· 情报分析与技术创新 ·

基于机器学习分类算法的高质量专利成果筛选研究

周一夫[1] 谭春辉[1] 江 婷[2*] 李玥澎[3] 毕慧婷[1] 汪红信[1]
(1. 华中师范大学信息管理学院，湖北 武汉 430079；2. 湖北技术交易所，湖北 武汉 430071；
3. 武汉大学信息管理学院，湖北 武汉 430072)

摘 要：[目的/意义] 基于客观数据形成一套自动筛选方法，对专利成果质量进行快速识别，为推动专利成果转化工作提供决策支持。[方法/过程] 首先，以专利成果的发明人数量、IPC 号数量等形式特征结合语义向量匹配度特征、专利成果质量标注结果，构建高质量专利成果筛选指标体系；其次，以“先进制造与自动化”领域为例，在专利之星平台检索该领域的发明专利作为专利文本数据来源，并以湖北省需求为例，将其相关的产业发展规划(宏观)和市场技术需求(微观)作为需求文本数据来源；随后，采用分词、去停、文本向量化等步骤对专利文本和需求文本进行处理，并整理形成训练集和测试集；最后，调用 8 种机器学习分类算法模型进行训练与评估，并对训练效果最优的算法展开应用测试，以验证筛选方法的可行性。[结果/结论] 结果显示，随机森林算法模型在选取的 8 类算法模型中整体表现最优，被用为高质量专利成果筛选方法中的内核分类算法。此外，本文提出的筛选方法对专利成果质量识别具备较强的可行性，能够结合不同省(市)的特定专利需求，快速地进行大批量专利成果的筛选，在一定程度上可有效降低人力、物力和财力成本的消耗。

关键词：专利成果筛选；高质量专利成果；机器学习；Doc2vec
DOI:10.3969/j.issn.1008-0821.2024.02.007
〔中图分类号〕G255.53 〔文献标识码〕A 〔文章编号〕1008-0821 (2024) 02-0081-11

Research on the Screening Method of High-quality Patent Results Based on Machine Learning Classification Algorithms

Zhou Yifu[1] Tan Chunhui[1] Jiang Ting[2*] Li Yuepeng[3] Bi Huiting[1] Wang Hongxin[1]
(1. School of Information Management, Central China Normal University, Wuhan 430079, China;
2. HuBei Technology Exchange, Wuhan 430071, China;
3. School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [Purpose/Significance] Based on objective data, the study forms a set of automatic screening methods to quickly identify the quality of patent results and provides decision support to promote the transformation of patent results. [Methodology/Process] Firstly, the study constructed a high-quality patent results screening index system with combining the formal features such as the number of inventors and the number of IPC numbers of patent results with the semantic vector matching degree features and the quality annotation results of patent results; Secondly, taking the field of “advanced manufacturing and automation” as an example, the study retrieved the invention patents in this field on the Patent Star platform as the source of patent text data, and took the demand of Hubei Province as an example, and took its relevant industrial development plan (macro) and market technology demand (micro) as the source of demand text data.; then, processed the patented text and the demanded text by using word separation, de-stopping, text vectorization and other steps, and or-

收稿日期：2023-05-18
基金项目：2022 年度华中师范大学基本科研业务费(人文社科类)交叉科学研究项目“基于大数据的科教智能评价与智慧服务模式研究”(项目编号：CCNU22JC031)。
作者简介：**周一夫**（1995-），男，博士研究生，研究方向：信息计量与科学评价、网络用户行为。**谭春辉**（1975-），男，教授，博士，博士生导师，研究方向：信息计量与科学评价、网络用户行为等。**李玥澎**（1998-），男，博士研究生，研究方向：信息计量与科学评价。**毕慧婷**（1998-），女，硕士研究生，研究方向：信息计量与科学评价。**汪红信**（2000-），女，硕士研究生，研究方向：信息计量与科学评价。
通讯作者：**江婷**（1985-），女，硕士，平台建设部高级专员，研究方向：科技管理，科技成果转化。

ganized to form a training set and a test set; finally, called eight machine learning classification algorithm model for training and evaluation, and tested the algorithm with the best training effect for application to verify the feasibility of the screening method. [Results/Conclusion] The results show that the random forest algorithm model has the best overall performance among the selected eight types of algorithm models, and is used as the kernel classification algorithm in the screening method of high-quality patent results. In addition, the screening method proposed in this paper has a strong feasibility for the quality identification of patent results and can combine the specific patent needs of different provinces (municipalities) to quickly screen large quantities of patent results, which face to a certain extent, effectively reduce the consumption of human, material and financial resources costs.

Key words: screening of patent results; high-quality patent results; machine learning; Doc2vec

科技创新是提高社会生产力和综合国力的重要战略支撑，促进科技成果转化、降低成果闲置率已成为当前全球多个国家(地区)科技发展的新要求。伴随综合国力的不断增强和科学技术的飞速发展，我国科技成果数量取得了新的突破，专利申请量和授权量跻身世界第一，但科技成果转化率和转化成效并未取得显著提升[1]。科技成果转化效率不高目前已成为制约我国跻身世界创新强国的一大障碍。2015 年 8 月，第十二届全国人民代表大会常务委员会第十六次会议修订了《中华人民共和国促进科技成果转化法》[2]，旨在为新形势下的科技成果转化活动提供保障和规范，凸显科技成果日益增长的经济价值和社会价值。进入“十四五”时期，我国科技成果转化整体情况虽然有所改善，但仍然存在着转化率不高、转化路径不清晰、供需匹配不明确等问题。为加快建设科技强国、实现高水平科技自立自强的目标规划，以习近平同志为核心的党中央高度重视科技创新工作，把促进科技成果转化摆在十分重要的位置进行谋划部署。2021 年 5 月，习近平总书记在中央全面深化改革委员会第十九次会议上明确提出“加快推动科技成果转化应用，加快建设高水平技术交易市场，加大金融投资对科技成果转化和产业化的支持”的要求[3]，进一步在国家层面明确推进科技成果转化相关工作的必要性。

