■ 教育数字化·高等教育大数据专题　　DOI:10.15998/j.cnki.issn1673-8012.2022.04.002

大数据与院校研究

林松月[1],刘　进[2]
(1.香港中文大学　教育学院,香港　999077;2.北京理工大学　人文与社会科学学院,北京　100081)

摘　要:院校研究具有运用大数据方法开展研究活动的天然优势,但已有研究尚未充分发掘适合大数据研究的条件和途径,尚未形成较好的大数据研究示范。用大数据思维替代传统研究思维、用大数据替代传统研究的有限数据、用大数据算法替代传统统计学算法是院校研究与大数据结合的基本技术原理。在深入讨论院校研究大数据方法运用原理的基础上,以高校大数据的最常见载体——校园一卡通为分析对象,对院校研究过程中各类大数据的分布、生成、存储、采集、使用等流程展开案例分析。研究发现:大数据方法对院校研究范式、过程、结果等具有颠覆性意义,大数据与院校研究具备良好的融合能力;院校研究大数据广泛分布于教育教学活动中,院校研究与大数据融合发展前景广阔。未来的院校研究应更加强调研究的科学化目标和实践导向,更好地营造证据导向、数据导向的研究场域,形成新的以大数据方法为准则的院校研究学术共同体。

关键词:大数据;院校研究;校园一卡通;大数据库;学术共同体

[中图分类号]G640　[文献标志码]A　[文章编号]1673-8012(2022)04-0007-13

进入第四次工业革命后,大数据和人工智能正在彻底改变社会科学研究方式,教育大数据研究结果直接应用于教育教学改革前景更为广阔。院校研究应抢抓大数据发展机遇,形成各类院校大数据库资源[1],提升院校研究科学化水平,进一步完善院校研究的理论、方法和公共知识体系。近年来,院校研究领域进入瓶颈期,在其他学科(如管理学、经济学、图书情报学、新闻学)和教育学其他领域迅速引入大数据方法的同时,旨在指导教育改革实践的院校研究仍处于"方法沉睡"阶段[2],尚未面向大数据时代实现研究意识、研究方法、研究技术转向,尚未激活各类大数据资源,部分大数据库仍处

修回日期:20220506
基金项目:国家自然科学基金面上项目"政府奖学金能否提升来华留学生质量:基于机器学习方法的'一带一路'国家因果推断"(71974012);国家自然科学基金面上项目"'一带一路'学术人才向中国流动的开放式'推-拉模型研究——人工智能方法的运用'"(71774015)
作者简介:林松月,女,河北邢台人,香港中文大学教育学院博士生,主要从事高等教育国际化研究;
通信作者:刘进,男,江苏东海人,北京理工大学人文学院副研究员,北京理工大学国际争端预防和解决研究院研究员,兰州大学格鲁吉亚研究中心特聘研究员,主要从事高等教育国际化和教育大数据与人工智能研究。
引用格式:林松月,刘进. 大数据与院校研究[J]. 重庆高教研究,2022,10(4):7-19.
Citation format: LIN Songyue, LIU Jin. Big data and institutional research[J]. Chongqing higher education research,2022,10(4):7-19.

于“数据封闭”和“数据孤岛”状态[3]。而这些问题的背后,是院校研究者尚未清晰了解教育大数据资源的现实分布、可能应用和广阔前景,一些研究文献虽然“鼓吹”在院校研究领域引入大数据资源,但大多为纸上谈兵,没有形成良好的研究示范。为进一步阐明院校研究领域的教育大数据生成、存储、采集、使用等流程,分析院校研究领域教育大数据的各类基础特征,展现教育大数据在院校研究领域的广阔应用前景,本研究将在理论分析的基础上引入部分研究实例展开阐述。

一、大数据在院校研究中的应用原理

(一)大数据提升院校研究科学化水平

相比于传统的各类研究方法,大数据在院校研究中的应用至少具有 3 个方面的颠覆性特征。

1. 对院校研究范式的颠覆性:基于证据的科学化、动态性研究范式

传统的院校研究大多基于已有的理论框架、研究假设或案例经验,展开研究设计、证据采集与结果分析。由于理论创新的难度较大,大多数研究沿用已有理论基础,或在已有理论基础上进行迁移、整合、嵌套或延伸,还有一些研究没有理论基础,只依据部分文献资料或实践案例整合形成各类研究假设。这种“理论/文献/案例—假设—验证”的基本研究范式,过度强化了已有理论、研究框架或实践案例的合法性,在一定程度上降低了院校研究理论创新的动机和可能性,很多院校发展规律受到理论束缚无法得以揭示。还有一些研究为理论而理论,为假设而假设,案例迁移使用牵强附会,存在研究理论的方便性使用、研究证据的选择性使用、研究工具的复杂化使用、研究结果的牵强性使用以及数学公式滥用、结构模型乱用等问题,进一步降低了院校研究的科学性。大数据研究则突破了这一传统研究范式,更多地基于科学证据本身做出研究判断,开展因果式研究而非相关性研究,进行各类院校规律预测并动态调整预测模型与预测结果(不断提高预测精度以逼近真实发生值),这将有望使院校研究逐步进入不必过于依赖理论、过于依赖假设、过于依赖传统案例或者决策经验的新阶段。这不仅大大提升了院校研究的科学化水平,通过公布数据源、研究代码等方法,推动院校研究过程透明化和结果公开化,重塑基于证据的院校研究共同体,而且可以降低院校研究难度,从各类理论研究、思辨研究、案例研究等逐步转向规范的大数据研究[4-5],提高研究的趣味性水平,探讨院校活动中的各类有趣现象。

2. 对院校研究过程的颠覆性:基于事实的全样本、客观性研究过程

一方面,传统的院校研究以问卷调查、访谈调查、质性研究、案例研究等方法为核心,研究过程只能实现对有限数据、有限资料的占有和分析,较容易导致研究结论的偏差。虽然一些研究试图通过借鉴统计学等方法降低抽样误差和选择性偏误,但研究者本人的研究方向偏好、个人研究习惯、院校研究经历、方法使用、数据获取、分析解释能力上的差异等,仍可能影响研究过程,降低研究成果质量。尤其是院校管理中很多问题本身就非常复杂[6],研究过程如果只采用少量资料容易导致“盲人摸象”,致使院校研究方案和决策出现偏差。另一方面,大多数已有的院校研究活动以探求各要素间相关关系为主要研究过程,但各类相关分析很容易遮蔽真实的教育问题发生机理。根据简单模型所形成的各类教育推导,也可能存在较大误差。大数据方法的引入,有可能为上述院校研究的过程缺陷带来弥补的转机。第一,大数据将使得院校研究过程数据化,基于事实数据开展研究,降低研究者个人偏好等对研究过程的影响;第二,大数据将尽可能呈现完全样本[7],有望回避或大幅降低因抽样等带来的研究误差;第三,大数据将可能使得研究过程透明化、动态化、可回溯,降低研究者主观学术不端或客观研究失误的可能性。

