



基于FSD模型的政府资助项目新兴主题探测与分析

徐路路¹ 靳 杨²

- (1. 南开大学 商学院信息资源管理系, 天津 300071;
2. 首都医科大学附属北京安贞医院, 北京 100029)

摘要:如何捕捉科技领域发展趋势并高效准确地追踪科研活动动态演变一直是研究人员关注的焦点。以美国国家科学基金会政府资助项目文本为分析数据源,综合运用主题模型及指标构建方法,探索文本结构特征并从资助金额、布局强度等多个维度分析,分析主题生命周期提出基于FSD模型的项目文本新兴主题探测方法。结果表明,该方法能够快速前瞻识别出新兴主题,形成主题—主题词—项目序列号的混合分布聚态集群,从新兴主题探测数量、探测质量及探测时间3个维度对比验证了新兴探测模型的优越性。

关键词:新兴主题;项目文本;FSD模型;预测分析

中图分类号:G31;G307 **文献标识码:**A **文章编号:**1002-0241(2019)02-0040-15

0 引言

目前全球范围内新一轮科技革命迅速发展,以大数据和信息处理技术为基础的物理量子、能源产业及纳米分析等领域技术群渗透融合交叉发展。2014年,习总书记在两院院士大会指出广大科技工作者应紧跟全球先进科技发展方向,发展科技必争方向以增强国家创新力与国际竞争力(人民网,2015),而我国政府实施制造强国战略第一个十年的行动纲领《中国制造2025》也指出把握未来科技趋势并寻求学科交叉融合的重要意义(中国政府网,2018)。因此,如何前瞻性识别出繁杂多样科技文献中的前沿主题,把握支撑重要领域科技创新的发展新脉搏并制定战略性科技政策,成为情报研究的重要任务(郑烨等,2017)。

目前很多学者以科技文献探测视角分析重大战略部署与发展态势,揭示交叉交融性与传统基础性科学的未来发展。但也存在一些问题,一方

面,众多研究围绕论文数据源展开,由于期刊论文的时滞性和计量指标不充分性,使得基于期刊论文的研究前沿识别存在科学性不足的问题(徐路路等,2018);另一方面,前沿识别领域对于新兴主题和热门主题存有界定不够明确、内涵要义混淆等缺陷,新兴主题是具有较大发展潜力、最新并具有灰色性、创新性和不确定性的主题,而热门主题是目前已广泛开展并深入研究的主题信息类型。传统文献计量方法多采用关键词统计方法,以词频大小判断主题强弱程度,但词频较高的研究主题往往是较为成熟的热门主题,而不是代表未来潜在发展新动向的新兴主题(许振亮等,2011)。

针对目前研究中存在的不足,本文借鉴应用于新闻媒体信息流中的TDT(topic detection and tracking,主题探测与跟踪)技术,提出一种基于FSD(first story detection,首次报道检测)模型的新兴主题探测与研判新方法,语义挖掘并识别政

收稿日期:2018-03-14

基金项目:国家社会科学基金重大项目(14ZDA063)

第一作者简介:徐路路(1991—),男,山东临沂人,南开大学商学院,博士生,研究方向:机器学习与知识发现。

通信作者:靳杨,mingzhenzi@foxmail.com

府资助项目文本中的新兴类型主题。政府资助项目蕴含着科技未来发展方向,是未来3年到5年不等一段时期内的科研规划和研究主题部署,具有重要的科学研究前沿信息。识别美国等科技密集地区和国家的新兴主题和前沿热点,势必在我国科技政策制定和优先发展领域筛选进行大势研判是提供有力的数据支撑,同时,本文尝试寻求FSD模型与新兴主题识别的应用结合点,以期为目前研究前沿相关领域研究提供一种新思路、新视角。

1 相关研究

1.1 新兴主题研究

目前文本挖掘和信息计量科学领域存在较多与新兴主题概念相近的研究主题术语,如研究前沿、研究热点等(Kessler, 1963; Kleinberg, 2003; Kontostathis et al, 2004; Mane et al, 2004),利用前沿主题信息把握主题迁移规律并进行知识创新、科学研究推进与辅助创新。但上述主题也存在新兴主题与热门主题以及知识基础等相关概念内涵及外在知识表达仍需进一步定义。

2001年,科学家Naohiro定义新兴主题概念并用多组关键词表征当前研究活动前沿信息(Matsumura et al, 2001)。2004年,学者Kontostathis进一步将新兴研究趋势(Emerging trend)定义为随时间增长其主题关注度和科研产出逐渐提高并引起科研人员探索兴趣热情的新兴主题领域(Kontostathis et al, 2004)。

近年来,新兴主题的识别与探测得到了相关学者与情报分析人员的广泛关注,并进一步丰富新兴主题研究内容和方法。2006年,学者Hoang提出新兴主题测量指标,利用论文来源、研究人员、主题属性、研究机构等外在属性确定主题的新颖度和受关注程度(Hoang, 2006)。2008年,殷蜀梅采用文献计量学、网络计量学以及时间序列分析模型等方法表达新兴主题的主题特征并构建新兴发展趋势评定技术框架(殷蜀梅, 2008)。2014年,葛

菲等提出基于关键词生命周期和引文分析的新兴主题识别方法,借助波士顿矩阵图以词频和增长率均值为原点对未来趋势主题进行可视化图谱分析(葛菲等, 2014)。2015年,黄鲁成等利用高频关键词共现及聚类探测生物材料新兴趋势,实验表明该方法具有较好的主题探测和分析能力(黄鲁成等, 2015)。2017年,段庆锋等提出基于替代计量学的监测指标运用社交媒体数据识别未来科学具有高度潜力和关注度的学科主题(段庆锋等, 2017)。

目前,从科技文献中识别和探测新兴主题取得了一定的研究成果,但也存在一些不足,如多数基于论文数据展开研究,只分析了主题外部特征(王贤文等, 2014),而未考虑研究主题内部的具体内容的变化,对主题前驱和后继发展揭示分析较多但不能刻画和线性分析主题脉络,不能判断主题在时间维度上最早出现的第一时间点,以上存在问题限制了新兴主题研究的准确性和有效性。同时,对于新兴主题的内涵尚需进一步界定,应进一步拓展和丰富新兴主题识别的分析数据源并针对特定数据文本进行特征分析。

1.2 TDT技术与FSD模型研究

TDT(话题检测与跟踪, topic detection and tracking)技术源于面向新闻事件识别与信息获取的处理技术,涉及机器学习、人工智能、自然语言等多学科以解决信息化时代数据超载和爆炸问题(王会珍等, 2006)。面向已知话题的跟踪任务评测体系的建立为后续未知话题探测以及短文本、流媒体及多模态信息流的探测提供解决方案(Allan, 2002)。目前,TDT技术在突发事件探测、网络舆情分析、信息安全等方面都取得较为广泛的应用但在科学研究前沿和新兴主题研究方面尚未展开(杨玉莲等, 2009; 张辉等, 2010; 张小明等, 2012)。

