DOI: 10.12120/bjutskxb202305138

融合知识网络嵌入特征的高价值专利预测

任海英, 孙闯闯
(北京工业大学 经济与管理学院, 北京 100124)

摘 要: 准确预测专利价值,尽早识别具有较高价值的专利对促进高价值专利技术的培育和发展,提前进行技术布局具有重要意义。基于知识重组和专利发明创造过程,划分专利价值特征,通过构建样本专利知识网络和领域先前知识网络,选取和计算知识网络嵌入特征来量化专利的新颖性和常规性,并将其与创新主体及专利申请特征加以融合,构建用于专利价值早期预测的指标体系,利用机器学习算法对处于申请早期的专利价值进行预测。以神经网络技术领域的专利进行实证研究。研究结果表明,融合知识网络嵌入特征的高价值专利预测模型 $F1$ 值达到 80%,预测结果具有有效性,并且知识网络嵌入特征尤其是网页排名(PageRank)和特征向量中心性等对预测高价值专利具有重要作用。

关键词: 高价值专利; 价值预测; 领域先前知识; 知识网络嵌入; 机器学习
中图分类号: G306 **文献标志码**: A **文章编号**: 1671-0398(2023)05-0138-15

在科技竞争日益激烈的知识经济时代,我国需要进一步提高创新能力和创新水平以实现科技自立自强[1]。《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》首次将高价值专利纳入经济社会发展的主要目标[2]。高价值专利通常是先进创新技术的核心,是科技高质量发展的关键要素。如果能从海量申请专利中尽早预测出高价值专利,可以为其技术培育和发展提供更多机会,提前瞄准核心技术和产品研发方向并进行重点布局,从而加快产品研发和占领市场前沿进度,对提升我国的技术创新水平具有重要意义。

学术界现有专利价值相关研究主要包括专利价值的界定和评估、影响因素分析以及早期预测。在评估识别方面,通常使用代表性专利指标或综合指标体系对已经产生技术、经济或法律价值的专利进行识别,属于一种事后评价;在影响因素方面,使用多元统计等方法探究专利价值产生机理,即如何受其自身、申请和运营等过程中的各类特征的影响。近年来,这两类研究衍生出对高价值专利的预测,即利用专利申请或授权的早期影响因素来预测其未来能否成为高价值专利。从专利价值的事后评价、影响机理到预测无疑是一种深化和扩展,但仍有许多影响因素和预测模型尚待开发。现有研究较少考虑领域先前知识和专利文本知识对专利价值的影响,这如同专利审查员在评审专利时不读专利文本而只看其外在特征,因此是不够全面的。专利价值具有内生性,并且一项发明是由创新主体在领域先前知识基础上发展和创造出来的,具有一定的新颖性和实用性才会被授权。而这种新颖性和实用性具有相对性,如何在领域先前知识层面表示出专利新颖性和常规性并将其应用到专利价值的早期预测中,是有现实意义的。

收稿日期: 2022-12-23
基金项目: 北京市自然科学基金面上项目(9192003)
作者简介: 任海英(1971—), 男, 北京工业大学经济与管理学院副教授。

本文首先对相关文献进行总结，分析现有研究的局限性，然后提出一种融合知识网络嵌入特征（knowledge network embeddedness）的专利价值早期预测方法，对某一领域的专利文献进行实例验证，试图为专利价值的早期预测提供一种新的思路和方法。

一、文献综述

（一）专利价值的界定和评估

目前，学术界和工业界对专利价值还没有较为统一的定义和界定标准。在学术研究中，学者们通常用专利授权后在维持阶段的指标作为专利价值代理指标开展相关研究，如被引次数[3]、维持年限[4-5]等。得益于专利数据的日益完善，专利拍卖价格[6]、异议涉诉[7]、转移转让[8]等指标也开始逐渐被广泛使用。对于以上结果特征，一般是基于特定的研究目的来研究各结果指标的影响因素或识别高价值专利[9]，满足不同的应用需求和为政策制定提供参考依据。在政策层面上，《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》强调了专利的“高价值”。2019 年 3 月，国家知识产权局将维持年限超过 10 年、获得国家科学技术奖或中国专利奖、实现较高质押融资金额等 5 类发明专利纳入了高价值发明专利统计范畴[10]，这对高价值专利的界定和代理指标的选取具有指导意义。程文银等[11]采用国知局对高价值专利的定义对我国高价值专利的特征、结构、创新主体和产学研合作等方面进行分析。从经济学角度，专利价值可以区分为私人价值和社会价值。例如，专利维持年限较多地反映了专利给所有者的收益，属于私人价值；而专利的被引次数作为对其他后续专利的启发和知识来源，具有更多的社会价值特性。

随着对专利价值研究和理解的深入，学者们从技术、法律、经济等多价值维度构建综合评价指标体系来评估和识别高价值专利。如王子焉等[12]从技术、法律、使用价值和网络平台服务价值构建指标体系，评估网络交易平台上的专利价值。慎金花[13]以及李娟等[14]从技术、经济和法律维度选取指标进行专利价值评估。杨登才等[15]从数量、质量（法律）以及价值（专利转化层面）三个维度设计指标，利用熵权法对高校专利质量进行测度。田雪姣等[16]构建“技术-市场-法律”三维度的核心专利评价指标体系，应用熵权-TOPSIS-德尔菲法方法对芯片制造领域专利数据进行核心技术识别，最终识别出 32 件芯片制造领域的核心技术。在这些综合评价体系的指标中，既有专利结果特征，也有如引用专利数、发明人数等运用先前知识和人才资本的专利指标，还有国际专利分类（International Patent Classification，IPC）数、权利要求数、文献页数等专利申请文本的特征，属于对专利价值的事后评价。

（二）专利价值影响因素

在专利价值评估和识别基础上，一些学者使用统计学习方法研究专利价值影响因素，探究专利价值产生机理。通常将专利结果特征作为专利价值代理变量，从专利创新主体、申请特征、结果特征等选取影响因素。郭状等[17]以专利引用滞后期作为专利价值代理变量，使用 Cox 比例风险回归方法探究专利家族、申请人类型等对专利价值的影响。马荣康等[18]用 Logit 回归模型研究了知识组合的多样性和新颖性对以专利被引次数代表的突破性发明形成的影响。荣雪云等[19]运用负二项模型检验发明者特征（发明者规模和发明者发明经验）、技术新颖性（知识重组新颖性和知识起源新颖性）和发明质量三者之间影响关系，发现知识重组新颖性会正向影响发明质量。冯仁涛[20]使用 Logistic 回归分析权利要求数、IPC 分类数、申请人类型、首项权利要求字数、文件页数、技术领域等对用维持年限超过 8 年表征专利质量的影响关系，发现首项权利要求字数对维持时间负向影响。杨武等[21]使用主客观赋权方法和 Cox 比例风险回归模型分析专利的权利要求数、引用专利数、被引次数、IPC 分类数和专利同族数对多个专利结果特征的影响进行分析。李睿等[22]研究证

