

欢迎按以下格式引用:刘锦行.高校科研评价的关键技术、模型构建与对策建议——基于大数据视角[J].长江大学学报(社会科学版),2022,45(2):119-124.

高校科研评价的关键技术、模型构建与对策建议

——基于大数据视角

刘锦行

(湖北工业大学 计算机学院,湖北 武汉 430068)

摘要:大数据的引入为高校科研评价提供了新的方法。文章基于大数据的视角,对高校科研评价的数据采集与挖掘、指标集与指标权集的构建、数据管理与计算平台、多模态可视化表达等方面的关键技术进行探讨并构建了相关模型。在大数据背景下,高校科研评价应加强基于大数据的科研质量评价理论研究,强化科研评价的“价值”取向,建立大数据库与搭建科研评价平台,立法规范与加强评价监督,发挥同行评价的作用。

关键词:大数据;高校科研评价;关键技术、模型构建

分类号:G644 **文献标识码:**A **文章编号:**1673—1395 (2022)02—0119—06

先进而科学的高校科研评价是推动高等教育高质量发展和促进科技创新的重要因素。充分利用大数据的资源,科学运用大数据技术评价高校科研质量和水平,对于促进高校对国家科技创新的贡献具有重要的意义。本文就在大数据背景下高校科研评价的关键技术、模型构建及对策建议进行一些探讨。

一、基于大数据的高校科研评价的关键技术与模型构建

(一)高校科研数据采集、挖掘等关键技术与运行机制

从海量的数据中采集和挖掘基本数据是高校科研评价的基础,探索数据采集、挖掘等关键技术与运行机制是做好高校科研评价的前提。

1.高校科研数据采集

高校科研数据来源广,主要有政府科技管理部

门、企业、大学和科技工作者个人公开的数据,有SCI、SSCI等检索系统的数据和中国知网(CNKI)等数据库数据以及其他多媒体与互联网上的数据。在采集的过程中,需要针对具体的业务需求对数据进行整理,例如对非法、虚假数据进行过滤,对不规范数据进行格式转换与数据规范化处理等。目前的数据采集主要有离线采集、实时采集和互联网采集三种主要方法。针对不同类型的数据,可分别采用相应的技术进行采集。

(1)结构化数据采集。对于来自关系数据库和excel等格式的项目、成果、经费等结构化数据,可采取离线采集、实时采集和网络蜘蛛(或数据机器人)采集、索引系统等技术综合采集。

(2)半结构数据采集。对于来自XML和JSON等格式的半结构化数据,如论文影响因子、他引量、ESI值等数据,可通过实时采集、R语言定义数据分

收稿日期:2021-11-28

基金项目:国家自然科学基金青年科学基金项目“融入情感分析的多模态社交媒体用户画像研究”(62106070);湖北省教育规划重点项目“大数据背景下的高校科研质量评价方法研究”(2020GA032)

作者简介:刘锦行(1986—),男,湖北仙桃人,讲师,博士,主要从事人工智能与大数据、用户画像与推荐系统研究。

析规则、正则表达式和语义分析互连网络方法综合采集^[1]。

(3)非结构化数据采集。非结构化数据也是有价值的数[2],对于来自文本文件、视频、音频和图片等格式呈现的人员、项目、成果、经费等非结构化数据,可通过搜索引擎和主题搜索引擎的互连网络采集、语义分析采集。

2. 高校科研数据预处理

在采集高校科研数据过程中,有许多数据特别是半结构化以及非结构化数据不可避免地会出现异常值、缺失值、拼写错误和不完整或者重复的数据等问题,比如期刊论文作者单位缺失、科技论文在不同期刊出现和期刊 IF 动态变化等问题,就必须对采集到的数据进行预处理,以解决数据质量问题。

数据预处理主要包含了数据清洗、数据融合、数据表征和数据降维等方面。一般通过箱线图等技术,SPSS、统计学方法或采用分类、聚类方法等技术,噪声平滑方法,卡方检验法等方法,对数据进行清洗、融合;采用数学函数对每个数据属性值进行映射,按比例缩放数据的属性值,对有关数据采取功效系数法进行无量纲处理,将数据进行转换,表征为适合于下一步处理的表示形式。同时,通过小波变换或主成分分析等降维方式,对数据进行降维,得到原数据的归约或“压缩”表示。

