doi:10.3772/j.issn.2095-915x.2021.05.009

基于 BERT 的农业领域前沿研究主题识别方法研究

开放科学（资源服务）标识码（OSID）

李松繁[1,2] 黄永[1,2] 杨金庆[1,2]

1. 武汉大学信息管理学院 武汉 430072；
2. 武汉大学信息检索与知识挖掘研究所 武汉 430072

摘要：[目的 / 意义] 为了快速准确地识别学科领域的前沿研究主题，为科研工作者提供精准的学科发展趋势预测服务。[方法 / 过程] 本文提出了一种基于 BERT 的前沿研究主题识别方法，结合本文改进的主题新颖度和提出的相关文献指标、主题发展态势指标，实现农业领域前沿研究主题的识别。[结果 / 结论] 基于农业领域期刊论文数据的实证研究发现，该方法在前沿研究主题的识别任务上效果显著，有效识别出了农业领域内的潜在前沿研究主题 1 个、热门前沿研究主题 2 个、衰退前沿研究主题 2 个。

关键词：前沿主题；BERT；文本聚类；前沿识别

中图分类号：G35

Research on Frontier Research Topic Recognition Method in Agriculture Field Based on BERT

LI Songfan[1,2] HUANG Yong[1,2] YANG Jinqing[1,2]

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072, China

Abstract: [Objective/Significance] In order to quickly and accurately identify the frontier research topics in the subject area, provide researcher with accurate subject development trend forecasting services. [Method/Process] This paper proposes a

基金项目 国家自然科学基金青年项目“主题演化中个体选择行为与整体主题分布的交互影响机制研究”（72004168）。

作者简介 李松繁（1996-），硕士研究生，研究方向为信息检索，知识挖掘，E-mail: sliongfan@whu.edu.cn；黄永（1991-），博士后，研究方向为信息检索、机器学习；杨金庆（1991-），博士研究生，研究方向为数据智能，信息组织与检索。

引用格式 李松繁 , 黄永 , 杨金庆 . 基于 BERT 的农业领域前沿研究主题识别方法研究 [J]. 情报工程 , 2021, 7(5): 100-114.

method for identifying frontier research topics based on BERT, combining the novelty of the topic improved in this paper, related literature indicators and topic development trend indicators, to realize the identification of frontier research topics in the agricultural field. [Result/Conclusion] Empirical study based on the data of journal papers in the agricultural field found that the method has significant effects on the task of identifying frontier research topics, effectively identifying 1 potential frontier research topic, two hot frontier research topics, and 2 recession frontier research topics in the agricultural field.
Keywords: Frontier topic; BERT; text clustering; frontier identification

引言

随着大数据时代的开启和快速发展，海量数据支撑下的科学研究为经济发展、产业形态转型、产业升级提供了强大的驱动力。作为科学研究中的重要数据组织形态，数量庞大的科技文献成为了重要的科研资产。在科研方法和科研数据快速更迭的背景下，如何从科技文献资源中快速准确地识别学科领域内的研究主题，成为了科研人员的亟待解决的重要问题。

“前沿性”是学科领域前沿研究主题的核心特征，实现对学科内的前沿研究主题挖掘，是情报工作者的重要研究内容。目前的学科前沿研究识别方法大多利用科技文献的内外部特征来构建“前沿性指标”。在外部特征方面，有学者利用科技文献的引文信息构建引文网络，揭示其中的学科发展脉络[1]；在内部特征方面，研究者利用科技文献的关键词，使用多种分析方法识别研究前沿[2]；也有学者使用热门文本分析工具“主题模型”，对科技文献内部的浅层语义特征进行抽取，挖掘出学科研究主题以进行前沿研究主题探测[3]。然而基于引文网络或共词分析的前沿研究识别方法只揭示了学科研究的结构关系，没有考虑到学科领域内研究内容的语义特征。基于主题模型的前沿研究主题挖掘方法也只从浅层语义分析了学科研究主题，同时主题模型在参数设置上较为复杂，抽取主题的可解释性也有待优化。

基于以上背景，本文立足于科技文献资源，利用时下热门的预训练模型对科学论文的文献内容进行深入挖掘分析，提出主题新颖度、相关文献指数、主题发展指数等多维前沿识别指标来构建一种前沿研究主题识别方法。

1 相关研究

“研究前沿”这一概念最早由科学计量之父普赖斯提出[4]，发展到如今已有大量国内外学者进行了相关研究。目前已有的前沿研究识别方法主要从定性分析和定量挖掘两方面进行。专家判断法是最为典型的定性分析方法，该方法利用学科领域内的专家拥有的大量知识和经验进行前沿研究的判断和预测。但该方法主观性较强，判断结果可能出现主观偏差[5]。

定量分析方法是学科前沿研究的主要分析方法，根据实现方法可将其分为基于引文分析的方法、基于关键词的分析方法和基于主题模型的识别方法等。

doi:10.3772/j.issn.2095-915x.2021.05.009

（1）基于引文的分析方法

基于引文的识别方法从科技文献彼此间的引用关系出发，通过构建引文网络，完成相关的可视化分析和前沿主题识别[6]。

1973 年 H Small[7] 最早提出用共被引分析法进行前沿主题识别研究。研究者利用文献共被引关系构建引用网络，进行聚类分析，划分关键节点，结合对关键节点的内容分析进行前沿识别。许振亮等[8]利用引文网络中关键被引文献的内容分析计量出前沿研究主题；潘黎等[9]基于 SSCI 高等教育学期刊绘制文献共被引网络图谱，识别国际高等教育的研究前沿； Huang 等[10]通过对比文献共被引和引文耦合分析方法在探测有机发光二极管（OLED）领域前沿研究时发现，引文耦合在前沿识别的数量和速度上有一定优势。