明了专利的各类引用均会正向影响专利维持年限。张亚峰等[23]使用1 704 项大学转让专利作为样本数据,以转让价格作为专利价值代理变量,使用回归分析发现专利价值是内生的,权利要求数、专利文献页数、海外同族专利数、独立权利要求字数、专利寿命对专利价值显著影响,发明人数量、审查时间、引用信息等与专利价值没有显著相关性。

技术维度的高价值专利主要指该专利的方案具有较大技术进步性,具有行业导向作用,专利的新颖性、创造性和实用性被许多学者看作是专利价值的最主要特征[24]。已有研究通过衡量专利技术主题或关键词新颖性来表示专利新颖性然后研究与专利价值的关系,主要是挖掘主题[25]识别出专利的新主题或者专利关键词新颖性[26],发现高新颖性主题、高常规性主题以及新颖关键词会正向影响专利价值。还有学者利用文本内容相似度[27]、IPC 组合[28]或是利用引文网络建立新链接[29]的时间来测度技术知识新颖性,研究结果表明会正向影响专利价值。郭颖等[30]从摘要中统计新兴术语的出现情况定量计量专利技术新兴度,然后使用回归模型分析与技术影响力间的关系,在纳米载药领域实证分析中发现技术新兴度会正向影响以专利被引次数表示的技术影响力。已有研究从不同角度初步证实了专利新颖性与专利价值间存在正向影响关系,但目前专利新颖性测度方法还存在着不足,如用技术主题来度量专利文本的新颖性可能还不够细致。另外,新颖性是一个相对概念,专利在刚申请时对于领域知识来讲可能是新颖的,因此应以专利申请时间为基准点,以领域先前知识为参照物来度量新颖性比较合理。本文认为,专利申请后的技术影响,IPC 与专利内容相比粒度较大,两者都不能精细地反映专利技术内容的新颖性,所以有必要从专利的技术内容本身出发,在本质上揭示出专利知识的新颖性。

(三)专利价值预测

专利价值预测是以专利的前因特征和申请特征作为输入变量,利用机器学习对一项或多项专利结果特征进行预测。Jiang 等[31]用专利的文本和结构化特征预测申请文本是否会被授权。马瑞敏等[32]在细分视角下,采用四年内被引用次数、同族专利数、IPC 前 4 位表示的专利宽度、权利要求数和引用科学论文数量表示的科学关联度 5 个专利特征作为模型输入变量,利用支持向量机对10 年内高被引专利进行预测。刘夏等[33]选择技术覆盖范围、同族专利数量、非专利文献的引用数量、合作专利、优先权相关、授权与否等作为输入变量,以专利被引次数作为输出指标,使用随机森林模型作为预测算法。王思培等[34]以同族成员数量、同族总被引数量等作为输入变量,以高端专利分析工具(Innography Advanced Analysis,Innography)中的专利强度作为专利价值指标,采用随机森林模型对潜在高价值专利进行了预测。Zhang 等[35]通过构建专利的后向引文网络、前向引文和综合引文网络分析专利引文网络中的多种网络结构特征如结构洞、网络密度和中心性等分析知识的生成和扩散,对专利价值进行早期预测。Choi 等[36]从专利数据中获得文献页数、从属权利要求数、家族数、优先权等 24 个变量和历史维持费用,使用前馈神经网络(FFN)预测专利的维持时间以评估专利的商业潜力。符川川等[37]利用自组织映射划分专利质量组,将专利家族数量、专利优先权等作为输入,利用支持向量机对专利质量进行分类预测。这些研究提出了一些行之有效的专利价值预测方法,但在输入变量方面尚未考虑先前领域知识和专利文本等可能进一步提升专利价值预测准确率的内在专利特征。另外一些研究使用了动态变化特征和结果特征作为输入变量,如四年内被引用次数、同族专利数、同族成员数量、同族总被引数量等专利家族指标。这些动态特征和结果指标在一定程度上已经度量了专利的价值,使得专利价值的预测具有后效性,难以在专利申请早期预测出潜在高价值专利。

(四)知识网络与专利价值

技术创新是一个知识、经验和技能的累计过程,创新主体在领域先前知识基础上进行整合、发展和创造,在此过程中,由知识元及其之间关系所构成的知识网络起到了重要作用。目前,在专利价值的研究中,主要是专利引文网络评估专利价值、专利影响力或者揭示技术发展路径和

识别关键技术[38-39]。Mariani 等[40]构建了专利引文网络，计算出年份调整的 PageRank 并作为专利价值预测模型的输入变量，用专家选出的历史高价值专利作为输出变量。于超等[41]探究了专利知识特质对其所处引文网络中心位置的影响，发现反映专利知识影响力的前向引用数量对专利点度中心度有显著正向影响，专利知识多样性会提升网络中的点度中心度和接近中心度。张克群等[42]以专利 H 指数作为专利价值代理变量，通过稀有事件 logit 回归和倾向得分匹配模型发现中心度、PageRank 值等对专利价值都具有正向影响。马荣康等[43]发现知识元素的网络中心度越低，占据结构洞位置越少，突破性技术发明形成的概率越高。网络嵌入（network embeddedness）可以反映行动者的社会网络地位以及与其他网络主体间的关系[44]，通常分为关系嵌入和结构嵌入两个维度。从关系嵌入维度上讲，强关系代表领域间的知识经常被整合利用，彼此间的联系比较紧密；弱关系则表示知识元之间的联系比较松散可以增强知识重组的灵活性[45]。在结构嵌入维度上，学者经常使用结构洞、中心性等特征衡量。在创新管理领域，通常使用知识网络嵌入来研究创新绩效和突破性创新。如李彦勇等[46]构建组织知识网络，发现知识网络中心势、结构洞与组织突破性创新绩效呈正相关的关系；刘嘉明等[47]从合作网络和知识网络的视角解析人工智能企业进行专利创新的路径，发现容易组合的知识元素对企业的专利创新数量具有先促进后抑制的作用。

已有研究通常使用专利的结果特征作为专利价值代理指标，对专利价值进行界定，进而对专利价值进行评估、影响因素识别或者预测研究，但在专利价值预测方法和专利技术价值的新颖性和常规性量化方面还存在不足。本文采用维持年限超过 8 年的专利作为高价值专利的界定标准，运用知识网络嵌入精细地计量专利的新颖性和常规性，并将其应用到专利价值的早期预测中。