3. 高校科研数据的挖掘

数据挖掘的任务有分类分析、聚类分析、关联分析、异常分析、特异群组分析和演变分析等。

常用的数据挖掘的方法有决策树法、关联规则法、神经网络法、遗传算法、模糊集和粗糙集法。决策树利用输入变量的值进行递归划分以预测因变量,通过构成决策树来预测在输入变量的大小关系之下预测值的分布概率,可用于判断一个未知事项的风险大小,为其可行性决策提供依据;关联规则主要用于发现事物之间的联系,这种联系完全基于历史数据,不同于人的大脑所理解的因果、相关等关系;人工神经网络是一种模仿自然生物神经网络处理信息和记忆信息过程的数据模型,本质上是一种由大量神经元连接而成的运算模型,每个神经元都是一个激励函数,节点之间通过权重连接。该网络依靠大量权值和阈值来进行学习、处理和记忆信息;遗传算法是一种基于遗传结合、遗传交叉变异及自然选择等遗传学规律和自然流变的随机并行搜索算法及机器学习方法,它的基本原理是适者生存法则;粗

糙集法是一种处理不精确性和模糊问题的现代数学工具^[3],在进行数据挖掘时,基于粗糙集原理对数据属性进行约简,通过约简操作降低属性的维数,规范成可支持决策的数据。

同时,针对高校科研数据背后的隐性数据、滞后性数据、长效性数据的挖掘,可分别采用相应的技术进行挖掘并提取有价值的信息。①隐性数据挖掘:可采用模式识别、机器学习、图谱构建与挖掘、事件分析法等技术,对基础研究中原创成果隐藏的科学价值进行挖掘,并预测产生的连锁效应对经济社会的科技贡献率等数据。拟采用语义分析法等办法对科研论文中隐藏的科学价值进行挖掘,洞察学术论文深层面的价值内涵^[1]。②滞后性数据挖掘:可采用回归分析、情感计算等技术,对高校科研在成果推广转化(含专利、软件著作权)经济效益数据进行挖掘;同样对高校教学研究成果等滞后性数据进行挖掘。③长效性数据挖掘:可采用关联规则、并行计算等技术,对高校科研人才在培养方面(研究生的培养和队伍成长)的贡献价值和促进学科建设(硕士点、博士点以及一流学科建设)的贡献价值进行挖掘。

4. 高校科研数据存储

在大数据处理流程中,数据存储与管理是十分重要的一环。当前大数据领域中,分布式文件系统的使用主要以 Hadoop HDFS 为主。HDFS 采用了冗余数据存储,增强了数据可靠性,加快了数据传输速度,除此之外,HDFS 还具有兼容的廉价设备、流数据读写、大数据集、简单的数据模型、强大的跨平台兼容性等特点。常用数据库技术大体有 SQL 关系数据库、高性能的非关系型的 NoSQL 数据库和 NewSQL 数据库。

(二)高校科研评价指标集构建模型与运行机制

通过大数据分析数据的特征和核心价值,揭示各类指标间关系,提取有价值的代表性指标,探索指标集构建模型与运行机制是关键。

1. 高校科研评价指标提取

从高校科研标准数据来源中提取高校科研指标是构建高校科研评价指标集的关键。可用神经网络技术构建科研评价指标提取模型,并通过数据训练优化提取模型。首先用德尔菲法从标准数据来源中选择部分评价指标,接着模拟生物神经网络,用神经网络方法,通过不断学习的复杂函数,找出适应评价的指标体系规律,构建科研评价指标提取模型,然后用选择到的评价指标作为样本大数据来训练并优化

