一种基于 LDA 模型的新兴主题识别与探测方法

吴东雪[a],沈桂兰[b]

(北京联合大学 a.应用文理学院;b.商务学院,北京 100191)

摘 要:新兴主题识别是科技研究领域识别新兴技术的重要方式,高效精准地识别新兴主题是早期辨识新兴技术研究方向的前提.提出一种基于 LDA 模型的新兴主题识别与趋势预测方法,通过 LDA 模型提取科技文献中的研究主题,构建主题强度、主题新颖度和复合主题关注度的指标体系识别新兴主题,采用 Prophet 模型预测新兴主题的主题强度,探测未来发展趋势.以智慧农业领域最近 14 年的科研文献为数据集,对提出的识别和探测方法进行验证,识别出了 5 个新兴主题,并预测了未来 3 年的发展趋势,同时验证所提方法的有效性.

关键词:主题识别;最优主题数;新兴主题识别指标;Prophet 模型

中图分类号:TP399 **文献标志码**:A **文章编号**:1000-2367(2024)02-0072-09

随着新一轮科技革命到来,产业急剧变革,全球技术竞争日趋激烈.新形势下,如何选择国家的科技发展战略,确定重点发展领域是各国政府面临的重要问题.掌握科学前沿理论和技术发展动态,可以在有限的资源支持下高效地推动科学技术进步.新兴主题是新出现的一组由多个关键词或词组表示的主题领域簇,代表着科学研究中极具发展潜力的研究方向或趋势[1].识别新兴主题是对科学前沿、技术前沿以经济和社会发展导向进行战略性探索的有效手段,新兴主题识别和预测方法的研究已经成为研究者关注的热点.

早期主要是通过科技文献的引文网络分析或关键词分析进行新兴主题识别[2-5],但引文网络分析时间滞后、缺少动态更新,关键词分析主题相对片面、缺少语义解读,识别出来的新兴主题往往准确性不高、可解释性不强,且无法预测新兴主题未来发展趋势.近年来,以潜在狄利克雷分布模型(latent dirichlet allocation,LDA)为代表的主题建模及其改进方法[6-8],以概率形式从科技文献中抽取大量主题信息,能高效快捷地挖掘科学领域中的主题方向而广受关注.周云泽等[9]对自动驾驶领域的专利和论文文献进行 LDA 主题建模识别该领域的新兴技术,吴胜男等[10]提出了 Co-LDA 主题模型和链路预测相结合的方法预测核心主题关联机会,张振青等[11]采用 PhraseLDA 模型对领域学科交叉主题进行识别,ALATTAR 等[12]将 LDA 模型看作一个过滤器,用按时间戳识别删除旧主题,保留新主题的方法识别新兴主题.但是上述研究忽略了 LDA 模型中最优主题数的设定,识别出主题往往存在较高的同质性,而且这些研究也未对识别出的新兴主题的未来发展趋势进行预测.另外,进行新兴主题识别时,还应该关注新兴主题的关键特征.新兴主题具有概念新、影响力大、增长快、成长潜力大的特点,PORTER 等[13]根据这些特点构建了新兴主题指标识别体系,但是缺乏针对 LDA 主题概率的指标量化表示.

收稿日期:2022-12-21;**修回日期**:2023-02-23.

基金项目:国家社科基金(21BXW057);北京联合大学人才强校优选项目(BPHR2019CZ03).

作者简介:吴东雪(1995-),女,河北唐山人,北京联合大学硕士研究生,研究方向为文本挖掘、竞争情报分析,E-mail:dongxue_wu@163.com.

通信作者:沈桂兰(1979-),女,河南信阳人,北京联合大学副教授,博士,研究方向为社交网络分析、数据挖掘研究,E-mail:guilan.shen@buu.edu.cn.

引用本文:吴东雪,沈桂兰.一种基于 LDA 模型的新兴主题识别与探测方法[J].河南师范大学学报(自然科学版),2024,52(2):72-80.(Wu Dongxue,Shen Guilan.An emerging topic identification and detection method based on LDA model[J].Journal of Henan Normal University(Natural Science Edition),2024,52(2):72-80.DOI:10.16366/j.cnki.1000-2367.2022.12.21.0001.)

为提升新兴主题的识别与趋势分析的准确性与有效性,本文提出一种基于 LDA 主题模型的新兴主题识别与探测的方法,主要贡献包括:(1)改进模型评估方法,优化 LDA 模型主题抽取结果;(2)构建新兴主题识别指标体系,设计适用于主题概率模型的指标量化计算方法;(3)在指标量化基础上,使用先知神经网络 Prophet 模型对新兴主题未来三年的趋势发展预测;以智慧农业领域科技文献为数据集进行仿真实验,验证了方法识别和预测的有效性和准确性.