二、研究设计

本文提出一种融合知识网络嵌入特征的专利价值早期预测方法，根据知识重组理论、专利新颖性和常规性的相对性，以焦点专利申请时间为基准点，以领域先前知识为参照物，通过挖掘专利文本，构建语义知识网络，不仅有利于识别新的知识元，将其嵌入到知识网络中，还能发现焦点专利在当前知识体系中的关系特征和位置特征。专利文本是由众多知识元依据句法关系组成的，以知识元为节点、知识元间的关系为连边可以构成知识网络[48]，通过网络嵌入的方式将样本专利知识与领域先前知识联系起来，在领域整体视角观测样本专利的“关系”和“地位”。根据复杂网络理论，知识元在知识网络中的分布并不均匀，其网络结构特征在一定程度上反映其重要性。例如，中心性代表了与其他节点直接相连的个数，反映了该知识元素与其他知识元素的组合潜力[29]，表现了该知识元的实用性。在研究专利价值时，本文考虑知识网络中知识元的新颖性和位置特征，通过选取和计算网络嵌入特征来准确量化专利新颖性和常规性，同时结合其他申请时可获取的静态特征对专利价值进行早期预测。

本文研究方案如图 1 所示。首先，在数据库检索选定技术领域的专利数据，对摘要文本进行预处理；对经过预处理的摘要利用自然语言处理技术抽取“输入—处理—输出”（Input，Process，and Output，IPO）结构[49]，分别构建样本专利知识网络和与之相对应的领域先前知识网络，将样本专利知识网络嵌入到领域先前知识网络中，计算表示样本专利新颖性和常规性的网络嵌入特征；同时，结合创新主体等前因特征和专利申请特征作为输入变量，选择维持年限作为区分高价值专利的代理指标，建立并训练基于机器学习算法的专利价值预测模型；最后，对模型效果进行评价和解释。

（一）知识网络构建

本文借鉴任海英等[49]构建知识网络的方法，通过抽取摘要文本的“输入—处理—输出”知识元信息，构建专利知识网络和领域先前知识网络，计算样本专利的知识网络嵌入特征，以申请时间

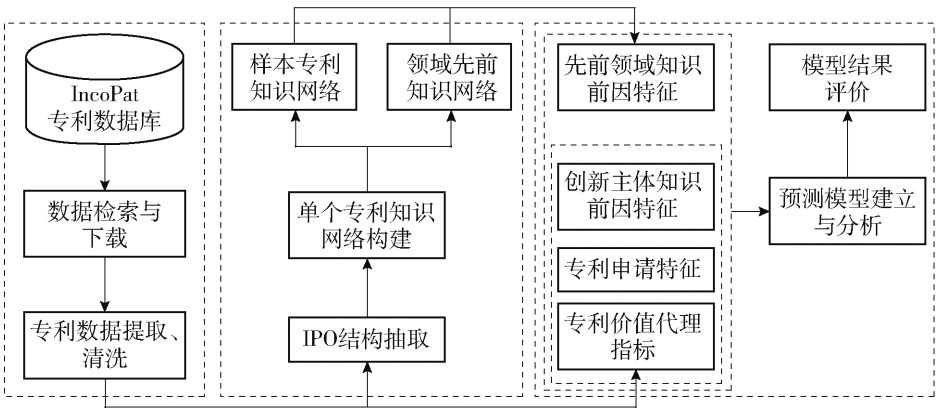

图 1　融合知识网络嵌入特征的高价值专利预测技术路线

为基准点，以领域知识为参照物，量化专利的新颖性和常规性。

1. 数据获取及文本预处理

（1）数据获取及清洗：根据焦点技术领域相关知识制定检索式，在数据库中检索并下载相关专利文本，解析专利文献，获取相关信息。

（2）文本预处理：对专利摘要删减无用字段、缩写词替换、指代消歧、分句等操作规范化专利摘要文本，完成预处理，方便后续使用自然语言处理技术抽取出关键信息。

2. 输入—处理—输出（IPO）结构抽取

（1）句法分析：对预处理过的摘要文本中每个单句进行句法分析，得到该单句的短语结构树（constituent tree），使用短语结构树描述词或词组之间的句法依存关系。

（2）抽取 IPO 结构三元组：根据短语结构树以及制定正则表达式来抽取 IPO 三元组。将名词块作为输入（I）或输出（O）要素，动词/介词/连词等组合作为处理要素（P）。根据设置的抽取规则使用正则表达式来提取短语结构树中的动词、介词和连词，并对相邻 P 要素合并。从短语结构树的根节点出发，根据 P 要素去搜寻其最近的父节点及子节点对应的 NP 块，记为 NP1 和 NP2 块并将它们作为输入要素（I）或输出要素（O），可以得到［NP1，P，NP2］，即为 IPO 结构三元组。

（3）规范化 IPO 结构：通过词形还原、首字母小写和去除停用词等操作对 IPO 结构进行规范化处理。

3. 专利知识网络构建

对于单篇专利，通过获得摘要中每个单句的 IPO 结构，从而获得整篇专利 IPO 边列表，去除重复的 IPO 边，将 I 和 O 要素作为知识网络节点，P 要素作为网络连边构建出加权无向单篇专利知识网络。

4. 领域先前知识网络构建

首先根据样本专利的申请时间来确定领域先前知识的时间范围，然后汇总申请时间之前技术领域内所有公开专利的 IPO 边列表，得到领域边列表集合，将 I 要素与 O 要素均相同的边合并，I 和 O 要素作为节点，P 要素作为连边，以 IPO 网络中各边所包含的 P 要素数量作为网络边权。经过上述步骤，完成加权无向的领域先前知识网络构建。

（二）知识网络嵌入特征选取和计算

设专利 i 在 t 年申请，则用 $t-1$ 及之前年份领域内公开的所有专利构建 i 的领域先前知识网络。专利 i 的网络嵌入特征是指 i 的专利知识网络嵌入到领域先前知识网络中所具有的网络特征和关系特征。已有研究表明，专利知识的新颖性或新兴性、常规性会对专利价值产生影响。假设单篇样本专利知识网络有 N 个节点，计算得出每个节点的网络嵌入特征，分别保存该专利 N 个节点