模型,最后用已构建的模型淘汰关联性不高的指标,提取出对高校科研质量和水平有重要影响的指标,得到反映高校科研质量和水平的代表性指标。

2. 高校科研评价指标集模型构建

可采用机器学习的聚类分析法对指标聚类,并构建高校科研评价指标多级多层次体系。

(1) 从评价指标中随机选择科研条件、科研过程、科研成果和科研管理等 K 个评价指标作为评价指标的聚类中心。

(2) 用 n 维空间词频数作为指标的坐标 X 和 Y , 该坐标到聚类中心的欧氏距离 $dist$ 见公式(1), 设定某个距离为阈值, 则阈值范围内的指标归并为一个簇:

$$dist(X,Y)=\sqrt{\sum_{i=1}^n(x_i-y_i)^2}$$

(1)

(3) 设定一个值 $r_{nk}=1$ 表示为指标在该聚类中心, 0 表示该指标不在该聚类中心。用 u_k 表示所有聚类中心 k 的数据点平均值, 计算每个聚类中心簇内所有点到该聚类中心距离的平均值 J , 见公式(2):

$$J=\sum_{k=1}^k\sum_{j=1}^nr_{nk}\parallel x_i-u_k\parallel^2$$

(2)

当 J 的值足够小, 说明该聚类中心簇内指标相关性大。若 J 值大于某一阈值, 则重新选择聚类中心重复步骤(2)和(3), 直到所有聚类中心不再发生改变。

(4) 根据模型运行结果整理多级多层次的高校科研评价指标集。

(三) 高校科研评价指标权集的构建模型与运行机制

构建指标权集模型与机制是高校科研评价的重点和难点。在构建科研指标评价体系后, 可用深度学习结合层次分析法构建、训练和优化高校科研评价指标权集模型。其运行机制如下:

1. 问题分析与模型假设

假设通过上面的研究确定了科研条件、科研过程、科研成果和科研管理等 T_1 、 T_2 、 T_3 和 T_4 4 个聚类为一级指标的递阶层次结构, 其中 T_2 有 T_{21} 、 T_{22} 、 T_{23} 、 \cdots 、 T_{26} 等 6 个二级指标。构建指标相对上层指标重要性的判断矩阵。根据判断矩阵, 用 R 语言计算每个判断矩阵的特征向量 W 、最大特征根 c 和一次性指标 CI , 结合随机一次性指标计算每个指标的特征向量。根据每个指标的特征向量, 用 R 语言计算并比较每个不同评价方案的决策组合向量。

2. 模型符号定义

对模型中出现的符号做如下定义, A : 代表模型的目标层; B_j : 代表准则层第 j 个指标 ($j=1, 2, 3, \cdots, 6$); C_i : 代表子准则层第 i 个指标名称 ($i=1, 2, 3, \cdots, 12$); D_q : 代表方案层第 q 个指标名称 ($q=1, 2$); 准则层对目标层的特征向量矩阵用 w_1 代表; 子准则层对准则层特征向量矩阵用 w_2 代表; 方案层对子准则层的特征向量矩阵用 w_3 代表; CI 代表一次性的指标; CR 代表随机一次性指标; Z 代表决策组合向量。

3. 建立科研评价指标权集模型

在递阶层次结构中确定指标 C_1, C_2, \cdots, C_6 对于准则层 B 相对的重要性即权重, 用两两比较方法确定权重。对于准则层 C 的指标 C_i 和 C_j 按重要程度依次递增赋 1 ~ 9 权重, 构造准则层 C 与其二级指标的成对比较判断矩阵。

4. 层次单排序及一致性检验

需要对验证构建出来的正互反判断矩阵 A 是否存在严重的非一致进行一致性检验, 以便确定是否接受 A 。一致性检验步骤如下:

首先采用 R 语言通过公式(3) 计算出一致性指标 CI 。然后搜寻对应的随机一致性指标 RI , 对 $n=1, \cdots, 9$ 给出 RI 的值, 见公式(4):

$$CI=(\lambda_{\max}-n)/(n-1)$$

(3)

$$RI=(\lambda_{\max}-n)/(n-1)$$

(4)

用随机的方式构建多个指标因子组成的矩阵, 从 1 ~ 9 及其倒数中随机抽取数字构建正互反矩阵, 得出最大特征根值, 并用计算出的 CI 和 RI 的值计算矩阵的一致性比例 CR , 见公式(5):

$$CR=CI/RI$$

(5)

