顾戈琦,李　瑾. 基于众包的农业大数据采集平台构建[J]. 江苏农业科学,2018,46(5):191 – 194.
doi:10.15889/j.issn.1002 – 1302.2018.05.051

基于众包的农业大数据采集平台构建

顾戈琦, 李　瑾
(北京农业信息技术研究中心/国家农业信息化工程技术研究中心/农业部农业信息技术重点实验室/
北京市农业物联网工程技术研究中心,北京 100097)

摘要:众包可以将非特定社会大众引入到农业大数据采集中,能有效扩充数据采集队伍,扩大数据采集范围。介绍了众包及农业大数据采集的内涵,针对参与众包大数据采集的农户、农场、农企的特性进行分析,并建立数据采集平台运行机制,设计平台多种数据接入、数据源质量评级、数据隐私分级等功能,分析众包农业大数据采集平台对现有农业数据采集工作的优势,为农业大数据的采集工作进一步提升助力。

关键词:众包;农业;大数据;数据采集;平台构建
中图分类号:S126　　**文献标志码**:A　　**文章编号**:1002 – 1302(2018)05 – 0191 – 03

众包即为打破原有体制限制,将原来须由系统内部工作人员将完整的任务置于开放平台上,使非特定的社会大众可以根据自己的能力选择适合自己的采集任务,而不须成为发布任务的单位中的一员[1-2]。现阶段,采集农业大数据多依托特定的政府部门、企事业单位建立有独立的数据采集团队进行特定农业数据的采集,采集到的数据部分进行公开分享,部分留于系统内部使用,这种采集方式具有采集成本高、采集队伍管理难度大等问题。众包农业大数据采集平台能打破不同单位间体系,将原来以特定体系为核心的任务完成方式转化成以特定任务为核心的网络化社会生产,只要具备数据采集能力的社会大众都可以参与农业大数据采集工作中,有效地扩充了数据采集队伍,扩大了数据采集覆盖范围[2-3],同时,应用先进的大数据技术,能有效减少在采集众包数据中产生的误差,在保证数据采集质量的前提下,降低采集成本、扩大采集范围。

1　众包及农业大数据采集

众包别称网络化社会生产,是指把过去由员工执行的工作任务,以自由、自愿的形式外包给非特定大众网络的做法,具有生产成本低、联动潜在生产资源、生产效率高以及满足用户个性化需求等优势[4-7]。众包具有组织开放性,众包发布者将公开发布需求,参与者不受组织边界的限制,无论是否属于发布者的组织,都可以参与解决众包问题,组织可以借助外部资源解决内部问题;众包具有地域分散性,众包发布者与参与者不受地理位置的限制,均可以通过信息技术手段沟通、讨论、解决问题,具有明显的个体分布特点[8];众包具有参与自主性,参与者根据自己的能力自主选择合适的众包需求,用“由下至上”的需求匹配模式代替“由上至下”的任务布置模式,大幅度提高了团队能力和任务需求的匹配程度。

最早于 1980 年由著名未来学家阿尔文·托夫勒提出大数据的概念[9-10],直到 2008 年以后,大数据的概念才逐步被认可,并被政府、企业以及学术界所广泛传播[11]。大数据有 5 个主要技术特点,可总结为 5V 特征:(1)大体量(volume),即可从数百太字节(terabyte,简称 TB)到数十数百拍字节(petabytes,简称 PB)、甚至艾字节(exabytes,简称 EB)的规模;(2)多样性(variety),即大数据包括各种格式和形态的数据;(3)时效性(velocity),即很多大数据需要在一定的时间限度下得到及时处理;(4)准确性(veracity),即处理结果要保证一定的准确性;(5)大价值(value),即大数据包含很多深度的价值,大数据分析挖掘和利用将带来巨大的商业价值[12-13]。农业大数据是指大数据技术、理念、思维在农业领域的应用,利用智慧化、智能化、网络化的现代信息技术,为农业生产、流通、消费过程服务[9,14]。农业大数据首先要解决的问题就是数据采集,只有采集到海量、多样、及时、准确的数据,农业大数据才能发掘出数据中的价值,更好地为农村农业发展、农业经济转型升级服务[15]。