的均值、中值和最大值，进而计算得到专利的知识网络嵌入特征。

假设样本专利 i 的知识网络为 S_i，各网络嵌入特征计算方法如下所示。

1. 较新节点数量

较新节点是指 S_i 中某个节点申请年份大于等于领域先前知识网络所有节点公开年份的 60% 分位数，该特征用来表示专利的新颖性或新兴性。拥有较新节点数量越多的专利其价值一般也越高，本文用较新节点数量的算术平方根表示。

$$\text{较新节点数量} = \sqrt{S_i \text{ 中较新节点数量}} \tag{1}$$

2. 新颖边数量

新颖边是指领域先前知识网络中不存在而 S_i 中存在的边。新颖边数量反映了样本专利的创新性和技术发展前沿，新颖边数量越多，表明专利的新颖性越强。本文用算术平方根表示新颖边数量。

$$\text{新颖边数量} = \sqrt{S_i \text{ 中新颖边数量}} \tag{2}$$

3. 边年份跨度

边年份跨度是用来衡量 S_i 中边出现的最晚年份与最早年份的差值。该特征用来度量专利知识的常规性，边的年份跨度较久说明该边连接知识元所代表的技术内容出现时间越早，可以认为该技术较为经典，比较实用。选择 S_i 中所有边的年份跨度均值、中值和最大值三种计量方式来表示专利边年份跨度。

4. 网页排名(PageRank)

张欣等[50]利用改进的 PageRank 算法识别出了核心专利。因此本文将其作为衡量专利常规性的特征。将样本知识网络嵌入到领域先前知识网络后，首先计算出 S_i 中每个节点的 PageRank 值，然后计算出所有节点的 PageRank 均值、中值和最大值作为专利价值的影响变量。

5. 征向量中心性

特征向量中心性反映 S_i 中节点的结构重要性，特征向量中心性越高说明越处于网络重要位置，表明 S_i 中核心节点越多。特征向量中心性对专利价值产生显著影响[39]。保存 S_i 嵌入先前知识网络的特征向量中心度的均值、中值和最大值作为输入变量。

(三)专利价值影响变量选取及专利价值划分

1. 专利前因特征选取

专利前因特征主要包括知识网络嵌入、创新主体和技术基础。知识网络嵌入特征已在上节进行详细表述。专利创新主体是指专利的发明人或申请人。发明人是为发明创造专利做出创造性贡献的人，申请人体现专利创造过程中的合作关系，反映专利申请的复杂性。本文选取发明人数量和申请人数量作为创新主体前因特征。

对于技术基础来讲，选择以显性知识表示的引用专利文献数量和引用非专利文献数量。两者可以分别表示与技术和前沿科学间的关联程度，引用的数量越多，说明联系越紧密，是比较常用的专利价值影响变量。

2. 专利申请特征选取

本文定义的专利申请特征是指专利申请时就具备的特征，包括摘要、说明书等文本知识和可以直接获取到的结构化特征，并且选取的特征是静态的，不随时间发生变化，便于准确地对专利申请或授权早期阶段专利价值进行预测，选取权利要求数量、技术覆盖范围、是否有代理机构、文献页数作为专利申请特征的变量。

本文选取用于专利价值预测的前因和申请特征共 19 个，如表 1 所示。

表 1　专利价值预测的输入变量

维度		编号	变量名称	意义
专利前因特征	创新主体知识	A_1	发明人数量	样本专利的发明人数量
		A_2	申请人数量	样本专利的申请人数量
	知识网络嵌入	B_1	较新节点数量	样本专利知识网络较新节点数量
		B_2	新颖边数量	样本专利知识网络新出现的边数量
		B_3	边年份跨度均值	边出现最早年份与最晚年份的差值
		B_4	边年份跨度中值	
		B_5	边年份跨度最大值	
		B_6	PageRank 均值	样本专利知识常规性
		B_7	PageRank 中值	
		B_8	PageRank 最大值	
		B_9	特征向量中心性均值	样本专利知识常规性
		B_{10}	特征向量中心性中值	
		B_{11}	特征向量中心性最大值	
	技术基础	C_1	引用专利数量	样本专利技术关联度
		C_2	非专利引用文献数量	样本专利科学关联度
专利申请特征		D_1	权利要求数量	样本专利权利要求数量
		D_2	技术覆盖范围	IPC 分类号数量（前 4 位）
		D_3	文献页数	样本专利文献页数
		D_4	是否有代理机构	样本专利是否有申请机构

为避免各输入变量间取值差异较大而对模型预测结果产生影响，本文采用 Min-Max 标准化方法对数据进行标准化处理来增强模型的稳定性。计算公式如下所示。

$$x_{\text{norm}} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

3. 专利价值划分

专利维持年限是比较公认的专利价值代理指标，只有专利权人获得收益大于成本时，才会继续维持专利的有效性，故通常使用维持年限认定其是否为高价值专利，在一定程度上表现出创新主体的私人价值。

（四）专利价值早期预测模型构建

本文采用 5 折交叉验证，按照 7∶3随机划分训练集和测试集，利用网格参数调优得到在测试集上效果最好的模型结果。利用多种机器学习算法进行专利价值的早期预测。

1. 预测模型构建

（1）逻辑回归（Logistics Regression，LR）。逻辑回归用于分类问题的基本思想是将训练数据转换成对应的结构化数值，将数据拟合进一个逻辑函数来估计属于某个类别的概率。逻辑回归的优点是计算消耗资源少，计算结果便于直接观测样本概率分布。

（2）支持向量机（Support Vector Machines，SVM）。支持向量机以统计学习理论为基础，基于结构风险最小化原理，将实际问题通过非线性变换转换到高维特征空间，学习并得到分类决策函数。

（3）决策树（Decision Tree，DT）。决策树通过建立一个树形图，利用一系列分类规则对样本点进行逐层判断和剪枝，从而实现分类，具有简单快速、计算过程透明等优点。

（4）随机森林（Random Forest，RF）。随机森林是以决策树为基分类器，组合多个决策树的集成分类器，可以缓解决策树过拟合的问题，对噪声和异常值不敏感。

（5）极端随机树（Extremely Randomized Trees，Extra Trees）。极端随机树与随机森林相似，是一种由多棵决策树构成的集成学习方法。随机森林采用随机采样，不能保证所有样本被充分利用，并且各决策树之间可能存在相似性。极端随机树可以弥补以上不足，具有抗噪能力强，训练时间效率高等优点。

（6）神经网络（Artificial Neural Network，ANN）。根据神经网络原理，设置输入变量和输出变量数分别为 N 和 M，依据不同实验设置不同个数的输入变量，输出为“高价值专利”和“非高价值专利”两类（$M=2$）。本文根据一般设计方案，中间层为 S 型正切函数，使用 softmax 激活函数，通过训练确定最后参数。

2. 模型评价方法

本文使用准确率（Accuracy）、$F1$ 和 AUC 共 3 个指标评价模型的性能。准确率（Accuracy）和 $F1$ 指标用混淆矩阵计算。

（1）准确率（Accuracy）和 $F1$

表 2 中 TP 表示模型预测为高价值专利中真实标签为高价值专利的数量；TN 表示模型预测为一般价值专利中真实标签为一般价值的专利数量；FP 表示模型预测为高价值专利中真实标签为一般价值专利的数量；FN 表示模型预测为一般价值专利中真实标签为高价值专利的数量。

