doi:10.3772/j.issn.2095-915x.2020.05.008

跨领域专利检索策略构建与实证研究

开放科学（资源服务）标识码（OSID）

高辰琛 刘琦岩 望俊成 张玄玄

中国科学技术信息研究所 北京 100038

摘要：由于跨领域专利具有跨越学科边界，融合多个领域的理论特征，且技术主题范围不明确，关键词涵盖范围大，目前传统的专利检索方法并不适用。本文在传统专利检索方法的基础上主要通过专利数据库选择、专利 IPC 分类号位置识别、关键词的确定和基于机器学习的专利文本识别与分类等四个步骤，并结合专家智慧实现对跨领域专利的检索，提出一种适用于跨领域专利的检索策略。同时，在“金融科技”领域进行实证研究与分析，证明了该策略的有效性，也为跨领域专利检索工作的开展提供借鉴。

关键词：跨领域专利；检索策略；IPC；关键词；文本分类

中图分类号：G354.2

Cross-disciplinary Patent Search Strategy Construction and Empirical Research

GAO Chenchen LIU Qiyan WANG Juncheng ZHANG Xuanxuan

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

Abstract: Because cross-disciplinary patents have the theoretical characteristics of crossing disciplinary boundaries and integrating multiple fields, and the scope of technical topics is not clear, and the scope of keywords is large, the current traditional patent search methods are not applicable. Based on the traditional patent search method, this article mainly adopts the four steps of patent database selection, patent IPC classification number position identification, keyword determination, and machine learning-based patent text recognition and classification, and combines expert wisdom to achieve cross-disciplinary patents

基金项目：中国科学技术信息研究所重点工作项目“金融大数据建设与知识服务（二期）- 金融科技知识图谱构建”（ZD2020-03）。

作者简介：高辰琛（1994-），硕士研究生，研究方向：科技战略与科技政策；刘琦岩（1964-），博士，研究员，研究方向：科技创新战略与政策研究；望俊成（1984-），博士，副研究员，研究方向：科技政策与科技管理、文本数据可视化、大数据治理，E-mail:wangjc@istic.ac.cn；张玄玄（1995-），硕士，研究方向：科技战略与科技政策。

search. Moreover, we propose a search strategy which are suitable for cross-domain patents. Meanwhile, empirical research and analysis in the field of "Fintech" proved the effectiveness of this strategy and provided reference for the development of cross-disciplinary patent search.

Keywords: Cross-disciplinary patent; search strategy; IPC; key words; text categorization

引言

在科学技术迅猛发展和经济全球化的时代背景下，专利能够相对全面完整地反映科学发展，特别是反映技术发展态势的专利信息就日渐成为重要的情报来源。专利的作用和影响在全世界日益显著，作为一种丰富的信息资源，专利数据被广泛的应用于技术创新、技术变革与技术管理研究中。随着产业模式不断创新和升级，技术的发展与创新速度达到了一个前所未有的巅峰，当代技术领域的进步极大地提高了对分散在各种信息来源中的组织知识进行管理的需求。单一学术研究或单一产业的发展已不足以应付企业所需，需要跨界整合的能力来应对外在变化，跨领域研究应运而生。跨领域研究在统筹资源、提高科技竞争力等方面具有先天优势，是企业技术创新的重要途径[1]。对于跨领域专利检索由于跨领域专利涵盖范围广，领域主题及关键词的确定存在难度。因此，通过合适的检索策略获得相对准确而全面的数据集，是亟待解决的问题，也是本文的研究重点。

