# 农业数据质量及评估方法探讨

李斐斐,张建华,朱孟帅,韩书庆,吴建寨*

(中国农业科学院农业信息研究所/农业部农业信息服务技术重点实验室,北京 100081)

**摘要** 在阐述农业数据质量内涵的基础上,对我国农业数据应用现状进行了分析,从数据流程角度选取了数据收集、数据存储、数据处理和数据管理4个影响因素,总结了单一准确性和多维综合性质量评估方法,并对数据质量检验的发展方向进行了展望,以期为提高我国农业数据质量提供参考。

**关键词** 农业现代化;农业数据;农业数据质量;质量评估方法

**中图分类号** S-058 **文献标识码** A **文章编号** 0517-6611(2017)36-0221-03

**Discussion on Agricultural Data Quality and Evaluation Methods**

**LI Fei-fei, ZHANG Jian-hua, ZHU Meng-shuai, WU Jian-zhai* et al** (Agricultural Information Institute of CAAS/Key Laboratory of Agri-information Service Technology, Ministry of Agriculture, Beijing 100081)

**Abstract** Based on expounding the connotation of agricultural data quality, we analyzed the application status of agricultural data in China. Four influencing factors of data collection, data storage, data processing and data management were selected from the point of view of data flow, we summarized the single accuracy and multidimensional comprehensive quality evaluation methods. And we forecast the development directions of data quality inspection methods, in order to provide references for improving the quality of agricultural data in China.

**Key words** Agricultural modernization; Agricultural data; Agricultural data quality; Quality evaluation methods

近年来,信息技术与经济社会、科学研究的交汇融合激发了数据量爆炸性增长,数据因其发现新知识、创造新价值、提升新能力的特点而成为国家基础性战略资源。我国是农业大国,传统农业农村数据积累量较大,而将传感器、移动通讯、数据库等现代信息技术引入农业生产、流通、消费等环节加速了数据量的跨级数增长。农业大数据是大数据理念、技术和方法在农业领域的实践[1],利用大数据驱动农业发展,有助于加快我国农业转型升级,提升国际竞争力和影响力。

数据包含数量和质量,二者相辅相成,没有质量的数量毫无意义,而数量又是质量的基础,质量的提升源于数量的积累。但是,随着农业数据资源开放性和多源性的增加,数据质量更加参差不齐,垃圾数据、错误数据、虚假数据等问题层出不穷,是我国农业面临"数据丰富、信息贫乏"困境的重要原因,不利于制定科学的决策。此外,与工商业不同,农业具有与时空密切相关、生产周期长、生产灾害难以弥补等特点,对数据质量具有更高的要求。

目前,数据用户着重于通过数据预处理等方式来提高数据质量,缺少对数据整体质量的评估,事实上,评估数据质量便于发现问题,合理控制影响因素,缩小误差。这不仅对于提高农业数据质量具有重要意义,而且有助于加强信息处理和知识发现,指导农业生产,满足农业农村发展的需要。因此,从数据科学的思想出发,明晰数据质量含义,把握我国农业数据应用现状,多角度揭示影响数据质量的因素,总结适用于农业数据的质量评估方法具有重要意义。

## 1 数据质量内涵

质量是一个多义词,它在质量管理学的定义为"一组固有特性满足要求的程度",质量的概念不是固定不变的,而是处于动态发展变化之中,与使用对象息息相关。在数据成为产品、可以提供服务功能后,数据领域引入了"质量"概念。数据质量最初是指数据的准确性,以误差大小作为衡量标准[2],这是一个相对狭义的定义。随着社会的发展和信息技术的进步,数据质量内涵不断扩展,准确性不再是评估其优劣的唯一指标,特别是从20世纪90年代起数据研究领域广泛采用多种质量维度,综合衡量数据情况。常用的质量维度有20多种[3],数据用户可根据需求灵活选取,在实际应用中,各质量维度重要性不一,始终没有统一的认识和标准[4-5],但它们并非完全孤立,而是相互关联的,其中数据准确性、一致性、完整性、可解释性等是基础性维度,其他维度可由这些维度推导得出,所以这些维度的高水平是其他可选维度质量的保障,也是数据质量的重要研究对象。

