

# 基于论文-专利的石墨烯领域硬科技创新技术主题识别研究<sup>①</sup>

张 楠<sup>②</sup> 赵 辉<sup>③</sup>

(中国科学技术信息研究所 北京 100038)

**摘要** 硬科技创新是原创性高精尖科技和实体经济的结合,对我国未来创新驱动发展起着巨大推动作用,有助于形成核心竞争力。识别硬科技创新主题在情报学领域仍是一个新课题,对其有效识别将有助于引导社会投资,及早挖掘新的经济增长点。本文以石墨烯领域为研究样本,结合学术型发明人的论文和专利,基于 LDA2Vec 主题模型和 K-means 聚类算法进行硬科技创新候选技术主题挖掘;通过发文强度、时序性分析筛选主题,根据学术型发明人主题关注度、高校及科研院所主题关注度、基金投资强度构建三维评价模型,筛选出最有可能成为硬科技创新的技术主题。结果表明,石墨烯电池等 7 个技术主题最有可能成为石墨烯领域内硬科技创新主题,为后续相关研究提供思路。

**关键词** 硬科技; LDA2Vec; 学术型发明人; 石墨烯

## 0 引言

我国经济发展中存在着人口红利逐渐消失、实体经济领域内投资不足、科技成果转化率不足等问题,因此学界提出硬科技(key & core technology)概念<sup>[1]</sup>。它是以实体经济为主的原创性高精尖科技,在研发过程中需要投入大量的资金和时间,分布于人工智能、航空航天、生物技术、光电芯片、信息技术、新材料、新能源、智能制造等 8 大领域中<sup>[2]</sup>,对我国未来创新驱动发展有着巨大推动作用,有助于形成核心竞争力。识别硬科技创新将有助于引导社会投资,助力相关产业发展,抢占经济制高点。

现有研究中硬科技创新技术主题的识别方法还未涉及,本研究从情报学角度融合论文与专利分析作出探索。

## 1 相关研究

### 1.1 硬科技的概念及特点

自“硬科技”概念提出以来,国内学者进行了多

角度研究,文献[3]提出硬科技创新的“硬”具体表现在这种创新能够提高生产效率、降低生产成本、提高产品质量。硬科技创新创业的实现要求 3 个背景,即研发和商业化同时产生、科学家和企业家紧密结合以及商业化得到持续风险投资<sup>[4]</sup>。文献[5]认为军用硬科技创新转民用也是一种发展路径。

国外与之对应的概念是由文献[6]提出的“深科技”,它包括先进材料、人工智能、生物工程、区块链、无人机和机器人、光电子、量子计算等 7 个领域,是比当前技术有更为显著进步的技术,其在商业上有巨大潜力,研发投入资金多、时间长。

表 1 为硬科技创新在技术、市场、人才 3 方面的特点。

硬科技创新与颠覆性创新、突破性创新有所区别。颠覆性创新是指以意想不到的方式取代现有主流技术的创新,低端或边缘市场往往是其切入点,简单、方便、便宜是其初始阶段特征。之后通过性能与功能不断改进与完善,取代已有技术开辟出新市场,形成新的价值体系<sup>[7]</sup>,技术的颠覆性能力是其主要

① 中央级公益性科研院所基本科研业务费专项(MS2020-02)资助项目。

② 女,1996 年生,硕士生;研究方向:信息与情报工程;E-mail: znozn5010@sina.com

③ 通信作者,E-mail: zhaoh@istic.ac.cn

(收稿日期:2020-11-02)

评价指标<sup>[8]</sup>;突破性创新指利用技术创新提升企业地位、重构市场格局的创新,突破性体现在其技术创

新强度上,常常基于新的科学理论产生<sup>[9]</sup>。

表1 硬科技创新特点

技术	原创性实体经济技术,研发投入大周期长,高准入技术壁垒 <sup>[2]</sup> 注重基础研究,主要由国立科研院所或高校承担研究任务 <sup>[10]</sup>
市场	回报周期长,高风险,但一旦推入市场能够带来巨大的经济效益和回报 对产业发展有较强的引领和支撑作用,面向战略性市场 <sup>[2]</sup>
人才	科技成果转化过程中科学家深度参与 <sup>[1]</sup> 一项硬科技创新成果的产生通常需要多种跨界人才的集聚 <sup>[4]</sup>