通过公式 5 测算准则层的判断矩阵 cr_0, cr_1, \cdots, cr_5 的一致性比例。

5. 层次总排序及一致性检验

对层次总排序做一致性检验, 判断各层次积累的非一致性是否引起分析结果严重的非一致性, 是必需的。设 B 层中与 A_j 相关因素的成对比较判断矩阵对应的单排序一致性指标为 $CI(j)$, 相应的平均随机一致性指标为 $RI(j)$, 用公式(6) 计算 B 层总排序随机一致性比例如下:

$$CR=\sum_{j=1}^mCI(j)a_j/\sum_{j=1}^mRI(j)a_j$$

(6)

通过 CR 值的比较, 可以获得比较好的评价效果。依此类推, 对同层次或不同层次的高校科研评价指标权集进行建模分析, 最后选择在多个模型中

R 语言计算 CR 值最高的模型作为高校科研评价指标权集模型。在完成科研评价指标权集模型后,用德尔菲法对指标权集进行校验。

(四)高校科研评价数据管理与计算平台构建方法

高校科研评价计算平台包含大数据管理平台和数据处理两部分。可采用过程化方法设计多线程 B/S 模式的高校科研评价平台。高校科研评价平台包含数据采集层、处理层、分析层、访问层和应用层。针对平台要具备评测结果的可视化输出及用户交互功能,采用移动端和 PC 两种方式构建,并开发基于 Android 和 IOS 两种版本的数据计算处理软件^[4]。

软件开发采用 Java 作为开发工具、Linux 为系统的开发运行平台。采用先抽象后细化的方法,用系统流程图、数据词典、数据流程图和程序流程图等方法依次对系统进行概要设计和详细设计。系统代码完成并编译后依次采用白盒测试和黑盒测试的方法对系统进行单元测试、递增式集成和系统测试等测试并上线运行。

(五)高校科研评价结果可视化表达

根据政府、高校、评价机构的需求将高校科研评测结果整理成数据表,并将这些数值通过丰富多彩的视觉形式(色彩、位置、形状、方向、纹理、尺寸、值等)表现出来。可以 Google Charts 的 HTML5 和 SVG 为基础,通过 IE、谷歌、360 等浏览器,创建可交互和可缩放的图表传递给用户,或以组合视觉结构形式传递给用户、或转换后以音频、视频等形式传递给用户,用户通过人机交互的技术和手段进行反向转换,了解评价结果数据表达的状况和问题^[5]。

上述五个方面构成了完整的技术路线和运行机制,整体流程如图 1。

二、基于大数据的高校科研评价的对策建议

(一)加快基于大数据的高校科研评价理论与评价体系的研究

高校科研评价需要制定一套完整的评价体系,包括评价的对象、评价的标的、评价的功能、评价的指标集和指标权集、评价的计算系统、评价流程等,这是评价的平台和工具,不然就没有办法操作。

但高校科研评价体系的制定及操作需要与时俱进的新理论、新的方法论来指导和规范。具备科学的、专业的、成熟的、且符合党和国家要求方向的评价理论是开展一切评价活动的基础,不管是传统的

评价,还是基于大数据环境下的评价都离不开理论的指导。大数据环境的高校科研评价,传统的评价理论可能落伍,急需理论的创新和发展。若没有大数据评价理论的指导,高校科研评价就会失去目标和方向,海量的数据采集、存贮,先进的处理技术和分析方法,可能会与高校科研评价形成“两张皮”,以至于无法无缝衔接。为此,必须加快对于基于大数据的高校科研评价理论的研究,并在此基础上制定高校科研评价体系。

(二)强化高校科研评价的“价值”取向

习近平总书记指出:“要改革科技评价制度,建立以科技创新质量、贡献、绩效为导向的分类评价体系,正确评价科技创新成果的科学价值、技术价值、经济价值、社会价值、文化价值”^[6]。为此,科研评价过程中应始终把握好评价的“价值”导向。对于高校科研评价而言,不仅要着眼于科研成果的“五大价值”,而且还要考察科研过程的人才培养价值,引领高校探求未知、创新科技、服务社会、培养人才。这应该是我们必须强化的高校科研水平评价的“指挥棒”,而不是“排行榜”。

