基于主题和指标特征的研究前沿识别系统设计与实现*

张辉 串丽敏 齐世杰 赵静娟 秦晓婧
(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、快速地识别研究前沿, 对于促进科技创新、攻克关键技术、推动学科发展和解决重大问题具有重要意义。以基金项目、学术论文资源为基础, 利用LDA主题模型、BERT模型、Word2Vec等方法对科技资源进行主题内容挖掘, 同时从新兴度、创新性、交叉性、关注度和中心性5个维度, 构建能够表征和识别前沿主题的指标体系, 并研发研究前沿识别与多维分析系统。该研究可为更加科学、准确、前瞻地识别科学研究前沿、分析演化路径提供具有应用价值的方法与工具。

关键词: LDA模型; 指标特征; 前沿识别; BERT模型; 演化路径
中图分类号: G353.1 DOI: 10.3772/j.issn.1673-2286.2023.07.002
引文格式: 张辉, 串丽敏, 齐世杰, 等. 基于主题和指标特征的研究前沿识别系统设计与实现[J]. 数字图书馆论坛, 2023 (7) : 9-18.

准确、快速地识别研究前沿, 对于促进科技创新、攻克关键技术、推动学科发展和解决重大问题具有重要意义。有效识别科技发展的研究前沿, 有助于科技人员、科研管理者和政策制定者准确把握科研的进展和方向, 从而将宝贵的人力、物力和财力投入最具战略意义的前沿领域, 以有限的资源来支持和推进科学进步。因此, 开展科技前沿识别研究对于洞察科研动向, 瞄准科技前沿领域, 制定科技创新发展战略规划和部署前瞻性研究方向具有重要意义。

在此背景下, 本研究立足基金项目、学术论文资源, 利用多种人工智能技术对资源内容进行深入挖掘分析, 从新兴度、创新性、交叉性、关注度和中心性5个维度构建前沿识别指标体系, 研发能够精准识别前沿主题并进行发展脉络演化分析的工具, 可为具体领域的科技创新提供重要的导向支撑。同时, 本研究也是情报学科高质量发展高度关注的重要内容。

1 研究前沿识别综述

“研究前沿”的概念由科学计量之父普赖斯[1]于1965年首次提出, 他认为活跃的研究前沿是引文网络中最近、被广泛引用的文献集合, 重点体现在核心文献的高被引程度。经过半个多世纪的发展, 国内外学者从不同角度出发, 提出了众多研究前沿识别方法, 主要分为基于专家主观判断的识别方法和基于数据客观分析的识别方法。

1.1 基于专家主观判断的识别方法

专家主观判断的识别方法以专家讨论分析为核

收稿日期: 2023-04-10
*本研究得到北京市农林科学院创新能力建设专项“智库型农业情报研究与服务能力提升项目”(编号: KJCX20230208)、“基于智能分析的作物育种关键技术识别与预测方法研究”(编号: QNJJ202308)、“基于多源数据融合的农业热点前沿主题识别与实证研究”(编号: KJCX20200403)资助。

心，涉及德尔菲调查、情景分析、现场研讨、在线讨论、通信调查、访谈等多种定性分析的方法，围绕领域研究前沿的趋势、创新点、局限性等进行定性评估，凭借专家的丰富经验和多维度的深入讨论交流，产生较为准确和具有启示性的结论和见解。众多国际组织、政府、科研管理机构、项目资助机构等经常组织相关领域的专家开展深入的研讨，通过一轮或多轮的咨询对话，获取有价值的建议和指导性意见。此外，一些权威学术网站、协会、团体或研究机构等发布的研究热点或前沿科学问题等多由领域专家遴选咨询和评议而得[2]。

基于专家主观判断的方法充分利用了领域专家的集体智慧，通过对各种不同思想、观点进行综合整理、归纳和分析来识别研究前沿。该方法发展历史悠久，已经比较成熟，是目前大多数国家和机构开展前沿科学技术识别、遴选和预测的主要方法。此类方法具有较高准确性，但是主观性强、耗时长、成本高，且受技术专家知识广度、深度的限制，容易产生预期或假设的偏差，使前沿识别难以收敛，最终导致无法有效支持决策。因此，前沿识别需要以基于数据客观分析的方法为补充。当前，随着数据密集型知识创新范式的出现和逐步深入，数据分析的作用日益凸显，已成为专家判断过程中重要的辅助方法。

1.2 基于数据客观分析的识别方法

随着科技文献的迅速积累和计算机技术的快速发展，基于数据客观分析的方法在科学研究前沿识别领域得到了广泛应用。近年来，如何快速、准确、自动地识别研究前沿是情报学领域的研究热点，现有的研究前沿识别方法可以分为基于引文、基于词语和基于主题模型三类，其原理和实现方法各异。

1.2.1 基于引文的识别方法

引文分析法是研究前沿识别中发展最早、理论基础最扎实、使用最广泛的方法，研究成果丰富。该方法通过对科技文献的引用关系进行分析，帮助人们了解研究领域内的前沿趋势和发展动态。不仅为科研人员提供了重要的科学研究参考，也为科技管理提供了决策支撑。引文分析法按引用类型，可以分为直接引用分析、共被引分析和文献耦合分析3种。每种方法都有其独特的应用场景和价值。