FSD模型(首次报道检测, first story detection task)是TDT中的一项子任务,用于检测新闻

媒体信息流中涌现的新闻报道,其主要任务是从具有时间顺序的报道流中发现第一篇报道所处的时空位置(Lo et al, 2002)。FSD模型的目标在于对第一次发生的新闻的探测与后续报道的预测分析,通过相似度阈值的计算自动实现新闻主题归类的智能化处理,从而发现该新闻语料出现的位置,有效识别和区分话题簇。如图1所示,不同区块代表不同新闻事件,有“地域”、“体育”、“电影明星”、“总统竞选”等,而FSD功能就是找到相同的事件为后续进行“总统竞选”话题的跟踪和预测分析做好语料筛选和预测工作。

FSD的核心思想是在时间维度上对后续相关报道进行相似度计算聚类分析并捕捉最新颖事件,对事件的前驱和后继发展及话题漂移现象进行研究。图2展示的FSD模型算法流程图。

FSD模型相似度阈值的计算及比较对新闻报道分类整合形成不同主题类型的话题簇,其部分模型设计中,媒体源信息流中处理得到不同类型主题 w ,而陆续出现报道 d 与已识别话题相关概率算示表达为:

$$P(T/d) = \frac{P(T) \times P(d/T)}{P(d)} \approx \prod_n \frac{P(w/T)}{P(w)} \quad (1)$$

$$P(w/T) = \alpha \times P(w/T) + (1 - \alpha) \times P(w) \quad (2)$$

式中: $P(T)$ 表示新闻报道与话题 T 相关先验分布概率; $P(w)$ 表示某个领域主题词 w 后验概率统计分布状态; $P(w/T)$ 表示主题词是先验主题报道 T 条件生成后的所属大小; $P(T/d)$ 指标则是衡量报道 T 和话题 d 的概率,即文本相似度计算数值。针对新

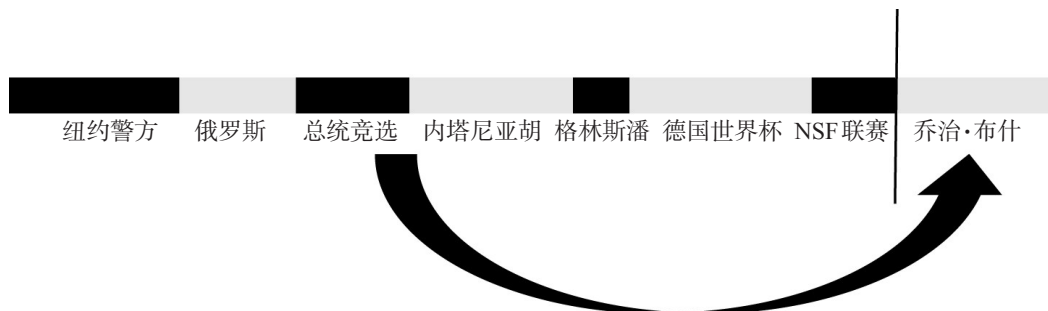


图1 FSD首次报道检测示意图

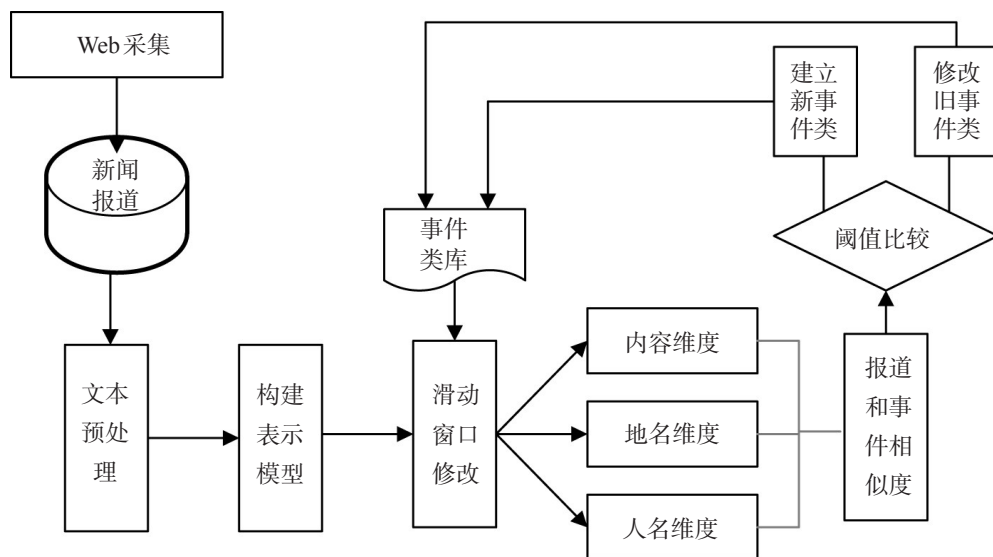


图2 FSD模型算法流程图

闻文本的文本特征,学者Allan(2002)采用线性差值法加入混合模型以降低主题重合度优化主题探测,具体如算式(2)所示。

为更为准确评价FSD模型的系统性能,科研人员设计漏检率和误检率2个指标来衡量并通过加权求和表达(张美珍, 2010),其评测算式如下:

$$C_{Det} = C_{Miss} \times P_{Miss} \times P_{target} + C_{FA} \times P_{FA} \times P_{non-target} \quad (3)$$

式中: C_{Det} 系统代价指标; C_{Miss} 和 C_{FA} 分别表示漏检率和误检率的代价系数,其数值为人设定; P_{target} 表示话题出现的统计概率; $P_{non-target} = 1 - P_{target}$ 。

随后,相关研究人员为使性能指标更有意义,对该公式进行一定改进(Elsayed et al, 2005),如下:

$$(C_{Det})_{Norm} = \frac{C_{Det}}{\min(C_{Miss} \times P_{target}, C_{FA} \times P_{non-target})} \quad (4)$$

石墨烯(graphene)作为一种碳原子单层纳米级纤维,是目前发现的最薄、强度最大、导电导热性能最强的平面结构新材料,在物理学、计算机、航空航天等领域都得到了长足的发展,具有广阔

的发展前景,因此科学家预言石墨烯产业是颠覆性新技术新产业。NSF(national science foundation, 美国国家科学基金会)是全球科研规模和资金投入最高也最有权权威性政府资助项目单位。

综上,本文拟以NSF石墨烯领域为分析数据来源,以美国国家基金委资助的石墨烯领域资助项目文本,结合政府资助项目文本所特有的资助强度、时间强度等文本特征,有效融合FSD模型设计思想,构建丰富的指标评价体系以判断新兴主题最新出现的第一时间点,为后续研究提供情报支持和参考。

