● 许学国，桂美增（上海大学管理学院，上海 200444）

基于GTM逆向映射的技术创新机会识别*
——以新能源汽车为例

摘 要：［目的/意义］技术创新机会识别能够帮助企业更好地实现技术布局战略，从而抢占技术制高点以形成科技创新的非对称优势。［方法/过程］文章提出了一种基于生成式拓扑映射（The Generative Topographic Mapping，GTM）的技术创新机会识别方法。结合深度学习、机器学习等方法从SCI论文及德温特专利数据入手进行技术创新机会识别。首先构建基于Bert-LSTM文本分类模型，实现对论文及专利数据的主题识别；然后运用自然语言处理（Natural Language Processing，NLP）技术中词性标注对论文和专利摘要文本进行挖掘，从而识别出不同研究领域的关键技术词；最后采用GTM算法绘制技术地图，通过对空白新兴技术点逆向映射进而实现对技术创新机会的有效识别。［结果/结论］以新能源汽车为例进行验证，结果显示基于Bert-LSTM文本分类模型的预测精度优于其他分类模型；基于论文和专利数据绘制的技术地图，能够直接呈现出不同数据源所涉及的研究范围，通过GTM算法逆向映射实现了快速、客观识别技术创新机会，从而为企业技术布局提供数据支撑。本研究为客观识别技术创新机会提供了新的解决思路，以期为相关研究者提供有益启示。

关键词：数据挖掘；深度学习；技术创新机会；技术地图；主题识别

DOI：10.16353/j.cnki.1000-7490.2021.06.021

引用格式：许学国，桂美增. 基于GTM逆向映射的技术创新机会识别——以新能源汽车为例［J］. 情报理论与实践，2021，44（6）：146-153，198.

Identifying Technology Innovation Opportunities Based on GTM Reverse Mapping: Taking the Case of New Energy Vehicles

Abstract: [Purpose/significance] Technological innovation opportunity identification can help enterprises to better realize the technology layout strategy, therefore to seize the technological high ground to form the asymmetric advantage of scientific and technological innovation. [Method/process] In this paper, an approach is offered to identify innovation opportunities based on The Generative Topographic Mapping (GTM). And the deep learning and machine learning methods are applied to identify technological innovation opportunities predicated on SCI papers and Derwent patent data. Firstly, we build a classification model based on Bert-LSTM to identify the topics of papers and patent data. Secondly, we use the natural language processing (NLP) technology to mine the abstracts of papers and patents to identify the key technology words in different research fields. Finally, we apply the GTM algorithm to draw a technology map, thus identifying the technology innovation opportunities effectively through the reverse mapping of the blank emerging technology dots. [Result/conclusion] This paper takes new energy vehicles as an example for validation, and the results show that the prediction accuracy of the Bert-LSTM text-based classification model is superior to other classification models. Further more, the technology map based on the papers and patent data can directly present the scope of research involved in different data sources, and the GTM algorithm reverse mapping enables rapid and objective identification of technology innovation opportunities, thus providing data support for enterprise technology layout. Ultimately, this study offers a new solution for identifying technology innovation opportunities objectively, with a view to providing useful insights for relevant researchers.

Keywords: data mining; deep learning; technological innovation opportunities; technology map; topic identification

* 本文为国家自然科学基金青年项目“关键利益相关者视角下新兴产业创新政策作用机制与仿真优化：以新能源汽车为例”的研究成果，项目编号：71704101。

技术创新是推动社会及经济发展的重要动力，论文与专利所蕴含的信息作为技术创新的知识载体越来越受到企业和政府的重视。近年来随着我国科技事业的蓬勃发展，技术创新机会识别逐渐受到学界及业界的关注。技术地图是挖掘技术信息必不可少的手段，它以简洁直观方式将海量数据之间的复杂关系展示出来，技术地图中的空白区域表示未来可能发展的新技术即技术创新机会[1]。通过对技术创新机会有效识别能够帮助企业集中优势资源攻克重点领域，打破技术桎梏。现阶段随着互联网、5G 与大数据的蓬勃发展，人工智能、深度学习等新一代信息技术的系统性突破预示着新一轮科技革命和产业革命的到来。目前基于数据驱动的决策、识别、战略研究越来越被欧美国家所推崇。在科技演化进程不断加快的背景下，运用新技术从海量数据中实现对技术创新机会的有效识别，无论对于国家还是研发机构的技术发展都具有重要意义。因此，本文借鉴国际上先进的数据挖掘方法，综合运用深度学习、自然语言处理等方法对论文和专利数据进行挖掘进而绘制技术地图，实现了从海量数据中识别技术创新机会。研究结果为企业技术研发指明方向，帮助企业在未来国际竞争中占据有利地势。

1 研究现状

技术创新机会识别是对科学、技术未来发展的系统性研究，其目标是从海量数据中挖掘出技术未来存在的发展方向。以往学者常采用专家意见法对技术机会进行识别分析，Chen 等采用德尔菲法对中国 2030 年可再生能源发展战略的关键技术要素进行识别[2]。Alexander 等运用德尔菲法对养老院信息技术进行评估，最后对处于成熟阶段技术的未来发展进行分析[3]。随着信息技术不断发展，学者们逐渐采用数据挖掘方法进行研究。数据挖掘是从海量数据中挖掘信息情报的理论和方法，是信息科学中关键技术和主要手段。通过对某一领域的数据挖掘，可以准确、全面了解该领域技术现状及发展趋势。

