

科学数据众包处理研究

赵江华^{1,2} 穆舒婷³ 王学志¹ 林青慧¹ 张 兮³ 周园春¹

¹(中国科学院计算机网络信息中心 北京 100190)

²(中国科学院大学 北京 100049)

³(天津大学管理与经济学部 天津 300072)

(zjh@cnic.cn)

Crowdsourcing-Based Scientific Data Processing

Zhao Jianghua^{1,2}, Mu Shuting³, Wang Xuezhi¹, Lin Qinghui¹, Zhang Xi³, and Zhou Yuanchun¹

¹(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190)

²(University of Chinese Academy of Sciences, Beijing 100049)

³(College of Management and Economics, Tianjin University, Tianjin 300072)

Abstract The ultimate goal of acquiring scientific data is to extract useful knowledge from the data according to specific needs and apply the knowledge to specific areas to help decision makers make decisions. As the volume of scientific data becomes larger, and the structure becomes more complex, such as semi or unstructured data, it is difficult to automatically process these data by computers. By incorporating human computing power in data processing, crowdsourcing has become one of the solutions for big scientific data processing. By analyzing the characteristics of crowdsourcing scientific data processing tasks to citizens, this paper studies three aspects, which are talent selection mechanism, task execution mode, and result assessment strategy. Then a series of crowdsourcing-based remote sensing imagery interpretation experiments are carried out. Results show that not only scientific data can be processed through crowdsourcing paradigm, but also by designing reasonable procedure, high-quality data can be obtained.

Key words crowdsourcing; scientific big data; data processing; talent selection; quality assessment

摘 要 获取科学数据的最终目的是根据具体需要从数据中提取有用的知识,并将这些知识应用到具体的领域中,帮助决策制定者制定决策. 由于科学数据规模越来越大,而且呈现结构复杂的特点,如半结构化或非结构化,难以通过计算机实现自动化处理. 众包通过高效调用人力资源,成为进行科学大数据众包处理的解决方案之一. 针对科学大数据众包处理的特点,围绕人才筛选机制、任务处理模式和结果评估策略 3 方面对科学数据众包体系进行研究,并通过地理空间数据云平台开展地学领域的基于众包的遥感影像信息提取实验. 研究表明,科学数据不仅能够通过众包模式来进行处理,而且通过合理的设计众包流程能够获得高质量的数据结果.

收稿日期:2016-11-15;修回日期:2016-12-30
基金项目:国家重点研发计划项目(2016YFB1000600,2016YFB0501900);国家自然科学基金项目(71571133);中国科学院战略性先导科技专项(XDA06010307)
This work was supported by the National Key Research Program of China (2016YFB1000600, 2016YFB0501900), the National Natural Science Foundation of China(71571133), and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06010307).
通信作者:周园春(zyc@cnic.cn)

关键词 众包;科学大数据;数据处理;人才筛选;质量评估

中图法分类号 TP391

科学数据是以解决世界上现存的社会和环境问题为驱动的^[1],必须由数据转化为信息,进而转化为知识才能够得到应用,从而体现其价值^[2].传统的科学数据处理方式为科学工作流.科学工作流系统可自动化科学数据处理任务的编排、执行、监控以及追踪^[3],清楚地说明计算任务和数据质检的关系,便于科学家定义多阶段的计算和数据处理通道,是管理科学数据处理过程的有用工具^[4].

随着数据获取技术的提高,以及计算机网络和数据存储技术的发展,海量科学数据资源的积累和传播成为可能.天文学、大气科学、基因组学、生命科学、地球与空间科学等学科正在产生越来越多的科学数据.据估计,到2020年,天文学产生的数据将达到60PB^[5].科学研究的时空范围越来越大,使得数据处理的工作量急剧增加,从而对科学大数据的高效处理和分析产生更大的需求.

科学大数据的处理和分析工作因其性质可分为自动化处理和非自动化处理.自动化的数据处理工作通常首先设定明确的目标和工作步骤,然后建立科学工作流^[6].对于求解规模大的问题,工作流的各个步骤可在复杂的分布式计算机系统上并行处理,如超级计算机、分布式集群系统、网格或云平台^[7].然而,由于大量的半结构化和非结构化的科学数据产生,使得科学大数据呈现结构复杂性的特点.这类数据往往难以通过计算机实现自动化处理,因此,传统的基于集群或云平台的科学工作流方法已经不能完全满足数据密集型的科学大数据处理.而且人工智能和机器学习等方法往往也需要人工的反馈来学习新的知识和做决策支持,受制于科学家有限的时间和精力,科学大数据处理面临极大的挑战.

众包作为一种分布式的群智计算模式,通过互联网高效地调用分布全球的人力资源,对于处理人类比较擅长而计算机难以自动化处理的任务有很大优势.因此,如何高效地利用人类和计算机组成的全球大脑网络^[8],来解决科学大数据快速精准的处理成为本文主要探索研究的方向.