3. 对院校研究结果的颠覆性:基于大数据的理论体系重构与研究结论重塑

一是大数据研究对于院校研究理论体系的颠覆性。传统院校研究的理论体系,一部分来自对教育哲学等反思性建构,另一部分来自各类传统教育研究的结论。同一教育问题往往存在多种理论解

释，甚至存在多种研究结果，往往导致院校决策活动在实践层面无所适从。进入大数据时代，院校研究初步具备了对各种教育理论体系进行检验和重构的数据库基础，将可能逐一检验传统院校研究理论的科学性与有效性，发现大量新的院校运行规律并凝练出全新的理论成果，新的经过大数据检验的理论成果将不仅更能指导教育实践，而且可能成为教育理论体系更新的主要合法性基础。二是大数据研究对于具体院校研究结论的颠覆性。院校研究活动内容繁复，涉及大量游离在教育理论体系之外的各领域研究要点，传统研究活动由于方法受限而在这些教育领域或要点中往往存在结论模糊、科学化水平不足等问题，导致院校决策无法直接采纳教育研究结论。而进入大数据时代，将有可能通过各类专项院校研究数据库建设，以及跨数据库的数据挖掘、碰撞、运算等，对院校研究中出现的新问题及时做出更加科学化的研判[8]，还有望通过大数据仿真，模拟不同院校决策路径的可能后果，大大提高院校研究结论采信的可能性，提高院校管理水平，降低院校改革成本。

以上分析的主要目的在于凸显大数据研究方法与院校研究结合可能产生的各类新趋势，无意于否定传统各类院校研究方法。事实上，除具有上述颠覆性特征之外，大数据方法与传统院校研究方法还具有融合性特征。比如，大数据方法与传统“小数据”院校研究方法如果充分融合，既可以从整体、宏观、全貌视角考察院校治理问题，也可以从细节、微观、案例视角分析问题成因，寻求对策，检验对策的现实有效性。一些传统的“小数据”方法，如统计方法等，也是大数据研究的基础工具。大数据方法与传统“非数据化”院校研究方法的融合可以取长补短，平衡量化研究与非量化研究，得到更有解释力的各类结论。

（二）大数据在院校研究中的技术原理

院校研究如能真正用好大数据方法，实现从传统方法向大数据方法的转向，则可迅速改善传统院校研究科学性不足等问题。通过大数据研究真正得到各类科学发现，逐步凝练而形成理论体系，有利于实现院校研究学者进行理论体系构建的“历史夙愿”。从技术原理的视角来看，大数据方法相比于传统研究方法，至少具有3个方面的技术特点。

1. 用大数据思维替代传统研究思维

院校研究的大数据思维，通常而言就是直接面向院校教育问题，运用大数据方法探寻问题内在规律和解决办法的思维方式。与传统院校研究思维不同，大数据思维有望摆脱传统院校研究思维惯性、理论预设和路径依赖，用数据说话思维取代思辨思维[7]，大数据思维取代普通数据思维，全样本思维取代小样本思维，数据因果推断思维取代相关分析思维。

2. 用大数据替代传统研究的有限数据

一些观点认为，教育大数据通常掌握在少数大的数据供应商或数据平台手中。本文认为，对于院校研究而言，其大数据目前绝大多数由院校内部掌握，无法通过公开渠道获得[9]。当前高等学校决策者和部分院校研究者尚未清晰认知大数据对于院校研究的重要价值，尚未打破院校内部的“数据孤岛”[10]，也尚未激活“沉睡”中的院校大数据。因此，应根据院校研究进行大数据的概念界定和数据采集，深化对于教育大数据来源和使用的认知[11]，充分挖掘存在于日常教育活动之中的各类大数据资源。

第一，全样本、小样本的连续追踪数据。虽然样本数不多，但构成长期、连续、多方位的数据追踪，涉及全体样本信息，完全可以成为院校研究的大数据资源。比如，本课题组对于北京市某“双一流”高校荣誉学院数百名学生展开大数据追踪，采集了学生的基础类数据，包括人口统计学指标、入校前各类档案信息、家庭信息、高考成绩信息等。在此基础上，从大一入学开始，对学生学习全过程进行追踪，定期进行问卷调查和访谈调查，采集所有学生学习结果类信息、校园卡刷卡信息、图书馆出入和借阅信息、学生专业选择和分流信息、攻读研究生或就业信息等，形成学生大数据闭环，据此可以对全体学生和个体学生的学习规律进行深入分析。例如，可以对学生高考表现、学习习惯、伙伴关系等与学

业表现的关系展开深入分析,发现各类有利于改善学生学习的规律。

第二,特定教育研究互联网数据的连续抓取。对于一些院校专题研究领域,传统大数据公司或互联网平台无暇进行专门整理,院校研究者可以根据研究需求对之进行专题数据抓取和分析,并进行数据挖掘匹配和大数据运算,最终形成教育研究专题大数据库。例如,本课题组长期致力于高校自主招生研究,曾对 2014—2019 年教育部阳光信息平台自主招生公示学生名单进行数据抓取,并对样本进行"中学层次"(省级示范性中学、市级示范性中学、普通中学)等各类数据匹配,形成了包含十多万条学生信息的大数据库。对于该数据库的分析,可以客观呈现近年来自主招生的基本情况,甚至可以对各地自主招生名额投放中的"地方保护主义"等问题展开细致分析[12-13],为深入开展自主招生研究提供大数据支撑。

第三,根据特定研究需求形成的定制式数据。院校研究者还可以根据特定的研究理论、研究方法或研究问题,创新形成各类数据库框架,通过数据采集和更新形成大数据库[14]。本课题组 2015 年根据研究需要,系统采集了 109 所"211 工程"高校所有中国科学院研究所和所有中国社会科学院研究所的教师(专业研究人员)简历,形成了包括 14 万余条学者简历信息的大数据库,分析了学术人才的学业流动、职业流动等各类规律。在此基础上,根据研究需要部分匹配了学者的学术产出等新数据指标,以此分析学术流动与学术产出的关系等各类研究问题,取得了良好效果[15-16]。课题组还根据特定院校需求,分析了该校师资力量现状、学科建设人才需求、人才引进方向、人才引进成效以及与同类型竞争大学的师资力量对比等。