一些突破性创新、颠覆性创新也可认为是硬科技创新<sup>[2]</sup>,本研究认为这3种创新之间有共性部分,但又有细微的差别。硬科技创新与突破性创新技术复杂度更高,颠覆性创新更依赖灵敏的市场嗅觉,可能从低端市场兴起。三者发起者角色不同,颠覆性创新中新兴进入者和小公司是主要来源<sup>[11]</sup>,突破性创新更易被大公司主导<sup>[12]</sup>,硬科技创新多由高校及科研院所和科学家所牵头的企业主导。

## 1.2 技术主题识别研究

“技术主题”概念本身没有明确定义,学者们根据各自研究目的、领域、颗粒度等对“技术主题”的研究也各不相同。但技术主题识别一般要经过特点分析、数据选择、定性定量评估的过程。比如文献[7]以颠覆性创新的颠覆性和市场制定识别路线,在论

文中找到候选技术主题,通过领域内专家组会和绘制技术路线图的形式,分析技术在市场发展的可能性,从中识别颠覆性创新。文献[13]提出跃迁指数概念,从知识跃迁和论文数增长两个方面来归纳突破性创新特点,时序性分析干细胞领域内新兴主题的引文状况,归纳子主题引文曲线特征,识别突破性创新。

在图情领域,技术主题识别一般以定量分析为主,专家意见定性分析为辅。定量分析中从数据源选择来说,技术主题识别可以从单一还是多数据源进行区分。单一数据源技术主题分析方法以论文或专利为研究样本挖掘技术主题,其识别方法如表2所示。值得一提的是不同于论文的引用目的,专利引用中通常也为了证明被引用专利的技术缺陷等问题。

表2 论文或专利主题识别方法

识别维度	识别层面	具体方法	特点
共现关系	基于词汇	关键词(德温特代码分类)共现	数据源单一对数据 信息挖掘较为局限
	基于作者	作者(专利权人)合著网络、作者(专利权人)共被引	
	.....	.....	
	科学知识 图谱方法	综合上述可视化呈现	
引用关系	基于 参考文献	频次统计、共被引分析、耦合分析	滞后性、不能深入 挖掘文本语义、 专利引用目的与 论文不同
文本挖掘	主题模型 分析方法	潜在语义分析(Latent Semantic Analysis, LSA)模型、概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)模型、文档主题生成(Latent Dirichlet Allocation, LDA)模型、词向量(Word to Vector, Word2vec)模型、文档词向量模型(Latent Dirichlet Allocation to vector, LDA2Vec)	从语义层面理解 技术主题能够 更加精确

题从而体现出本专利技术的进步,所以专利引用不能与论文引用同等看待<sup>[14]</sup>。

但如今单一数据源越来越不能展现领域内技术发展状况,论文与专利之间存在紧密关联,挖掘二者

之间的关联关系能够反映研究主题的科学和技术成果,论文与专利之间的关联分析方法如表 3 所示,根据关联层级可分为基于外部与内部关系两种,本研究也选择论文-专利关联分析方法挖掘技术主题。

表 3 论文-专利关联分析方法

关联层级	方法	作用
外部	引用关系识别 <sup>[15]</sup>	探讨基础研究与应用研究之间的关系以及其知识转移状况
	主题相似度测量 <sup>[16]</sup>	发现不同数据源之间的主题关联
内部	基于共性字段混合进行文本挖掘主题分析 <sup>[17]</sup>	随着科学与技术边界越来越模糊,相较于单一数据源对于技术主题识别更具参照性

### 1.3 主题模型算法研究

传统文本聚类方法存在向量维度过高、数据稀疏等问题,因此主题模型算法应运而生。主题模型算法发展过程如表 4 所示,每一阶段后续算法都对

前者进行改进,近年来比较流行的算法为融合了 LDA 和 Word2Vec 优点的 LDA2Vec 主题模型算法,它既能够考虑文本上下文含义又能够考虑潜在语义关系,是一种比较理想的模型。