在技术创新机会识别领域常用的数据挖掘方法有 4 种，一是通过引用关系来进行分析，Woo 依据专利关键词空间向量和专利引用次数运用 K 值临近算法实现了对技术开发过程中早期创意的价值识别[4]。Yoon 运用专利文献关键词和引文数据构建专利地图实现了技术发展路径预测[1]。二是针对文本信息进行挖掘，运用 NLP 技术对文献的文本信息进行分析。Rezaeian 等以通风技术为例采用文本挖掘与聚类分析相结合的方法对知识距离进行分析，进而实现技术创新机会识别[5]。许学国等运用深度学习方法对 SCI 论文摘要文本进行挖掘，识别了机器人技术的主要技术领域并对各领域未来发展进行分析[6]。三是共现分析，施萧萧等采用共现分析方法对国内外颠覆性技术发展现状及未来趋势进行分析[7]。Park 采用 IPC 共现信息构建技术知识流网络，运用社会网络分析方法确定技术的核心性与中介性，从而帮助国家对技术研发进行合理性规划[8]。四是采用技术空白点，技术空白点是指现有论文或专利中还未涉及的技术组合。以往学者常采用主成分分析法（Principal Component Analysis，PCA）和自组织映射方法（Self Organizing Map，SOM）实现技术空白点识别。由于这两种方法严重依赖于研究人员主观经验，其识别结果存在因人而异的情况，因此客观性较差。Son 等[9]提出了新的技术空白点识别方法——生成式拓扑映射，该方法克服了以往方法存在的主观识别和解释技术空白不足的缺陷。我国学者吴菲菲等[10]运用 GTM 算法绘制专利地图实现了陆地无线接入技术标准的空白点识别，结果验证了该方法可以很好地应用于技术创新机会识别。

可见，现阶段国内外学者对技术创新机会识别已进行了深入研究，而针对技术创新机会识别的方法研究却略显不足。首先不同文献类型表征了技术发展的不同阶段，SCI 论文能够表征技术研究的基础阶段，而专利数据表征技术研究试验发展阶段[11]，已有文献常常采用一种数据进行分析，鲜有学者运用多维数据进行综合挖掘。此外，技术创新机会识别对新技术开发至关重要，目前国内公开文献中还没有针对新能源汽车技术创新机会识别的研究成果。最后，近年来随着深度学习在自然语言处理领域取得飞速发展，借助深层神经网络，对文本集合中所蕴含的知识进行学习，进而通过文本分析实现数据挖掘取得了一定成果，但运用深度学习从语义分析角度实现技术创新机会识别的研究仍比较少。基于此本文以新能源汽车为例，从 SCI 论文和专利数据入手，综合运用深度学习、自然语言处理和技术地图等方法实现技术创新机会识别。该方法首先通过基于 Bert-LSTM 文本分类模型，对论文及专利研究主题进行识别，然后采用自然语言处理中的词性识别技术结合专家意见识别出各研究领域的关键技术词，最后运用 GTM 算法绘制技术地图通过对技术空白点逆向映射实现对技术创新机会的识别。本研究为技术创新机会识别提供了新的解决思路，以期为相关研究者提供有益启示。

2 模型方法

2.1 基于 Bert-LSTM 文本分类模型

2.1.1 训练集和测试集的构建 首先从 Web of Science 核心论文数据库与德温特专利数据库中检索下载新能源汽车领域的全部论文及专利数据。随后采用 Python 语言对数据进行清洗，主要包括剔除空白与重复数据，保证数据一致性。然后构建文本分类的训练集与测试集，本文根据查

阅2019年中国工程前沿报告、2019年世界新能源汽车大会报告并结合相关学者研究成果[12-13]，最终确定了新能源汽车领域8项具有发展潜力的技术领域。由于原始数据较大，完全采用人工筛选的方式较难实现且较为耗时，因此本文采用特征词初筛与人工标注结合的方法筛选训练样本，详细过程如图1所示。首先通过咨询专家确定各技术领域1～2个特征词。随后采用Python语言将专利与论文摘要文本数据与特征词进行匹配，剔除摘要中不包含特征词的数据，获得初始数据集。然后采用两人独立背靠背的数据标注方法，通过阅读初选数据样本的摘要及标题对其所属类别进行标注，获得数据样本。最后依照8∶2的比例从已标注好的数据样本中随机抽取数据，获得训练所需的训练集与测试集。

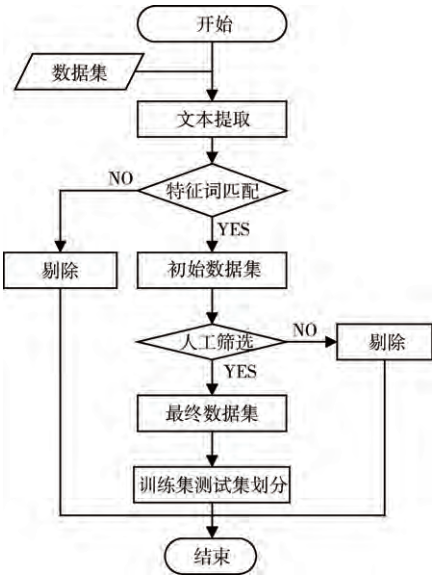

图1 训练集与测试集选取流程图

2.1.2 基于Bert-LSTM文本分类模型 Bert（Bidirectional Encoder Representation from Transformers）是谷歌公司AI团队在2018年10月发布的一种基于深度学习的语言表示模型。研究表明通过预训练的词嵌入方法比传统文本编码方式更有利于NLP分析[14]。常用的预训练模型有OpenAI GPT、ELMo和Bert等，ELMo通过构建词嵌入动态调整的双向神经网络，从而实现了上下文特征信息提取。Bert相比于ELMo模型进一步拓宽了词向量的泛化能力，增强了字向量的语义表示能力，因此Bert的性能更为优异[15]。