1　基于 LDA 模型的主题抽取

LDA 是词-主题-文档三层结构的经典概率生成模型[14],以"主题"为语义中介,将词与文档连接起来,认为每一个文档集都是一组潜在主题的集合,其原理如图 1 所示.LDA 通过引入 α、β 两个参数作为 Dirichlet 分布的超参数,分别生成主题的多项分布 Θ 和词的多项分布 φ,对于文档集 M 中的一篇有 N 个词的文档 m,主题分布 Θ 中的主题 Z 以某个概率选定了这个文档,同时从主题 Z 对应的词分布 φ 中选中某个概率词 W,过程重复 N 次,就产生了文档 m.在这个过程中,通过文档-主题概率分布和主题-词概率分布实现词-主题-文档的语义结构关联.

LDA 模型的概率主题分布可以抽取科技文献中潜在主题信息,其中最优主题数目的确定对主题抽取至关重要.困惑度是经典的最优主题数目判定指标,该指标关注模型泛化能力,实际应用中存在提取主题数量较大,主题间相似度高的问题.为平衡模型的泛化能力以及主题抽取效果,提出了 Perplexity-Var(*P-V*)指标,在进行困惑度计算的基础上辅以主题方差、复合困惑度和主题相似度来确定最优主题数[15],计算公式如下:

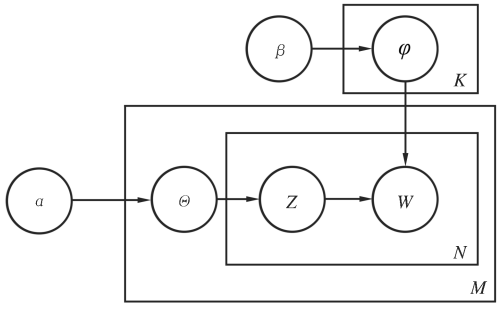

图1　LDA模型原理图

Fig.1　Schematic diagram of LDA model

$$P\text{-}V=\frac{P(D)}{\mathrm{Var}(T)},\tag{1}$$

其中,D 是文本数据集,$P(D)$是数据集的困惑度,T 是数据集中 LDA 抽取的主题,$\mathrm{Var}(T)$是数据集的主题方差.P 越小,LDA 的泛化能力越好.主题方差越大,LDA 主题抽取的效果越好,同时 $P\text{-}V$ 指标就越小.综上,当 $P\text{-}V$ 指标最小时,其主题数对应的 LDA 主题模型识别主题的效果最优.

困惑度 $P(D)$的基本思想是给测试集赋予较高概率值的语言模型泛化效果更好,困惑度越小,模型对新文本具有越好的预测作用,计算公式如下:

$$P(D)=\exp\{-\frac{\sum_{d=1}^{M}\lg\ p(w_d)}{\sum_{d=1}^{M}N_d}\},\tag{2}$$

其中,语料库中的数据集 D 中共M 篇文档,w_d 表示文档d 中的词,$p(w_d)$ 即文档中词w_d 产生的概率,N_d 表示每篇文档 d 中的单词数量.

主题方差 $\mathrm{Var}(T)$是各个主题分别与其均值之间的距离平方和的平均数,衡量了主题之间的稳定性和差异性,当 $\mathrm{Var}(T)$越大时,主题间的差异性越大,主题区分度越好,计算公式如下:

$$\mathrm{Var}(T)=\frac{\sum_{i=1}^{K}[D_{\mathrm{JS}}^{(T_i,\bar{\varphi})}]^2}{K},\tag{3}$$

其中,T 表示 LDA 抽取的主题,$\bar{\varphi}$ 为主题-词概率分布均值,D_{JS}表示 JS 散度(Jensen-Shannon divergence),度量各个主题和其均值之间的偏离程度,K 表示主题数目.

2　新兴主题识别指标体系

新兴主题通常是研究内容上具有较高的新颖性,具有一定的话题规模,并能够吸引新的学者进行研究的

主题.构建了包含主题强度、主题新颖度、复合主题关注度的新兴主题识别指标体系.

