DOI 10.15302/J-SSCAE-2023.07.005

面向智慧税务的大数据知识工程技术及应用

郑庆华[1,2]，师斌[2,3*]，董博[2,4]

（1. 西安交通大学，西安 710049；2. 陕西省大数据知识工程重点实验室，西安 710049；3. 西安交通大学计算机科学与技术学院，西安 710049；4. 西安交通大学继续教育学院，西安 710049）

摘要：税收在国家治理中发挥着基础性、支撑性作用，实现智慧税务是政府在数字时代转型的必然要求，因而梳理智慧税务中的关键问题并探讨发展思路兼具理论研究与实践应用价值。本文分析了我国智慧税务领域的发展现状及面临挑战，提出了以“数据知识化、知识体系化、知识可推理”为核心的大数据知识工程解决方案，构建了由知识源层、知识提取层、知识图谱层、知识推理层、应用层组成的“五层”技术架构；结合大数据知识工程在智慧税务领域中的代表性应用案例，如知识驱动的税收优惠计算、可解释的税收风险识别、税收政策智能化决策支持、智慧问税，探讨了所提方案的局限性并论述了进一步的研究方向。从数据、技术、生态三方面出发，形成了规范涉税数据、健全国家数据共享 / 开放 / 保障体系，融合并更新信息学科成果，完善面向智慧税务的大数据知识工程应用系统，推动大数据知识工程技术的标准建设与人才培养等发展建议，以期为基于大数据知识工程的智慧税务高质量发展研究提供参考。

关键词：智慧税务；知识工程；大数据；知识图谱；知识推理

中图分类号：TP319　　**文献标识码：**A

Technologies and Applications of Big Data Knowledge Engineering for Smart Taxation Systems

Zheng Qinghua[1,2], Shi Bin[2,3*], Dong Bo[2,4]

(1. Xi’an Jiaotong University, Xi’an 710049, China; 2. Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi’an 710049, China; 3. School of Computer Science and Technology, Xi’an Jiaotong University, Xi’an 710049, China; 4. School of Continuing Education, Xi’an Jiaotong University, Xi’an 710049, China)

Abstract: Taxation is vital for national governance, and the digital transformation of governments necessitates smart taxation. Therefore, analyzing the key issues and exploring the development ideas for smart taxation is of both theoretical and practical values. In this study, following an analysis of the development status and challenges facing China’s intelligent taxation field, we proposed a big data knowledge engineering approach that emphasizes data knowledgeization, knowledge systematization, and knowledge reasonability, and developed a five-layer technical architecture that consists of knowledge sources, knowledge extraction, knowledge mapping, knowledge reasoning, and application layers. After elaborating the representative application scenarios including knowledge-driven tax preference calculation, interpretable tax risk identification, intelligent decision support for tax policies, and smart tax questioning, we investigated the limitations of the proposed approach and further discussed the directions for future research. Furthermore, we proposed the following development suggestions in terms of data, technology, and ecology: (1) standardizing tax-related information and improving the national data sharing, opening, and guarantee system; (2) integrating the achievements of various information

收稿日期：2022-07-30；**修回日期：**2022-09-29

通讯作者：*师斌，西安交通大学计算机科学与技术学院副教授，研究方向为知识工程、云计算；E-mail: shibin@xjtu.edu.cn

资助项目：国家自然科学基金项目(62250009, 61721002)；中国工程科技知识中心项目(CKCEST-2022-1-40)

本刊网址：www.engineering.org.cn/ch/journal/sscae

disciplines and improving the application system of big data knowledge engineering for smart taxation; and (3) promoting talent training and the development of technical standards for big data knowledge engineering.
Keywords: smart taxation; knowledge engineering; big data; knowledge graph; knowledge reasoning

一、前言

知识是人工智能（AI）的动力[1]，基于知识工程方法开展机器学习、拓展开放性智能应用的实践由来已久。随着互联网、大数据技术的发展，大数据知识工程这一新范式应运而生，旨在运用知识工程理念与方法，从大数据中获取、验证、表征蕴含的知识，进行知识推理和应用，形成解决大数据背景下实际工程问题的方法[2]。大数据知识工程的应用，一方面使得知识的大规模获取成为可能，解决了传统知识工程面临的人工成本偏高、专家经验局限等问题；另一方面将海量低质的大数据转化为人类可理解，机器可表示、可计算的结构化知识体系，降低了机器学习对训练样本的依赖度，增强了泛化能力，为解决大数据背景下场景动态、规则与边界未知的复杂工程问题提供了可能性[3]。有观点认为，大数据知识工程是信息化迈向智能化的必由之路[4]，大数据知识工程作为决定未来经济发展的颠覆性技术之一，产业前景广阔[5]。

目前，教育、政务、税务、交通、医疗、金融等行业，经过数十年的信息化建设积累了大量数据，都面临如何将大数据转化为结构化知识体系并进行应用的问题，信息化到智能化的共性需求强烈。国内外机构致力于大数据知识工程研究，如国际商业机器公司的Watson商业AI认知系统，谷歌公司用于增强搜索引擎能力的知识图谱技术，微软公司的大规模概念知识图谱Probase[6]，卡内基梅隆大学的实时知识库学习系统NELL[7]；中国科学院数学与系统科学研究院的知识工程语言设计[8]，清华大学的科研人员社会网络挖掘系统ArnetMiner[9]，西安交通大学的在线教育开放知识获取系统[10]，中国科学院计算技术研究所的的开放网络知识库[11]。

值得指出的是，虽然大数据知识工程技术的领域应用发展迅速，但是在智慧税务方向的应用仍处于前期探索的起步阶段。智慧税务是AI技术与税收治理深度融合的交叉方向，现实需求与社会效益极为突出。自1994年实施“金税工程”以来，我国积累了工商、税务等领域的数万亿条数据，这些数据位置分散、结构无序、关联复杂、动态变化；依靠传统的专家系统、机器学习方法，很难刻画涉税数据中的本质规律，面临着无法动态演化、不可解释的问题。在智慧税务领域中引入大数据知识工程的理论与方法，将为解决上述问题奠定坚实基础，对提升税收治理水平具有重要意义。在政府治理向数字时代转型、“互联网+”战略积极推进的背景下，探索面向智慧税务的大数据知识工程技术兼具理论意义和实践价值。