基于此本文构建Bert-LSTM文本分类模型，从而实现对论文及专利的主题识别，该模型结构如图2所示。首先使用谷歌预训练完成的Bert模型（BERT-Base，Uncased: L-12_H-768_A-12）（https://github.com/google-research/bert）对训练样本进行预训练，从而获取Bert模型对输入句子的向量化结果。随后将句子的向量化结果输入到单层

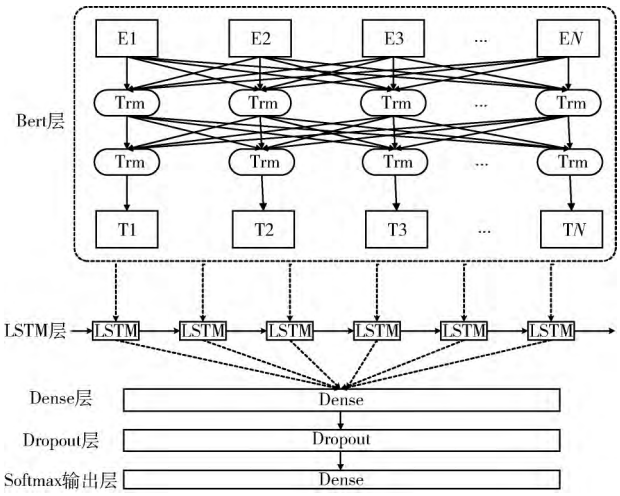

图2 Bert-LSTM模型示意图

LSTM，然后添加全连接层和减少过拟合的droupout层。最后添加激活函数为“softmax”的全连接层获得最终识别结果。在训练完成后，利用测试集数据对模型性能进行测试。模型在训练过程中冻结Bert模型参数，只对LSTM和全连接层的参数进行训练，从而使模型在训练过程中既保证预测精度又可以相对减少训练参数，进而缩短训练时间。

2.2 基于数据挖掘的技术创新机会识别

2.2.1 词性标注 词性标注（Part-of-speech Tagging）是指对每个分词确定词性的过程，是一种文本语义挖掘方法。最初采用规则方法进行词性标注，但该方法存在构造规则耗时及脱离语境无法准确选择的问题。现阶段随着统计学在计算语言学中的不断发展，SVM、最大熵和深度学习等方法也被用于词性标注，并取得了不错的效果。

通过词性标注，每一个词都有了词性信息，如RB表示副词、NN表示名词。这些信息可以成为计算机识别的类别标签，该方法可以快速准确地识别出文本中含量最多的信息，实现各种格式的文本语义挖掘。基于此，本文通过词性标注对论文及专利摘要进行文本挖掘，从而快速识别出技术领域中关键技术词，进而实现技术创新机会识别。

2.2.2 基于技术地图的技术创新机会识别 随着技术竞争的日趋激烈，技术地图已经成为各国技术挖掘的有效工具[16]。技术地图是由技术数据信息构成，通过算法将原始数据从高维数据空间映射到低维正则网格上，地图中空白点表示该点并没有与之对应的技术组合，即为技术空白点，通过对技术空白点分析可以更为有效实现技术创新机会识别。

采用技术地图进行空白点识别，关键是绘制出客观的技术地图并对空白点进行科学解读。PCA、SOM及GTM

(见图3) 均能绘制出客观科学的专利地图，但在空白技术点解读方面，PCA 与 SOM 均需要借助专业人员进行主观判断。而 GTM 算法具有反向映射技术地图的能力，采用 GTM 绘制出的技术地图能够将地图中的空白技术点反向映射到原始技术空间，进而获得空白技术的技术组合(见图4)。