2.1 识别指标

2.1.1 主题强度

主题强度是一种抽象的属性,可以用不同的量化计算方法,如按照包含主题的论文数量计算以及根据发文量和被引量计算等[16],因 LDA 模型提取的主题是文档-主题概率形式,直接采用主题文档数量方式量化计算存在偏差,这里定义主题强度(T_s)为该主题在某时间节点内所有的文档-主题概率的总和.

$$T_s=(\sum_{i=1}^{n}p_i), \tag{4}$$

其中,p_i 表示主题 s 内的第 i 个文档的文档-主题概率.

2.1.2 主题新颖度

主题新颖度一般根据主题的年份信息确定,某个主题的年份越新,其新颖性越高.选取各个主题下概率大于等于 10%的文档作为主题的支持文档[17],用主题内所包含文档的平均年份作为主题新颖度 N_s 的度量,即:

$$N_s=\frac{\sum_{i=1}^{n}y_i}{n}, \tag{5}$$

其中,n 表示主题 s 内文献数量;y_i 表示第 i 篇论文的发表年份.

2.1.3 复合主题关注度

主题关注度是测度主题对研究者的吸引力大小,可以用相关文献指数[18]表示,即主题相关的文献数量与对应年份下平均主题相关文献数量的比值来表示,计算公式如下:

$$\theta_s=\frac{d_s}{M_t}, \tag{6}$$

其中,d_s 代表主题 s 的相关文献数量;$M_t=\frac{c_t}{n}$ 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.

主题关注度也可以表现为作者关注指数,用主题相关的作者数量与对应年份下平均主题相关作者数量的比值来表示,计算公式如下:

$$\gamma_s=\frac{n_{ts}}{N_t}, \tag{7}$$

其中,n_{ts} 表示时间窗口 t 下某个主题 s 所有作者数量和,N_t 表示时间窗口 t 内平均主题相关作者数量,$N_t=\frac{A_t}{n}$,A_t 为 t 年的相关作者总数量,n 为 t 年的主题数.

复合主题关注度 T'_s 综合考虑指标的变异性和冲突性,将相关文献指数和作者关注数进行复合加权,计算公式如下:

$$T'_s=\alpha\theta_s+\beta\gamma_s, \tag{8}$$

其中,权重 α 和权重 β 由 CRITIC 客观赋权法[19]确定.

2.2 识别主题类型判定

根据主题强度和主题新颖度的值,将识别出主题进行类型划分,包括新兴主题、潜在新兴主题、非成长型主题和热门主题 4 类,以主题强度和主题新颖度值的均值为原点,绘制出主题类型的判定坐标系,如图 2 所示.

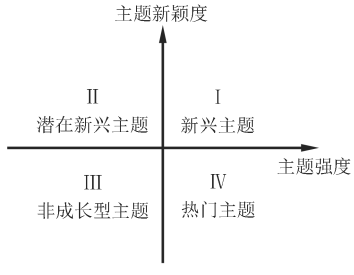

图2 主题类型判定坐标系

Fig.2 Theme type determination coordinate system

在判定坐标系中,第一象限是新兴主题,具有一定的话题规模,研究内容具有较高的新颖性;第二象限的是潜在新兴主题,主题强度不高,但具备一定的主题新颖度,有吸引研究者进入研究,具有成为新兴主题的潜力;第三象限是非成

长型主题,具有一定的成长停滞性,其话题规模和新颖度都较小,吸引研究者关注的潜力也相对较小;第四象限的是热门主题,具有较大的话题规模,但主题新颖度相对较低,对研究者的吸引度相对低.

3　新兴主题的识别与预测

新兴主题识别与预测的方法流程如图 3 所示,主要包括数据获取和预处理、基于 LDA 主题模型的主题提取、识别指标计算、新兴主题识别与预测.

3.1　新兴主题的识别

新兴主题识别的具体的步骤如下:

步骤 1　准备领域科技文献数据集.

步骤 2　构建领域词典,提取论文摘要进行分词、去停用词操作,并构建文档数据集的词频—逆文档频率(term frequency-inverse document frequency,TF-IDF)模型,计算每篇文档的 TF-IDF 向量值.

步骤 3　训练不同主题数的 LDA 主题模型,并计算模型对应的 Perplexity-Var 指标,选取令 Perplexity-Var 值最低的主题数为最优主题数.

步骤 4　根据最优主题数进行 LDA 主题建模,提取领域内研究主题.利用主题-词概率分布确定主题高概率主题词进行主题内容解读.

步骤 5　利用文档-主题概率分布确定主题所属文档,计算各个主题的主题强度和主题新颖度及所有主题强度和主题新颖度的均值.

步骤 6　以主题的强度和新颖度的均值为坐标轴的原点,绘制主题类型判定坐标系.

