情报学报　2023 年 6 月　第 42 卷　第 6 期
Journal of the China Society for Scientific and Technical Information, Jun. 2023, 42(6): 663-680

DOI: 10.3772/j.issn.1000-0135.2023.06.003

一种基于特征拼接、标签迁移及深度学习组合的专利价值评估方法

赵雪峰[1]，胡瑾瑾[1]，吴德林[1]，吴伟伟[2]，孙安东[3]，赵　涛[4]

（1. 哈尔滨工业大学（深圳）经济管理学院，深圳　518000；2. 哈尔滨工业大学经济与管理学院，哈尔滨　150000；
3. 深圳沃德知识产权代理事务所，深圳　518000；4. 深圳盈峰知识产权咨询有限公司，深圳　518000）

摘　要　专利价值评估对打击非正常申请、净化市场环境具有重要的现实意义。本文以特征拼接、标签迁移及深度学习组合为中心构建专利价值评估方法，并基于2010—2020年广东省专利申请探究评估方法实际表现，引入多组对比模型进行实验分析。研究结果表明：①拼接著录事项信息，可构建出技术特征显现更强的专利研究对象，避免因指标类研究对象未足够体现专利技术本质而引发评估准确率不高的现象；②以专利法律视角量化出更具专利价值代表性的价值标签体系，在延展专利标签体系研究深度的同时，解决因引用率、下载访问量等传统标签与专利实际价值不匹配而造成的价值评估错误问题；③以高精细度词向量为构建原理，组建以BERT（bidirectional encoder representations from transformers）及LSTM（long short-term memory）为核心的专利价值评估模型，有效解决传统模型特征因提取能力不足而产生评估准确率偏低的弊端。本文从研究对象有效性、标签体系性及模型构建评估率三个方面提出优化改进策略，为专利价值评估提供了新工具，具有较强的实际应用价值。

关键词　特征拼接；BERT；标签迁移；专利；深度学习

Patent Valuation Method Based on a Combination of Feature Stitching, Label Migration, and Deep Learning

Zhao Xuefeng[1], Hu Jinjin[1], Wu Delin[1], Wu Weiwei[2], Sun Andong[3] and Zhao Tao[4]

(1. *School of Economics and Management, Harbin Institute of Technology, Shenzhen, Shenzhen*　518000;
2. *School of Management, Harbin Institute of Technology, Harbin*　150000; 3. *Shenzhen Ward Intellectual Property Agency, Shenzhen*　518000; 4. *Shenzhen Yingfeng Intellectual Property Consulting Co., Ltd, Shenzhen*　518000)

Abstract: Patent value evaluation is of great practical significance in cracking down on abnormal applications and purifying the market environment. This study uses feature combination, label migration, and deep learning combination to construct a patent valuation method and explores actual performance based on patents in Guangdong Province from 2010 to 2020. Several sets of comparative models are introduced for experimental analysis. Our findings reveal the following conclusions. (1) Stitching together the information of bibliographic documents can construct more powerful patented research objects with technical characteristics, thereby overcoming the phenomenon that evaluation accuracy is not high owing to

收稿日期：2022-07-08；**修回日期**：2022-10-24

基金项目：国家自然科学基金面上项目“大数据能力驱动的突破性技术创新行为触发路径与演化机制”（72072047）；教育部人文社会科学研究青年基金项目“基于企业行为理论的创业导向对突破性技术创新行为影响的传导机制研究”（20YJC630090）；中央高校基本科研业务费专项资金项目“面向新动能塑造的技术管理能力对突破性创新行为的作用机制研究”（HIT.HSS.ESD202310）。

作者简介：赵雪峰，男，1993年生，博士研究生，研究方向为知识产权管理及专利价值分析；胡瑾瑾，女，1997年生，博士研究生，研究方向为数字经济；吴德林，男，1963年生，博士，教授，博士生导师，研究方向为信息有用性；吴伟伟，男，1978年生，博士，教授，博士生导师，研究方向为技术管理与创新管理，E-mail：wuweiwei@hit.edu.cn；孙安东，男，1992年生，本科，研究方向为专利价值分析；赵涛，男，1996年生，本科，研究方向为专利价值分析。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

the insufficient reflection of the nature of patented technology. (2) We can quantify a more representative patent value from the patent law. While extending the research depth, the mismatch between traditional labels and the actual value of the patent is also resolved. (3) A patent value evaluation model with bidirectional encoder representations from transformers (BERT) and long short-term memory (LSTM) is established based on the construction principle of high-precision word vector, effectively solving the disadvantages of low evaluation accuracy caused by the lack of feature extraction ability of traditional models. This study has a strong application value and presents improvement strategies from the three aspects of research object effectiveness, label system, and model evaluation rate, providing a new tool for patent value evaluation.
Keywords: feature combination; BERT; label migration; patent; deep learning

0 引 言

专利是知识产权核心制度，具有创造性、新颖性和技术公开的特点[1-3]，以及提升公民及企业研发积极性、促进国家科技发展等重要作用[4-7]。我国自1984年实施《中华人民共和国专利法》至今，专利综合发展指数以每年11.8%的增速跃升至全球第八，成绩斐然[8-9]。在此期间，为强化公民及企业的知识产权保护意识，我国出台了一系列减免补贴的专利申请鼓励政策，导致了申请量与创新科研水平不相符的现像，简称“非正常申请”[10-12]。非正常申请导致我国知识产权环境指数低下，影响了综合发展指数的进一步提高，因此，充分评估每件专利的技术价值，进而遏制异常申请、净化市场环境，已刻不容缓[13-14]。

目前，已有一批优异的专利价值评估方法，根据其研究对象的不同，主要分为离散型评估和连续型评估。

离散型评估主要以发明人及申请（专利权）人数量、研发资金占比、研发人员规模等彼此离散且互不关联的专利指标作为专利价值的研究对象。刘夏等[6]以2010—2011年国家知识产权局受理的85万余件专利申请为研究对象，抓取申请文档中以及相关引文的特征信息，搭建完整的随机森林模型，对后续被引情况进行机器学习及预测；刘大勇等[15]通过追踪2006—2014年中国230个城市的发明专利，构建了城市层面的科技成果转化指标；Liu等[16]以离散网络构建专利转换预测方法发现高质量专利；李治东等[17]用专利申请、授权和无效三个阶段构建新的指标体系，并对指标体系识别方法中主观性较强的特点，采用主观与客观相结合的熵权层次分析法对指标进行赋值，增加了指标间的客观分析，并对指标选取和赋值方法的可行性和科学性进行了验证；Trappey等[18]将可识别离散指标关联性的主成分分析（principal component analysis，PCA）融合至深度神经网络（deep neural networks，DNN），得到可从海量专利中自动识别高价值专利的方法；Wang等[19]基于图概率模型及知识表示学习构建TDLDA（time-based dynamic latent Dirichlet allocation）模型，并利用TDLDA探究区块链对专利价值的贡献。

连续型评估主要以包括文字或图片的著录事项书、专利申请文件、说明书附图等在内的数据为专利价值研究对象。Huang等[20]引入基于公共词典构造的先验知识网络，得到可在海量专利申请文件中发现高价值专利关键词的新型TextRank模型；Chung等[21]采用CNN（convolutional neural network）网络及LSTM（long short-term memory）构建高质量专利挖掘模型，并将三级专利等级作为分类标准，进而捕获专利文本的语义特征，达到提高高质量专利评估准确率的目的；Zhu等[22]将专利标题和摘要作为输入数据，利用词嵌入技术对输入数据进行分割和矢量化，构建对称分层卷积神经网络完成专利价值识别；Wu等[23]利用前馈神经网络和自注意机制组成实体识别网络，该网络可提取专利上下文信息，进而提高专利价值评估准确率；Ni等[24]将经过训练的双向LSTM神经网络集成至SAM-IDM模型中，从而提高高质量专利的识别效率。

综上可见，以往学者提出了丰富的方法与模型，拓展了研究深度，但目前专利价值评估依然存在以下问题待克服。

（1）研究对象的技术特征显现不足。目前专利价值的研究对象主要集中于离散特征或单一文本，虽然可适当体现专利价值性，但就专利价值本质是由技术创造性及新颖性决定的而言[25-26]，权利要求数、申请类型、引用量等离散指标或单一文本并未足够深入体现技术本质。由于不具备代表技术特征实质性的能力，若仅将离散特征或单一文本作为研究对象，则容易因专利价值特征显现力不够，引发评估准确率不高的问题。

（2）价值标签无法准确表示专利价值。标签是

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一种量化专利技术价值的手段，如 Chung 等[21]以引用率为标准，将专利价值分为 A、B、C 三种等级。但专利所承载的技术因技术方向不同，往往具有冷门、热门之分，热门技术受众广、研究人员多，其引用率、下载量等指标也远大于其他方向技术，但其专利价值并不一定高于冷门方向专利。因此，传统标签体系无法准确代表专利价值，直接影响价值评估的公正性。

（3）评估模型特征提取能力较弱。现有研究主要将 XGBoost（extreme gradient boosting）、SVM（support vector machines）、CNN 等表现优异的机器学习或神经网络作为评估模型[27]，虽然具备专利价值评估能力，但模型构造相对简单，无法充分提取出专利的技术特征，导致价值评估准确率有待提升。

针对上述问题，本文提出由特征拼接、标签迁移及深度学习组合而构建得到的专利价值评估方法。首先，通过特征拼接方法，重组专利著录事项所记载的技术实施过程、申请（专利权）人等信息，进而构建得到技术特征显现力更强的研究对象；其次，将标签体系构建视角迁移至专利所经历的法律状态，通过法律状态的流程严谨性和技术专业性，量化出更具专利价值代表性的法律类专利价值标签；最后，通过组合表现优异的深度学习模型，增强对研究对象的技术特征提取能力，从而提高专利价值评估准确率。总的来说，特征拼接、标签迁移及深度学习组合分别从特征、标签及价值评估模型三个方面提出改进，可有效提高专利价值评估的准确率。

1　构建专利价值评估方法

本文以特征拼接、标签迁移及模型组合对应改善研究对象、标签体系及评估模型为原理，依次详述构建过程。

由图 1 可知，首先，特征拼接和标签迁移在数据构建阶段对应生成研究对象和价值标签；其次，在模型组合阶段，连接 BERT（bidirectional encoder representations from transformers）与 LSTM；最后，利用研究对象和价值标签训练并测试组合模型，从而实现专利价值评估。

1.1　特征拼接

连续型文本相近语段间一般具有逻辑关联性[28]，特别是围绕专利申请所生成的著录事项书，依据法律严谨性要求，其著录事项信息之间具有更强的逻辑关联，因此，从著录事项书中提取重要特征，并通过特征拼接组合生成新的研究对象用于评估专利价值，可有效增强高价值专利的特征显现，提高评估准确度。