本文以面向智慧税务的大数据知识工程技术为研究对象，着重从方法论角度剖析并阐述大数据知识工程在智慧税务中的实践思路。梳理智慧税务需求及面临挑战，凝练大数据知识工程解决方案涉及的关键问题；提出面向智慧税务的大数据知识工程定义、架构与关键技术；结合大数据知识工程在智慧税务中的实践成果与应用案例，探讨所提方案的局限性以及后续研究方向。从数据、技术、生态等方面形成发展建议，以期为加速实现基于大数据知识工程的智慧税务建设提供参考。

二、智慧税务的需求、挑战与大数据知识工程解决方案

（一）智慧税务的内涵和需求

税收是国家最为重要的收入来源，发挥着组织财政收入、调节经济活动、监督经济运行等作用[12]，在支撑经济社会高质量发展方面作用突出。高效征管税收是政府治理能力的标志，实现智慧税务是政府在数字时代转型的必然要求。《关于进一步深化税收征管改革的意见》（2021年）要求，“十四五”时期基本建成功能强大的智慧税务体系，形成一流的智能化行政应用系统，全面提高税务执法、服务、监管等能力[13]。建设和实施智慧税务，需要在加强税务部门税收监管能力、优化纳税服务的基础上，提供智能化的决策管理能力[14]。

从加强税收监管的需求角度看，智慧税务有助于精准监管和精确执法，充分利用税收大数据等现代化技术来识别潜在的涉税违规企业，支持税务部

门有效管控企业的违法风险，降低偷税漏税带来的财政损失，同时避免打扰诚信纳税人。对于涉税违规企业，支持形成识别结果和相关的证据链，维护税务执法过程的可信性、公信力和执行力。

从优化纳税服务的需求角度看，智慧税务将为纳税人提供智能化的办税系统，支持税收征管方式“收税”“报税”“算税”的逐步升级。例如，自动匹配纳税人数据和相关政策法规、自动计算税额、自动填写申报，可显著减少纳税人的时间及心理成本，保障各类税收政策的应享尽享[15]，利于营造省心、舒心、放心的办税缴费环境。可针对纳税人的个性化需求提供精准化服务，如涉税问题的智能问答、政策法规的精准推送，让纳税人享受专业、便捷的涉税服务。

从赋能决策管理的需求角度看，智慧税务将在支持政府决策和政策制定方面发挥基础性作用。在现阶段，通过税收政策法规与纳税人数据的精准匹配等技术，实现了对税收政策效果的细粒度预测与量化，直接辅助决策层面的管理部署。随着数字政府的全面建设，智慧税务技术水平的进一步提升，预测各类经济主体跟随政策的行为变化有望成为现实。针对拟发布政策进行潜在影响的综合分析，从税收政策层面促进社会的正向发展，使税收的国家财政收入、宏观经济管理“调节器”作用发挥更为充分。

（二）智慧税务面临的挑战

经过多年的税务信息化建设，我国税务治理成效显著；尤其在“金税工程”三期“以票控税”理念的驱动下，形成了运行稳健、功能完备的税收信息体系，有效整合了税务部门的海量涉税数据。这是进一步探讨智慧税务的基础前提。不可忽视的是，从涉税数据的有效整合到智慧利用依然存在鸿沟，实现智慧税务挑战极大。

1. 海量、多源数据关联融合

为了使计算机程序全面“理解”涉税业务，需要完成数据碎片化向体系化的转变，这一过程存在“散”“杂”“乱”难题。“散”指数据信息空间分散、内容片面，“杂”表现为模态多样、关联复杂，“乱”体现在跨域交叉、线索凌乱。例如，38、38 度、用电 38 度，一个月用电 38 度、建国酒店一个月用电 38 度、酒店行业平均用电量大于 1000 度等数据，即具有上述特点。然而，随着原始数据中融入更多的数据和知识，相关的可解释性、可理解性逐步增强，智能化水平稳步提升。多源数据从封闭式转向开放式，实现数据的关联与融合，是智慧税务面临的首要挑战。

2. 算法和规则的自动演化、学习及更新

传统上数据驱动的机器学习方法，多针对场景单一且规则事先明确的问题（如语音识别、人脸识别），也需依赖数据同分布假设。智慧税务则面临数据模式变化、数据特征变化、任务目标变化等动态场景，如“偷逃骗税”行为已经从个体化转向团伙化、专业化、隐蔽化，而现有的风险防控系统较多依赖专家制定规则及指标，更新迭代周期长，对新型违法手段的适应性不佳。对于税收政策的自动匹配功能，关键在于税收政策与纳税人的双向精准匹配。在新型冠状病毒感染疫情发生后，国家和地方税务部门及时发布的税收优惠政策超过 100 条；已有算法如何适应快速变化的政策法规并反映纳税人经营变化情况，也是技术挑战。实现算法与规则的自动演化、学习及更新，是智慧税务能力建设的重要内容。

3. 破解可解释性难题

AI 技术应用面临可解释性难题，深度学习自带的“端到端”黑匣子特性导致过程和结果都是不可解释的。税务服务、监管和执法建立在严格的法律条文和道德规范之上，必然要求过程透明、结果符合逻辑认知。以“偷逃骗税”识别模型为例，如果只能根据深度学习结果来认定“偷逃骗税”行为，而没有出具证据链，则认定结果难以解释，也就导致执法过程质疑多、求证难，直接影响可信性、公信力和执行力。破解可解释性难题是实现智慧税务有效应用的重要挑战。