2 基于FSD模型的新兴主题探测方法

2.1 新兴主题

新兴主题在时间维度上并非是一成不变的,不同时期内研究前沿主题内部从属的主题词是不断变化的,宏观表现为主题的新生、成长、成熟、衰老等形式,与生命周期曲线引入、发展、成熟、衰老等发生发展过程存在一定共性。图3反映生命周期与主题周期不同时期的对应关系以及主题周期不同阶段所呈现的外在特征。

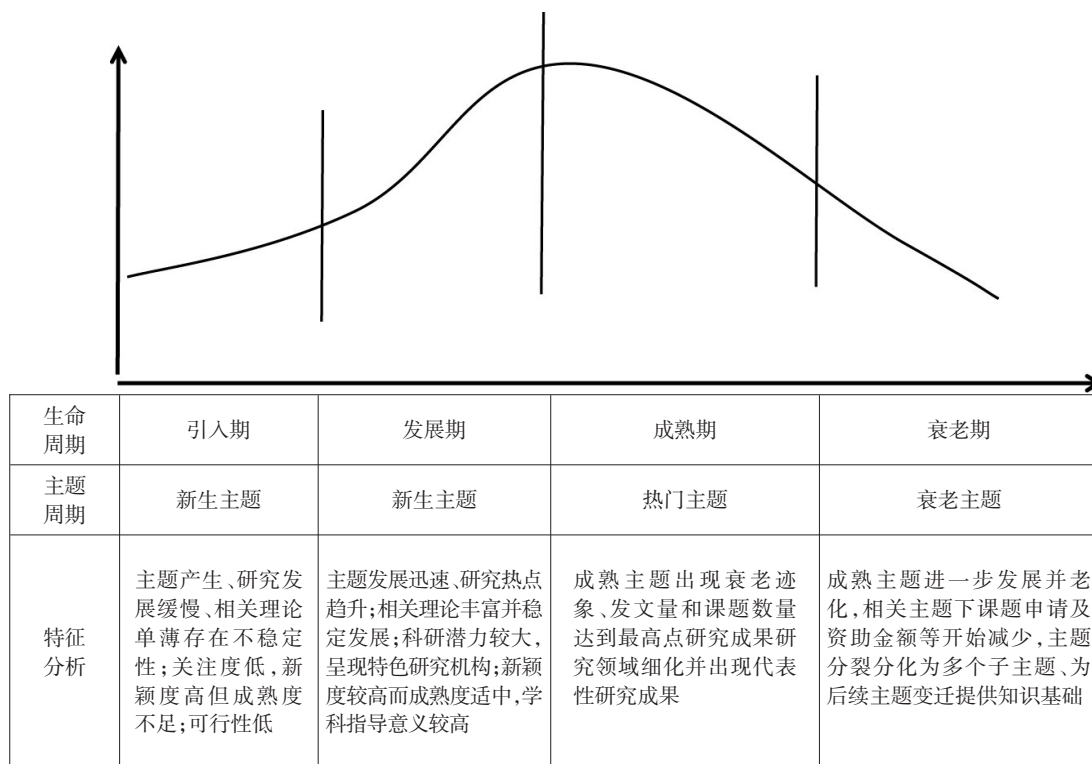


图3 生命周期与主题周期比较分析图

本文认为新兴主题是指具备成为热门主题的潜在可能性但尚未广泛开展研究的主题,其显著特点是新颖性、创新性和巨大发展潜力,新兴主题未来发展可能成为热门主题。新兴主题作为热门主题的前期演化主题类型,比热门主题出现时间更早,而其前瞻价值和探索性亦高于热门主题,因此,对于新兴主题的探测与研究可以更好地进行领域研究前沿预测分析。

本文根据主题生命周期发展演化规律从主题发展的潜在时期和突破时期定义和分析新兴主题:

①利用生命曲线表征主题演化过程可以得出,作为起始未引起人们广泛关注的新兴主题需具备潜在阶段。该阶段,如 t 时期内主题研究热度处于上升期,存在研究机构和研究人员数量较少、政府资助项目主题数较少、资助金额较少等特点,该主题研究年龄较新、研究力度和产出水平较低。

②新兴主题还应具备突破阶段。具备潜在阶段的研究主题是否为新兴主题还需要经过两式的判别,即存在 $t+1$ 时期,该主题的研究热度超过同时期不同主题的平均研究热度,资助金额、资助时长、主题强度等特征要素要高于平均研究水平。

经 $t+1$ 时期后,该主题后续发展为广泛研究的热门主题。

新兴主题在不仅主题生命周期上表现出明显的阶段特征,也具有明显的属性特征,如新兴主题的高成长潜力度、高关联度等(黄鲁成等, 2015)。欧盟在未来和新兴技术项目(future and emerging technologies, FTE)中指出其开放式研究的新兴领域的研究主题应满足基础性、长期性、灰色性、风险性等属性特点。因此,本文在前期研究的基础上,定义了新兴主题5大关键属性并对其相关要素参数评价和实现途径进行总结归纳,见表1。

2.2 新兴主题探测方法构建

FSD模型在计算机领域对新闻文本的应用研究与前沿探测领域针对科技文献进行新兴主题识别和演化分析过程存在研究目的和研究方法的相似性,因此,本文将计算机领域FSD模型和科学学中新兴主题探测有机融合,提出基于FSD模型的新兴主题探测模型。

(1) FSD模型针对新闻媒体信息流不同文本特征采用差异性表示模型,注重细粒度时间切片以客观细致地进行主题探测与新闻追踪。因此,

表1 新兴主题属性特征分析

关键属性	内容	指标评价	作用	实现途径
基础性	指新兴主题需具备初期理论基础和实验基础,并进行科研活动再创新	主题相似度 (<i>Similarity</i> 大于 0 表示有初期理论基础)	分析不同时期文本内容主题的相似度	选取Jaccard相似度系数、Pearson相关系数等,进行相似度计算
长期性	指新兴主题需具备完整主题生命周期,有较长的研究时间跨度	阶段资助时长及周期资助时长	分析其主题持续资助的时间,以排除短暂资助但后续衰老的主题	通过利用项目文本的资助起止时间得到资助时长,如NSF、NIH等文本
突破性	指新兴主题应具备突破阶段,即存在 t 时期其研究水平要高于主题平均研究水平	主题研究水平并与基准值比较分析	分析该主题同时期所处的研究水平并判断发展趋势	综合考量不同时期主题特征要素,进行多维尺度分析,寻找主题突破点
新颖性	指新兴主题与前期或当期研究领域的相似性及创新力度	主题聚类分析及相似度计算	分析某研究主题与其他主题研究内容的差异性并判断其新颖性	通过自然语言处理技术计算当期主题与其他主题相似度,评估其原始创新能力
跨学科	指新兴主题在多领域、多学科的分布特征,存在的研究主题的学科交叉程度	学科交叉度、资助机构合作度等测度指标	分析跨学科主题研究分析,识别潜在跨学科领域或潜在跨学科研究团队	通过资助机构、作者、目标文献、参考文献等学科类别进行跨学科测度