1 跨领域研究含义与特征

1.1 跨领域研究定义

目前，跨领域研究的概念界定尚为统一，但学者们的观点基本一致。2005 年美国国家科学院促进跨学科研究委员认为跨领域研究是团队或个人进行的一种研究模式，它整合了来自两个或多个学科或专业知识机构的信息，数据，技术，工具，观点，概念或理论，以增进基础理解或解决问题其解决方案超出了单个学科或研究实践领域的范围，是一种整合了概念或理论，工具或技术，来自不同知识体系的信息或数据的研究模式[2]。学者陈虹[3]认为跨领域研究是将纷繁复杂的理论、知识、方法、人员、机构有效协同统一的有机整体，在不同学科之间存在一种联系，这种联系使得问题的分析遵循一定的规律，相互兼容，在研究中逐渐形成统一的价值观或看法。本文认为，跨领域研究不仅是将两个或多个领域融合在一起来创建一个产品，也是对思想和方法的整合和综合。例如，医学领域 3D 打印生物技术整合“化学工程”与“机械工具”领域专利，通过对材料的改进来提高 3D 生物打印的效率。

1.2 跨领域研究特征

跨领域研究特征主要从“多样性、均匀性、差异性和一致性”来体现[4]。“多样性”是指跨领域技术所涉及的领域种类的数量有两类或多类；“均匀性”是指跨领域所涉及的各种技术所占的分量；“差异性”是指不同技术领域

doi:10.3772/j.issn.2095-915x.2020.05.008

在理论、方法、数据、工具等方面的差别；“一致性”是指所涉及的各技术领域之间联系的紧密程度[5]。本文认为跨领域特征主要为：跨越学科边界，运用不同领域的理论方法解决某些特定问题；参与解决问题的领域各方合作产生同一成果；出现新兴产业，并产生新知识和新领域。

2 国内外专利检索方法研究状况

目前，国外专利检索中应用最广泛的技术是查询重组（Query Reformulation，QRE），目的是通过减少或扩展检索项，将输入的检索项转化为带有权重的检索项，以提高相关文档的可检索性。通常通过查询约简（Query Reduction，QR）、查询扩展（Query Expansion，QE）以及两者交叉使用的方式进行。查询约简（QR）：从全部检索项中选择代表词的子集并用作最终检索的词。基于位置的方法是最常用的方法，运用专利文档中特定部分或其中的一部分的术语，或者选择其他方法为检索项匹配更高的权重。另一种方法是基于 IPC 的方法，利用 IPC 定义中的术语作为检索项的词典或停用词列表。查询扩展（QE）：在此类别中，伪相关反馈（Pseudo Relevance Feedback，PRF）方法最为突出，提取全部检索项中有代表性的术语进行检索，将含有排名较高术语的文档假设为最佳结果，并用于扩展检索项。另外，也可以运用基于语义的查询扩展方法通过将检索项扩展为具有类似含义的术语（例如同义词）来进行检索工作。此外，运用 PageRank 算法计算专利文档的得分，估算专利文档的被访问可能性，并将此概率用作每个文档的基于引文的得分。Fujii[6] 最早在 2007 年提出一种基于文本和引文的检索方法，同时结合 PageRank 算法，对无效专利进行检索并取得很好的效果。Tannebaum[7] 等通过对检索人员的查询日志分析，从检索词的主要来源、扩展词的应用、运算符的应用等方面对检索系统进行优化。

国内专利检索人员在专利检索策略运用方面主要集中在对于 IPC 的选择、关键词的确定方面，通过查询专利检索工具，如专利公报、文摘、索引、期刊、手册、述评等内容确定目标专利的分类号或关键词。随着计算机技术发展，众多学者开始通过计算机技术对专利文档进行批量操作，运用自然语言处理、人工智能等技术提高专利检索效率。程晓静等[8]将自然语言处理技术应用在药物专利检索工作中，运用半自动化的手段将专利文本描述转化为 GSCCT 格式。孙志飞[9]通过增加发明专利相关关键词权重的方式，同时结合用户反馈与审核结果等非文献数据，对语义检索系统进行改进。颜端武等[10]研究了基于双语词典和歧义消解的中英双语专利信息检索方法，通过基于双语词典的提问式翻译实现双语专利检索，基于潜语义分析的提问式消歧策略进行歧义消解，结合文本表示模型构建检索表达式，并在专利数据库中进行匹配，得到检索结果。刘梦兰等[11]提出基于词向量的专利自动扩展查询方法，以词向量为基础构建关键词查询网络，依据稠密子图发现算法整合扩展词，并通过实证证明提出的算法能够保证扩展词集获取的灵活性和有效性，进一步提高专利检索的召回率。