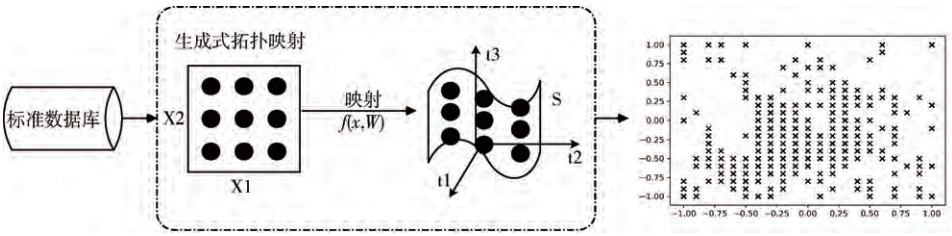

图3 生成式拓扑映射过程

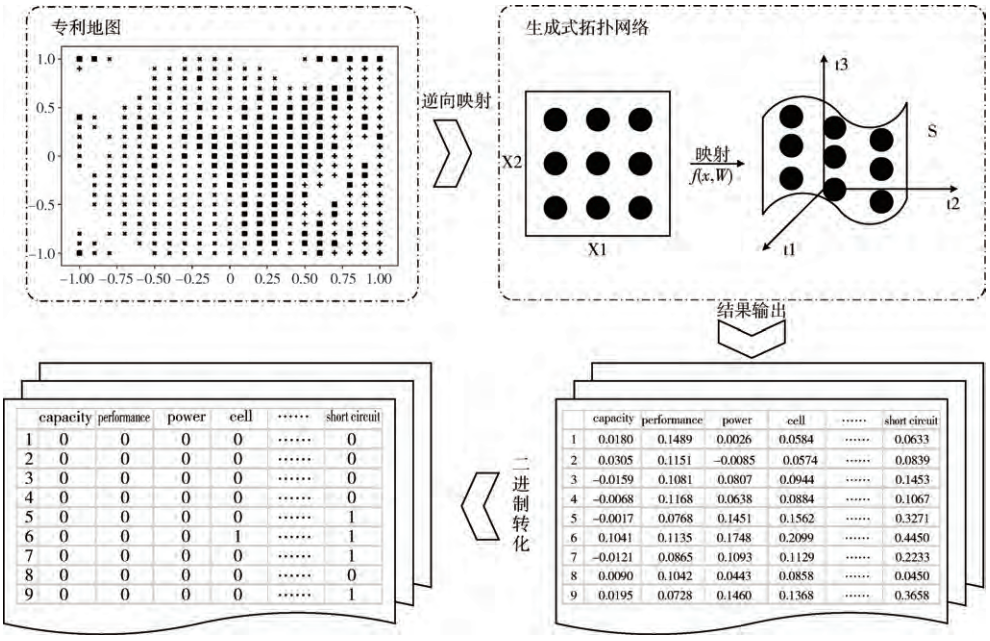

图4 GTM 逆向映射示例

由于 GTM 算法具有反向映射技术地图的能力，采用该方法绘制出的技术地图能够将空白技术点反向映射到原始技术空间，进而获得空白技术组合，其结果更具客观性[10]。基于此本文创新性采用 GTM 算法对 SCI 论文数据和专利数据进行综合分析，进而绘制出同时反映论文数据和专利数据的技术地图，最终获得更为客观、全面与科学的技术空白点，通过逆向映射实现对技术创新机会的识别。

3 实验及分析

3.1 数据收集与预处理

3.1.1 **数据采集** Web of Science 数据库是国际公认的反映科学研究水准的数据库，其中包括 SCI、SSCI 等引文索引数据库，Derwent Innovations Index (DII) 专利数据库。本文采用 TS = (New energy vehicle OR New energy automobile OR Battery Electric Vehicle OR Battery Electric automobile OR pure electric vehicle OR pure electric automobile OR hybrid electric vehicle OR hybrid electric automobile) 检索式在 SCI 核心论文数据库和 DII 专利数据库中分别进行检索，时间设置为 1968—2019 年，检索时间为 2020 年 8 月 11 日。最终获得“Article”论文 15966 篇和 71064 项专利(包含专利族) 数据。

DII 数据库提供的专利信息包含专利族，专利族中的基础专利是某组织最早申请的专利，随后该组织继续申请的相似技术信息专利都会归为该专利族中。由于现阶段鲜有可以直接对 DII 专利族进行处理的软件，为了保证研究结果科学性与严谨性，本文使用 Python 语言编写程序实现对德温特专利族数据的提取与处理，从而获得更加准确、全面的专利信息(共获得 169280 项)，并将处理结果用于之后的专利分析中。

3.1.2 **文本数据预处理** 文本数据预处理是文本挖掘的基础工作，通过对摘要文本数据的预处理，实现了摘要数据的规范化和结构化，使其能够作为 NLP 的输入数据，数据预处理的好坏直接影响了文本分析的结果质量，因此摘要预处理就显得尤为重要。本文使用 Python 语言对提取到的 SCI 论文与专利摘要进行分词，除去停用词等预处理，进而保证模型结果质量。

3.1.3 **训练集与测试集构建** 本次实验选择新能源汽车领域的论文及专利数据进行方法有效性验证。依据《2019 年中国工程前沿报告》《2019 年世界新能源汽车大会报告》及相关学者的研究成果[12-13]，并结合专家意见，最终选择了 8 项新能源汽车未来核心技术领域“锂离子电池”“燃料电池”“无线充电”“快速充电”“逆变器”

“电池管理系统”“镍氢电池”和“永磁电机”进行基于Bert-LSTM的文本分类模型实验。为了加快人工筛选实验训练集及测试集，结合专家意见对各技术领域选择了具有代表性的特征词（见表1），通过特征词与论文及专利摘要文本信息匹配，进而实现对待分析文本数据筛选，最终获得实验的初始数据集。

表1 核心技术领域与特征词

序号	技术领域	特征词
1	锂离子电池	Lithium ion Battery；Li-ion Battery
2	燃料电池	Fuel Cell；Fuel Cells
3	无线充电	Wireless Charge；Wireless Charging
4	快速充电	Fast Charging；Quick Charge
5	逆变器	Inverter
6	电池管理系统	BMS；Battery Management System
7	镍氢电池	NiMH Battery；NI-MH Battery
8	永磁电机	PM Motor；Permanent Magnet Motor

获得初始数据集后，本文采用2人独立背靠背的数据标注方法对初选数据的所属类别进行人工标注，为了保证识别结果的准确性，将剔除的不相关噪声文件加入其他研究领域进行训练。各个研究领域的数据样本，训练集与测试集详情如表2所示。

表2 数据样本、训练集、测试集

技术领域	数据源	数据样本	训练集	测试集
锂离子电池	论文	465	371	94
	专利	617	480	137
燃料电池	论文	536	434	102
	专利	728	577	151
无线充电	论文	275	221	54
	专利	616	485	131
快速充电	论文	336	268	68
	专利	543	434	109
逆变器	论文	324	260	64
	专利	497	384	113
电池管理系统	论文	411	329	82
	专利	568	426	142
镍氢电池	论文	300	240	60
	专利	536	429	107
永磁电机	论文	366	292	74
	专利	531	415	116
其他	论文	776	627	149
	专利	845	672	173
合计	论文	3789	3042	747
	专利	5481	4302	1179

3.2 文本分类结果对比分析

文本分类常用的分类模型有决策树（Decision Tree，DT）、支持向量机（SVM）、贝叶斯（Native Bayes，NB）和集成算法Bagging、Xgboost。为了对比分析本文构建的Bert-LSTM文本分类模型的分类效果，使用决策树、支持向量机、贝叶斯、Bagging和Xgboost算法构建文本分类模型。进而对各文本分类模型的分类效果进行评估。