表 4 主题模型算法发展

模型名称	提出时间	基础思路	局限性
LSA <sup>[18]</sup>	1990	文本映射向量空间	不考虑词汇在文本中出现概率
PLSA <sup>[19]</sup>	1999	基于概率统计模型将文本映射到向量空间	不能对新加入样本进行概率分布,数据样本多时容易过度拟合
LDA <sup>[20]</sup>	2003	引入狄利克雷先验概率挖掘潜在语义关系	不考虑文本上下文含义
Word2Vec <sup>[21]</sup>	2013	神经网络模型,包含词袋和 Skip-Gram 两种架构,关注词汇上下文共现	不挖掘潜在语义
LDA2vec <sup>[22]</sup>	2016	既考虑文本上下文含义又考虑潜在语义关系	/

## 2 基于论文-专利的硬科技创新技术主题识别方法构建

### 2.1 数据选择

硬科技创新不仅要求扎实的基础研究,更要求商业化实现技术成果,所以对论文和专利的选取非常有必要,且硬科技创新企业要求创始人有很强的科学背景,需要科学家领袖支撑。硬科技创新创业团队中需要有学识渊博、精于实验设计、拥有长期行业操作经验的科学领军人物<sup>[2]</sup>。他们同时拥有论文和专利产出,文献[23]称这类学者为学术型发明人(academic inventors)。他们的学术成果有统一的知识基础,能够从论文层面反映出技术的发展阶段

和趋势,不同技术流派的关注度、参与度等,从专利层面反映相应制备技术和工艺路径的发展成熟度、拟产业化的竞争态势等。文献[24]以艾滋病药物治疗整合酶抑制剂研究领域为样本,识别出其中学术型发明人,通过他们的专利构建论文与专利之间的引用知识网络,证明他们对领域内的知识流通起到重要作用,是领域内高影响力作者。

本研究中通过中心性指标和发文强度识别领域内核心作者,从中找出同时拥有论文和专利产出的学者为学术型发明人,并考虑作者的唯一性识别,文献[25]认为论文和专利中同属同一机构为同一作者。文献[26]利用 Soundex 设计算法“姓名游戏(Name Game)”,以 USPTO 内的金融领域专利作为研究样本证明该思路的可行性。因此本研究也根据

机构来唯一性识别学术型发明人。

## 2.2 数据处理

本研究使用 LDA2Vec 主题算法识别学术型发明人的论文和专利摘要主题,该算法对数据质量要求较高,需进行分词和词性标注、去停用词、词干提取等数据处理过程和大规模主题训练。此外,删除专利摘要字段中如“声明(Claim)”、“版权声明(如 XXX Elsevier……reserved 字样)”、“优点(ADVANTAGE)”等说明字段。

## 2.3 候选硬科技创新技术主题识别方法设计

硬科技创新特点之一是研究时间跨度较长,所以先对论文和专利的语料按照年份进行时间切片处理识别主题,时间间隔为 1 年,观察主题在时间上的分布。

LDA2Vec 主题模型算法既能提高计算词间相似度的准确率,也在一定程度上扩充文本的语义特征,能生成具有可解释性的概率矩阵,是一种比较理想的主题模型算法。文献[27]采集包括农业、经济、工程等在内的 14 万多篇论文摘要,利用 LDA2Vec 主题模型训练数据集,之后再用常规的聚类方法 K-means 聚类,并最终证明两种聚类方法的结合效果要优于单独用主题模型聚类的方式。因此本研究也将采用该方法使用 LDA2Vec 主题模型算法识别每一年份下主题,并使用 K-means 聚类,合并相似主题,观察主题在时间上的连续性。

## 2.4 硬科技创新候选主题筛选方法设计

根据硬科技创新特点构建能够匹配某一技术主题下的学术型发明人技术主题关注度、技术主题基金投入强度、高校及科研院所关注度的三维评价模型,使其能综合反映技术主题的硬科技创新特性。