立足科研评价的“价值”取向,是构建成熟的高校科研质量和水平评价体系的根本,高校科研评价的“价值”导向是高校科研评价主体在对科研评价认识实践的过程中,产生的具有主导性的价值观,即评价主体对高校科研评价价值标准所取的方向,它表达的是评价主体对高校科研评价的价值立场和价值追求^[7],它对科技工作者乃至高校科技工作具有强烈的导向作用。可以说,正确的价值取向是构建科学合理的高校科研评价的根本。在实践中,我们不难发现,根据不同的价值取向所构建的标准是完全不同的。如果我们的评价取向侧重工具价值取向,把评价纯粹当做考核科技工作者工作量的手段,必然会导致科技工作者追求数量,忽视质量。久而久之,功利主义泛滥,个人主义盛行就在所难免,科研质量和水平就很难保证^[8],科研质量和水平的提高和改进就会大打折扣,科技对经济和社会乃至国防建设的发展的贡献将会受到严重影响,科技创新就无从谈起,教科结合、培养人才无法落地。为此,我们应该确立以科研“创新价值”为导向的科研质量和水平评价体系。这种价值取向具体应该体现在“鼓励原创,探求未知;创新技术,服务社会;促进教学,培养人才;反思自身,坚守使命”^[9]等方面。国家和教育、科技管理部门要根据这一价值取向,构建我国高校科研方向和目标、质量和水平的评价体系,科研

