

利用人工智能技术挖掘高层次创新人才 ——以专利数据为例

胡寅骏^{1,2} 王 茜¹ 徐旻昕¹ 刘晋元¹

(1. 上海市研发公共服务平台管理中心, 上海 200235; 2. 上海科技发展有限公司, 上海 200235)

摘要: 人才是国力强盛最重要的战略资源, 如何从各类数据资源中挖掘人才、专家已成为各国科技发展的一个重要课题。利用人工智能技术, 通过扩充检索关键词、判别专利领域、评价专利和创新人才等3个步骤对专利数据进行挖掘, 以人工智能领域、脑科学与神经科学领域的创新人才挖掘结果为例, 对挖掘结果的可视化和宏观分析揭示利用人工智能与专利数据挖掘创新人才的可行性和有用性。

关键词: 人工智能; 专利; 自然语言处理; 数据挖掘; 人才挖掘

中图分类号: G202; TP391

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2020.03.010

Mining High-level Innovation Talents with Artificial Intelligence Technology —A Case Study of Patent Data

HU Yinjun^{1,2}, WANG Qian¹, XU Minxin¹, LIU Jinyuan¹

(1. Shanghai R&D Public Service Platform, Shanghai 200235; 2. Shanghai Science and Technology Development Co. Ltd, Shanghai 200235)

Abstract: Talent is the most important strategic resource for strengthening the national power. How to mine talents and experts from all kinds of data resources has become an important topic for the development of science and technology in various countries. This paper uses the artificial intelligence (AI) technology to mine the patent data through the three processes of expanding search keywords, judging technical fields, evaluating pieces of patent and talents. Then it takes the results of mining innovative talents in artificial intelligence, brain science and neuroscience field as an example to reveal the feasibility and usefulness of using the AI and patent data mining for innovative talents through the visualization and macro analysis of mining results.

Keywords: artificial intelligence, patent, natural language processing, data mining, talent mining

0 引言

人才挖掘是数据挖掘的一个分支, 也是集统计学、机器学习、数据库、人工智能等学科为

一体的课题。目前, 人才挖掘已引起学者们的广泛关注。张腊梅等^[1]基于主题模型提出了一种挖掘微博中有专业知识背景人才的方法, 并验证了该方法的有效性; 杜晖等^[2]提出了一种基于文献

作者简介: 胡寅骏 (1986—), 男, 上海市研发公共服务平台管理中心硕士, 研究方向: 数据挖掘、自然语言处理; 王茜 (1984—), 女, 上海市研发公共服务平台管理中心工程师, 研究方向: 科技资源管理; 徐旻昕 (1986—), 男, 上海市研发公共服务平台管理中心工程师, 研究方向: 数据加工、数据管理; 刘晋元 (1978—), 男, 上海市研发公共服务平台管理中心硕士, 研究方向: 数据理论、数据技术 (通信作者)。

收稿时间: 2019年6月19日。

数据,利用知识网络 and 知识关联理论进行人才挖掘、构建领域专家库系统的方法;毛进等^[3]以论文文本为线索,提供了一种利用重叠K-Means聚类方法对研究领域的专家进行聚类的方法,并通过识别专家专长、聚集专家和可视化表示构建了研究领域专家图谱。田瑞强等^[4]提出了专利文献是体现技术创新的重要载体,并提供了一种基于专利数据的科技人才识别方法。

专利数据作为数据挖掘的数据源具有贴近产业、基于技术领域分类、代表技术创新前沿等独特的优势,然而对其的挖掘目前还远远不够。主要体现在技术手段单一导致搜索关键词覆盖不够全面、专利及人才领域判别脱离实际产业、对人才创新能力评价不够准确等方面。