为使本研究中新兴主题探测更为准确,本文基于FSD模型提出针对项目文本的摘要、资助受体单位、主题项目布局、资助开始与终止时间等信息设计多种特征,构建针对特定文本类型的新兴主题探测公式。

(2) 在处理新闻文本时,不同要素的数值及数值单位不尽相同,因此,FSD在综合不同要素指标进行统一分析时采用了归一化开销的方法,使得要素参数有效融合,该方法值得本文借鉴并制定项目文本综合评价模型。

(3) 本文借鉴FSD模型在新闻识别中对首次报道事件的探测原理,尝试寻求政府资助项目中的新兴主题出现的第一时间点,从而展开对于新兴主题识别和追踪起始时间点因此从探测时间、主题及数据等不同角度审视本文方法的探测有效性。

(4) 相似度阈值的大小直接影响FSD系统性能,阈值设定过低,会增加误检率(PFA);阈值设定过高,则会增加漏检率($PMiss$),需要反复调整和设计。因此,本研究基于FSD模型设计理论,提出动态阈值调控法,选取不同阈值进行多组实验,阈值过高导致主题关联关系构建困难而阈值过低则不易发现新兴主题,引起需多次尝试寻求阈值最优值。

2.2.1 基于FSD模型的项目文本新兴主题探测公式

本文借鉴FSD模型中思想及新主题相似度研判公式,根据文本分析加入资助金额、平均资助金额、布局强度、资助时长以及对应同时期平均研究水平等不同文本特征参数综合判定新兴主题,提出石墨烯项目文本新兴主题探测算式,计算如下:

$$DV_t = \alpha \times \frac{TI_t^z}{ATI_t} + \beta \times \frac{FAI_t^z}{AFAI_t} + \gamma \times \frac{FTI_t^z}{AFTI_t} + \chi \times \frac{LSI_t}{ALSI_t} + (1 - \alpha - \beta - \gamma - \chi) \times \frac{NI_t}{ANI_t} \quad (5)$$

式中: $\alpha, \beta, \gamma, \chi$ 为调谐系数; TI_t^z 表示子时期 t

时刻的主题强度值, ATI_t 表示平均主题强度,可通过并行LDA主题概率识别模型得到; DV_t 是综合考虑项目文本主要特征要素的新兴主题探测值(detection value, DV_t), LSI_t 为项目布局强度, NI_t 指标则主要考虑文本主题新颖性而设置指标,尝试构建基金项目文本指标体系。具体参数构建如下。

2.2.2 资助金额参数指标(funding amount index, FAI)

假设依据1:在政府资助项目中,和其他研究主题相比,某一研究主题的资助金额越高一定程度上说明该主题研究难度和研究价值越大,代表该领域内的重要研发方向和科技创新突破点,越可能是该时间段内研究前沿和热点。

$$FAI_t^z = \frac{Sum_t^{FA}(Z)}{PC_t(Z)} \quad (6)$$

式中: $Sum_t^{FA}(Z)$ 表示 t 时间段内围绕主题 Z 的政府资助项目的累计资助金额总和; $PC_t(Z)$ (program count, PC) 表示并行LDA识别后同主题的项目资助的个数; FAI_t^z 表示对总体主题金额与数量比值化处理得到单个项目资助水平,资助金额越大说明科技政策制定者认为该主题在未来具有更好的研究价值,一定程度反映该主题的研究潜力和前瞻价值,该数值较小则说明该主题资助意义或者潜在研究价值较小不值得进一步开展广泛研究。

2.2.3 平均资助金额参数指标(average funding amount index, AFAI)

假设依据2:根据2.2.2中提出的资助金额参数并参考Elsayed等(2005)文中杜教授基线VDP的概念,本文提出平均资助金额比参数是对2.2.2求和均值处理,反映了 t 时间段内所有主题资助金额平均比例。

$$AFAI_t = \frac{FAI_t^z}{TC_t} \quad (7)$$

式中: FAI_t^z 表示资助金额参数,反映单主题下单项目的资助金额; TC_t 表示一定时期内政府资助项目中主题数量的总和(topic count, TC); $AFAI_t$ 表示特定时间段内多个主题资助金额参数的平均值。

利用上述参数的比较分析,可以判定该主题目前资助水平的高低水平和主题发展趋势,如 t 时刻 FAI_t^z 小于 $AFAI_t$ 的研究主题,则说明该主题低于平均水平,属于新兴主题潜在阶段;而在 $t+1$ 时刻大于后者,该主题高于平均值,该时间点为新兴主题突破点,也很有可能为发展潜力的新兴主题。

2.2.4 资助时长参数指标(funding time index,FTI)

假设依据3:政府项目资助文本中具有资助时长特征,资助时长表示项目审批者考虑研究难度、研究风险及创新度基础上给与项目起始和完成期限,资助时间越长一定程度上代表了国家重点攻关的决心和攻坚克难的预算期限。该指标一定程度反映了未来科技布局的重心和难点所在,因而具有更强的科研价值和科研创新。

$$FTI_t^z = \frac{Sum_t^{FT}(Z)}{PC_t(Z)} \quad (8)$$

式中: $Sum_t^{FT}(Z)$ 表示在 t 子时期内同一研究主题分布下资助时长; $PC_t(Z)$ 与 2.2.2 指标含义相同,均表示项目数量级; FTI_t^z 表示 t 时间段单一资助项目政府时长。

2.2.5 平均资助时间参数指标(average funding time index,AFTI)

假设依据4:本文提出的平均资助时长参数是对资助时长参数的求和均值处理,该参数反映了 t 时间段内所有主题的资助时长的平均水平。通过某一主题资助时长与该参数的大小关系比较,可以看出该主题在整体资助时间中所占的位置,进而同时时间维度判断出哪些主题是新兴主题,哪些相对较为成熟。

$$AFTI_t = \frac{FTI_t^z}{TC_t} \quad (9)$$

式中: FTI_t^z 表示资助时长参数,反映单主题下单项目的资助时间; TC_t 表示一定时期内政府资助项目中主题数量的总和(topic count, TC); $AFTI_t$ 表示子时期内的资助金额参数的平均值。

2.2.6 布局空间指标(layout space index,LSI)

假设依据5:若基金项目申请量较大,存在研究机构众多、研究理论较为成熟的情况,体现出布局数量大、布局分散等特点。本文认为该主题在未来一段时间创新和发展的潜在可能性较低,潜在布局空间较小。布局空间指标反映了研究主题在未来一定时间片段下布局和发展的空间,该指标越大,则说明潜在发展和研究意义较大。

$$LSI_t = \frac{1}{\sum_{i=1}^n num(doc)} \quad (10)$$

式中: $\sum_{i=1}^n num(doc)_i$ 表示 y 年主题 z 的项目申请的数量总和,该参数越大说明项目申请基数众多、研究机构分散等,一定程度反映该主题较为成熟和老旧,因此本文在讨论面向未来的新兴主题时考虑布局空间指标 LSI_t (layout space index),并去其倒数数值表示未来主题在研究布局的潜在空间大小。平均布局空间指标则反映了该基金项目当年度的平均探测水平。