目前，跨领域专利的检索方法并未形成系

统，一般而言，学者们针对某个特定领域提出相应的检索策略。例如，陈琼娣 [12] 总结了“绿色技术”检索过程中，IPC 分类号与关键词检索的劣势，提出运用 USPTO 中“环境友好型”技术专利分类索引的检索策略，并进行修正，最终得到精确检索结果；张嶷 [13] 运用基于词频分析与比较分析的专利检索策略，对电动汽车领域的专利进行检索研究与分析。

3 跨领域专利检索策略构建

跨领域专利的检索策略，是为了解决 IPC 分类号检索和关键词检索在多领域专利检索应用中存在的问题。本文根据国际专利分类标准的特点，融合 IPC 位置识别方法、关键词收集以及基于机器学习的专利文本识别与分类等三个主要流程步骤，从而构建跨领域专利检索策略，如图 1 所示。

3.1 确定专利检索数据库

正确选择检索数据库很重要，这个环节是信息收集、分析、研究的起点，由此确定决策的方向、确定研究对象。检索数据库是否符合研究课题的需求，直接影响到检索结果的可靠性和准确性；充分而正确地理解检索需求，是正确选择检索数据库的前提 [14]。面对众多专利数据库，通常从几个方面进行考量。(1) 数据库中专利数据覆盖的全面性是专利文献被检索到的前提，数据是所有分析工作的基础。因此，专利数据库中收录的专利数据量是选择的重要依据。(2) 数据库中提供专利信息的全文浏览、下载的方法以及格式。(3) 如果需要分析大量专利数据时，通常要将专利数据导入数据分析工具或者自己建立一个与技术相关的数据库，根据需求选择合适的下载文件格式。

3.2 跨领域专利IPC位置识别

对于跨越不同领域的专利来说，关键词的确定和选择是非常困难的。因此，专利检索人员意识到要想提高检索的准确度并节约时间，需要有一个统一完整的分类系统帮助限定词义的范围。专利分类号对每种技术作了明确的分类，因而大部分的专利检索都是通过专利分类号来完成。但是，由于跨领域利的特殊性，依靠传统的专利分类号进行检索依然很难达到应有的效果。确定专利分类号通常有两种途径，(1) 在德温特数据库中检索各领域关键词的德温特手工代码，从而得到相应的 IPC 位置分布。(2) 将产生的新兴领域的关键词作为检索词，在专利数据库中进行粗检，得到粗略范围 IPC 分布。

3.3 跨领域专利检索关键词收集

确定检索需求后应结合检索的技术领域特点，在没有准确的专利分类号的情况下，则可以通过关键词检索这一重要手段。对于关键词的选择方法，(1) 根据新兴领域的关键词描述在文献数据库中进行相关文献检索；(2) 从新兴领域的不同组成领域进行领域关键词的获取；(3) 专业咨询机构发布的行业报告；(4) 从新兴领域自身出发收集领域关键词。对于关键词形式的表达，应当考虑一词多义、词语不同形式的表达。在使用英文单词做检索词进行检索时，要考虑单词的词性、名词的单复数形式、单词的拼写方式、错拼单词的情况。同时，也需要正