（1）直接引用分析。直接引用分析法由Garfield[3]于1963年提出，该方法利用科技文献中的直接引用关系，成为分析评估科学研究影响力的关键方法。Shibata等[4]使用拓扑聚类方法将引文网络划分为聚类，跟踪论文在每个聚类中的位置，并用每个聚类的特征术语构建可视化引文网络，以此检测科学出版物引文网络中的新兴研究前沿。Lu等[5]提出基于引文网络的边介数聚类方法，对旅游文献进行分组，并将每组所有文章的关键词和主路径上文章的主要主题进行整合，找出研究重点和前沿。然而，构成引文关联需要一定时间窗口以确保文章聚类效果合理与有效，因此该方法并未被广泛采用。

（2）共被引分析。基于文献共被引关系的研究前沿探测方法由Small[6]于1973年提出，该学者认为相比直接引用分析法，共被引分析更能从客观的角度体现随时间变化的科学研究能力和认知结构。White等[7]利用作者共被引分析，识别有影响力的作者及论文，并通过可视化方式揭示细粒度的学科领域研究内容。Morris等[8]通过期刊论文基础数据集，绘制可视化的时间线，以识别期刊论文中的前沿内容和发现时间变化规律等。共被引聚类是一种基于先验知识的方法，它以某个领域中的高被引论文为研究对象，并利用论文之间的共被引关系进行聚类分析。但是，如果该领域中未出现高被引论文，那么共被引聚类就无法准确地识别该领域的研究前沿。

（3）文献耦合分析。文献耦合理论的基本出发点为当两篇文献共同引用了一篇或多篇文献时，它们之间必然存在相关关系。而且，两篇文献引用的共同文献数量越多，它们之间的相关程度就越高，也就越有可能属于相同研究主题或领域。因此，文献耦合理论通过分析文献之间的共同引用关系，可以揭示文献之间的相互联系和研究热点。Glänzel等[9]利用文献耦合技术，通过设定文献数量和链接强度的阈值，筛选领域文献传播网络中的核心文献，并识别热门研究主题。Huang等[10]采用文献耦合和滑动窗口法对2000—2009年有机发光二极管（OLED）领域的研究前沿进行探索，追踪研究前沿的产生、增长、衰退和消失。基于文献耦合分析的方法能够避免引用时滞问题，但是较为依赖权威、高质量的文献，因此识别的结果多为研究热点而不是研究前沿。

1.2.2 基于词语的识别方法

针对文献中的关键词、主题词进行统计分析，利用词频分析法和共词分析法进行研究前沿识别。词频分析法基于文献计量学，通过统计分析主题词的词频或词频变化来识别前沿主题。Kleinberg[11]提出突发词检测算法，通过检测短期词频，识别突然出现的主题词，以揭示领域热点发展趋势。马红鸽等[12]利用CiteSpace工具进行词频和主题聚类分析，绘制文献关键词聚类知识图谱，探测国内外人工智能对劳动力供给影响的热点问题与研究前沿。Peters等[13]提出共词映射方法，基于对多维缩放的单词共现数据矩阵的直接应用，探索化学工程领域研究进展和重点问题。徐畅等[14]通过对关键词的共词分析、聚类分析和战略坐标图分析，总结人工智能研究的热点、潜在发展方向和研究前沿。

在实际应用中，词频分析法以词频变化分析研究前沿，而共词分析法更加注重词语间的共现强度，并且通常与聚类分析结合使用，以实现对研究前沿的识别。共词分析法具有直观、灵活等特性，因此得到广泛的应用。

1.2.3 基于主题模型的识别方法

主题分析法是利用自然语言处理技术对文本进行主题抽取的方法，通过识别文档中的主题来分析文档的语义内容。相比其他研究前沿识别方法，主题分析法能够对非结构化的文本（如摘要、全文）进行有效分析，因此在近年来备受关注。LDA（Latent Dirichlet Allocation）[15]模型是主题模型中的典型代表，利用词袋特征代表文档，不考虑文档中的单词顺序以及时间动态等问题，实现对潜在的语义关系和主题信息的挖掘。作为非监督机器学习方法，主题分析法受到学者的广泛关注与使用，如：周云泽等[16]利用LDA主题模型识别多源数据主题，用Word2Vec模型进行主题合并，同时使用主题强度、主题新颖度指标判别新兴研究技术；林晗等[17]针对LDA的拓展模型HLDA开展研究，同时结合文本相似度因子，提高前沿主题识别的准确性与可靠性；刘自强等[18]利用LDA模型进行基金、论文研究主题探测，同时综合新兴度、关注度指标和战略坐标图等，在主题扩散演化滞后效应测度结果基础上进行研究前沿识别。基于引文和词语的识别方法忽略了文本语义内容和语义关系，主题模型在很大程度上弥补了这个不足，从文本语义理解角度探测研究前沿。

综上所述，现有研究利用多种方法进行前沿主题识别，但仍存在文本语义理解深度不够、多维度指标普适性不强以及识别结果分析较为单一等问题，同时欠缺基于具体应用场景的实际应用系统设计。因此，本研究以基金项目、学术论文资源为基础，利用LDA主题模型、BERT（Bidirectional Encoder Representations from Transformers）模型、Word2Vec等方法对科技资源进行深入挖掘分析，结合研究前沿主题的共性特征，从新兴度、创新性、交叉性、关注度和中心性5个维度，构建能够表征和识别前沿主题的指标体系，以此增强识别的准确性和客观性。在此基础上，研发研究前沿识别与分析系统，探索出一套更加科学、全面、精准，具有应用性的方法与工具。