基于引文分析的前沿研究识别方法起步较早，至今仍有广泛应用。但由于一篇文献需要经过一段较长的时间积累被引次数，导致此方法存在时间滞后性，无法及时获取前沿研究。同时引文耦合关系在施引文献发表时就已经确定，使得该方法缺乏动态发展性。

（2）基于关键词的识别方法

基于关键词的前沿识别方法，是以学科研究关键词为出发点，从词汇角度进行相关前沿探测，如词频分析法、共词分析法等。

词频分析法通过统计文献主题词的词频或随着时间变化的词频变化率来完成前沿主题识别。研究者通常将词频突然增长的“突发词”作为前沿主题词。J Kleinberg[11]提出的突变检测算法可以用于探测一个学科领域内突然增长的研究兴趣，基于此可以完成突发词的检测和识别。

共词分析法对同时出现在同一文献中的词汇对进行分析，它将不同词语进行连接，弥补了词频分析法中主题词孤立的缺陷，能反映出学科领域内知识结构的变化，反映学科概念和主题的增长规律[12]。章成志等[13]采用主题聚类方法，以包含时间信息的学术论文为数据集进行主题聚类，归纳出某一学科领域的研究热点和研究趋势。侯海燕等[14]将共被引与共词分析相结合的方法，利用知识图谱，得出了科学计量学领域的前沿课题及重点研究方向。

基于关键词的前沿研究识别方法，能够从微观层面挖掘文本间的结构信息，但缺乏对文本内容语义信息的挖掘。

（3）基于主题模型的识别方法

以隐狄利克雷分布（Latent Dirichlet Allocation，LDA）算法[15]为代表的主题模型，用无监督学习的方式对全文本进行语义结构和聚类分析，从文本中抽取有价值的主题及主题关键词分布。基于主题模型的前沿主题识别方法，在一定程度上弥补了引文分析和关键词分析的不足，使得前沿主题包含了更多的文本语义信息。范云满等[16]基于 LDA 主题模型，构建了主题新颖度、作者发文量、文章被引量结合的识别指标，对新兴主题进行探测。朱茂然等[17]通过不同时间窗口下的相似主题比例分布和主题 - 词汇分布，分别解释主题强度的变化和主题内容的变化，并对情报学领域的前沿主题进行识别和主题演化分析。杨金庆等[18]使用 LDA 主题模型完成多源科技文献的主题抽取和主题相似度计算，寻求多源科技文献主题的最优匹配组合，完成多源科技文献的时滞性计算。

基于主题模型的前沿研究识别方法，通常需要超参数调优来发掘主题，主题模型得到的结果解释性程度不高，难以直观理解主题含义，对于文本的语义理解也只停留在浅层语义挖掘上，无法获得文本的深层语义。

综上，学科领域前沿主题识别方法已经有了一定程度的研究，但多是从科技文献的结构关系出发进行前沿研究识别，或对科技文献进行了浅层语义分析，缺乏对文献内容的深层语义挖掘，同时前沿研究主题的可解释性也不够突出。因此，本文拟利用期刊论文数据，运用文本句嵌入构建、文本聚类、主题关键词抽取等方法，深度挖掘数据内部语义信息，同时提出主题新颖度、相关文献指数、主题发展态势指数等多维前沿识别指标，构建一种基于 BERT[19] 的学科领域前沿研究主题识别方法。

2 研究方法

本文提出的基于 BERT 的学科领域前沿研究主题识别方法设计思路如图（1）所示。首先，从科技文献数据库中收集期刊论文数据，构建待分析内容语料库；其次，对语料库数据按照发表年份划分时间窗口，将其归类到不同时间窗口下；再次对数据进行预处理，使用 BERT 模型构建文本的句嵌入集合，并在此基础上使用文本聚类算法进行文本聚类，抽取聚类簇中的重要主题词作为该类簇的主题表示；最后，计算研究主题间的相似度，构建多维前沿识别指标，识别潜在前沿研究主题、热门前沿研究主题和衰退前沿研究主题。

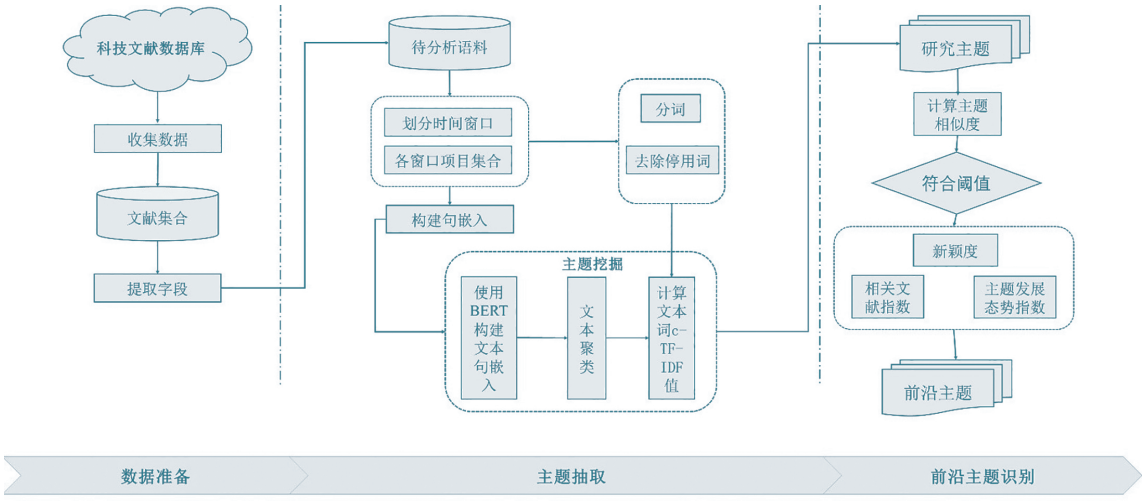

图 1 技术路线图

（1）数据获取

本文选择科技文献数据库作为数据来源，收集科技文献的核心数据，利用发表年份、摘要等字段进行分析研究。

（2）数据预处理

数据预处理主要包括划分时间窗口和文本数据的分词、去除停用词。以科技文献的不同发表年份作为时间窗口，将其划分到不同时间

doi:10.3772/j.issn.2095-915x.2021.05.009

窗口下，便于抽取每个时间窗口下的文献主题。同时对数据进行分词，将分词结果运用于聚类簇的主题词抽取。为了降低程序计算量，提高处理效率，本文在进行主题抽取之前对文本停用词进行剔除，使得科技文献主题抽取的结果更加准确，更加接近文献的真实主题。