2　农业大数据众包采集平台模式设计

2.1　平台众包对象

2.1.1　农户　农户受限于自身技术水平,应用数据指导生产的能力较弱,但因其具有人数众多、时间相对充裕、生产经验较为丰富、收入偏低等特点,在众包农业大数据采集中可以作为广泛的数据采集源。农户利用闲散时间上传相关数据信息,并结合其丰富的生产经验,对数据的准确定期进行人工审查,同时,由于其收入偏低,数据采集费用也相对较低。

2.1.2　合作社　农业合作社具有一定的规模及资金实力和技术能力,每天都会产生大量生产、销售数据,如对这些数据进行汇总分析可产生巨大的价值。同时,合作社具有初步应用数据能力但大多没有专业的数据分析人员,无法针对数据进行深入分析进而指导生产,但可以通过数据共享交换数据

收稿日期:2016 – 10 – 18
基金项目:北京农林科学院创新能力建设专项;工程院咨询课题(编号:2016 – ZD – 03 – 04);北京市自然科学基金面上项目(编号:9162006)。
作者简介:顾戈琦(1988—),男,江苏泰兴人,硕士,助理研究员,研究方向为农业农村信息化。E – mail:gugeqi@qq.com。
通信作者:李　瑾,博士,研究员,研究方向为农业农村信息化。E – mail:lij@nercita.org.cn。

服务的方式,使合作社参与到众包农业大数据采集中。

2.1.3 农业企业 农业企业是指围绕农业生产、流通、消费各环节提供增值服务的企业,其生产经营具有较强的专业性。企业内部大多建有信息管理系统,具有一定的数据意识和数据分析能力,数据对于企业生产经营效率提升较为显著,故其使用数据的意愿较为强烈。在众包农业大数据采集中,一方面可以将企业信息系统中的数据进行脱敏采集,交换对应的数据,另一方面可以让企业支付一定的费用,获取其需要的目标数据。

2.2 平台机制

2.2.1 多源采集机制 平台集合农业合作社、农业企业、个体农户等多种采集主体,通过传感器直采、信息系统接入、农户手机上报等多种采集方式,采集生产环境、生命信息、农田变量信息、农产品市场经济等多种类型的数据,广泛采集农业相关数据,实现多来源、多类型数据的全覆盖。

2.2.2 多重校验机制 平台采用多重校验机制,不同质量级别的数据源对应不同的数据检验方法,评级低的数据源须进行多次、多种校验。不同来源的数据通过智能算法进行交叉校验,对于部分质量不达标的数据会进行二次人工审核。在使用数据的过程中,用户也可以对数据进行审查,如有误,可提交纠错,实现多层次、多方法的数据校验。

2.2.3 用户激励机制 平台可根据用户采集数据的数量、质量、时效性等特征,将用户采集的数据统一转化成数据分,用户可以使用自己的数据分交换平台上的原始数据、数据分析报告等数据服务或者直接交换现金,使不同的数据采集用户都可以在平台上获取有效激励。

2.3 平台设计思路

众包农业大数据采集平台利用众包的思想,转变数据采集工作思路与采集人员队伍建设,将传统的独立成体系的数据采集队伍打散,将普通社会大众纳入到农业大数据采集队伍中,每一个普通社会大众利用闲散时间就可自主参与农业大数据采集工作,平台利用大数据技术进行交叉校验、结合人工数据检验,可以有效保证数据质量,数据需求方也可根据自身需求发布数据采集任务,减少自建数据采集队伍的成本。这样既可以扩大数据采集范围,又可以降低数据采集成本,能有效地提高农业大数据采集效率。