由于文本数据是一种非结构化的数据，为了使计算机能够处理文本信息，需要将文本数据转化为便于计算机处理的结构化数据。因此使用LDA主题模型对专利和论文摘要文本进行主题分析，提取出摘要中的特征词，作为各机器学习分类算法的特征向量。LDA模型参数依据摘要文本主题困惑度、模型一致性得分[17]最终确定主题数K为2，特征词为20。参数α、β设置为经验值，分别是$\alpha = 50/K$，$\beta = 0.1$[18]，各分类模型准确率、召回率、$F1$值见表3。

表3 文本分类模型平均准确率、平均召回率、平均$F1$值

数据源	分类模型	准确率	召回率	$F1$值
论文	LDA-DT	0.781	0.775	0.778
	LDA-NB	0.713	0.577	0.638
	LDA-SVM	0.777	0.754	0.765
	LDA-Bagging	0.801	0.793	0.797
	LDA-Xgboost	0.802	0.799	0.800
	BERT-LSTM	0.812	0.806	0.809
专利	LDA-DT	0.753	0.752	0.752
	LDA-NB	0.778	0.728	0.752
	LDA-SVM	0.788	0.778	0.783
	LDA-Bagging	0.775	0.764	0.769
	LDA-Xgboost	0.797	0.785	0.791
	BERT-LSTM	0.843	0.832	0.837

根据表3可以发现，本文采用的Bert-LSTM文本分类模型无论在论文数据和专利数据，该模型的预测准确率，召回率及$F1$值均优于其他常用的文本分类模型。结果显示本文提出的Bert-LSTM文本分类模型能够很好地适用于论文与专利主题分类，该方法提升了文本分类精度，未来能够很好地应用于文献主题分类领域。

本文采用最优Bert-LSTM文本分类模型对新能源汽车领域所有SCI论文及专利数据进行主题识别，进而获得各技术领域的已公开发表的SCI论文与专利情况，具体结果如表4所示。

表4 研究领域详情

序号	技术领域	论文	专利
1	锂离子电池	2139	15390
2	燃料电池	1751	7402
3	无线充电	432	2549
4	快速充电	542	2282
5	逆变器	891	8784
6	电池管理系统	1099	1418
7	镍氢电池	570	4871
8	永磁电机	785	2775
9	其他	7757	123810

3.3 基于数据挖掘的技术创新机会识别

根据1996—2019年各研究领域SCI论文发表和专利

申请数据，本文绘制出各研究领域逐年论文发表和专利申请的时间序列图（如图5、图6所示）。从图中可以发现，各技术领域论文发表逐年递增，由于专利申请到公示的周期较长，导致各技术领域的专利申请数量出现了近几年骤减的状况。为了更加深入挖掘各技术领域未来发展方向，本文以论文及专利申请数量最多的锂离子电池为例，对锂离子电池技术创新机会进行识别，进而分析出该技术未来发展方向。

3.3.1 技术词提取 词性标注是自然语言处理过程中重要的步骤之一，通过词性标注可以实现对SCI论文摘要及专利摘要的文本挖掘，进而识别出各研究领域的技术特征词。本文首先使用Python语言运用自然语言处理中常用的工具包（NLTK）对SCI论文摘要及专利摘要文本进行词性标注，随后筛选出SCI论文摘要及专利摘要文本中的名词与专有名词，并统计各技术名词出现频率。然后根据各名词出现频率并结合专家意见最终筛选出技术领域的关键技术词，表5是以锂离子电池技术为例筛选出的该领域100个关键技术词。

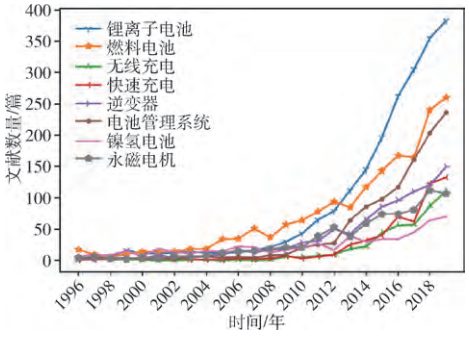

图5 各研究领域论文数量时间序列

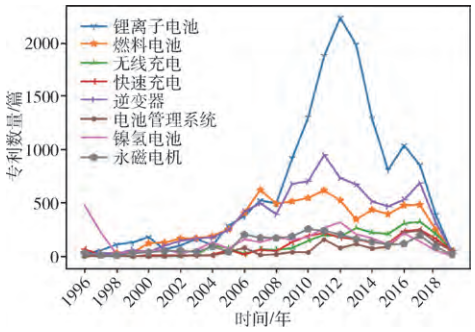

图6 各研究领域专利数量时间序列

表5 锂离子电池技术关键技术词

编号	技术词	编号	技术词	编号	技术词	编号	技术词	编号	技术词
1	the state of charge	21	lithium-ion	41	cycle life	61	retention	81	capacity
2	lithium-ion battery	22	current collector	42	performance	62	estimate	82	body
3	positive electrode	23	battery pack	43	temperature	63	storage	83	power
4	thermal stability	24	state of charge	44	application	64	algorithm	84	cell
5	reversible capacity	25	energy storage	45	electrode	65	layer	85	mixture
6	battery capacity	26	soc estimation	46	simulation	66	heat	86	rate
7	discharge capacity	27	hybrid vehicle	47	carbon	67	apparatus	87	voltage
8	maximum temperature	28	power density	48	compound	68	motor	88	system
9	lithium ion battery	29	short circuit	49	technology	69	device	89	cathode
10	battery module	30	metal oxide	50	lithium ion	70	separator	90	lithium
11	solid electrolyte	31	battery cell	51	control	71	element	91	discharge
12	secondary battery	32	electric vehicle	52	degradation	72	housing	92	estimation
13	hybrid electric vehicle	33	lithium battery	53	anode	73	collector	93	safety
14	high energy density	34	battery model	54	impedance	74	solution	94	pressure
15	electric power storage	35	rate capability	55	electrolyte	75	plate	95	surface
16	electrochemical cell	36	specific capacity	56	li-ion	76	particle	96	material
17	electrochemical performance	37	cycling stability	57	density	77	film	97	structure
18	plug-in hybrid electric vehicle	38	energy density	58	efficiency	78	case	98	charge
19	lithium secondary battery	39	heat generation	59	manufacturing	79	polymer	99	soc
20	lithium-ion secondary battery	40	high capacity	60	demand	80	assembly	100	cycle