2 研究前沿识别系统设计

2.1 设计原则

根据实际情况和未来发展需要，遵循以下原则设计研究前沿识别系统。①易理解原则。系统设计的前沿识别指标体系应该结构清晰，指标描述言简意赅，易于用户理解和使用。②高效性原则。系统功能页面应该设计简约，突出数据挖掘与前沿识别的主要信息。导航系统在层次清晰的同时应该方便浏览者对相关信息和服务的访问。③可扩展性原则。系统设计应该考虑未来研究的需要，各个功能模块间的耦合度小，便于数据资源和指标项的扩展，并平滑地与其他应用系统接口对接。

2.2 系统架构

基于论文、项目等数据的研究前沿识别与分析的需求，设计和构建基于主题和指标特征的研究前沿识别系统。该系统包括基础层、数据层、分析层、应用层等4个部分，体系架构如图1所示。

基础层包括主机与服务器、网络与安全设备、存储与备份设备、高性能计算设备等硬件，是支撑系统运行与算法计算的基础设施。

数据层包括数据处理和数据存储库，为系统的应用提供数据支撑。数据处理流程涉及筛选、清洗、转换、集成等一系列操作，通过对原始数据中存在的缺失、偏差等缺陷问题进行分析与处理，完成数据清洗与规范化。数据存储库包括应用数据库和实验数据库：科

技文献数据、领域词表等数据经过预处理，存储到应用数据库；词语功能识别模型的训练数据经过预处理，存储到实验数据库。

分析层包括主题挖掘、研究前沿主题挖掘、主题演化分析、词语功能识别，是整个系统的核心单元。通过集成指标体系、数据挖掘算法与模型，实现研究前沿识别与分析。

（1）主题挖掘。基于LDA主题模型的文本挖掘算法进行主题聚类，提取表示主题的特征词。

（2）研究前沿主题挖掘。构建多维度前沿识别指标体系，设计新兴度、创新性、交叉性、关注度、中心性等5个维度，通过设计各指标项计算方法及权重分配方法，实现对前沿主题内容的识别。

（3）主题演化分析。针对前沿主题识别结果，按发表年份将主题划分到不同时间窗口下，通过对主题词的聚类及主题相似度计算，挖掘主题在不同时间下的演变和发展关系。

（4）词语功能识别。通过构建词语功能分类训练数据集，利用BERT模型对训练数据集进行训练，获取词语功能分类模型。在此基础上，利用训练好的模型对主题词、关键词进行计算推理，识别词语功能类别。