doi:10.3772/j.issn.2095-915x.2020.05.008

确使用各种截词符从而保证关键词的完整性，关键词的含义上是指应当考虑关键词的同义词、近义词、上位词、下位词等形式。

3.4 专利文本分类与识别

文本分类是基于自然语言处理技术框架下对单词、句子、段落等文本类别标签作出判断的一种技术，属于监督学习方法[15]。文本分类是将文本作为分类对象，需要预先定义文本类别，最后根据设置好的分类判断需要分类的文本内容[16]。具体涵盖五个方面内容。(1) 文本预处理：预处理是为了将原始数据进行规范格式的处理，完成噪声的去除，并进行分词处理，同时去掉连词、介词、助词等对文本分类工作无意义的停用词。(2) 文本转化为向量模型：在词袋模型[17]中，运用 TF-IDF（Term Frequency-Inverse Document Frequency，词频 - 逆文本频率）根据词出现的频率设置计算权重。(3) 特征选择：主要是先对特征进行降维，通过空间降维提取特征，可以保证所提取的特征符合文本语义描述，降低特征矩阵的维度，可以避免计算机超负荷计算工作，通过对高维特征矩阵进行空间降维，以避免原始数据过于庞大而造成计算机计算能力超负荷运行。(4) 构建分类训练样本：首先，通过专家阅读原始数据的方式对文本类别设置标签，然后构建分类训练样本，通常运用的分类方法包括 SVM、K 近邻法、决策树、随机森林等方法。(5) 分类性能评价：文本分类性能评价根据数据集、评价指标、评价测试策略进行评价。

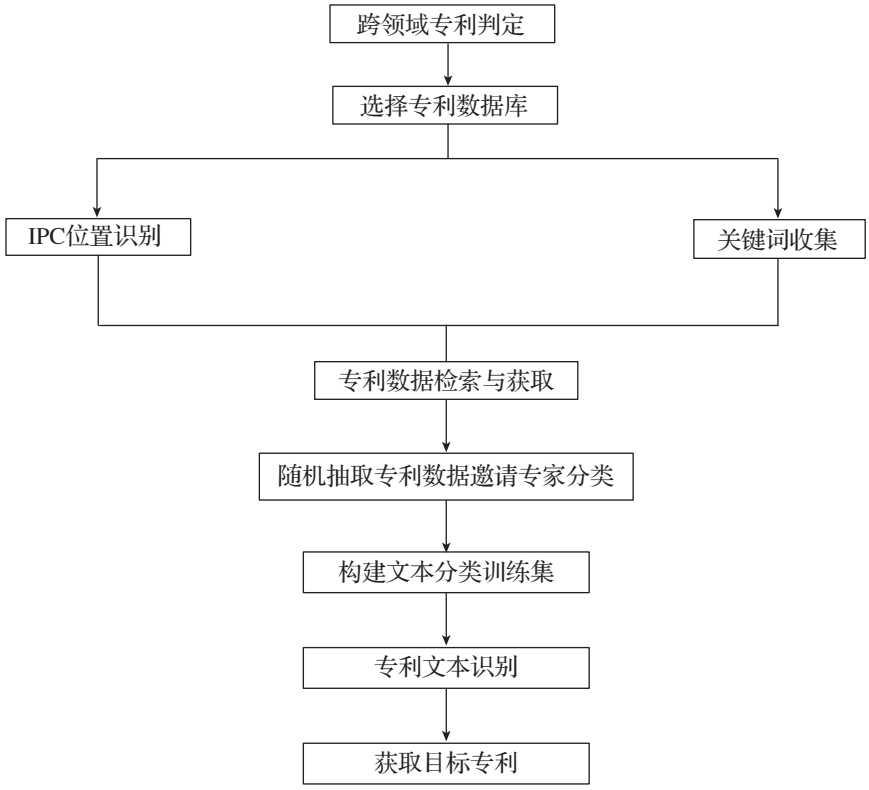

图 1 跨领域专利检索策略

4 基于“金融科技”领域的跨领域专利检索策略实证研究

4.1 实证研究领域

本文选择金融科技领域进行实证研究，金融科技最早由国际金融稳定理事会提出，通过技术带动金融创新，对金融市场以及金融服务业务供给产生重大影响的新兴业务模式、新技术应用、新产品服务。因此，金融科技被认为是新兴领域，对社会发展有积极作用，具有跨领域的典型特征，是金融和科技的结合，也是现代技术在金融行业的运用，即通过利用各类科技手段如大数据、区块链、云计算、人工智能等新兴前沿技术创新传统金融行业所提供的产品和服务，也是科技领域与金融领域的一次跨领域融合，为传统金融行业带来冲击的同时，也为金融业的发展带来新的机遇。在科技和金融二元融合与相互渗透的背景下，新业态、新机构、新产品的快速出现正加速推进金融市场的重构。