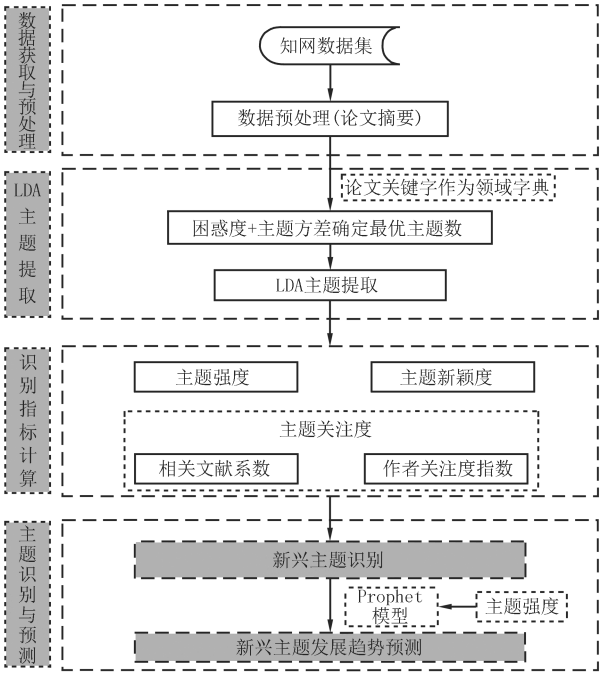

图3　新兴主题识别和预测流程

Fig.3　Emerging topic identification and prediction process

步骤 7　根据主题强度值和主题新颖度值将每个主题划分到对应的象限.位于第一象限的新兴主题和第二象限的潜在新兴主题是要关注的主题.

步骤 8　计算复合关注度的权重 α 和 β,确定复合关注度的计算方法.

步骤 9　计算所有主题关注度均值,对步骤 7 中筛选出的主题进行二次筛选,筛选出大于主题关注度均值的主题,作为最终新兴主题的识别结果.

3.2　新兴主题的预测

Prophet 先知神经网络[20]是目前时间序列分析[21]的热门工具,与 ARIMA[22]模型、LSTM[23]神经网络模型等主流的时间序列模型相比,Prophet 模型具有自动性好、可解释性强、可扩展性强、训练速度快等优点.作为一个加法模型,其假设观测变量的规律满足如下公式:

$$y(t)=g(t)+s(t)+h(t)+\varepsilon_t, \tag{9}$$

其中,$g(t)$ 为非周期性的增长的趋势项,$s(t)$ 是周期因素项,$h(t)$ 为节假日因素项,ε_t 是满足正态分布的误差项.

模型训练中趋势项 $g(t)$选择分段线性进行趋势预测,不考虑周期因素和节假日因素的影响,分段线性函数满足以下公式:

$$g(t)=(k+a(t)^{\mathrm{T}}\delta)t+(m+a(t)^{\mathrm{T}}\gamma), \tag{10}$$

其中 $k+a(t)^{\mathrm{T}}\delta$ 表示增长速率,$m+a(t)^{\mathrm{T}}\gamma$ 表示线性的偏移.考虑到时间序列中可能的突变点,引入了指示函数 $a(t)$,δ 和 γ 表示突变点对趋势函数的斜率和偏移量影响的大小,参数 T 控制趋势灵活度.

具体的预测步骤如下:

步骤 1 以年为时间片计算新兴主题的年度主题强度,组成主题强度序列;

步骤 2 将主题强度序列划分已知序列和待预测序列;

步骤 3 设置 Prophet 模型参数,构建预测模型,利用 R-squared 和平均绝对误差指标分别验证模型拟合度和预测准确率;

步骤 4 设定模型准确率阈值,如果模型准确率大于阈值,则使用该模型对未来 3 年的主题强度进行迭代预测;

步骤 5 如果模型准确率小于阈值,则重新训练模型.

4 新兴主题识别与探测的应用研究

将提出的方法应用到智慧农业技术领域,识别该领域的新兴主题并预测未来的发展趋势.

4.1 数据集及预处理操作

在中国知网中以“智慧农业”“农业物联网”为主题检索词进行智慧农业技术领域的期刊文献检索,时间跨度为 2009—2022 年,获得 1 710 篇科技文献.

预处理操作主要针对文献摘要进行的文本预处理,首先使用正则表达式剔除掉论文摘要中的非中文字符,包括特殊符号、数字、标点、英文字符等,然后在以论文的关键词作为领域字典的基础上,进行分词、去停用词处理,最后采用 TF-IDF 模型对语料进行向量化处理.