本文拟以专利数据为分析对象,将人工智能技术应用于人才挖掘。具体来说,本文创新性地将自然语言处理中的词向量技术以及人工智能中的搜索排序算法、编辑距离技术、特征工程技术、机器学习等方法有机地结合起来,以专利数据作为数据源,通过收集相关专利文本、建立创新人才领域判断模型、评价专利与挖掘创新人才3个步骤对高层次人才进行挖掘,并以人工智能领域、脑科学与神经科学领域的实际挖掘过程和结果对数据挖掘成果进行分析和讨论。作为挖掘对象,本文所涉及的高层次人才是指特定产业领域中的尖端创新人才,其特征是具有较多的专利申请量且所申请专利质量较高。

1 数据与方法

可用于挖掘科创人才的数据来源十分广泛。科技论文数据能够通过被引用量等体现一个人才在相关领域的学术水平;获奖信息可以彰显学术人才的科研实力、学术地位及其突出贡献;人才的微博粉丝及微博被转发数量则体现了其社会影响力。如果仅对这些数据采用不同的方法对人才、专家进行了挖掘,势必产生侧重挖掘基础研究型学术人才、难以覆盖到产业中的创新人才的局限性,并且由于科技论文多采用的是学科分类体系,还会产生在确定专家所属技术领域时分

类粒度粗、信息更新滞后、自主可控性差等问题。例如,采用科技论文数据作为数据源,在作者未将“物联网”写入关键词的情况下只通过其论文关键词和学科分类进行判断,将难以判断其为物联网相关的人才。这时就需要通过技术手段为该人才增加一个“物联网”的标签。余丽等^[5]已经注意到此类局限性,并提出了利用深度学习对ACL(Association for Computational Linguistics)论文集自动标注细颗粒度标签的方法。

与上述这些数据相比,专利文本数据除了其本身具有贴近产业、基于技术领域分类、代表技术创新前沿等优势以外,其与人工智能技术的亲和力也是其他数据不可比拟的:专利文本数据包含丰富的半结构化数据,部分特征被提取后可以直接用来建模;专利是天然的平行语料库,大大降低了构建多语言模型的难度。因此,将人工智能的机器学习和自然语言处理技术与专利文本数据挖掘相结合既具有必要性也具有必然性。

使用人工智能技术对专利数据进行挖掘的优势可以归结为以下4个方面:一是运用机器学习技术,能够降低数据加工、标注的人工成本,提高处理专利文本数据的效率,并能够大幅降低目标专利漏检的可能性。二是运用自然语言处理技术,能够高效获得特定领域创新人才,并解决专利IPC分类与行业不完全吻合的问题。三是机器学习和自然语言处理技术相结合,能够实现机器自动评价、将相近领域自动归类、知识推理构建创新人才竞争合作关系网络等,深入挖掘数据背后的潜在关系。四是专利数据的数量和质量能够满足利用人工智能进行挖掘的基本条件,并有充足的数据来验证挖掘创新人才过程中人工智能模型的性能。

因此,本研究以中国国家知识产权局提供的103个国家1.3亿件(截至2019年8月)专利文本为数据源,使用人工智能技术对其进行数据挖掘。为确保本文的严谨性,在此对数据挖掘的概念进行如下界定:数据挖掘是数据分析的一种,数据分析是一个信息技术领域的概念,理论上任

何对数据进行计算、处理从而得出一些有益结论的过程都能够称作数据分析。从数据本身的复杂程度、对数据进行处理的复杂程度和深度以及从工程化的角度来看,可以将数据分析分为:基于抽样的数据统计、数据挖掘、大数据科学。对全球专利文本数据进行计算处理挖掘得到创新专家属于数据挖掘领域。

本文所使用的人工智能技术是指人工智能产业链中的技术架构层,包括通用技术(如机器学习、深度学习、自然处理技术、传感器)、算法模型(如支持向量机、聚类、各种神经网络模型等)、算法框架(如TensorFlow、Caffe、DMTK等)。

在专利挖掘创新人才的过程中,人工智能技术主要应用于以下几个方面:一是利用自然语言处理中的词向量技术来扩充检索领域专家相关数据的关键词,以保证所获得数据的覆盖程度;二是利用搜索排序算法替代传统的h指数,以抵消或降低在人才创新能力评价过程中由于自引用产生的引用得分,并提高被得分文献引用而产生的引用得分;三是通过编辑距离技术,解决文献中的笔误纠正和同一实体不同表述的标准化问题;四是利用特征工程技术和机器学习的方法,解决前述创新人才的技术领域分类问题。