3 平台功能设计

3.1 系统接入功能

政府机关、科研单位、农业企业及部分农业合作社多已建有管理信息系统,这些信息系统覆盖气象、农产品市场价格、生产环境、土肥配方等领域,包含从政府宏观层面到企业微观层面的信息,但由于功能设计、应用技术、数据结构等原因,系统与系统之间的数据相互孤立,平台通过建立通用开放接口,连通多种类型信息系统接入数据,数据源可根据接入数据的质量与数量获取相应的金钱收入或交换对应的数据服务。

3.2 物联网设施数据直采功能

无线射频识别(radio frequency identification,简称 RFID)技术、空气温湿度传感器、土壤温湿度传感器等物联网设施在农业领域应用逐渐深入,采集到海量生产环境、物流、产品溯源等信息,平台建有物联网数据采集模块直接接入物联网信息采集硬件设备,直接读取硬件设备采集的多种信息,减少信息采集中间环节,减少物联网设施安装、软件系统构建成本。同时,数据源可根据接入数据的质量与数量获取相应的金钱收入或交换对应的数据服务。

3.3 数据人工直采功能

在农业生产各个环节中,很多数据的采集还须依赖人工进行,现阶段采集手段多为人工记录,然后统一上传到特定的信息系统中,部分地区还使用原始的人工纸笔记录,逐级上报的信息采集手段,平台建有移动信息采集端,可以安装到信息采集人员的手机上,也可以适配移动扫码枪、移动电子秤等移动信息采集端,及时、完整地将采集到的信息汇集到平台中,减少时间延误和上报过程中的误差。人工直采信息员可以是企业、政府等有组织的信息员,也可以是普通个人用户,可以在私人手机上安装信息采集端上传数据,根据接入数据的质量与数量获取相应的金钱收入或交换对应的数据服务。

3.4 网络数据抓取功能

互联网包含海量数据,很多与农业直接相关的数据,如农产品价格、农产品供需、气象、政策法规等数据,还有很多与农业间接相关数据,例如宏观经济、市民生活、交通物流等数据,在大数据技术支持下,间接数据可以作为直接数据应用的有力补充,提高数据应用效果。平台建有互联网数据爬虫,广泛采集互联网农业直接相关和间接相关的各类数据,构建农业综合数据库。

3.5 数据源质量评级功能

根据数据源的获取方式、接入渠道,对数据源进行分级,如是物联网设施直采数据,政府、科研单位、知名企业、大型农场信息系统接入数据以及有组织的人工直采数据,评级较高,进行简单清洗统一结构即可接入平台;网络抓取数据、零散的人工上报数据、小型信息化水平较低单位的信息系统接入数据,则评级较低,须进行数据清洗校验接入平台,同时保留原始数据供用户深入分析。高级别的数据可以减少数据清洗校验的环节,提高数据采集的时效性,同时,用户也可以参考评级分类,选择适合自己的数据。

3.6 数据隐私评级功能

数据具有隐私性,部分隐私程度高的数据只能供给特定用户使用,例如部分政府数据只能供给特定的研究机构使用,部分企业数据也无法做到完全公开。平台提供数据隐私评级功能,数据提供方可以在接入平台的时候,选择自己的数据隐私评级,保护自己的数据权益,这样才能让更多的数据源接入采集平台。

3.7 数据智能清洗汇总功能

将数据采集到大数据平台之后须进行简单的清洗,首先剔除格式错误、乱码数据等形式错误,然后针对异源同类数据进行校验,如来源不同的同类数据出现不同,则标注数据存入异常数据库中,再将异源同类数据进行合并汇总,减少数据重复。

3.8 数据人工纠错功能

受限于现阶段的数据清洗技术单纯的计算机无法高效准确地清洗所有数据,平台同时开放人工数据审核功能,用户可以根据自己的特点申请分级审核资格,在获取分级审核资格之后,针对目标数据进行人工审核,可根据审核工作量、审核挑出的错误数,获取相应收入。