4.2 智慧农业领域研究主题提取

对预处理后的文献语料首先进行 LDA 主题建模,为使抽取的主题和主题词更具代表性,设置主题分布的先验参数 $\alpha=0.01$,词分布的先验参数 $\eta=0.01$,然后利用 Perplexity-Var 的计算选取最优主题数.计算了主题数量为 1 到 20 之间的 Perplexity-Var 指标值,结果如图 4 所示,可以看出,当主题数量为 9 时,Perplexity-Var 值最小,模型泛化能力与主题区分度相对较好.

每个主题提取了前 20 个关键词,对应的主题-词表如表 1 所示.

表 1 主题-词表

Tab. 1 Theme-vocabulary

主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8	主题 9
农机	平台	系统	物联网	数据	建设	服务	农业物联网	科技
企业	数据	实现	应用	生产	农产品	气象	智慧	高级阶段
南京	网络	信息化	快速	分析	现代农业	生态环境	模式	融合
合作	中心	控制	智能监测	问题	农村	经营	基地	中联重科
装备	专家	数据	知识	互联网	农民	农业装备	应用	产业
国际	应用	传感器	生产	促进	服务	服务体系	蔬菜	广东
机械化	集成	生产	效率	经济	打造	流程	种植	共享
农场	广西	远程	现代农业	推动	创新	优化	改革	温度
共同	实验	采集	无人机	传统	战略	制度	示范	农业产业
作业	需求	管理	传感	转型	经济	合理	服务中心	升级

4.3 新兴主题的识别与发展趋势预测

4.3.1 主题强度和新颖度的计算

对利用 LDA 提取的 9 个主题分别计算出其主题强度的值和主题新颖度的值,主题强度的均值为 0.11,大于均值的有主题 3、主题 4、主题 5、主题 6、主题 8;说明智慧农业领域的技术较新,各个主题新颖度普遍较高,但略有差异,大于均值的有主题 1、主题 4、主题 5、主题 7 和主题 8.

以主题强度和新颖度的均值为原点，绘制主题分布坐标系，对每个主题进行分类划分，结果如图 5 所示. 主题 4、主题 5、主题 8 的位于第一象限，为新兴主题；主题 1 和主题 7 位于第二象限，为潜在新兴主题；位于第三象限的主题 2 和主题 9 是非成长型主题；位于第四象限的主题 3 和主题 6 为热门主题.

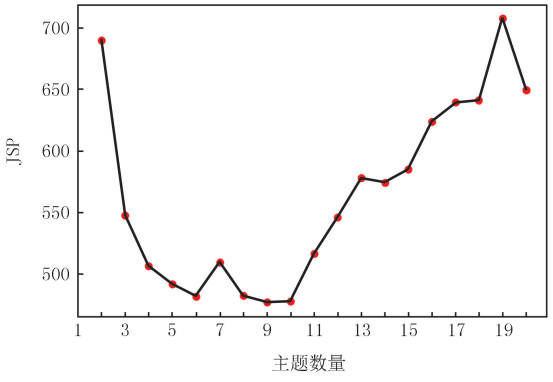

图4　不同数量主题对应的Perplexity-Var值

Fig. 4　Perplexity-Var values corresponding to different number of themes

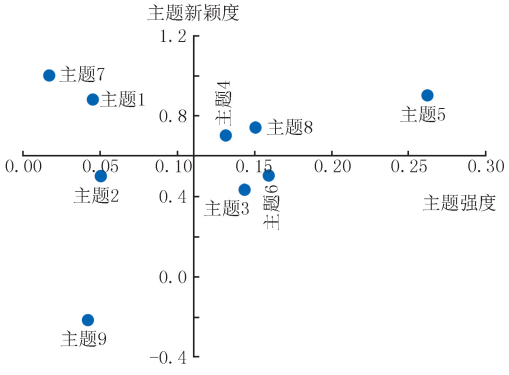

图5　主题的类别划分

Fig. 5　Classification of themes

4.3.2　复合主题关注度计算

在确定新兴主题和潜在新兴主题后，计算复合主题关注度对新兴主题进行二次筛选，主题相关文献指数和主题作者关注指数随年份变化的结果如图 6 和图 7 所示.利用 CRITIC 客观赋权法计算得到的相关文献指数和作者关注指数的权重，如表 2 所示.复合的主题关注度计算结果如表 3 所示，平均主题关注度约为 9.16，高于均值的有主题 3～6 以及主题 8(黑色字体).