著录事项书记载了专利申请、审查、授权及权利丧失等阶段下专利信息的产生、变更及消亡的全生命周期轨迹[29]。其中，专利信息包括申请方式、申请号、公布日期、分类号、发明名称、专利权人、年费缴纳及海关备案等[30]。

由图 2 可知，申请阶段已包括“申请号、公布号、公布日、申请人、发明人、地址、摘要及摘要附图”等复杂的著录事项信息，如何从庞大复杂的专利信息中提取可显著表现专利价值的特征，是本文首先需要解决的问题。

《专利审查指南》是专利法及其实施细则的具体化，详细规定了著录事项书对专利信息的记载标准和流程规则。本文参照 2021 年最新版《专利审查指南》[29]对著录事项书中各专利信息的阐述，以对专利价值影响力为原则，将各专利信息划分为普适型、离散特征型、连续特征型及标签型四大类。其中，标签型专利信息用于后续标签细化，普适型、离散特征型、连续特征型用于形成专利价值的研究对象。

表 1 展示了以普适型、离散特征型、连续特征

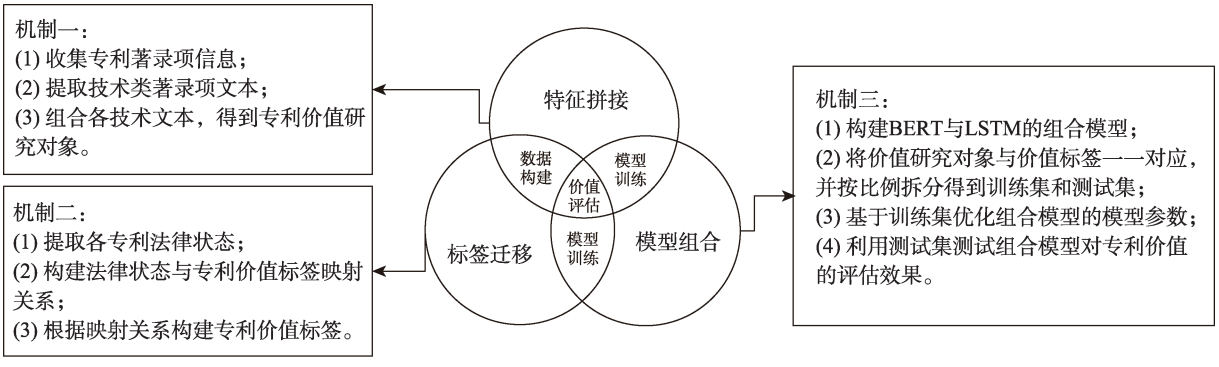

图 1　专利价值评估的构建过程

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net


(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106686082 A
(43)申请公布日 2017.05.17

(21)申请号 201611249851.5

(22)申请日 2016.12.29

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 孙勐 陈安伟

(51)Int.Cl.
H04L 29/08(2006.01)

权利要求书2页 说明书6页 附图5页

(54)发明名称
存储资源调整方法及管理节点

(57)摘要
本发明提供一种存储资源调整方法及管理节点。所述管理节点连接于业务节点与存储设备之间。所述管理节点执行所述方法以接收业务节点发送的负载指标，所述负载指标为所述业务节点采集的业务节点在运行应用时的性能参数，并在原始采集数据库中查找所述业务节点所运行应用对应的存储区域的至少一个存储指标的采集值，所述原始采集数据库为所述存储设备周期性采集的存储设备中各存储区域的存储指标的值，所述存储指标为每个存储区域的性能参数，然后对每个存储指标的采集值进行分析，确定需要调整的存储指标，再对所确定的需要调整的存储指标进行调整。本发明可以根据业务节点运行的应用的需求对存储设备的存储资源进行调整。

图2 著录事项部分专利信息在申请公示阶段的显示

表1 基于著录事项的专利信息划分

类别	编号	专利信息	解释或示例
普适型	1	申请号	CN202011000056.9
	2	公布日期	2020/11/10
	3	联系人	张三
离散特征型	1	专利申请类别	发明、实用新型、外观设计
	2	优先权	具有/不具有/部分具有
	3	权利要求数量	最少为1,平均为10
连续特征型	1	专利权人	平安科技(深圳)有限公司
	2	发明名称	病灶检测分析方法、装置、电子设备及计算机存储介质
	3	摘要	本发明公开了一种车损识别方法及服务器,该方法包括:接收用户通过第一终端发出的定损请求……
	4	所属技术领域	人工智能技术领域

型为依据划分出的部分专利信息。其中，普适型专利信息又称必要信息，具有标识专利申请唯一性或记录专利非技术信息的作用。例如，表1中每份专利均对应互不相同且唯一的申请号，具有索引标识的作用。此外，联系人具有配合国家知识产权局推进审查阶段需提供实物展示、优先权证明等事宜的作用。由此可见，普适型信息具有普遍存在于每件专利、但又因不具备技术属性而无法体现专利价值的特征，因此，本文将普适型信息用于标识索引专利，以便于后续专利价值评估在训练及测试阶段的效果展示。

离散特征型信息，又称指标专利信息，结合引言所述，其主要以数量、类别作为展现形式，且彼此离散不直接关联，如专利申请类别包括的发明、实用新型及外观设计三种类别。因《专利审查指南》规定权利要求的数量不少于1项，且大于10项额外收费，产生了权利要求数平均为10项的现象。Nagler等[26]、Yu等[31]研究结果显示，发明专利创造性远大于实用新型与外观设计，权利要求数量正向影响专利价值等，故离散特征可适当体现专利价值性，但追述专利价值本质是由所记录技术的创造性及新颖性决定发现的[27]，因此权利要求数、申请类型、引用量等离散指标并未足够深入体现技术本质特征。以申请号CN201010033923.9发明专利为例，虽然其权利要求数高达12项，但若深入了解主权利要求中所记录的技术实施步骤——“一种动画播放方法，其特征在于，包括：预设附加动画的图元在附加动画各帧下的观看属性，更新所述附加动画的图元在附加动画当前帧下的观看属性”，可发现主权所记录实施步骤仅在于在动画图元中添加观看属性，并未足够体现专利价值，因而被驳回。因此，综合来说，以离散特征作为研究对象在一定程度上可达到评估专利价值的目的[32]，但由于其并不具备代表技术特征实质性的能力，若仅将其作为研究对象，则容易因专利价值特征显现力不够导致评估准确率不高的问题。

连续特征型信息，又称连续专利特征，其主要以文本或图片等作为展现形式。相比于离散特征，连续特征具有更强的上下关联性和逻辑关系，特别是在专利领域中，体现专利价值的技术特征一般均以连续文本的形式记录在著录事项书及法律申请文件中。就表1著录事项书所记载的专利申请文件摘要来说，《专利审查指南》规定，“摘要应当写明发明名称和所属技术领域，并清楚地反映所要解决的技术问题、解决该问题的技术方案的要点以及主要用途，且不能超过300字”。由此可发现，摘要以简明、扼要的连续文本提纲挈领出专利技术特征。以申请号CN201611249851.5的发明专利为例，其对应的摘要为，“本发明提供一种存储资源调整方法及管理节点。所述管理节点连接于业务节点与存储设备之间。所述管理节点执行所述方法以接收业务

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

节点发送的负载指标，所述负载指标为所述业务节点采集的业务节点在运行应用时的性能参数，并在原始采集数据库中查找所述业务节点所运行应用对应的存储区域的至少一个存储指标的采集值，所述原始采集数据库为所述存储设备周期性采集的存储设备中各存储区域的存储指标的值，所述存储指标为每个存储区域的性能参数，然后对每个存储指标的采集值进行分析，确定需要调整的存储指标，再对所确定的需要调整的存储指标进行调整。本发明可以根据业务节点运行应用的需求对存储设备的存储资源进行调整”。可见其具有强烈的专利技术特征显现功能，充分揭露了此次申请的技术背景、实施细节及用途。因此，通过分析摘要所显现的技术特征，探究专利价值性具有重要的研究价值。

此外，参照《中华人民共和国专利法》《中华人民共和国专利法实施细则》关于发明名称的定义“发明名称应当简短、准确地表明发明专利申请要求保护的主题和类型，且一般不得超过 25 个字”可知，发明名称可言简意赅地表达专利申请的技术类型。因此，通过发明名称可有效识别专利所属技术领域，进而判断专利价值。另外，王玲等[33]研究表明，专利申请（专利权）人对于专利质量及专利技术方向具有重要影响。因此，将上述三种具有强烈特征表现类型的连续文本，以“专利申请（专利权）人+发明名称+摘要”的形式拼接得到用于专利价值评估的研究对象。

1.2　标签迁移

标签是为量化专利价值而实施的一种可视化手段。Chung 等[21]以每年被引用数为标准，将专利分为 A、B、C 三个标签等级，其中 A 等级专利的被引数排名前 20%，B 等级专利排名前 21%~60%，C 等级排名为末尾 40%。

不同学者对标签定义有所差别，但已达成共识的是，标签的作用是有效识别不同专利所承载的技术价值，且应在价值评估模型具备较强评估能力的基础上，最大可能地精细化，从而提升专利与专利之间的价值差异性，故 A、B、C 三种类型的专利价值标签等级所体现的专利价值差异性能力应强于仅有 A、B 两种价值标签的等级。

因此，本文在实现特征拼接的基础上，进一步细化出多种标签，从而提高专利价值评估精细度。标签细化的手段较多，如上述 Chung 等[21]从大众视角出发，利用被引用数衡量专利价值，其被引用数越高，证明专利所记载的技术已被公众承认，其专利价值在理论上也越高。

由图 3 可知，虽然该方法已被证实具有可行性，但专利所承载的技术因方向不同，往往具有冷门、热门之分，而热门技术如计算机通信，在深度学习、5G 网络的技术革新下，其受众广，研究人员多，其引用率也自然远大于其他方向技术。显而易见，因受众原因所产生的高引用率专利，其价值性不一定高于其他方向专利。

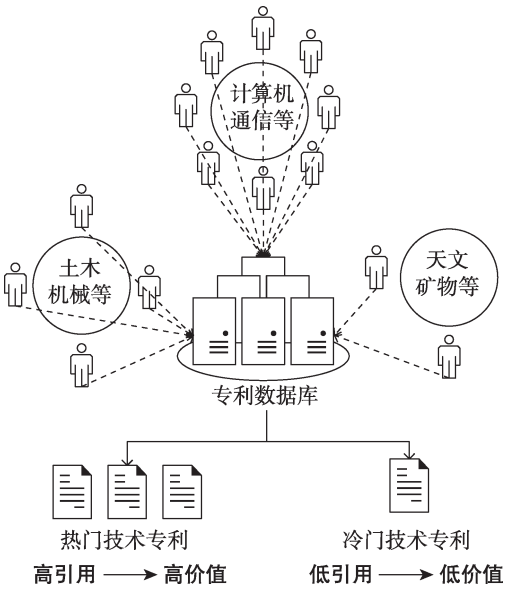

图 3　被引用数衡量专利价值的可视化过程

Marco 等[34]通过实证分析发现，包括审查步骤、审查强度等法律审查流程的变化会影响专利质量；Mossinghoff 等[35]也证实，不规范的审查制度会不可避免地影响专利价值；Rai[36]研究表明，法律规定的专利许可和诉讼成本的增加会导致专利质量下降，从而削弱社会创造积极性；其他学者也多次阐述审查制度、诉讼、无效、专利许可及海关备案等法律流程对专利质量的影响[37-39]。由此可见，专利法律状态与专利质量具有内在关系。