（3）研究主题抽取

本文利用 BERT 预训练模型对不同时间窗口下的原始文本数据进行句嵌入构建。BERT 模型是通过自监督学习，从大规模语料中获得与具体任务无关的、独立的模型，它能够体现某一个词在上下文中的深度语义特征。

获得文本的句嵌入集合后，本文使用 HDBSCAN 文本聚类算法 [20] 对句嵌入集合进行聚类分析，获得文本聚类簇。HDBSCAN 算法是一种基于密度和基于层次的文本聚类算法，它不用人工设置主题数目，只用设置最小生成聚类集合的大小，算法可以自动推荐最优的聚类结果，同时为类簇中的每个文本分配主题标签。

本文使用 c-TF-IDF 算法挖掘聚类簇中的重要主题词。c-TF-IDF 算法是从 TF-IDF 算法中衍生出的基于聚类集合的 TF-IDF 方法，该方法应用于多个聚类集合，将每个集合的所有文档合并为一个文档。然后，针对每个聚类簇 i，提取单词的频率 t，除以单词总数 w。接着将所有类别 m 中未合并的文档总数除以所有聚类簇 i 的单词频率总和。c-TF-IDF 可以表示为公式（1）：

$$c-TF-IDF_i=\frac{t_i}{w_i}\times\log\frac{m}{\sum_j^n t_j} \quad (1)$$

通过公式（1）计算文本类簇中词汇对类簇的重要程度，选择最重要的词汇作为该类簇的主题词，完成主题抽取。

（4）研究主题向量构建与相似度计算

BERT 模型基于句子级别的语料进行训练，在使用时接受 1~2 句话作为输入。为了能够得到完整的主题向量化表示，本文将每个主题包含的主题词进行拼接，按照特定的格式输入到 BERT 模型中，以得到主题向量。

获得主题的向量化表示后，本文使用点积余弦相似度计算不同主题之间的相似度。余弦相似度通过计算两个向量之间的夹角大小来测度向量相似性，余弦相似度值越接近 1，说明两主题相似性越高，如公式（2）所示：

$$\cos\theta=\frac{a-b}{\|a\|\cdot\|b\|}=\frac{\sum_{i=1}^n A_iB_i}{\sqrt{\sum_{i=1}^n A_i^2}\sqrt{\sum_{i=1}^n B_i^2}} \quad (2)$$

公式（2）中，θ 代表向量夹角，n 代表向量维度，A_i 代表向量 A 中第 i 个值，B_i 代表向量 B 中第 i 个值。

本文通过计算不同时间窗口下的子主题相似度来获得学科领域的所有父主题，过程如下：以时间窗口第一年为初始年份，该年下各主题为初始主题，计算初始主题与其他年份下各子主题的相似度，获得初始主题与其他年份主题的相似度列表。接着对相似度列表进行筛选，相似度值最高且高于某一阈值的主题可以划分为同一主题。完成所有年份的计算后，同一类型的子主题集合组成同一父主题。如果初始年份往后各年中存在子主题不属于之前划分的父主题，则以该主题为新的初始点，重复上述过程，获得新的父主题。所有的父主题构成了学科领域主题集合。

（5）前沿研究主题识别指标

在完成学科领域所有主题识别的基础上，

本文参考已有研究对主题新颖度指标进行改进，并结合本文提出的相关文献指数、主题发展态势指标以衡量主题的前沿性。

① 新颖度

主题新颖度指标主要用于探测学科内某主题的新颖程度。冯佳等 [3] 采用“某主题的平均发文时间”来计算主题新颖度，时间距今越近，主题新颖度越高。考虑到本文使用的领域主题抽取方法，本文将前文提到的时间窗口按照时间顺序切分为两个时期，将学科内某个父主题下子主题首次出现的年份作为该父主题的出现年份，以出现年份所属的时期来判断主题新颖度，时期越早，该主题新颖度越低；时期越晚，该主题新颖度越高。

②相关文献指数

领域内某主题是否拥有较高的研究热度，可以通过关注该主题的研究者数量或其他类似指标进行测度。曾海娇等 [21] 统计研究主题的作者数量来衡量主题的受关注度。本文在使用 HDBSCAN 算法在每个时间窗口下进行文本聚类时，每个文本已经被分配主题标签，这些文本可以被视为与子主题相关的文本。因此，本文采用相关文献指数来表征主题的研究热度。可以利用与主题相关的文本数量与对应年份下平均主题相关文本量的比值来表示主题的相关文献指数：

$$\theta_j = \frac{d_j}{M_t} \qquad (3)$$

公式（3）中 θ_j 代表子主题 j 的相关文献指数值，也即父主题在时间窗口 t 下的相关文献指数；$M_t = \frac{C_t}{n}$ 代表时间窗口 t 下平均主题相关文本量，C_t 为 t 年的相关文本总数，n 为 t 年的主题数；d_j 代表主题 j 的相关文献数量。

在同一年份中，相关文献指数的标准值为 1，如果主题相关文献指数大于标准值，表示该主题受到关注较多，是热门主题；如果主题相关文献指数小于标准值，表示该主题受到关注较少，是冷门主题。