表 2　混淆矩阵

	实际正类	实际负类	总计
预测正类	TP	FP	TP + FP
预测负类	FN	TN	FN + TN
总计	TP + FN	FP + TN	TP + FN + FP + TN

准确率（Accuracy）是指对于给定的测试集，预测模型正确预测的样本数量与全部样本数量的比值。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$F1$ 在分类模型性能评估的实际应用中，需要综合考虑模型的精确率和召回率，通过对二者的加权调和平均作为评价指标；在本文中，精准率表示实际维持年限大于等于 8 年的专利数量占预测结果的比例；召回率表示被预测出的维持年限大于等于 8 年的专利占实际专利维持年限大于等于 8 年的比例。

$$F1 = \frac{\text{TP}}{\text{TP} + \dfrac{\text{FN} + \text{FP}}{2}} \tag{5}$$

（2）AUC（Area Under the Curve）

AUC 是通过绘制 ROC 曲线并计算其下面积来计算的，可以反映分类器性能，AUC 越大，分类器性能越好，计算公式如下所示。

$$\text{AUC} = \frac{\sum_{i \in \text{positive}} \text{rank}_i - M(M+1)/2}{M \times N} \tag{6}$$

在本文中维持年限大于等于 8 年的专利作为正样本集合，其余为负样本集合。其中，positive 表示正样本，M 为正样本个数，N 为负样本个数；rank_i 表示第 i 条正样本的序号。

三、实证研究

（一）数据说明

本文以合享（IncoPat）全球专利文献数据库为数据源，以神经网络技术领域专利作为研究对象，通过查阅相关文献资料确定检索式为“IPC - LOW =（G06N3/02）”，检索截止时间为 2021 年 12 月 31 日，共得到神经网络领域自 1985 年以来的 6 000 余篇发明专利。综合考虑国家知识产权局对高价值发明专利的统计标准、数据可获取性以及指标被认可程度，选用维持年限作为真实专利价值的代理指标，选取 2012—2013 年发明申请专利作为样本专利。为便于专利价值的分类预测，将维持年限大于等于 8 年的标记为 1，代表高价值专利；小于 8 年的记为 0，代表非高价值专利。删去存在缺失值的专利数据，得到 2 190 项样本专利，按照维持年限进行分类标注，共有 1196 项高价值专利和 994 项非高价值专利。通过构建知识网络并计算知识网络嵌入特征，得到研究的全部变量。使用 Min-Max 方法对数据进行标准化处理，消除量纲对预测结果的影响。

本文对总体样本专利、高价值专利子样本和非高价值专利子样本计算均值、方差并进行独立样本检验。从表 3 可以看出，高价值专利的技术覆盖范围、最大特征向量中心性、较新节点数量、新颖边数量等特征均值都高于非高价值专利。高价值专利与非高价值专利在大部分特征上存在显著性差异，如权利要求数量、文献页数、引用专利数量、最大特征向量中心性、新颖边数量等。

表 3　专利价值预测变量统计分析

特征	总样本		高价值专利样本		非高价值专利样本		平均值等同性 t 检验	
	均值	方差	均值	方差	均值	方差	t	Sig.（双尾）
权利要求数量	18.477	473.821	16.730	219.277	20.579	772.522	4.136	0.000
申请人数量	1.351	0.948	1.401	1.109	1.290	0.748	2.675	0.008
技术覆盖范围	1.767	0.761	1.882	0.943	1.629	0.508	6.838	0.000
发明人数量	3.528	6.174	3.652	6.952	3.378	5.204	2.606	0.009
文献页数	27.925	673.661	24.940	371.044	31.517	1 014.873	5.950	0.000
引用专利数量	8.504	692.083	11.630	882.643	4.742	437.525	6.151	0.000
非专利引用文献数量	6.495	405.137	9.007	578.303	3.474	180.419	6.464	0.000
新颖边数量	1.632	0.137	1.644	0.125	1.618	0.152	1.670	0.095
边年份跨度中值	0.053	0.614	0.052	0.498	0.054	0.755	0.060	0.952
边年份跨度最大值	2.248	37.547	2.090	35.455	2.438	40.037	1.321	0.187
边年份跨度均值	0.392	1.670	0.355	1.462	0.436	1.918	1.455	0.146
特征向量中心性中值	0.003	2.06×10^{-4}	0.003	9.30×10^{-5}	0.004	9.30×10^{-5}	2.481	0.013
特征向量中心性最大值	0.111	0.011	0.116	0.012	0.104	0.011	2.610	0.009
特征向量中心性均值	0.018	4.70×10^{-4}	0.018	4.06×10^{-4}	0.018	0.001	0.750	0.454
PageRank 中值	1.27×10^{-4}	1.04×10^{-7}	1.12×10^{-4}	6.27×10^{-8}	1.44×10^{-4}	6.27×10^{-8}	2.282	0.023
PageRank 最大值	0.003	4.00×10^{-6}	0.003	5.00×10^{-6}	0.002	5.00×10^{-6}	1.427	0.154
PageRank 均值	4.59×10^{-4}	2.05×10^{-7}	4.63×10^{-4}	1.88×10^{-7}	4.55×10^{-4}	1.88×10^{-7}	0.416	0.677
较新节点数量	2.906	1.068	2.927	1.007	2.881	1.141	1.047	0.295

（二）实验设计及结果分析

1. 实验设计

在训练集和测试集上分别使用上述算法训练并评价模型预测效果。为了更好地了解和分析知识网络嵌入特征对预测结果的影响，本文设置两组对比实验，分别为不考虑网络嵌入特征和考虑全部前因特征和申请特征，分析比较网络嵌入特征对高价值专利预测的影响，如表 4 所示。表 5 是两组特征所构建模型在测试集上的性能。

表 4　实验设计

实验	特征维度	特征选取
实验一	不考虑网络嵌入特征	A_1、A_2、C_1、C_2、$D_1 \sim D_4$
实验二	考虑全部前因特征和申请特征	A_1、A_2、C_1、C_2、$B_1 \sim B_{11}$、$D_1 \sim D_4$

表 5　基于维持年限的高价值专利预测模型性能对比

方法	实验一			实验二		
	准确率（Accuracy）	*F*1	AUC	准确率（Accuracy）	*F*1	AUC
逻辑回归（LG）	0.64	0.68	0.63	0.71	0.68	0.66
支持向量机（SVM）	0.66	0.72	0.73	0.73	0.75	0.76
决策树（DT）	0.67	0.71	0.71	0.72	0.76	0.76
随机森林（RF）	0.75	0.75	0.76	0.82	0.83	0.82
极端随机树（ExtraTrees）	0.76	0.78	0.76	0.81	0.84	0.84
神经网络（ANN）	0.75	0.76	0.75	0.80	0.82	0.79