## 2 我国农业数据应用现状

数据是驱动农业现代化发展的重要力量,是连接农业生产、经营、消费、市场、贸易等环节的关键。数据的有效应用,一方面可以全息立体反映农业全过程,促进相关要素之间的联系,另一方面还可以通过数据间关联特征,预测未来,提前做好准备,应对行业变化。然而,现阶段我国农业数据发展水平并不均衡,主要侧重于农业生产、安全监管、市场调配等方面的应用。

在农业生产方面,农业数据的应用主要体现在精准生产、作物育种、灾害防御方面。①在精准生产中,利用农业物联网、通讯技术,实时获取环境中的温湿度、风速、二氧化碳以及土壤水分、电导率、矿物质含量等指标,并与农作物各阶段生长规律相结合,完成精准施肥、浇水等农耕活动,以实现资源最节约、效益最大化。②在作物育种中,通过大数据技术和生物技术获取更完整、准确的生物基因组数据,挑选出具有特定形状的基因组(如高钙、抗氧化、抗敏等),提高育种效率,弥补传统杂交育种工作中偶然性大、成功率低的缺

**作者简介** 李斐斐(1992—),女,河南安阳人,硕士研究生,研究方向:农业数据质量。*通讯作者,副研究员,博士,从事农业信息分析。

**收稿日期** 2017-09-18

点[6]。③在灾害预防中,利用历史气象数据建立相关自然灾害、病虫灾害模型,预测未来某时间点可能出现的意外。一方面,指导农户合理避开减产作物种植,或做好预防措施;另一方面,辅助制定农业保险政策,降低农户损失[7-8]。刘祖建等[9]对 1991—2010 年的 2 代稻飞虱发生情况和气象资料进行相关分析,已建立了成虫始盛期、若虫高峰期、发生程度及发生面积的预测模型,效果良好。

在安全监管方面,农业数据能有效促进农产品安全监管。传统农产品生产、流通、消费、存储过程中存在渠道复杂、信息紊乱、监管不透明等问题,安全控制难度极大。基于 RFID 射频、二维码等技术的农产品溯源体系,能将农产品生长、流通过程中的环境指标、地理信息、仓储信息等其他数据实时收集、存储、处理并用可视化方式展示,方便终端消费者全面获取“从田间到餐桌”过程中有关的产地、种植人、施肥量、农药用量、病害虫灾、采摘时间等,提高食品安全监管效率。

在市场调配方面,农业数据能驱动商业模式创新,完善市场调配。传统农产品市场信息不对称现象明显,供需不平衡情况也十分广泛,“田头贱、摊头不贱”“蒜你狠、姜你军”等问题层出不穷。在大数据技术支持下,农产品电商平台可以将生产者与消费者快速、精准地衔接、匹配。一方面,通过连续分析消费者在不同节气和温度下的购买习惯,实现精准订货、存储和配货,统筹不同区域农产品生产;另一方面,利用农业监测预警技术,分析各种农产品的交易情况、价格波动,提前发布市场信号,有效通过信息引导市场,应对市场变化。

## 3 农业数据质量的影响因素

农业数据在为农业发展创造重大机遇的同时,也带来了巨大的挑战,主要体现在对数据质量有更高要求。从数据流程,即数据生命周期角度来探讨影响数据质量的因素,大致可以分为数据收集、数据存储、数据处理、数据管理 4 个阶段。

**3.1 数据收集** 数据收集是数据生命周期的开始,对数据质量起决定性作用,若收集到的数据错误、不一致、滞后甚至无效,数据质量就无从谈起。农业数据来源广、种类多[10],选择合适的收集方式至关重要,传统农业统计以普查、抽样调查、重点调查或行政记录获取数据,易出现数据模糊、精度损失、记录不完备等问题,而现代农业已经将物联网、互联网、遥感技术[11-12]引入,极大改变了传统数据的采集模式,在系统交互过程中能获取更加具体细化的数据,但成本相对较高,目前主要应用在规模化、标准化的科研基地以及农业企业等单位中,普通用户短时间内难以普及。

**3.2 数据存储** 数据存储是保障数据质量水平的重要环节,数据存储紊乱会影响数据的使用效率,从而降低数据质量。目前,农业各业务数据以结构化为主,存储在传统的关系型数据库中,而半结构化数据和非结构化数据则需先转化成结构化数据才能得到有效存储。在异质数据转化的过程中,若转化不当对各质量维度的影响很大,特别是农业数据数量更庞大、结构更复杂、变化更快,出错率更高,所以突破异质数据转换、集成与调度技术[13],尽可能消除数据整合过程中出现的不兼容、精度损失等问题,完善大数据环境下的数据库建设十分必要。