3.9　数据订单悬赏功能

虽然大数据采集平台广泛采集各类农业数据,但部分数据无法满足需求,用户可以根据自己的数据需求进行订单化数据悬赏,鼓励其他数据源分享数据,鼓励个人用户积极参与数据采集工作,既可省去自建数据采集队伍的高昂成本,也可获取急需的重要数据。

3.10　数据交易功能

数据拥有方可以将自有数据放在大数据采集平台上进行交易。

4　平台优势

4.1　众包数据采集体系

现有的数据采集体系大多为政府、科研单位、企业等为自身目标建立的完整的数据采集系统,数据采集人员多为该单位雇佣人员,同时,由于体系限制,特定系统工作人员只能采集该系统所需数据,大多数数据采集人员的工作量远没有达到饱和状态,导致了数据采集队伍重叠,数据采集能力浪费等问题。基于众包的原理,众包农业大数据采集平台打破原有建立的完整数据采集队伍进行数据采集的模式,汇集社会各界力量,使每个具有数据采集能力的人都可以参与到数据采集工作中,以数据采集目标为核心进行数据采集工作。

4.2　多类型海量数据采集

现有数据多分散地存储于不同的信息系统、数据库中,由于部门限制、商业利益等原因不能完整有效的公开,在原始数据的基础上部分公开数据进行了数据整合,处理之后的数据,很多宝贵的细节信息会丢失,导致深入分析的价值大幅降低。通过开放信息系统接口的方式,众包农业大数据采集平台使现存于各个信息系统、数据库的数据能够便捷、广泛地汇集到平台中,通过物联网设备直采、人工采集数据直采功能,快速、高效地将原始数据采集到平台中,保留丰富的原始数据细节。

4.3　多层级分类保护数据隐私

现有数据采集平台无法根据数据的隐私程度进行数据隐私分类,但许多政府、企业单位的数据由于数据隐私性、数据敏感性等多种原因无法对全部使用者开放,由于无法控制数据传播和使用范围,这些单位选择了完全不开放数据。平台提供数据隐私评级功能,允许数据发布者选择数据分享隐私级别,使用户可以选择数据分享的受众范围,使部分具有机密性的数据只能被部分用户访问、使用,最大限度的保护数据源的隐私,使更多的政府、企业愿意将自己的数据在平台上分享。同时,通过数据源分级的机制,用户可以自行甄别数据源的质量,信息分析能力强的用户可以选择原始数据进行深度分析,使信息分析能力弱的用户可以选择经过初步处理的数据应用,以满足不同人群的需求。

4.4　多源数据交叉验证

平台在采集端进行广泛的数据接入,不仅可以接入现有数据库、信息系统中进行初步加工的数据,还可以直接接入物联网设施、人工直接采集的数据,这些数据不仅存在数据结构不同、采集误差、传输误差等系统问题,还由于众包数据采集队伍构成人员复杂、数据采集水平高低等导致的采集专业性、采集连续性等人员问题。平台通过数据挖掘、人工智能等技术进行数据交叉验证、补全,可以有效减少单一数据采集系统存在的系统性错误,剔除异源同类型数据中存在的错误,可以减少众包采集人员采集到的数据误差。

4.5　数据质量人工校验

平台不仅通过数据挖掘、人工智能等计算机技术进行自动化交叉验证,还开放了人工数据验证功能,具有一定数据识别能力的人可以在平台上申请人工数据校验资格,具有数据校验资格后,利用空闲时间进行数据人工查错,如果找到错误数据并进行有效更正,即可获得查错奖励,这样在数据校验层面上也利用众包的思想汇集社会各界力量,用人工的方式进行数据校验可以发现机器无法发现的更为细致的数据错误。