2. 结果分析

由表 5 实验一和实验二结果可知，融合知识网络嵌入特征后，各类模型预测效果均有较大提升。实验二与实验一相比准确率和 *F*1 值提高 5% 左右，说明知识网络嵌入特征的加入，可以显著提高高价值专利预测的准确率、*F*1 和 AUC 等。此外，在两组实验中，基于树模型的结果如随机森林（RF）、极端随机树（ExtraTrees）的准确率及 *F*1 值均超过 75%，表明分类器具有较好的泛化能力和实际应用价值。

为了更好地理解每个输入变量对预测变量（专利价值）的影响，鉴于夏普利加性解释（SHAP）的理论支持[51-52]和稳定性，本文选择 SHAP 对实验二模型结果进行解释。将训练好的极端随机树（ExtraTrees）作为预测模型，输入到 SHAP 中，计算出每个输入变量对于预测变量的 SHAP 值，分析输入变量的重要性和对预测变量的影响方向（详见图 2）。

具体来看，（a）子图按照每个输入变量的 SHAP 绝对值的平均值从高到低排序，反映了每个输入变量对维持年限的影响程度，可以看出权利要求数、引用专利数量、文献页数、代理机构、非专利文献引用数量对维持年限表征的高价值专利（class 1）和一般价值专利（class 0）都有重要影响，且影响程度逐渐降低；知识网络嵌入特征中的特征向量中心性和新颖边数量对维持年限产生重要影响。（b）子图是密度图，一个点代表了一个样本，帮助理解每个输入变量对预测变量的影响方向。*x* 轴是 SHAP 值，表示输入变量对模型的影响，SHAP 值大于 0，说明是正向积极影响，反之则为负向消极影响。在本文中，正向表示高价值专利，负向表示一般价值专利。颜色代表该特征值的大小，越靠近红色代表输入变量值越大，越靠近蓝色越小。由（b）子图可知，在维持年限预测模型中，申请特征中的权利要求数量、引用专利数量、文献页数、代理机构等对维持年限的影响程度逐渐降低；

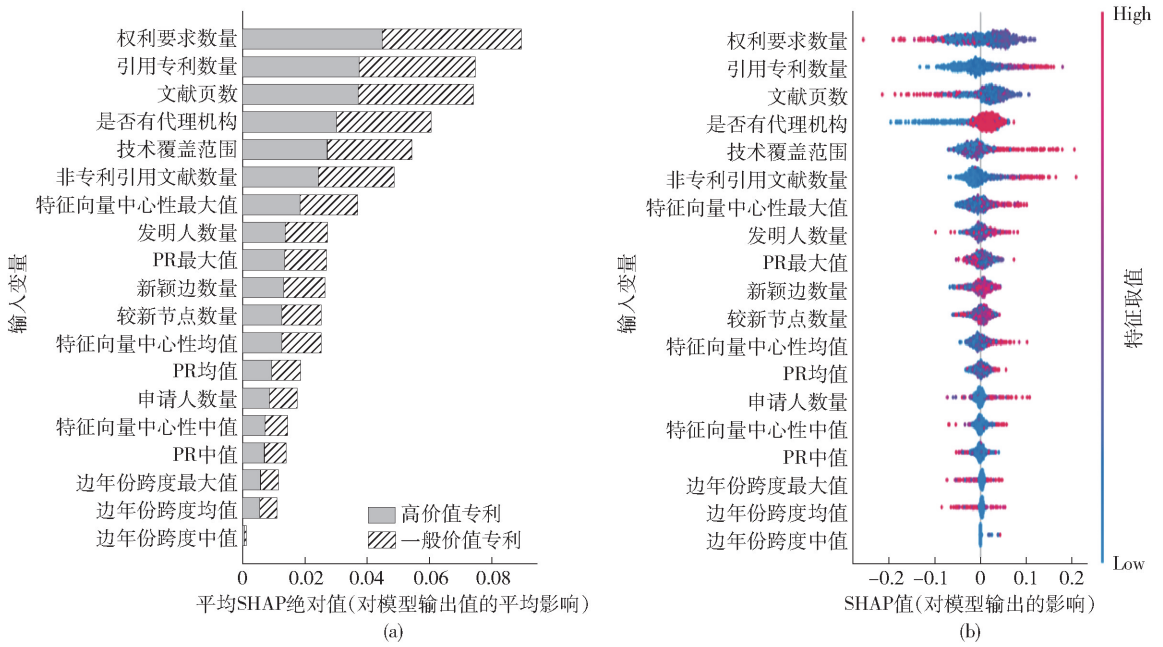

图 2　维持年限的变量重要性排序和变量对模型预测影响

知识网络嵌入特征中的特征向量中心性和新颖边数量会对维持年限产生重要影响。权利要求数量、文献页数可以在一定程度上反映出专利技术的广泛性和复杂性，但是对维持年限显著负向影响；代理机构会正向影响专利的维持年限，因为代理机构的参与通常会以较为恰当的词汇和表达方式完善专利申请文件。技术覆盖范围、引用专利数量和非专利引用文献数量分别表示了专利的实用性和与先进技术知识、科学知识联系的紧密联系，会正向影响维持年限。本文用来表示专利常规性的特征向量中心性、专利知识新颖性中的较新节点数量和新颖边数量会正向影响专利的维持年限，这与先前分析和以往研究的结论是一致的。专利知识元的 PageRank 值和特征向量中心性可以识别出知识网络中的核心节点，特征的值越大，说明该专利具备的知识元不仅与领域中的其他重要知识元素相连接，且其本身也具有较高的中心性。在专利文本中，一些专有名词能在很大程度上代表专利的创新内容，因此新颖和较新的知识元表示了新的知识和创意，它们数量越多代表创新性越强，对知识网络增长的贡献可能也越大。该结论进一步验证了高价值专利在成文时已具有较高价值。

在对维持年限的实证分析中发现，是否有代理机构的影响程度相对比较大，并且权利要求数量和文献页数对维持年限显著负向影响，这个发现与现在的主流观点有出入。本文认为，这可能是因为文献页数和权利要求数量容易受到代理机构影响的原因。代理机构通过撰写专利文件以较为专业的术语完善专利技术表达，增加文献页数，并且可以通过拆分技术特征的方式增加权利要求数量，这种情况下，专利价值并不一定就高，从而不会正向影响专利价值。