促进科技成果转化是推动经济社会发展和适应国际竞争形势的迫切需要，同时也是科技成果应用于生产实践的重要支撑。专利成果作为科技成果转化体系中的重要组成部分，如何从海量专利中筛选出适应市场需求的高质量成果，然后有针对性地促进其转化？解决这一问题不仅有利于提升专利成果转化成效，同时对我国经济和科技长期高质量发展具有重要战略意义。

1 文献回顾

1.1 专利成果转化相关研究

专利成果转化是一项由政府引导规范、多方主体参与的活动，其具体内涵是指新技术、新发明经过试验、开发、应用和推广，实现商品化和产业化，最终实现经济价值的过程[4]。目前已有不同领域的学者从多角度对其展开深层次剖析，主要聚焦于 3 个方面：①专利成果转化政策研究。自改革开放以来，我国各级政府为推动专利成果转化工作，先后颁布了系列政策并逐步形成一个较为健全的政策体系[5]，学者们也从政策组态效应[6]、政策文本量化[7]、政策优化策略[8]、政策实施效能[9]等不同视角对专利转化活动进行研究，旨在为后续政策出台实施、修订完善、执行落实等环节提供决策支持，达到加速科技成果转化的目的；国外对专利成果转化政策关注较少，相关研究从政策对专利申请量的影响[10]、专利保护效果与质量提升的影响[11]等方向进行了探讨；②专利成果转化现状与对策。在专利转化过程当中，由于涉及多领域和面临多重复杂环境，难免遇到各种困境与阻力，因此，不少学者以不同领域的转化主体为研究对象，对高校[12]、国防[13]、国企[14]等主体的转化现状进行梳理，明晰转化过程中遇到的困境，并从体制机制改革[15]、交易成本模型探究[16]等方面有针对性地提出对策和建议；③专利成果转化绩效研究。转化绩效是衡量成果从理论应用到实际的一项重要指标，已有学者采用层次分析[17-18]、理论归纳[19]等方法构建评价指标体系，对成果转化绩效进行了评价研究；也有科研人员通过数据包络分析模型[20]、社会网络[21]、面板数据模型[22]等视角，探究相关因素对转化绩效的影响程度，以期有效推进创新驱动发展和提升专利成果转化绩效。

1.2 专利成果筛选相关研究

随着专利成果数量的逐年激增，对高质量专利成果的筛选显得尤为关键，传统人工筛选已无法满足海量专利的不断累积，因此吸引了国内外不少学者对专利成果筛选展开相关探究。通过文献梳理后发现，关于专利成果筛选方法主要有计量学识别[23-24]、引证关系识别[25-26]、主题模型识别[27-28]、机器学习算法识别等。鉴于当前专利成果数量规模，同时相较于其他专利筛选方法，机器学习算法识别具备高效迅速、精度可增长性、结果一一映射性等优点，已经成为目前主流专利成果筛选方法，如 Krestel R 等[29]通过总结 40 篇使用深度学习框架对专利分类的文献，发现相关研究仍处于起步阶段，同时预计专利分析的方法将由经典机器学习逐步朝深度学习的方向发展；Liu B C 等[30]提出了由自组织映射(SOM)、核主成分分析(KPCA)和支持向量机(SVM)组成的机器学习组合模型，并将其应用于生物医药产业专利质量预测；Hu Y F 等[31]在现有三维专利价值评价指标的基础上增加了跨境维度指标，采用随机森林、决策树等机器学习算法对可转让专利进行识别，研究发现，机器学习方法能够较好地支持海量数据中可转让专利的识别；张彪等[32]基于技术的新颖性、独特性和重要性 3 个维度来构建相关指标，采用 K 近邻、逻辑回归等 7 种机器学习算法对高价值专利进行筛选；吴洁等[33]基于专利形式特征并结合专利文本特征生成的专利—核心词汇网络，通过搭建图卷积网络对高质量专利进行自动识别；付振康等[34]从专利寿命视角切入，选取影响专利寿命的相关因素作为识别指标，选用 5 种深度学习模型对专利寿命进行预测，然后通过设置阈值的方式识别核心专利。

1.3 简要述评

能否有效筛选是专利成果成功转化的关键环节之一，从现有研究来看，仍有可继续深化之处：①早期由于信息数据等资源相对匮乏、技术手段不够完善等原因，用客观数据进行专利筛选的方法不够成熟，可能存在主观性较强、组织过程复杂、成本花费较大等弊端；②目前专利成果识别、探测方法主要研究来源多数依靠单点预测和自身形式特征的分析，未能较好地结合地域发展规划与市场需求，难以保证专利成果识别、探测技术可成功应用到转化过程中，可能由于需求适应性不足导致专利转化失败。

为了弥补上述不足，本文在已有研究的基础上，进一步将专利成果形式特征与市场需求相结合，综合运用文本挖掘、专利计量、机器学习等方法，对专利文本、需求文本等材料进行处理与分析，以期形成一套基于客观数据的自动筛选方法，在一定程度上克服主观性较强、人财物力花费较大的弊端，为筛选高质量专利成果、促进转化提供一种可行思路。

2 研究设计

2.1 研究框架

本文总体思路如下：第一，选定特定领域检索专利成果并整理其形式特征，按照一定规则对其质量进行人工标注，同时提取专利摘要形成摘要文本；第二，检索并摘取相应领域产业发展规划(宏观)和市场技术需求(微观)形成需求文本；第三，对摘要文本和需求文本进行 Jieba 分词和去停用词的处理，得到实验语料集；第四，运用 Doc2vec 模型将处理后的专利摘要文本和需求文本进行向量化表示，并计算专利摘要语义向量与需求语义合向量之间的余弦相似度，以得到“语义向量匹配度”特征；第五，综合整理专利成果的形式特征、“语义向量匹配度”特征和质量类别标签，编写 Python 程序调用机器学习算法模型进行训练与评估，选取性能最优的分类算法模型作为高质量专利成果筛选方法中的内核分类算法；第六，对筛选方法进行应用测试，以验证筛选方法的可行性。基于上述研究思路，本文制定的研究框架如图 1 所示。