质量与水平一定有所改观和提升,对社会的进步和人类的文明将会有更大的贡献。

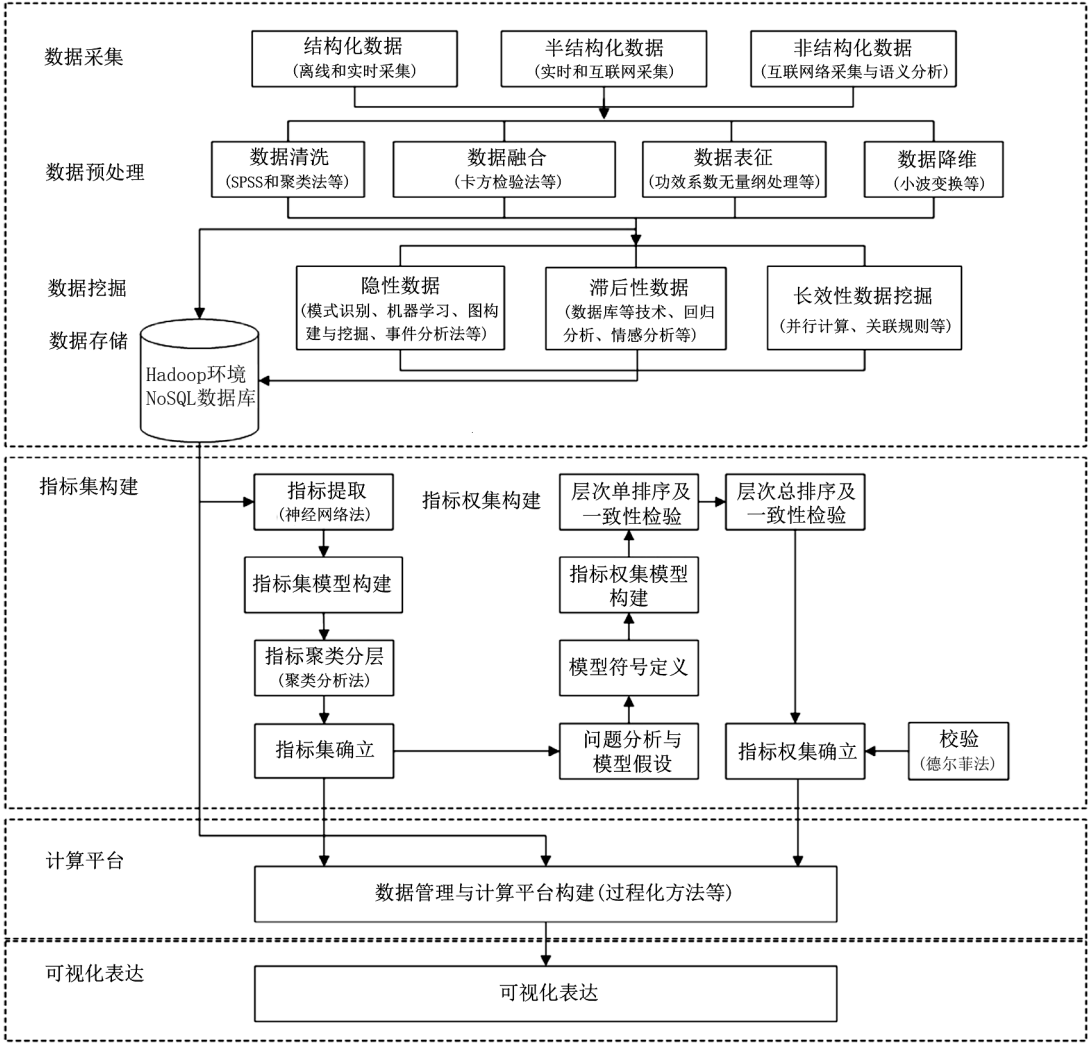


图 1 技术路线

(三)建立大数据库,搭建科研评价平台

大数据对于高校科研评价的实践应用还仅仅只是开始,任重道远,而当下的任务是加强高校科研评价的基础性建设,即全国统一的大数据信息资源库的建设,尽可能完备地将高校科研从立项到成果鉴定、应用全流程、从队伍到条件、经费全方位的相关信息集中在一个统一的高校科研信息平台中,并建立统一的分类和元数据标准等等。

为此,特提出如下几点设想:①建立国家层面的权威高校科研大数据库。建议教育部设立高校科研大数据库,收集和整理全国高校所有的科研人员和队伍、科研项目与经费、科研条件与平台、科研进展、科研成果、科技转化等方面的数据,特别是收集和认可科技成果方面的论文、专著、专利、获奖等方面的数据。②建立面向全国的高校科研评价管理网站。

网站公布各项科研数据,设立项目成果库,项目评审系统、项目推广应用管理系统等内容,为高校科研评价管理提供有效平台,对把控高校科研项目起到很好的作用。③规范高校科研网站管理,要求高校发布真实信息,公布可靠数据。④加强对民间机构发布数据的校正、厘清。一些民间机构出于趋利考虑和自身责任的异化,可能会出现所发布的数据不真实、不系统、不严谨的数据现象。这样就需要有关部门及时予以干预,帮助这些民间机构端正态度,指导他们校正数据,厘清真伪。⑤充分利用国外、国内知名数据库,为高校科研质量评价提供参考依据。第一,在高校科研质量评价中,还要注重利用国外知名数据库。比如世界著名检索性数据库美国 SCI(《科学引文索引》)、美国 EI(《工程索引》)、美国 CA(《化学文摘》)、英国 SA;或 INSPEC(《科学文摘》)、俄罗斯

AJ(《文摘杂志》)、日本 CBST(《科学技术文献速报》)。第二,国内知名数据库。主要指国内知名中文检索:《中国期刊网全文数据库》(CNKI)、《维普中文科技期刊数据库》和《万方数据库资源系统数字化期刊》。另外,还有超星图书馆、书生之家图书馆等。