3. 实验方法验证

本文为验证方法的科学性和灵活性，从专利交易这种能够体现专利商业价值的场景出发，选择专利交易（转让或者许可）作为专利价值代理指标，并使用上述方法进行早期预测，获取 2 190 项样本专利的交易数据，其中发生过交易的有 733 项，将其记为 1 表示高价值专利；未发生交易的有 1 457 项，将其记为 0，表示一般专利。由于专利价值分类数据不均衡，为保证预测模型质量，采用 SMOTE（synthetic minority over-sampling technique）过采样方法[53]，对少数类样本人工合成新样本添加到数据集中，达到平衡正负类样本数据分布的目的。预测模型在测试集上的性能指标如表 6 所示。

表 6　专利交易预测模型的性能指标对比

方法	实验一			实验二		
	准确率（Accuracy）	*F*1	AUC	准确率（Accuracy）	*F*1	AUC
逻辑回归（LG）	0.69	0.67	0.71	0.73	0.68	0.72
支持向量机（SVM）	0.71	0.68	0.72	0.71	0.69	0.73
决策树（DT）	0.75	0.73	0.74	0.74	0.71	0.73
随机森林（RF）	0.81	0.76	0.82	0.87	0.83	0.85
极端随机树（ExtraTrees）	0.82	0.75	0.83	0.85	0.83	0.86
神经网络（ANN）	0.79	0.76	0.81	0.82	0.81	0.82

由表 6 预测性能的对比可知，对专利交易数据来讲融合知识网络嵌入特征和专利申请特征的预测模型依然是最优的。实验二的预测准确率、*F*1 和 AUC 提高了 3% 以上，并且基于树模型的极端随机树（ExtraTrees）表现最优，其次是随机森林（RF）。最佳预测模型的准确度在 80% 以上，具有较好的预测性能和实用价值，从而验证了预测模型的科学性和灵活性。

四、结论

面向海量申请专利，尽早预测识别出高价值专利是促进核心技术发展的重要一环。本文提出一种新的方法量化专利新颖性和常规性，并结合其他专利常规特征对专利价值进行早期预测。首先，根据知识重组理论，考虑到专利技术新颖性和常规性的相对性，以专利申请时间为基准点、领域先前知识为参照物，通过挖掘专利文本知识分别构建了样本专利知识网络和领域先前知识网络，将样本专利知识网络嵌入到领域先前知识网络中表示样本专利知识在领域中的“地位”和“关系”，选取和计算知识网络嵌入特征来量化样本专利的新颖性和常规性；然后，结合在专利申请或授权的阶段就可获取的静态特征，构建了用于高价值专利早期预测的输入变量体系，利用机器学习算法对专利价值进行早期预测；最后，使用神经网络技术领域专利数据，选用维持年限作为专利价值代理指标。研究发现，知识网络嵌入常规性特征确实会正向影响以维持年限表征的专利价值，且能提高预测模型性能，并使用专利交易数据作为模型输出指标进行了方法验证，结果同样也证明了这些结论。

本文的研究结果和方法具有一定的现实和理论意义。在现实应用层面，一方面，早期高价值专利的识别为创新主体的专利申请、商业化等专利管理决策提供指导建议，为其技术培育和发展赢得更多的时间，创造更多机会和技术创新空间，从而帮助企业和国家尽早抢占市场发展先机和主动权；另一方面，能帮助创新主体更有针对性地增加关键专利技术的资源配置，避免盲目投入。在理论层面上，以专利申请时间为基准点，以领域先前知识为参照物，以知识网络嵌入新颖性和常规性量化样本专利的新颖性和常规性，一方面能够提高这些特征变量的细致程度和解释性，另一方面也扩展了专利价值的特征维度，为专利价值研究提供了新的思路，也丰富了知识网络与专利价值之间关系的分析内容。

参考文献：

[1] 柳卸林，马瑞俊迪，刘建华. 中国离科技强国有多远[J]. 科学学研究，2020，38(10)：1754-1767.

[2] 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要[EB/OL].

(2021-03-13)[2022-10-28]. http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.
[3] GRIMALDI M, CRICELLI L. Indexes of patent value: a systematic literature review and classification[J]. Knowledge Management Research & Practice, 2019(1): 1-20.
[4] 冯仁涛. 基于专利文献的专利维持时间影响因素分析[J]. 情报杂志, 2020, 39(7): 202-207
[5] 肖冰. 基于法定保护期的专利维持时间影响因素研究[J]. 科学学研究, 2017, 35(11): 1652-1658.
[6] 李黎明, 张敏, 李小娟. 引证网络专利质量对专利拍卖经济价值的影响效应研究[J]. 情报杂志, 2021, 40(10): 115-121.
[7] 张杰, 孙超, 翟东升, 等. 基于诉讼专利的专利质量评价方法研究[J]. 科研管理, 2018, 39(5): 138-146.
[8] 魏太琛, 刘敏榕, 陈振标. 高校专利技术转移转化价值影响因素实证分析——基于 11 所一流高校专利转移转化数据[J]. 图书情报工作, 2022, 66(9): 103-116.
[9] 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.
[10] 彭小宝, 陈文清. 我国高价值发明专利界定标准研究[J]. 科技与法律(中英文), 2021(6): 58-64.
[11] 程文银, 胡鞍钢, 陈雪丽. 知识产权强国背景下中国高价值专利发展: 测度与实证分析[J]. 北京工业大学学报(社会科学版), 2022, 22(5): 1-12.
[12] 王子焉, 倪渊, 张健. 基于灰色关联分析-随机森林回归的网络平台专利价值评估方法研究[J]. 情报理论与实践, 2019, 42(10): 109-116.
[13] 慎金花, 刘玥, 张更平. 单项专利价值的评估与定量评估指标体系的构建——基于邻域粗糙集与果蝇优化神经网络的单项专利价值评估[J]. 大学图书馆学报, 2020, 38(3): 48-56.
[14] 李娟, 李保安, 方晗, 等. 基于 AHP-熵权法的发明专利价值评估——以丰田开放专利为例[J]. 情报杂志, 2020, 39(5): 59-63.
[15] 杨登才, 李国正. 高校专利质量评价体系重构与测度——基于 23 所高校的实证分析[J]. 北京工业大学学报(社会科学版), 2021, 21(2): 109-121.
[16] 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.
[17] 郭状, 余翔. 基于我国人工智能专利数据的专利价值影响因素分析[J]. 情报杂志, 2020, 39(9): 88-94.
[18] 马荣康, 王艺棠. 知识组合多样性、新颖性与突破性发明形成[J]. 科学学研究, 2020, 38(2): 313-322.
[19] 荣雪云, 杨中楷, 徐鑫. 发明者特征、技术新颖性和发明质量关系研究[J]. 科学学与科学技术管理, 2020, 41(10): 93-104.
[20] 冯仁涛. 基于专利文献的专利维持时间影响因素分析[J]. 情报杂志, 2020, 39(7): 202-207
[21] 杨武, 孙世强, 陈培. 技术锁定视角下的专利价值影响因素分析[J]. 科学学研究, 2022, 40(6): 1024-1033.
[22] 李睿, 王堂蓉, 龙瑞. 专利引证与专利维持时间的相关性实证[J]. 情报杂志, 2022, 41(7): 71-76.
[23] 张亚峰, 李黎明. 专利价值再认识: 大学专利转让的实证研究[J]. 科学学研究, 2022, 40(9): 1608-1620.
[24] 刘妍. 专利价值评估研究综述与趋势展望[J]. 图书情报工作, 2022, 66(15): 127-139.
[25] LU Q, CHESBROUGH H. Measuring open innovation practices through topic modelling: revisiting their impact on firm financial performance[J]. Technovation, 2022, 114: 102434.
[26] ARTS S, HOU J, GOMEZ J C. Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures[J]. Research Policy, 2021, 50(2):