（三）大数据知识工程解决方案及关键问题

大数据知识工程方法可从海量数据中获取知识、表示知识，基于知识进行推理计算，这是破解智慧税务建设和应用挑战的有效手段。数据 – 信息 – 知识 – 智慧（DIKW）模型描述了从数据、信息、知识到智慧的不断增值过程[16]，大数据知识工程有待研究的问题也可纳入 DIKW 模型。基于此思路，本文识别出智慧税务大数据知识工程解决方案中的 3 个关键问题（见图 1）。

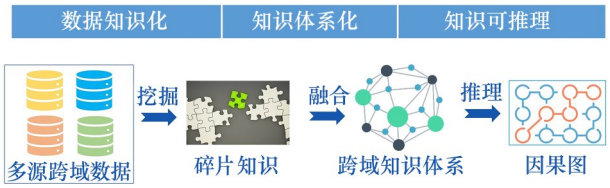

图 1　智慧税务大数据知识工程的关键问题

1. 数据知识化

非结构化数据到结构化数据的转变过程包括：从海量涉税数据中抽取碎片知识，通过指代消解、实体消歧、实体统一来进行碎片知识的量质转换，经由表征学习方法将碎片知识映射到低维子空间。这个过程对“散”“杂”“乱”的多源碎片化知识进行抽取及表征，是智慧税务大数据知识工程的第1个关键问题。

2. 知识体系化

结构化知识到体系化知识的转变过程包括：涉税的碎片知识通过知识推理及更新而不断融入新知识，挖掘知识前后因果关系，对已表征的碎片知识进行非线性融合。这个过程对已表征的多源碎片化知识经非线性融合后生成体系化知识，是智慧税务大数据知识工程的第2个关键问题。

3. 知识可推理

传统的机器推理在场景静态、规则明确、边界确定等问题上表现较好[17,18]。智慧税务场景中的问题多为动态、规则事先未知、边界未知，因而采用传统的机器学习方法存在可解释性[19]、组合爆炸[20]等不足。在大数据知识工程中融合符号推理和深度学习[21]，才能实现知识体系到智慧化的转变。利用体系化知识进行可解释的机器推理并用于税务场景复杂问题求解，是智慧税务大数据知识工程的第3个关键问题。

三、面向智慧税务的大数据知识工程定义与架构

面向智慧税务的大数据知识工程定义为：基于海量数据、计算能力、群智计算等大数据技术，利用知识工程的理念和方法，从政策法规、报表、发票、工商、海关等数亿条规模的税务相关数据中获取蕴含的法规、经济、行业等知识；对提炼的知识进行推理和应用，解决税务领域面临的智能化决策支撑、可解释的税收监管等关键难题，全面提高税务治理的智能化水平。大数据知识工程融合了大数据、知识工程等技术，具有诸多优势。从知识的角度看，自动获取并融合跨源、跨域、跨模态大数据中蕴含的知识，减少传统方法所需的高额人工成本，破解碎片化数据的片面性问题。从模型的角度看，支持实现“数据驱动+知识引导”的新型机器学习模型，以传统的数据驱动模型训练为基础，将符号化知识嵌入到数值化表示的深度学习模型中，提高模型的可解释性，降低模型对特定规则和样本的依赖性。

大数据知识工程在智慧税务系统中的应用架构包含五方面：知识源层、知识提取层、知识图谱层、知识推理层、应用层（见图2）；知识源层、知识提取层解决数据知识化的问题，知识图谱层解决知识体系化的问题，知识推理层、应用层解决知识可推理的问题。具体而言，知识源层汇集税收、金融、海关、工商等多源和多模态的数据，进行预处理并传输；知识提取层利用实体抽取、关系抽取等技术，从海量、异构知识源中抽取碎片知识；知识图谱层重在挖掘碎片知识之间的内在联系，构建单一知识图谱并将不同知识图谱融合为体系化的知识森林；知识推理层借助知识图谱推理技术，在知识图谱上进行推理计算与图谱演化（更新）；应用层基于以上技术实现智慧税务应用，如智慧问税、智慧管税、智慧办税等。

（一）知识源层

知识源层主要汇集了税收、金融、海关等多源大数据。鉴于数据的海量属性，难以进行数据汇总和统一处理，需要采取分布式计算方式来处理数据。大数据批处理架构以Hadoop为代表，形成了以分布式文件系统、分布式计算框架、分布式数据库为核心的完整生态系统；涵盖海量数据分布式存储、管理与计算的完整流程，为税务大数据的分布式高效存储与处理提供了基础性支持。

（二）知识提取层

传统的知识工程依赖专家经验获取知识，而大数据知识工程侧重在多源、海量、异构的税收大数据中自动化提取碎片知识。知识提取层面向税务领域的结构化 / 非结构化多模态数据，如交易数据、