因此，为克服利用被引用数细化标签所带来的问题，本文将大众视角迁移至法律视角，巧妙地从法律角度寻找出与专利技术方向严格对应的专业人员或团队，通过著录事项书确定专业人士处理专利所得出的法律状态，并进一步以法律状态匹配专利价值，消除专利高引用率无法严格代表高价值性的弊端。

由图 4 可见，发明专利法律状态的变化过程极其严谨且复杂，从递交、初审、公开、实审直至专利权终止等环节，涉及众多知识产权从业人员、技

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

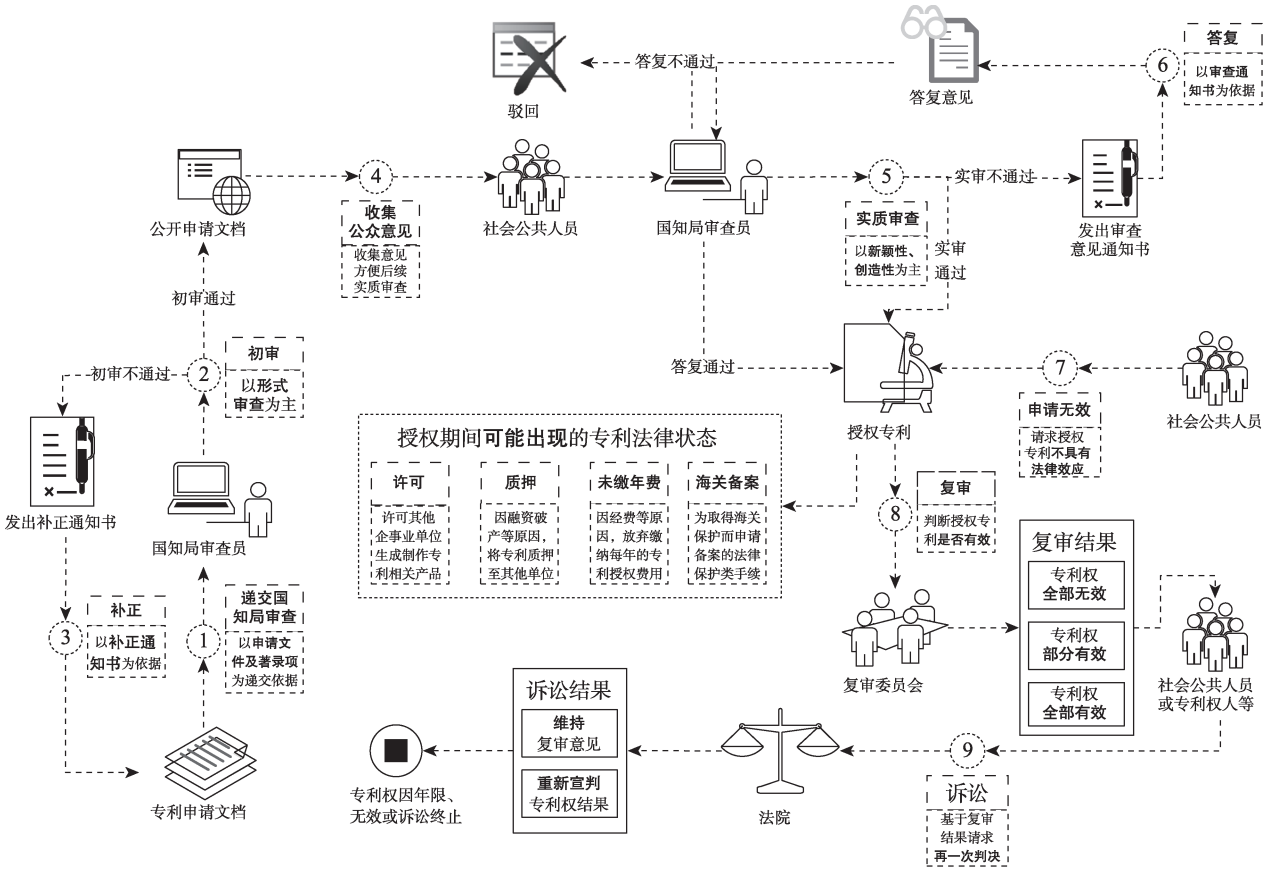

图4 发明专利法律状态的全生命周期过程

术研发团队及公众的监督及审查，因此，整个专利法律状态变更过程恰好也是专利价值性被反复验证的过程。

以申请号CN201110089122.9、发明名称“电动独轮自行车”专利为例，技术发明人将“电动独轮自行车”相关申请文件及著录事项等文档递交至国家知识产权局，并于2011年10月8日生成申请号后，审查员启动初审程序，并未发现文档存在形式错误，故同年12月14日公开；在公开期间，关心电动车或自行车领域的公众人群及企事业单位学习“电动独轮自行车”的技术实施手段，并反馈意见至国家知识产权局；当审查员开启实质审查时，通过公众反馈意见及专利数据库，详细比对“电动独轮自行车”技术的新颖性和创造性，发现其满足专利法授权要求，并于2013年8月21日授予专利权，同年11月5日，常州爱尔威智能科技有限公司认为其不具备授权标准，请求复审委员会无效该专利，但2014年7月16日复审委员会宣判，通过再一次审查认为，“电动独轮自行车”专利具备专利权，维持第一次审查结果，该专利授权状态也一直维持至今。

自2011年10月开始申请至2014年7月申请无效宣判的近三年时间内，“电动独轮自行车”专利虽经审查员、公众人群、竞争对手、复审委员会等一系列专业审查，但最终仍维持专利权有效，其专利的高价值性早已不言而喻。

综合来说，专利法律流程包括初审、实审、无效、诉讼等严谨复杂环节，且因不同专利申请保护技术的不同，在一系列流程中，所关联的技术申请人、审查员、公众人群等也都具备与所保护技术对等的研究或从业背景。因此，在法律流程严谨性和人员-技术严格对应性的作用下，所产生的法律标签可有效代表专利价值，高价值专利因自身技术已被同行业人员反复证实，具备新颖性和创造性，自初审至诉讼仍能维持有效，而低价值专利则因技术简单、仿造等难以通过实审。由此可见，本文巧妙地将大众视角迁移至法律视角，通过法律严谨、复杂的流程设计，寻找出与专利技术方向严格对应的专业人员，进而通过著录事项书确定在各法律阶段下专业人员处理专利所得出的法律状态，并进一步以法律状态匹配出专利价值，从而消除因被引用数、下载量等专利指标衡量专利价值受限于技术热

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

度，从业人员数量等原因导致其无法严格代表专利价值的弊端。

1.3　模型组合

当完成特征拼接及标签迁移细化得到研究对象与标签后，需构建表现优异的价值评估模型以提取研究对象的技术特征，从而基于技术特征与标签的对应关系完成专利价值评估。

本文主要以 BERT 和 LSTM 构建专利价值评估模型，简称 BLModel（spiral patent value evaluation model based on LSTM and BERT）。其中，BERT 是一种基于多头注意力机制（multi-head attention）和 Mask 方法的预训练语言表征模型，其可精确提取并表示专利技术特征[40-42]；LSTM 是改良版循环神经网络，用于解决在梯度反向传播过程中因逐步缩减而产生专利技术信息丢失的问题[43-45]。由于高价值专利具有启发研究人员、引导技术发展、提升我国专利市场环境的重要作用，因此，以摘要、专利权人拼接的研究对象必然包括重要的技术特征，若在价值评估时对拼接文本的词向量操作细腻度不高，则会直接影响识别准确率，而 BERT、LSTM 恰恰具有强大的专利技术特征提取能力，可以提高词向量细腻度。

由图 5 可知，BLModel 先以 BERT 为核心提取专利研究对象的技术特征，再将技术特征以词向量化形式输入 LSTM 模型进行线性激活，从而预测专利标签完成价值评估。

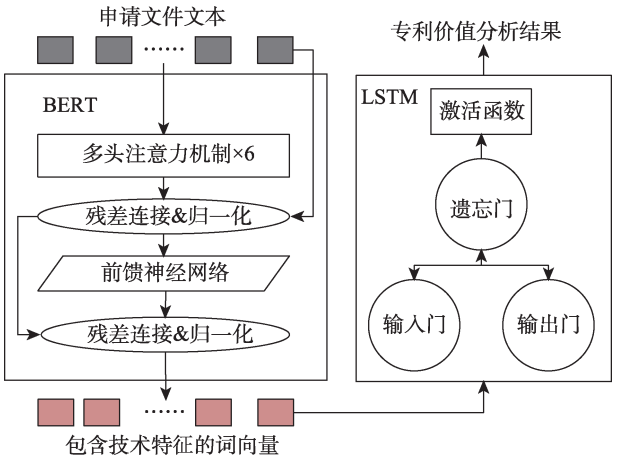

图 5　BLModel 模型结构

如图 6 所示，BERT 包括多头注意力机制、残差连接、前馈神经网络等，作为核心部分的多头注意力机制，由多个 Self-Attention Layer 组成，Self-Attention Layer 的层数需依照研究复杂度确定。需要

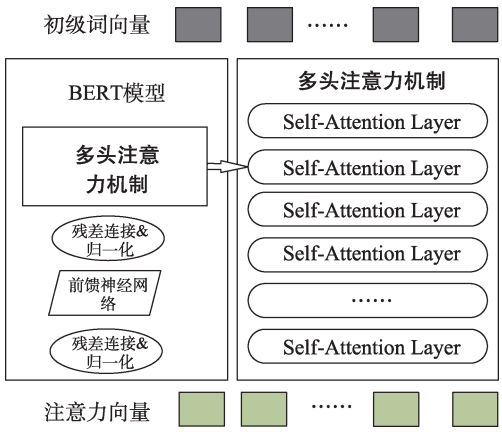

图 6　BERT 结构

说明的是，每个 Self-Attention Layer 对拼接对象执行自注意力运算均是并行处理，且不同的 Self-Attention Layer 之间互不影响[36]。因此，根据 Self-Attention Layer 的层数会生成对应数量的注意力向量。如图 6 所示，若有 6 层 Self-Attention Layer，则 1 组初级词向量经 6 层 Self-Attention Layer 的多头注意力机制计算，会生成 6 组注意力向量。