应用层通过图、表等形式展示数据分析结果，涉及主题分布、研究前沿主题发现、主题演化路径、主题词功能识别、统计分析等。

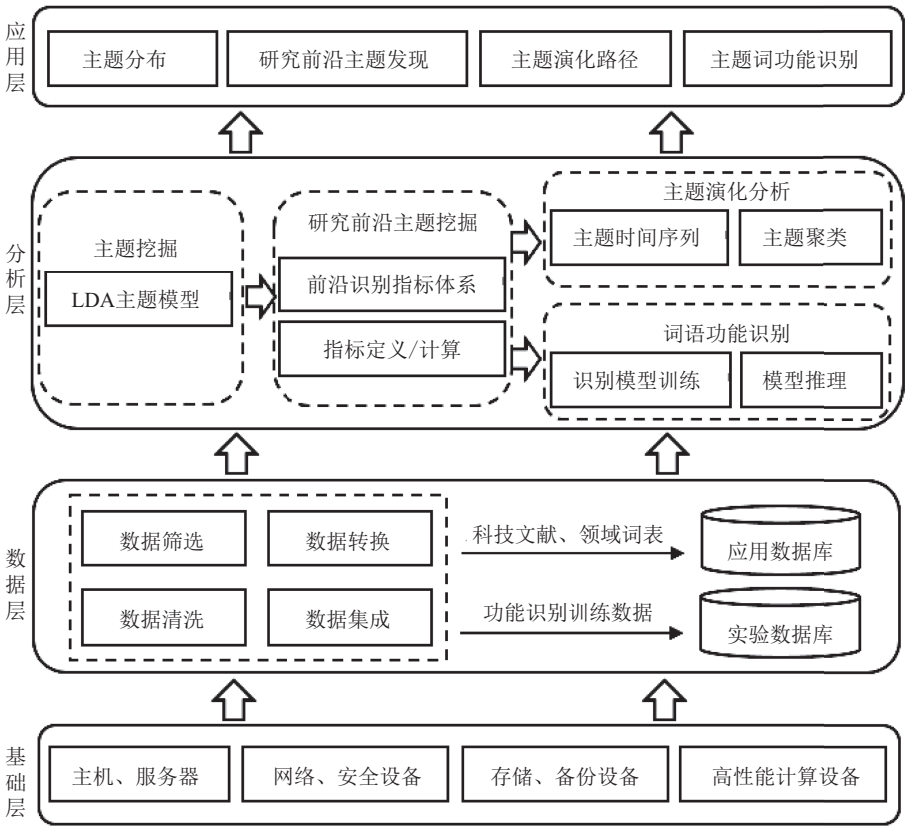


图1 研究前沿识别系统架构

3 关键技术

3.1 主题模型

主题模型是一种以非监督学习的方式对文本集的隐含语义结构进行聚类的统计模型，广泛应用于文本潜在语义关系和主题信息挖掘。LDA模型具有对海量异构文本数据建模的优势，是当下的主流方法。通过运用LDA模型，可以实现研究主题的提取。在主题提取的基础上，LDA模型进一步在科技文献知识挖掘、科学研究热点发现与主题演化分析、新兴前沿主题探测等研究方向得到了广泛的应用。因此，利用LDA模型进行研究前沿主题挖掘，其实现流程如图2所示，其中：α表示主题分布θ的先验分布参数，β表示主题词分布φ的先验分布参数；M表示文档数量，K表示主题数量；z和w分别表示LDA模型抽取生成的主题和最终确定的主题

词，N_m表示文档m的词语总数。

首先，根据概率p选择文档m，从α中抽样生成文档m的主题分布θ_m，并从θ_m中抽取文档m的主题$z_{m,n}$。其次，从β中抽样生成主题$z_{m,n}$对应的主题词分布φ_k。最后，从φ_k中抽样生成主题词$w_{m,n}$。LDA模型语料库的生成概率如式（1）所示。

$$p(w,z|\alpha,\beta)=\prod_{k=1}^{K}\frac{\Delta(\varphi_k+\beta)}{\Delta\beta}\prod_{m=1}^{M}\frac{\Delta(\theta_m+\alpha)}{\Delta\alpha} \quad (1)$$

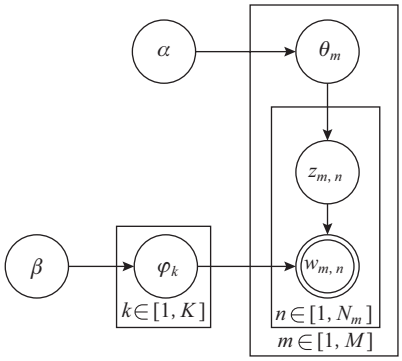


图2 LDA模型结构

利用LDA主题模型进行主题挖掘时，需要确定数据集包含的主题数量，采用困惑度指标P确定最优主题数量。困惑度表示文档所属的主题的不确定性：当困惑度曲线逐渐趋于稳定状态时，主题数量取值为最优主题数量。其计算公式如式（2）所示。

$$P(D)=\exp\left[-\frac{\sum_{m=1}^{M}\log p(w_m)}{\sum_{m=1}^{M}N_m}\right] \quad (2)$$

式中：D代表包含M篇文档的测试数据集；$p(w_m)$代表测试数据集中每一个主题词出现的概率。

3.2 前沿识别指标

在前人研究成果的基础上，构建一套完善的前沿识别指标体系，以科学、有效地衡量主题的前沿性。新兴度、创新性、交叉性、关注度和中心性是从项目和论文中均可提取的能够表征前沿主题的5个共性特征。基金项目是围绕前瞻性科技规划组织实施的研究任务，体现了当前及今后一定时期的具体部署，是科学研究关注的焦点。论文是科技创新研究成果的载体。项目和论文是首选的前沿主题识别的基础信息（数据）来源。针对项目和论文两种数据源，同一维度的评价指标项略有不同，各指标具体含义和计算方法如下。

（1）新兴度。新兴度是指研究主题新颖性。从时间维度上看，越晚出现的主题越容易包含最新的研究内容，也更有可能成为研究前沿。

①对项目数据，用主题中项目的平均立项年份来表征新兴度，计算方法为某主题中所有项目的立项年份之和除以项目数量。

②对论文数据，根据论文平均出版年份、参考文献平均出版年份和施引文献平均出版年份进行新兴度评价，具体计算方法为某主题中所有论文的发表年份之和除以论文数量。

（2）创新性。创新性是指内容创新性，突出强调研究内容的突破性与引领性。主题下的主题词越新颖，主题的创新性越高。因此，通过检测主题词成为突发词概率，判断主题创新性。采用Kleinberg突发词检测算法，使用无限状态自动机对时间序列数据进行建模，时间序列数据状态的转变标志着突发事件的出现，主题创新性为突发词概率值之和。

（3）交叉性。交叉性是指学科交叉的广度。多学科的交叉促使科学研究成果跨领域应用，使创新性影响的出现概率加大。项目在申报时会标注所属研究领域的信息，论文在发表时会标注研究领域。因此，主题内平均每个项目（或者每篇论文）包含的研究领域数量可以表征主题交叉性，交叉性计算方法为某主题下所有项目（论文）的研究领域数量除以项目（论文）数量。

（4）关注度。关注度是指研究内容在时间跨度内的受关注程度，受关注程度高的内容能够代表当前阶段领域发展的水平，或能够影响领域未来的发展趋势。因此，针对主题的关注度进行度量，分析主题的前瞻性。

①对项目是否受到关注和支持，可以用项目获得的资助时间长短、获得的资金数额高低、项目数量和经费的变化幅度以及项目数量的占比来反映。因此，主要用以下4个指标来衡量项目关注度。