3. 用大数据算法替代传统统计学算法

传统统计学算法广泛应用于院校研究领域,其基本原理在于构建统计模型、降低抽样误差、形成相关性研究结论。而进入大数据研究时代,院校研究将有望用大数据算法替代传统统计学算法,进行全样本计算而不是样本估算,进行精确计算而不是模糊推算,进行因果关系计算甚至超级计算而不是相关关系演算。例如,本课题组 2019 年运用北京市车辆交通信息数据,测算教育是否为造成北京市交通拥堵的关键因素,涉及对几十万台汽车的大数据分析。该数据库每 15 秒更新一次车辆经纬度、速度等近百项数据,数百天的数据观测便形成包含数亿条数据的大数据资源库,通过超级计算可以还原北京市交通拥堵全过程,对教育因素的影响进行精准分析。再比如,本课题组基于上文提及的某荣誉学院全样本数据,通过识别学生整个学期内在食堂的刷卡记录,对每名学生整个学期早起行为进行全方位、持续性监测,共涉及几十万条刷卡记录。用更为客观、持续的事实大数据展开大数据算法,取代基于问卷调查等传统方法获取的主观性较强、小样本的横截面数据,可以展开更为科学、有效的研究设计,获取更为客观、精准的研究结论。将大数据思维运用到院校研究中,不仅可以大幅提升院校研究各项结论的有效性与科学性,而且将进一步拓宽院校研究的研究视角与分析思维。

二、大数据在院校研究中的应用举例

为更清晰地呈现大数据方法在院校研究中的使用,更好地展现大数据在院校研究中的各类优缺点,本文以某高校开展的"学生学习行为与学习效果评价"这一院校委托任务为案例,以高等学校最为常见的教育大数据来源——校园一卡通数据为例,对具体研究过程进行实证展示。

大数据与院校研究的结合涉及几个关键问题:一是数据库的构建;二是数据挖掘与匹配;三是大数据运算以及研究结论的形成。具体到本项委托研究任务,在大数据库建设过程中,分别存在学生基本信息数据、学生成绩数据和校园一卡通数据。除前两类传统技术数据外,校园一卡通数据包括学生消费金额、时间、地点(例如食堂、超市)等数据,转账充值数据,图书馆进出与借阅记录数据,寝室进出记录,浴室使用时间与频次数据,体育馆运动项目数据,乘坐校车出行时间与频率数据等。三类数

据共同形成新的研究大数据库，可以对大学生的学习、消费、饮食、读书、健康、卫生、就寝、运动、出行等进行全方位分析，以此探讨学生学业表现与学习行为之间的关联。

本研究的研究对象为北京某“双一流”高校某荣誉学院学生，收集了该学院 4 个年级共 499 名学生一学期产生的一卡通数据，共计 93 万余条。主要包括：(1)学生基础信息，包括姓名、性别、学号、班级、宿舍楼、宿舍号、校园卡号等；(2)一卡通消费及相关信息，包括一卡通使用时间、金额、流水类别、卡机编号、站点编号、充值记录等；(3)图书借阅信息，包括学生证件号、书籍信息(题名、出版社、ISBN 号、索书号)、借书日期、应还和还书日期、借阅登记、累计借书等。在一卡通数据基础上，进一步调取了研究对象的其他各类数据资源并进行了数据匹配，主要包括：(1)学生家庭背景信息，包括民族、生源地、政治面貌、家庭住址、父母工作性质等；(2)学生高考相关信息，包括学生高考总成绩、各科目成绩、生源地、毕业中学、入学类型(高考、保送)等；(3)大学学业表现信息，包括各学科成绩、总学分绩点、大学四六级成绩，出国情况、出国交流项目、出国交流时间、国外学校名称、最终是否出国深造等信息，学生获奖情况、论文发表等；(4)学习困难学生具体信息，包括学生挂科次数、挂科科目及具体分数，辅导员谈话次数、谈话记录，心理咨询情况，分流后学业发展等；(5)学生毕业信息，主要涉及学生毕业去向，如保研、出国、工作等。数据采集、清洗与匹配过程如图 1。

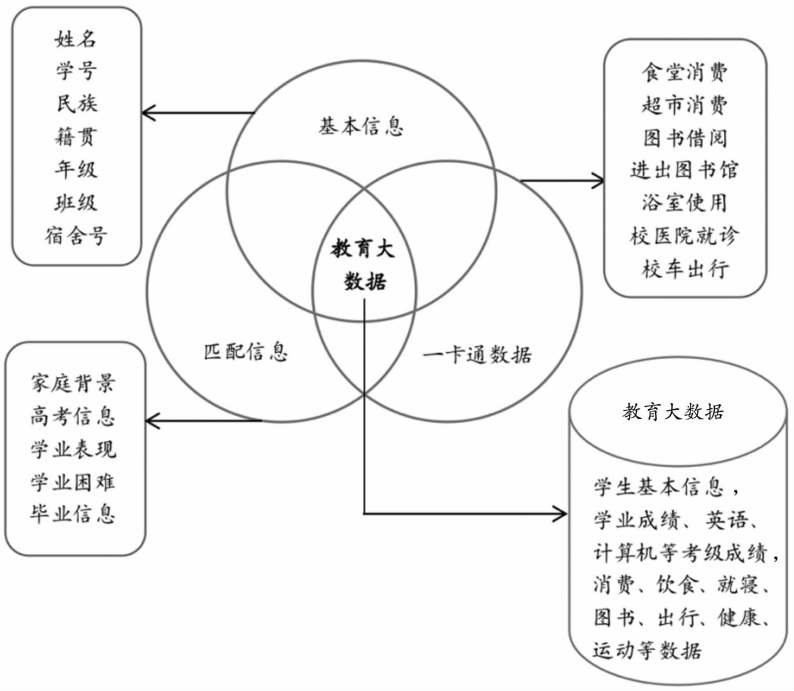


图 1　教育大数据采集、清洗与匹配过程

本研究数据采集涉及校内多个部门，主要包括学生工作处、招生就业处、教务部、学生事务中心、图书馆等。数据采集的遗憾之处在于，缺乏来自校医院的数据(以此考察学生的健康状况)和来自校车管理中心的数据(以此考查学生跨校区流动情况)。虽然如此，各类数据汇总已超过 100 万条，综合使用这些数据，可以分析学生的日常行为，可以对学生学业表现等展开综合研究，甚至可以进行有效的学生行为画像。

(一)研究举例 1：早起行为与学业表现的关系研究

高校通常鼓励学生早起进行学习活动，例如以早操、升旗、早读等方式鼓励学生早起。但学术界已有研究尚未发现早起与学业表现之间的直接关联，即早起行为是否真的能够提升学业表现。为此，本文以上述教育大数据库为依托，对学生早起行为进行大数据刻画。具体步骤是：(1)基于一卡通数