2.2 关键过程

2.2.1 专利成果形式特征选取

专利成果的发明人、IPC 分类号、同族专利、引用文献、实质审查时间等均为专利成果的形式特征，现有研究已经证实这些形式特征能够在一定程度上反映出专利成果的价值[35]。根据形式特征指代价值的不同，本文将其划分为技术价值特征和法律价值特征两类：

1）专利成果技术价值特征指标

专利成果的技术价值特征主要反映其所承载的技术内容的先进性、应用前景等，作为专利文献所承载的核心内容，技术价值应当纳入专利成果评价

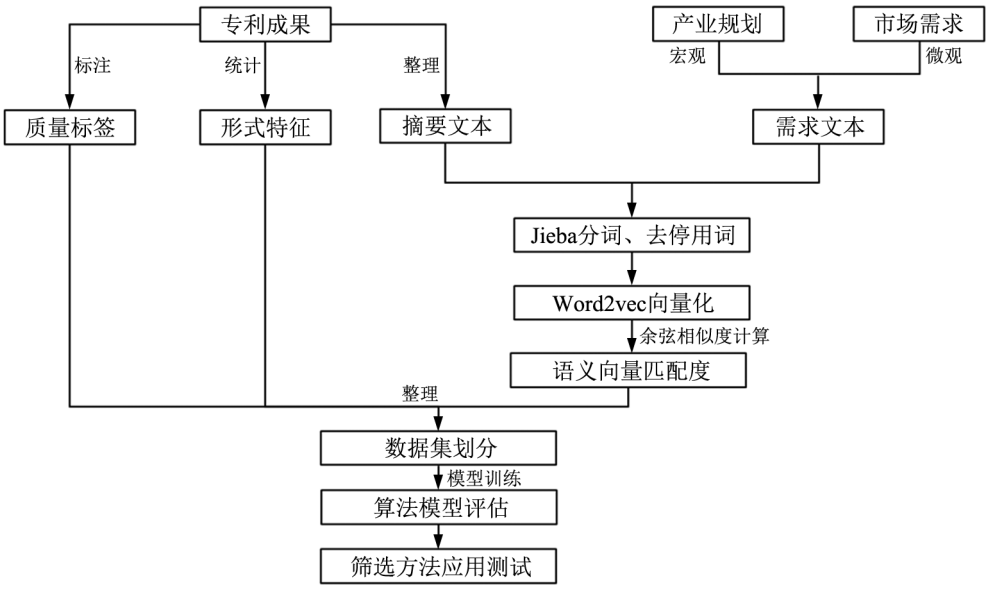

图 1　整体研究框架图

Fig. 1　Overall Research Framework Diagram

指标体系中[36]。借此，本文选取的专利技术价值特征指标包括发明人数量、技术分类(IPC 号)数量、知识产权局引证数量。其中发明人数量反映专利成果研究团队的规模；技术分类(IPC 号)数量反映专利成果涵盖技术领域的规模；知识产权局引证次数反映专利成果融合其他专利成果的规模。

2）专利成果法律价值特征指标

专利权是一种受国家法律保护的知识产权，专利的法律保护稳定性、侵权可判定性等均会影响该专利的转化质量与转化效率。结合本研究总体目标，借鉴冉从敬等[37]、许鑫等[38]的研究成果，选用说明书页数、实质审查时间两项指标分别反映专利成果的细节描述程度和保护强度。

2.2.2　专利成果语义特征与需求语义特征匹配度的计算

为全面评估专利成果价值，本文将专利成果内容文本向量化，通过成果语义向量与需求语义向量进行相似性匹配评估专利成果内容质量，帮助构建更加适应市场发展需求的高质量专利成果筛选方法，处理步骤如下：

1）语义特征提取与向量表示

Doc2vec 最早于 2014 年由谷歌公司的 Quoc Le 和 Tomas Mikolov 提出，是一种非监督式深度学习方法，其主要思想是将句子或段落转化为空间向量。Doc2vec 是 Word2vec 的延伸与拓展，其中 Doc2vec 在 Word2vec 的基础上增加了段落向量，并分别从 Word2vec 中 CBOW 和 Skip-gram 架构的基础上衍生出 PV-DM 和 PV-DBOW 两种训练架构(其训练方式如图 2 所示)。在 PV-DM 架构中，训练语料中每个段落都有唯一的 id(即 Paragaph id)，在训练过程中 Paragaph id 与其他单词(W)一样，首先被映射成相同维度的向量，但是被存储在不同的向量空间当中；在之后一个段落的若干次训练过程中，Paragaph id 保持不变，词向量与段落向量进行累加或连接来预测句子中的下一个词语(也相当于每次在预测单词的概率时，都利用了整个段落或句子的语义)，其处理方式类似于 Word2vec 中的 CBOW 架构。与 PV-DM 架构利用上下文与段落预测词语不同的是，在 PV-DBOW 架构中，首先直接将段落向量作为输入单元(但忽略其上下文之间的关系)；然后在每次迭代的过程中从文本中采样得到一个窗口，再从这个窗口中随机采样一个单词作为预测任务进行预测，其处理方式与 Word2vec 中的 Skip-gram 架构较为相似。

向量具备空间和大小双重属性，向量的加法可用平行四边形法则来进行描述，如图 3 所示，其中 F1、F2 表示两个不同的共点向量，它们邻边的夹角线 F 合表示合向量的大小和方向。在本文研究过程中，由于涉及语义向量匹配度的计算，因此需要对需求语义向量进行合向量的计算，用以表征整体需求特征。具体处理步骤如下：首先，采用 Doc2vec 模型将处理后的需求文本逐条转化为向量