利用识别出的关键技术词，对论文及专利摘要文本数据进行处理，构建二进制表示的关键技术词表示向量。当某项论文或专利的摘要文本中包含了所选定的关键技术词，则该技术词向量中对应的元素值为1，否则为0，最终形成论文与专利数据的技术词矩阵。

3.3.2 基于GTM技术创新机会识别 为了对各技术领域进行深入挖掘，识别出各领域未来技术发展机会，本文使用GTM算法绘制技术地图。在应用GTM之前，需要确定模型参数，本研究参考以往学者研究经验选用21* 21维的高斯函数作为基函数，矩阵中每个函数的中心都位于21* 21的网格中，基函数的宽度值为两个相邻函数最短距离的2倍，正则化参数设为0.001。根据上述参数，将获得的技术词矩阵输入到使用Python语言编译GTM算法中，绘制技术地图。

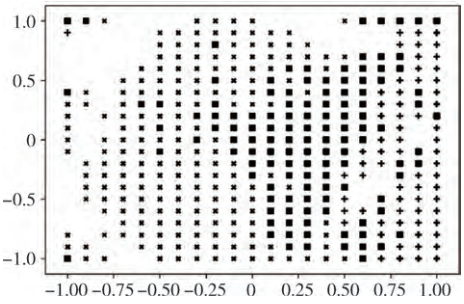

图7 锂离子电池技术地图

由于本文创新性地在专利数据基础上结合论文数据绘制技术地图，因此使用GTM算法所形成的技术地图与以往仅采用专利数据形成的技术地图有所不同。以往采用专利数据所形成的技术地图，地图中仅有专利技术点和空白点两种情况。本文运用论文和专利数据所形成的技术地图中存在体现试验发展阶段的专利技术点（“▪”）、基础研究阶段的论文研究点（“+”）、交叉点（“■”）和空白点四种情况。图7是以锂离子电池技术为例绘制出的技术地图，可以发现锂离子电池技术中部分技术正处于基础研究与试验发展并行的阶段，但仍有不少技术正处于基础研究或试验发展阶段。

Yoon等研究发现，在GTM映射的技术地图中与已有技术相邻的空白点有较大概率发展成为下一期的新兴技术[9]。基于此本文将技术地图中周围三面已有技术点的空白点进行标注（用“•”表示）作为识别出的空白新兴技术点（见图8），这些点很有可能成为未来技术的发展方向。

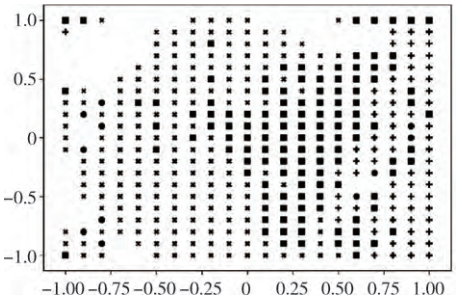

图8 锂离子电池技术地图

根据识别出的空白新兴技术点，为了获得各点所表征的技术组合，需要将空白新兴技术点分布作为一个新的空间矢量，通过GTM算法反向映射到原始数据空间（见图4）。每个空白新兴技术点对应图中表格中的一行，表格中的数值是经过GTM算法逆向映射得到的结果，代表每个空白新兴技术点所对应技术出现的概率。由于原始技术矢量是由二进制数值构成，所以反向绘图得到的结果要转换成二进制的数值，通过将这些数值与设定的阈值比较，进而获得最终结果。

目前还没有一个确定的标准来设定阈值，一般根据研究目的来确定阈值，如果阈值小，专利空白区域包括的技术领域就多，如果阈值大，空白区域的技术领域就少。本文借鉴以往论文将阈值设为0.2得到最终的技术矢量矩阵[19]，矩阵中1表示该空白区域出现对应的技术，0表示未出现。根据矩阵可以获得每个空白新兴技术点包含的技术词，从而确定该技术所涉及的领域。

3.3.3 识别结果分析 表6是锂离子电池技术通过GTM对空白技术点逆向映射识别出的技术创新机会。每一个技术创新机会中可能涉及多个关键技术词，这些关键技术词可能是未来技术创新的关键点，技术研发人员可以通过它们来判断技术未来发展方向，进行技术创新。例如空白技术点4所涉及的关键技术词有：power（功率），cell（电池），temperature（温度），voltage（电压），system（系统），surface（表面），heat（热），degradation（降解），

