

教育数据卡片在人工智能知识库建设中的应用实践

陆凌霄

宁波市教育服务与电化教育中心 浙江宁波 315000

摘要: 基于教育数据中心,运用“数据卡片体系”理论,结合教育行业的特点,阐述了教育数据卡片体系的概念,提出了教育数据卡片体系的构建方法,设计了相应的管理体系,并通过应用案例进行了实现。实践表明,构建教育数据卡片体系对于人工智能知识库的建设价值,促进教育领域创新应用的转变是行之有效的。

关键词: 人工智能;知识库;卡片体系

近年来,随着教育信息化水平的逐步提高,人工智能技术在教育领域的应用日益增加,把海量数据治理成为标准化知识库是支撑人工智能应用场景建设的核心关键,而数据治理最有效途径就是构建教育数据卡片体系。为了加强教育系统内教育知识库的管理,促进教育数据助力人工智能应用创新,保护管理机构、学校、老师和学生等的合法权益,保障教育领域的数字化改革,深化智慧教育建设,推进市域教育数据治理体系和治理能力现代化。教育系统内数据的收集、归集、存储、加工、传输、共享、开放等数据治理工作日趋成熟。人工智能对教育数据资产的利用便利性要求日益提高,从本质上讲是一些采集、汇聚所获得的教育数据资源。要让这些数据资产发挥价值就需要把它转化成一类对教育管理有价值输出的产物^[1],实现上层业务的创新应用,真正将数据变为实现业务价值的创新利器^[2]。

1. 教育数据卡片的建设背景

党的二十大将教育作为全面建设社会主义现代化国家的基础性、战略性支撑进行系统谋划,极具战略意义和深远影响^[3]。目前,我国正全面实施教育治理,提升我国教育水平,教育治理也是实现教育现代化、建设教育强国必须要面对和解决的重大理论与实践问题。在新时代的背景下,随着互联网特别是移动互联网发展,社会治理模式正在从单向管理转向双向互动,从线下转向线上线下融合,从单纯的政府监管向更加注重社会协同治理转变。对于我国各省市而言,如何综合运用数字化技术、数字化思维、数字化认知,对教育治理的体制机制、组织架构、方式流程、手段工具进行全方位系统性重塑,高效构建教育治理新平台、新机制、新模式,高水平推进教育服务治理现代化,是推进教育治理

数字化改革的重要问题。

人工智能可以把众多数据资源进行整合,借助专业的硬件系统和软件程序对这些数据信息进行分析与整理,不仅可以确保分类具有合理性,还可完成储存管理和索引建立,实现教育领域知识库的建设,随着我国引进并推广数据库技术,其在我国生产建设的应用范围变得更加广泛^[4]。知识库技术应用于教育治理,可以便于教育信息数据的有效管理,明确教育体制机制和组织架构,是教育治理数字化改革的重要技术手段。

2. 教育人工智能知识库建设现状

我国各地区教育知识库体系并不完善,普遍存在一些问题:首先,并发访问性较低。由于部分教育知识库信息是非格式化的数据,数据具有对象复杂、存储分散和时空同步等特点,导致知识库接口存储实时访问效率低,高并发访问不稳定;其次,数据安全问题。数据泄露会严重侵害个人信息主体权益,影响国家安全、经济发展和社会稳定。全国教育系统的信息化,积累了大量个人信息甚至是个人敏感信息,海量教育数据面临着严峻的数据安全风险^[5]。虽然各地教育厅对数据安全和个人信息保护原则性要求提出了工作指南,但尚未明确技术手段建设的具体要求与规范标准,无法有效推动教育数据中心开展数据安全技术手段的研发与应用^[6];最后,数据分类分级和质量评估效率较低。数据分类分级管理是数据实现共享和开放最基础的工作^[7],由于数据信息体量极大且具有多源异构型,数据分类和评估费时费力。

3. 教育数据卡片体系定义

教育数据卡片是对教育数据某个维度特征的描述,能反应特定维度的显著特征。传统的卡片生成方法主要采用经

验模式结合特征分析来实现卡片体系的搭建和应用,而人工智能技术的发展为卡片的计算、存储和应用提供了高效的技术支持^[8]。

教育数据卡片体系是根据教育管理的要求,分业务(部门)、分类别、分层级建立的一套知识卡片体系,包括静态属性和动态属性卡片,主要用于开展以卡片为维度的数据深度分析、以卡片为维度的业务应用提升、以卡片为维度的跨部门管理创新^[9]。