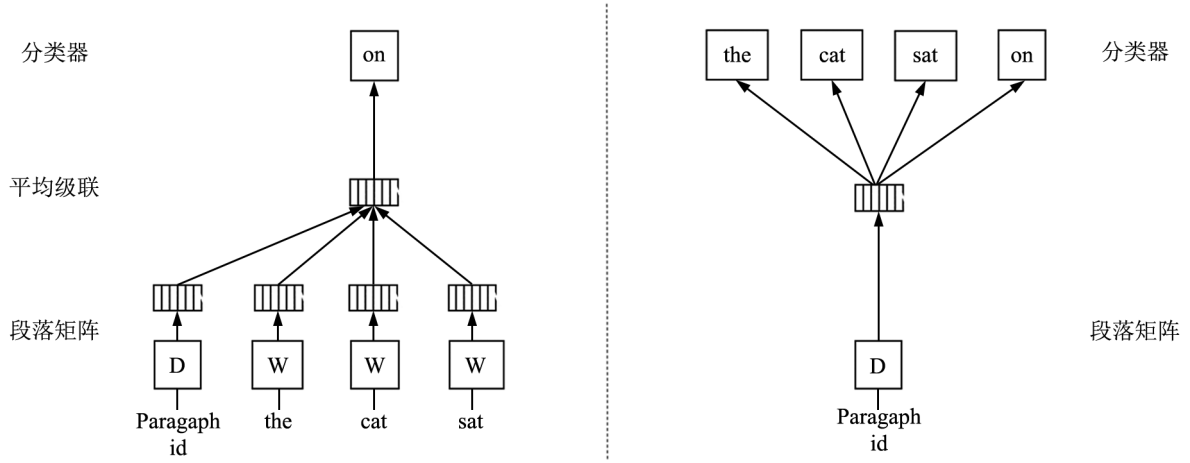

图 2 PV-DM 模型和 PV-DBOW 模型训练架构

Fig. 2 Training Architecture of PV-DM Model and PV-DBOW Model

表示；然后编写 Python 程序，借鉴平行四边形法则思想，对上述需求语义向量进行求和得到需求语义合向量，为下一步分析奠定基础。

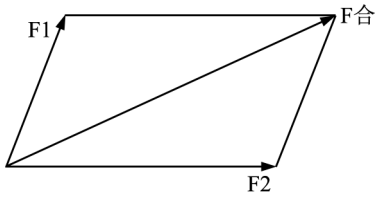

图 3 平行四边形法则

Fig. 3 The Parallelogram Law

基于上述分析，同时综合考虑本文数据规模和数据特征后，本文决定采用 Word2vec 向量模型中的 PV-DM 训练架构来进行文本向量的训练。具体处理过程如下：首先将经过分词和去停处理后的专利摘要文本逐条转化为专利语义向量；随后摘取产业发展规划(宏观层面)和市场技术需求(微观层面)相关文本，同样按照专利摘要文本处理方式将其转化为若干需求语义向量，并将这些语义向量求和得到需求语义合向量。

2）专利成果语义特征与需求语义特征匹配

将上文处理完成的专利语义向量逐个与需求语义合向量进行余弦相似度的计算，如式（1），得到“语义向量匹配度”特征。

$$similarity=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|} \tag{1}$$

其中 A、B 分别代表专利摘要语义向量和需求语义合向量，$\cos(\theta)$代表夹角度数，角度越小，余弦相似度越高。

2.2.3 专利成果质量类别标注

机器学习主要可分为监督学习和非监督学习两类，其主要区别在于是否使用人工标注数据集加以训练和测试。区别于非监督学习，在监督学习中每个实例的输入对象(数据特征)和输出值(监督信号)都是一一对应的，因此在监督学习中往往需要人工对数据集进行标注。此外，监督学习算法是通过分析尽可能多地训练数据，并由此产生推断的功能，从而对后续未知实例的标签进行预测。在本文的研究中，将整理好的专利形式特征和“语义向量匹配度”特征作为监督学习过程中的数据特征，将人工对专利成果质量的标注作为监督学习模型的监督信号，二者共同形成模型训练所需的数据集。

国家知识产权局于 2021 年 3 月首次将战略性新兴产业的发明专利、在海外有同族专利权的发明专利、维持年限超过 10 年的发明专利、实现较高质押融资金额的发明专利、获得国家科学技术奖或中国专利奖的发明专利定义为高价值发明专利[39]。但关于高质量专利的定义，目前学界尚未形成统一观点，如吴洁等[33]将专利维持年限用来表征专利质量；许鑫等[38]认为，高质量专利应具备较强市场竞争力和权力稳定性等特性；徐明等[40]则认为，专利质量的评价标准应当包括技术进步性和经济效益性。基于上述分析同时考虑数据可获取性及后续研究需要，本文参考并制定了以下专利质量标注规则，如表 1 所示。

2.2.4 高质量专利成果筛选方法构建

在构建高质量专利成果筛选方法前，需要对上

表 1 专利质量类别标注规则

Tab. 1 Rules for Categorizing Patent Quality

标 签	标注规则
1(高质量)	①该专利曾获国家或省(市)专利奖项[39] ②该专利在其发明单位的科技管理部门中被公布为已转化的高质量专利[40] ③该专利的专利权人发生转让[38,40]
0(低质量)	①该专利已失效且未发生专利权人转让[38,40]

表 2 模型评估指标计算公式

Tab. 2 Formula for Calculating Model Evaluation Metrics

指标名称	计算公式
准确率(A)	$A=\frac{TP+TN}{TP+TN+FP+FN}$
精确率(P)	$P=\frac{TP}{TP+FP}$
召回率(R)	$R=\frac{TN}{TN+FN}$
$F1$ 值	$F1=\frac{2*P*R}{P+R}$

述处理完成的专利成果形式特征、“语义向量匹配度”特征、专利质量类别标签进行汇总整理，按照8:2的比例划分为训练集和测试集，同时在划分过程中保证训练集和测试集均等比例包含 0 类和 1 类标签。完成数据集的划分后，本文进行高质量专利成果筛选方法的构建，构建过程主要包括：机器学习分类算法的选取与训练、机器学习分类算法训练效果评估、筛选方法应用测试 3 个部分。

1）机器学习分类算法的选取与训练

在机器学习中，常用分类算法模型主要有：K 近邻、线性支持向量机、逻辑回归、朴素贝叶斯、决策树、梯度提升决策树、随机森林、多层感知机等，鉴于不同算法各有优缺点，本文编写 Python 程序，分别调用 Sklearn 中集成的 8 类机器学习分类算法对训练集进行训练，然后依据训练结果对测试集进行测试，从中挑选出效果最优的分类算法模型。