2 结构与分析

对创新人才的挖掘有3个步骤:收集相关专利文本、建立创新人才领域判断模型、评价专利与挖掘创新人才。

2.1 收集相关专利文本

主要采用自然语言处理技术,其核心技术点

和分步骤是:(1)通过深度学习的词嵌入(Word embedding)技术^[9]和借助外部语义资源抽取关键词的上下位词来扩充收集专利用的关键词;(2)通过中英文版本的专利和点信息的计算获得关键词对应的英文单词;(3)整理中英文关键词并通过利用大数据技术中常见的工具“存储平台”和“计算引擎”对全球专利中包含上述关键词的文献进行抽取,构成专利池(图1)。其中,扩充关键词主要是针对传统专利检索行业中容易产生漏检的问题而设置的环节,如全球范围内摘要字段中包含“人工智能”只有7000余件,而事实上与人工智能有关的专利已远超60万件。这是由于专利的记述内容更偏向于具体的技术点,比起“人工智能”这样的较为抽象的概念,更倾向于出现“机器学习”“深度学习”,甚至是“机器学习”分支方向中一个具体的方法如“支持向量机”等词汇。

2.2 建立专利领域判别模型

专利领域判别的过程,其本质就是利用人工智能技术对专利进行分类的一个过程。由于从业务需求及前置分析中已确立领域及领域分支方向的名称,本研究中采用监督学习方法。领域判别的例子如图2所示:(1)通过IPC分类号粗略地对专利进行是否属于该领域的简单判断;(2)在考虑数据量、偏离度(Bias)等要素的基础上抽取种子专利;(3)按照事先确定好的分支方向对种子专利进行人工标注;(4)通过利用词嵌入将专利摘要转换为专利向量,并和人工标注结果结合采用SVM分类器建立专利领域判别模型;(5)利用GridSearch自动评测、调整模型,并应用于专利池中所有中文专利;(6)通过平行语料库的