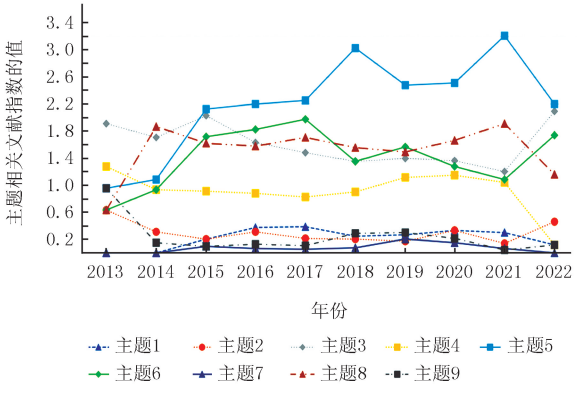

图6　主题相关文献指数

Fig. 6　Theme related literature index

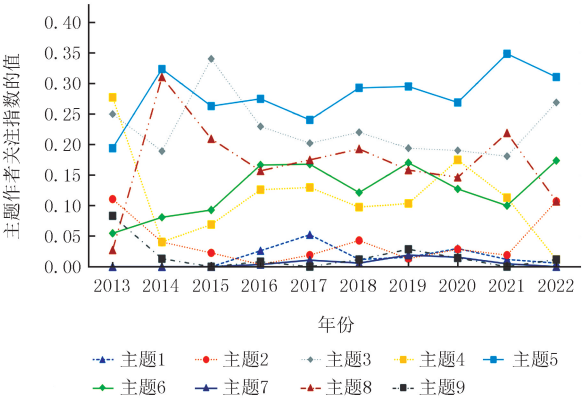

图7　主题作者关注指数

Fig. 7　Theme author attention index

表 2　CRITIC 权重计算结果

Tab. 2　CRITIC weight calculation results

指数	指标变异性	指标冲突性	信息量	权重/%
相关文献指数	7.726	0.019	0.147	49.21
作者关注指数	7.973	0.019	0.152	50.79

4.3.3　新兴主题的识别

通过包括主题强度、主题新颖度和复合主题关注度在内的新兴主题识别指标体系的筛选，最终确定主题 4(物联网技术在智慧农业的探索应用)、主题 5(智慧农业下的数据来源以及利用)、主题 8(示范基地的带动以及由单点到面的推广)为新兴主题；主题 1(吸引企业助力发展智慧农业)和主题 7(智能装备以及服务体系优化)为潜在新兴主题.

从主题识别内容上看，涵盖了在智慧农业的试点示范阶段，研究者对于智慧农业在综合运用移动互联

网、物联网、智能控制、无线传感等现代信息技术上的探索以及如何有效推广和助力智慧农业发展上的探索.

表 3 主题关注度计算

Tab. 3 Calculation of topic attention

类别	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8	主题 9
相关文献指数	2.237	2.997	16.155	9.145	22.049	14.098	0.727	15.179	2.414
作者关注指数	1.348	3.177	17.862	9.213	22.617	10.337	0.554	13.542	1.353
主题关注度	1.792	3.087	**17.009**	**9.179**	**22.333**	**12.216**	0.640	**14.361**	1.884

4.3.4 新兴主题发展趋势预测

利用 Prophet 模型进行 2022 年至 2024 年的 3 年发展趋势预测.

为了验证模型性能,先进行样本内预测,利用训练集中 2020 年之前的主题强度序列进行 2021 年的预测.模型训练中的趋势增长模型选择分段线性函数进行预测,即设置参数 growth='linear',不考虑周期因素和节假日因素的影响,设置 weekly_seasonality = False,daily_seasonality = False,其他参数使用模型默认.

模型的拟合度优劣采用 R 方(R-squared)进行衡量,R 方越接近于 1,模型拟合度越好.选择平均绝对误差(mean absolute error,MAE)进行预测偏差评估,MAE 值越小,预测效果越好.出于对主题强度发展偏差的考量,设定 MAE 值小于 10 时,模型预测有效.

2021 年的主题预测结果如表 4 所示,其中 R 方为 0.995,说明模型拟合效果优良;MAE 值为 6.97,在设定的阈值范围内;主题 4、主题 1 和主题 7 的实际值均在预测区间内,主题 5 和主题 8 在预测区间外,但根据其主题内容判断数据融合应用以及在政策推动下,主题 5 和主题 8 近期获得的关注度较大,有超出模型预测区间的可能.总体来看,Prophet 模型可以进行 2022—2024 年的趋势预测.