2）机器学习分类算法训练效果评估

准确率(Accurate Rate)、精确率(Precision Rate)、召回率(Recall Rate)和 F1 值是机器学习研究中常用的模型评估指标。准确率为所有正确分类的专利文档数目与全部专利文档数的比率，能够较为直观、全面地衡量机器学习算法的识别和分类效果；精确率为准确分类专利文档数与所有预测为该类文档数的比率；召回率为准确分类文档数与实际文档数的比率；F1 值为精确度和召回率的调和平均数，是同时考虑精确度和召回率的综合性评价指标，上述指标计算公式如表 2 所示。基于上述分析，本文将从 8 类算法中评估出准确率、召回率和 F1 值最优的分类算法模型，用于高质量专利成果的筛选。

3）高质量专利成果筛选方法应用测试

所有机器学习分类算法完成测试并计算其准确率、精准率、召回率和 F1 值后，经人工综合判断选取其中性能最优的算法模型作为高质量专利成果筛选的分类算法。选取的分类算法与 Word2vec 向量模型等处理程序共同构成高质量专利成果筛选方法，并编写 Python 执行程序加载完成预处理的测试数据，完成应用测试。

3 实证分析

3.1 数据采集与处理

由于各省市战略目标定位与经济发展状况存在着一定程度的差异，导致其对科技成果也有着不同的需求，借此，本文将以湖北省为例，选取相关产业规划和市场需求文本，结合专利成果自身形式特征，进行高质量专利成果筛选方法的研究。在中央出台的《关于新时代推动中部地区高质量发展的意见》中，“坚持创新发展，构建以先进制造业为支撑的现代产业体系”被摆在首位，湖北省委第十一届九次全会中提出构建“51020”现代产业体系的设想，旨在发挥好湖北制造业大省的优势，为实现全省高质量发展奠定产业基础。基于上述分析，本文选取“先进制造与自动化”领域为研究样本，采集该领域相关数据开展研究。

湖北省“51020”现代产业体系发展规划和湖北省制造业发展“十四五”规划是湖北省推进制造业发展目标的集中体现，对科技成果发展方向具有指导意义。借此，本文摘取湖北省“51020”现代产业体系发展规划和湖北省制造业发展“十四五”规划两份文件中相关文本作为宏观层面的需求文本来源语料。

个人或企业在生产实践中遇到的难以攻克的技术问题时，往往需要吸收市面上的专利成果加以利

用。考虑到专利成果转化往往具有一定的时滞性，故本文编写 Python 程序爬取科惠网（http://www.51kehui.com/#/）中于 2019 年 1 月 1 日—2021 年 12 月 31 日登记的“先进制造与自动化”领域技术需求，经过人工逐条筛选剔除重复或无效数据后，最终获得 429 条有效数据作为微观层面的需求文本原始语料。

发明专利是新技术的重要展现形式。本文在专利之星检索平台（https://www.patentstar.com.cn/）中检索研究所需专利数据。根据研究需要对照 IPC 部类表检索 F 部类（机械工程，照明，加热，武器，爆破），检索时间窗口为授权日在 2019 年 1 月 1 日—2021 年 12 月 31 日的发明专利，检索时间为 2022 年 7 月 14 日，然后根据其主分类号剔除“先进制造与自动化”领域应用较少的 F21（照明）、F41（武器）、F42（爆破）二级类目专利，同时为保证数据的完整性，本文还剔除了不公告发明人的专利数据。筛选完成后的数据分别整理其摘要、发明人数量、IPC 号数量、知识产权局引证数量、说明书页数和实质审查时间，并严格按照上文表 1 制定的标注规则，人工逐条对专利成果质量进行 0（低质量）和 1（高质量）两类标注，处理完成后最终获得4 730条有效数据，其中 1 标签共2 147条，0 标签共2 583条。

3.2 高质量专利成果筛选方法构建

3.2.1 文本向量处理

为保证最终得到的文本向量具备可比性，本文统一设计处理程序，具体操作步骤如下：首先对专利摘要文本和需求文本进行分词和去停用词的处理；其次将处理好的专利摘要文本和需求文本按条目分别整理至 csv 文件中；随后利用 Doc2vec 对文本数据进行统一训练，使其转化为向量表示（程序参数设置为：向量维数 vector_size 设置为 100 维；最小语词忽略阈值 min_count 设置为 2；迭代次数 epochs 设置为 10），并将需求文本词向量求和得到需求语义合向量；最后将专利文本向量逐条与需求语义合向量进行余弦相似度的计算，得到“语义向量匹配度”特征，并整理汇总专利成果形式特征与“语义向量匹配度”特征（数据集示例如表 3 所示），编写 Python 程序进行模型的训练与评估。

表 3 数据集示例

Tab. 3 Example Dataset

类别标签	语义向量匹配度	发明人数量	IPC 号数量	知识产权局引证数	说明书页数	实质审查时间（天）
0	-0.0835348	1	3	0	25	503
0	-0.0556715	2	6	9	9	550
……			……			
1	0.1463725	1	8	7	13	196
1	-0.1141956	4	5	7	16	208
……			……			

3.2.2 算法模型训练与评估

为提高模型预测性能，本文在训练模型前首先调用 sklearn 中集成的网格搜索方法（GridSearchCV）对不同参数（参数组合）进行交叉验证以确定最优参数，按照最优参数构建模型训练方法，各模型最终选取的参数如表 4 所示。得到最优参数（参数组合）后，编写 Python 程序调用 8 类机器学习算法模型按照各自最优参数组合对训练集进行训练，并对测试集进行预测检验，各模型预测性能如表 5 所示。

从表 5 中可以看出，随机森林和决策树两种分类算法模型的整体准确率为 0.85，居所有测试算法模型首位，但从其内部的小类指标角度来看，随机森林算法模型的各项指标值分布相对更加均衡，这表明该模型对于专利成果质量的预测能力更加稳定，故本文决定选用随机森林算法模型作为高质量专利成果筛选方法中的内核分类算法。