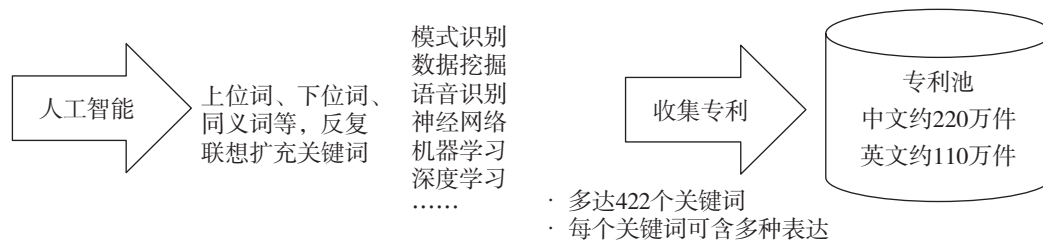


图1 收集相关专利的流程

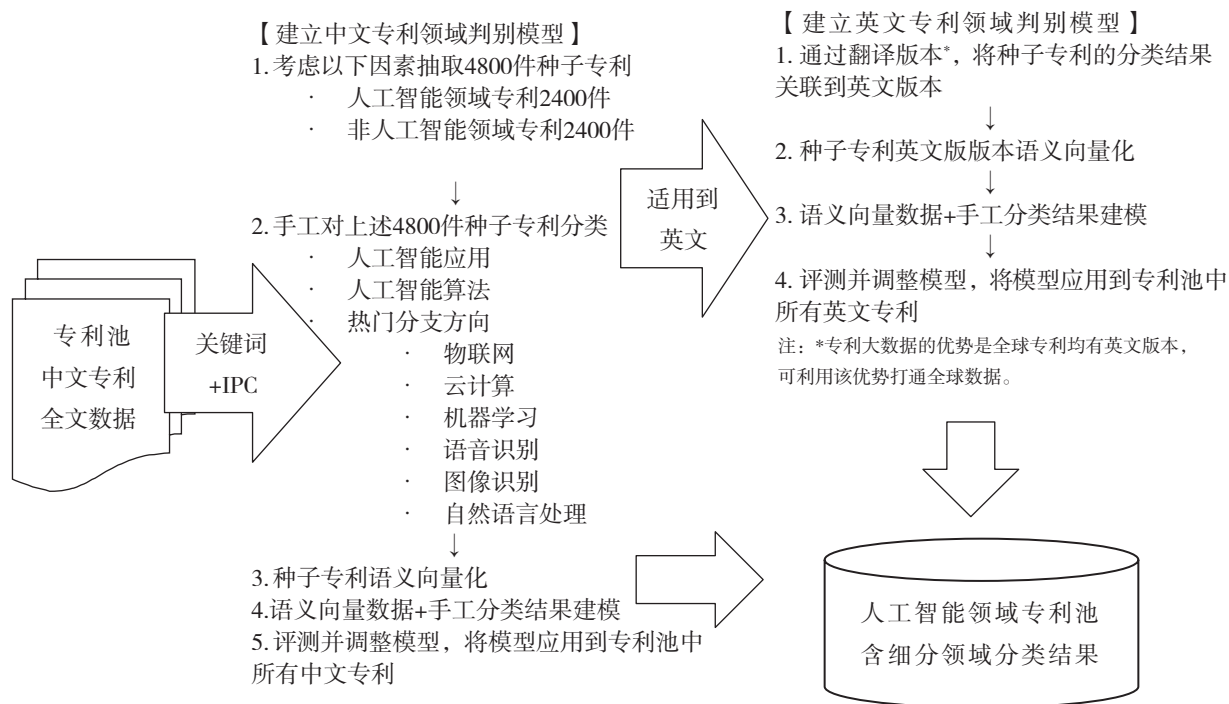


图2 人工智能领域判别流程

翻译版本，将模型自适应到英文专利分类的任务中。

利用词嵌入技术将专利向量化的公式如下：

$$V_p = \frac{1}{|K|} \sum_{k \in K} V_k \quad (1)$$

其中， V_p 为专利的向量表示， K 为通过计算TF-IDF获得的得分最高的前20个名词（含名词短语）， V_k 为单词 k 的向量表示。

2.3 评价专利与挖掘创新人才

本研究中通过对人才所拥有专利的评价结果对人才的创新能力进行评价并进行人才挖掘。其具体过程主要包括评价专利、计算人才创新能力得分和挖掘人才。

在评价专利的过程中，中文专利采用了国内知识产权行业最具权威的PatentRank算法（<http://patentrank.cnpat.com.cn/view/sample.htm>），从专利的法律价值、技术价值、经济价值3个维度进行分值化计算；英文专利则采用引证与被引证得分和多国申请数量得分两个维度的加权平均进行评价。其中，引证与被引证的得分主要参考了智能搜索引擎技术中经典的PageRank算法^[7]。式（2）

是多国申请的综合得分计算方法。

$$F(p) = \begin{cases} f(c) - 1 + b(c), & 1 \leq f(c) < 5 \\ 4 + b(c), & f(c) \geq 5 \end{cases} \quad (2)$$

其中， $F(p)$ 为专利多国申请数量得分； $f(c)$ 为五国专利局（中国、美国、欧洲专利局、日本、韩国）专利申请国数； $b(c)$ 代表五国专利局以外是否有申请别国的专利（有申请为1，无申请则为0）。

人才创新能力得分指的是将评价后的专利按照<申请人（即工作单位），人才>的维度进行聚合后，通过计算获得的人才专利得分均值。本研究中，考虑到人才的专利过少会导致数据偏离（误判人才领域或人才创新能力得分不准确），事先通过专利数量阈值对人才进行了初步筛选。再根据人才的创新能力得分由高到低进行排序，并按照需求挑选出相应人数的创新人才。

