发展农业大数据的主要问题及主要任务

郭承坤[1,2],刘延忠[3*],陈英义[1,2],孙敏[1,2],屠星月[1,2]　(1. 中国农业大学信息与电气工程学院,北京 100083;2. 农业部农业信息获取技术重点实验室,北京 100083; 3. 山东省农业科学院科技信息研究所,山东济南 250100)

摘要　大数据的应用已经成为各领域的研究热点。大数据理念和技术在我国农业领域应用方面具有一定的特殊性。重点研究分析大数据应用在我国农业领域时可能遇到的主要问题,从应用过程分析,包括数据获取过程中的数据量化和数据共享问题,数据处理过程中的预处理和元数据产生问题,数据分析解释过程中的客观性问题。针对上述问题,提出了发展农业大数据的3大任务,包括农业数据整合、农业大数据平台构建和多元研究团队培养。

关键词　农业大数据;多源异构数据整合;大数据平台
中图分类号　S126　　**文献标识码**　A　　**文章编号**　0517 - 6611(2014)27 - 09642 - 04

Major Issues and Missions in Agricultural Big Data

GUO Cheng-kun[1,2], LIU Yan-zhong[3*], CHEN Ying-yi[1,2] et al　(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083; 2. Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture, Beijing 100083; 3. Institute of Information Technology, Shandong Academy of Agricultural Sciences, Jinan, Shandong 250100)

Abstract　Application of big data has become a research hotspot in various fields. The application of big data concept and technology in agriculture has a certain particularity. The issues are concluded by analyzing the process of big data application. The first is quantification and sharing in data acquisition, the second is preprocessing and metadata generation in dispose of data and the third is objectivity in analyzing and interpretation of data. In order to resolve the issues, the three major missions are proposed including agricultural data integration, agricultural big data platform construction and multi-team cultivating.

Key words　Agricultural big data; Multi-sourced heterogeneous integration; Big data platform

早在1980年,著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中,将大数据热情地赞颂为"第三次浪潮的华彩乐章"。近年来,大数据技术已经在商业、金融、通信等领域得到了广泛应用。目前,关于大数据的概念还没有统一的定义,但其核心概念和思想是一致的。现在通常用"5V"理论解释大数据,即数据量大(volume)、处理速度快(velocity)、数据类型多(variety)、价值大(value)、精度高(veracity)[1-2]。另外,一些媒体、学者还指出大数据具有1C(complexity)的特性,即指数据结构复杂,需要新技术来满足异构数据统一接入和实时数据处理方面的需求。

随着精准农业、智慧农业、农业物联网的快速发展,传统农业向信息化、智能化农业转型,农业各环节、各领域中的信息也呈井喷式增长,从而为大数据技术在农业领域的应用提供了数据基础。农业大数据即运用大数据理念、技术和方法,解决农业或涉农领域数据的采集、存储、计算与应用等一系列问题[3],为粮食安全、农业生态环境监测、农业精细生产、农产品安全管理与溯源、生物品种感知以及农业科研等农业管理与研究提供科学支撑。

虽然大数据的理念和技术具有一定的普适性,但是运用到农业领域时,又有其特殊性。相比于商业、工业、公共卫生等其他行业,农业数据涵盖面广、数据源复杂,使得大数据思维和技术在农业中的推广面临很多挑战,也是目前为止,大数据在农业领域中还没有成熟运用的原因。笔者将农业大数据发展中的主要问题归结为以下3点:数据有效性问题、数据误解和数据等价性问题,并针对这些问题,结合我国农业现状,提出了发展农业大数据的3大任务,包括农业数据整合、多元研究团队培养、农业大数据平台构建。下面进行详述。

1　发展农业大数据面临的主要问题

1.1　数据获取问题　数据是应用大数据技术的根本基础,我国是世界人口大国,也是农业大国,理应拥有庞大的数据资源。但是实际存储下来的数据总量仅仅是北美的7%、日本的60%[4]。其中能被有效利用的数据则更少,通过研究分析发现,该问题主要是由于数据获取过程量化能力低与管理过程中数据共享少造成的。