3.2.3 高质量专利成果筛选方法应用测试

基于上文分析结果，本文选取 Doc2vec 向量模型、随机森林算法模型等程序，结合分词与去停用

表 4 算法模型最优参数组合

Tab. 4 Optimal Combination of Algorithm Model Parameters

模型名称	最优参数(参数组合)
K 近邻	K = 5
线性支持向量机	C = 0. 2 Gamma = 0. 01
逻辑回归	Penality = l2 C = 0. 5 Max_iter = 100 Solver = lbfgs
朴素贝叶斯	Alpha = 0. 01
决策树	Max_depth = 4
梯度提升决策树	Max_deepth = 4 n_estimators = 50 learning_rate = 0. 01
随机森林	Max_deepth = 2 n_estimators = 150
多层感知机	Activation = relu Alpha = 0. 05 Hidden_layer_sizes = [10,10] Max_iter = 1500

词等操作来构建应用测试模型并完成测试。应用测试模型包含两个脚本程序和一个专利特征库，两个脚本程序分别为：①txt_to_vector.py，其功能为处理文本向量并构建专利成果的“语义向量匹配度”特征；②prediction.py，其功能为应用随机森林模型，根据前期测试所得的最优参数对完成处理的专利成果质量进行判断。专利特征库中存有前期算法模型训练过程中形成的“语义向量匹配度”特征和专利成果的形式特征与质量标签。

为完成筛选方法的应用测试，同时考虑到需求文本的时效性，本文按照上文相同检索方式，在授权日在 2022 年 1 月 1 日—8 月 1 日专利中，随机挑选 10 条已转化(高质量)和 10 条未转化(低质量)的专利数据进行应用测试。具体处理过程如下：首先运行文本处理程序 txt_to_vector.py 将上述 20 条专利成果的摘要文本处理为语义向量，完成处理后逐条计算各项专利成果语义向量与需求语义合向量之间的余弦相似度，形成“语义向量匹配度”特征；随后将专利形式特征与“语义向量匹配度”特征进行汇总整理，整理结果如表 6 所示；最后运行 prediction.py 加载处理好的应用测试数据集，完成预测后根据其类别标签的不同输出预测结果，程序运行完成后输出与预测结果如图 4 所示(其中前 10 条为高质量专利，后 10 条为低质量专利)。

表 5 算法模型预测性能

Tab. 5 Predictive Performance of Algorithm Models

算法模型	K 近邻	线性支持向量机	逻辑回归	朴素贝叶斯	决策树	梯度提升决策树	随机森林	多层感知机
0 类别精准率	0. 80	0. 81	0. 68	0. 99	0. 84	0. 81	0. 87	0. 92
0 类别召回率	0. 77	0. 79	0. 83	0. 66	0. 89	0. 89	0. 85	0. 78
0 类别 F1 值	0. 78	0. 80	0. 75	0. 79	0. 86	0. 85	0. 86	0. 85
1 类别精准率	0. 74	0. 76	0. 72	0. 71	0. 86	0. 86	0. 82	0. 78
1 类别召回率	0. 76	0. 78	0. 54	0. 99	0. 80	0. 75	0. 85	0. 92
1 类别 F1 值	0. 75	0. 77	0. 62	0. 83	0. 83	0. 80	0. 83	0. 84
整体准确率	0. 77	0. 79	0. 70	0. 81	0. 85	0. 83	0. 85	0. 84

从图 4 中的预测结果来看，在 20 条应用测试专利中共有 16 条专利质量被正确预测(其中包含 9 条高质量专利和 7 条低质量专利)，整体识别准确率达到 0. 8，识别效果较好。此外，结合表 6 结果来看，随机森林算法模型的泛化性能为 0. 85，在实际应用测试中表现为 0. 8，训练效果和测试效果十分接近，这表明上文提出的高质量专利筛选方法具备一定的可靠性与稳定性，可考虑应用于后续大规模专利数据质量预测工作，辅助人工进行专利筛选，在一定程度上能够降低人力、物力、财力的消耗，提升专利筛选效率，从而达到促进专利成果高效转化的目的。

4 结论启示

本文基于“先进制造与自动化”领域专利数据，综合运用专利计量、自然语言处理、机器学习

表 6 整理后的应用测试数据集

Tab. 6 Refined Application Testing Dataset

专利申请号	语义向量匹配度	发明人数量	IPC 号数量	知识产权局引证数	说明书页数	实质审查时间（天）
CN202111383734. 9	-0. 0327716	1	4	4	17	60
CN202111058422. 0	0. 1130819	1	6	5	15	144
CN202111118162. 1	-0. 0574852	1	6	5	13	147
			……			
CN202010573213. 9	-0. 2141659	3	3	4	9	118
CN202080042776. 3	0. 1270986	5	6	5	11	603
CN202111181889. 4	-0. 0092674	1	4	0	9	568

图 4 应用测试数据集预测结果

Fig. 4 Predicted Results of Application Testing Dataset

等方法，结合湖北省自身特色需求，将专利形式特征和“语义向量匹配度”特征相结合，对专利成果质量的识别进行探索。研究发现，随机森林算法模型在选取的 8 种算法模型中，整体识别准确率和内部各小类指标综合表现最优，故本文选取随机森林算法模型作为高质量专利成果筛选方法中的内核分类算法，并结合 Doc2vec 向量模型等处理程序完成筛选方法的构建。此外，经过实证测试，本文提出的筛选方法基于客观数据综合考虑了专利成果的形式特征、地域发展规划与技术市场需求，能够较好地对专利质量进行预测，不仅有利于后续专利筛选工作的实际开展，同时还能为各省市政府相关部门提供决策支持，帮助其较为快速和全面地掌握专利成果整体质量情况，进而推动专利成果加速转化为生产力，助力经济高质量发展。

为更好地促进专利成果成功转化，提升成果利用效率，使其高效服务产业和经济发展规划的需要，本文提出以下启示：

1）规范数据采集，构建并不断完善各类数据库。当前各类专利信息和市场需求信息类型繁多且存在大量非结构化数据，在进行高质量专利成果筛选时往往需要花费大量时间和人力去进行数据采集和整理。因此，各省市主管部门（科技局、技术交易所等）可考虑与高校、企业进行合作，定期安排专人负责采集和整理专利成果数据、产业规划数据和市场需求数据，并将其按照专业领域或 IPC 号分类存储，构建并不断完善专利成果供应库、产业规划库、市场需求库、高质量专利成果库等特色数据库，并结合参考本文提出的筛选方法，实现专利成果供需关系的动态匹配，促进转化效率的提升，进而达到科技助力经济发展的效果。