平均资助时长：项目所有资助时长之和除以项目数量。

平均资助强度：项目所有资助经费之和除以项目数量。

平均增长率：$0.5x+0.5y$，其中x为项目数量年度增长率，y为项目经费年度增长率。

主题强度：某一主题的项目数量之和除以所有主题项目数量。

②对论文，可用平均被引频次、论文数量的变化幅

度以及论文数量的占比来反映论文关注度，主要用以下3个指标来衡量。

平均被引频次：某一主题论文的总被引频次除以该主题的论文数量。

平均增长率：主题论文数量年度增长率之和除以统计时间（年）。

主题强度：某一主题论文数量除以所有主题论文数量。

（5）中心性。中心性可度量节点在网络中的重要性。选取主题词作为节点，通过评价核心节点，揭示主题研究热点。某主题下主题词共现网络的网络聚类中心度越高，则主题中心性越高。具体计算方法为：首先，获取一个主题的所有主题词；其次，统计两个主题词在同一篇论文中的出现频次，构建主题词共现矩阵，并以共现矩阵中的词语为节点、以词语共现频次为边的权重构建无向图；最后，对无向图进行迭代计算，得到主题网络聚类中心度。

3.3 词语功能识别方法

词语功能即词语在特定上下文语境中承载的语义功能，常见的词语功能包括问题、方法、对象、数据、指标、工具等[19]。主题词、关键词是主题分析的核心词语，通过对其进行语义功能识别，能够从语义层面捕获更为精细的主题间差异，揭示不同研究主题背后的知识交叉与融合情况，辅助分析主题的演化路径及未来的发展趋势。

研究对象、问题、方法是体现研究内容价值的重要功能元素。因此，选择设定研究对象、设定研究问题、设定研究方法和其他4种功能类型。同时，核心词语多由单词或短语组成，因此，将词语功能识别转为短文本分类任务，构建基于BERT向量表示[20]的词语功能识别模型，在理解语义的基础上实现主题词、关键词词语功能判定。BERT模型结构如图3所示，其中：T表示Transformer输出结果，Trm表示双向Transformer编码器，$\boldsymbol{E}$表示编码向量。

（1）词语向量表示。使用BERT模型进行词语向量表示，将全部单词、短语看作短文本形式，具体流程如图4所示。首先，对短文本进行预处理，操作包括词性还原、单复数转换、连接词切分、分词等。其次，使用BERT模型对短语进行特征提取，获得短语的特征向量表示，对短语中每个单词的特征向量进行拼接。最后，得到短文本向量表示。

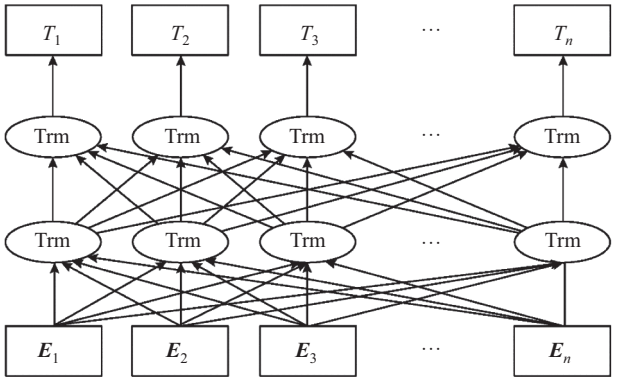


图3 BERT模型结构

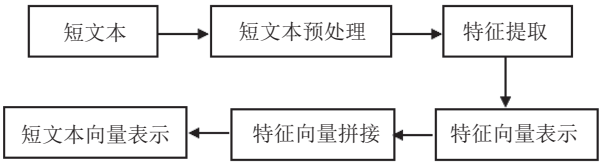


图4 短文本向量表示过程

（2）基于BERT的词语功能识别算法，得到词语功能识别算法实现步骤如下。

算法输入：初始短文本训练集$\boldsymbol{T}=\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$，$i$=1、2、…、$N$。其中：$\boldsymbol{x}_i$代表第$i$个训练样本内容对应的短文本向量，$y_i$代表第$i$个训练样本内容对应的分类类别。

算法输出：文本分类类别。

第1步：通过预处理训练集$\boldsymbol{T}$，得到处理结果$\boldsymbol{T}'=\{(\boldsymbol{x}_1', y_1'), (\boldsymbol{x}_2', y_2'), \cdots, (\boldsymbol{x}_N', y_N')\}$。

第2步：基于BERT模型对训练集$\boldsymbol{T}'$进行模型微调，通过模型输出得到$\boldsymbol{T}'$对应的特征表示$\boldsymbol{V}=(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N)$。其中：$\boldsymbol{v}_i$代表$\boldsymbol{x}_i'$对应的特征向量。

第3步：将第2步得到的特征表示$\boldsymbol{V}$输入Softmax回归模型，并进行模型训练，最终输出文本对应的类别。

4 研究前沿识别系统实现

围绕前沿主题挖掘与分析的实际需求，研究前沿识别系统可实现主题聚类、前沿识别、主题演化分析、词语功能识别、主题统计分析等主要功能。系统开发环境为Ubuntu 18.04、Python 3.8，开发工具为PyCharm、WebStorm，采用MongoDB数据库进行数据存储。选择农业资源环境领域相关数据作为实验数据：选择美国农业部国家粮食与农业研究所（NIFA）的基金项目数据，通过分类维度筛选相关项目，开展实证研