**3.3 数据处理** 数据处理是提高数据质量的有效手段,包括数据更新、预处理、提取、分析等。①要满足数据质量维度自身要求,就数据时效性和价值性而言,温室控制中对温湿度、二氧化碳含量数据若更新不及时将导致环境调节滞后,影响农作物产量,数据价值性骤降;②加强数据预处理、提取、分析,尽管在数据采集、存储中都规范了流程,但仍会存在不准确、不一致、不完整的数据,降低数据挖掘效率和精确率,所以对数据进行分类或分组前的优化、排序是十分必要的。

**3.4 数据管理** 数据管理是干扰数据质量的外界因素,这里特指各种人为操作。数据收集、存储和处理侧重于从技术上规避问题,而数据管理旨在从人为角度分析影响准确性、一致性、完整性等质量维度的因素。一方面,数据收集时基层统计人员统计过于随意,上级领导为追求政绩会伪造数据,数据汇总时横向或纵向沟通不畅更会造成数据重复统计,增加冗余;另一方面,数据基本存储在数据库中,数据生命周期中数据库管理员都担负着重要职责,在设计存储架构时要充分考虑数据不兼容、不一致等问题。

## 4 农业数据质量评估方法

数据质量评估能够预先发现数据问题,为改善数据质量提供指导,是数据质量研究过程中的重要环节。现有研究多为框架理论,评估方法相对统一,主要围绕每个质量维度下数据指标的结构或内容展开。笔者总结了农业领域易出现的生产数据紊乱、价格数据不平衡等问题,结合国内外提出的模型方法,大致归纳为定性分析、定量分析 2 种。

**4.1 定性评估** 定性分析是以用户需求为中心的主观评价法,基于一定的评价准则,综合评判农业数据集的“好”与“坏”,评价结果可用等级制、百分制或其他方法表示,应用范围较广。传统的定性分析方法包括用户反馈法、专家评议法、第三方评测法,分别以数据用户需求、专家经验知识、特定信息需求为核心进行评估,这些方法适用于小样本数据,难以满足大数据在评估效率和准确率等方面的要求。当数据样本较大时,可以将目标质量维度简单归纳,根据需求进一步分解为更小的单位,直接或间接地评估其内在质量,如分析数据现实世界、信息世界的对应关系[14],分析数据更新频度等来判断数据的准确性、完整性、一致性、及时性等[15],还可以将研究视角拓宽至相关环境数据。此外,也可利用主观数据质量参数和客观数据质量指示器等其他合理的方法[16]进行研究。定性分析的方法简单易用,但评价结果比较模糊,缺乏客观、量化的分析。

**4.2 定量分析** 定量分析是以数据为中心的客观评价法,根据需求制定合理规则集[17],将目标质量维度进行量化和重现,评价结果直接用数字表示。根据评估的维度数量,可分为单维度准确性评估、多维度综合性评估。

**4.2.1** 单一准确性评估方法。早期有关数据质量的研究主要针对数据准确性,一般采用统计学模型分析,比较经典的方法包括逻辑关系检验法、核算数据重估法、计量模型分析法、统计分布检验法、调查误差评估法等。

逻辑关系检验法分为比较逻辑检验法和相关逻辑检查法,主要从横向或纵向角度粗略地检查统计指标之间存在的恒等、包含和相关关系,如各省农业产值之和与全国农业产值之和不一致。核算数据重估法是对逻辑关系检验法的拓展,主要从统计核算的角度出发评估农业生产数据、农产品价格数据或者行业增加值。计量模型分析法能通过建立计量经济模型,量化更复杂的相关关系,对相关指标的数据质量进行评估,但它一般是基于数据完全真实的假设上。统计分布检验法是根据统计总体的个体都服从特定的函数分布的性质,如正太分布等,若待评估数据集符合特定分布,则初步认为数据准确率高。调查误差评估法主要分析数据中所包含的误差进行评估,包括抽样误差和非抽样误差,对于非抽样误差可以用其他指标间接分析,也能用对统计数据执行二次抽样调查,并与前者进行对比核查。