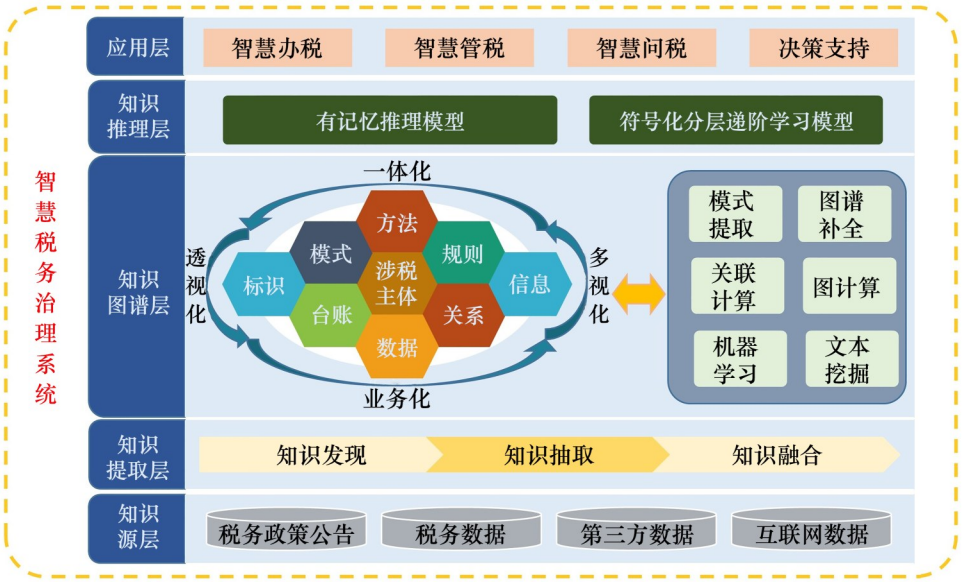

图2 面向智慧税务的大数据知识工程技术架构

发票数据、关联新闻报道、税务法规文件等；应用实体抽取、关系抽取、知识融合等技术，完成碎片知识的自动化提取。

1. 实体抽取

实体抽取又称命名实体识别，用于从文本中抽取实体信息元素[22]，如行业属性、纳税人属性、税种信息、涉税约束等；完成质量将极大影响后续关系抽取任务的效率与准确性，是知识抽取中的基础性任务。鉴于税务实体类型的多样性，可采用深度“端到端”实体抽取模型构建复杂实体自动抽取方案[23]，据此精准抽取税务实体。

2. 关系抽取

关系抽取是知识提取的重要子任务，从非结构化数据中抽取出两个或多个实体之间的语义关系[24]。关系抽取和实体抽取密切相关，通常在识别出数据中的实体后再抽取实体之间可能存在的关系。为了降低人工构建特征的困难，可利用基于深度学习的关系抽取方法并借助弱监督学习，减少对数据标注的需求[25]，从而降低海量税务数据关系抽取的人工成本。

3. 知识融合

知识融合主要解决税务大数据的多源、异构特性导致的共指难题（即实例名与指代实体不匹配），从而实现不同实体之间的信息交互，确保知识的关联与融合[26]。按照应用场景的不同，分为基于自然语言处理的匹配、基于本体结构的匹配、基于实例的机器学习等方法。鉴于税务场景数据的跨源、跨域、跨模态特性，面向智慧税务的知识融合方法需综合使用多种方法才能提高融合结果的质量。

（三）知识图谱层

知识图谱层旨在将提取的碎片知识转换为可被计算机直接表征与处理的知识图谱[27]。面向智慧税务的大数据知识工程需要构建3类知识图谱。

1. 税法知识图谱

税法知识图谱借助实体识别、关系抽取等技术，将税务政策法规、经济知识、金融知识等专家知识自动转化为知识图谱的形式，构建税务领域知识全景；用于指导其他图谱构建以及图谱上的推理计算，对部署可解释性的智慧税务应用具有重要价值。

2. 案例知识图谱

案例知识图谱将涉税案例表示为由实体（如企业、个人、银行），各种时序关系（如交易、支付、缴税）构成的动态知识网络[28]，为自动识别与抽取特定场景（如“偷逃骗税”行为模式）提供支持。需要构建满足图谱定义的数据模型，明确数据表征范围和具体形式；在实体层面，如案例类实体、相关的一般性实体，对关键属性进行规范化表征；在关系层面，对各类实体之间的关联关系进行梳理，形成明确定义。

3. 纳税人知识图谱

纳税人知识图谱利用知识森林技术[4]，将蕴含在跨源、跨域、跨模态税务大数据中的碎片知识融合为知识图谱，实现碎片知识的体系化表达。基于知识森林构建的纳税人知识图谱，适应层次化、主题化、跨领域知识表示的需求，可克服税务大数据知识碎片化、知识“散”“杂”“乱”难题。

（四）知识推理层

知识推理指利用已掌握的知识求解问题或推演新知识的过程。知识推理层借助税务知识图谱进行推理计算，支持构建智慧税务应用，可克服现有深度学习面临的可解释性难题。

1. 有记忆推理模型

有记忆推理模型借鉴人脑处理信息过载问题的方式，提高神经网络的信息处理能力；引入额外的外部记忆模块来优化神经网络的记忆结构，提高神经网络存储信息的容量[29]。为了处理税收场景中的复杂数据特征与信息，运用以长短期记忆网络为代表的循环神经网络、神经图灵机、可微计算机等，形成稳健的智慧税务知识推理机制。

2. 符号化分层递阶学习模型

符号化分层递阶学习模型的核心在于“分层递阶可控+符号化知识驱动”，即应用介科学理论，将多层次、多尺度、动态时空关联的复杂数据系统划分为若干介区域，进而形成分层递阶结构；结合各个介区域的功能与状态特点，构建内嵌物理学或社会学知识的符号化控制机制（如常识、规则）[30]。在智慧税务应用构建过程中，按照税务领域知识属性，合理分解深度学习模型，实现分层可控制、人工可干预、结果可回溯，从而破解现有深度学习模型因不可分解性导致的可解释性问题。

（五）应用层

在智慧税务大数据知识工程的技术框架中，智慧税务场景主要有办税、管税、决策支持等基本应用。

1. 智慧办税

智慧税务系统以海量的涉税数据为基础。① 主动识别输入端办税人员或办税企业信息，借助系统的推理层进行计算推理，自动地将个人或企业的基本信息、涉税数据与当前税收政策进行适配，实现税务申报的自动填写。② 快速精准地推送税收优惠政策、涉税风险提醒，提供精细化的办税服务，满足纳税人的个性化需求，使得税收业务办理过程更为轻松。③ 提供智慧问税服务，如纳税人可通过语音、文字、图片等形式咨询智慧税务系统，系统则根据纳税人的问题进行智能判断与精确引导，自动化、高契合度地提供答复。

2. 智慧管税

智慧税务系统汇集企业或个人不同源的涉税数据（如资金、发票、合同等信息）并进行全面整合，将智能决策分析系统、财税行业特征知识相结合，对涉税企业实体、涉税行为进行全面管控。运用风险预测模型，识别潜在的涉税违规企业，增强税务部门管控涉税违法风险的能力。