究；以Web of Science核心合集为基础，通过学科分类提取论文数据。

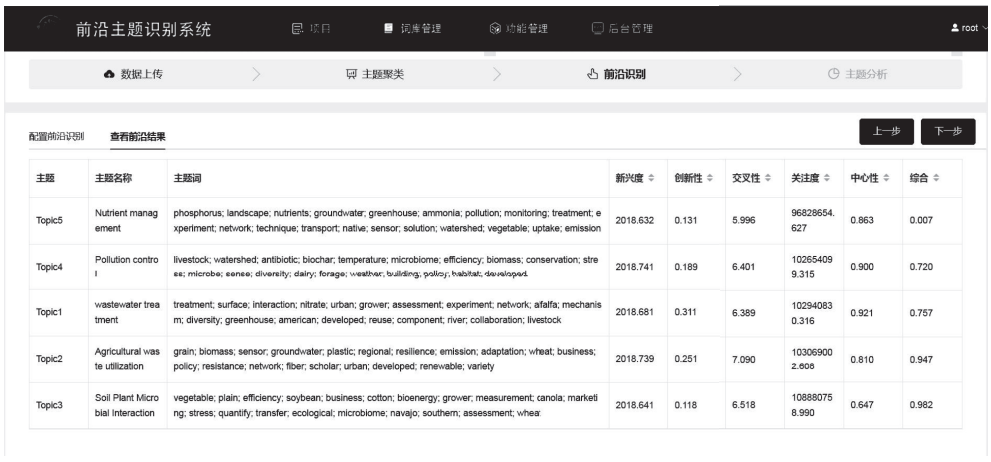
4.1 主题聚类

主题聚类模块主要包括配置主题聚类参数、查看主题结果和主题分布。①配置主题聚类参数是LDA模型求解中较为关键的步骤，用户可选择文本标题、摘要、关键词等3种类型进行文本主题聚类，并提供两种模式（固定模式、区间模式）进行主题数量配置和最佳主题数量确定。同时，用户根据实际需求配置输出结果中的主题词数量，并通过更改模型训练频次调整模型性能和准确率。②查看主题结果。用户可查看每个主题下的主题词及概率，同时可对主题进行命名以及导出保存主题内容。③查看主题分布。基于pyLDAvis工具包构建Web交互式主题可视化模块，供用户直观分析主题分布情况。同时，通过手动调整lambda参数，可查看主题特有关键词和高频关键词，进一步分析主题突出的内容。如图5所示，针对上传到系统内的2016—2021年美国农业资源环境领域SCI论文进行主题聚类参数配置，聚类计算完成后，可查看主题最佳聚类结果与主题分布。

图5 主题聚类参数配置

4.2 前沿识别

基于前沿识别指标体系和主题对应的原始数据集，通过设定各指标权重及指标计算方法，系统对主题内容进行判断，按照新兴度、创新性、交叉性、关注度、中心性5个维度以及综合得分进行展示，用户根据计算结果分值选取相应内容进行对前沿主题的深入分析与解读。如图6所示，在主题聚类结果基础上，通过配置前沿指标项，系统自动计算得出5个主题的指标分值，并展示主题基本信息。



主题	主题名称	主题词	新兴度	创新性	交叉性	关注度	中心性	综合
Topic5	Nutrient management	phosphorus; landscape; nutrients; groundwater; greenhouse; ammonia; pollution; monitoring; treatment; experiment; network; technique; transport; native; sensor; solution; watershed; vegetable; uptake; emission	2018.632	0.131	5.996	96828654.627	0.863	0.007
Topic4	Pollution control	livestock; watershed; antibiotic; biochar; temperature; microbiome; efficiency; biomass; conservation; stress; microbe; sense; diversity; dairy; forage; weather; building; policy; habitat; developed	2018.741	0.189	6.401	10265409.315	0.900	0.720
Topic1	wastewater treatment	treatment; surface; interaction; nitrate; urban; grower; assessment; experiment; network; alfalfa; mechanism; diversity; greenhouse; american; developed; reuse; component; river; collaboration; livestock	2018.681	0.311	6.389	10294083.0.316	0.921	0.757
Topic2	Agricultural waste utilization	grain; biomass; sensor; groundwater; plastic; regional; resilience; emission; adaptation; wheat; business; policy; resistance; network; fiber; scholar; urban; developed; renewable; variety	2018.739	0.251	7.090	10306900.2.606	0.810	0.947
Topic3	Soil Plant Microbial Interaction	vegetable; plain; efficiency; soybean; business; cotton; bioenergy; grower; measurement; canola; marketing; stress; quantify; transfer; ecological; microbiome; navajo; southern; assessment; wheat	2018.641	0.118	6.518	10888075.8.990	0.647	0.982

图6 前沿识别结果

4.3 主题演化分析

主题演化分析描述主题随时间的发展过程，展示不同时间下主题之间的关联性，并从中发现主题之间的演变关系。系统提供主题词、主题关键词（原始数据集中的关键词字段内容）两种内容形式的主题演化分析，处理流程如下。