4.2 确定专利数据库

本文选择德温特数据库（Derwent Innovations Index）为检索数据库，专利数据覆盖范围广，包括将德温特世界专利索引 (Derwent World Patents Index) 和专利引文索引 (Patents Citation Index)。用户通过该数据库不仅可以检索专利信息，还可以检索到专利的引用情况，收录了来自世界各地超过 52 家专利授予机构提供的增值专利信息，涵盖 3,050 多万项发明并记录于 6500 万篇的专利文献中，每周更新并回溯至 1963 年；提供浏览和下载的内容全面，每条记录除了包含相关的同族专利信息，还包括由各个行业的技术专家进行重新编写的专利信息，如描述性的标题和摘要、新颖性、优点等。

4.3 确定IPC分类号

对于 IPC 的确定，分两个步骤进行：首先因为金融科技（Fintech）是金融（Finance）与科技（Techonoly）的合成词，所以分别从金融相关词汇角度和科技相关角度，通过德温特手工代码检索“FINANCIAL”、“AI”、“CLOUD COMPUTING”等与金融科技相关的词汇，如表 1、2 所示，最终得到国际专利分类号分布情况如图 2 所示。

表 1 德温特手工代码

手工代码	解释
T01-J05A1	FINANCIAL
T01-N01A	FINANCIAL/BUSINESS
T01-N01A1	FINANCIAL TECHNOLOGY SYSTEMS
W02-F10J	FOR ACCESS TO FINANCIAL NETWORK
W03-A16C5J	FOR ACCESS TO FINANCIAL NETWORK
T01-N01D3A	CLOUD COMPUTING SERVICES
T01-J16	ARTIFICIAL INTELLIGENCE (AI)

本文还通过中国知网专利检索库、万方数据检索库，以“金融科技”“数字货币”“比特币”“众筹”等金融科技领域的关键词术语进行专利检索，并阅读检索结果验证得出上述分类结果比较准确。为了在初步检索时，获得更全面涵盖金融科技专利，并确定首先选择“G06Q*”、“H04L*”两组。

doi:10.3772/j.issn.2095-915x.2020.05.008

表 2 主要国际专利分类号分布

国际专利分类代码	记录	百分比%	国际专利分类代码	记录	百分比%
G06Q-040/00	35291	11.674	H04L-029/08	24239	8.018
G06Q-030/02	34818	11.517	G06F-017/30	21466	7.101
G06Q-030/00	29798	9.857	H04L-029/06	18309	6.056
G06F-017/60	24730	8.18	G06Q-020/40	16199	5.358
G06Q-030/06	24576	8.129	G06Q-040/02	14978	4.955