3. 决策支持

智慧税务系统对税务大数据进行规模化处理，过程及时高效，可挖掘数据中的重要知识。通过税收政策法规、纳税人数据的精准匹配，将税收政策效果进行量化，以模拟计算税收的方式规避各类涉税政策的潜在冲突，实现税收政策体系的综合优化；保障税收在调节经济形势方面充分发挥杠杆作用，为税务政策制定、涉税应急决策提供数据支撑。

四、面向智慧税务的大数据知识工程典型应用

（一）知识驱动的税收优惠计算

我国经济发展已经由高速增长阶段转入高质量发展阶段。大规模、实质性减税降费，有利于减轻企业负担、促进实体经济转型升级，是税收政策着眼经济转型进行的重要转变[15]。在企业实际办税过程中，经常面临优惠政策应享而未享、不应享而享等问题；通过智能化手段，为企业办税提供辅助支持，实现优惠政策应享尽享，提高企业端税收遵从度，成为发展亟需。

税收优惠计算的本质在于税收政策与纳税人的双向精准匹配。针对性提出了知识工程驱动的税收优惠计算框架（见图3），用于应对税收政策文本、纳税人经营情况实时变化的挑战。① 通过知识提取层，从税收政策文本中抽取多类条件（由文本描述，包括行业属性、纳税人属性、税种信息、涉税

约束等），构建规则知识库，采用知识融合技术对知识库中的规则进行重复合并、失效剪裁。② 以领域知识为指导，进行规则编码，构建决策表。③ 结合实际业务需求，从源数据中获取纳税人数据，由规则计算引擎筛选出符合条件的目标数据。

知识工程驱动的税收优惠计算主要由4个组件构成：应享优惠资格检测，用于检测纳税人是否有资格享受某项优惠；可优惠税额计算，在纳税人具备资格享受某项优惠的条件下计算出优惠额度；申报智能指导，在纳税人具备资格享受某项优惠的条件下检测出需提供的申报材料；条件缺失预警，在纳税人不具备资格享受某项优惠的条件下判别出未满足的条件。

（二）可解释的税收风险识别

近年来，税务部门采用了以“信用+风险”为基础的新型税收风险监管机制，在提升税收遵从度方面成效明显，但针对部分重点行业（如成品油、文娱影视）的税收风险识别工作仍需加强[31]。增强税收风险识别水平，可有效减少税收损失并规范经济秩序，而面向税收大数据开展税收风险识别工作极具挑战性。运用智慧税务大数据知识工程支撑税收风险识别，是亟待解决的问题。

应用大数据知识工程开展税收风险识别涉及四方面。① 通过知识提取层获取不同来源和形态的多维度知识，对获取的知识进行全方位、多维度表示。例如，从企业资金流、发票流、合同流、物流中抽取碎片化知识，结合财税行业特征知识，构建面向税务部门的财税知识库。② 在知识推理层，通过智能风险分析模型对涉税风险进行智能监控，辨识可疑企业的潜在违法风险线索，运用符号化表征代数，将风险线索依据时序、依赖、因果等关系进行动态融合，生成推理路径和证据链，提高涉税违法行为稽查结果的可解释性。③ 针对不同来源渠道的线索信息以及融合生成的证据链，归纳差异性，实现可视化分析；提供线索及证据链交互分析工作台，辅助税务部门完成案件分析、报告在线编制。④ 支持税收监管从“以票管税”向“以数治税”精准转变，深入分析企业涉税违法行为，精准发现涉税违法线索，为税务执行提供及时准确支持。

以国家税务总局公布的某违法企业为例，阐述税收风险识别的应用成效（见图4）。某企业进口原油、产出汽油，本应缴纳燃油税后将汽油售予加油站，实现产品正常流通；但与其他违法企业组成团伙，通过多轮“加工”“交易”“运输”形式拉长链条，在这一过程中逐渐模糊品名、增大损耗，多以沥青等低税率化工品进行缴税，从而达到逃避纳税义务的目的。这一过程中的链条多级协同、分散嫌疑，使各环节表面合规、难以察觉。智能风险分析模型主要从已有的财税知识库中识别风险线索，如资金闭环、控股互锁、购销背离、假冒进出口等，随后经碎片线索融合得到完整的推理证据链，从而有效甄别出上述涉税违法行为。可解释的税收风险识别，提高了税务人员高效精准开展涉税监管工作的能力，显著提升了财税领域智慧化建设水平，产生了显著的社会效益。