除了具有并行计算、互不干扰的优点之外，Self-Attention Layer 还能有效捕捉拼接对象中各专利词语的顺序关系，并筛选出重要的技术特征信息，继而聚焦此类技术特征信息，为后续提高价值专利识别准确率提供前置基础。

图 7 显示了基于 Self-Attention Layer 的自注意力运算，以拼接对象 $z^1,z^2,\cdots,z^i,\cdots,z^n$ 中的 z^1 为例，

$$a^1 = w^i z^1 \tag{1}$$

求得 a^1 后，利用 a^1 初始化出问题矩阵 $\boldsymbol{q}^1$、键值矩阵 $\boldsymbol{k}^1$ 及搭配矩阵 $\boldsymbol{v}^1$，即

$$\boldsymbol{q}^1 = w^q a^1 \tag{2}$$

$$\boldsymbol{k}^1 = w^k a^1 \tag{3}$$

$$\boldsymbol{v}^1 = w^v a^1 \tag{4}$$

需要说明的是，a^1 对应 $\boldsymbol{q}^1,\boldsymbol{k}^1,\boldsymbol{v}^1$，$a^2$ 对应 $\boldsymbol{q}^2,\boldsymbol{k}^2,\boldsymbol{v}^2$，$a^i$ 对应 $\boldsymbol{q}^i,\boldsymbol{k}^i,\boldsymbol{v}^i$。以问题矩阵 $\boldsymbol{q}^1$ 为例，分别计算其与所有键值矩阵 $\boldsymbol{k}^1,\boldsymbol{k}^2,\cdots,\boldsymbol{k}^i,\cdots,\boldsymbol{k}^n$ 的乘积，即

$$\alpha_{11} = \frac{\boldsymbol{q}^1 \times \boldsymbol{k}^1}{\sqrt{d}} \tag{5}$$

$$\alpha_{12} = \frac{\boldsymbol{q}^1 \times \boldsymbol{k}^2}{\sqrt{d}} \tag{6}$$

$$\cdots$$

$$\alpha_{1i} = \frac{\boldsymbol{q}^1 \times \boldsymbol{k}^i}{\sqrt{d}} \tag{7}$$

其中，d 为问题矩阵与键值矩阵的矩阵维度。利用 softmax 函数激活每个 α_{1i}，得到 $\hat{\alpha}_{11},\hat{\alpha}_{12},\cdots,\hat{\alpha}_{1i}$，然后，

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

将每个 $\hat{\alpha}_{11},\hat{\alpha}_{12},\cdots,\hat{\alpha}_{1i}$ 与搭配矩阵 $\boldsymbol{v}^1,\boldsymbol{v}^2,\cdots,\boldsymbol{v}^i,\cdots,\boldsymbol{v}^n$ 相乘得到第一个注意力向量 $\boldsymbol{b}^1$，即

$$\boldsymbol{b}^1 = \sum_{i=1}^{n} \hat{\alpha}_{1i}\boldsymbol{v}^i \tag{8}$$

根据上文可知，注意力向量 $\boldsymbol{b}^1$ 是以问题矩阵 $\boldsymbol{q}^1$ 与每个键值矩阵相乘为基础，通过 softmax 函数激活并与搭配矩阵相乘求和得到的矩阵向量。以此类推，可依次计算得到与 $\boldsymbol{q}^2$ 对应的注意力向量 $\boldsymbol{b}^2$、与 $\boldsymbol{q}^i$ 对应的注意力向量 $\boldsymbol{b}^i$，从而汇总得到专利词语集的注意力向量集 $\boldsymbol{b}^1,\boldsymbol{b}^2,\cdots,\boldsymbol{b}^i,\cdots,\boldsymbol{b}^n$。

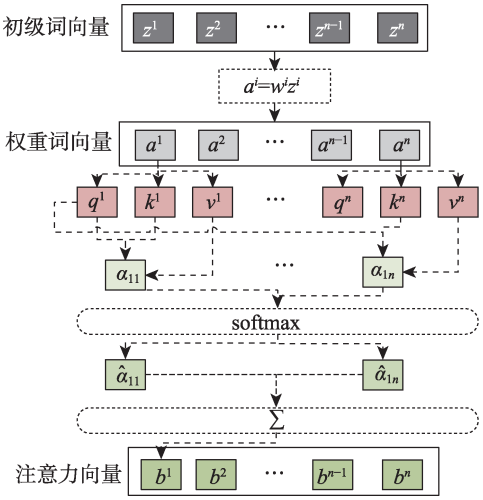

图 7 基于 Self-Attention Layer 的自注意力运算

结合图 6 可知，Multi-Head Attention 包括多层 Self-Attention Layer，每个 Self-Attention Layer 均生成注意力向量集 $\boldsymbol{b}^1,\boldsymbol{b}^2,\cdots,\boldsymbol{b}^i,\cdots,\boldsymbol{b}^n$。当 Multi-Head Attention 具有 6 层 Self-Attention Layer 时，对应生成 6 组注意力向量集 $\boldsymbol{b}^1,\boldsymbol{b}^2,\cdots,\boldsymbol{b}^i,\cdots,\boldsymbol{b}^n$。

结合图 5，当得到多组注意力向量集后，依次将每组注意力向量集与拼接对象执行残差连接及归一化后输入前馈神经网络执行映射计算，并将映射值再次执行残差连接及归一化操作，得到标准词向量组。

最后，本文将经过词向量化所得到的标准向量组输入 LSTM 执行线性激活，得到对应的价值专利识别结果。LSTM 属于改进的循环神经网络，其独特的记忆及遗忘技巧，使 LSTM 具有识别文本特征并自动记忆、可充分挖掘特征关联信息的能力[28]，因此，本文将其作为衔接在 BERT 后的专利价值识别模型。

LSTM 由若干个单元组成，每个单元均与其他单元前后连接。图 8 以其中一个单元为例，展示输入门、输出门及遗忘门之间的数据交互过程。为方便解释各门控的运行机制，假设偏置均为零的前提下接收专利词语 α，在完成专利词语激活处理后，当前单元会根据前连接单元的输出信息，分别计算输入门与输出门在当前单元的存储信息。一般情况下，输入门与输出门均采用 Sigmoid 函数，因 Sigmoid 函数值域为[0,1]，故可将存储信息直接映射为 0 或 1。

此外，遗忘门中存储与前连接单元相关的记忆信息 β。在当前单元中，利用遗忘门计算是否选择遗忘前连接单元所存储的记忆信息 β：若选择遗忘，则当前单元的记忆信息更新为 $f(\alpha)t(\gamma_1)$；若选择不遗忘，则当前单元记忆信息为 $\beta t(\gamma_2)+f(\alpha)t(\gamma_1)$。三大门控单元之间相互传导共享信息，使 LSTM 具有更强的技术特征识别及自动记忆特征的能力，从而有效提高专利价值识别准确率。

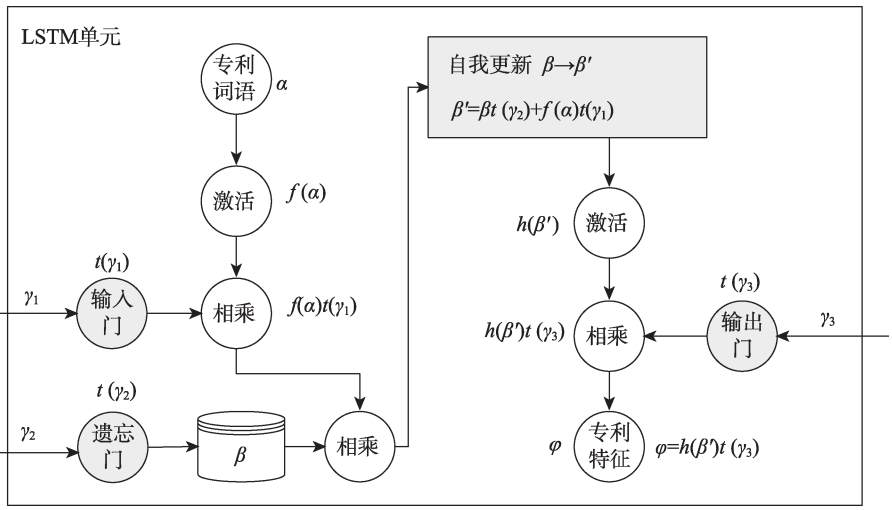

图 8 LSTM 结构

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

综合来说，BLModel 将 BERT 对专利技术特征强大的提取能力与 LSTM 可防止技术特征消散的优点巧妙结合，提高了专利价值评估的准确率。

2　实验验证

2.1　实验设计

本文所述专利价值评估是对特征、标签及模型的全面改进。首先，拼接著录事项书中包括摘要、申请人等文本实现特征重组；其次，提取著录事项书各专利法律状态，并分类得到价值标签；最后，利用 BLModel 从拼接文本中提取技术特征，从而预测专利价值。

专利价值评估过程如图 9 所示，整个评估过程可分为五个步骤，具体如下：

Step1. 获取并拆分数据。从佰腾、智慧芽等专利数据库中，按照申请号提取每件专利所对应的著录事项书，其中著录事项书应包括每件专利的发明名称、申请号、申请（专利权）人、摘要、法律状态等信息。此外，因“特征拼接”与“标签迁移”对应不同类型数据，因此需将所获取的著录事项按照数据类型，对应拆分为包括申请（专利权）人、摘要及发明名称等信息的待拼接文本，并记录专利从申请、审查乃至诉讼阶段的法律状态文本。

Step2. 拼接、预处理并拆分文本。由于待拼接文本所包括的摘要、发明名称等彼此离散，未满足本文对数据的使用要求，因此，首先以“发明名称+申请（专利权）人+摘要”等拼接方法拼接得到整体文本；其次，为提高专利价值评估准确率，去除包括停用词、高频词、标点符号等非正常词；最后，将文本拆分为训练集与测试集，分别用于优化 BLModel 参数及测试。

Step3. 标签迁移。根据图 4，首先将标签制定视角从引用量关注度等迁移至专利法律，以专利所在法律阶段提取符合实验的法律状态；其次，以是否授权作为法律状态的分界点，将已授权（或已驳回）后的法律状态量化为专利价值数值，并映射价值数值得到价值标签。

Step4. 训练模型。需要强调的是，因 BLModel 是基于 LSTM、BERT 等深度学习模型构建的，根据 Patil 等[45]、Zaki 等[46]、Wang 等[47]、Moon 等[48]研究结论，深度学习模型的表现力很大程度取决于模型训练，相比于其他深度学习模型，LSTM、BERT 内部参数更庞大复杂。以谷歌实验室所提供的中文基础 BERT 来说，其模型参数已达 300 兆[47]，为保证 BLModel 的价值评估准确率，本文根据图 9 中的⑦模型训练及⑧模型测试构建出训练验证图（图 10）。