(四)立法规范,加强评价监督

通过立法规范高校科研质量评价,并加强其过程监督,是搞好高校科研质量评价工作的保证。

立足于高校科研质量评价标准,做好科研质量评价,从我国高校科研质量评价的现实状况看,当下可在两个方面推进:其一,在《中华人民共和国高等教育法》增加科研评价的条款,从宏观层面明确高校科研评价的意义、作用、地位和原则,指导高校科研质量评价。其二,在《关于改进科学技术评价工作的决定》的基础上,结合大数据时代科研评价面临的新情况和新问题,出台专门的科研评价法或高校科研质量评价法,调整政府与高校、高校科研评价评论主体与客体、队伍与科研成果等的各种关系,明确各方的权利和义务及监督机制、责任追究。规定科研评价的价值取向、评价的目标、评价机构、隐私保护、信息安全、信息泄露、伦理道德、安全保密等。把各类高校科研评价纳入法制化管理的轨道,从而保证科研评价的精确性、权威性和公正性,促进科研质量和水平的提高。特别是要从部门规章上升到法律层面需要诸多努力,是一个复杂的过程。为做好立法工作,我们必须把握如下几点:从评价目标上看,我们需要明确高校科研评价工作中的主要问题,倡导质量第一,克服功利主义和浮夸浮躁心理,营造科技创新的氛围,正确引导高校科研质量评价工作;从评价内容上看,提倡务实评价,建立立足国情并与国际接轨的评价内容,逐步完善各类评价指标体系;从评价方法上看,加强具体指导,明确职能定位,规范科研质量评价方法;从程序上看,杜绝任何形式的学术不端行为的出现,避免过繁过重和虚假的科研质量评价活动;从评价标准上看,以“科学、合理、可行”为原则,区别不同评价对象,区分各类评价标准,坚决反对浮夸作风和短视行为,客观评价非主流、非共识、非名人的科研结果,营造良好的创新文化。

(五)规避“数据万能”的陷阱,发挥同行评价的作用

大数据应用不当容易沦为“数据万能”的陷阱,在精准、科学运用大数据手段的同时,同行评价不能丢。

其一,我们面对的数据是海量的,类型也是多样的。从数据来源分类,可分为社会的数据、通过传感器收集的来自物理空间的数据和网络空间的数据,但这些数据并不一定是真实可信的^[10]。面对劣质的、信度不高的数据,大数据技术可能有其局限性,如果处理不好,大数据的优势就会变成劣势。充分发挥同行评价的作用,弥补大数据及大数据技术的缺陷,控制好信度和效用,是必不可少的。

其二,基于大数据及大数据技术的评价实质上是量化评价,无法消除定量评价本身带来的固有问题。如大数据具有强大的预测功能^{[11](P12)},不可避免地会存在从已知事实推理预测未知现象的可能,在进行相关性分析和数据挖掘过程中也容易受主观主义的影响^[12]。在人文社会科学研究水平和质量评价方面,其信度与效度难测量、理论方法不可逆,其评价更不能简单地用量化来评价其成果的真正价值。

其三,由于科研成果涉及很强的专业性、学术性,仅通过量化评价会遇到很多盲点和障碍,这些专业性、学术性很强的问题通过同行专家来评价可能最合适。

参考文献:

- [1]刘在洲.大数据应用于高校科研评价的价值意蕴与适用构想[J].科技管理研究,2021(4).
- [2]李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域:大数据的研究现状与科学思考[J].中国科学院院刊,2012(6).
- [3]田芸.稀疏交叉熵粗糙集 DDRBM-DNNS 高校科研能力评估[J].数学的实践与认识,2016(23).
- [4]李振,周东岱,刘娜,等.教育大数据的平台构建与关键实现技术[J].现代教育技术,2018(1).
- [5]刘三女牙,周东波,李浩,等.基于地图的教育大数据可视分析方法探讨[J].电化教育研究,2018(7).
- [6]习近平.瞄准世界科技前沿引领科技发展方向 抢占先机迎难而上建设世界科技强国——习近平在两院院士大会上的讲话[N].人民日报,2018-05-29.
- [7]刘在洲,张云婷.高校科研质量评价价值取向的反思与重构[J].科技进步与对策,2015(4).
- [8]左清.高校科研质量管理中的价值诉求及实现[J].现代大学教育,2010(1).
- [9]张云婷,刘在洲.高校科研质量评价的价值取向[J].长江大学学报(社会科学版),2014(9).
- [10]胡弼成,王祖霖.“数据”对教育的作用、挑战及教育变革趋势——大数据时代教育变革的最新研究进展综述[J].现代大学教育,2015(4).
- [11](英)舍恩伯格,库克耶.大数据时代[M].周涛,译.杭州:浙江人民出版社,2012.
- [12]张安淇,李元旭.大数据时代科学评价面临的变革与坚守——以人文社会科学为例[J].情报杂志,2018(9).

责任编辑 刘玉成 E-mail:770533213@qq.com