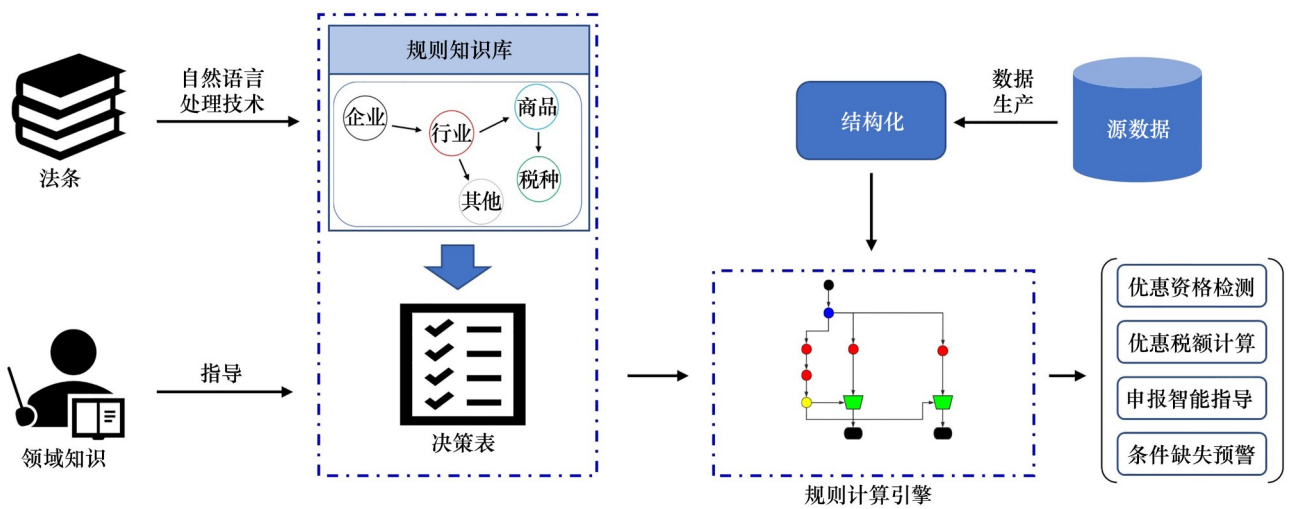

图3　大数据知识工程驱动的税收优惠计算框架

（三）税收政策智能化决策支持

税收政策是政府因调节市场经济主体的物质利益、给予强制性激励等需求而发布的一系列规章制度，在引导社会正向发展的同时，需尽量规避政策之间潜在的冲突。随着税务信息化建设的逐步完善，各级税务部门积累了海量的税务数据；获取其中蕴含的法律信息等知识并开展推理和应用，可以反哺和支持税务政策制定，提高涉税管理决策水平。如何利用税务数据中的知识为涉税决策提供科学支撑，成为税务信息化建设方面新的更高要求。

在推理层面提出了文本–知识“可计算”、知识–知识库“可计算”、知识库–数据“可计算”等准则。针对涉税决策支持的“可计算”特征，构建了基于大数据知识工程的税收政策决策支持框架（见图5），打通了税务征收从成文条例到可视化、定性分析的全链条，并对拟出台政策的可能影响进行分析、模拟、预警。① 通过知识提取层，在文本–知识“可计算”准则的指导下，从税收政策法规文本中抽取结构化数据（如涉税实体、实体关系），以知识融合解决共指与对齐等问题；将非结构化的税收政策法规文本转变为结构化的税务知识库，通过知识推理及更新实现动态演化，支持新知识的持续融入。② 在知识–知识库“可计算”准则的指导下，利用知识库中蕴含的高度规则性、逻辑性，采用Text2SQL等技术，将税务知识库编译为可执行代码库。③ 在知识库–数据“可计算”准则的指导下，由知识源层实时提供海量纳税人数据，投入可执行代码库并根据时序变化计算涉税金额；结合税务业务规则，触发税务决策路径，生成“减免税款超限，建议驳回”“降低原材料成本X%，预期刺激居民消费Y%”等分析结果。

（四）智慧问税

更高水平的纳税服务，在不断改善办税便捷性之外，需要在财税知识精准搜索与推送、针对财税知识咨询的优质答复等方面着手，支持企业合理享受税收优惠，减少因法律知识缺乏而面临的经营风险；实现内部涉税处理、外部涉税需求的快速联动，满足企业多样化的问税需求。智慧问税服务主要分为检索、推荐、问答三部分，特色在于包含问答、文章、会话等多种内容形式，涵盖税务业务、法规政策、系统使用、咨询解答等全量内容的综合性服务。

构建了财税知识管理方面的知识搜索引擎。知识搜索引擎是知识管理的实现工具，承担了财税知识汇聚、发现、分类、聚类、门户构建等功能。财税知识管理系统以搜索引擎、分类体系、专家网络为主要内容，其中财税知识特指税务领域的综合知识，如税务业务知识、法规、政策、文件、约定、文章、多媒体等。

构建了面向纳税人、税务人员的个性化知识推荐引擎。知识推荐引擎按照不同场景，指定选择某

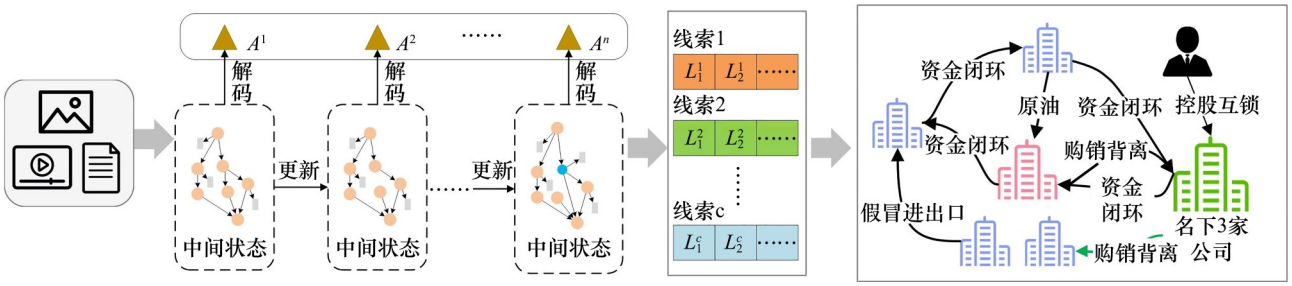

图4　某违法企业税收风险识别案例

图5　基于大数据知识工程的税收政策决策支持框架图

些税务领域的综合知识；可向税务人员推荐相关知识，也可依据纳税人画像将关联的个性化知识（如新闻发布会信息、税收政策、税收服务）进行定向推送

构建了知识问答引擎，用于精确理解问题并高质量应答。按照不同的场景，指定选择部分税务领域的综合知识来提供多轮对话问答能力；具有税务领域专业特色，可按需设定敏感规则、情绪识别规则等。企业可根据答复采取必要措施，合理降低因财税专业和法律知识不足可能面临的经营风险；通过问答来理解和享受政策让利，有利于税企双赢。

五、面向智慧税务的大数据知识工程后续研究方向

受限于大数据知识工程技术的当前发展水平，本研究提出的智慧税务系统应用架构存在局限性。例如，对于税收政策的智能化决策支持，现有方案在预测政策实施后征纳双方税收数据变化等直接关联问题方面具有良好表现，而在预测政策实施后国家经济、企业行为、居民消费等间接关联问题方面的稳定性有所不足，也就制约了大数据知识工程在保障涉税决策管理方面的可用性。因此，面向智慧税务的大数据知识工程，后续研究需着重关注以下三方面。