表6 锂离子电池技术创新机会

序号	技术组合
1	capacity; performance; power; cell; temperature; surface; carbon; density
2	performance; power; cell; temperature; surface; carbon; density
3	performance; power; cell; temperature; system; surface; heat; carbon; lithium-ion battery
4	power; cell; temperature; voltage; system; surface; heat; degradation; separator; pressure; lithium-ion battery; battery module
5	capacity; power; lithium; structure; battery pack; heat; simulation; degradation; li-ion; secondary battery; positive electrode; lithium ion; element; collector; plate; lithium ion battery; lithium-ion secondary battery
6	capacity; power; lithium; material; structure; battery pack; heat; simulation; degradation; li-ion; secondary battery; positive electrode; lithium ion; device; element; lithium secondary battery; solution; lithium ion battery; lithium-ion secondary battery
7	power; material; structure; charge; simulation; degradation; impedance; li-ion; algorithm; secondary battery; positive electrode; lithium ion; device; separator; element; lithium secondary battery; solution; current collector; film; manufacturing; short circuit
8	power; material; structure; charge; simulation; degradation; li-ion; algorithm; secondary battery; positive electrode; lithium ion; device; element; compound; control; housing; manufacturing; assembly; short circuit
9	capacity; power; lithium; material; structure; battery pack; heat; simulation; degradation; li-ion; secondary battery; positive electrode; lithium ion; element; lithium secondary battery; solution; plate; lithium ion battery; lithium-ion secondary battery
10	power; material; structure; charge; simulation; degradation; impedance; li-ion; algorithm; secondary battery; positive electrode; lithium ion; device; separator; element; lithium secondary battery; solution; film; manufacturing; short circuit

separator (分离器), pressure (压力), lithium-ion battery (锂离子电池), battery module (电池模块), 同样其余 9 个空白技术点也可以采用这样的方式进行分析。根据学者研究发现[20-21] 锂离子电池技术未来研发方向将集中于安全性、续航能力、薄膜研究、电池组电路设计等领域，运用新材料、新工艺进行技术突破。其中高能量密度和高功率电池技术将会是未来研究焦点。对本文识别出的技术空白点进行分析可以发现，其中空白技术点 7、10 涉及薄膜研究，空白技术点 1、5 与高能量密度技术相关，空白技术点 2、3 属于高功率电池技术相关研究，空白技术点 4 涉及电池组电路设计，空白技术点 8 涉及新材料相关研究。该结果印证了 GTM 在技术创新机会识别方面表现良好，可以很好地应用于技术创新机会识别领域。

基于 GTM 算法实现了快速从专利及论文数据中识别出空白技术所涉及的具体研究方向，避免了人工识别空白点的个体差异，但空白点以技术词组合方式体现，这些技术词组合是否有意义，能否成为未来的技术机会仍需要专家判断。当然，基于 GTM 的技术空白点识别为专家研发方向的选择提供了分析依据，帮助技术研发人员快速发现技术创新机会的关键点与方向，从而实现新技术研发。

4 结论与不足

在当今技术快速发展的新时期，有效实现技术创新机会识别，有利于国家和企业管理者识别出技术的未来发展方向，从而调整发展战略，为技术竞争占据有利态势。基于此，本文从论文及专利数据入手，综合运用深度学习、自然语言处理和技术地图等方法实现技术创新机会识别。该方法首先通过基于 Bert-LSTM 文本分类模型，对论文及专利研究主题进行识别，然后采用自然语言处理中的词性识别结合专家意见识别出各研究领域的关键技术词，随后运用 GTM 绘制技术地图通过逆向映射实现对技术创新机会识别。最后以新能源汽车技术为例，验证了本文提出的技术创新机会识别方法。

研究结果表明本文提出的 Bert-LSTM 文本分类模型精度优于其他分类模型，因此可以很好地应用于文献主题识别领域，为基于人工智能、大数据分析的技术识别提供了一种新的解决思路与方法。基于 GTM 算法绘制的技术地图，能够直观显示论文及专利相关信息，改善了单一维度信息进行技术创新机会识别时信息缺失的缺陷，从而使得识别结果更加全面、科学。通过 GTM 逆向映射能够有效实现技术创新机会识别，为我国其他技术领域技术创新机会识别提供了一定的借鉴意义。

当然，本文的研究还存在一定不足，这将成为后续的研究重点。首先本文只完成了技术创新机会的识别，并没有提供技术创新机会的技术解决方案；其次本文仅使用了论文与专利信息进行机会识别，因此研究结果仅体现了基础研究阶段与研究试验阶段情况，在实际分析时仍要结合其他相关维度信息。未来将会围绕这些问题继续开展研究。□

参考文献

[1] YOON B, MAGEE C L. Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction [J]. Technological Forecasting and Social Change, 2018, 132: 105-117.

[2] CHEN K, REN Z, MU S, et al. Integrating the Delphi survey into scenario planning for China's renewable energy development strategy towards 2030 [J]. Technological Forecasting and Social Change, 2020, 158: 1-12.

[3] ALEXANDER GL, DEROCHE C, POWELL K, et al. Forecasting content and stage in a nursing home information technology maturity instrument using a Delphi method [J]. Journal of Medical Systems, 2020, 44 (3): 60-68.

[4] WOO H, YEOM J, LEE C. Screening early stage ideas in technology development processes: a text mining and K-nearest neighbours approach using patent information [J]. Technology Analysis & Strategic Management, 2019, 31 (5): 532-545.

[5] REZAEIAN M, MONTAZERI H, LOONEN R C G M. Science foresight using life-cycle analysis, text mining and clustering: a case study on natural ventilation [J]. Technological Forecasting and Social Change, 2017, 118: 270-280.

[6] 许学国，桂美增. 基于深度学习的技术预测方法——以机器人技术为例 [J]. 情报杂志，2020，39 (8)：53-62.

[7] 施萧萧，张庆普. 基于共词分析的国外颠覆性创新研究现状及发展趋势 [J]. 情报学报，2017，36 (7)：748-759.

[8] PARK H, YOON J. Assessing coreness and intermediarity of technology sectors using patent co-classification analysis: the case of Korean National R&D [J]. Scientometrics, 2014, 98 (2): 853-890.

[9] SON C, SUH Y, JEON J, et al. Development of a Gtm-based patent map for identifying patent vacuums [J]. Expert Systems with Applications, 2012, 39 (3): 2489-2500.