2.2.7 新颖度指标(novelty index,NI)

假设依据6:新兴主题不同于热门主题,具有创新性、灰色性和新颖性,越新兴的主题与历史主题尤其是热门主题的主题相似度越小,涌现主题词越新颖则之前出现的情况越少,情报学和计算机理论常用相似度表达主题的相似和重合程度。在此基础上本文提出新颖度指标,该指标对主题相似度值倒数处理,数值越大表示此时期新颖度越高,未来成为新兴主题潜力越大。

$$NI_t = \frac{1}{sim(topic_t, topic_{t-1})} \quad (11)$$

相比传统新兴主题探测方法中仅仅以主题强度单一参数判定主题,本文提出的探测公式及参数考

虑更为全面,把资助强度和新颖度指数、布局强度等考虑到主题探测中使得探测结果更具科学性和客观合理性。同时,以比值的形式描述主题强度、资助强度和时间强度,进行归一化开销可以消除单位的不同所带来的麻烦。新兴主题探测值 DV_i 与平均探测值(数学推导可得其值为1)做比较,可以判断出哪些是成熟主题,哪些是新兴主题。

2.3 实验流程

本文结合政府资助项目数据的关键特征要素和主题内容,融合新闻媒体应用广泛的TDT相关技术,提出基于FSD模型的新兴主题探测方法。

第一步:数据收集。根据研究目标选定数据库,构建检索式,获取近13年美国国家科学基金会(NSF, national science foundation)政府资助项目文本,完成数据准备整理工作。

第二步:数据加工处理。利用正则表达式过滤标点数字等无价值字符串、POS词汇标注、句子抽取与停用词剔除等数据处理工作,形成单粒度特征词汇集合。

第三步:主题识别。利用主题模型对项目文本建模分析并识别主题,结合主题探测与追踪技术,加入时间维度和阈值判定进行主题探测,拟选用包含MALLET机器学习工具包的Ktime实验分析平台。

第四步:新兴主题探测与判别。具体分为:

(1) 政府资助项目文本特征要素分析,统计并计算时间维度上资助强度、新颖度、布局空间等多重文本特征要素。

(2) 结合文本主题基本参数并构建文本特征参数,建立针对于项目文本的新兴主题探测综合判定公式,提出新兴主题探测值 DV_i 。

(3) 利用探测公式完成FSD模型新兴主题识别与跟踪分析。

第五步:实验方法对比验证分析。根据上述方法和步骤实现新兴主题的探测,利用基准方法与

本文设计方法从探测时间、探测数量及探测质量3个维度验证本实验方法的有效性和科学性。

第六步:石墨烯领域发展新动向。深入挖掘美国政府资助项目文本中蕴含面向未来发展的石墨烯领域重要研发方向和科技主题分布,捕捉最新的科研活动信息为我国该领域发展提供数据分析支持。

3 实证研究

3.1 实验环境和数据集

(1) 硬件:Windows7旗舰版64位操作系统、i5-4590 CPU、8GRAM、500G HardDrive;

(2) 软件:Ktime实验分析平台;

(3) 数据库:美国国家科学基金会(NSF, national science foundation)基金项目数据;时间跨度:2004年—2016年;检索式:Keyword="graphene" or "Graphene";检索结果:716项。

3.2 数据集分析与预处理

检索数据集时间跨度13年,2004—2016年NSF政府资助石墨烯领域项目数量总体趋势呈现增长趋势,尤其是2009年以后项目数量增幅明显,说明石墨烯领域逐渐成为美国国家科技工作人员关注的热点。石墨烯项目数据分布如图4所示。

新兴主题探测在于第一时间发现具有较大潜力而未引起广泛关注的主题,因此,将子时期单位设置为一年可较早识别短时间内突发主题词。2004—2007年项目资助数量较少,在LDA主题识别中主题表征效果较差,因而实验起始年份选择为2008年。

3.3 主题识别实验

3.3.1 参数设置及实验流程

本部分实验采用主题概率识别方法识别项目文本主题。PLDA是基于Gibbs sampling近似分布并行框架的LDA模型,收敛效果与文本主题识别准确度较高。相关参数: $No\ of\ topic$ 表示主题数,设置为10; $No\ of\ words\ per\ topic$ 表示每个主题的

词数,实验设置为10; α 取值范围为0.1到1之间,表示文档主题 k 的优先权重数值越小表示文本数据呈现稀疏分布状态,实验设置为0.4; β 为Dirichlet

$Prior$ 先验参数取值0.1与1之间,表示单词 w 在主题中权重分布,实验设置为0.6; $No\ of\ iteration$ 表示迭代次数,设置为1000实验趋于函数收敛稳定状态; $No\ of\ thread$ 表示模型处理线程数可增加模型运行和处理速率,实验设为5。上述实验设置决定主题随机抽取数量、平滑系数等并直接影响实验效果。主题复杂度表示模型对于文本内容的表达能力,为避免主题过度表达和主题冗余需进行主题复杂度实验并设置主题数量为10,单个主题词表达为10。PLDA主题模型识别流程如表2所示。

3.3.2 项目文本主题表征

实验得到主题—主题词—项目序列号的混合分布聚态集群,采用非监督机器学习方式识别出项目文本中潜藏的主题信息。不同主题包含不同主题词和对应权重,每个项目文本数据都附加时间标签,利用PLDA模型主题分布集群可以得到主题与政府资助项目对应关系和结果,为后续新兴主题判别提供基础。表3展示的是各个子时期政府资助项目主题及其所对应的项目数量。在NSF数据集中每个项目都有唯一的项目号,建立多维度映射关系,可以找出每个政府资助项目所对应的主题及资助金额、资助起始时间及结束时间以及分布特点等项目文本基本特征,如表4所示,表示2015年政府资助项目相关文本特征要素及所对应的主题,该实验结果为后续新兴主题的探测提