**4.2.2** 多维综合性评估方法。多维综合性评估是对单一准确性评估的进一步拓展,评估对象包括数据基础维度和其他可选维度,评估方法是建立合理的评价模型,而模型的核心是如何有效度量数据的不精确、不完整、不一致等程度。

目前,农业数据大多以结构化方式存储在关系数据库中,数据各质量维度的度量大多采用数据库技术或数据挖掘技术。在基于数据库技术的方法中,学者广泛应用函数依赖关系分析数据集,函数依赖是指在关系数据库 R 中 2 个属性集合 X、Y 属性值之间的约束关系,如实体完整性、参照完整性、用户定义完整性等,用户依据既定的函数依赖,利用 SQL 命令批量筛选目标数据,如根据语法上相同或相似的不同记录可能代表现实世界同一实体的原理,用排序—合并、建立索引的方法检测违反完整性的重复记录,还可以统计属性字段缺失的记录,得到数据集的完整率、一致率等,有效量化数据集各维度质量;在基于数据挖掘技术的方法中,各质量维度的量化方法不同,用户可根据数据特点,采用聚类、分类、关联规则或自定义算法进行有限次迭代循环,筛选并统计符合用户需求的记录数,如基于距离的相似度计算、基于信息内容的语义相似度测度等。与数据库分析数据相比,它能动态计算属性相应的权重,客观性更强、处理效率快、精度更高。此外,还可以用信息熵、逆文献频率加权法等进行计算。

根据各质量维度的度量结果对数据集进行评估时,大致可分为以下 3 个层次。①根据度量结果直接对数据集进行评价,如农作物基因组数据的准确率、一致率、完整率分别为 78%、90% 和 95%,数据完整率较高,但若准确率更重要时,就难以突出重要质量维度,有时无法满足用户需求。②将目标质量维度进行分类,如分为核心维度和一般维度,黄莺等[18]在研究元数据质量时建立了一个四维核心模型,该模型由 2 层组成,一层是与数据内在质量密切相关的固定维度,另一层与数据外部环境联系较强的可选维度,其中第一层重要性更高。这种方法使研究对象主次有别,客观性更高。③构建综合数据评估模型,模型可以是简单的线性关系,也可以是复杂的多项式等关系[19],主要采用加权法(约束加权法、属性加权法、维度加权法等)给不同的质量维度赋予相应权重。针对农业数据非平衡问题,王晓华等[20]提出一个数据质量评估体系,用基于属性加权的缺失评估算法、非平衡离群评估算法进行数据缺失、离群评估,缺失评估算法的权重由基于类分布的属性加权求得,可靠性更高,该评估体系已经在马铃薯销售量和销售额中表现出良好的适用性。

实际应用中,为了使评估更加合理,充分发挥二者优势,可以将定性和定量分析结合使用,常用的方法包括层次分析法、模糊综合评价、灰色聚类法等。

## 5　数据质量评估方法发展

农业现代化进程中,农业也进入了大数据时代,各种监测网点及网络信息平台相继建立,数据环境愈加复杂,数据多源异构特点明显,同时错误、无效及过时数据也更多。为了提高大数据的应用价值,质量评估方法需要具有更高的效率和精确率,今后主要从适应分布式数据环境、加强知识发现、降低响应时间度等方面进行发展。

大数据质量评估方法要适应分布式数据存储环境。多源异构的农业数据主要存储在分布式数据库中,但分布式数据库的不同节点间多通过 Web 等方式连接,每个节点仅包含部分数据,数据类型、结构往往存在差异,传统函数依赖通用性差,为提高数据可迁移性,应明晰数据本质,挖掘数据间存在的异同,重新建立约束机制,以数据不一致性为例,京东和淘宝平台上相同的农产品在数据库存储中可能存在栏目、主题、约束、类型、结构、指代不一致等问题,可以建立基于层次概率判定的 Web 不一致数据自动发现算法。

大数据质量评估方法要加强知识发现能力。由于农业行业的特殊性,将传统数据库和基于专家知识的知识规则库融合使用,能深入洞悉数据特征,描述更加复杂和多样化的约束算法规则,全面判断数据质量。施建平等[21]据此建立了农田土壤自动识别和动态勘察的规则库,完成数据质量相关的背景和方法信息检验(检验样地代码一致性、长期采样地管理数据、标准物质测定准确度等检验)和数据检验(土壤微量元素等指标的阈值检验、统计检验、关联检验等)。