据构建早起值概念,根据学生校园卡早餐消费时间和地点,倒推学生的起床时间。首先定义上午 6:00~10:00 为早饭时间范围(10:00 之后可能为学生午饭时间),进一步确定 6:00~7:00、7:01~8:00、8:01~9:00、9:01~10:00 为 4 个早起的时间界定范围,并分别赋值为 4、3、2、1(即起床时间越早,分值越高)。(2)在此基础上,构建早起值公式 $Y=X_1+X_2+\cdots+X_n$。其中,X 为学生当天在早饭时间范围内在食堂最早一条刷卡信息的赋值(由于早饭时间范围内可能出现多次刷卡信息,例如去不同窗口购买不同类别的食物。因此,提取同一人同一天早饭时间范围第一条刷卡记录作为衡量其早起时间的数据,剔除早饭时间范围内其他刷卡记录),n 代表学生本学期在早饭时间范围内进入食堂刷卡的天数,将 n 个 X 值的总和定义为该生的早起值。(3)剔除部分“早起极低值”(可能跟个人生活习惯有关)后,对学生的早起值与其专业学分绩点进行相关性分析,结果显示,学生早起值越高,其专业学分绩点也越高。这在一定程度上表明,早起行为与学业表现具有正相关关系,即具有早起习惯的学生有更好的学业表现倾向。

此外,在此基础上,通过分析各年级学生早起情况差异进行学生早起行为划分。结果可以归纳为 4 类,按照早起值从高到低分别为:大二年级为“早起勤奋年级”,大一年级为“早起良好年级”,大四年级为“早起一般年级”,大三年级为“早起不佳年级”(见表 1)。即大一、大二年级的学生早起表现良好,而大三、大四年级的学生早起表现不佳,反映出不同年级生活习惯与学习习惯的差异。此案例中“学生年级越高,早起习惯相对越差”的问题应引起重视,其原因可能为低年级学生在一定程度上仍保持高中勤奋学习习惯,而随着年级增长该习惯逐渐消磨。据此结果,院校应引导学生形成长期的良好的生活和学习习惯。教育大数据在学生早起行为中的应用流程如图 2。

表 1　各年级早起行为的教育大数据特质推断

年级	学生人数	匹配人数	平均早起值	平均早起次数	教育大数据特质推断
大四	75	70	76.5	41	早起一般年级
大三	53	48	149.6	38	早起不佳年级
大二	79	75	166.0	76	早起勤奋年级
大一	292	286	165.0	61	早起良好年级

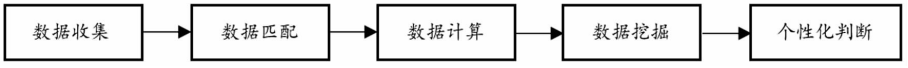


图 2　教育大数据在学生早起行为中的应用流程图

(二)研究举例 2:学生独立行为与同伴关系研究

同伴关系对于大学生个体发展以及高校人才培养与教育管理具有重要意义[17]。然而由于同伴关系较难测量的固有特点,关于大学生同伴关系的已有研究或进行理论性评述,或通过各类问卷调查进行测量,在大学生自主式填答“受欢迎程度”“肯定与欣赏”“亲密与交流”等测量同伴关系的相关问题时,会由于个人评判标准不同而在一定程度上导致主观性过强等问题,不利于研究结论的客观性呈现。以校园卡为依托的院校研究大数据恰好可为此类研究带来全新的求证机会与分析视角。

以本研究开展的对学生独立行为分析的研究为例,基于学生在食堂刷卡信息模拟学生独立行为,通过分析是否与室友一起进餐进行独立性评价,并据此分析学生独立行为与学业表现等方面的关系。基于研究需要,从校园卡大数据信息中筛选出 563 683 条食堂刷卡记录作为研究大学生独立行为与同伴关系的数据样本。首先,运用 Python 语言完成以下预处理过程:第一,抓取同一校区、同一寝室学生在每天三餐时间内的消费记录(该校为四人寝规格,每个寝室共 4 人;定义早饭时间为 6:00~10:00,午饭时间为 11:00~14:00,晚饭时间为 16:00~19:00);第二,以寝室为单位,分别筛选同一天三

餐时间范围内每人的第一条刷卡信息（同上，由于在同一进餐时间范围内同一人可能会在不同窗口产生多次刷卡记录，因此剔除同一进餐时间范围内其他刷卡记录）；第三，观测每位学生该条刷卡信息时间前后 10 分钟内是否有其他 3 位室友在同一食堂的消费记录，每有一位室友在同一地点进行过消费则计分为 1（共 3 个室友，因此同一进餐时间范围内每人最低得分为 0，最高得分为 3）。据此，得到每人每餐的独立性评价，按照此流程对每位学生进行整个学期的独立得分计算。分数越低，说明其独立性越强，即与室友亲密关系越弱；分数越高，独立性越弱，即与室友亲密关系越强。

研究结论表明，总体来看，独立性较强即与同伴关系较疏远的学生占比较大。据此，根据每位学生的独立得分可以对寝室进行归类。其中，“单人独立–三人抱团”型的寝室占比最高，占比接近一半；其次为“两人独立–双人抱团”型，占比为 36.61%；再次为“三人独立–单人亲密”型（即三人相对独立，另外一人与三人关系相对都较为亲密），占比为 12.24%；“四人分别独立”型寝室占比最小，仅为 1.61%。同时，通过分析学生独立得分与学业成绩的关系发现，对于男生而言，独立性评价与学业表现、学业失败显著相关，独立性越弱、亲密关系越强，其成绩均分越高，挂科次数、挂科率越低，而对于女生而言则无显著相关关系。此外，通过分析不同类别学生独立得分还发现，女生独立性较男生更高；社会经济地位越高，独立性越强；汉族学生独立性比少数民族学生强；单亲家庭、非教师家庭学生独立性较强；大四学生独立性最强，大一新生亲密性最强。研究分析部分举例见图 3。据此可以深入探究大学生同伴关系与学业表现之间的内在联系，以一卡通大数据为依托可以大幅提升各类院校研究结论的科学性与客观性。

