

[45] QI H X，WU Z Y，ZHANG L，et al. Monitoring of peanut leaves chlorophyll content based on drone-based multispectral image feature extraction[J]. Computers and Electronics in Agriculture，2021，187：106292.

[46] CHENG Y J，ZHANG K，CHAO L J，et al. A comprehensive drought index based on remote sensing data and nested copulas for monitoring meteorological and agroecological droughts：a case study on the Qinghai-Tibet Plateau[J]. Environmental Modelling & Software，2023，161：105629.

[47] QIAN X M，QI H W，SHANG S H，et al. Deep learning-based near-real-time monitoring of autumn irrigation extent at sub-pixel scale in a large irrigation district[J]. Agricultural Water Management，2023，284：108335.

[48] GARCÍA-MARTÍNEZ H，FLORES-MAGDALENO H，KHALIL-GARDEZI A，et al. Digital count of corn plants using images taken by unmanned aerial vehicles and cross correlation of templates[J]. Agronomy，2020，10（4）：469.

[49] KUMAR D，KUSHWAHA S，DELVENTO C，et al. Affordable phenotyping of winter wheat under field and controlled conditions for drought tolerance[J]. Agronomy，2020，10（6）：882.

[50] CHO B H，LEE K B，HONG Y，et al. Determination of internal quality indices in oriental melon using snapshot-type hyperspectral image and machine learning model[J]. Agronomy，2022，12（9）：2236.

[51] MURADYAN V，TEPANOSYAN G，ASMARYAN S，et al. Estimating Mo，Cu，Ni，Cd contents in the crop leaves growing on small land plots using satellite data[J]. Communications in Soil Science and Plant Analysis，2020，51（11）：1457-1468.

[52] YANG M，XU X G，LI Z Y，et al. Remote sensing prescription for rice nitrogen fertilizer recommendation based on improved NFOA model[J]. Agronomy，2022，12（8）：1804-1810.

[53] LI L，HASSAN M A，YANG S R，et al. Development of image-based wheat spike counter through a faster R-CNN algorithm and application for genetic studies[J]. The Crop Journal，2022，10（5）：1303-1311.

[54] 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.

[55] 应继锋，刘定富，赵健. 第5代（5G）作物育种技术体系[J]. 中国种业，2020（10）：1-3.

[56] 孙巍，李周晶，马晓敏，等. 2023全球农业研究热点前沿分析解读[J]. 农学学报，2024，14（3）：5-9.

[57] 汪海，赖锦盛，王海洋，等. 作物智能设计育种：自然变异的智能组合和人工变异的智能创制[J]. 中国农业科技导报，2022，24（6）：1-8.

作者简介

齐世杰，女，硕士，助理研究员，研究方向：科学计量学、文本挖掘。
串丽敏，女，博士，副研究员，通信作者，研究方向：智能知识服务技术，E-mail：Chuanll@agri.ac.cn。
赵静娟，女，硕士，副研究员，研究方向：科技情报研究。
张辉，男，硕士，助理研究员，研究方向：智能知识服务技术。
贾倩，女，硕士，助理研究员，研究方向：科技情报分析技术。

Emerging Cross Topic Recognition in the Field Based on Topic Models: Taking Intelligent Crop Breeding as an Example

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

Abstract: Accurately identifying cutting-edge interdisciplinary topics helps to understand the development context of disciplines, explore key development directions in the field, and provide references for future innovative and breakthrough research. The article proposes a method for identifying emerging cross topics. Firstly, a method for calculating the interdisciplinary degree of a paper is proposed, which combines disciplinary diversity and cohesion. By using this method, papers with high interdisciplinary degree are selected to obtain potential datasets. Then, the study utilizes an improved LDA model combined with domain dictionaries to identify research topics. Finally, by constructing a multidimensional emerging topic measurement model that integrates novelty, breakthrough, and influence, the study identifies emerging cross topics. This study conducts empirical analysis in the field of intelligent crop breeding, identifies 4 emerging cross topics, and validates the effectiveness of the method through literature analysis. The research findings offer valuable insights for the research and application of identifying emerging cross topics based on scientific papers.

Keywords: Interdisciplinary Research; Emerging Topic Recognition; Topic Modeling; Intelligent Crop Breeding

（责任编辑：王玮）