根据图 10 可知，本文先利用 BLModel 预测出与专利训练集对应的预测标签，然后计算预测标签与价值标签的误差值，并通过误差值启动优化函数不断调节 BLModel 的模型参数，直至误差值最终小于预设阈值，表示以特征拼接所构建的专利训练集及标签细化所得到的价值标签可用于 BLModel 的训练，且训练评估准确率满足要求。

Step5. 测试评估。Step4 表示在训练阶段可符合专利

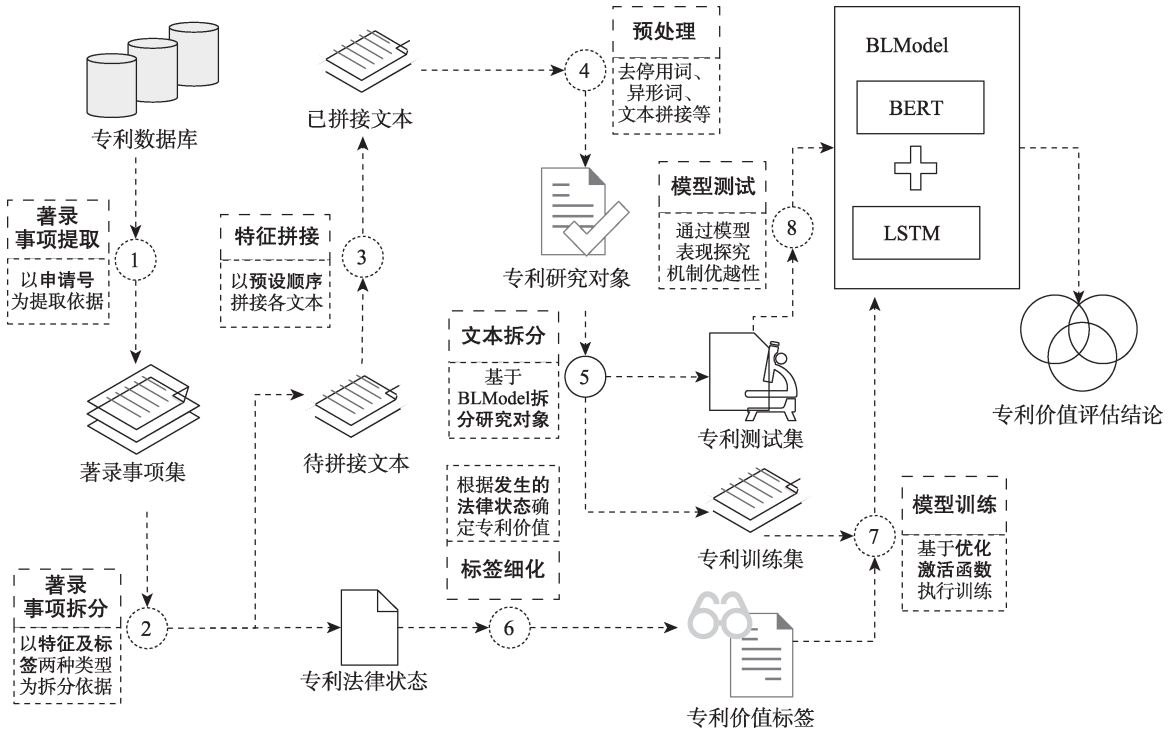

图 9　专利价值评估过程

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

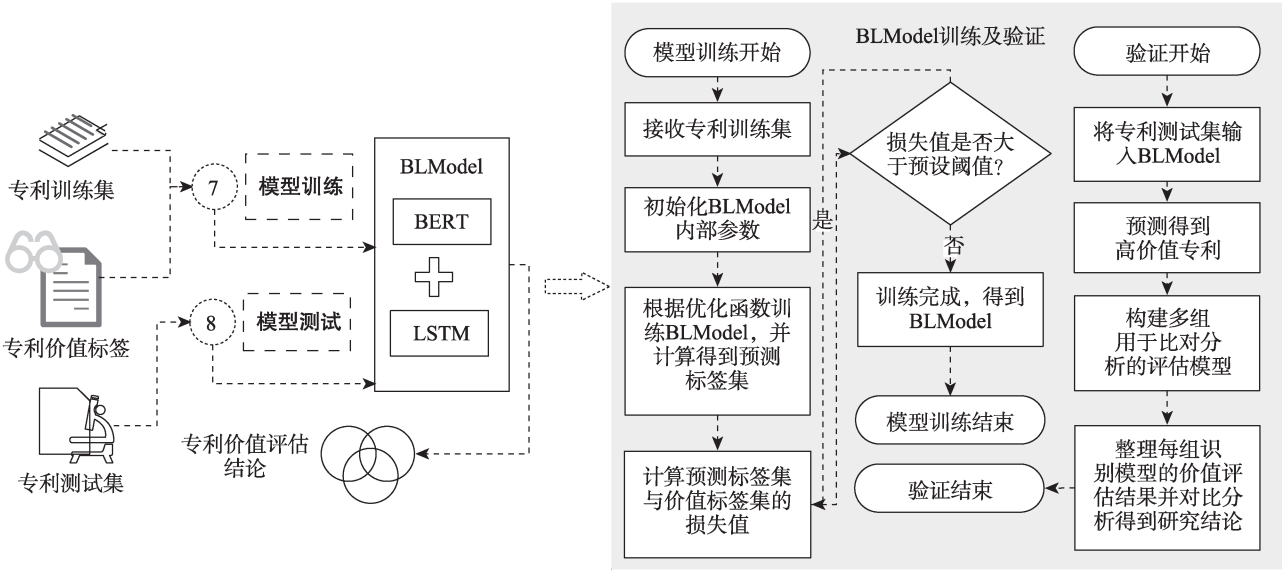

图 10 模型训练及验证

价值评估要求，为继续测试模型的鲁棒性，需进一步利用专利测试集测试 BLModel 表现，如图 10 所示，为体现测试对比效果，还同时引入多组其他专利价值评估模型用于对比分析，从而完成模型测试。

2.2 实验数据构建

由于衡量本文所提出的专利价值评估的贡献需以专利训练集、测试集与价值标签为数据基础，因此，本文先利用“特征拼接”得到训练与测试集，再结合“标签迁移”生成价值标签。

由表 2 和 Step1 可知，首先，从佰腾、智慧芽等专利数据库中采集著录事项信息；其次，将著录事项拆分为待拼接文本及法律状态文本，分别用于特征拼接及标签迁移；最后，利用待拼接文本构建得到训练集、测试集，并将法律状态文本映射成价值标签。

其中，待拼接文本包括发明名称、申请（专利权）人及摘要，由上述特征拼接原理可知，按照发明名称+申请（专利权）人+摘要的拼接顺序组合可得到实验数据，将实验数据按照预设比例拆分为训练集及测试集。此外，由于训练测试集内标点符号、停用词及高频词会直接影响基于深度学习构建的 BLModel 的评估表现[18,48]，因此，需清洗文本以提高评估质量。本文的清洗操作主要包括分词、去标点、去停用词、异形词、高频词、语段重组等。以表 2 中的申请号 CN201010104157.0 为例，去除待拼接文本中“一种、的、处于”等停用词、“方法、用户、设备”等高频词及标点符号后，即可得到供后续研究使用的训练测试集。

需要强调的是，一方面，科学技术迭代更新的速度快、周期短，年代久远的技术对现今技术发展的贡献价值相对较小，因此，本文仅采集申请号为 2010 年后的专利作为实验数据，以避免因技术产生年限过长而降低专利价值的问题[49]；另一方面，我国专利申请量巨大，难以以全国范围为采集单位，国家知识产权局发布的数据[50]表明，广东省每年的专利申请量占全国的 15% 以上，具有突出代表性，因此，本文以 2010 年以后的广东省专利申请作为研究对象，经数据下载、剔除、清洗、对齐、筛选等操作后，共得到 257360 份专利。为探究这 257360 份专利是否符合实验数据要求，防止专利因技术重复性过高而影响专利价值评估，本文按照申请时间将 257360 份专利分为 10 组，其中第一组申请时间为 2010—2011 年，第二组为 2011—2012 年，以此类推。

图 11 展示了 2010 年、2015 年及 2020 年广东省专利核心技术关键字的词频。可以发现，2010 年，因发展相对滞后，该阶段技术主要以“PCB、控制器、基站、组件、服务器”等传统制造业为核心；2015 年，伴随如“报文、终端、细胞、活性”等技术关键词的高频显现，专利热点已逐步渗透到信息安全、生物医药、网页开发等行业；2020 年，人工智能、半导体等热门技术呈现“井喷式”发展，与之相关的“算法、面板、稳定性、验证器”等高频技术术语成为主角。由此可见，在 2010 年至 2020 年这 10 年间，广东省专利的技术方向逐渐从制造业迁移至大数据、人工智能领域，与我国科技发展主流一致。因此，本文利用上述 257360 份专利构建

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2　部分待拼接文本与法律状态信息

编号	申请号	待拼接文本			法律状态
		发明名称	申请(专利权)人	摘要	
1	CN202011298429.5	一种交互式控制方法、装置及设备	深圳市房多多网络科技有限公司	本发明实施例涉及网络远程控制技术领域，公开了一种交互式控制方法、装置及设备，该方法包括：通过HTTP/HTTPS连接接收第一客户端发起的交互请求；根据交互请求获取第一客户端的登录账号和操作信息；若所述登录账号同时登录第二客户端，则将所述第二客户端作为目标客户端；生成包含所述操作信息的远程交互消息，通过Web Socket连接向所述目标客户端发送所述远程交互消息，以使所述目标客户端根据所述远程交互消息进行操作。通过上述方式，本发明实施例能够大大降低跨多个客户端操作的操作用时，提高工作效率，提升用户体验。	驳回
2	CN201110089122.9	电动独轮自行车	林鸿贵	本发明提出了一种站立式的电动独轮自行车，该电动独轮自行车主要是改变了以前业内提出的独轮自行车的结构，在踏脚板的上方设置了一个与车体相连的靠腿板，它的作用是使操作者的腿能够靠住车体，除了踏脚板上的蹬足处以外，增加了一个基本固定的连接点，使操作者的每条腿与车体之间有了两个连接点后，操作者就可站稳在踏脚板上，稳定地操控车子了。此外，本发明还增加了一些友好的人机界面的设计：设计了可折叠式的踏脚板；轮子外加设了轮子护罩，以免衣物卷入轮中；轮子上方加设了为方便携带与搬运用的提手。本发明所提出的电动独轮自行车，稳定性较好，易于操纵，也使初学者易于掌握其驾驶技术。可以成为一种个人代步的交通工具。	授权|诉讼|无效程序|权利转移|海关备案|口头审理
3	CN202011330654.2	基于云计算的分布式软件开发编译方法及软件开发平台	深圳君南信息系统有限公司	本申请实施例提供一种基于云计算的分布式软件开发编译方法及软件开发平台，通过编译依赖关系匹配软件功能分区的目标源编译特征与软件功能分区的分布式编译描述信息，以综合软件功能分区的分布式编译软件包信息与软件功能分区的分布式编译描述信息，提取到丰富的软件功能分区的编译依赖关系，从而便于在编译过程中基于编译依赖的关系进行高效地分布式编译；另外，通过软件功能分区的目标源编译特征，对软件功能分区进行源码文件编译，得到软件功能分区的当前编译文件，从而便于模块化单元编译，减少编译失败后的编译重启时间。	授权|权利转移
4	CN201010104157.0	组件显示处理方法和用户设备	华为终端有限公司	本发明实施例提供一种组件显示处理方法和用户设备。方法包括：获取组件处于待处理状态的指示消息；根据指示消息对容器中显示在屏幕上的显示区域进行缩小处理，以使屏幕在所述显示区域缩小后空余出的区域显示所述容器的隐藏区域，容器包括容纳组件的显示区域和隐藏区域。用户设备包括获取模块，用于获取组件处于待处理状态的指示消息；处理模块，用于根据指示消息对容器中显示在屏幕上的显示区域进行缩小处理，以使屏幕在显示区域缩小后空余出的区域显示所述容器的隐藏区域，容器包括容纳组件的显示区域和隐藏区域。本发明实施例在用户需要对组件进行处理时，可以提示用户可以对显示区域和隐藏区域中的组件进行处理。	授权|诉讼|无效程序|权利转移|海关备案|口头审理