3 挖掘结果讨论

为进一步检验利用人工智能技术挖掘创新人才的有效性，本文以利用专利数据挖掘人工智能领域、脑科学与神经科学领域的创新人才为例，

应用所提出的方法对创新人才进行挖掘，并对挖掘结果进行分析。

3.1 各国研发创新的侧重点

为更好地找到各国在人工智能领域、脑科学与神经科学领域研发创新中的侧重点，本研究将利用人工智能技术中的自然语言处理理解技术对论文摘要进行语义向量化，并利用机器学习的方法对专利进行分类。所述方法较传统学科分类、IPC分类的优势在于：可以灵活地制定更贴近产业、实际应用的分类方法，并且利用机器学习模型的分类过程是全自动的。

图3与图4展示了利用上述方法对主要国家人工智能领域、脑科学与神经科学领域的分类结果及其可视化展现，包括原理（或算法）、应用占比（用气泡图展现）及各分支方向占比（用雷达图展现）。其中，人工智能领域的分支方向包括物联网、云计算、机器学习、自然语言处理、图像识别、语音识别；脑科学与神经科学领域的分支方向包括医疗设备、医药、神经网络、阿尔茨海默症、帕金森、VR/AR、脑电波、记忆力与智力发育、ADHD。由于领域判别采用了有监督学习的SVM而非聚类方法，其占比为项目在该领域所有专利中的占比，且各项目占比总和未必等于1（如人工智能领域在中国申请的专利中算法和应用占比分别为57%和54%）。

图3中的气泡图显示了人工智能领域主要创新国家/地区的基本信息和该国该地区是偏向应

用方面还是偏向算法（理论方面）。气泡的大小代表专利数量。可以看到，从专利总数来讲，美国>中国>日本>欧专局。纵向的y轴坐标代表该国家/地区人工智能专利中应用型专利的所占比重。从中可见，中国专利占比最高，也就是说在中国人工智能专利中，有50%~60%都是应用型的；美国专利占比最低，其应用型专利仅占其人工智能专利总数的30%~40%。横向的x轴坐标代表该国家/地区人工智能专利中算法型专利的所占比重。从中可见，美国最靠右，数值最高，有接近80%都是算法型专利；而中国最靠左，显示算法型专利在中国人工智能专利中仅占60%左右。这个气泡图直观地展示了在人工智能领域中国申请的专利偏向于应用型，这也印证了德勤在2018年发布的《中国人工智能产业白皮书》中主要发现的“中国企业价值链布局侧重技术层和应用层，对需要长周期的基础层关注度较小”^[8]；与之形成对比的是美国申请的专利偏向于算法理论。图3中的雷达图部分则显示了美国在云计算方向上的霸主地位（卡方检验， $p<0.05$ ），而中国在物联网方向略有优势（卡方检验， $p<0.05$ ）。

图4中的气泡图显示了脑科学与神经科学领域主要创新国家的基本信息和该国是偏向应用方面还是技术原理（理论方面）。与人工智能气泡图的表示方法相同，气泡的大小代表专利数量。从图4中可以直观地看出，从脑科学与神经科学领