1.1.1　数据量化能力低。农业普查目前还是我国获得农作物产量数据、农产品市场价格数据等重要数据的主要途径。而农业普查获得的数据极易受人为因素影响。例如,由于普查人员各人利益的原因或业务素质的原因,可能导致数据源头出现质量问题;普查机构审核控制不严密,决策部门制定的普查方案、普查体系瑕疵[5],也会影响农业普查所得的数据质量。

随着互联网、移动互联网和物联网技术在农业领域的普及,由机器提供的数据,将成为大数据的主要来源。其中,物联网技术是实现"一切皆可量化"的重要技术,农业物联网的核心是采集农业生产过程中影响动植物生长的温度、湿度、光照、土壤状况、水质状况、气象状况等信息进行加工、传输和利用,为农业生产在各个阶段的精准管理和预测预警提供信息支持。目前,我国农业物联网技术主要应用在蔬菜大棚种植、牲畜养殖、水产养殖等高端农产品领域,其应用范围有限。其他农业物联网应用方向,如农业环境监测、智能化节水灌溉、智能饲料投喂、动植物疾病远程诊断大都在试验阶

基金项目　山东省自主创新专项(2012CX90204)。
作者简介　郭承坤(1990 -),男,内蒙古呼和浩特人,硕士研究生。*通讯作者,副教授,硕士生导师,从事农业信息化研究。
收稿日期　2014-08-07

段,还没有大规模的商业应用,获得的数据量小,通常掌握在少数研究机构和农业相关企业中。

1.1.2 数据共享量少。在市场经济条件下,农业的分散经营和生产模式,使得农业生产很难在全国范围内形成统一规划,农业信息也分散在各类不同的涉农网站及研究管理机构数据库中。但是由于体质和利益等原因,这些数据相互之间缺乏统一标准和规范,在功能上不能关联互补、信息不能共享互换、信息与业务流程和应用相互脱节,形成了所谓的"信息孤岛"。

例如,在各地的农业信息网站可以查到地方农产品批发市场的粮食、油料、糖烟茶、蔬菜、果品、药材、植物油、畜禽产品、水产品当天的价格信息,但是无法查阅或下载农产品价格历史数据。其他农业数据,如农作物长势数据、病虫害数据、农产品供应、需求数据等,目前主要是以半结构化或非结构化的形式分散存储在一些农业信息平台上,或者在一些研究机构的数据库中,难以形成规模,以大数据的方法进行利用。

另一方面,目前农产品质量安全追溯系统的应用范围较大,是农业数据的重要来源。国内较有影响力的农产品溯源系统主要有上海食用农副产品质量安全信息查询系统、北京市农业局食用食品(蔬菜)质量安全追溯、世纪三农"食品安全追溯管理系统"、中国牛肉全程质量安全追溯管理系统、国家蔬菜质量安全追溯体系[6]。然而,它们从识别码、存储信息到网络查询系统等各方面都不完全统一,所针对的食品对象也不尽相同。由于开发商不同,其溯源信息的存储未能贯通也不能达到共享,无法进行跨系统查询。

1.2 数据处理与管理问题 在小数据时代的背景下研究农业,要求数据精确可靠,所使用的数学模型也比较复杂。许多学者希望将这些模型、方法直接用在大数据上。在一些情况下是可行的,但是通常会遇到下述两个问题,即数据预处理和元数据产生的问题。

1.2.1 数据预处理。在数据量不大的情况下,容易使要求数据尽量满足规范和需求。而农业大数据包含大量多源异构数据,且数据质量参差不齐。因此在分析运用数据之前,有必要对数据进行预处理。针对大量结构化、半结构化、非结构化数据需要进行数据整合,使其满足使用需求,能够完成数据之间的交互和协同。同时,还需对价值不大,或不感兴趣的数据,以及故障数据、异常数据进行剔除和清理。由于农业分散经营,个体差异大,因此如何整合清理来自不同数据源的数据,并使其有效地应用大数据分析,是发展农业大数据面临的巨大挑战。