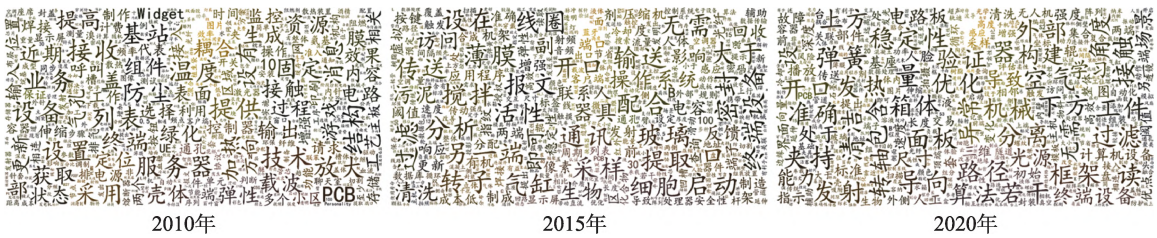

图 11　2010—2011 年广东省专利技术核心关键字

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

训练集与测试集，符合对专利价值评估的实验数据要求。

进一步地，价值标签可区分专利的技术价值。由图4可知，因专利包括递交、初审、公开、实审等多个复杂严谨的法律流程，其所记载技术涉及知识产权从业人员、技术研发团队及公众等多方监督，故在流程严谨性及人员专业性的作用下所产生的法律标签可有效代表专利价值，达到精细化区分专利价值的目的。

专利具有法律保护资格的核心依据是该专利通过形式审查与实质审查后被授予专利权。被授权专利在法律定义上表示所记载的技术方案具备新创性，由此任何专利的价值性研究基础也来源于其授权状态的判断。相反地，若该专利未达到授权标准而被驳回，则表明该专利所揭示的技术方案已被其他企业或个人预先申请取得专利权，或其所揭示的技术方案仅为简单排列组合，未达到远超预期的技术效果。因此，驳回专利不具备法律保护资格，更无法从法律角度承认其价值性。综合来说，是否授权是从法律角度判断专利价值最基础且核心的判断依据，因此，本文将授权专利价值赋予+1值，将驳回专利价值赋予−1值。

此外，如一案双申、放弃专利权、未缴年费、海关备案等其他专利状态，本文从利益角度衡量其专利价值性，对于可提高经济利润可能性的法律状态设置为+0.5值；反之，对于降低经济利润可能性的法律状态设置为−0.5值。例如，若申请人取得专利权后又提出放弃专利权，原则上是因为申请人主观判断所取得专利权的专利的可实施性较低，无法创造满足预期的经济收益，相比于每年依然需缴纳专利年费的开支来说，主动放弃专利权可及时缩减专利权人的财务开支。

本文依次遍历257360份专利的每个法律状态，结合图4汇总得到各法律状态的价值性，如表3所示。

从表3可见，法律状态可分为申请前、申请后

表3 法律状态及价值性

编号	法律状态名称	法律状态释义	是否人为主观干预而生成的法律状态	所属法律阶段	法律状态价值
1	一案双申	基于同一件专利申请同时提出发明专利申请与实用新型专利申请	是	递交国知局审查	若授权，+0.5
2	撤回	专利申请人在提出申请后授予专利权前，将专利申请收回的手段	是	申请后授权(驳回)前	0
3	公开	发明专利在申请日后、授权前的一定时间向公众公开的手段	否	申请后授权(驳回)前	0
4	实质审查	国家专利局对申请专利新颖性、创造性、实用性等实质性内容所作的审查	否	申请后授权(驳回)前	0
5	复审	专利申请被驳回时给予申请人的一条救济途径	是	驳回后	+0.5
6	放弃	专利权人放弃专利权的直接处分权利	是	授权后	−0.5
7	无效程序	经第三人申请、国家知识产权局专利局复审和无效审理部审查，对已授予的专利作出无效的决定	是	授权后	+0.5
8	未缴年费	专利权人自被授予专利权的当年开始，在专利权有效期内未逐年向专利局缴纳费用的结果	是	授权后	−0.5
9	海关备案	指专利权人为寻求海关对其专利权实施保护而制定的一种备案手段	是	授权后	+0.5
10	许可	专指专利技术所有人或其授权人许可他人在一定期限、一定地区、以一定方式实施其所拥有的专利，并向他人收取使用费用的规定	是	授权后	+0.5
11	诉讼	专利权人、第三人等对专利机关有关专利权的决定不服，而向法院提起诉讼的一种解决专利权纠纷的方式	是	授权后	+0.5
12	部分无效	专利权被宣告无效的法律后果是被宣告无效的专利权视为自始即不存在，部分被宣告无效的，则不影响其他部分的效力	是	授权后	+0.5
13	驳回	对不符合专利法规定的专利申请予以驳回不授予专利权的一种决定	是	授权(驳回)中	−1.0
14	授权	授予法律保护专利权的发明和实用新型，应当具备新颖性、创造性和实用性	是	授权(驳回)中	+1.0

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

授权（驳回）前、授权（驳回）中和授权（驳回）后四个阶段，但由于申请前、申请后授权（驳回）前阶段并未完成专利实质审查，无法确定专利相比于其他专利是否具备新颖性与创造性，无法衡量专利价值，因此，本文剔除仅处于申请前、申请后授权（驳回）前阶段的所有专利，以是否授权作为衡量专利价值的分界点，将驳回专利价值设置为−1，“授权”专利价值设置为+1，在授权（驳回）中阶段前的法律状态均置 0，在授权（驳回）中后的法律状态均以驳回或授权两种法律状态为研究基础，按照释义分为包括许可、诉讼等的积极状态以及放弃、未缴年费的消极状态，上下浮动 0.5。

此外，每个法律状态可按照是否人为干预划分为可避免及不可避免，其中实质审查、公开两种法律状态是每件发明专利在授权前均需执行的不可避免流程，不以人为干预而改变，因此具有专利普适性，无法通过实质审查|公开区分出专利价值，但可利用其识别专利所处的法律阶段，从而达到筛选实验数据的目的。以申请号 CN202010020281.2、名称为“一种用于太阳能电池板加工的生产线”的发明专利为例，其法律状态为“实质审查|公开|一案双申”，而实质审查|公开属于申请后授权（驳回）前法律阶段，并未涉及实质审查，因此，不满足专利数据需进入授权（驳回）中或授权（驳回）后法律阶段的实验数据要求。对比之下，申请号 CN201510003631.3、名称为“高真空度旋转密封机构”的发明专利，法律状态为“授权|一案双申”，显而易见，其处于授权（驳回）中阶段，满足实验数据要求，故通过表 3 各专利价值求和得到总价值=1.0（授权）+0.5（一案双申）。

为方便后续 BLModel 训练与测试，本文将专利总价值映射为不同价值等级。为防止因专利状态过多出现总价值过高而导致专利价值等级溢出的现象，设定“无效、复审”“放弃、未缴年费”“许可、海关备案、一案双审”及“诉讼、部分无效”同时出现时，其价值不叠加，依然只取 0.5。

专利价值标签示例如表 4 所示，以申请号 CN201310001270.X、发明名称“冲牙器”为例，法律状态为“授权|诉讼|无效|海关备案|一案双申”，但由于“海关备案、一案双申”同时出现时价值依然仅取 0.5，则总价值为 1.0（授权）+0.5（一案双申或海关备案）+0.5（诉讼）+0.5（无效），对应价值标签 A。

综上所述，实验数据包括训练集、测试集与价值标签，是特征拼接及标签迁移的重要体现，也对后续 BLModel 训练产生重要影响。本文首先在规避年限对专利价值影响及全国范围采样困难的前提下，利用特征拼接得到训练集与测试集，并基于词频随年限的变化过程，验证数据集的可行性。其次，结合专利呈现的流程严谨性及人员专业性的特点，通过标签迁移生成法律状态，并将法律状态量化及映射成价值标签，有效解决传统标签受限于技术热度、从业人员数量等产生专利价值表示不准确的问题。故本节所述实验数据充分体现了特征拼接及标签迁移的优越性，也为 BLModel 后续评估打好了基础。

2.3　实验程序部署

由于深度学习具有计算量大、复杂度高等特点，难以通过普通程序实现，因此，本文结合 Pytorch 部署实验程序。Pytorch 是目前最受欢迎的深度学习框架之一，具有灵活性高、运行速度快、代码部署清晰等优点[51-52]，可提供便捷的功能类函数及神经网络实现类，包括 Feedforward Neural Network、Skip-Connect 及归一化函数等，通过继承封装这类功能函数及实现类，可形成端到端的专利价

表 4　专利价值标签示例

编号	申请号	发明名称	法律状态	总价值	价值区间	价值标签
1	CN201510003417.8	半自动烧录装置及烧录方法	驳回	−1.0	[−1.0,+0.5]	C
2	CN201510014351.2	一种丁腈橡胶/聚氯乙烯复合材料及其制备方法	授权|未缴年费	−0.5		
3	CN201710000143.6	一种自动部署操作系统的方法	授权	+1.0	[+1.0,+1.5]	B
4	CN201310012680.4	界面物的移动方法及支持界面物移动的装置	授权|海关备案	+1.5		
5	CN201510010409.6	一种饼干爬坡理料运输机构	授权|无效|一案双申	+2.0	[+2.0,+4.0]	A
6	CN201310001270.X	冲牙器	授权|诉讼|无效|海关备案|一案双申	+2.5		