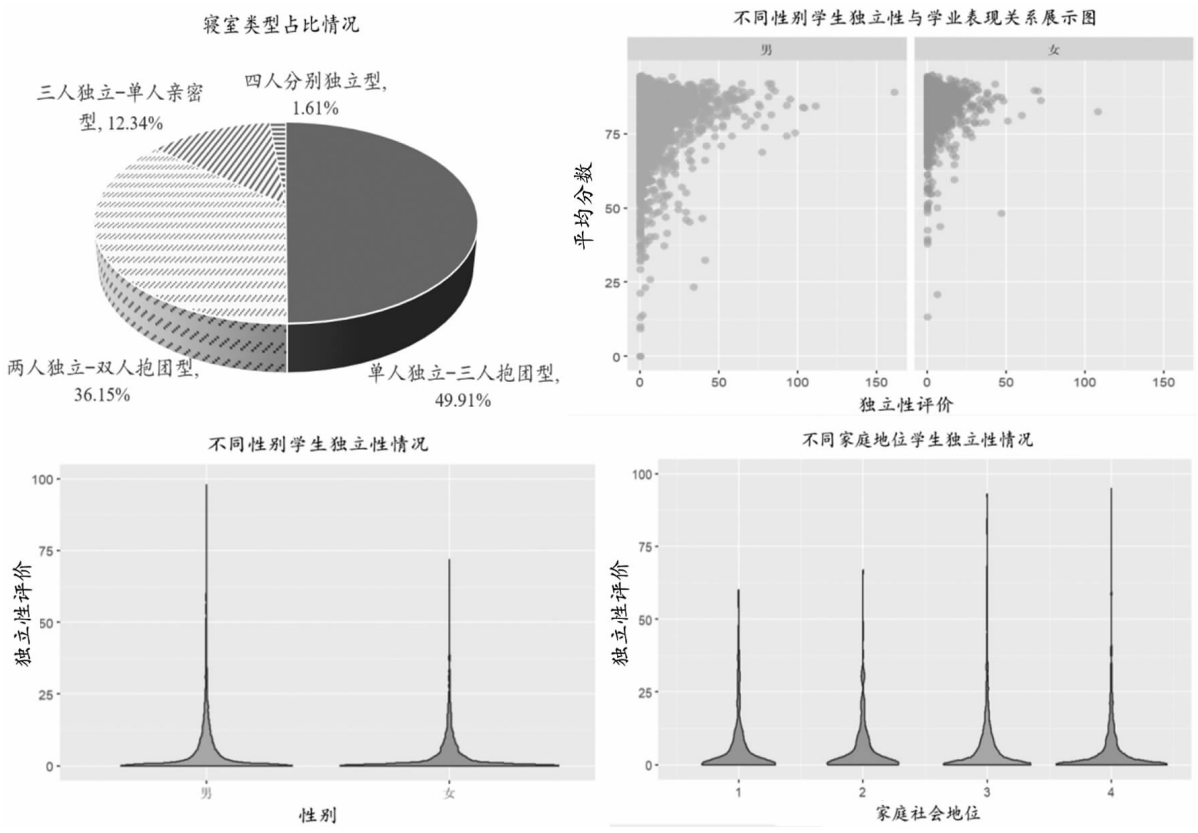


图 3　大学生独立行为与亲密关系研究分析部分举例

（三）研究举例 3：生源地与消费水平的关系研究

受既有认知与以往研究结论的束缚，西部生源学生家庭收入水平相对较低因而消费水平也较低往往成为固有认知，大数据研究或将颠覆以往的惯有结论和认知。在上述教育大数据库的基础之上，

本文以某年级生源信息为例,研究不同地域生源的消费水平是否存在差异。具体步骤为:(1)按照国家统计局的相关标准,将生源地域划分为东部、中部、西部(分别标为 1、2、3)3 个地理区域。(2)通过教育大数据库对每位学生的家庭年收入与一学期校园一卡通消费情况(包括该学期内该生在不同卡机如食堂、超市、浴室、校车等各站点的所有消费记录)进行精准匹配,并统计每位学生该学期的总消费值。在此基础上,剔除部分极低特殊值(可能与个人付款习惯有关),对该学期所有学生校园一卡通消费情况的地域差异进行分析。各数据之间的精准匹配样例见表 2。(3)通过对整学期学生一卡通消费大数据的匹配与分析发现,学生家庭年收入具有明显地域差异,西部生源家庭年收入整体水平较东、中部生源低,但学生在校整体消费水平并无明显地域差异,且家庭收入较低的学生消费水平高于家庭收入较高的学生等情况也并不罕见。虽然学生的消费面较为广泛,一卡通消费只是其中一类,但作为学生校园生活的最主要消费表现,其在一定程度上反映的学生消费水平具有较高代表性。通过跟踪一学期学生消费情况,得到西部生源并非消费水平较低的结论,颠覆了以往对西部生源的部分刻板印象。

表 2　学生地域、家庭年收入、消费信息的精准匹配举例

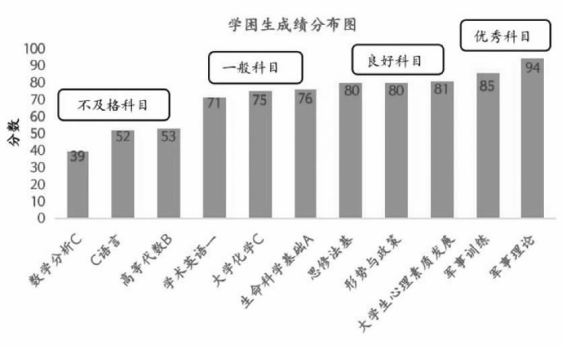
序号	学生	地域	家庭年收入(万元)	消费(元)	序号	学生	地域	家庭年收入(万元)	消费(元)
1	保＊＊	3	10	3 517	11	武＊＊	1	20	1 311
2	晁＊＊	1	30	5 776	12	王＊＊	3	15	3 250
3	单＊＊	2	15	3 909	13	徐＊＊	3	9	4 131
4	韩＊＊	1	5	3 696	14	杨＊＊	1	20	3 050
5	胡＊＊	3	30	2 663	15	尹＊＊	1	30	3 023
6	金＊＊	2	10	1 650	16	孙＊＊	1	15	3 401
7	李＊＊	1	6	2 552	17	孙＊＊	2	10	2 354
8	卢＊＊	1	10	2 825	18	张＊＊	3	9	4 050
9	吕＊＊	1	8	3 610	19	王＊＊	1	10	2 812
10	吴＊＊	3	5	2 498	20	王＊＊	2	10	4 380