大数据质量评估方法要减少系统响应时间。数据规模的增大降低数据处理效率、增加系统响应时间是现阶段存在的重要问题,为减少数据处理过程中的时间消耗,一方面可以选择 MapReduce 分布式计算框架、分布式内存计算系统、分布式流计算系统等性能较好的模型或系统;另一方面,要化繁为简,降低算法复杂度,如在满足复杂多样的约束规则的同时,利用并行函数依赖和剪枝等方式。

## 6　结论

农业大数据时代已经来临,农业数据能全面揭示我国农业现状、突出问题和主要矛盾,是反映我国农业基本状况、生产方式、动力源泉的重要依据。对数据质量进行评估能宏观

(下转第 258 页)

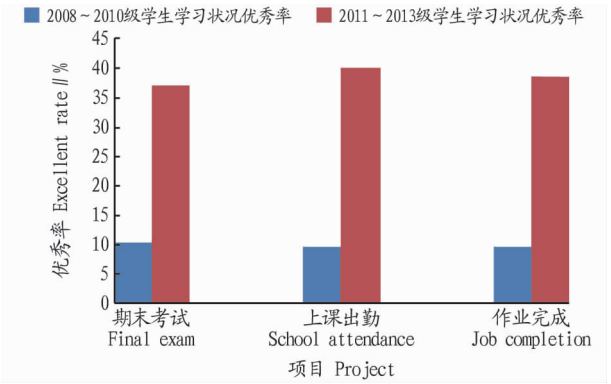


**图 1 2008～2010 级与 2011～2013 级学生学习优秀率对比**

**Fig. 1 Comparison of excellent rate of 2008 – 2010 grade students and 2011 – 2013 grade students**

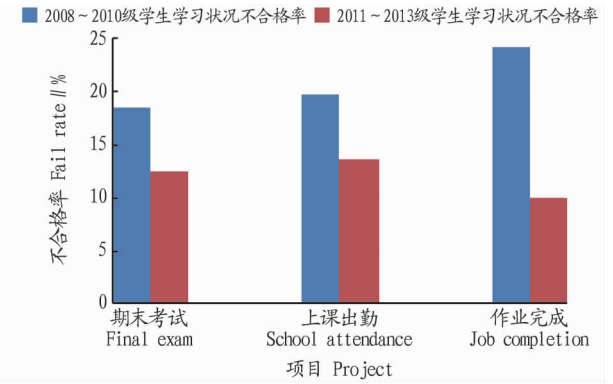


**图 2 2008～2010 级与 2011～2013 级学生学习不合格率对比**

**Fig. 2 Comparison of fail rate of 2008 – 2010 grade students and 2011 – 2013 grade students**

## 参考文献

[1] 于晓明.《结构力学》教学实践的体会和建议[J]. 科技信息,2012(35):200.
[2] 蔡东升,刘荣桂. 土木工程大类专业结构力学教学探讨[J]. 高等建筑教育,2012,21(4):62 – 65.
[3] 夏江涛. 基于应用型本科的结构力学教学体会与思考[J]. 攀枝花学院学报,2009,26(6):106 – 108.
[4] 李娟娟. 论我校教学竞赛体系的构建[J]. 怀化学院学报,2013,32(2):120 – 122.
[5] 杨柳春,汝宇林,徐俐. 基于工作过程教学模式的教师实践教学竞赛方案设计与实践[J]. 兰州石化职业技术学院学报,2009,9(1):61 – 64.
[6] 冯露,亢一澜,王志勇,等. 基于问题学习的探究式教学改革实践[J]. 高等工程教育研究,2013(4):180 – 184.
[7] 王有鹏. 精心组织课堂竞赛活动:让教学生机勃勃的课堂活动组织艺术之四[J]. 中学政治教学参考,2008(9):23 – 24.