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

值评估，弱化中间过程，提升专利价值评估流畅性[46]。

实验部署主要包括以“特征拼接及标签迁移”为主的文本处理、BLModel训练及BLModel调用三个部分。根据图12可知，首先，文本处理程序作为专利文本的入口程序，包括jieba、nltk等文本预处理包，包中可调用split、cut、stopswords等函数实现分词、去停用词等操作；其次，训练程序主要基于Pytorch的torch包，通过继承torch包下nn.Module类，并引入LSTM、BERT Tokenizer、BERT_pretrain等模型，构建损失函数及优化函数引导参数迭代更新，直至达到训练要求，保存优化后的参数集得到BLModel；最后，BLModel及衔接程序预测专利测试集中各专利价值等级，完成价值评估。

总的来说，部署阶段需先基于Python语言调用多种类型的数据分析包，以完成数据清理及维度转变等处理步骤，然后结合LSTM、BERT等原理继承实现Pytorch中预定义的nn.Module类，从而构建得到专利价值评估程序。

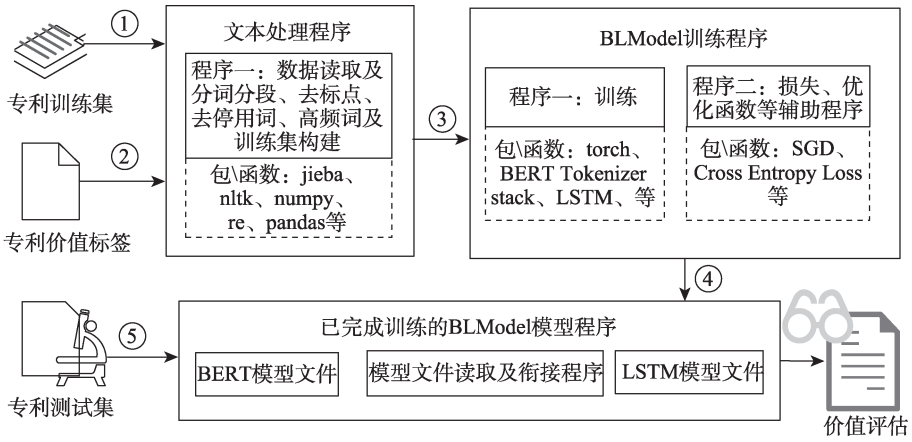

图12 实验部署架构

2.4 模型训练

在上述专利价值评估程序成功构建的基础上，本文首先从佰腾、智慧芽等专利数据库提取每件专利的著录事项书，其中著录事项书包括专利的发明名称、申请号、申请（专利权）人、摘要、法律状态等信息；其次，根据特征拼接原则，拼接包括摘要、发明名称及申请（专利权）人等离散文本，得到用于执行模型训练及测试的训练文本共257360份；再其次，根据标签迁移原则，从著录事项书中提取专利所经历的法律状态，并以是否授权作为法律状态的分界点，将已授权（或已驳回）后的法律状态量化为专利价值数值，并映射专利价值数值为A、B或C三种专利价值标签，实现专利价值量化；最后，将包括法律价值标签的训练文本输入BLModel，利用BLModel的多头注意力机制和各类门控单元实现对训练集的专利价值预测，得到预测价值标签，进而根据交叉熵损失函数计算预测价值标签与法律价值标签的交叉熵损失值。其中，可利用交叉熵损失值确定模型的专利价值评估准确率。

具体的训练过程：首先，将257360份包括法律价值标签的拼接文本分为训练集及测试集。其中，训练集205888份用于调整优化BLModel模型参数，测试集51472份用于后续模型比对实验，验证实际表现。其次，设置训练周期epochs=3，训练批次数batch_size=16，学习率learning_rate=1e-3，并构建当第1000次训练loss值若无提升则退出的训练条件。最后，调用cross_entropy函数计算loss值，设置Adam为BLModel参数优化器等后启动训练。需要强调的是，因BERT模型结构复杂参数庞大，难以利用普通算力设备支持其完整训练，故本文先从Hugging Face社区（https://huggingface.co/）下载预训练完成的BERT模型，然后采用fine-tune方法微调BLModel参数，最终得到BLModel训练中的loss值变化，如图13所示。

图13将每个step所对应的损失值均显示至坐标轴中。在[0,40]区间段，随step增加，训练损失值逐渐收敛至0.1以下，当step达到40后，训练损失值小范围波动，表明BLModel参数已收敛并至训练结束，显然采用fine-tune微调BLModel的训练方法会加速参数寻优速度，快速完成BLModel的模型训练。

2.5 测试及对比

为测试专利价值评估方法的实际表现，本文整

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

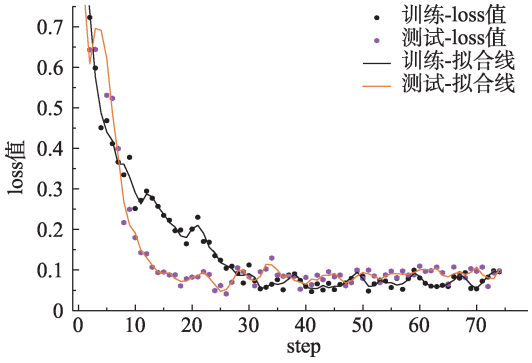

图 13　BLModel 训练中的 loss 值变化

理出 2018—2020 年广州、深圳、佛山及东莞处于实质审查阶段的专利申请 8670 份，利用已经过特征拼接和标签迁移训练完成的 BLModel，依次对每份专利进行价值分类，得到的部分结果如表 5 所示。

由表 5 可见，不同专利因所记录技术的差异性，均被智能化地识别出不同的价值等级。进一步构建对比实验，探究本文方法的优势。一方面，根据 Chung 等[21]和 Zhu 等[22]的研究结论，引入表现优异的卷积神经网络、XGBoost 等模型；另一方面，为探究 BLModel 的组合设计是否具有优势，拆分 BLModel 并尾接激活函数，得到 LSTM-Linear 和 BERT-Linear 两种对比模型，经实验得到的对比结果如表 6 所示。

表 6 共展示了 7 组模型，为能统一对比各模型训练准确率和测试准确率的差异，均使用本文根据法律状态而量化的价值标签（见表 4）。编号 1~2 组为机器学习模型，其研究对象以发明人数量、权利要求数等专利指标为主。编号 3~7 组为深度学习模型，研究对象以文本为主，为体现特征拼接的优点，编号 3~4 组模型仅使用摘要作为研究对象，编号 5~7 组使用拼接文本作为研究对象。由此可见，以法律标签衡量专利价值为基础，通过拼接文本研究专利价值，发现组合模型 BLModel 在训练准确率和测试准确率上均优异于其他模型。图 14 进一步直观显示了各模型表现。

结合图 14 和表 6 分析，以研究对象来说，不同研究对象适用不同价值分析模型，如文本适用深度神经网络，指标数据适用机器学习模型。对于专利价值分析的实际表现，需结合具体模型结构具体分析，如采用指标数据的 XGBoost 虽为传统机器学习，但合理构建树深度等参数，其表现反而优异于

表 5　部分实质审查专利的价值分类

编号	申请号	发明名称	申请(专利权)人	价值等级
1	CN202080008700.9	摄像模组及计算机设备	闻泰科技(深圳)有限公司	B
2	CN202011369480.0	一种具有翻转功能的连续弯管机	佛山市天箭机械设备有限公司	C
3	CN202011378785.8	燃烧器及厨房器具	广东美的厨房电器制造有限公司	B
4	CN202011380501.9	边缘网关和综合能源系统	广东电网有限责任公司	C
5	CN202011384852.7	六氟化硫击穿电压试验装置	广东电网有限责任公司东莞供电局	C
6	CN201810747197.3	探测参考信号传输	华为技术有限公司	A
7	CN201810144866.8	不对称网络地址封装	华为技术有限公司	A
8	CN201910658351.4	一种音源设备及其控制方法、耳机	广州市智专信息科技有限公司	B
9	CN202011379037.1	一种装配式无人值守智慧环控机房	深圳市豪斯特力科技有限公司	C
10	CN202011378902.0	信息推送系统和信息推送方法	深圳市博云慧科技有限公司	C
11	CN202011369868.0	通信系统和方法	南方电网数字电网研究院有限公司	C

表 6　基于螺旋机制的对比结果

编号	模型名称	研究对象	研究对象示例	价值标签	训练准确率(%)	测试准确率(%)
1	SVM	指标数据	权利要求数、技术领域等	法律状态量化标签	69.2	58.7
2	XGBoost				75.3	65.2
3	CNN-Linear	单一文本	摘要或发明名称或权利要求		65.7	54.1
4	CNN-LSTM				77.8	67.2
5	LSTM-Linear	拼接文本	摘要+申请人+发明名称		72.2	65.8
6	BERT-Linear				75.1	67.3
7	BLModel				79.3	69.7

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

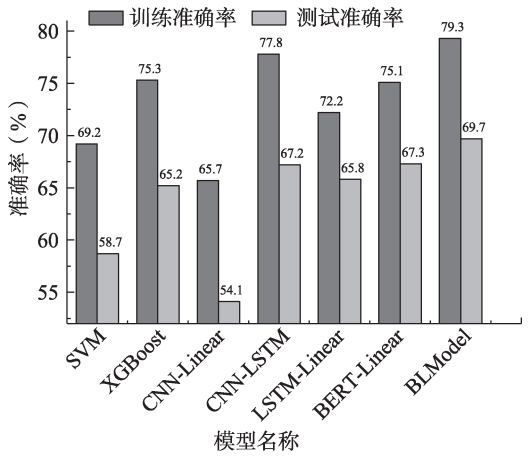

图 14 各模型训练及测试准确率

卷积神经网络 CNN-Linear。可以肯定的是，将拼接文本作为研究对象可提高专利价值分析准确率，拼接手段组合不同的单一文本可重构出技术特征显现更强的研究对象，从而解决指标数据或单一文本未足够体现技术本质而引发评估准确率不高的问题。

以模型结构来说，不同模型结构会直接影响专利分析准确率，如采用同一组指标特征的 SVM 和 XGBoost，在训练和测试阶段均具有较大差别；但组合模型的分析表现明显优异于单一模型，如拆分 BLModel 所得到的 LSTM-Linear 及 BERT-Linear，由于对拼接文本的技术特征提取能力不足，导致价值专利识别精细度明显低于 BLModel 和 CNN-LSTM。此外，BERT 模型设计更成熟，预训练时内部参数寻优比 CNN 更加准确，从而产生 BLModel 表现优异于 CNN-LSTM 的现象。

3 结 论

专利价值评估对我国遏制异常申请、净化市场环境具有重要现实意义。本文以特征拼接、标签迁移及深度学习组合为中心构建专利价值评估方法，并基于 2010—2020 年广东省专利探究其在专利价值评估的实际表现，最后引入多组对比模型进行实验分析，得到如下主要贡献。