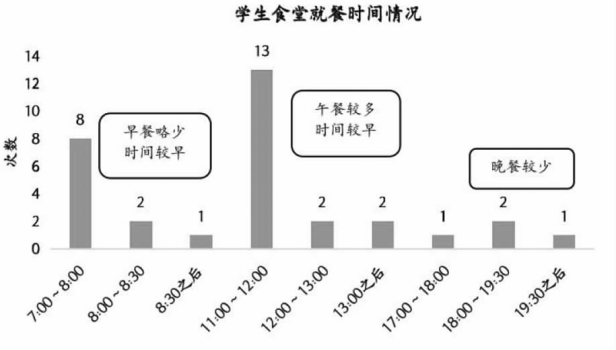
(四)研究举例 4:学生行为刻画和数据画像

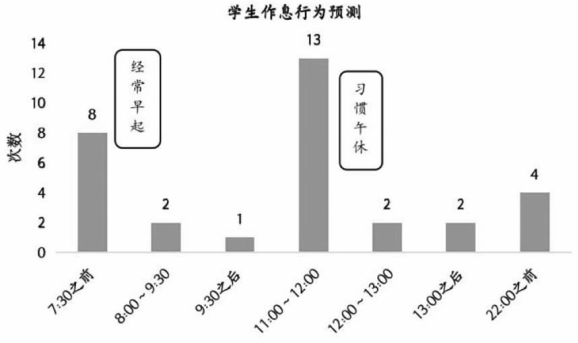
借助大数据库,还可在学困生的研究上发挥大数据的特色优势,全方位考察学困生的行为表现,进一步为学困生精准画像[18]。具体步骤是:(1)借助教务处学生历年各科目考试成绩数据,分析其学业成绩,整理其优秀科目、良好科目、不及格科目;(2)匹配该学生一学期的校园一卡通在食堂的刷卡信息,分析其每日用餐特征,并归纳总结其一学期内在食堂就餐的规律;(3)通过学生一卡通数据监测其作息规律,推测该生作息习惯;(4)通过图书借阅数据,监测分析该生的阅读偏好和习惯。通过对多个端口的教育大数据进行整体归纳和总结,对个人的学习行为和学业表现进行精确的个性化分析。

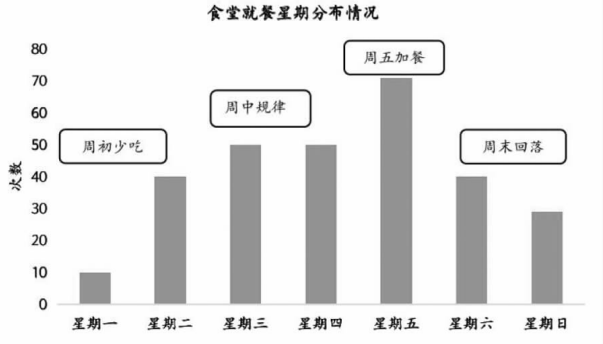
以某学困生为例,依托其校园一卡通数据对其学业表现行为进行分析。分析结果如图 4。该学困生基础学业成绩信息为:一学期中,3 门课程不及格,为表现不佳科目;3 门课程为 70~80 分,为表现一般科目;4 门课程为 80~85 分,为表现良好科目;1 门课程在 90 分以上,为表现优秀科目。匹配该学生一个月的校园一卡通数据可知,该学生一日就餐中,具有“早餐频次略少、时间较早,午餐频次较多、时间较早,晚餐很少”的特征。此外,根据其就餐时间也可发现其作息规律。例如,根据其早餐时间可推测该生习惯早起,属于“勤奋早鸟型”,午餐大多分布在 11~12 点,属于较早的午餐时间,可推

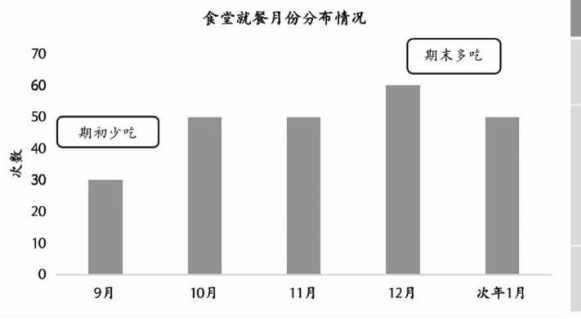
测其具有午休习惯，或可判断其倾向于避开用餐高峰期。分析该学困生一周内在食堂就餐情况，可归纳该生为“周初少吃、周中规律、周五加餐、周末回落”的类型（可能原因为，作为每周工作日第一天，周一时学生尚未回归至正常作息，而周末在食堂就餐频次回落可能与学生外出或以外卖形式就餐有关）。分析该学生整个学期每月刷卡就餐情况，可知其 12 月份在食堂就餐频次最高，可归纳该生为“期末多吃食堂型”或“期末饮食规律型”（这可能与期末临近，学生需要认真备考，从而饮食作息趋于规律有一定关系）。未来可进一步对学生进出寝室的一卡通刷卡数据进行研究，分析和推测其是否存在晚睡行为等。此外，通过该学困生校园一卡通在图书馆终端的数据分析，可得知其借阅习惯，该生该学期只借阅了两本图书，均与学习相关（第一本为“如何高效学习”，反映该生有意提升学习效率；第二本为专业工具书，反映该生有意提高专业知识水平），借书行为发生在上半学期末与下半学期初，在一定程度上反映出该生在学期初与学期末抱有“认真学习”的心态，但学期中并无借阅行为，且借阅时间均较短（或可推测该生在学期初与学期末更易激发学习斗志，但并未养成长期的良好学习习惯）。由此总结该生借阅的个性特征为“学期初/末借书”“借书少”“借阅时间短”以及“借书偏好为工具书”，同时在具体时间上还可发现该生偏好下午进出图书馆的特点。











一卡通学生借阅行为分析				
学年借阅量	借阅日期	还书日期	借阅书目	进馆时间
2本	2018-11-27	2018-12-03	如何高效学习	18:14:50 (借) 15:19:12 (还)
	2019-02-21	2019-03-25	C语言程序设计	15:08:44 (借) 16:44:47 (还)
读书少	借阅时间短		工具书	下午进馆特点

图 4　某学困生校园一卡通数据反映学业表现行为分析图举例

再以另一学困生为例，通过数据整合刻画其精准画像（如图 5）。基础信息为：保同学，女，2018

级,汉族,青海人;家庭年收入 10 万元;高考成绩 577 分,其中数学 133 分,理综 212 分。基于此可清晰得知其来自西部省份,数学基础尚可,理科基础较差。一学期学业行为表现为:早起得分 231 分(2018 级学生平均早起值为 165 分),早起次数 85 次;借阅图书 2 本,借阅类型为工具书。一学期生活习惯表现为:校园卡刷卡消费约 3 500 元,属于中等偏低水平;日常饮食习惯较为规律。据此可推测该生“勤奋早起”“消费偏低”“借阅量少”“饮食规律”等各项学习与生活特征。基于此,通过分析其入学前基本信息和入学后行为表现,可以部分解释造成该生学习困难的因素大致为:学业基础较为薄弱、家庭环境优势欠缺、努力程度相对不足、不同地域固有差异(例如,大数据表明,西部地区学生在英语、计算机等科目上与东中部学生存在显著差异,且本课题组前期研究表明,西部学生存在学习心理压力较大、自我效能感较低等问题[19])。