表 4 2021 年值与预测区间对比

Tab. 4 Comparison of 2021 values and forecast intervals

类别	主题 4	主题 5	主题 8	主题 1	主题 7
预测区间	[37.7,49.8]	[75.1,106.2]	[43.5,59.1]	[12.7,17.0]	[4.5,6.8]
预测值	46.46	90.90	51.29	14.96	5.63
实际值	49.70	111.70	59.82	16.71	6.18
MAE 值	—	—	6.97	—	—
R 方	—	—	0.995	—	—

将预测区间设为 2022—2024,预测结果如表 5 所示,可以看出 5 个主题在未来 3 年的主题强度区间值均呈逐渐上升趋势,表明这些主题会持续获得领域内研究者的关注和探究.

表 5 2022—2024 年预测区间

Tab. 5 Forecast interval for 2022—2024

年份	主题 4	主题 5	主题 8	主题 1	主题 7
2022	[41.0,51.7]	[81.1,112.2]	[47.3,61.9]	[13.7,18.0]	[5.1,7.5]
2023	[44.2,55.1]	[87.4,117.1]	[50.1,65.1]	[14.9,19.4]	[5.6,8.0]
2024	[47.6,59.1]	[94.5,124.2]	[53.8,69.0]	[16.2,20.6]	[6.2,8.1]

5 结 语

围绕新兴主题识别和探测,首先对新兴主题识别的研究以及主要方法进行了梳理,然后基于 LDA 主题模型,利用 Perplexity-Var 指标确定的最优主题数进行主题抽取,最后通过新兴主题识别指标体系的筛选识别出新兴主题,并利用 Prophet 模型对新兴主题未来发展趋势进行预测.以智慧农业领域的文献数据为实验数据集,经过实验验证,最终确定了 3 个新兴主题和 2 个潜在新兴主题,反映了当前智慧农业领域的研究发展前沿及未来 3 年的发展趋势.

构建的包含主题抽取最优数目确定、识别指标体系优化以及利用 Prophet 模型进行趋势分析的新兴主

题的识别与趋势预测方法是对新兴主题识别和预测进行的有益探索，实验结果较好地反映了智慧农业领域内的新兴主题及发展趋势，表明识别和预测方法的有效性，能够达到优化和探索新兴主题识别和趋势分析的目的.

当前研究尚存在一定的不足.首先，限于篇幅和研究精力，数据源只选择了科技论文文献，未考虑专利文献数据、基金项目数据、网评文本数据等；新兴指标测量上只考虑了文献本身的发表年份、作者、关键词、摘要等文本内容特征，忽略了文献之间的引文特征.在以后的研究中，数据源可采用论文、专利等多源数据从不同角度反映领域主题的发展情况；指标识别体系上可从文本内容特征、结构特征、引用特征等多角度进行指标构建以更好、更全面、更客观地进行新兴主题识别；此外，在新兴主题的趋势发展分析上也可尝试用不同的参数设置进行趋势优化探索.

参　考　文　献

[1] 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.

[2] LI H Y，CUI L，CUI M，et al.Active research fields of acupuncture research：a document co-citation clustering analysis of acupuncture literature[J].Alternative Therapies in Health and Medicine，2010，16(6)：38-45.

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

[4] 陈新亚，李艳.近 20 年来我国教育技术研究的热点与前沿：基于 7 种 CSSCI 期刊的文献计量分析[J].现代教育技术，2020，30(12)：12-19.
CHEN X Y，LI Y.The hotspots and frontiers of Chinese educational technology research in the lastest 20 years：based on the bibliometric analysis of 7 CSSCI journals[J].Modern Educational Technology，2020，30(12)：12-19.

[5] 曹琨，吴新年，靳军宝等.基于共词和 Node2Vec 表示学习的新兴技术识别方法[J/OL].[2023-10-10].http://kns.cnki.net/kcms/detail/10.1478.G2.20221125.1824.012.html.

[6] VAYANSKY I，KUMAR S A P.A review of topic modeling methods[J].Information Systems，2020，94：101582.

[7] 严宇宇，陶煜波，林海.基于层次狄利克雷过程的交互式主题建模[J].软件学报，2016，27(5)：1114-1126.
YAN Y Y，TAO Y B，LIN H.Interactive topic modeling based on hierarchical dirichlet process[J].Journal of Software，2016，27(5)：1114-1126.

[8] WANG J Y，ZHANG X L.Deep NMF topic modeling[J].Neurocomputing，2023，515：157-173.