（1）根据主题聚类结果，将主题下原始数据集按时间切片划分，一个主题可能会出现在多个时间切片中。

（2）分别对每个时间切片下的主题词、关键词进行聚类，每个聚类表示为演化图中的一个节点，节点名称为聚类中心词。

（3）计算相邻时间切片内两两主题之间的余弦相似度，并以余弦相似度表示主题演化的概率值。

（4）设置主题演化概率值的阈值，在主题演化路径中仅保留概率值大于阈值的边。

（5）根据时间切片、主题节点、主题演化概率值，绘制主题演化路径图。

系统采用K均值聚类机器学习算法（KMeans）对主题词进行聚类。首先，采用Word2Vec预训练模型获取主题词的向量表示；其次，将上述向量表示作为KMeans的输入，并进行聚类学习，选择最佳聚类模型；最后，获取最佳聚类模型中的聚类中心及聚类结果，并使用聚类中的词语向量的平均值表示当前聚类向量，使用聚类中心词表示当前聚类名称。

如图7、图8所示，用户配置需要分析的主题及数据，系统计算生成主题演化结果，并采用河流图形式进行可视化展示。

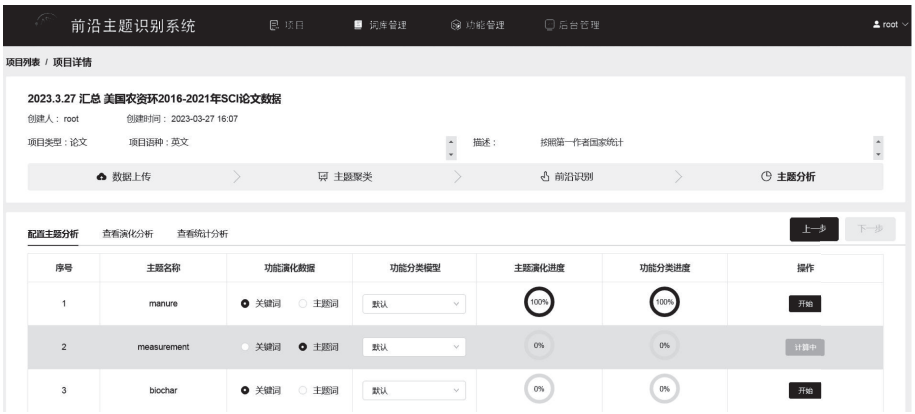


图7 主题演化分析分类配置

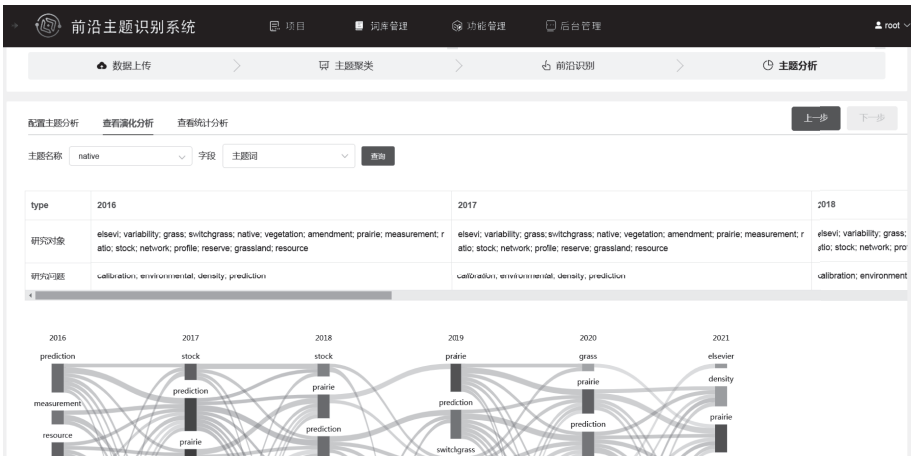


图8 主题演化分析分类结果

4.4 词语功能识别

将用户上传的功能词标注数据作为训练语料，对功能识别模型进行训练，并保存模型。在训练参数设定上，将BERT模型最大序列长度设置为512。[CLS]向量后接入的全连接层用于计算Sigmoid函数，设置

MLP隐层向量为128。模型微调训练时，batch_size设置为20，学习速率设置为0.000 1，warm_up比例设置为0.1；使用偏移修正的Adam优化算法以使微调过程更快收敛并获得更小的loss；训练过程中通过调整训练步数来提升模型学习效果，最终设置学习步数为10 000步。BERT网络共12层，其中底层的网络往往学习到比较通用的特征信息，而顶层的网络一般会学习到具体下游任务的特征信息。因此，在模型微调训练时，保留底部1~3层的BERT权重参数；对于顶部4~12层的BERT权重参数，使用预训练参数进行初始化并进行训练学习。

模型推理过程中，用户选择已保存的训练模型对当前主题下的主题词或关键词进行模型推理。首先，加载保存的模型；其次，加载主题数据（主题词或关键词），并对数据进行预处理和向量表示；最后，将准备好的数据输入模型进行计算，得到功能分类结果。