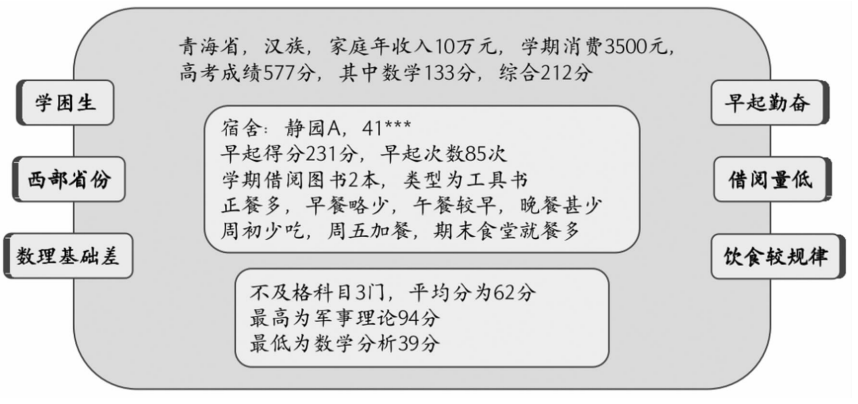


图 5　某学困生画像举例

三、大数据在院校研究中的应用展望

当前,人类社会已经逐步进入大数据和人工智能时代,院校研究活动应适时更新研究方法,这对于提高院校研究的科学化水平、更深层次推动院校研究与教育实践融合乃至重塑院校研究的合法性地位等都具有重要意义。本文对教育大数据研究方法的基本原理进行了初步探讨,选取了院校办学活动中最为常见的校园一卡通数据和最为常见的院校委托任务学生学习评价进行的大数据研究案例展示,研究显示教育大数据分布广泛,教育大数据研究潜力巨大,基于教育大数据方法的院校研究转向前景非常广阔。

通过本项研究,至少可以对大数据与院校研究的结合作如下总结:

第一,院校研究的大数据资源来自哪里。本研究显示,院校研究的大数据资源并非神秘莫测,而是广泛分布在教育教学活动中。校园一卡通是教育机构最为常见的大数据载体,在一卡通数据基础上,本文采集了学生家庭信息数据、入校前与入校后学业表现数据、各类行为习惯数据等,形成了有效的教育研究大数据库资源。本课题组拥有多份院校研究大数据资源,之所以用“一卡通”数据举例,是为了说明即使是校园里最为常见的数据,也可广泛用于院校大数据研究活动。事实上,除“一卡通”数据外,高等学校还包含文字、语音、视频等各类大数据资源,当前大量可用于院校研究的大数据资源呈现零星、散乱、沉睡状态,需要院校研究者构建大数据框架,进行收集、加工、整理、计算,形成各类院校研究大数据库[20-21]。这些院校研究的大数据资源像“老中医”,不仅不会过时,而且积累时间越久,数据维度越丰富,数据挖掘越充分,越可能更科学地指导院校研究和具体实践。

第二,院校研究如何使用各类大数据资源。研究显示,和普通问卷、访谈、质性研究素材不同,院

校研究的大数据资源使用具有多目的性和多功能性特征，但仍然主要遵循院校研究以问题为导向的数据使用基本原则，仍然按照“研究问题/委托任务—数据收集—数据分析与结论提出”的基本流程展开研究活动。所不同的是，一旦数据库生成，则可能形成大量研究目标之外的“副产品”，这将有利于新的理论创新与实践创新。以本文为例，本文最早的研究任务是分析大学生学业失败的影响因素，但研究活动结束后，还大量呈现了其他各类研究结论。大数据库通常具有海量性、多样性、高塑性和异变性等特征，一项研究任务完成后，随着大数据库的持续更新，以及新的数据变量加入，或者对传统数据变量的再赋权、再加工、再计算，也可能满足更多的研究需求。事实上，除校园一卡通数据之外，本项目团队还帮助某校建立了包含教师教学信息、学生选课信息、学生评教信息、学生运动表现信息（体能测试结果）等在内的更大的院校研究大数据库，这将有利于更全面、更深入、更系统地推进院校研究与大数据的融合。比如，传统的学生对教师评教数据通常并不准确，无法真实判断教师教学质量。通过上述大数据库，则可以进行新的大数据评教。以“大学物理”为例，对于A教师的“大学物理”授课水平评价，除传统学生评教分数等观测点外，还可以分析参加A教师课程学习的学生进入后续与大学物理应用相关的其他课程学习的大数据表现，反向倒推A教师“大学物理”的授课质量。

第三，院校研究如何科学呈现大数据研究结果。相比于传统的研究方法，大数据方法在院校研究中的使用重点在于数据本身，数据库建设与研究目标的一致性决定了研究的有效性，数据质量则决定了研究质量。在此过程中，核心工作是数据资源的探寻、收集、加工和运算，但传统SPSS等封闭式数据统计类研究工具存在局限，需要引入Python，MATLAB等研究工具，甚至在一项研究活动中需要运用多种工具。也因为各种统计类、绘图类研究工具的加入，基于大数据的院校研究结果呈现出可视化程度更高、直观性更强、美观度更好、更能直接指导教育改革等特点[22]。以本文校园一卡通数据为例，研究首次发现的西部学生因学业基础不佳、入校后学业困难的问题已经引起项目委托单位的注意，并采取相应举措加强对该部分学生的学业指导和学习帮扶。这显示出，基于大数据方法的院校研究活动，研究设计科学性更强，数据占有量更大，研究结论可靠性更高，研究发现更多为因果式推断，研究结论的呈现将更可能推动教育实践改革，并加快从学术成果到实践成果转换的速度。

第四，院校研究用好大数据方法的难点。和传统研究方法一样，教育大数据研究方法也存在自身的缺陷和不足，教育大数据研究的难点在于保证数据库构建的过程和最终研究结果的有效性。从本文所举例的校园一卡通大数据研究活动来看，基于大数据方法推进院校研究的思路相对简单。难点一在于数据收集、数据清洗整理等环节。相比于传统研究，本研究的数据收集过程复杂度更高，涉及在学校不同部门之间进行沟通协调（这也反映出学校层面的数据库建设滞后[23]），有些教育管理部门缺乏大数据思维和能力，甚至因保管不力导致部分数据丢失，因此需要加强推进院校研究资源的共享与信息采集的规范化[24]。与此同时，大数据库建设过程中的数据加工整理是工作量最大的环节，看起来简单的研究结论，背后却牵涉不同数据库资源的整合、对无效数据的剔除、对数据画像的标准设定以及海量的数据运算。难点二在于与传统理论与研究的对话。大数据研究方法直接切入研究问题本身，并未遵循“理论（文献）—假设—数据—发现”这一基本逻辑，研究结论往往“就事论事”，容易引发研究活动理论性不强等质疑。本文认为，大数据研究有可能将院校研究带入碎片化阶段，如何通过一个个小的教育问题的发现和解决、重构院校理论体系，是大数据时代院校研究需要深入思考的问题。

第五，在院校研究活动中推广大数据方法的建议。本文认为，院校研究应尽快转向大数据方法，应更加强调院校研究的科学化目标、实践导向和问题解决能力[25-26]，应更好地营造证据导向、数据导向的院校研究场域，加强教育大数据相关立法以降低数据获得门槛，形成新的以大数据方法为准则的院校研究学术共同体，引导院校决策规划、相关学术论文发表、院校调查报告、教育教学改革等院校研