104144.
[27] 任海英，邵文，李欣. 基于专利内容新颖性和常规性的突破性发明影响因素和研发策略分析[J]. 情报杂志，2019，38(2)：56-63.
[28] 王萍萍，王毅. 技术新颖性从何而来？——基于纳米技术专利的分析[J]. 管理工程学报，2020，34(6)：79-89.
[29] VEUGELERS R，WANG J. Scientific novelty and technological impact[J]. Research Policy，2019，48(6)：1362-1372.
[30] 郭颖，王明星，段炜钰. 专利的技术新兴度与其技术影响力间关系研究[J]. 科学学研究，2022，40(6)：1034-1043.
[31] JIANG H，FAN S，ZHANG N，et al. Deep learning for predicting patent application outcome：the fusion of text and network embeddings[J]. Journal of Informetrics，2023，17(2)：101402.
[32] 马瑞敏，尉心渊. 技术领域细分视角下核心专利预测研究[J]. 情报学报，2017(12)：1279-1289.
[33] 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(4)：402-410.
[34] 王思培，韩涛. 基于随机森林算法的潜在高价值专利预测方法研究[J]. 情报科学，2020(5)：120.
[35] FENG Z，JIANG G，HE X. Patent citations and value：through the lens of a social network approach[J]. International Journal of Management and Network Economics，2018，4(2)：115-143.
[36] CHOI J，JEONG B，YOON J，et al. A novel approach to evaluating the business potential of intellectual properties：a machine learning-based predictive analysis of patent lifetime[J]. Computers & Industrial Engineering，2020，145：106544.
[37] 符川川，陈国华，袁勤俭. 基于机器学习的专利质量分析与分类预测研究——以区块链技术专利为例[J]. 现代情报，2021.
[38] 巩永强，王超，王锐，等. 复杂网络视角下的核心专利识别研究[J]. 情报理论与实践，2022(10)：103-113.
[39] SONG H，HOU J，ZHANG Y. The measurements and determinants of patent technological value：lifetime，strength，breadth，and dispersion from the technology diffusion perspective[J]. Journal of Informetrics，2023，17(1)：101370.
[40] MARIANI M S，MEDO，MATÚ，LAFOND F. Early identification of important patents：design and validation of citation network metrics[J]. Technological Forecasting and Social Change，2019：644－654.
[41] 于超，王涛，苏信宁，等. 引证关系视角下知识特质对“明星”专利形成的影响研究[J]. 软科学，2021，35(2)：19-25.
[42] 张克群，项星星，张婷，等. 识别高被引专利——基于稀有事件 Logit 与倾向得分匹配模型[J]. 图书馆论坛，2021，41(6)：67-74.
[43] 马荣康，陶雪蕾，李少敏，等. 知识元素网络搜索与突破性技术发明形成[J]. 科学学研究，2021，39(05)：794-804.
[44] 魏江，徐蕾. 知识网络双重嵌入、知识整合与集群企业创新能力[J]. 管理科学学报，2014，17(2)：34-47.
[45] 王巍，孙笑明，崔文田. 社会网络视角下的知识搜索和知识扩散研究述评与展望[J]. 科学学与科学技术管理，2020，41(6)：36-54.
[46] 李彦勇，林润辉. 知识网络结构、跨界搜索对组织突破性创新的影响：美国人工智能技术领域专利的分析[J]. 科技管理研究，2020，40(23)：204-212.
[47] 刘嘉明，闵超，严笑然. 合作网络和知识网络对 AI 企业专利创新的影响[J]. 图书馆论坛，2022(9)：132-144.
[48] 王巍，李德鸿，侯天雨，等. 多重网络视角下突破性技术创新的研究述评与展望[J]. 科学学与科

学技术管理，2022，43(10)：83-102.
[49] 任海英，李真. 基于输入输出型 SAO 网络的核心技术链识别方法研究——以量子计算领域为例[J]. 图书情报工作，2021，65(19)：117-129.
[50] 张欣，马瑞敏. 基于改进 PageRank 算法的核心专利发现研究[J]. 图书情报工作，2018，62(10)：106-115.
[51] LUNDBERG S M，ERION G，CHEN H，et al. From local explanations to global understanding with explainable AI for trees[J]. Nature Machine Intelligence，2020，2(1)：56-67.
[52] LUNDBERG S M，LEE S I. A unified approach to interpreting model predictions[C]. Advances in neural information processing systems，2017：4765-4774.
[53] CHAWLA N V，BOWYER K W，HALL L O，et al. SMOTE：synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research，2002，16(1)：321-357.

Prediction of High-value Patents by Incorporating Knowledge Network Embeddedness

REN Haiying，SUN Chuangchuang
（School of Economics and Management，Beijing University of Technology，Beijing 100124，China）

Abstract：Accurate prediction of patent value and early identification of patents with high value are of great significance to promote the cultivation of high-value patents and technical layout. Based on knowledge reorganization and patent invention creation process，the article selects and designs indicators of the knowledge network embeddedness to represent the association between sample patents′ knowledge and the prior knowledge of their domain. By having integrated the characteristics of innovation actors and patent application，a variety of machine learning models are built to predict the value of the patents at early stage of their application. The high-value patents in the field of neural networks are studied empirically，and the F1 value of the proposed high-value patent prediction model reaches 80%，and the prediction results are effective. Meanwhile，knowledge network embeddedness（especially PageRank and eigenvector centrality）plays an important role in predicting high-value patents.

Key words：high-value patent；value prediction；domain prior knowledge；knowledge network embeddedness；machine learning

（责任编辑：刘 凡）