③主题发展态势指标

本文认为识别前沿主题，要把握前沿主题的发展趋势，直观了解学科领域中研究主题的发展过程、规律和态势。因此本文提出“主题发展态势指标”来衡量主题的前沿发展态势。本文首先计算父主题在不同时间窗口上的相关文献指数。然后将时间窗口 t 上该主题的相关文献指数 θ_j^t 与前一个时间窗口 t–1 的相关文献 θ_j^{t-1} 指数相加取平均值，将该平均值作为时间窗口的主题发展指数，并将当前时间窗口 t 的相关文献指数更新为平均值，以表示该主题随着时间推移的发展。主题发展指数计算如公式（4）：

$$\delta_i^t = \frac{\theta_i^t - \theta_i^{t-1}}{2} \qquad (4)$$

公式（4）中，δ_i^t 代表主题 i 在时间窗口 t 的主题发展指数，θ_i^t 代表主题 i 在时间窗口 t 的相关文献指数。

（6）前沿研究主题识别

根据前文所述前沿指标计算结果的不同，本文按照图（2）所示的识别逻辑，将领域前沿研究主题分为三类：

①潜在前沿研究主题

该类研究主题具有较高的主题新颖度，且最新相关文献指数高于标准值，同时主题发展呈明显上升趋势。这表明此类主题是近期出现的研究主题，且逐渐受到了科研工作者的关注

doi:10.3772/j.issn.2095-915x.2021.05.009

和重视，快速拥有了较高的研究热度。本文将该类主题定义为“潜在前沿研究主题”。

②热门前沿研究主题

该类研究主题不一定具有较高的主题新颖度，但自出现起其相关文献指数就维持在较高的水平，且主题发展没有明显下降趋势。这表明此类主题一直拥有较高的研究热度，且暂时没有下降的趋势。本文将该类主题定义为“热门前沿研究主题”。

③衰退前沿研究主题

该类研究主题新颖度较低，且在刚出现时拥有较高的初始相关文献指数，随着时间的发展该类主题相关文献指数逐年下降，最新相关文献指数低于标准值，主题发展呈明显下降趋势。这表明该主题研究热度已经有较长的研究年限，研究内容已无法体现学科领域的前沿知识。本文将该类主题定义为“衰退前沿研究主题”。

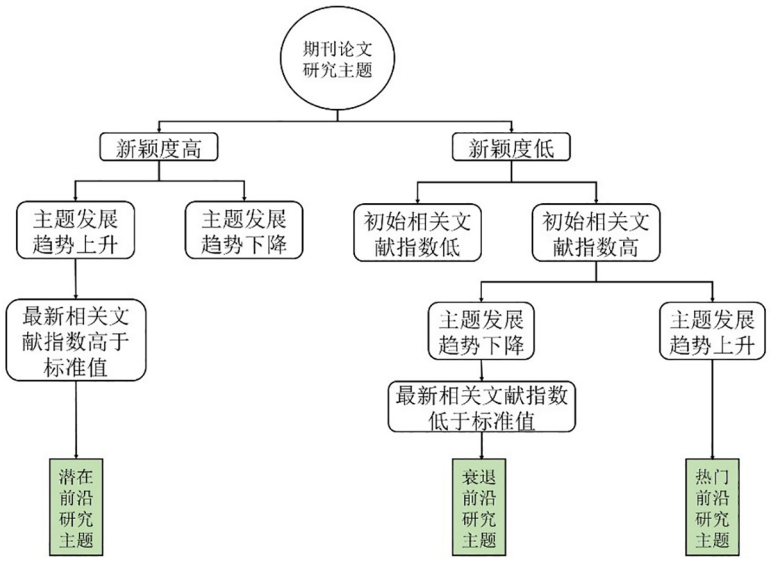

图 2 前沿研究主题划分思路

3 实证研究

3.1 实验环境

本文实验环境如表 1 所示。

表 1 实验环境与配置

实验环境	环境配置
操作系统	Ubuntu 20.04
GPU	NVIDIA GeForce RTX 3090
内存	32G
编程语言	Python 3.7
深度学习框架	TensorFlow

3.2 数据预处理

本文以农业领域为例对基于预训练模型的前沿研究主题识别方法进行实证。选取中国知网数据库收集农业领域相关期刊论文的标题、发表年份和摘要数据，共 57301 条，时间跨度为 2008-2018 年。对上述数据按发表年份进行时间窗口切片，每个时间窗口下的摘要数据进行分词、去除停用词。

3.3 研究主题抽取

在自然语言处理任务中，不存在一个在所

有可能的 NLP 任务上都表现出色的通用模型。本文选择在语义相似度计算、聚类方面表现良好的基于 BERT 的语言模型 Sentence-Transformers，并选择支持中文文本句嵌入构建的模型“distiluse-base-multilingual-cased-v1”进行文本句嵌入的构建。

获取到文本的句嵌入表征结果后，使用 HDBSCAN 算法进行文档聚类，以寻找到文本集合中的相似文档集合。本文设置 HDBSCAN 算法的相似度量方式为欧式度量，最小聚类大小为 30。

在生成文本聚类集合后，使用 c-TF-IDF 算法计算聚类集合中词汇的重要程度，取前 20 个词作为每个类簇的主题词，如 2008 年主题 2 的主题词为“基因 , 表达 , 序列 , 仔猪 , 扩增 , 引物 , 蛋白 , 病毒 , 遗传 , 克隆 , 载体 , 基因组 , 断奶 , 疫苗 , 检测 , 片段 , 重组 , 质粒 , 氨基酸 , 抗体”。从主题词可以分析出该主题研究内容为“基因技术与牲畜养殖应用”。值得注意的是，在文本聚类和主题抽取结果中，主题标签为“-1”的主题应该被排除，因为该主题表示的聚类结果被聚类模型视为“噪声”，其中掺杂了很多未被识别的主题，很难被人解读。