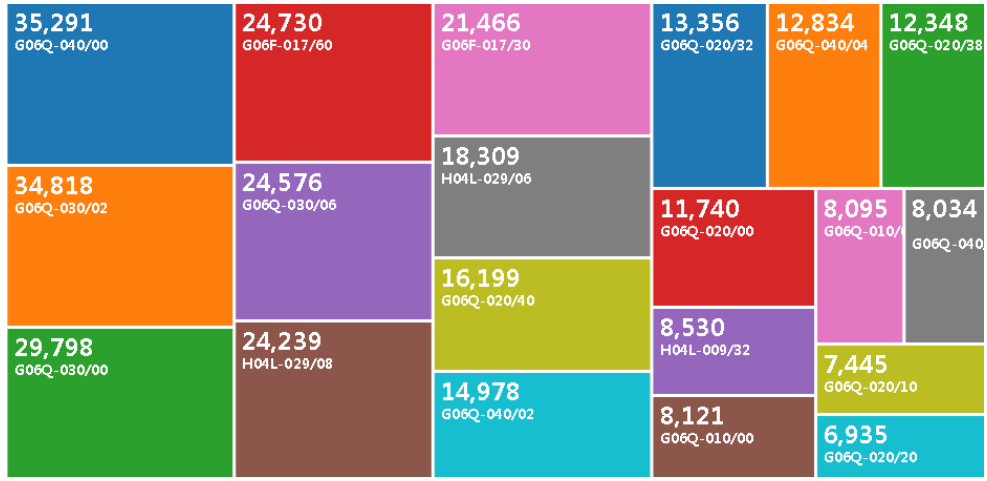

图 2 主要国际专利分类号分布

4.4 确定检索关键词

作为跨领域的典型代表，金融科技涵盖范围非常广，包括支付、银行理财、保险、证券、消费金融、供应链金融、基金等业务[19]，为了能够收集较为全面的检索词，主要通过三种途径进行关键词的收集：(1) 从国内外知名媒体、咨询机构等的报道、咨询报告中收集到金融科技相关的词汇，涵盖金融科技创新企业及众多常用金融科技常用术语（如 IOT、bitcoin、digit cash 等）；(2) CMC Markets(英国一家知名的衍生品经纪商) 的网站上给出的金融专业术语合集（如 Account、Electronic Currency、Margin 等）；(3) 金融科技底层技术（如 Internet、Internet of things、Big Data、AI 等）。在此基础上，加入部分最近被确认为金融科技专业术语的词汇（如 Digital Wallet、Cryptocurrency、Crowdfunding 等），包括金融科技创新企业名单及报告中的常用术语，以及最终形成 300 个与金融科技相关的关键词集合。

表 3 词频分布

排名	关键词	频次	排名	关键词	频次
1	Bank	576	6	Internet	240
2	Inclusive Finance	302	7	Internet Finance	168
3	Big data	292	8	Payment	110
4	AI	276	9	Currency	72
5	Block chain	250	10	Insurance	54

4.5 专利文本获取与识别

(1) 专利数据的获取

在德温特数据库中进行专利检索操作，将收集整理的关键词通过编辑检索式方式进行专利检索，运用关键词检索策略进行检索词的修饰，防止因同义词出现的检索结果冗余的情况，排除了在其专利文件标题、摘要或权利要求中不包含任何列表中金融词汇的专利，以达到更好的检索结果，并确定国际专利分类号“G06Q*”、“H04L*”，时间限定在 2016-2019 年，共检索得到 19447 条记录，数据内容包括每项专利的标题、所有权、摘要和权利要求等全部详细数据，得到最终的专利分析数据集。

(2) 专利文本分类模型构建

将空格、标点为标志对原始文本进行切分，使句子变为有序的词项；去除在停用词词表中的词项，如 a、an、the、that 等；使用 Stanford 词性还原工具进一步整合相关词汇，减少英文语境下时态、语气词对词项的影响；运用词袋模型将句子转换为向量表示的模型，该模型仅关注每个单词在文本中出现的次数。传统词袋模型以词频对句子进行向量化展示，在面对诸如提取关键词等问题时，面临高频词并不是关键词的矛盾问题（如“the”在句子中可能出现次数最多，却不是能够区分句子的明显特征词）。而运用 TF 进行词袋模型的构建，则可避免单一的词频产生的影响，即一个单词在文本中出现次数越多，在分类时被认为价值越小，增强了单词将所在文本与其他文本区分开的能力，适用于解决跨领域专利的分类问题（如“NOVELTY”在德温特专利数据库中是摘要项的一个分项，出现在所有的专利数据，其权重应该较小）[20]，其计算公式如下所示。