### (1) 学术型发明人技术主题关注度

该指数越高代表学术型发明人对该主题的关注度越高,能够引导领域内其他学者跟进研究,其越有可能成为硬科技创新技术主题。该指数计算如式(1)所示。

$$Q_d = \frac{\sum_{i=1}^T Q_i}{T} \quad (1)$$

式中,  $Q_d$  表示某主题在学术型发明人数据集下的发文强度,  $T$  代表数据集数量, 每条论文或专利的摘要

数据视为 1 条,  $Q_i$  代表某主题数据集数量频次。

### (2) 技术主题基金投入强度

硬科技创新成果的产生需要长期的资金投入支持,技术专家往往需要数十年的精力才能使成果有所突破,虽然技术探索初期无法从经济收益的角度评估技术价值,但是投资资本密集度也可反映出其技术价值。我国基础研究主要由政府资助,据统计基础研究科研经费占所有科研经费的约 5.2%, 高校和科研院所占基础研究支出的 93.8%<sup>[28]</sup>。但目前 80% 的社会资本都投入了互联网行业<sup>[1]</sup>, 其他行业的投资是远远不够的, 所以目前硬科技创新投资还是以基础研究投资为主。

因此本研究将某一主题下的论文资金资助数量作为重要衡量指标之一,该指标计算如式(2)所示。

$$F_d = \frac{\sum_{i=1}^P F_i}{P} \quad (2)$$

其中  $F_d$  代表某主题下资金资助强度,  $F_p$  代表该主题下某篇论文的资金资助数量,  $P$  为论文总数。

### (3) 高校及科研院所关注度

中国目前科技力量主要被高校及科研院所支撑<sup>[1]</sup>, 他们的研究并不追求短期经济效益, 多聚焦国家战略重点任务实现和成果转化, 因此高校及科研院所研究成果转化一直是硬科技创新研究成果转化的主流方式之一。本研究认为某一主题专利下高校及科研院所关注度可作为一个评估指标, 其计算如式(3)所示。

$$C = \frac{\sum_{i=1}^R R_i}{\sum_{i=1}^T I_i} \quad (3)$$

其中,  $C$  表示高校及科研机构关注度,  $R_i$  为某主题下第  $i$  篇专利下的高校及科研院所数量,  $I_i$  为某主题下第  $i$  篇专利下的高校及科研院所总数量。

因此可构建如图 1 所示的硬科技创新识别流程。

## 3 实证研究——以石墨烯领域为例

### 3.1 数据选择与处理

在 Web of Science 数据库以检索式(TS = “graphene”)进行检索,语种为英语,文献类型为 Article,

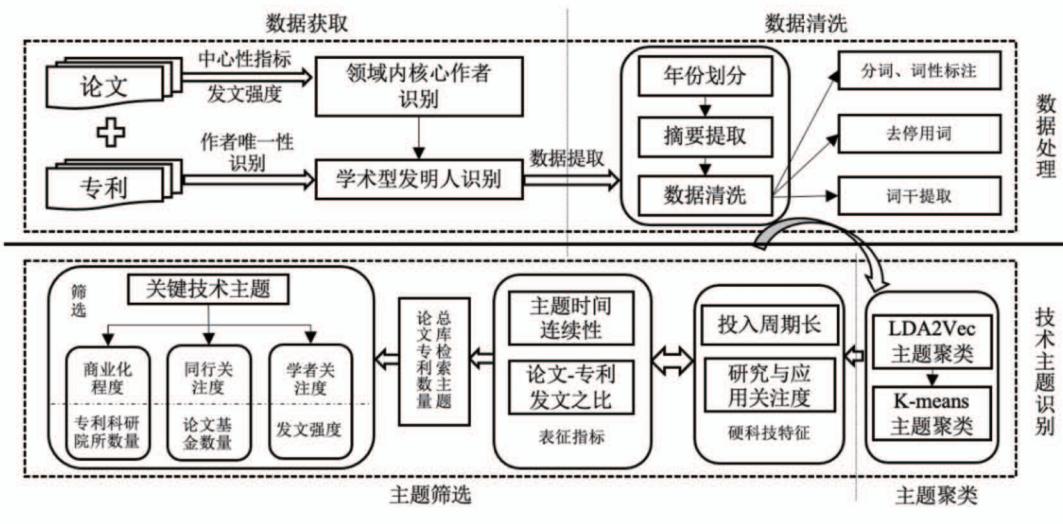


图 1 硬科技创新识别流程

检索库设定为 SCI-EXPANDED、SSCI、CPCI-S、CPCI-SSH, 时间跨度为 2009 – 2018 年, 检索日期为 2020 年 7 月 3 日, 共获取 140 823 篇论文, DII 数据库收集专利 84 696 篇, 能够包含该领域内大部分研究。