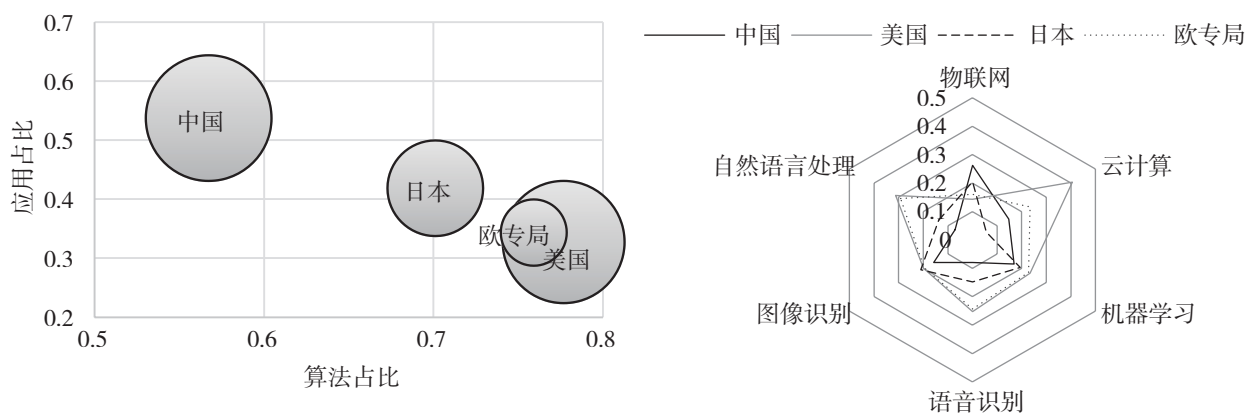


图3 人工智能领域主要创新国家/地区在领域分支方向上的侧重点

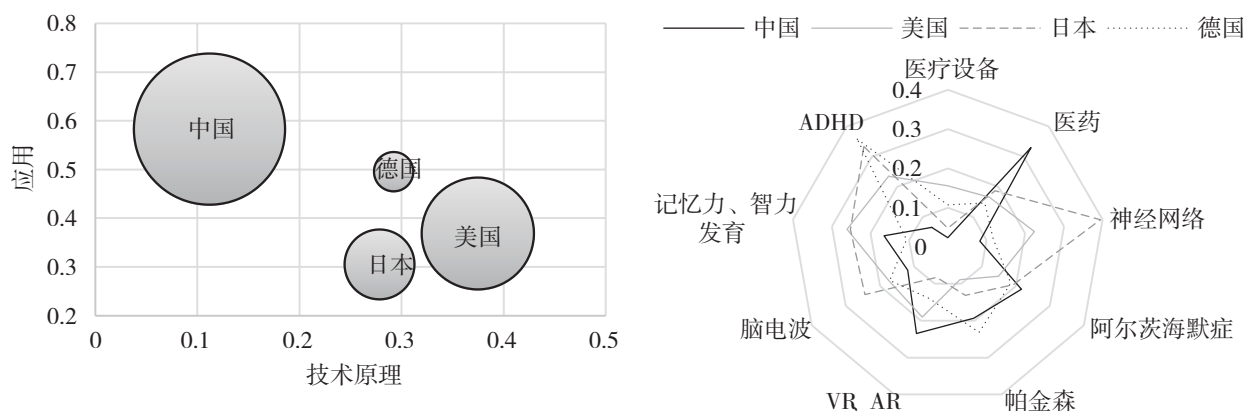


图4 脑科学与神经科学领域主要创新国家/地区在领域分支方向上的侧重点

域的专利总数来讲,中国>美国>日本>德国。纵向的y轴坐标代表该国家脑科学与神经科学领域的专利总数中应用型专利的所占比重。从中可见,中国的占比最高,也就是说在中国脑科学与神经科学领域专利中,有60%都是应用型的;日本的占比最低,应用型专利仅占30%左右。横向的x轴坐标代表该国家脑科学与神经科学领域技术原理专利的所占比重。从中可见,美国数值最高,有接近40%都是技术原理型专利;而中国数值最低,技术原理型专利仅占比10%左右。这个气泡图直观地展示了在脑科学与神经科学领域,依然是中国偏向于应用而美国侧重于技术原理(理论方面)。雷达图则进一步展现了中国在医药和VR/AR方向上具有优势(卡方检验, $p<0.05$),美国较为均衡,德国侧重于帕金森、阿尔茨海默症、ADHD方向上的研究创新(卡方检验, $p<0.05$),日本则在脑科学的人工智能神经网络上有突出表现(卡方检验, $p<0.05$)。

3.2 创新人才分布

在对创新人才专利进行聚合以获得人才创新能力得分的过程中,存在申请人(即发明人工作单位)表述不统一(如“鸿海精密工业股份有限公司”,由于笔误收入数据库时被录入为“鸭海精密工业股份有限公司”)或人名表述不一(如被称为“机器学习之父”的加州大学伯克利分校教授“Michael I. Jordan”或因专利代理人不同而被写成“Michael Jordan”)。在本研究中,利用人工智能技术中在姓名和实体命名(Named