3.4 不同时间窗口间主题相似度计算

完成文本聚类集合的主题词抽取后，本文使用 BERT 预训练模型生成每个主题的主题向量，并计算不同时间窗口下两两主题之间的相似度。本文在设置相似度阈值的前提下，为了提升不同主题之间相似度与阈值差异的显著性，利用公式（5）来计算某一个时间窗口下某一主题与另一时间窗口下所有主题相似度与阈值差异的显著性指标：

$$\rho_i^j = \frac{s_i^j - T}{|S_{max} - T|} \qquad (5)$$

公式（5）中，ρ_i^j表示某一时间窗口下主题 i 与另一时间窗口下主题 j 相似度与阈值差异的显著性指标，s_i^j表示主题 i 与主题 j 的相似度值，T 表示阈值，S_{max} 表示主题 i 与主题 j 所属时间窗口下所有主题间相似度的最大值。

通过以上过程，当主题间相似度值为最大且大于阈值时，显著性指标为 1，前文所述“相似度值最高且高于某一阈值的主题”可以变换为寻找显著性指标值为 1 的主题。

本文将相似度阈值设置为 0.97，计算不同年份下不同主题间的相似度显著性值，将显著性指标为 1 的主题划分为学科领域内的同一个主题。图 3 展示了 2008 年主题与 2009 年主题之间相似度显著性指标热力图，图中坐标轴以“年份”+“主题标签”命名。

根据相似度显著性指标计算结果，本文对所有两两主题之间显著性指标为 1 的主题进行统计，总结得到农业领域期刊论文数据 2008-2018 年的领域主题共 15 个，如表 2 所示。

3.5 前沿主题识别指标计算

通过上一小节的领域主题抽取结果，结合每个主题在不同年份的相关文献和相关文献指数计算方法，得到每个领域主题在不同年份的相关文献指数如表 3 所示。表中每个主题第一个非零相关文献指数对应的年份为该主题第一次出现的年份。

doi:10.3772/j.issn.2095-915x.2021.05.009

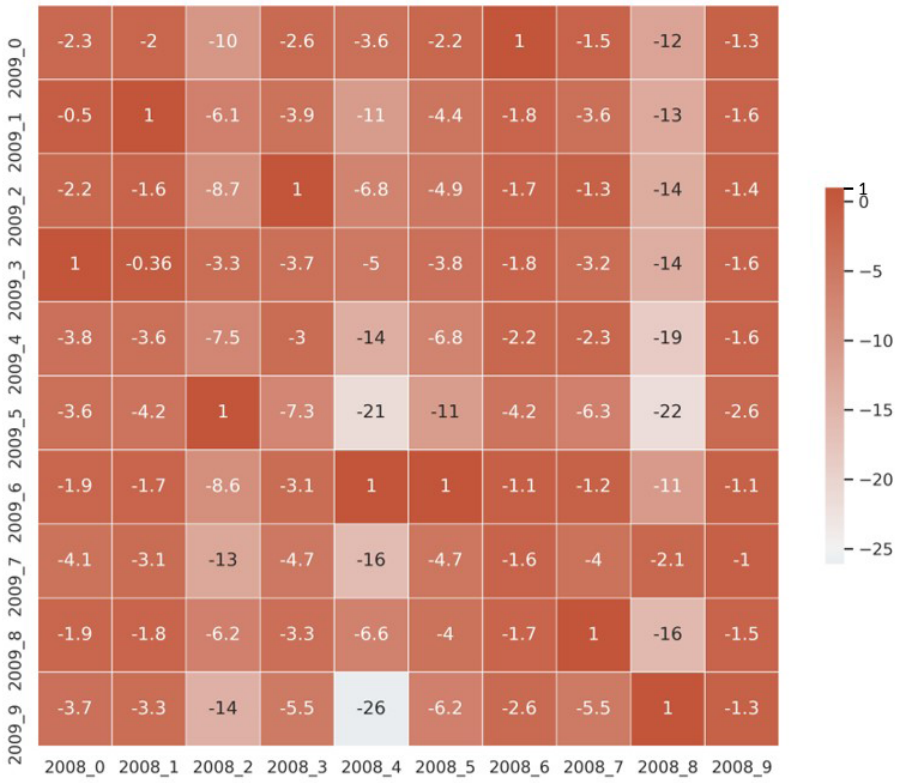

图 **3** 期刊论文主题相似度计算结果（部分）

表 **2** 农业领域科学论文研究主题（**2008-2018** 年）

主题序号	主题内容
1	家禽饲养研究
2	基因技术应用于牲畜相关研究
3	鱼类培育研究
4	防治病虫害相关研究
5	豆类植物优质培育研究
6	水稻等粮食作物的栽培研究
7	果树培育及果实品质提升相关研究
8	植株的栽培与育苗研究
9	林业资源保护研究
10	土壤侵蚀及防止水土流失相关研究
11	细菌真菌的培养及抑制相关研究
12	森林群落保护及多样性研究
13	瓜果培育及提高产量相关研究
14	茶叶种植及园林设计
15	草场保护及防止牧场退化相关研究