教育数据卡片体系的建立能为上层应用提供统一的卡片数据目录和卡片调用接口,沉淀上层应用制作的模型卡片,实现高价值数据卡片共享复用体系,形成卡片应用管理的闭环流程,从而实现教育人工智能的价值和创新应用的快速落地。

4. 教育数据卡片体系的构建

随着人工智能的发展,卡片体系已被互联网企业广泛使用。通过特征集合并关联打卡片对象,对分析对象生成画像来深度挖掘对象的价值。教育数据卡片主要可以分为事实特征卡片、统计数据卡片和预测分析卡片三大类。如基于老师的基本信息、兴趣爱好等数据可以对老师进行不同特征的卡片划分。而基于时序特征的培训情况、获奖情况、科研情况、学生评价情况等行为数据,结合老师的特征化卡片和时间轴,可以构建面向老师在培训、奖惩、科研、评价等多维度的成长数据链及统计类数据卡片。事实特征卡片和统计类数据卡片是基于现有的数据计算得到的卡片,而预测分析卡片是在事实特征卡片和统计类数据卡片的基础上,进一步构建预测分析模型衍生得出的预测卡片。这类卡片能更好地对未来教育的发展趋势提供决策支撑。

教育数据卡片体系的构建必须以教育数据中心已有的可靠、可信、可控的数据为依托,从教育管理的各个维度为各类对象建立卡片化的描述体系。

教育数据卡片体系模型的设计主要包括三个步骤:

(1) 部门业务流程的调研梳理:对各教育业务部门进行走访调研,了解熟悉、并梳理各部门的业务管理流程,理解部门的决策流程,熟悉部门的核心业务需求,明确业务数据和学生、老师及管理者的行为。

(2) 明确部门卡片分类:根据各部门的业务管理目标,对教育管理流程所涉及的业务数据和用户行为等属性进行卡片分类,建立不超过 3 级的卡片类别体系。

(3) 建立数据卡片体系模型:在卡片类别体系的基础上,根据时效性将卡片分为静态属性(长期甚至永远都不会发生改变)和动态属性(需要定期地更新,保证卡片的有效性)卡片,或者从数据获取维度来分为事实特征卡片(从原始数据中提取)、统计数据卡片(定义规则,建立统计模型来计算得出卡片实例)和预测分析卡片(预测行为、喜好或趋势)。整个模型可以包括业务类型、规则策略、管理方式、统计维度、算法挖掘、权限归属等类别,并最终形成完整的教育数据卡片体系模型。

在建立教育数据卡片体系模型后,还需要借助教育数据卡片管理系统来实现对卡片体系的定期更新维护,包括卡片的提出、卡片的生成、卡片的审批和卡片的执行四个过程,确保卡片体系能及时满足各教育主管部门新业务、新数据、新目标的需要。另外,教育数据卡片体系的设计过程中要实现卡片的最小颗粒度要触达到具体业务事实数据,同时支持对应卡片实例的规则自定义;不同的卡片可以相互组合为新的上级卡片来更好地表述对象间的不同层级和不同视角关系。

5. 教育数据卡片算力需求

(1) 数据卡片需要云计算的算力支持,利用云计算高灵活性、可扩展性和高性价比等,需要通过网络“云”将巨大的数据计算处理程序分解成无数小程序,然后,通过多部服务器组成的系统进行处理和分析这些数据得到结果并组成数据卡片。通过这项技术,可以在很短的时间内(几秒钟)完成对数以万计的数据的处理,从而达到强大的数据服务。现阶段所说的云计算已经不单单是一种分布式计算,而是具有分布式计算、效用计算、负载均衡、并行计算、网络存储、热备份冗余和虚拟化等计算机技术混合演进并跃升的优势,大大提高数据卡片建立的效率。

(2) 数据卡片需要基于 REDIS 的分布式缓存架构,Redis 是一款开源免费的高性能 Key-Value 数据库。与关系型知识库 Oracle、Microsoft SQLServer、MySQL 不同,Redis 属于 NoSQL 数据库(非关系数据库),Redis 不使用表,其知识库也不会预定义或者强制去要求用户对 Redis 存储的不同数据进行关联。对于海量的数据要高效的建立数据卡片,如果直接从数据库查询,就要消耗大量的数据库请求。如果将这些数据储存在内存中,要用的时候,直接从内存调取,不仅可以大大节省系统直接读取磁盘来获得数据的 IO 开销,