Entity)匹配任务中有较大优势的Jaro-Winkler Distance算法^[9-10],事先将不同表述的申请人进行标准化处理。

为获得高层次的创新人才,首先对中文专利和英文专利分别采用国内专利评价权威的PatentRank算法和搜索引擎技术中经典的PageRank算法对专利质量进行评价,再通过计算创新人才专利进行聚合后的均值来获得创新人才的得分。

表1和表2分别显示了人工智能领域、脑科学领域(在本领域至少有2篇专利并且人才创新能力评价得分大于0的)的创新人才分布。通过统计数据可以发现:就专利数据所体现的特定领域创新人才而言,在人工智能领域,中国的人才占比位居全球第一,比第二位的美国多出8个百分点;在脑科学与神经科学领域,美国的人才占比则比中国高出3个百分点。

在创新人才人均申请专利方面,在人工智能领域,中国为0.88件/人,美国为1.01件/人;在脑科学与神经科学领域,中国为1.93件/人,美国为0.91件/人。值得关注的是,较其他国家人均专利申请而言,中国在脑科学与神经科学领域的创新人才人均专利申请量较多,接近2件/人。这是由于在脑科学领域中有如阿尔茨海默症等病症相关的分支方向,而在中国申请的专利中含有大量针对这些症状的中药专利申请。遗憾的是,这些专利多为质量不高的“量产”专利。李慧等^[11]指出中药专利质量有待提升的观点,并提

出了一系列的应对措施和方案。

综上所述，在挖掘创新人才、构建创新人才库的过程中，对专利、人才创新能力的评价是必不可少的一个环节。而这个评价过程通过用算法实现自动化计算来替代人工评价，所呈现的结果会更加客观，更贴近真实。

3.3 人工智能技术标注人才产业领域标签

本研究旨在利用人工智能技术挖掘创新人才，其核心技术点在于：利用深度学习技术来扩展检索关键词，利用机器学习模型来判别专利所属领域及其细分方向，利用自然语言处理技术对人才进行归一化，并利用搜索排序技术对专利、创新人才进行评价。人工智能的人才挖掘部分结果示例（排名相对靠前、工作单位不重复且能在互联网上找到对比信息的示例）如表 3，通过利

用人工智能的技术手段，可以突破传统学科分类体系带来的局限，为人才打上更贴合产业的领域标签。特别是在全球人才竞争加剧的当今，中国更需要如原松下电器产业株式会社（2019 年起已加入北京澎思智能科技有限公司）的申省梅团队这样的全球范围高层次创新人才加盟。而通过人工智能技术挖掘创新人才，能帮助国内高校、科研院所和科技型企业更快、更准确地找到这些细分领域的高层次人才。

4 结语

本研究讨论了一种融合人工智能技术和专利数据进行创新人才挖掘的解决方案。以专利数据为例，描述了半自动构建创新人才平台过程中遇到的实际问题，并运用人工智能、自然语言处理

表 1 人工智能领域创新人才分布

国家	人才数/万人	占比/%
中国	24.5	36
美国	18.7	28
日本	7.5	11
韩国	4.5	7
德国	2.4	3
加拿大	2.0	3
其他	7.8	12

表 2 脑科学与神经科学领域创新人才分布

国家	人才数/千人	占比/%
美国	15.9	22
中国	13.7	19
澳大利亚	5.1	7
韩国	4.9	7
加拿大	4.7	6
日本	4.4	6
其他	24.3	33

表 3 人工智能领域部分高层次创新人才示例

工作单位及创新人才	领域标签	专利质量得分（排名）
华为技术有限公司：郑志彬	应用型、算法型、物联网	63.50（1）
清华大学：徐恪	应用型、算法型、云计算	61.81（10）
中国科学院计算技术研究所：高文	算法型、自然语言处理	58.64（52）
松下电器产业株式会社：申省梅	机器学习、图像识别	58.25（69）
微软公司：黄学东	算法型、语音识别	55.34（271）