[10] 吴菲菲，米兰，黄鲁成. 以技术标准为导向的企业研发方向识别与评估 [J]. 科学学研究，2018，36 (10)：1837-1847.

[11] MARTINO J P. A review of selected recent advances in technological forecasting [J]. Technological Forecasting and Social Change, 2003, 70 (8): 719-733.

[12] MINGGAO O, JIUYU D, HUEI P, et al. Progress review of US-China joint research on advanced technologies for Plug-in electric vehicles [J]. Science China-technological Sciences, 2018, 61 (10): 1431-1445.

(下转第 198 页)

mentation. Cham: Springer, 2016: 213-231.

[74] LEE S, SHON T. Open source intelligence base cyber threat inspection framework for critical infrastructures [C] //FTC2016-Future Technologies Conference, San Francisco, 2016: 1030-1033.

[75] TSAI M H, WANG M H, YANG W C, et al. Uncovering internal threats based on open-source intelligence [C] // International Computer Symposium. Springer, Singapore, 2019: 618-624.

[76] CASCAVILLA G, BEATO F, BURATTIN A, et al. OSSINT-open source social network intelligence an efficient and effective way to uncover "private" information in OSN profiles [J]. Online Social Networks and Media, 2018 (6): 58-68.

[77] 胡雅萍, 洪方. 社交媒体情报研究 [J]. 情报杂志, 2018, 37 (3): 15-21.

[78] KPOZEHOUEN E B, CHEN X, ZHU M, et al. Using open-source intelligence to detect early signals of COVID-19 in China, descriptive study [J]. JMIR Public Health and Survlance, 2020, 6 (3): e18939.

[79] Allied Market Research. Open source intelligence market [EB/OL]. (2020-05) [2020-09-10]. https://www.alliedmarketresearch.com/open-source-intelligence-market.

[80] ELDRIDGE C, HOBBS C, MORAN M. Fusing algorithms and analysts: open-source intelligence in the age of "Big Data" [J]. Intelligence and National Security, 2018, 33 (3): 391-406.

[81] MILLER B H. Open source intelligence (OSINT): an oxymoron? [J]. International Journal of Intelligence and Counterintelligence, 2018, 31 (4): 702-719.

[82] BAYERL P S, AKHGAR B. Surveillance and falsification implications for open source intelligence investigations [J]. Communications of the ACM, 2015, 58 (8): 62-69.

[83] TASSI P. How ISIS terrorists may have used PlayStation4 to discuss and plan attacks [EB/OL]. (2015-11-14) [2020-09-30]. https://www.forbes.com/sites/insertcoin/2015/11/14/why-the-paris-isis-terrorists-used-ps4-to-plan-attacks/#4add57877055.

作者简介： 董尹，男，1981年生，博士，副教授。研究方向：情报信号，公开源情报，物流与供应链管理。刘千里，男，1983年生，博士，讲师。研究方向：竞争情报，安全情报。胡雅萍，女，1988年生，博士，副教授，研究方向：情报决策，竞争情报。宋继伟，男，1981年生，博士，助理研究员。研究方向：竞争情报，安全情报，海疆情报。赵小康，男，1983年生，博士，副教授。研究方向：竞争情报，警务情报，安全情报。

作者贡献声明： 董尹，整体框架结构设计、论文撰写。刘千里，研究思路商讨。胡雅萍，关键词共现分析。宋继伟，文献收集、研究思路商讨。赵小康，文献收集、研究思路商讨。

录用日期： 2020－12－09

(上接第153页)

[13] 郭本海, 陆文茜, 王涵, 等. 基于关键技术链的新能源汽车产业政策分解及政策效力测度 [J]. 中国人口·资源与环境, 2019, 29 (8): 76-86.

[14] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing [J]. IEEE Computational Intelligence Magazine, 2018, 13 (3): 55-75.

[15] DEVLIN J, CHANG M, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [J]. Arxiv Preprint Arxiv: 1810.04805, 2018.

[16] TRAPPEY A J C, CHEN P P J, TRAPPEY V C, et al. A machine learning approach for solar power technology review and patent evolution analysis [J]. Applied Sciences-basel, 2019, 9 (7): 1-25.

[17] STEVENS K, KEGELMEYER P, ANDRZEJEWSKI D, et al. Exploring topic coherence over many models and many topics [C] //Empirical Methods in Natural Language Processing, 2012: 952-961.

[18] SUN Y, KIRTONIA S. Identifying Regional characteristics of transportation research with transport research international documentation (trid) data [J]. Transportation Research Part A-policy and Practice, 2020, 137: 111-130.

[19] 吴菲菲, 陈明, 黄鲁成. 基于GTM的3D生物打印专利技术空白点识别 [J]. 情报杂志, 2015, 34 (3): 58-64.

[20] 关鹏, 王曰芬, 傅柱. 基于LDA的主题语义演化分析方法研究——以锂离子电池领域为例 [J]. 数据分析与知识发现, 2019, 3 (7): 61-72.

[21] 文亚, 黄学杰, 朱春丽. 我国国立科研机构全链条式创新的模式研究——以中国科学院物理研究所的锂离子电池研究为例 [J]. 中国科学院院刊, 2019, 34 (12): 1450-1457.

作者简介： 许学国 (ORCID: 0000-0001-9898-3233)，男，1967年生，博士，教授，博士生导师。研究方向：创新与知识管理，技术创新管理。桂美增 (ORCID: 0000-0003-4810-8996，通信作者)，男，1992年生，博士生。研究方向：技术创新管理，机器学习，数据挖掘。

作者贡献声明： 许学国，论文构思，论文修改。桂美增，论文构思，论文撰写，论文修改。

录用日期： 2020－12－07