逆文本频率 Inverse Document Frequence:

$$\mathrm{IDF}(w)=\log\left(\frac{\mathrm{N}}{\mathrm{n}(w)}\right)$$

TD-IDF 计算公式：

$$\text{TF-IDF}(w)=\text{TF}(w)*\text{IDF}(w)$$

在本文中，将文本特征项用 w 表示，文档集中的文本总数用 N 表示。TF 表示如果一个词在文本中出现的次数较多，则可以用来描述该文本的主要信息，IDF 表示如果一个词在少量文本中大量出现，则更有利于描述文本信息。

有监督机器学习能够进行学习的基本保障之一是设置标签的训练样本与未设置标签的专利文本数据集应为独立同分布的，即进行标签设置的训练样本是随机抽取且相互独立的。本文邀请金融领域专家对随机抽取的专利文本进行标签设置，保证了训练样本的准确性。

(3) 专利文本识别

跨领域专利文本识别的最后一步就是使用机器学习方法结合已设置标签的训练样本找出最优的机器学习模型，并应用于全部专利数据集实现对相关专利的识别与分类。本文运用在文本多分类问题中常用的机器学习算法，如支持向量机（Support Vector Machine, SVM）、决策树 (Decision Tree）、随机森林 (Random forest) 等，通过最结果的比较来选择最优的方法，得到较好的多分类效果。过程中运用多种方法对金融科技专利数据的特征进行学习，同时对专利文本向量构建算法，以及不同算法的参数等进行调整设置，达到的准确性如下表 4 和表 5 所示。选取标注样本的 80% 作为训练集，

doi:10.3772/j.issn.2095-915x.2020.05.008

20% 作为测试集。训练结果，如表 4 所示，各方法表现较好，SVM、随机森林方法准确率较高，准确率超过 99%，而决策树准确率相对较差，平均准确率仅 75% 左右。

表 4　训练集机器学习准确率

	LibSVM	Random Forest	Decision Tree
Overrall Accuracy	98.89%	99.37%	75.73%

测试结果如表 5 所示，各机器学习算法表现维持在 70% 左右，随机森林方法准确率达到 75.20%，决策树准确率最低为 71.83%，结合训练集和测试集的结果，因此选择随机森林的机器学习方法进行全文本分类。

表 5　验证集机器学习准确率

	LibSVM	Random Forest	Decision Tree
Overrall Accuracy	72.23%	75.20%	71.83%

随机森林模型主要思想来源于 Boosting、Ada-Boost 和 Bagging 等算法，是多个决策树即通过多个弱学习器的结合，达到强学习器的效果 [21]，核心思想是以多棵决策树为基础的集成分类器，通过采取多个不同的训练样本子集来加大分类模型之间的相异性，从而能够提高该模型的泛化能力以及预测能力，具有精度高、参数少、性能稳定的特点 [22]。由于其良好的性能，进而被广泛应用到生物信息、医学研究、语言建模、文本分类等实际领域，并在这些领域中利用随机森林方法都取得了不错结果。此外，通过实验数据可知，随机森林算法在专利文本分类识别工作中的准确率最高，故本文选择随机森林算法进行跨领域专利的分类与识别。

5　总结

本文根据跨领域研究的特征，总结了跨领域专利的特征，例如跨领域专利的边界不清晰、技术主题和关键词范围广。从专利数据库的选择、IPC 位置的识别、关键词的确定、专利文本的分类与识别四个方面，提出适用于跨领域专利的检索策略。同时，本文通过邀请专家对专利文本进行标签设置，以达到更好的专利文本识别效果。此外，本文以“金融科技”领域为实证进行研究分析，从而验证了本文提出的跨领域专利检索策略的可行性，并为跨领域专利检索工作在 IPC 位置识别以及关键词关键词收集等方面提供有意义的参考与借鉴。

但本文也存在一定的局限性，(1) 在 IPC 分类号位置识别时，选取范围较大，增大后期专利文本处理的工作量；(2) 关键词收集与处理采取人工编辑方式，增加检索人员的工作强度。在未来的研究工作中，本文将尝试借助计算机完成 IPC 分类号识别以及关键词的批量处理。

参考文献

[1] 王晓宇 , 苗红 , 王芳 . 技术知识的跨领域应用及潜在技术方案的识别 [J]. 图书情报工作 , 2016, 60(23):87-96.