的技术解决这些问题。通过对挖掘结果的分析,验证了所提方法不仅能够压缩挖掘创新人才的成本,而且能够利用挖掘结果进行宏观分析,达到助力相关机构进行科研管理、技术转化的目的。

首先,本研究创新地将自然语言处理中的词向量技术、人工智能中的搜索排序算法、编辑距离技术、特征工程技术、机器学习等多种人工智能技术有机结合,将其应用于人才挖掘过程中,有效地解决了传统数据挖掘技术中相关搜索关键词覆盖不全面、人才领域判别脱离实际产业、人才创新能力评价不准确、用于验证挖掘手法效果的数据不够充足等问题。其次,本研究创新地以专利文本数据为例进行挖掘,进一步论证了机器学习和自然语言处理技术对专利文本数据进行数据挖掘有着其他技术不可比拟的优势,并解决了以往人才挖掘研究多侧重于基础研究学术方面而忽视或无法实现对产业中的创新人才进行挖掘的问题。最后,本研究利用深度学习技术和机器学习模型判别专利所属领域及其细分方向、利用自然语言处理技术对人才进行归一化、利用搜索排序技术对专利和创新人才进行评价,论证了该方法具有突破传统学科分类体系的局限,为人才打上更贴合产业的领域标签的优势,能够有效助力产业和人才的精准匹配。

今后将在以下两个方面对研究成果进行进一步优化:一是对模型进行改进,PageRank算法存在倾向于发表时间久的局限性^[12],可以考虑改进评价方法。如付常雷等^[13]提到以学术知识图谱为基础,根据知识点的“热度”“新颖度”“权威度”三维指标来筛选“创新种子”。二是将专利文本数据与论文文献、获奖头衔、海外交流、同行评价等多维度数据相结合,建立大规模科技创新人才图谱,助力全球高层次科技专家信息平台的构建。

参考文献

- [1] 张腊梅,黄威靖,陈薇,等. EMTM: 微博中与主题相关的专家挖掘方法[J]. 计算机研究与发展, 2015, 52(11): 2517-2526.
- [2] 杜晖,邱均平. 领域专家库系统构建研究[J]. 情报学报, 2014(10):1022-1031.
- [3] 毛进,李纲. 一种基于OKM的研究领域专家图谱构建方法[J]. 图书情报工作, 2014(14):34-40.
- [4] 田瑞强,刘伊颖,姚长青,等. 基于专利文献的创新科技人才识别研究[J]. 情报杂志, 2018(8):71-77.
- [5] 余丽,钱力,付常雷,等. 基于深度学习的文本中细粒度知识元抽取方法研究[J]. 数据分析与知识发现, 2019, 3(1): 38-45.
- [6] MIKOLOV T, SUTSKEVER I, KAI C, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.
- [7] PAGE L. The PageRank citation ranking: Bringing order to the Web[J]. Stanford Digital Libraries Working Paper, 1998, 9(1):1-14.
- [8] 中国人工智能产业白皮书[EB/OL]. (2018-11-13) [2019-08-26]. <https://www.deloitte.com/content/dam/Deloitte/cn/Documents/innovation/deloitte-cn-innovation-ai-whitepaper-zh-181126.pdf>.
- [9] COHEN W W, RAVIKUMAR P, FIENBERG S E. A comparison of string metrics for matching names and records[C]. Kdd Workshop on Data Cleaning & Object Consolidation, 2008.
- [10] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.
- [11] 李慧,冯晓慧. 中药专利质量的界定与提升[J]. 中草药, 2016, 47(16):2960-2964.
- [12] 段庆峰,朱东华,汪雪锋. 基于改进PageRank算法的引文文献排序方法[J]. 情报理论与实践, 2012, 35(1): 115-119.
- [13] 付常雷,钱力,张华平,等. 基于深度学习的创新主题智能挖掘算法研究[J]. 数据分析与知识发现, 2019, 3(1):50-58.