4.6　定制化数据采集

现阶段,数据使用方大多只能在现有的数据中选择自己需要的数据进行使用,对于没有现成数据的情况,如果实力雄厚可以自建数据采集队伍,定向采集目标数据,但对大多数用户来说,无法建立自己的数据采集队伍,只能通过估算等方式获取近似数据。众包农业大数据采集平台具有定制化数据采集功能,数据需求方可以根据自己的需求按照数据采集的难度、数量、频率等标准发布数据采集任务,数据采集者可以领取任务进行数据采集工作,这样数据需求方只须专注于自己的数据需求而不用再为此建立一支数据采集队伍,相应的数据获取成本也会大幅度降低。

5　总结

应用农业大数据对农业生产效率提高具有重要价值和意义,大数据得以有效应用的前提就是广泛采集多源多类型的农业数据。传统的农业数据采集系统多有部门限制,采集队伍管理难度大、数据采集成本高,限制了农业数据采集的广泛性和普遍性,众包农业大数据采集平台结合互联网领域应用广泛的众包思想,将普通的社会大众都转化成数据采集员、数据质量校验员,有效地扩充了数据采集员队伍,扩大了数据采集覆盖范围,降低了数据采集的成本,为农业大数据的深度应用打下坚实的数据基础。

参考文献:

[1]魏拴成. 众包的理念以及我国企业众包商业模式设计[J]. 技术经济与管理研究,2010(1):36-39.

[2]赵景明,时永梅. 图书馆众包模式的理论与实践研究[J]. 图书馆理论与实践,2011(8):12-13.

[3]刘文华,阮值华. 众包:让消费者参与创新[J]. 经营与管理,2009(7):67-69.

[4]Jonassen D H. Learning to solve problems:an instructional design guide[M]. New york:John Wiley and Sons,2004.

[5]Terwiesch C, Xu Y. Innovation contests, open innovation, and multiagent problem solving[J]. Management Science,2008,54(9):1529-1543.

[6]Trompette P,Chanal V,Pelissier C. Crowdsourcing as a way to access external knowledge for innovation[C]//24 th EGOS Colloquium, Amsterdam,2008:1-29.

[7]Whitla P. Crowdsourcing and its application in marketing activities [J]. Contemporary Management Research,2009,5(1):15-28.

[8]Guido J,英介鈴木. Inside Cisco's search for the next big idea[J]. Harvard Business Review,2010,35(4):64-71.

苏玉宁,姜　艺,陈贺胜,等. 基于 Ontology 的农业科学领域知识库构建[J]. 江苏农业科学,2018,46(5):194 – 198.
doi:10.15889/j.issn.1002 – 1302.2018.05.052

基于 Ontology 的农业科学领域知识库构建

苏玉宁[1], 姜　艺[2], 陈贺胜[3], 朱俊武[2]
(1. 扬州大学农学院,江苏扬州 225000; 2. 扬州大学信息工程学院,江苏扬州 225009; 3. 扬州大学物理科学与技术学院,江苏扬州 225002)

摘要:本体(ontology)的概念起源于哲学领域,在农业信息检索领域中,由于本体可用以解决知识概念表示和知识组织体系方面的问题,因此本体概念引起了农学界专家的高度关注。发达国家在农业科学领域已经建成一些很成熟的领域本体库,并得到了实际应用,为加快这方面的工作,我国在“十一五”计划中,将开展以网络农业信息资源组织为主的农业本体构建技术研究列入其中,因此,农业本体研究是响应国家号召,大力发展和提高我国农业技术和服务水平的重要措施。以农业科学领域中的油料作物——油菜为主要对象,构建农业领域本体知识库。首先在中文科技期刊全文数据库中检索包含油菜的论文题目作为基本语料,然后对检索到的题目利用汉语词法分析系统(institute of computing technology,Chinese lexical analysis system,简称 ICTCLAS)进行分词分割,按词频出现频率筛选农业科学领域的关键词语并进行定义和细化,最后利用 Protégé 软件构建农业科学领域知识库。构建简单的农业科学领域本体库模型,目的是为农业领域实现网络信息快速检索和提高农业信息共享水平打下坚实基础,同时为开展以农业本体服务为目的的农业本体论研究与应用作初步探索。