5 结论与展望

设计与实现的基于主题和指标特征的研究前沿识别系统是一个功能多样、流程化、可扩展的系统。具有以下优势。①系统功能丰富多样，操作简单易上手。系统提供数据加工与展示、数据挖掘、语料个性化配置、分析结果可视化展示等功能，可满足不同领域人员进行前沿主题识别与分析的需求。同时，系统页面设计简单整洁，功能模块划分清晰，易于用户操作使用。②系统流程化设计，形成完整数据挖掘体系。系统按照数据源编辑、主题挖掘、前沿识别、前沿主题分析的业务流程，按步操作、层次清晰，形成一套完整的前沿主题识别与分析流程体系。③系统可扩展性较好，易于后期功能升级。系统设计的各个功能模块间的耦合度小，易于后期增加分析所需的数据资源类型和前沿识别指标项等，方便系统功能的完善与提升。

综上，本研究利用LDA主题模型、前沿识别指标体系、BERT模型等技术与方法设计实现研究前沿识别系统，该系统能够较好地满足领域研究前沿主题识别与深度分析的需求。未来，系统将在数据挖掘模型优化、数据资源类型扩充、前沿识别指标体系完善等方面进行改进和丰富，为研究领域的发展趋势研判、战略部署、科技创新等提供重要支撑。

参考文献

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

[2] 曾文，李辉，李荣. 基于多源数据的前沿科学领域与新兴研究方向识别和遴选方法研究[C]//2018年北京科学技术情报学会学术年会——智慧科技发展情报服务先行论坛论文集. 2018：2-10.

[3] GARFIELD E. Citation indexes in sociological and historical research[J]. American Documentation，1963，14（4）：289-291.

[4] SHIBATA N，KAJIKAWA Y，TAKEDA Y，et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications[J]. Technovation，2008，28（11）：758-775.

[5] LU L Y Y，LIU J S. A novel approach to identify research fronts of tourism literature[C]//2015 Portland International Conference on Management of Engineering and Technology（PICMET），August 2-6，2015，Portland，OR，USA. New York：IEEE Press，2015：2211-2217.

[6] 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.

[7] WHITE H D，MCCAIN K W. Visualizing a discipline：an author co-citation analysis of information science，1972-1995[J]. Journal of the American Society for Information Science，1998，49（4）：327-355.

[8] MORRIS S A，YEN G，WU Z，et al. Time line visualization of research fronts[J]. Journal of the American Society for Information Science and Technology，2003，54（5）：413-422.

[9] GLÄNZEL W，CZERWON H J. A new methodological approach to bibliographic coupling and its application to the national，regional and institutional level[J]. Scientometrics，1996，37（2）：195-221.

[10] HUANG M H，CHANG C P. Detecting research fronts in OLED field using bibliographic coupling with sliding window[J]. Scientometrics，2014，98（3）：1721-1744.

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

[12] 马红鸽，莫正晖. 人工智能对劳动力供给的影响问题研究：基于Cite Space科学知识图谱（2010—2020）[J]. 重庆理工大学学报（社会科学），2022，36（4）：118-132.

[13] PETERS H P F, VAN RAAN A F J. Co-word-based science maps of chemical engineering. part I: representations by direct multidimensional scaling[J]. Research Policy, 1993, 22(1): 23-45.

[14] 徐畅, 管开轩, 宋昱晓, 等. 文献计量视角下全球人工智能领域研究态势与热点分析[J]. 科技促进发展, 2021, 17(11): 1968-1977.

[15] BLEI D M, NG A, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(4/5): 993-1022.

[16] 周云泽, 闵超. 基于LDA模型与共享语义空间的新兴技术识别: 以自动驾驶汽车为例[J]. 数据分析与知识发现, 2022, 6(S1): 55-66.

[17] 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.

[18] 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.

[19] 程齐凯. 学术文本的词汇功能识别[D]. 武汉: 武汉大学, 2015.

[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2023-03-01]. https://arxiv.org/abs/1810.04805.

作者简介

张辉，男，硕士，助理研究员，研究方向：知识组织与数据挖掘。
串丽敏，女，博士，副研究员，通信作者，研究方向：农业科技情报，E-mail：chuanlm@agri.ac.cn。
齐世杰，女，硕士，助理研究员，研究方向：农业科技情报。
赵静娟，女，硕士，副研究员，研究方向：农业科技情报。
秦晓婧，女，硕士，助理研究员，研究方向：农业知识管理与情报研究。

Design and Implementation of a Research Frontier Identification System Based on Theme and Index Characteristics

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

Abstract: Accurately and quickly identifying research frontiers is of great significance for promoting technological innovation, grasping key technologies, promoting disciplinary progress, and solving major problems. Based on fund projects and academic paper resources, this paper uses the LDA topic model, BERT model, Word2Vec, and other methods to mine the subject content of scientific and technological resources. At the same time, from the five dimensions of emerging, innovative, interdisciplinary, attention, and centrality, this paper constructs an index system that can characterize and identify cutting-edge themes, and develops a cutting-edge identification and multidimensional analysis system as well. This study provides valuable methods and tools for identifying scientific research frontiers, analyzing evolutionary paths, and providing more scientific, accurate, and forward-looking insights.

Keywords: LDA Model; Index Characteristic; Frontier Identification; BERT Model; Evolutionary Path

（责任编辑：王玮）