（一）高阶多跳推理

现有的大数据知识工程方法，在融合多种深度学习模型与大量符号知识时面临组合爆炸难题，导致应对“税收政策变更，对纳税人产生影响，进而对相关行业产生影响，最终对国民经济产生影响”这类高阶多跳推理问题的效果不理想。引入神经符号网络，将一阶逻辑运算符号转换为可微神经几何操作，指导海量涉税符号知识的高阶多跳推理，解决当前技术应用面临的组合爆炸、知识规模庞大等问题，显著增强智慧税务对于涉税管理决策支持的能力。进一步地，探索规则运算、概率逻辑的结合，模拟多跳推理答案的不确定性，提高推理结果的可信性与安全性，对于涉税管理决策等风险敏感应用极为关键。

（二）反事实推理

反事实推理指对过去已发生的事实进行否定而重新表征，以建构一种可能性假设的思维活动，诸如“假如……那么……”之类的问题可称为反事实问题。反事实逻辑推断的能力是大数据知识工程中智能化水平的重要体现。反事实推理对于智慧税务中的政策效应分析、政策风险评估等应用具有关键作用，引入后可显著增强模型的推理决策能力、结果可解释性，甚至推动智慧税务中诸多智能化应用的变革。可借鉴潜在结果框架、结构因果模型等相对成熟的反事实因果推断框架，解决反事实推理在鲁棒性、动态性方面的不足，拓展在因果图不完整、数据特征缺失等更复杂推理场景中的应用。

（三）知识差异化表征机制

现有的大数据知识工程方法，对不同类型的知识通常采用同样的表征机制。与之相反，人脑对不同类型的知识由不同的部位进行存储和表征，如事实性知识、情景知识主要由海马体负责编码表征，程序性知识主要由小脑负责，各自的工作机制差异明显。可借鉴脑科学的研究进展，探索采用不同类型知识的差异化表征机制，用于提升大数据知识工程在决策管理支持等智慧税务下游任务中的表现。借鉴AI领域中的“认知图谱”、认知神经科学领域中的“双过程理论”、脑科学领域中的“海马体”记忆回放机制等研究成果，厘清大脑使用的知识编码方式对后续知识推理或行为决策的作用机制；探索新型神经网络架构，兼顾记忆具体细节、提取抽象结构与共性特征，设计更接近人脑机制的知识编码、记忆与推理模型。

六、面向智慧税务的大数据知识工程发展建议

（一）规范涉税数据，健全国家数据共享 / 开放 / 保障体系

税收数据的质量与标准是大数据知识工程应用于税收分析业务的条件和基础，智慧税务系统建设需重点开展税收数据的规范统一化处理。在采集分散于各个平台和终端的涉税数据时，应采取统一的规范标准，同步进行数据校验和纠错，实质性提升税务数据的基础质量。逐步规范并合理拓展税务数

据的共享范围，融入更多“非税”数据，如加强与银行、公安、社保、质监、统计、银行等部门的合作，打破数据壁垒，为智慧税务高质量发展提供进一步的数据支撑。建议推动建设涉税数据安全保障系统，采用密钥等基础技术对涉税人员提供身份认证、数据使用等安全服务，充分保障涉税信息在采集、运输、存储、使用时的安全性，为智慧税务应用提供可靠支撑。

（二）融合并更新信息学科成果，完善面向智慧税务的大数据知识工程应用系统

大数据知识工程方法的应用系统建设，尤其是面向智慧税务场景时，需融合海量、多源数据进行推理计算，技术性保障各方数据的安全和隐私问题。融合并更新包括区块链、联邦学习等技术在内的信息学科研究成果，与智慧税务应用场景深度结合，可信、高效地开展数据协同计算。推动建设基于区块链的纳税服务平台，发展税务知识联邦理论框架体系，融合AI、分布式计算、密码学等学科知识，逐步满足数据“可用不可见”的实际需求。在更高水平上开展涉税知识的共创和共享，为优化面向智慧税务的大数据知识工程应用筑牢基础。

（三）推动大数据知识工程技术的标准建设与人才培养

大数据知识工程技术及其在各领域的应用发展迅速，但相关的技术标准研究和制定处于初级阶段。大数据知识工程相关术语、适用准则等标准缺失，不利于大数据知识工程在智慧税务领域中的规范开发与落地应用。建议结合文中的“五层”技术模型，从概念理解与沟通、关键技术研究与实施、行业落地应用三方面出发，建立大数据知识工程标准体系架构。此外，高校、科研院所依托新工科建设、卓越工程人才计划等项目，加快制定大数据知识工程技术人才培养方案，在理论拓展、技术应用、工程管理等方面投入必要资源，批量培育应用型的大数据知识工程技术人才，为智慧税务领域高质量发展、信息技术产业竞争力提升提供基础支撑。

利益冲突声明

本文作者在此声明彼此之间不存在任何利益冲突或财务冲突。

Received date: July 30, 2022; **Revised date:** September 29, 2022

Corresponding author: Shi Bin is an associate professor from the School of Computer Science and Technology of Xi'an Jiaotong University. His major research fields include knowledge engineering, cloud computing. E-mail: shibin@xjtu.edu.cn

Funding project: National Natural Science Foundation of China (62250009, 61721002); China Engineering Science and Technology Knowledge Center Project (CKCEST-2022-1-40)

参考文献

[1] Schreiber G, Akkermans H, Anjewierden A. Knowledge engineering and management: The common KADS methodology [M]. Cambridge: MIT Press, 2000.

[2] Wu X D, Chen H H, Wu G Q, et al. Knowledge engineering with big data [J]. IEEE Intelligent Systems, 2015, 30(5): 46–55.

[3] 吴信东, 靳小龙, 陈欢欢. 大数据知识工程研究进展与发展趋势 [J]. 中国计算机学会通讯, 2021, 17(6): 1–8.
Wu X D, Jin X L, Chen H H. Research progress and development trend of big data knowledge engineering [J]. Communications of the CCF, 2021, 17(6): 1–8

[4] 郑庆华, 刘均, 魏笔凡, 等. 知识森林: 理论、方法与实践 [M]. 北京: 科学出版社, 2021.
Zheng Q H, Liu J, Wei B F, et al. Knowledge forest: Theory, methodology, and application [M]. Beijing: Science Press, 2021.