1.2.2 元数据。数据清理完成后,就需要建立"元数据",即用来描述数据的数据。元数据的主要内容是数据的来源、采集方式、采集时间、采集人等。不同的行业或项目的研究目的不同,需要不同的元数据格式,如何设计适用于农业的元数据格式,将成为一个研究难点。农业大数据的多样性、复杂性、多源异构性,要求多模态的数据管理处理方式,而元数据的建立是数据处理与管理的重要依据。例如,管理与处理农业教学视频及农产品品种图片、专家建议语音的过程中,需要对各类数据的不同特征进行描述,满足使用者分析查询的要求。

1.3 数据分析解释客观性问题 面对农业领域的大数据,任何有效的大数据工具和方法都可以对其进行分析。但是不同的机构有各自的标准和规范,其解释数据的结果也必然带有主观性,造成"数据偏见"。

因此需要考虑用以分析的数据能否代表"客观事实",分析人员在清洗数据时,是否会将"不利"数据忽略,在得出统计结论时,是否被人为地忽略掉一些重要结论。从技术层面看,分析人员可以使用多种免费的工具如 R、Hadoop、Pig 等,结合具体的统计分析方法,将数据"塑造"成其预期的模式[7]。其次,在数值结果或者图表产生后,对数据的解释过程中,也不免受到分析人员的主观意见影响,从而使数据分析结果偏离客观现实,难以真实科学地反映科学事实。

2 发展农业大数据的主要任务

2.1 农业数据整合 进入大数据时代,面对更多、更杂的数据,研究者在处理数据时,思维需要首先发生重大的变化。在研究农业科学时,不应再追求小范围内的精确数据,而应接受数据的多样性、混杂性。曾经认为是废弃的数据,也有其价值。如在农业物联网应用场景中,由于传感器异常导致的错误数据,会影响农业物联网的商业应用和传统农业科学的研究,但是将此类异常数据作为传感器故障诊断研究的样本数据,当其数据量足以作为研究故障诊断的"全体数据"时,将发挥重要作用。

在小数据时代,统计学家收集样本的时候,会制定一整套的策略来减少错误发生的概率。在收集好样本后,还需检查是否有系统性偏差的发生。在统计结果发布前,检验结果是否在误差范围内。这些策略的实施,需要制定各种数据格式、协议,需要经过专门训练的专家来采集样本。这些工作即使在少量数据的时候,也耗费巨大。由于农业环境的复杂性、研究对象的多样性,在大规模数据的基础上保持数据的质量和一致性在目前是不现实的。随着农业物联网技术和数据处理技术的进步,这种不准确性会逐渐减弱,但是一定会长期存在。所以,接受数据的不准确性是发展农业大数据时需要坚持的思想。

数据融合技术在整合农业数据时将发挥重大作用,数据融合的目的简而言之就是将来自多个传感器或多源信息进行综合处理,从而得到更为准确、可靠的结论[8]。数据融合技术可以融合来自同一平台的或者不同平台的多传感器数据。按照数据抽象层次分类的话,数据融合技术可以分为像素级融合、特征级融合和决策级融合 3 类。目前,对于数据融合的研究,多是根据实际应用问题,使用"定制"的融合方案,还缺乏统一的理论框架和融合模型。建立适用于农业领域的数据融合模型也是发展农业大数据的一个重要任务。