首先通过论文寻找领域内核心作者, 利用 CiteSpace 选取中心性指数  $\geq 0.05$  且发文量  $\geq 70$  的学者作为主要研究对象, 共计 49 位学者 4881 篇论文。

接下来识别学术型发明人, 通过对数据的大量观察发现, 专利的命名方式为姓(全拼) + 名(简写), 如作者“Kim, Nam Hoon”, 在专利中其命名方式为“KIM N H”, 再根据论文和专利中机构比对来唯一性识别重名。利用 MySQL 整理出学术型发明人如表 5 所示, 共计 34 位学者, 论文 3781 篇, 专利 802 个。最后使用 Python 的 NLTK 工具包来辅助实现清洗摘要文本, 汇总得到 4583 条数据。

表 5 学术型发明人汇总

序号	论文			专利		
	姓名	数量/篇	机构	姓名	机构	数量/个
1	Kim, Nam Hoon	117	Chonbuk Natl Univ	KIM N H	UNIV CHONBUK NAT IND COOP FOUND	4
2	Huang, Ying	114	Northwestern Polytech Univ	HUANG Y	UNIV NORTHWESTERN POLYTECHNICAL	15
...	.....	.....	.....	.....	.....	.....
34	Yan, Xingbin	70	Chinese Acad Sci	YAN X	CHINESE ACAD SCI	5

### 3.2 候选主题聚类

利用 LDA2Vec 主题模型识别每一时间切片下的技术主题, 抽取主题关键词并作出文档矩阵, 选取困惑度最低时的主题数目作为实验参考参数, 以 2013 年为例, 其困惑度曲线如图 3 所示, 其中主题数为 15 时困惑度最低, 约为 1.62, 因此主题数可设置为 15。

该模型运行还需设置其他参数, 如-notopics(困

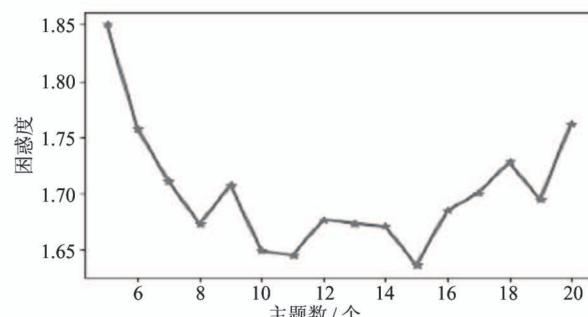


图 3 困惑度曲线(以 2013 年为例)

惑度最低时主题数)、-niter(需要训练数据数)、-twords(某一主题下关键词个数),得到每年主题词汇总,以2009年为例的主题词汇总如表7所示。

利用K-means算法聚类上述子主题。以轮廓系

数最低时最佳聚类个数设置参数,评分如图4所示,其中“\*”代表每个主题数下的具体分数值,从图中可看出,随着主题数增长分数逐渐降低,主题数约为14分时评分数最高,约为0.067。

表7 主题词汇总(以2009年为例)

年份	序号	主题词
2009	0	sheet, organic, electrode, substrate, transparent, display, light, resistance, large, process, low, efficiency, flexible, investigated, prepared
	1	graphene, sheets, layer, oxide, film, conductive, composite, electrically, making, Andor, thin, ceramic, matrix, comprising, dispersed
	...	.....
	12	observed, band, intensity, also, function, bands, strain, due, g, peak, ev, atomic, transitions, phonon, mode

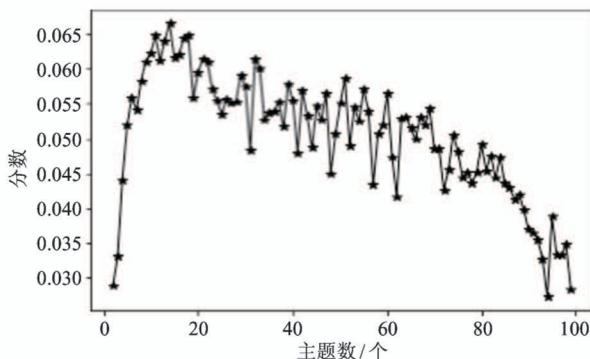


图4 主题数K-means评分

设置主题数为14,最终推断每个主题的含义汇总如表8所示。

### 3.3 硬科技创新技术主题筛选

根据硬科技创新研究周期长这一特点可排除主题11。继续检索其他主题下的论文与专利概况。以主题7(石墨烯氧化还原反应及其电催化作用)为例,该主题下共有4个子主题分别为2013-0、2014-0、2015-0、2018-0,根据子主题下主题词构建检索式,选取主题词中权重较高的词语作为检索词,并