表 **3** 期刊论文主题相关文献指数表

年份	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	**0.98**	1.17	0.97	1.17	0.96	0.95	0.78	1.12	1.41	1.51	0.87
2	**1.09**	1.15	0.85	1.63	1.21	0.98	2.07	1.21	1.38	1.56	1.45
3	**1.74**	1.07	1.83	0.86	0.72	1.32	0.92	0.88	1.20	0.77	1.23
4	**0.51**	0.90	0.75	0.72	1.12	0.36	0.54	0.00	0.00	0.00	0.00
5	**0.55**	0.00	0.00	0.00	0.48	0.34	0.00	0.41	0.00	0.00	0.00
6	**1.22**	1.01	0.83	0.84	0.00	0.32	0.54	0.87	0.65	0.89	0.55
7	**0.62**	1.00	0.73	0.00	1.19	0.52	1.07	1.37	1.20	1.15	0.99
8	**1.82**	0.86	1.30	0.71	2.12	0.00	0.65	1.21	1.09	0.79	1.00
9	**1.01**	0.88	0.85	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	**0.46**	0.00	1.23	0.00	0.52	0.00	1.07	1.44	0.00	0.00	0.00
11	0.00	**0.71**	0.67	0.92	0.00	0.31	0.00	0.00	0.56	0.00	0.00
12	0.00	**1.26**	0.00	1.24	0.00	4.59	1.74	1.12	1.14	0.98	1.18
13	0.00	0.00	0.00	0.00	**0.34**	0.31	0.00	0.38	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.48**	0.68	0.50
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.65**	1.58

根据上表计算得到的领域主题相关文献指数，结合主题发展态势指标，计算出每个领域主题的主题发展态势指数如表 4 所示。

表 **4** 期刊论文主题发展态势指数表

年份	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	0.98	1.07	1.02	1.09	1.02	0.99	0.88	1	1.2	1.35	1.11
2	1.1	1.125	0.99	1.31	1.26	1.12	1.56	1.39	1.38	1.47	1.46
3	1.73	1.4	1.61	1.23	0.98	1.15	1.03	0.95	1.08	0.92	1.8
4	0.51	0.7	0.72	0.72	0.92	0.64	0.59	0.3	0.15	0	0
5	0.54	0.27	0.135	0.07	0.28	0.31	0.15	0.28	0.14	0	0
6	1.2	1.1	0.96	0.9	0.45	1.2	0.88	0.88	0.77	0.83	0.69
7	0.62	0.81	0.77	0.385	0.8	0.66	0.86	1.16	1.16	1.15	1.07
8	1.82	1.34	1.32	1.01	1.56	0.78	0.72	0.96	1.03	0.91	1
9	1	0.94	0.89	0.9	0.45	0.225	0.1	0	0	0	0
10	0.46	0.23	0.73	0.36	0.44	0.22	0.65	1.04	0.52	0.26	0.13
11	0	0.71	0.69	0.81	0.4	0.35	0.175	0.09	0.323	0.16	0
12	0	1.26	0.63	0.94	0.57	2.575	2.16	1.64	1.39	1.18	1.18
13	0	0	0	0	0.34	0.32	0.16	0.27	0.13	0	0
14	0	0	0	0	0	0	0	0	0.48	0.58	0.54
15	0	0	0	0	0	0	0	0	0	0.65	1.12

doi:10.3772/j.issn.2095-915x.2021.05.009

根据主题发展态势指数表，绘制领域内所有主题的发展态势折线图，如图 4 所示。

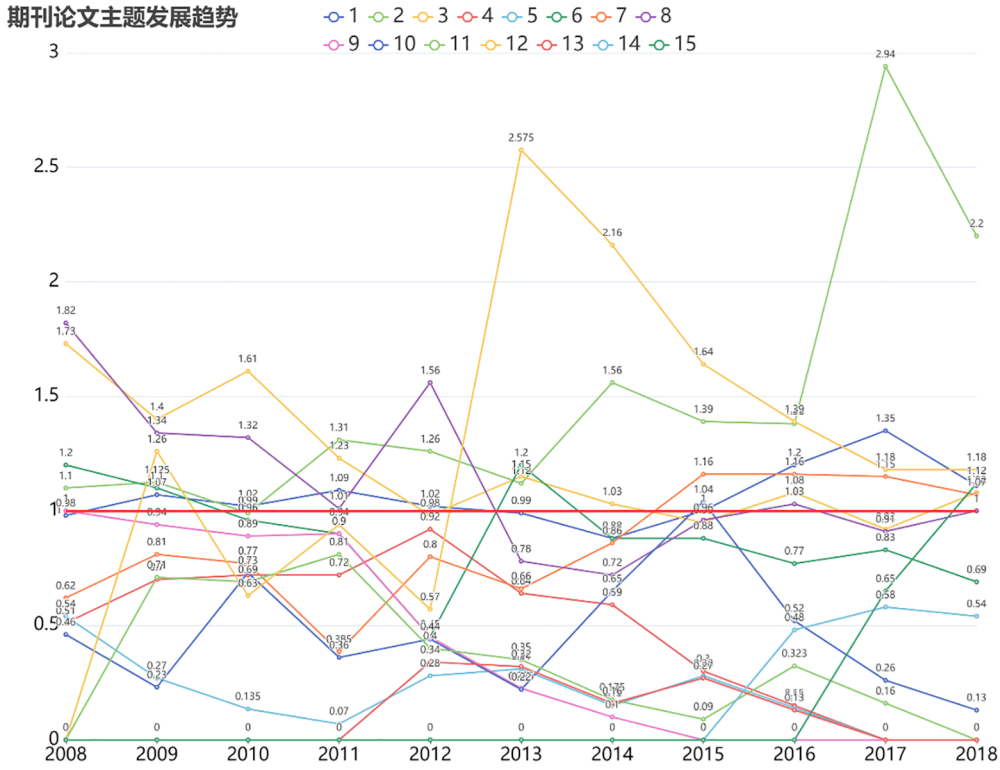

图 4 科学论文主题发展趋势

3.6 前沿研究主题识别结果分析

通过对农业领域期刊论文的文档内容进行主题识别，进而根据前文计算的新颖度、相关文献指数、主题发展指数，按照图 2 给出的识别逻辑，识别出农业领域前沿研究主题。在识别过程中，本文将时间窗口按时间顺序进行排列，划分为两个时期，主题首次出现年份落在前 50% 时期的标记为“新颖度低”，反之则为“新颖度高”；计算每个主题的主题发展指数拟合线性方程，方程斜率大于或等于 0 表示主题上升或稳定型发展，反之则为衰退型发展；将初始相关文献指数大于 1 的主题标记为“初始热门主题”，将近期相关文献指数小于等于 1 的主题标记为“冷门主题”，即该类主题近期研究热度较低。根据上述标记方法，对农业领域期刊论文主题前沿指标进行计算，结果总结如表 5 所示。