关键词:本体;农业科学;油菜;汉语词法分析系统(ICTCLAS);中图法
中图分类号: S126　　**文献标志码**: A　　**文章编号**:1002 – 1302(2018)05 – 0194 – 05

本体(ontology)的概念起源于哲学领域[1],20 世纪 90 年代以来,本体的概念被逐步引入人工智能、图书情报和知识工程等领域。由于本体通过对词语的严格定义和词之间的关系来确定词汇的精确含义,因此用本体建立的词汇模型可以让机器理解 Web 页面的语义,可以解决语义层次上 Web 信息共享和交换,因此在语义 Web 中,本体具有非常重要的地位。在农业信息检索领域中,由于本体可用以解决知识概念表示和知识组织体系方面的问题,因此本体概念引起了农学界专家的高度关注。在农业科学领域,发达国家已经建成一些很成熟的领域本体并使其得到了实际应用。为加快这方面的工作,我国在“十一五”计划中,将开展以网络农业信息资源组织为主的农业本体构建技术研究列入其中。因此,农业本体研究是响应国家号召,大力发展和提高我国农业技术和服务水平的重要措施。

1　农业科学领域关键词语的获得

1.1　关键词语的来源

要建立一个农业科学领域本体库(或知识库),这样做的目的是为实现农业信息知识 Web 页面自由检索。领域本体库必须描述关键概念以及概念与概念之间的关系。因此,建立领域本体库的首要工作是列举出农业科学领域中的所有关键概念。农业科学领域知识十分庞大,为保证建立的本体库在检索时的查全率与查准率比较理想,领域本体创建就要尽量包括本领域中尽可能多地概念,尽可能多的把本领域中一些重要概念都包括进去,只有这样才能建立起一个实际可用的农业科学领域本体库。因此,本体概念的选择至关重要。为了使本研究选择的概念具有代表性、说服力,本研究采用的关键词语均来自网络中文科技期刊全文数据库,这是由于科技期刊中的论文大多数为本领域专家的研究成果,他们对本领域的概念比较熟悉和了解,也是本领域专业术语的解释者和本领域未来发展局势的判定者。关键词语是在中文科技期刊全文数据库中检索从 1989 年至 2015 年收录在农学学科,题名或关键字中包含油菜的论文题目。数据库显示共有符合条件的 19 721 条记录结果,本研究摘录论文题目中包含油菜的共 11 242 条记录,作为建立农业科学领域本体库的分词对象。论文题目检索部分结果如表 1 所示。

选择油菜作为检索条件是因为根据骨架法构建本体的流

收稿日期:2016 – 10 – 11
基金项目:江苏高校品牌专业建设工程一期项目(编号:PPZY2015A060)
作者简介:苏玉宁(1975—),女,陕西宝鸡人,硕士研究生,馆员,研究方向为计算机技术。E – mail:ynsu@yzu.edu.cn。

[9]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,2013,15(6):63 – 71.
[10]许世卫. 农业大数据与农产品监测预警[J]. 中国农业科技导报,2014,16(5):14 – 20.
[11]王文生,郭雷风. 农业大数据及其应用展望[J]. 江苏农业科学,2015,43(9):1 – 5.
[12]陶雪娇,胡晓峰,刘　洋. 大数据研究综述[J]. 系统仿真学报,2013,25(增刊 1):142 – 146.
[13]涂新莉,刘　波,林伟伟. 大数据研究综述[J]. 计算机应用研究,2014,31(6):1612 – 1616.
[14]张浩然,李中良,邹腾飞,等. 农业大数据综述[J]. 计算机科学,2014,41(11A):387 – 392.
[15]郭承坤,刘延忠,陈英义,等. 发展农业大数据的主要问题及主要任务[J]. 安徽农业科学,2014,42(27):9642 – 9645.