表8 主题K-means聚类主题

序号	包含类别	含义
0	2009-0、2009-5、2010-4、2010-7、2011-5、2012-1、2014-4、2014-10、2015-2、2017-3、2018-8、2018-11	石墨烯薄膜制备
1	2009-1、2009-11、2010-1、2010-2、2010-3、2011-2、2011-6、2012-4、2013-3、2013-4、2013-9、2013-12、2013-13、2014-1、2014-6、2014-9、2015-5、2015-8、2016-6、2017-5、2017-8、2018-10、2018-14	多层石墨烯复合材料制备
...	.....	.....
13	2012-5、2013-14、2014-11、2015-4、2016-7、2017-9、2018-6	石墨烯基复合材料及其热学特性

查阅补充简写词汇,如“Graphene Oxide(GO,氧化石墨烯)”,“Chemical Modified Graphene(CMG,化学改性石墨烯)”。在检索式构建过程中发现主题6提取出的关键词无意义词汇较多,结果可能有偶然

性而被排除。查阅汇总后以每个主题下的论文和专利数量均值为基准将上述主题分为4种,大于均值可视为多(坐标为正),小于均值可视为少(坐标为负),归一化处理后展示在如图5所示的象限图上。

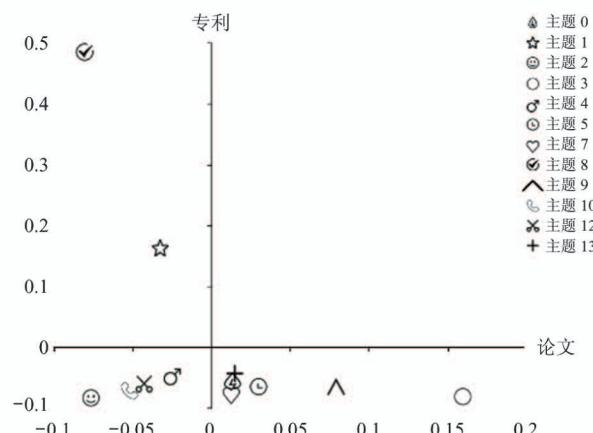


图 5 文献数量分布

(1) 第 1 象限, 论文多、专利多。该类型主题基础研究和应用都较为成熟, 属于领域内的热点问题, 未找到相关。

(2) 第 2 象限, 论文少、专利多。有主题 1、主题 8, 该象限下主题应用研究较为成熟。主题 8 的专利数远远大于论文数, 考虑其偶然性大予以排除。

(3) 第 3 象限, 论文少、专利少。有主题 2、主题 4、主题 10、主题 12。大部分学者们对于该主题关注度不够, 但在未来仍具有发展潜力。

(4) 第 4 象限, 论文多、专利少。有主题 0、主题 3、主题 5、主题 7、主题 9、主题 13。该类型主题基础研究较为深入应用研究较少, 有较大的发展潜力。

继续分析上述主题。测算各个主题专利下的学术型发明人技术主题关注度、技术主题基金投入强度、高校及科研院所关注度, 将 3 个维度的数据经过归一化处理后展示在如图 6 所示的硬科技创新技术主题预测三维模型图上。

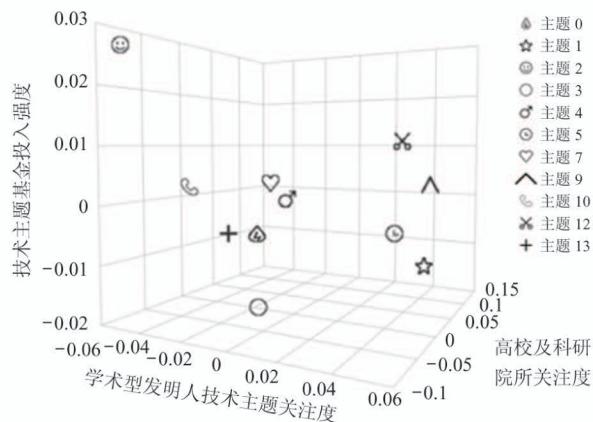


图 6 技术预测三维模型散点图

从图中可以看出, 3 项指标占比均较低的主题有主题 0(石墨烯薄膜制备)、主题 3(氧化石墨烯表征检测)、主题 13(石墨烯基复合材料及其热学特性), 可排除本研究所挖掘的硬科技创新技术主题。主题 4(化学沉积法(CVD)制备石墨烯)和主题 10(石墨烯电子结构)各项指标比较平均, 也可排除本研究所关注的研究主题。