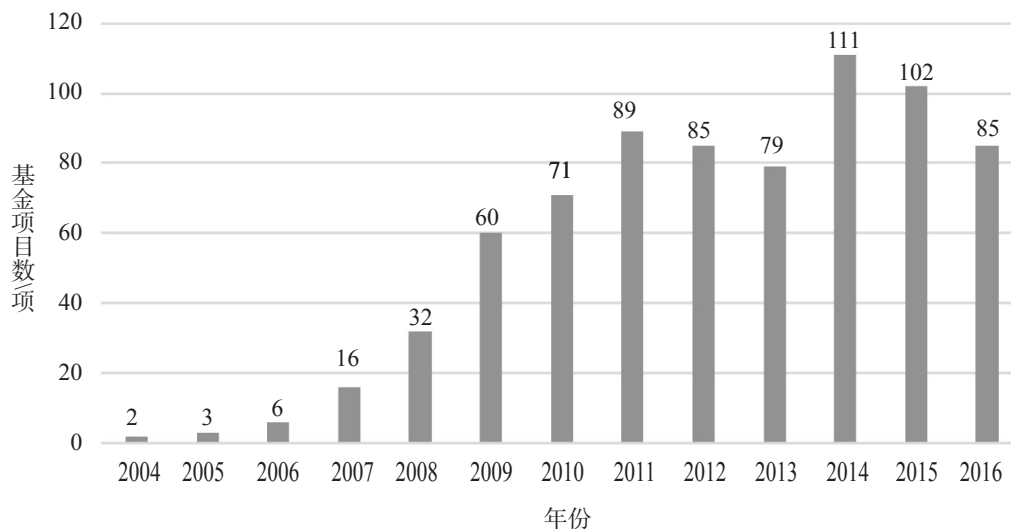


图4 2004—2016年NSF石墨烯领域资助项目数量分布

表2 PLDA主题识别流程

流程步骤
Step1.数据集准备并从节点库调取语法分析器 String to Document
Step2.构建词袋生成器 Bag of Word Creator
Step3.抽取关键词,将文档向量表示降维处理 $v = (x_1, x_2, x_3, \dots, x_n)$
Step4.构建主题抽取器并调整参数设置,进行主题建模 $Topic = \{w_1, w_2, w_3, \dots, w_n\}$
Step5.实现主题-主题词-项目号表征,建立映射关系
Step1.数据集准备并从节点库调取语法分析器 String to Document

供奠定基础。

得到子时期模型识别结果后建立不同区间主题关联从而确定在时间序列中的主题前继后驱状态,同一时期识别10个不同的主题分别用权重最高的主题词,按权重值从大到小表征得到该区间主题识别,共计9个子时间段,具体见表5。

3.3.3 计算探测参数并绘制探测表格

本部分对2008—2016年9个子时期的项目文本主题的资助强度、时间强度、布局强度等进行量化分析并计算参数具体数值,综合衡量项目文本的特征变量,为后续计算新兴主题测度值做好基础工作,如表6展示2014年度项目中主题的相关参数计算。

表3 部分主题词对应子时期资助文本数量/项

子时期	OPTICAL	ELECTRONIC	MENBRANCE	RESEARCH	...	Amount
2011	114	214	333	137	...	89
2012	125	260	203	194	...	85
2013	116	149	177	176	...	79
2014	322	373	309	660	...	111
2015	1051	387	219	270	...	102

表4 部分项目文本特征要素主题表征(2015年资助项目)

Award Number	Start Date	State	Award Instrument	End Date	Awarded Amount	编号
1760041	03/11/2015	AZ	Standard Grant	01/31/2015	\$333 976.00	topic2
1761437	04/01/2015	CA	Standard Grant	07/31/2019	\$509 118.00	topic4
1780164	07/15/2015	OH	Continuing grant	07/31/2017	\$145 000.00	topic9
1899710	07/31/2015	AZ	Continuing grant	09/30/2016	\$708 470.00	topic0
1967107	12/15/2015	UT	Standard Grant	05/31/2020	\$804 117.00	topic4

表5 项目文本各子时期主题表征

子时期	子时期主题词
2008年	properties proposed models matter boron research project international graphene research
2009年	thermal carbon research graphene quantum research growth chemistry nanofibers carbon
2010年	uantum protection graphene transport materials energy project integrated research graphene
2011年	electronic adsorption system nano carbon devices dispersions platform matter materials
2012年	nanomaterials systems growth materials graphene thermal synthesis nanomaterials chemistry project
2013年	clay energy nanotube graphene carbon graphene methane quantum nanomaterials thermal
2014年	graphene instrument nanowires thermal electronic algorithms optical changes energy methaneq
2015年	materials nitride theoretical engineering electronic battery chemical nano carbon devices
2016年	system materials energy project carbon research nano diaphragm ware uantum

表6 2014项目主题相关参数(部分)

编号	项目数	主题强度	平均主题强度	主题资助额度	单项目资助额度	平均资助金额	单项目时间
topic0	11	373	351.3	3 324 227	302 202.5	387 489.3	3.3724
topic1	5	211	351.3	1 268 767	253 753.4	387 489.3	3.4997
topic2	9	298	351.3	2 189 638	243 293.1	387 489.3	2.7680
topic3	6	290	351.3	2 110 495	351 749.2	387 489.3	3.8142
topic4	5	188	351.3	2 491 099	498 219.8	387 489.3	3.4323
topic5	15	462	351.3	4 144 676	276 311.7	387 489.3	3.0889
topic6	13	309	351.3	2 887 615	222 124.2	387 489.3	1.9639

进行主题相似度计算,确定不同时间片段的主题演变过程。基于FSD模型阈值设计相关思想,提出动态阈值调控法,当阈值设置为0.4时实验相关最佳,可有效反映不同时间片段主题变化:某主题与前一主题相似度阈值大于0.4表示该主题为不同时间下的同一主题;当相似度阈值小于0.4时,表示该时间片段为新兴主题最新出现的第一时间点,进而,10个主题在不同时间的具体数值。

3.4 实验结果

根据本文3.2.1提出的基于FSD模型项目文本新兴主题探测公式,归一化处理并设置调谐系数求得新兴主题探测值 DV_i ,若 DV_i 小于1表示该主题探测值低于平均水平,若 DV_i 大于1则表示该主题探测值高于平均值,进而判断出哪些是成熟主题,哪些是新兴主题。探测结果见表7。

3.5 实验结果分析

3.5.1 探测数量

利用本文提出的新兴主题探测方法及模型公式,可以得到政府资助项目文本中蕴含的新兴主题,为更好表征该方法在项目文本新兴主题探测的效果,对实验结果更好地分析,本部分对传统主题识别方法(基准方法)探测主题进行对比分析。基准方法主题探测结果见表8。

通过对比分析,基于FSD模型的新兴主题探测得到的新兴主题数为4个,而传统的新兴主题识别方法探测到得2个。从新兴主题的数量来看,本文提出的探测方法要优于传统方法,见表9。

3.5.2 探测质量

利用主题探测方法得到新兴主题的质量用孤点主题数来衡量。孤点主题是指分析数据源受噪

表7 新兴主题公式探测结果

编号	08年	09年	10年	11年	12年	13年	14年	15年	16年	主题类型
topic0		0.6499	0.319367	0.819853	0.675011	1.07302	1.01502	1.69705	1.01579	新兴主题
topic1	1.0702	1.8620	1.99136	2.23416	1.04857	2.26261	2.42974	3.58721	1.38371	热门主题
topic2	3.2810	2.5759	1.12199	1.41182	2.39583	0.798531	0.423793	0.589531	0.799324	衰老主题
topic3	1.0985	2.5786	1.06413	0.851418	2.39583	1.07302	0.703419	1.18339	0.549821	噪音数据
topic4				0.892114	0.675011	2.14895	2.50246	1.95712	1.79609	新兴主题
topic5			0.380233	0.197045	0.531045	0.893205	0.033062	0.717097	0.242628	潜在新兴
topic6	0.8370	2.7920	1.99136	1.41182	0.281458	0.84818	1.44242	0.970248	0.952771	噪音数据
topic7	1.3735	2.7942	1.20131	1.30842	0.750168	0.798531	0.237931	0.958953	1.38371	衰老主题
topic8				0.8921	0.8146	1.9732	2.4297	1.4698	1.4118	新兴主题
topic9	0.8370	0.6499	0.8023	0.4118	1.3241	1.8952	1.0150	1.6971	1.3837	新兴主题