2）加强引导效应，改革管理体制与反馈机制。从政府层面来看，应当积极引导专利成果申报和市场需求登记，一方面，在政策上支持专利成果的申请、审查和审批流程，强化企业（或成果持有人）的知识产权意识；另一方面，引导需求方积极登记，准确精练地表达自身技术需求。此外，科技局、技术交易所等科技成果转化主管部门可考虑在机器筛选的基础上，辅助组织人工进行随机检验，并根据检验结果及时反馈更新相关数据库，从而进一步提升高质量专利成果识别的准确率。从科研院所或企业层面来看，成果申报方应以解决实际问题为研究

导向，注重提升专利成果技术质量与撰写质量，加强产学研“一体化”协作，避免出现研发资源浪费和成果闲置。

5 结 语

世界新一轮科技革命为科技成果涌现创造了新的机遇，推进成果转化已成为大国博弈的战略选择。本文提出的筛选方法为实现自动化筛选高质量专利成果提供了参考方案，能够有助于识别具有潜在发展前景的专利成果，帮助科技主管部门精准施策，推动构建精准高效的专利成果转化机制。但同时本文也存在着一些有待完善之处，例如：①本文以湖北省为例，将其相关产业规划和技术需求与专利成果进行语义向量匹配度的计算，在需求文本数据范围的考量上可能有所欠缺，但本文主要提供一种筛选思路，旨在为各省(市)结合区域特色需求，从大规模专利中筛选出符合自身需求发展的高质量成果提供参考借鉴，以期降低专利成果闲置率并助力经济发展；②本文所构建的科技成果筛选方法是基于客观数据构建得来，虽然能够在一定程度上减轻相关人员的低级重复工作量并辅助决策，但缺少专家知识与经验的支持，未来将考虑在筛选过程中融入专家判断结果作为“案例语义特征”进行辅助筛选，以期更好地为专利成果转化工作提供决策支持。

参 考 文 献

[1] 贾雷坡，张志旻，唐隆华. 中国高校和科研机构科技成果转化的问题与对策研究 [J]. 中国科学基金，2022，36 (2)：309-315.

[2] 中华人民共和国国务院. 全国人民代表大会常务委员会关于修改《中华人民共和国促进科技成果转化法》的决定 [EB/OL]. http://www.gov.cn/xinwen/2015-08/30/content_2922110.htm，2022-07-13.

[3] 中华人民共和国国务院. 习近平主持召开中央全面深化改革委员会会议 [EB/OL]. http://www.gov.cn/xinwen/2021-05/21/content_5610228.htm，2022-07-13.

[4] 张静，徐海龙，王宏伟. 面向科技成果转化的服务需求研究 [J]. 中国科技论坛，2022，(9)：25-33.

[5] 方齐，谢洪明. 科技成果转化政策供给与政策协调的组态效应 [J]. 科学学研究，2022，40 (6)：991-1000.

[6] 张玉华，李茂洲，杨旭森. 基于主题模型的地方科技成果转化政策组态效应研究 [J]. 中国科技论坛，2022，(5)：11-20，30.

[7] 史童，杨水利，王春婧，等. 科技成果转化政策的量化评价——基于PMC指数模型 [J]. 科学管理研究，2020，38 (4)：29-33.

[8] 李巧莎，吴宇. 科技成果转化的政策优化策略 [J]. 宏观经济管理，2021，(10)：69-76.

[9] 钱学程，赵辉. 科技成果转化政策实施效果评价研究——以北京市为例 [J]. 科技管理研究，2019，39 (15)：48-55.

[10] Sousa M J，Jamil G，Walter C E，et al. Big Data Analytics on Patents for Innovation Public Policies [J]. Expert Systems，2021，40 (1)：12673.

[11] Lu Y X，Lai C C. Effects of Patent Policy on Growth and Inequality：A Perspective of Exogenous and Endogenous Quality Improvements [J]. MPRA Paper，2021.

[12] 孟祥利，曹源，王巨汉. 高校科技成果转化的困境与对策 [J]. 中国高校科技，2020，(9)：94-96.

[13] 尹岩青，李杏军. 国防科技成果转化的现状与问题研究 [J]. 科学管理研究，2017，35 (5)：26-29.

[14] 何丽敏，刘海波，许可. 国有资产管理视角下央企科技成果转化制度困境及突破对策 [J]. 济南大学学报 (社会科学版)，2022，32 (3)：102-110.

[15] Fan X，Hua F. Solving the Dilemma of Transferring and Transforming Scientific and Technological Achievements Through Systematic and Mechanism Reforms [C] //AEIC Academic Exchange Information Centre (China). Atlantis Press，2019：678-686.

[16] Lei K，Zhang M. Research on Transformation Mode of Scientific and Technological Achievements with Different Transaction Costs：Taking the Electric Power Industry as an Example [J]. MATEC Web of Conferences，2021，336 (1)：09016.

[17] 卜伟，郑园园，陈军冰. 江苏高校科技创新政策绩效评价——基于层次分析-熵值法和K-means聚类分析法 [J]. 科技管理研究，2022，42 (24)：118-124.

[18] Liang Z，Anni Y. Design of Performance Evaluation System for Transformation of Patent Achievements in Colleges and Universities Based on AHP [C] //CIPAE 2021：2021 2nd International Conference on Computers，Information Processing and Advanced Education，2021.

[19] 张浩，霍国庆，汪明月，等. 科技成果转化的战略绩效评价——基于国家科学技术进步奖成果的实证研究 [J]. 科学学与科学技术管理，2020，41 (8)：7-25.

[20] Wang X，Liu Y，Chen L. Innovation Efficiency Evaluation Based on a Two-Stage DEA Model With Shared-Input：A Case of Patent-Intensive Industry in China [J]. IEEE Transactions on Engineering Management，2021，(99)：1-15.

[21] 郭颖，段炜钰，孟婧，等. 中国科学院产学研合作网络特征对其科技成果转化绩效的影响 [J]. 中国科技论坛，2022，(5)：81-89.