[9] 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.
ZHOU Y Z，MIN C.Identifying emerging technology with LDA model and shared semantic space—case study of autonomous vehicles[J].Data Analysis and Knowledge Discovery，2022，6(S1)：55-66.

[10] 吴胜男，田若楠，蒲虹君，等.基于社交媒体的医药领域关联主题预测方法研究[J].数据分析与知识发现，2021，5(12)：98-109.
WU S N，TIAN R N，PU H J，et al.Predicting related medical topics from social media[J].Data Analysis and Knowledge Discovery，2021，5(12)：98-109.

[11] 张振青，孙巍.基于特征测度和 PhraseLDA 模型的领域学科交叉主题识别研究：以纳米技术的农业环境应用领域为例[J].数据分析与知识发现，2023，7(7)：32-45.
ZHANG Z Q，SUN W.Interdisciplinary subject recognition based on feature measurement and PhraseLDA model—case study of nanotechnology in agricultural environment[J].Data Analysis and Knowledge Discovery，2023，7(7)：32-45.

[12] ALATTAR F，SHAALAN K.Emerging research topic detection using filtered-LDA[J].AI，2021，2(4)：578-599.

[13] PORTER A L，GARNER J，CARLEY S F，et al.Emergence scoring to identify frontier R&D topics and key players[J].Technological Forecasting and Social Change，2019，146：628-643.

[14] BLEI D M，NG A Y，JORDAN M I.Latent dirichlet allocation[J].Journal of Machine Learning Research，2003，3：993-1022.

[15] 关鹏，王曰芬.科技情报分析中 LDA 主题模型最优主题数确定方法研究[J].现代图书情报技术，2016(9)：42-50.
GUAN P，WANG Y F.Identifying optimal topic numbers from sci-tech information with LDA model[J].New Technology of Library and Information Service，2016(9)：42-50.

[16] 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.
BAI J Y，YAN D W，CHEN Q.Trend prediction of emerging topics based on topic model and curve fitting[J].Information Studies：Theory & Application，2020，43(7)：130-136.

[17] MANN G S，MIMNO D，MCCALLUM A.Bibliometric impact measures leveraging topic analysis[C]//Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries.New York：ACM，2006：65-74.

[18] 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.

LI S F,HUANG Y,YANG J Q.Research on frontier research topic recognition method in agriculture field based on BERT[J].Technology Intelligence Engineering,2021,7(5):100-114.

[19] 颜惠琴,牛万红,韩惠丽.基于主成分分析构建指标权重的客观赋权法[J].济南大学学报(自然科学版),2017,31(6):519-523.
YAN H Q,NIU W H,HAN H L.Objective weight method based on principal component analysis to establish index weight[J].Journal of University of Jinan(Science and Technology),2017,31(6):519-523.

[20] TAYLOR S J,LETHAM B.Forecasting at scale[J].The American Statistician,2018,72(1):37-45.

[21] HOSSAIN M M,ANWAR A H M F,GARG N,et al.Monthly rainfall prediction at catchment level with the facebook prophet model using observed and CMIP5 decadal data[J].Hydrology,2022,9(6):111.

[22] HAN F S,ZHANG C X,ZHU D L,et al.Talent cultivation of new ventures by seasonal autoregressive integrated moving average back propagation under deep learning[J].Frontiers in Psychology,2022,13:785301.

[23] FENG S F,FENG Y.A dual-staged attention based conversion-gated long short term memory for multivariable time series prediction[J].IEEE Access,2021,10:368-379.

An emerging topic identification and detection method based on LDA model

Wu Dongxue[a], Shen Guilan[b]

(a. College of Applied Arts and Sciences; b. Bussiness College, Beijing Union University, Beijing 100191, China)

Abstract: Emerging topic identification is an important way to identify emerging technologies in the field of science and technology research, and efficient and accurate identification of emerging topics is the premise of early identificating emerging technology research direction. An emerging topic identification and trend prediction method based on LDA model is proposed. It extracts research topics from scientific literature by LDA model, constructs an index system of topic strength, topic novelty and composite topic attention to identify emerging topics, and uses Prophet model training to predict topic strength of emerging topics. Based on the data set of scientific research literature in the field of smart agriculture in the last 14 years, the proposed recognition and detection methods are verified. Five emerging topics are identified, and the development trend in the following three years is predicted. The validity of the proposed methods is verified.

Keywords: topic identification; optimal topiccount; emerging topic identification indicators; Prophet model

[责任编校 陈留院 赵晓华]