(上接第 223 页)

把握数据可用性,制定科学决策,更好地服务政府部门,帮助农民合理规避农业风险,平衡市场和生产者的供应关系等。传统的数据质量评估方法相对成熟,在处理小样本数据集时表现出较高的处理效率和精准率,但是,在面对海量多源多模态农业数据时,现有评估方法还是难以满足对计算速率、数据种类等方面的要求,针对未来更加复杂,也更加开放的农业环境,今后应做好以下方面工作:加强算法在数据适用性、可扩展性,特别是共享性方面的研究,提高数据处理能力;建立农业大数据环境下更全面的评估标准、量化方式;完善在评估数据之后提高数据质量的策略。

## 参考文献

[1] 孙忠富,杜克明,郑飞翔,等. 大数据在智慧农业中研究与应用展望[J]. 中国农业科技导报,2013,16(6):63 – 71.
[2] WINKLER W E. Methods for evaluating and creating data quality[J]. Information system,2004,29(7):531 – 550.
[3] HUANG K T,LEE Y W,WANG R Y. Quality information and knowledge management[M]. New Jersey:Prentice Hall,1998.
[4] 黄刚,袁满,吴秀英,等. 元数据驱动的数据质量评估体系架构研究[J]. 计算机工程与应用,2013,49(8):114 – 119.
[5] BRUCE T R,HILLMAN D I. The Continuum of Metadata Quality:Defining,Expressing,Exploiting [C]//HILLMANN D I,WEATBROOKS E L. Metadata in Practice. Chicago:American Library Association,2004:238 – 256.
[6] RADAUER C,BREITENEDER H. Pollen allergens are restricted to few protein families and show distinct patterns of species distribution[J]. J Allergy Clin Immunol,2006,117(1):141 – 147.
[7] TAO F L,ZHANG S,ZHANG Z. Changes in rice disasters across China in recent decades and the meteorological and agronomic causes[J]. Regional Environ Change,2013,13(4):743 – 759.
[8] LIU X W,FEIKE T,SHAO L W,et al. Effects of different irrigation regimes on soil compaction in a winter wheat-summer maize cropping system in the North China Plain[J]. Catena,2016,137:70 – 76.
[9] 刘祖建,陈冰,陈蔚烨,等. 广东省西南部稻飞虱发生期和发生程度的气象预测模型[J]. 中国农业气象,2013,34(2):204 – 209.
[10] BROWN J C,KASTENS J H,COUTINHO A C,et al. Classifying multiyear agricultural land use data from Mato Grosso using time – series MODIS vegetation index data[J]. Remote sensing of environment,2013,130(4):39 – 50.
[11] 戈锦文,肖璐. 农业统计存在的问题及变革趋向[J]. 统计与决策,2016(18):188 – 189.
[12] JIAO L Z,DONG D M,ZHENG W G,et al. Research on fiber-optic etching method for evanescent wave sensors[J]. Optik-international journal for light and electron optics,2013,124(8):740 – 743.
[13] 马茜,谷峪,张天成,等. 一种基于数据质量的异构多源多模态感知数据获取方法[J]. 计算机学报,2013,36(10):2120 – 2131.
[14] WAND Y,WANG R Y. Anchoring data quality dimensions in ontological foundations[J]. Communication of the ACM,1996,39(11):86 – 95.
[15] WANG R Y,KON H B,MADNICK S E. Data quality requirements analysis and modeling[C]//Proc of Ninth ICDE. [s. l. ]:[s. n. ],1993.
[16] AEBI D,PERROCHON L. Towards improving data quality[C]//Proceedings of the International Conference on Information Systems and Management of Data. [s. l. ]:[s. n. ],1993:273 – 281.
[17] 杨青云,赵培英,杨冬青,等. 数据质量评估方法研究[J]. 计算机工程与应用,2004,40(9):3 – 4,15.
[18] 黄莺,李建阳. 元数据质量评估方法及模型研究[J]. 图书馆学研究,2013(12):52 – 56,51.
[19] 杨青云,赵培英,杨冬青,等. 数据质量评估方法研究[J]. 计算机工程与应用,2004,40(9):3 – 4,15.
[20] 王晓华,苏宏业,渠瑜,等. 面向电信欠费挖掘的数据质量评估策略研究[J]. 计算机工程与应用,2011,47(12):220 – 224,233.
[21] 施建平,沈志宏,苏贤明,等. 基于知识规则的数据质量检验方法在农田土壤监测中的应用[J]. 科研信息化技术与应用,2012,3(2):53 – 61.

**本刊提示** 文稿题名下写清作者及其工作单位名称、邮政编码;第一页地脚注明第一作者简介,格式如下:“作者简介:姓名(出生年—),性别,籍贯,学历,职称或职务,研究方向”。