各项指数较高的主题有主题 9(石墨烯电池, 包括超级电容器、锂电池、电极材料、离子电池、石墨烯-聚苯胺复合材料等)和主题 12(类石墨烯 2D 材料, 以二硫化钼和基于 2D 材料的异质材料为主), 这两项主题已获得较高关注, 最有可能是硬科技创新技术主题; 识别出 1 项高基金投入技术主题, 即主题 2(过渡金属二卤化物), 在未来有可能成为硬科技创新技术主题, 需持续关注其后续专利产出分析; 2 项高校及科研院所高关注度技术主题, 即主题 5(石墨烯电化学检测)和主题 7(石墨烯氧化还原反应及其电催化作用), 该类主题技术复杂度高, 在未来有可能成为硬科技创新技术主题; 1 项学术型发明人高关注度技术主题, 即主题 1(多层石墨烯复合材料制备), 在未来有可能成为硬科技创新技术主题。

## 4 结 论

硬科技创新是以实体经济为主的原创性高精尖科技, 具有在科技成果转化过程中科学家深度参与、以科研院所为主导、技术复杂度较高、需要长期资金和时间投入的特点。本文从情报学角度构建技术主题识别方法, 以学术型发明人的论文和专利作为数据源, 利用 LDA2Vec 主题模型和 K-means 算法进行硬科技创新主题聚类得出候选主题, 之后按照技术主题关注度、时序性分析等指标筛选, 并通过技术主题的学术型发明人关注度、基金投入强度、高校及科研院所关注度等指标构建关键技术识别三维模型, 识别出石墨烯领域内最有可能成为硬科技创新的技术主题。

石墨烯电池类研究和类石墨烯 2D 材料类研究各项识别指标均较高, 其在学术型发明人群体中关

注度高,专利中科院所参与程度高,且获得了较多基金投入,判断该类主题为硬科技创新技术主题,可作为短期内的投资关注技术主题参考;多层石墨烯复合材料类研究的学术型发明人关注度高,相关资金投入还未深入跟进,在未来有可能成为硬科技创新技术主题,可作为中长期投资技术主题参考;石墨烯的电子结构特性类研究和石墨烯的氧化还原反应及其电催化作用类研究的应用研究高校和科研院所关注度高,技术复杂度较高,在未来有可能成为硬科技创新技术主题,可作为中长期投资技术主题参考;石墨烯过渡金属二卤化物类研究获得了较多基金投入,但是还未有更多高校和科研院所专利化成果,在未来有可能成为硬科技创新技术主题,需持续关注后续专利成果作进一步评估,可作为长期投资技术主题参考。