（1）拼接著录事项信息，以构建技术特征显现更强的研究对象。著录事项书记载自申请、审查、授权及权利丧失等阶段下专利技术的全生命周期轨迹，是技术价值高低的重要记录载体，故本文将研究对象从授权量、申请量、发明占比等指标数据转移至著录事项书，结合知识产权法律知识，拼接包括摘要、发明名称及申请（专利权）人等著录事项的信息，得到技术特征显现更强的研究对象，从而解决指标类研究对象未足够体现专利的技术本质而引发后续评估准确率不高的问题。

（2）迁移至专利法律视角，以量化出更具专利价值代表性的法律类价值标签体系。为解决因引用数或下载访问量等价值标签不匹配实际专利价值的问题，本文巧妙地以法律角度寻找出与专利技术严格对应的专业人员和法律流程，并通过著录事项书确定专业人员处理法律流程所得出的法律状态，进一步以法律状态映射出价值标签体系，在延展专利标签体系研究深度的同时，解决因引用率、下载访问量等与专利价值不匹配而造成价值评估错误的问题。

（3）组合 BERT 及 LSTM 得到 BLModel，以提升对研究对象的技术特征提取能力。本文以词向量的高精细度为构建原理，组建以 BERT 为核心的 LSTM 专利质量评估模型，通过 BERT 所包括的多头注意力机制和 MASK 方法，可从专利研究对象中精确提取并表示出专利技术特征。与其他评估模型对比发现，BLModel 可有效解决因特征提取能力不足而产生评估准确率偏低的问题，为专利价值评估提供了新工具。

综上所述，本文结合专利法律知识，基于特征拼接方法组装新的专利价值研究对象，以专利法律状态视角构建新的价值标签体系，并组合各深度学习模型得到新的专利价值评估模型，在研究对象有效性、标签体系性及模型构建评估率三个方面均提出优化改进策略，具有较强的实际应用价值。

参 考 文 献

[1] Wang J J, Ye F Y. Probing into the interactions between papers and patents of new CRISPR/CAS9 technology: a citation comparison[J]. Journal of Informetrics, 2021, 15(4): 101189.

[2] Schwartz H M. Global secular stagnation and the rise of intellectual property monopoly[J]. Review of International Political Economy, 2022, 29(5): 1448-1476.

[3] 王格格, 刘树林. 国际专利分类号间的知识流动与技术间知识溢出测度——基于中国发明授权专利数据[J]. 情报学报, 2020, 39(11): 1162-1170.

[4] Kuhn J M, Teodorescu M H M. The track one pilot program: who benefits from prioritized patent examination?[J]. Strategic Entrepreneurship Journal, 2021, 15(2): 185-208.

[5] Feng J, Jaravel X. Crafting intellectual property rights: implications for patent assertion entities, litigation, and innovation[J]. American Economic Journal: Applied Economics, 2020, 12(1):

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

140-181.

[6] 刘夏, 黄灿, 余骁锋. 基于机器学习模型的专利质量预测初探[J]. 情报学报, 2019, 38(4): 402-410.

[7] Ljungberg D, Bourelos E, McKelvey M. Academic inventors, technological profiles and patent value: an analysis of academic patents owned by swedish-based firms[J]. Industry and Innovation, 2013, 20(5): 473-487.

[8] 李睿, 赵峰. 届满专利与无效专利的施引特征对比及其情报学意义[J]. 情报学报, 2016, 35(6): 586-596.

[9] Hou B J, Zhang Y M, Hong J, et al. New knowledge and regional entrepreneurship: the role of intellectual property protection in China[J]. Knowledge Management Research & Practice, 2021: 1-15.

[10] 杨思思, 戴磊, 郝屹. 专利经济价值度通用评估方法研究[J]. 情报学报, 2018, 37(1): 52-60.

[11] Song X Y, Huang X H, Qing T. Intellectual property rights protection and quality upgrading: evidence from China[J]. Economic Modelling, 2021, 103: 105602.

[12] Wu H C, Chen H Y, Lee K Y. Unveiling the core technology structure for companies through patent information[J]. Technological Forecasting and Social Change, 2010, 77(7): 1167-1178.

[13] Lai K K, Chen H C, Chang Y H, et al. A structured MPA approach to explore technological core competence, knowledge flow, and technology development through social network patentometrics[J]. Journal of Knowledge Management, 2021, 25(2): 402-432.

[14] Trappey A J C, Trappey C V, Wu J L, et al. Intelligent compilation of patent summaries using machine learning and natural language processing techniques[J]. Advanced Engineering Informatics, 2020, 43: 101027.

[15] 刘大勇, 孟悄然, 段文斌. 科技成果转化对经济新动能培育的影响机制——基于230个城市专利转化的观测与实证分析[J]. 管理科学学报, 2021, 24(7): 49-65.

[16] Liu W D, Qiao W B, Wang Y, et al. Patent transformation opportunity to realize patent value: discussion about the conditions to be used or exchanged[J]. Information Processing & Management, 2021, 58(4): 102582.

[17] 李治东, 熊焰, 方曦. 基于熵权层次分析法的核心专利识别应用研究[J]. 情报学报, 2016, 35(10): 1101-1109.

[18] Trappey A J C, Trappey C V, Govindarajan U H, et al. Patent value analysis using deep learning models—the case of IoT technology mining for the manufacturing industry[J]. IEEE Transactions on Engineering Management, 2021, 68(5): 1334-1346.

[19] Wang J L, Fan Y, Zhang H, et al. Technology hotspot tracking: topic discovery and evolution of China's blockchain patents based on a dynamic LDA model[J]. Symmetry, 2021, 13(3): 415.

[20] Huang Z X, Xie Z P. A patent keywords extraction method using TextRank model with prior public knowledge[J]. Complex & Intelligent Systems, 2022, 8(1): 1-12.

[21] Chung P, Sohn S Y. Early detection of valuable patents using a deep learning model: case of semiconductor industry[J]. Technological Forecasting and Social Change, 2020, 158: 120146.

[22] Zhu H M, He C H, Fang Y, et al. Patent automatic classification based on symmetric hierarchical convolution neural network[J]. Symmetry, 2020, 12(2): 186.

[23] Wu H Q, Shen G Q, Lin X, et al. A transformer-based deep learning model for recognizing communication-oriented entities from patents of ICT in construction[J]. Automation in Construction, 2021, 125: 103608.

[24] Ni X, Samet A, Cavallucci D. Similarity-based approach for inventive design solutions assistance[J]. Journal of Intelligent Manufacturing, 2022, 33(6): 1681-1698.

[25] 李睿, 周维, 容军凤, 等. 高价值企业专利的被引特征分析——以世界500强企业专利为例[J]. 情报学报, 2015, 34(9): 899-911.

[26] Nagler M, Sorg S. The disciplinary effect of post-grant review-causal evidence from European patent opposition[J]. Research Policy, 2020, 49(3): 103915.

[27] Jeon D, Ahn J M, Kim J, et al. A doc2vec and local outlier factor approach to measuring the novelty of patents[J]. Technological Forecasting and Social Change, 2022, 174: 121294.

[28] 陈亮. 面向专利分析的Patent Classification LDA模型[J]. 情报学报, 2016, 35(8): 864-874.

[29] 专利审查指南[EB/OL]. [2022-07-01]. http://www.cypatent.com/cn/sczn.htm.

[30] 李雨峰. 论专利公开与排他利益的动态平衡[J]. 知识产权, 2019, 29(9): 3-10.

[31] Yu L P, Duan Y L, Fan T T. Innovation performance of new products in China's high-technology industry[J]. International Journal of Production Economics, 2020, 219: 204-215.

[32] Di Gennaro G, Buonanno A, Palmieri F A N. Considerations about learning word2vec[J]. The Journal of Supercomputing, 2021, 77(11): 12320-12335.

[33] 王玲, 李文昌, 赵梦. 不同类型专利权人的专利失效影响因素研究[J]. 科技管理研究, 2021, 41(19): 149-154.

[34] Marco A C, Sarnoff J D, de Grazia C A W. Patent claims and patent scope[J]. Research Policy, 2019, 48(9): 103790.

[35] Mossinghoff G J, Kuo V S. Post-grant review of patents: enhancing the quality of the fuel of interest[J]. Idea, 2003, 43: 83.

[36] Rai A K. Improving (software) patent quality through the administrative process[J]. Houston Law Review, 2013, 51(2): 503-543.

[37] Novelli E. An examination of the antecedents and implications of patent scope[J]. Research Policy, 2015, 44(2): 493-507.

[38] Kim Y K, Park S T. Patent litigation research trends and trend analysis[J]. Journal of Computational and Theoretical Nanoscience, 2021, 18(5): 1485-1489.

[39] U. S. Patent and Trademark Office, U. S. Department of Com-

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

merce. Patent litigation and USPTO trials: implications for patent examination quality[R]. Alexandria: United States Patent and Trademark Office, 2015.

[40] Sun F, Liu J, Wu J, et al. BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2019: 1441-1450.

[41] Wan C X, Li B. Financial causal sentence recognition based on BERT-CNN text classification[J]. The Journal of Supercomputing, 2022, 78(5): 6503-6527.

[42] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2017: 6000-6010.

[43] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[44] Nguyen H D, Tran K P, Thomassey S, et al. Forecasting and anomaly detection approaches using LSTM and LSTM autoencoder techniques with the applications in supply chain management[J]. International Journal of Information Management, 2021, 57: 102282.

[45] Patil A, Viquerat J, Larcher A, et al. Robust deep learning for emulating turbulent viscosities[J]. Physics of Fluids, 2021, 33(10): 105118.

[46] Zaki G, Gudla P R, Lee K, et al. A deep learning pipeline for nucleus segmentation[J]. Cytometry Part A, 2020, 97(12): 1248-1264.

[47] Wang X, Wang K, Lian S G. A survey on face data augmentation for the training of deep neural networks[J]. Neural Computing and Applications, 2020, 32(19): 15503-15531.

[48] Moon T, Son J E. Knowledge transfer for adapting pre-trained deep neural models to predict different greenhouse environments based on a low quantity of data[J]. Computers and Electronics in Agriculture, 2021, 185: 106136.

[49] Veugelers R, Wang J. Scientific novelty and technological impact [J]. Research Policy, 2019, 48(6): 1362-1372.

[50] 国家知识产权局 . 国内专利授权年度状况(2019 年)[R/OL]// 2019 知识产权统计年报 . [2022-07-01]. https://www.cnipa.gov.cn/tjxx/jianbao/year2019/b/b2.html.

[51] Zimmer L, Lindauer M, Hutter F. Auto-pytorch: multi-fidelity MetaLearning for efficient and robust AutoDL[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(9): 3079-3090.

[52] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library[C]// Proceedings of the 32nd Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2020: 7994-8005.

（责任编辑 王克平）

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net