究活动不断向大数据方法靠拢,提高院校研究的问题针对性、过程透明性、研究趣味性和结论可用性。与此同时,应注意大数据使用规范,用数据说话的同时避免唯数据等问题的出现。

本文撰写的初衷是抛砖引玉,通过对某项院校研究大数据研究活动的剖析,深化院校研究者对于大数据研究方法的认知,帮助一些院校研究者克服对于大数据研究神秘化、技术化、困难化的畏惧,推动营造“处处皆有大数据、人人用好大数据”的新院校研究场域,带动院校研究大数据方法的推广和使用。但本文研究过程中涉及的数据样本仍然有限,相关研究发现仍停留在早期阶段,大数据方法运用仍然存在不规范、不深入、不系统等问题,院校研究大数据方法论的系统构建,需要海量的类似于本文的案例支撑[27],亟待院校研究理论与实践界行动起来,共同推动本项研究改革。

参考文献:

[1] 赵炬明. 科学管理与院校研究[J]. 高等教育研究,2007(7):55-58.

[2] 刘进. 人工智能如何使教育研究走向科学[J]. 高等工程教育研究,2020(1):106-117.

[3] 汤贝贝,薛彦华. 大数据背景下高等教育治理转型:机遇、挑战与应对策略[J]. 重庆高教研究,2019,7(2):77-86.

[4] 张俊超. 院校研究如何通过数据分析为大学管理决策服务:“院校研究数据分析的对象、内容和方法”研讨会暨 2013 年中国院校研究会年会综述[J]. 高等教育研究,2013,34(8):105-109.

[5] 陈廷柱,孙丽芝. 变革中的高等教育及其对高等教育研究的挑战:中国高等教育学会高等教育学专业委员会 2013 年学术年会综述[J]. 高等教育研究,2013,34(12):102-105.

[6] 张德祥,别敦荣,周光礼,等. 加强新时代教育科学研究工作(笔谈)[J]. 中国高教研究,2019(12):1-9.

[7] 张俊超. 大数据时代的院校研究与大学管理[J]. 高等工程教育研究,2014(1):128-135.

[8] 周光礼. 相似的挑战、不同的逻辑:院校研究的中国化[J]. 高等工程教育研究,2017(2):118-121,133.

[9] 刘献君,余东升,陈敏,等. 高教研究转型迫在眉睫[N]. 光明日报,2016-03-17(15).

[10] 彭雪涛. 美国高校数据治理及其借鉴[J]. 电化教育研究,2017,38(6):76-81.

[11] 陈武元. 高教研究如何补齐“短板”冲出“困境”[N]. 中国科学报,2019-11-20(04).

[12] 刘进,陈建. 中国高校自主招生地方保护主义的大数据分析[J]. 上海教育科研,2016(5):5-10.

[13] 刘进,陈健,杜娟. 弱势地区自主招生参与的公平问题研究[J]. 高校教育管理,2016,10(3):54-59.

[14] 周川,蔡国春,王全林,等. 院校研究:高等教育研究的新领域[J]. 高等教育研究,2003(3):46-51.

[15] 刘进,哈梦颖. 世界一流大学学术人才向中国流动的规律分析:“一带一路”视角[J]. 比较教育研究,2017,39(11):26-33.

[16] 刘进. 学术职业流动:中日对比研究:中国 M 大学与日本 N 大学的教师流动情况实证分析[J]. 中国高教研究,2019(4):55-63.

[17] 李佳哲,元静,胡咏梅. 本科生宿舍同伴关系的测量及其异质性研究:基于对某高校教育学专业本科生的调查[J]. 重庆高教研究,2019,7(6):116-128.

[18] 巫芯宇. “新四科”背景下大学生信息素养教育质量提升路径探究[J]. 重庆文理学院学报(社会科学版),2022,41(1):114-126,140.

[19] 刘进,林松月,王艺蒙,等. 西部生源大学生学业表现的多维对比分析:基于 B 大学 N 学院学生的实证研究[J]. 重庆高教研究,2019,7(4):12-29.

[20] 李兴华,刘俊学,罗元云. MOOC 背景下教学流程再造的内涵与路径[J]. 重庆高教研究,2017,5(1):18-22.

[21] 刘献君,赵炬明,陈敏. 加强院校研究:高等学校改革和发展的必然要求[J]. 高等教育研究,2002(2):54-58.

[22] 侯雨欣,王冲. 大学生信用评价指标体系构建与优化:基于高校管理视角[J]. 四川师范大学学报(社会科学版),2022,49(3):130-137.

[23] 雷洪德,黄敏. 从美国院校研究的三重三轻看中国院校研究的挑战与突破[J]. 高等工程教育研究,2017(2):128-130.

[24] 余东升,陈廷柱. “院校研究与现代大学管理”国际学术研讨会综述[J]. 高等教育研究,2005(1):103-105.

[25] 周光礼,莫甲凤. 高等教育智库及其学术研究风格:中国著名高等教育研究机构的学术转型[J]. 高等工程教育研究,2014(6):45-57.

[26] 别敦荣. 加强中国的院校研究理论建设 助推高校提升办学水平[J]. 中国高教研究,2016(10):24-26,77.
[27] 张应强. 我国院校研究的进展、问题与前景[J]. 高等教育研究,2011,32(12):40-45.

(编辑:杨慷慨　校对:张　腾)

Big Data and Institutional Research

LIN Songyue[1], LIU Jin[2]
(1. *Faculty of Education*, *The Chinese University of Hong Kong*, *Hong Kong* 999077, *China*;
2. *School of Humanities and Social Sciences*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: Institutional research has the natural advantage of using big data methods to carry out research activities, but the existing research has not fully explored the conditions and ways suitable for big data research, and it has not formed a good big data research demonstration. Replacing the traditional research thinking with big data thinking, replacing the limited data of traditional research with big data, and replacing the traditional statistical algorithm with big data algorithm are the basic technical principles of the combination of college research and big data. Based on the in-depth discussion of the application principle of big data research methods in colleges and universities, taking the most common carrier of big data in colleges and universities, all-in-one campus card, as the analysis object, a case analysis was made on the distribution, generation, storage, collection, using and other processes of all kinds of big data in the process of college research. The study shows that the big data method has subversive significance for the paradigm, process and results of institutional research, and it also has good integration ability between big data and institutional research; big data in institutional research is widely distributed in education and teaching activities, and the integration of institutional research and big data has broad prospects. Future institutional research should put more emphasis on the scientific goal and practice orientation of research, better create an evidence-oriented and data-oriented research field, and form a new academic community of institutional research based on big data methods.

Key words: big data; institutional research; campus card; big database; academic community