数据融合系统主要包括 3 个组件:输入数据预处理模块、输入数据集融合和过滤模块、数据集后处理模块。系统运行的主要流程如图 1 所示。

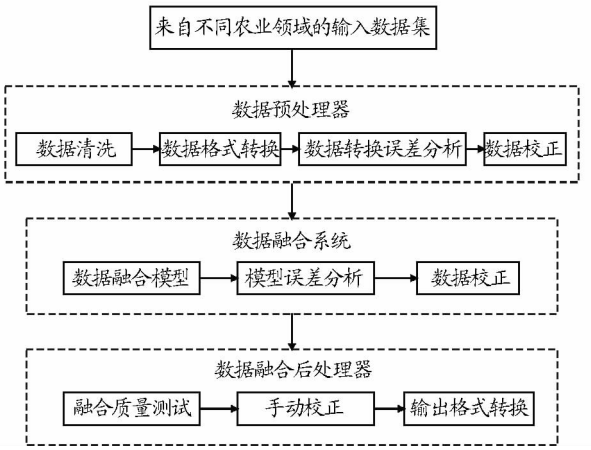

图 1 农业大数据融合流程

预处理组件首先需要对数据冗余进行清理,然后将异构数据格式转换为统一的数据格式标准,在数据格式转换后,需要对可能产生的误差进行分析和校正。预处理器的输出将作为融合系统的输入,首先输入到数据融合模型中。融合模型的研究是整个数据融合系统的重点,目前主要使用的方法有产生式规则、模糊逻辑、神经网络等。模型处理后的结果可能存在系统误差,所以需要对模型结果进行误差分析和校正。在数据融合后处理器中,对融合结果进行实际验证和手动校正,然后将输出结果转换为统一的数据交换格式。

2.2 农业大数据平台搭建 为全面、规范、及时地采集农业数据,并在第一时间对大量的、异构的农业数据进行分析处理,需要建立一个农业大数据平台。首先,需要统一数据采集、上报接口。在技术上,可以基于 SOAP(Simple Object Access Protocol)协议,建立一套适合于农业的数据交换协议,通过该数据交换协议,可以将原来已经广泛存在于 Internet 上的农业信息通过主动或者被动方式收集起来,还可以为以后的数据采集、上报提供一套标准接口。在大数据平台中,还应该使用 Web Service 等技术,向外提供一套标准数据访问接口,其他农业类网站、政府、研究机构可以通过这一接口访问到平台中的数据。

在数据处理方面,为能及时快速地处理大量的、多源的农业数据,需要探索并应用目前业界广为流行的分布式计算以及分布式存储系统,如 Hadoop + Hbase 的分布式文件系统架构。在此之上,为使平台业务与数据操作相隔离,帮助农业研究者专心于业务领域,而不是复杂的大数据操作,平台应该基于 MVC(Model + View + Controller)的设计架构,将对文件系统的 MapReduce 操作封装到 Model 层里。在 View 层和 Controller 层提供更多的可扩展性和可配置性,以使从事不同农业领域的农业工作者、数据上报人员、数据分析人员可以根据自身需求,定制平台中的相关功能。

在业务方面,农业大数据平台应该尽可能覆盖我国已经发展较好且已有一定信息化基础的农业产业,如粮食作物、经济作物、果树种植、蔬菜种植、林木花卉、畜禽养殖、水产养殖、农产品物流等至少 8 项农业产业;提供包括农情(苗情、墒情、灾情、病虫害)、市场(价格、供求)、农业资讯(新闻、行业信息)等信息服务;基于大数据技术,研发智能化的决策支持系统,可提供大数据分析成果发布和决策管理信息发布,为科研机构、政府等农业管理者提供技术和决策支持,为农业从业者提供个性化的生产指导。