[22] 马大来，叶红. 供给侧结构性改革视角下中国科技成果转化绩效研究——基于空间面板数据模型的实证分析 [J]. 重庆大学学报 (社会科学版)，2020，26 (1)：45-60.

[23] Noh H, Lee S. What Constitutes a Promising Technology in the Era of Open Innovation? An Investigation of Patent Potential from Multiple Perspectives [J]. Technological Forecasting and Social Change, 2020, 157: 120046.

[24] 田雪姣, 鲍新中, 杨大飞, 等. 基于熵权-TOPSIS-德尔菲法的核心技术识别研究——以芯片产业技术为例 [J]. 情报杂志, 2022, 41 (8): 69-74, 86.

[25] 赵蓉英, 李新来, 李丹阳. 专利引证视角下的核心专利研究——以人工智能领域为例 [J]. 情报理论与实践, 2019, 42 (3): 78-84.

[26] Barberá-Tomás D, Jiménez-Sáez F, Castelló-Molina I. Mapping the Importance of the Real World: The Validity of Connectivity Analysis of Patent Citations Networks [J]. Research Policy, 2011, 40 (3): 473-486.

[27] Jaehyun C, Dongsik J, Sunghae J, et al. A Predictive Model of Technology Transfer Using Patent Analysis [J]. Sustainability, 2015, 7 (12): 16175-16195.

[28] 杨恒, 王曰芬, 张露. 基于核心专利技术主题识别与演化分析的技术预测 [J]. 情报杂志, 2022, 41 (7): 49-56.

[29] Krestel R, Chikkamath R, Hewel C, et al. A Survey on Deep Learning for Patent Analysis [J]. World Patent Information, 2021, 65: 102035.

[30] Liu B C, Lai M Z, Wu J L, et al. Patent Analysis and Classification Prediction of Biomedicine Industry: Som-kpca-svm Model [J]. Multimedia Tools And Applications, 2020, 79 (15-16): 10177-10197.

[31] Hu Y F, Yang S W, Shi A J, et al. Research on Transferable Patent Recognition Based on Machine Learning [C] //Proceedings of 2021 4th International Conference on Data Science and Information Technology, 2021.

[32] 张彪, 吴红, 高道斌, 等. 基于潜在高被引论文与高价值专利的创新前沿识别研究 [J]. 图书情报工作, 2022, 66 (18): 72-83.

[33] 吴洁, 桂亮, 刘鹏. 基于图卷积网络的高质量专利自动识别方案研究 [J]. 情报杂志, 2022, 41 (1): 88-95, 124.

[34] 付振康, 柳炳祥, 鄢春根, 等. 专利寿命视角下的核心专利识别模型构建 [J]. 图书馆论坛, 2023, 43 (1): 112-120.

[35] Schettino F, Sterlacchini A, Venturini F. Inventive Productivity and Patent Quality: Evidence from Italian Inventors [J]. Journal of Policy Modeling. 2013, 35 (6): 1043-1056.

[36] Albert M B, Avery D, Narin F, et al. Direct Validation of Citation Counts as Indicators of Industrially Important Patents [J]. Research Policy, 1991, 20 (3): 251-259.

[37] 冉从敬, 宋凯. 高校可转化专利识别模型构建——以人工智能领域为例 [J]. 情报理论与实践, 2020, 43 (11): 79-85.

[38] 许鑫, 赵文华, 姚占雷. 多维视角的高质量专利识别及其应用研究 [J]. 现代情报, 2019, 39 (11): 13-22, 45.

[39] 中华人民共和国中央人民政府. 国家知识产权局关于印发《推动知识产权高质量发展年度工作指引(2021)》的通知 [EB/OL]. https://www.cnipa.gov.cn/art/2021/3/5/art_75_157233.html, 2022-08-01.

[40] 徐明, 陈亮. 基于文献综述视角的专利质量理论研究 [J]. 情报杂志, 2018, 37 (12): 28-35.

(责任编辑: 郭沫含)

(上接第 42 页)

[23] 郭庙琴. 基于用户需求的科学数据与科技文献关联模式研究 [D]. 长春: 东北师范大学, 2021.

[24] 黄筱瑾. 基于元数据的科学数据与科技文献关联研究 [J]. 情报理论与实践, 2013, 36 (7): 27-30.

[25] 黄筱瑾. 基于内容特征的科学数据与科技文献关联研究 [J]. 现代情报, 2018, 38 (1): 56-59.

[26] 邱春艳. 期刊文献与科学数据的关联服务研究 [J]. 情报资料工作, 2014, (2): 63-66.

[27] Elixir Data Platform [EB/OL]. https://elixir-europe.org/platforms/data, 2023-03-10.

[28] OpenAIRE. OpenAIRE - research Graph [EB/OL]. https://graph.openaire.eu, 2023-05-20

[29] 卫军朝. 科学文献与科学数据关联实践研究——以 Elsevier 为例 [J]. 国家图书馆学刊, 2017, 26 (3): 93-101.

[30] 杨宁, 文奕, 张鑫, 等. 高能物理科学数据与科技文献关联研究 [J]. 图书馆学研究, 2019, (1): 47-52.

[31] 陈帅印, 刘桂锋, 刘琼. 基于元数据的数据论文与期刊论文的关联研究——以全球变化科学研究数据出版系统为例 [J]. 数字图书馆论坛, 2022, (8): 11-18.

[32] Dryad Executive Director to Pursue New Opportunities [EB/OL]. https://blog.datadryad.org/2018/02/, 2023-05-12

[33] 潘有能, 吕晶晶, 丁楠. 基于 Labeled-LDA 模型的科学数据与科技文献关联识别研究——以生物医学领域为例 [J/OL]. 情报科学: 1-13 [2023-06-04]. http://kns.cnki.net/kcms/detail/22.1264.G2.20230509.1613.034.html.

[34] Elsevier and PANGAEA Link Contents for Easier Access to Full Earth System Research [EB/OL]. https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-and-pangaea-link-contents-for-easier-access-to-full-earth-system-research, 2023-04-10.

[35] DRYAD [EB/OL]. Our platform. https://datadryad.org/stash/our_platform, 2023-03-10.

(责任编辑: 杨丰侨)