根据上述结算结果，本节对探测得到的各类前沿研究主题进行分析：

（1）潜在前沿研究主题

潜在前沿研究主题为主题 15，该主题相关前沿指数如图 5 所示。

主题 15 首次出现年份为 2017 年，属于近期阶段，新颖度较高。该主题相关文献指数在 2018 年达到了 1.58，显著高于标准值，同时该主题一经出现便有着显著的上升趋势，因此属于潜在前沿研究主题。

从研究内容看，主题 15 主要研究内容为“草场保护及防止牧场退化”，这与近年来国家强

调的“绿水青山就是金山银山”较为符合。在国家大力倡导环境保护的情况下，该主题所代表的研究方向成为了一个潜在研究主题，具有良好的研究前景。

表 5 农业领域期刊论文前沿指标计算结果

主题	新颖度	线性方程斜率	初始相关文献指数	近期相关文献指数	识别结果
1	低	0.0181	0.98	0.87	非前沿研究主题
2	低	0.0437	1.1	1.45	热门前沿研究主题
3	低	-0.0661	1.73	1.23	非前沿研究主题
4	低	-0.0748	0.51	0.00	非前沿研究主题
5	低	-0.0316	0.54	0.00	非前沿研究主题
6	低	-0.0346	1.21	0.55	衰退前沿研究主题
7	低	0.0581	0.62	1.07	非前沿研究主题
8	低	-0.0694	1.82	1.00	衰退前沿研究主题
9	低	-0.1235	1.00	0.00	非前沿研究主题
10	低	-0.0054	0.46	0.00	非前沿研究主题
11	低	-0.0827	0.71	0.00	非前沿研究主题
12	低	0.0495	1.26	1.18	热门前沿研究主题
13	低	-0.06	0.34	0.00	非前沿研究主题
14	高	0.03	0.48	0.5	非前沿研究主题
15	高	0.47	0.65	1.58	潜在前沿研究主题

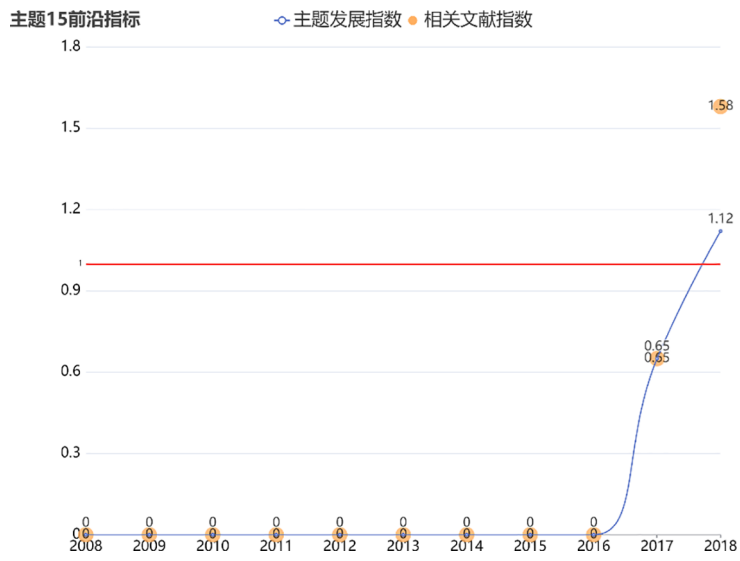

图 5 主题 15 前沿指标

（2）热门前沿研究主题

热门前沿研究主题为主题 2 和主题 12。以主题 2 为例，该主题相关前沿指数如图 6 所示。

doi:10.3772/j.issn.2095-915x.2021.05.009

图 6 主题 2 前沿指标

主题 2 新颖度较低，但除了 2010 年以外，该主题相关文献指数始终显著高于标准值，同时主题发展态势也呈上升趋势，因此将该主题划分为“热门前沿研究主题”。

主题 2 的主要研究内容为“基因技术应用于牲畜相关研究”。从研究内容看在农业及生命科学领域，基于基因技术的研究始终是研究者的工作重点，该主题将基因技术应用于牲畜，进行牲畜性状的选择，以及利用牲畜生产制造各种贵重药物，属于高科技研究，同时具有极高的经济意义，所以研究热度经久不衰。