该研究设计了农业大数据平台(图 2),Web 网站接收到需要增删改查数据的请求后,将操作数据的请求发送给 HBase 的 HMaster,HMaster 负责管理所有的 HRegion Server(ZooKeeper 用于保存 root region 地址和跟踪 region 服务器),而 HRegion Server 又管理了多个 HRegion。在物理上,HRegion 被分为了 3 个部分:Hmemcache、Hlog、HStore,分别存储缓存、日志和持久层。在持久层中,每个 Store 实例包含了 1 个或多个 StoreFile 实例,它们是实际数据存储文件 HFile 的轻量级封装,而实际存储文件的功能是由 HFile 实现的。Hbase 的 HFile 基于 Hadoop 的 TFile 类,对于持久层的操作将被该类作为一个 MapReduce 请求通过 Client 提交到 Hadoop 的 JobTracker 端,最后到达数据的存储位置 DataNode。Hadoop 内部的数据处理过程已经超过该研究的讨论范围,详细内容可在相关书籍中查阅。

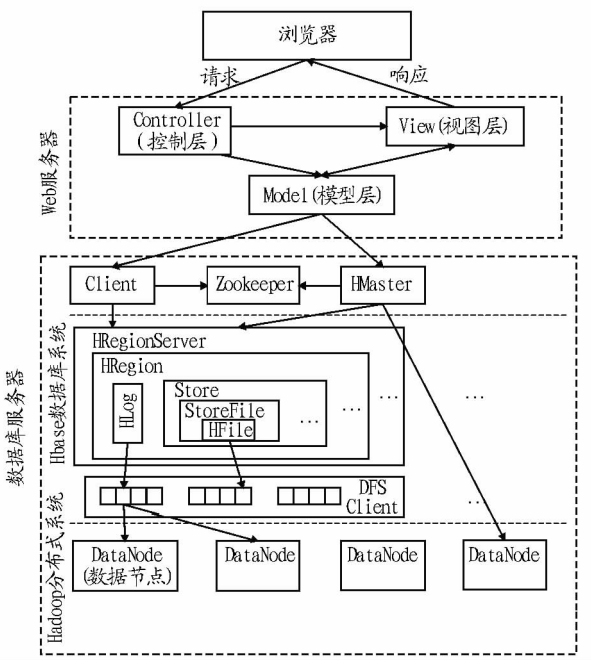

图 2 农业大数据平台系统结构

2.3 多元团队培养 建立在相关关系分析的基础上进行预测是大数据的核心。传统农业科学的研究方法基于小而精的数据样本,然后通过机理性的研究对某机理模型中的参数进行校正,或者采用启发式算法对某特定参数进行预测。两种方法各有优劣。机理模型的研究方法需要提前做出假设,然后通过数据进行验证,其解释性更好,但是容易受固有思维和偏见的影响,无法发现新的规律。启发式算法可以不受机理模型的局限,随着样本的增大,可以逐渐接近真实情况;但是限于数据量的限制和计算的复杂度,通常考虑的因素较少,如在预测养殖水体的溶氧时,输入模型的参数通常只有水溶氧、入水温度、池水温度、室温、水深、pH、盐度和气压等,而实际影响溶氧的因素还包括动植物、气象等各种其他因

素。所以传统启发式算法的结果在解释性上必然受限。

如上所述,农业科学中因果关系的探究较为复杂。但是,利用大数据技术进行相关关系的分析可以突破这些局限。大数据的相关关系分析法更准确、更快,而且不易受偏见的影响,可以使人们更好地认识农业领域,甚至发现新的知识。但是,大数据本身并不具有自解释性,只是为理论研究提供了"客观真理"。因此,可以在这个"客观真理"的前提下,提出更合理的假设,指导因果关系的研究。

对于农业中这些错综复杂的因果关系的研究,需要多学科配合的团队。农业专家、传感器及传感器网络工程师、气象学家、IT、统计分析人员[6]都是团队中不可缺少的成员。收集的数据越多、越全面,可以发掘到的相关关系就越多,而对其进行解释的难度就越大。所以,在将大数据技术应用于农业领域时,构建一个多元的学科团队是十分必要的。