注:图中加粗数字表示 DV_i 高于1

表8 基于基准方法的新兴主题探测结果(主题强度值)

编号	08年	09年	10年	11年	12年	13年	14年	15年	16年	主题类型
Topic0		108	107	114	125	116	462	1 051	333	新兴主题
Topic1	258	910	379	333	152	933	400	261	1 208	热门主题
Topic2	224	193	379	472	921	177	309	349	175	噪音主题
Topic3	224	138	175	158	266	221	400	300	102	噪音主题
Topic4				237	373	221	660	260	288	新兴主题
Topic5			68	214	260	149	343	387	200	潜在新兴
Topic6	76	910	189	137	194	177	660	270	287	噪音主题
Topic7	258	910	379	333	203	177	309	219	156	衰老主题
Topic8				237	194	91	462	191	472	噪音主题
Topic9	76	138	68	472	203	221	660	261	174	噪音主题

注:图中加粗数字表示主题强度高于1

声影响较大、探测数值不稳定的主题,即为噪音数据主题,该类主题在时间维度上波动较大,不能有效反映新兴主题真实变化水平和演化过程,因此,可利用孤点主题数量来判断主题探测质量的好坏,孤点主题数较低说明探测得到新兴主题质量较好,具体情况见表10。

由表可得,基准方法仅以主题强度为单一衡量指标,受噪声影响较大,不能充分反映项目文本的具体内容,如Topic9若是采用基准方法定义为噪音主题因其12年和15年主题强度(203,261)均低于平均水平(247.5,347),但研究过程中发现该主题的资助时长和资助强度(4.25年,573 830美元)均大于当年平均值(3.34年,484 501美元)。因此,基准方法存在不足,而利用本文提出的方法得到2012年和2015年新兴主题探测值分别为1.324和1.697,均高于当年平均值,因此,本文提出的新兴主题探测方法质量较高,能客观准确得反映政府资助项目文本中的主题特征属性。

3.5.3 探测时间

探测政府资助项目文本中的新兴主题的时间对于学科领域抢占科研先机、把握科技发展脉搏尤为重要,因此,尝试如何更快更早地揭示某领域最新动向一直是情报分析人员重点研究方法。本文提出基于FSD模型的新兴主题模型相对于基准方法,可以较快得识别项目文本中蕴含的新兴主题,如主题Topic0用上述2种方法探测均为新兴主题,利用基准方法探测得到的时间为2014年,而本文提出的方法其探测时间为2012年。

3.6 新兴主题发展趋势分析

基于本文提出的新兴主题探测方法及项目文本探测公式,共探测得到石墨烯领域4个新兴主题(topic0、topic4、topic8、topic9),本部分将对石墨烯领域相关研究方向及主题进行发展趋势预测分析。

新兴主题 Topic0——石墨烯光学特性研究(optical)及工艺应用研究。主要探索光与石墨烯增强交互式作用如金属—石墨烯微结构、波导—石墨烯结构(structure)等方式的研究;进一步拓展应用石墨烯电子(electron)元件应用,开发基于石墨烯和硅波导光电器件以及透明电极触摸屏等器件;探究石墨烯材料全反射结构光学(optics)性能与厚度、光数据以及超灵敏单细胞传感等方向研究;偏振吸收石墨烯层数测量与扫描成像(formation)等细微方向的研究成为新兴研究主题,该项目资助呈现明显增长趋势,未来我国可根据自身发展需求对石墨烯光学领域开展相应研究。

新兴主题 topic4——石墨烯基础性能与工艺(technology)研究和工业化应用(application)开展。该主题主要围绕石墨烯电学、光学与机械等基础性能的实验与工业界相融合,如石墨烯薄膜改进润滑和减小摩擦(friction)效能研究,与金属氧化物以及其他复合材料活物在超级电容器方面应用以提高石墨烯和电容活性物质协同作用机制;石墨烯与金属离子聚合物复合材料制备人工肌肉等生物仿生(biomimetic)技术研究,包括磺酸化石墨烯薄膜亲水性研究探测;石墨烯与树脂、乳液等

表9 探测数量对比

方法	主题	主题数	探测度
基准方法(主题强度)	topic0、topic4	2个	一般
基于FSD模型探测方法	topic0、topic4、topic8、topic9	4个	较好

表10 探测质量对比

方法	孤点主题	孤点主题数	质量对比
基准方法(主题强度)	topic2、topic3、topic6、topic8、topic9	5个	一般
基于FSD模型探测方法	topic3、topic6	2个	较好

成膜材料(membrane)及溶剂相互作用研究,探索在功能涂料和特殊涂料(coating)的应用研究;作为导体材料与半导体及改性导电塑料(plastic)的开展与研究。该主题具有较强的灵活性和应用场景,有效探索和寻找有价值的石墨烯应用空间对于基础研究的开展也具有较强的带动作用,在未来的研究中逐渐增强说明国家科技政策制定者对于工业化应用的强调和重视,未来发展前景良好。

新兴主题 topic8——石墨烯复合物的性能(property)探究。该研究主题由传统石墨烯制备工艺演化而来,主要围绕石墨烯聚合复合材料(composite)生成如石墨烯层状聚合物、功能化聚合物以及填充聚合物3种类型以拓展石墨烯在液晶元器件、催化载体以及能量存储(energy)等场景的应用;在传统制备工艺基础上创新聚合方法比如等离子增强化学沉积、溶剂热法等,提升石墨烯材料热耗散及热导率等热学特性,探索石墨烯与无机纳米材料性能研究减少石墨烯片层薄膜(membrane)相互作用提高性能稳定性,制备工艺有水热法、溶胶凝胶法以及热蒸发法等方法;石墨烯表面特性(surface)化学催化性能研究、石墨烯卷曲限阈效应研究等。探究石墨烯表面催化对于多种碳催化研究具有一定积极意义。此主题近年来生命力较为旺盛,属于未来热门主题的潜力较大,而围绕传统方法和工艺的改进提升一直以来成为学者们关注的焦点,我国应重视石墨烯基础研究,为其他应用及性能研究提出理论基础。