[2] Futures Initiative. Facilitating Interdisciplinary Research[J]. Facilitating Interdiplinary Research, 2005(48).

[3] 陈虹 , 黄斌 , 杨桂栓 . 跨领域研究的识别、管理与评价研究 [J]. 科技与经济 , 2016, 29(4):11-15.

[4] Rafols I, Meyer M. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience[J]. Scientometrics, 2010, 82(2):263-287.

[5] 黄鲁成 , 郭彦丽 , 吴菲菲 , 等 . 新兴技术跨领域评

价方法研究——以 3D 打印技术为例 [J]. 中国科技论坛 , 2015(5):42-47.

[6] Fujii A. Enhancing patent retrieval by citation analysis[C]. SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007. ACM, 2007.

[7] Tannebaum W, Mahdabi P, Rauber A. Effect of Log-Based Query Term Expansion on Retrieval Effectiveness in Patent Searching[A]. Lecture Notes in Computer Science[M]. Springer, Cham, 2015:300-305.

[8] 程晓静 , 程文堂 . 自然语言处理技术在药物专利检索中的应用研究 [J]. 计算机与应用化学 , 2005, 22(2):143-147.

[9] 孙志飞 . 语义检索在专利文献检索中的应用及改进 [J]. 信息技术 , 2014(5):135-137.

[10] 颜端武 , 任婷 , 陶志恒 . 基于双语词典和歧义消解的中英双语专利信息检索研究 [J]. 情报理论与实践 , 2018, 41(2):138-142+154.

[11] 刘梦兰 , 刘斌 , 彭智勇 . 基于词向量的专利自动扩展查询研究 [J]. 计算机工程与科学 , 2017, 39(12):2297-2305.

[12] 陈琼娣 , 余翔 . USPTO“绿色技术”专利检索策略研究 [J]. 现代情报 , 2012, 32(8):27-31+36.

[13] 张嶷 , 汪雪锋 , 郭颖 , 等 . 中国专利数据检索策略研究 [J]. 科学学研究 , 2011, 29(6):833-839.

[14] 华薇娜 . 网络信息检索策略的设计与实施的探讨——基于网络数据库信息检索各环节的实例分析 [J]. 图书馆论坛 , 2008, 28(6):111-114+178.

[15] 王丁 . 基于机器学习的文本分类技术分析与研究 [J]. 科技创新导报 , 2020, 17(8):90+92.

[16] Basu T, Murthy C A. Effective Text Classification by a Supervised Feature Selection Approach[C]. IEEE International Conference on Data Mining Workshops. Brussels, Belgium. 2013:918-925.

[17] Lebanon G, Mao Y, Dillon J. The locally weighted bag of words framework for document representation[J]. Journal of Machine Learning Research, 2007, 8(8): 2405-2441.

[18] 谷政 , 石岿然 . 金融科技助力防控金融风险研究 [J]. 审计与经济研究 , 2020, 35(1):16-17+11.

[19] 艾瑞咨询 . 2019 年中国金融科技行业研究报告 [R]. 上海 : 上海艾瑞市场咨询股份有限公司 , 2019.

[20] 徐璐 , 卢小宾 , 杨冠灿 . 金融科技专利识别与分类方法构建及应用 [J]. 图书情报工作 , 2020, 64(11):87-95.

[21] 卢婧 , 冯仲科 . 运用随机森林模型对北京市林分蓄积生长量的预测 [J]. 东北林业大学学报 , 2020, 48(5):7-11.

[22] 王猛 , 张新长 , 王家耀 , 孙颖 , 箭鸽 , 潘翠红 . 结合随机森林面向对象的森林资源分类 [J]. 测绘学报 , 2020, 49(2):235-244.