（3）衰退前沿研究主题

衰退前沿研究主题为主题 6 和主题 8。以主题 8 为例，该主题前沿指数如图 7 所示。

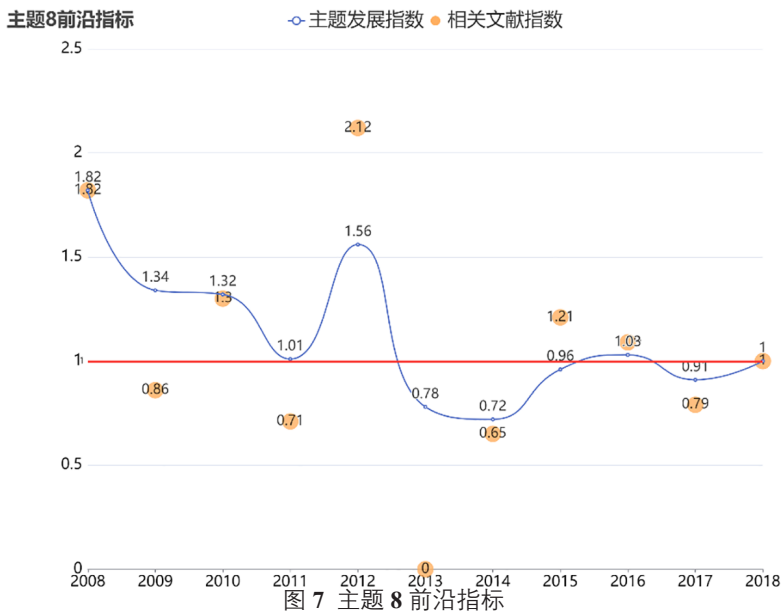

图 7 主题 8 前沿指标

该主题首次出现为 2008 年，新颖度很低，初始相关文献指数处于高水平，但随即逐年下降，主题发展态势也呈明显下降趋势，因此为“衰退前沿研究主题”。该主题的研究内容为“植株的栽培与育苗研究”。在农业领域，对于植物植株的栽培、扦插技术有着较长的研究年月，相关研究的难度不高，研究条件也十分成熟，因此已经发展成为成熟的研究主题，其研究热度逐年下降。

4 总结

围绕学科领域前沿研究主题识别这一研究主题，本文首先梳理了现有前沿研究识别方法，然后运用预训练模型、文本聚类技术、主题挖掘技术、可视化分析技术等，构建了学科领域前沿研究主题识别指标。通过主题向量构建和主题相似度计算，完成领域主题抽取和前沿研究主题识别，并利用农业领域期刊论文数据进行实证研究，识别出潜在前沿研究主题 1 个、热门前沿研究主题 2 个、衰退前沿研究主题 2 个，结果表明本文提出的前沿研究主题识别方法具有显著可行性。

然而，该方法仍然存在一定不足。首先，本方法使用了基本的中文语言模型进行文本句嵌入的构建。未来可考虑针对特定任务使用效果更佳的语言模型，或针对特定领域语料训练自有的预训练模型。其次，本方法仅使用了论文数据作为数据源，今后可考虑综合多种数据源进行前沿研究主题识别。最后，由于缺乏统一的前沿研究主题识别的评价指标体系，本文提出的前沿研究主题识别方法在结果评价方面有所欠缺，未来研究中需要进一步探讨相关评价方法。

参 考 文 献

[1] 许振亮 , 郭晓川等 . 国际技术创新研究前沿的科学计量学分析 [J]. 图书情报工作 , 2011, 55(8):49-53.

[2] 杨选辉 , 杜心雨 , 蔡志强 . 基于突变检测与共词分析的深阅读新兴趋势分析 [J]. 图书馆建设 , 2018(5):48-53.

[3] 冯佳 , 张云秋 . 基于 LDA 和本体的科学前沿识别与分析方法研究 [J]. 情报理论与实践 , 2017, 40(8):49-54.

[4] De Solla Price D J. Networks of scientific papers[J]. Science,1965, 149(3683): 510-515.

[5] Bengisu M, Nekhili R. Forecasting emerging technologies with the aid of science and technology databases[J]. Technological forecasting & social change, 2006, 73(7) : 835-844.

[6] 杨金庆 , 魏雨晗 , 黄圣智 , 等 . 基于科技文献的新兴主题识别研究综述 [J]. 情报科学 , 2020, 38(8):159-163+177.

[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] 许振亮 , 郭晓川 . 国际技术创新研究前沿的科学计量学分析 [J]. 图书情报工作 , 2011, 55(8):49-53.

[9] 潘黎 , 侯剑华 . 国际高等教育研究的热点主题和研究前沿——基于 8 种 SSCI 高等教育学期刊 2000—2011 年文献共被引网络图谱的分析 [J]. 教育研究 , 2012, 33(6):136-143.

[10] Huang M H, Chang C P. A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation[J]. Scientometrics, 2014.

[11] Kleinburge J. Bursty and Hierarchical Structure in Streams[C]. Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and

doi:10.3772/j.issn.2095-915x.2021.05.009

data mining, July 2002. ACM, 2002:91–102.

[12] 王晓光 . 科学知识网络的形成与演化 (Ⅱ): 共词网络可视化与增长动力学 [J]. 情报学报 , 2009, 28(4):599-605.

[13] 章成志 , 梁勇 . 基于主题聚类的学科研究热点及其趋势监测方法 [J]. 情报学报 , 2010, 29(2):342-349.

[14] 侯海燕 , 刘则渊 , 栾春娟 . 基于知识图谱的国际科学计量学研究前沿计量分析 [J]. 科研管理 , 2009, 30(1):164-170.

[15] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[16] 范云满 , 马建霞 . 基于 LDA 与新兴主题特征分析的新兴主题探测研究 [J]. 情报学报 , 2014, 33(7):698-711.

[17] 朱茂然 , 王奕磊 , 高松 , 等 . 基于 LDA 模型的主题演化分析 : 以情报学文献为例 [J]. 北京工业大学学报 , 2018, 44(7):1047-1053.

[18] 杨金庆 , 陆伟 , 吴乐艳 . 面向学科新兴主题探测的多源科技文献时滞计算及启示——以农业学科领域为例 [J]. 情报学报 , 2021, 40(1):21-29.

[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[20] Campello R J G B, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates[A]. In: Pei J, Tseng V S, Cao L., Motoda H., Xu G. Advances in Knowledge Discovery and Data Mining[C]. PAKDD 2013 Lecture Notes in Computer Science, 2013. Berlin: Springer, 2013.

[21] 曾海娇 , 孙巍 . 基于专利与论文关联的潜在科学前沿识别——以生物农药领域为例 [J]. 农业展望 , 2020, 16(9):93-100.