[5] Manyika J, Chui M, Bughin J, et al. Disruptive technologies: Advances that will transform life, business, and the global economy [M]. San Francisco: McKinsey Global Institute, 2013.

[6] Adams T. Google and the future of search: Amit singhal and the knowledge graph [J]. The Guardian, 2013, 19: 1–8.

[7] Wang Z Y, Huang J M , Li H S, et al. Probase: A universal knowledge base for semantic search [R]. Beijing: Microsoft Research Asia, 2010.

[8] Mitchell T, Cohen W, Hruschka E, et al. Never-ending learning [J]. Communications of the ACM, 2018, 61(5): 103–115.

[9] Lu R Q, Jia C Y, Zhang S F, et al. An exact data mining method for finding center strings and all their instances [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(4): 509–522.

[10] Tang J, Zhang J, Zhang D, et al. ArnetMiner: An expertise oriented search system for web community [EB/OL]. (2007-06-15) [2022-05-15]. https://ceur-ws.org/Vol-295/paper01.pdf.

[11] Wei B F, Liu J, Ma J, et al. Motif-based hyponym relation extraction from Wikipedia hyperlinks [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(10): 2507–2519.

[12] 游家兴, 柳颖, 杨莎莉. 智慧税务助力高质量发展的实践与探索 [J]. 税务研究, 2022 (7): 64–69.
You J X, Liu Y, Yang S L. Practice and exploration of intelligent taxation to help high-quality development [J]. Taxation Research, 2022 (7): 64–69.

[13] 张靖. 深化数字技术运用 推动智慧税务建设 [J]. 税务研究, 2022 (5): 128–130.
Zhang J. Deepen the application of digital technology and promote the construction of smart taxation [J]. Taxation Research, 2022 (5): 128–130.

[14] 孙存一, 谭荣华. 简析大数据支撑下的“互联网+智慧税务” [J]. 税务研究, 2018 (4): 104–107.

Sun C Y, Tan R H. Brief analysis of “Internet + intelligent taxation” supported by big data [J]. Taxation Research, 2018 (4): 104–107.

[15] 余红艳, 孙丽, 刘亚利. 减税政策: 动因追溯、制度约束与路向选择 [J]. 税务研究, 2022 (7): 32–37.
Yu H Y, Sun L, Liu Y L. Tax reduction policy: Motive tracing, institutional constraint and direction choice [J]. Tax Research, 2022 (7): 32–37.

[16] Ackoff R L. From data to wisdom [J]. Journal of Applied Systems Analysis, 1989, 16(1): 3–9.

[17] Shaw J, Rudzicz F, Jamieson T, et al. Artificial intelligence and the implementation challenge [J]. Journal of Medical Internet Research, 2019, 21(7): 1–12.

[18] 丁梦远, 兰旭光, 彭茹, 等. 机器推理的进展与展望 [J]. 模式识别与人工智能, 2021, 34(1): 1–13.
Ding M Y, Lan X G, Peng R, et al. Progress and prospect of machine reasoning [J]. Pattern Recognition and Artificial Intelligence, 2021, 34(1): 1–13.

[19] 张钹, 朱军, 苏航. 迈向第三代人工智能 [J]. 中国科学: 信息科学, 2020, 50(9): 1281–1302.
Zhang B, Zhu J, Su H. Towards the third generation of artificial intelligence [J]. Scientia Sinica Information, 2020, 50(9): 1281–1302.

[20] Marcus G. The next decade in AI: Four steps towards robust artificial intelligence [EB/OL]. (2020-02-14)[2022-06-15]. https://arxiv.org/abs/2002.06177.

[21] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436–444.

[22] 姜磊, 刘琦, 赵肄江, 等. 面向知识图谱的信息抽取技术综述 [J]. 计算机系统应用, 2022, 31(7): 46–54.
Jiang L, Liu Q, Zhao Y J, et al. Review on information extraction techniques for knowledge graph [J]. Computer Systems & Applications, 2022, 31(7): 46–54.

[23] Li J, Sun A X, Han J L, et al. A survey on deep learning for named entity recognition [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50–70.

[24] 程华龄, 陈艳平, 杨卫哲, 等. 基于多维语义映射的关系抽取方法研究 [J]. 计算机科学, 2022, 49(11): 206–211.
Cheng H L, Chen Y P, Yang W Z, et al. Relation extraction based on multidimensional demantic mapping [J]. Computer Science, 2022, 49(11): 206–211.

[25] Feng J, Huang M L, Zhao L, et al. Reinforcement learning for relation classification from noisy data [C]. New Orleans: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, 2018.

[26] Cohen W W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records [C]. Austin: International Conference on Knowledge Discovery and Data Mining, 2003.

[27] Nguyen H L, Vu D T, Jung J J. Knowledge graph fusion for smart systems: A survey [J]. Information Fusion, 2020, 61: 56–70.

[28] Althoff T, Dong X L, Murphy K, et al. Timemachine: Timeline generation for knowledge-base entities [C]. Sydney: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[29] Vrandecic D, Krotzsch M. Wikidata: A free collaborative knowledgebase [J]. Communications of the ACM, 2014, 57(10): 78–85.

[30] Guo L, Wu J, Li J H. Complexity at mesoscales: A common challenge in developing artificial intelligence [J]. Engineering, 2019, 5(5): 924–929.

[31] 王军. 以深入开展“学查改”专项工作为契机 扎实推动习近平经济思想在税务系统落地生根 [J]. 中国税务, 2022 (7): 7–9.
Wang J. Taking the opportunity of the special work of “learning, investigation and reform” to promote Xi Jinping’s economic thought in the taxation system [J]. China Taxation, 2022 (7): 7–9.