新兴主题 topic9——石墨烯半导体薄膜探测设备(detector)的研制与技术开展,主要围绕石墨烯复合物探测设备(equipment)、低噪声特性碳纳米管复合光探测器研发应用以及新型半导体(semiconductor)石墨烯肖特基探测设备研究等,探究受带隙以及耗尽层对石墨烯材料偏置电压设备(volt-

age)的相互作用机制,对于石墨烯探测设备灵敏度(sensibility)、光能量吸收以及弱光探测能力(detect)的研究;多重量子结构和窄带半导体探测器制备与开发应用;石墨烯太赫兹(terahertz)等纳米带探测器的仿真与结构优化,具体围绕多种方法展开如有限元分子结构力方法等多维度分析探测器性能。石墨烯高精度新型红外探测器、硅光电探测器等具有广阔的市场应用价值,目前该主题发展势头良好,研究趋势和研究热情较高,属于较大发展潜力和空间的新兴主题之一。

4 讨 论

目前研究中数据源单一、科技文献数据源未能有效拓展以及利用项目文本存在指标体系构建不足等诸多问题制约了学界对于前沿探测的科学性,因此,本文综合分析资助强度、资助时长、新颖性以及布局强度等特征提出针对政府资助项目文本的新兴主题探测公式;然后,利用PLDA模型识别文本主题并利用动态阈值调控法进行相似度动态阈值实验;继而,探测得到新兴主题并寻找到该主题最早出现的第一时间点,在原有研究基础上进一步丰富项目文本指标体系;最后,通过该方法与基准方法中探测时间、质量及数量3个维度进行对比分析及NSF资助石墨烯领域的实证研究验证本文提出新兴主题探测方法的有效性。

同时本文也存在不足之处,在项目文本特征参数提出利用新颖度以及布局量等指标量化项目文本特征,但围绕项目文本指标体系的构建仍需进一步完善;其次,进行领域内新兴主题探测受噪音数据影响较大,应考虑减少离群数据的扭转偏离并进一步拓展分析数据源。在下一步工作中,本研究将进一步考虑项目文本其他特征数据并尝试综合论文数据、政府科技报告、专利数据等多源数据交叉融合,以使得新兴主题探测更加准确。

参考文献

- 段庆锋,潘小换. 2017. 利用社交媒体识别学科新兴主题研究[J]. 情报学报,(12):1216-1223.
- 葛菲,谭宗颖. 2013. 学科领域主题新兴趋势探测方法研究:基于关键词生命周期和引文分析[J]. 情报理论与实践,36(9):78-82.
- 黄鲁成,唐月强,吴菲菲,等. 2015. 基于文献多属性测度的新兴主题识别方法研究[J]. 科学学与科学技术管理,(2):34-43.
- 黄鲁成,王静静,李欣,等. 2015. 基于文献计量的新兴趋势分析:以生物材料为例[J]. 情报杂志,(7):58-64.
- 人民网. 2015. 从“跟跑者”向“并行者”“领跑者”转变[EB/OL]. <http://theory.people.com.cn/n/2015/0906/c40531-27546326.html>, 09-06.
- 王会珍,朱靖波,季铎,等. 2006. 基于反馈学习自适应的中文话题追踪[J]. 中文信息学报,20(3):92-98.
- 王贤文,毛文莉,王治. 2014. 基于论文下载数据的科研新趋势实时探测与追踪[J]. 科学学与科学技术管理,35(4):3-9.
- 徐路路,王效岳,白如江. 2018. 基于PLDA模型与多数据源融合相关性分析的新兴主题探测研究:以石墨烯领域为例[J]. 情报理论与实践,41(4):63-69+43.
- 许振亮,郭晓川. 2011. 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,55(8):49-53.
- 杨玉莲,谢磊. 2009. 基于子词链的中文新闻广播故事自动分割[J]. 计算机应用研究,26(2):189-192.
- 殷蜀梅. 2008. 判断新兴研究趋势的技术框架研究[J]. 图书情报知识,(3):76-80.
- 张辉,周敬民,王亮,等. 2010. 基于三维文档向量的自适应话题追踪器模型[J]. 中文信息学报,24(5):70-76.
- 张美珍. 2010. 话题检测与跟踪算法的研究[D]. 北京:北京交通大学.
- 张小明,李舟军,巢文涵. 2012. 基于增量型聚类的自动话题检测研究[J]. 软件学报,23(6):1578-1587.
- 郑烨,杨若愚,刘遥. 2017. 科技创新中的政府角色研究进展与理论框架构建:基于文献计量与扎根思想的视角[J]. 科学学与科学技术管理, 38(8):46-61.
- 中国政府网. 2018. 中国制造2025[EB/OL]. <http://www.gov.cn/zhuanti/2016/MadeinChina2025-plan/index.htm>, 08-01.
- Allan J. 2002. Topic Detection and Tracking Pilot Study[M]. Boston: Springer.
- Elsayed T, Oard D W. 2005. On evaluation of adaptive topic tracking systems[C]. Salvador: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
- Hoang L M. 2006. Emerging Trend Detection from Scientific Online Documents[R]. Ishikawa: Japan Advanced Institute of Science and Technology.
- Kessler M M. 2003. Bibliographic coupling between scientific papers[J]. Journal of the American Society for Information Science & Technology,14(1):10-25.
- Kleinberg J. 2003. Bursty and hierarchical structure in streams[J]. Data Mining & Knowledge Discovery,7(4):373-397.
- Kontostathis A, Galitsky L M, Pottenger W M, et al. 2004. A survey of emerging trend detection in textual data mining // Berry M W. Survey of Text Mining Clustering Classification & Retrieval[M]. New York: Springer.
- Lo Y Y, Gauvain J L. 2002. The LIMSIS topic tracking system for TDT 2002[C]. Gaithersburg: Topic Detection & Tracking Workshop.
- Mane K K, Bärner K. 2004. Mapping topics and topic bursts in PNAS[J]. Proceedings of the National Academy of Sciences, 101(S1):5287-5290.
- Matsumura N, Ohsawa Y, Ishizuka M. 2001. Discovery of emerging topics between communities on WWW[C]. Maebashi: Asia-Pacific Conference on Web Intelligence: Research and Development.

Emerging Topics Detection and Analysis of Government Funded Projects based on FSD Model

XU Lulu¹, JIN Yang²

(1. Department of Information Resources Management, School of Business, Nankai University, Tianjin 300071, China; 2. Beijing AnZhen Hospital Affiliated to Capital University of Medical Sciences, Beijing 100029, China)

Abstract: How to capture the development trend of science and technology and track the dynamic evolution of scientific research activities efficiently and accurately has been the focus of researchers. Using the text of NSF government-funded projects as the analysis data source, this paper synthetically uses thematic models and index construction methods, explores text structure features and analyzes multi dimension of funding amount and layout intensity, and analyzes the new topic detection method of project text based on FSD model. The results show that this method can quickly and quickly identify new topics and form a mixed distribution cluster of theme words, project, and sequence numbers, the superiority of the new detection model is verified by comparing the three dimensions of detection quantity, detection quality and detection time.

Key words: emerging topics; government funded project text; FSD model; prediction analysis