## 参考文献

- [1] 米磊.“硬科技”创业的黄金时代[J].中国高新区,2016(13):26-29
- [2] 国务院发展研究中心国际经济研究所,中科创星,麻省理工科技评论,等.2019年中国硬科技产业投资发展白皮书[R].西安:中国科学院西安光学精密机械研究所,2019:1-193
- [3] 邵景峰,王希尧.西安硬科技产业投入与产出效率分析与对策研究[J].价值工程,2019,38(2):18-21
- [4] 长城战略研究所.“硬科技”创业趋势研究(上)[J].新材料产业,2019(2):67-71
- [5] 田富强.西安国家中心城市硬科技知识产权军兼民机制[J].情报杂志,2018,37(4):62-68
- [6] Arnaud De La Tour, Soussan P, Harlé N, et al. From Tech to Deep Tech[R]. Boston: The Boston Consulting Group, 2017:11-18
- [7] Kostoff R N, Boylan R, Simons G R. Disruptive technology roadmaps [J]. *Technological Forecasting and Social Change*, 2004, 71(1/2):141-159
- [8] 苏鹏,苏成,潘云涛.基于历史案例的颠覆性技术特征分析[J].中国科技论坛,2019(8):1-9
- [9] Dewar R D, Dutton J E. The adoption of radical and incremental innovations: an empirical analysis [J]. *Management Science*, 1986, 32(11):1422-1433
- [10] 郑南磊.补足“硬”科技短板[N].证券时报,2018-07-25(A09)
- [11] Govindarajan V, Kopalle P K. Disruptiveness of innovations: measurement and an assessment of reliability and validity[J]. *Strategic Management Journal*, 2006, 27(2):189-199
- [12] Chandy R K, Tellis G J. The incumbent's curse? incumbency, size, and radical product innovation[J]. *Journal of Marketing*, 2000, 64(3):1-17
- [13] 曹艺文,许海云,武华维,等.基于引文曲线拟合的新兴技术主题的突破性预测——以干细胞领域为例[J].图书情报工作,2020,64(5):100-113
- [14] 赵黎明,高杨,韩宇.专利引文分析在知识转移机制研究中的应用[J].科学学研究,2002,20(3):297-300
- [15] Shibata N, Kajikawa R, Sakata R. Extracting the commercialization gap between science and technology case study of a solar cell[J]. *Technological Forecasting and Social Change*, 2010, 77(7): 1147-1155
- [16] Wang M Y, Fang S C, Chang Y H. Exploring technological opportunities by mining the gaps between science and technology: microalgal biofuels[J]. *Technological Forecasting and Social Change*, 2015, 92(5): 182-195
- [17] 刘小玲,谭宗颖.新兴技术主题识别方法研究进展[J].图书情报工作,2020,64(11):145-152
- [18] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. *Journal of the American Society for Information Science*, 1990, 41(6): 391-407
- [19] Thomas Hofmann. Probabilistic latent semantic analysis[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, 1999: 50-57
- [20] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. *Journal of Machine Learning Research*, 2001, 3(4/5): 993-1022
- [21] 周练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济, 2015(2):145-148
- [22] Moody C E. Mixing dirichlet topic models and word embeddings to make lda2vec[J]. arXiv:1605.02019, 2016
- [23] Margherita B, Stefano B, Francesco L, et al. Networks of inventors and the role of academia: an exploration of Italian patent data[J]. *Research Policy*, 2004, 33(1):127-145
- [24] Winnink J J, Tijssen R J W. R&D dynamics and scientific

- ic breakthroughs in HIV/AIDS drugs development; the case of integrase inhibitors [J]. *Scientometrics*, 2014, 101 (1): 1-16
- [25] Boyack K W, Klavans R. Measuring science-technology interaction using rare inventor-author names [J]. *Journal of Informetrics*, 2008, 2(3): 173-182
- [26] Trajtenberg M, Shiff G, Melamed R. The ‘Names Game’: harnessing inventors’ patent data for economic research NBER Working Papers [J]. *Expert Opinion on Therapeutic Patents*, 2006, 7(11): 1297-1306
- [27] Aytug O. Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering [J]. *IEEE Access*, 2019(7): 145614-145633
- [28] 赵腾宇, 裴瑞敏, 杨国梁. 中国科技研发经费体系的发展与现状 [J]. 科技导报, 2019, 37(18): 98-108

## Identification of key & core technology innovation based on patent and paper data in graphene field

Zhang Nan, Zhao Hui

(Institute of Scientific and Technical Information of China, Beijing 100038)

### Abstract

The key & core technology innovation is a combination of real economy and advanced technology. It also plays an important role in promoting future innovation-driven development and forming core competitiveness. At present, identifying the key & core technology innovation is still a new topic in the field of information science. Effect identification of key & core technology innovation will help to guide the social investment, and find out new economic growth point. Graphene field is used as the study sample, the technical theme is identified based on the LDA2Vec theme model which used papers and patents of the academic inventors. The themes are selected by the intensity and timing analysis, and a 3D model is made according to the attention of academic leaders, the degree of commercialization of scientific research institutions and the intensity of the fund amount. It is concluded that 7 technical topics including graphene batteries are most likely to become the key & core technology innovation topics, providing ideas for subsequent related researches.

**Key words:** key & core technology, LDA2Vec, academic leader, graphene