3　结论和展望

大数据、物联网、云计算等信息技术已经在军事、商业中得到了较为广泛的应用,这些新一代的信息技术正在深刻地改变着人们的生产和生活方式。作为我国支柱产业的农业,也正在经历着向信息化和智能化方向的转变。该研究总结了在农业中使用大数据技术可能遇到的 3 大问题:数据、技术、思维。在大数据时代,数据、技术、思维是 3 大核心竞争力。三者必居其一,才有可能发挥大数据技术的优势[9]。要在大数据时代发展农业,可以将数据、技术、思维比作大数据时代的生产资料、生产工具与生产者。三者互为条件,协调发展,才能保证大数据在农业领域能得到充分的应用。今后的研究可以遵循该研究提出的整合农业数据,构建多元团队,建立农业大数据平台的 3 个农业大数据发展思路和方法,融合来自农业中不同领域的数据,结合各领域专家知识和大数据分析工具,提高农业信息化和智能化水平。

参考文献

[1] 孟小峰, 慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 2013(1):146 - 169.

[2] LUO S,WANG Z,WANG Z. Big-data analytics: Challenges, key technologies and prospects[J]. ZTE Communications, 2013(2):11 - 17.

[3] 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(6):63 - 71.

[4] 温孚江. 农业大数据研究的战略意义与协同机制[J]. 高等农业教育, 2013(11):3 - 6.

[5] 霍蓉. 浅谈影响农业普查数据质量控制的因素与对策[J]. 青海统计, 2008(3):31 - 33.

[6] 陈华. 食品质量溯源系统的现状及发展建议[J]. 湖南农业科学, 2010(21):87.

[7] LUDENA,DENNIS A,AHRARY,et al. Big data's risks and opportunities for ICT agriculture[C]//Advanced Applied Informatics (IIAIAAI), 2013 IIAI International Conference. Los Alamitos, CA,2013:116 - 120.

[8] 高翔, 王勇. 数据融合技术综述[J]. 计算机自动测量与控制, 2002(11):706 - 709.

[9] MAYER-SCHÖNBERGER V,CUKIER K. Big data: A revolution that will transform how we live, work, and think[M]. Houghton Mifflin Harcourt, USA, 2014.

(上接第 9641 页)

随着热线不断发展,热线本身积累大量信息资源,同时需要不断整合社会各界农业科技信息以满足用户需求,因此需要建设独立自有数据库,并通过建立标准的共享数据接口实现社会资源按需共享。

IVR 交互式语音应答系统是北京 12316 农业服务热线的一个重要组成部分。随着北京 12316 农业服务热线不断深入地开展,科技服务的 IVR 语音导航系统(图 7)也需要不断更新和升级。

随着技术的不断更新,新的 IVR 编辑器和执行器不断推出,新的 IVR 编辑器和执行器不仅在编辑 IVR 的效率上有了很大的提高,而且在 IVR 的执行速度和效果上也有很大的优化。北京 12316 农业服务热线二期将采用新的 IVR 编辑器和执行器[6]。

随着北京 12316 农业服务热线的科技服务的不断深入,科技服务的内容也在不断变化,相应的科技服务信息的数据库结构也在不断变化,为此 IVR 导航结构也在不断变化。

参考文献

[1] 岳进,尚明瑞. 甘肃省 12316"三农"服务平台建设的技术体系架构及其实现目标[J]. 农业科技与信息,2012(24):21 - 24.

[2] 许娓. 呼叫中心的核心技术及组成[J]. 电脑与信息技术,2001(6):15 - 18.

[3] 徐庆征. 呼叫中心技术及其发展浅述[J]. 江西蓝天学院学报,2006(1):49 - 51.

[4] 潘山,潘鲁萍,冯良清. Web Services 技术在呼叫中心系统中的应用研究[J]. 制造业自动化,2012,34(9):14 - 16,23.

[5] 张翠丽. 基于统一受理的农业呼叫中心解决方案[J]. 计算机应用与软件,2006(10):31 - 32.

[6] 李沐华,周志鹏. 农业信息服务呼叫中心的设计与实现[J]. 计算机与现代化,2011(2):70 - 72.