提高服务器的资源利用率，还能极大地提升速度。Redis 支持数据的持久化（包括 AOF 和 RDB 两种模式），可以将内存中的数据保存在磁盘中，重启的时候可以再次加载进行使用，性能与可靠性兼顾。在人工智能应用场景下，数据规模可达数十 G，甚至数百 G，而物理机的资源却是有限的，内存无法无限扩充，且单实例 Redis 并发性能也有限，对于海量的数据还需要创建 Redis 集群来处理。该技术在数据卡片建立过程中主要用于数据、分析结果的缓存，数据储存到内存中，要用的时候，直接从内存调取，可以大大节省系统直接读取磁盘来获得数据的 IO 开销，提高服务器的资源利用率，还能极大地提升数据卡片建立的速度。

6. 教育数据卡片体系在知识库中的应用与实践

当前优质教育资源相对短缺，中高职、初中、小学等入学依然是社会关注的焦点，本文以建设人工智能体进行学校招生情况互动为主要目标来阐述数据卡片体系的具体应用方法，促进招生入学体制机制的创新，提高教育管理部门的办事效率和用户服务的满意度。该应用例子中共包括 2 个一级卡片、6 个二级卡片和 29 个三级卡片。

整个招生入学数据卡片体系的处理，主要由数据预处理、数据建模和数据分析三大块组成。数据预处理包括数据缺失值处理、异常值处理、变量变换、数据离散化等过程，如采用均值填充，随机森林方法，最小值填充等方法来解决缺失值问题。对于连续属性值进行离散处理，如家庭经济情况可以通过聚类分析方法划分为优越、良好与贫困等离散值。对于预测分析卡片则利用机器学习方法建立相应模型来实现卡片的计算，如设计求解损失函数极值，将损失函数泰勒展开到二阶函数来进行计算如公式 1，且在损失函数中加入了正则化项等方法来设计卡片预警模型。

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad \dots\dots \text{公式(1)}$$

↑ ↑
梯度提升算法损失 正则项

基于上述过程可以建立面向学生和老师的各种应用场景数据卡片体系，以及管理部门的管理效率、服务效率、服务质量等的绩效评估数据卡片，来展现现代化教育的多样性、便捷性、共享性、时效性、互动性、丰富性。

7. 结论与展望

教育人工智能作为一类教育资源，基于“卡片数据”技术理论，通过对其进行分析研究，构建教育数据卡片体系，呈现多维度、多视觉、系统性的教育数字画像，一方面有助于形成教育人工智能分析挖掘的方法和技术，另一方面帮助教育从业人员和管理者更好地洞察和理解教育规律。因此，不断提高教育数据的广度、数据的质量、数据分析模型的科学性、卡片设计的合理性，将是未来教育智能化技术发展的重要方向。

参考文献：

- [1] 唐树莺, 胡波, 崔易, 李铁瑞. 基于人工智能和知识图谱的零售客户标签体系设计[J]. 全国流通经济, 2021(02):29-31. DOI:10.16834/j.cnki.issn1009-5292.2021.02.009.
- [2] 李阿勇, 税雪, 宋志伟. 数据标签研究与应用[J]. 电力人工智能, 2020,23(06):69-74. DOI:10.19317/j.cnki.1008-083x.2020.06.010.
- [3] 教育部发展规划司. 《2022 年全国教育事业发展的基本情况》. 2022
- [4] 王钊. 知识库技术在信息管理中的应用[J]. 现代工业经济和信息化, 2023,13(01):142-144.
- [5] 马建立. 基于《数据安全法》的教育数据分类分级研究[J]. 天津电大学报, 2023,27(01):55-60.
- [6] 沈玉青. 数据安全, 护航钢铁产业高质量发展[N]. 江苏经济报, 2023-03-28(A01).
- [7] 高潮. 人工智能时代用户消费型数据的分类分级隐私保护策略研究[J]. 通信技术, 2016, 36(9): 10-12+17.
- [8] 周成祖, 吴文, 蔡晓强. 基于分类分级的数据安全防控策略研究[J]. 数据与计算发展前沿, 2023,5(01):128-135.
- [9] 赵永国. 人工智能挖掘的用户画像人才标签体系生成方法[J]. 信息技术, 2021(12):142-146. DOI:10.13274/j.cnki.hdzj.2021.12.025.

课题：本论文研究也作为 2024 年浙江省教育信息化研究课题《数据驱动“五育并举”教育智评模式的探索与实践》的成果。