本文首先对科学大数据众包处理方法、众包管理机制和质量控制研究现状进行总结,分析了基于众包的科学数据处理体系所面临的关键问题和挑战,并对其进行研究;然后通过在地学领域开展基于众包的科学数据处理实验,来对本文提出的解决方

案进行验证;最后对本文进行了总结,并对未来的研究方向进行了展望.

1 相关工作

1.1 数据众包处理方法

众包是由英文合成词 Crowdsourcing 一词翻译而来,最早由 Howe 提出^[9],即将过去由公司内部员工执行的工作,以自由自愿的形式交给大众去完成的做法.作为一种分布式的问题解决和生产模式,众包模式在科学数据处理领域有着极大的优势:1)众包模式打破了专职科学家和业余科学家的界限,不仅使有一技之长的任何个人都可以投入自己的精力、时间和技能参与到科学数据处理和知识创造中,并获得回报,还使得科学家有更多时间和精力投入到对专业水平要求更高的研究创新中^[10];2)公众的参与使得众包成为知识传播的绝佳方式,基于众包的科学数据处理可以提高科学研究的影响力和公众的科学素养^[11];3)研究表明依靠达成一致共识的公众来生产费时且难以自动化,但有科学价值的数据产品,不仅可降低时间和金钱成本,而且在提供有更多训练或评价反馈的情况下,众包得到的数据结果比专家结果更高^[12];4)基于众包的科学数据处理使公众成为劳动力、技能、计算能力甚至是资金支持的重要来源^[13],不仅改变了专职科学家的工作模式,还可降低成本,优化科研资源配置.

根据数据处理内容和公众参与形式,科学数据众包处理方法主要有协作集成式、竞赛选择式和微任务市场3类.

协作集成式是由科学家或研究人员将科学数据处理任务分解为简单任务,公众无需具备任何专业知识即可参与.Zooniverse 即是一个典型案例,科学家首先将科学数据处理工作设计成框选图像内容和添加标记等可重复进行的简单任务,然后在 Zooniverse 平台上发布,公众通过平台所提供的交互可视化平台,在接受一个简单培训后,即可进行分类和标注.例如发布在 Zooniverse 平台上的 Galaxy Zoo 项目,将全球的85万志愿者组织起来,对海量星系图片进行识别和分类,以寻找行星和查看天文物体.已有50多篇科研文章基于该项目产生的数据发表^[14].此类众包方式多采用投票方式进行数据质量的控制,

即将一个任务分发给多人处理,选择多数人相同的结果为正确结果.但由于此类众包任务对参与数据处理的人数要求较高,且任务分解工作复杂,同时还需进行后期的数据结果的分析和整理,因而时间成本较高.

竞赛选择式是将定义好的科学数据处理任务进行在线发布,公众提交相应的解决方案,任务发布方在对解决方案进行分析和评价后,选择最优的方案,并给予方案提供者相应的报酬.竞赛选择式众包平台有 Kaggle^[15], Crowdforge^[16] 和 Innocentive^[17] 等.竞赛类任务通常比较复杂,需要公众具备相关技能并投入大量时间,因此对参与者的要求较高,任务报酬也较高,不适合难以自动化的大规模科学数据处理工作.

微任务市场是首先将大任务划分为多个小任务,然后基于第三方平台,将小任务分配给不同的人处理^[6].此类任务简单、独立,通常需要较少时间和知识就可以完成,报酬通常也非常低,例如识别图片或视频中的物体.而且处理结果往往容易验证,因此可以利用最广大的劳动力资源,以低成本获得很好的结果. Amazon 的 Mechanical Turk(MTurk)即是典型的微任务平台,该平台发布的任务多种多样,包括输入、修改、验证给定信息、比较、分类、信息检索、知识综合、认知实验、判断和决策、用户界面评价等.大量研究表明 MTurk 是一个非常有用的数据采集和评价平台,但由于目前科学数据处理工作复杂性越来越高,有些任务难以进行分解,因此科学数据处理的众包体系需要进一步的完善.

1.2 众包管理机制

众包管理机制可以定义为通过有效地引导用户参与和使用他们的知识与技能来实现特定业务的过程. Saxton 等人认为,与传统企业管理不同的是,众包平台的管理重点不是增强员工的潜在技能和留住员工,而是发现和利用参与者的潜在技能并且吸引更多有才能的人加入^[1].影响用户持续参与众包的原因有很多,很多学者在这方面进行了有意义的研究.

Saxton 等人重点研究了 3 种管理控制系统:酬金方案、建立信任系统和投票评价机制^[1]. Brabham 通过对 Istockphoto 和 Threadless 这 2 个众包社区进行研究发现,影响用户参与众包社区的主要动机有金钱激励、提高个人技能、获得趣味和自我成就感,此外,对社区的认同感和热爱也是用户持续参与

众包行为的重要原因^[18].可以看出,影响用户持续参与行为的因素主要有财务激励和内在激励 2 种.

Lakhani 等人的研究表明金钱和其他奖励是个人参与众包交易的关键动机.除了金钱报酬外,奖励的形式有很多,包括积分、奖品、虚拟货币等^[19]. DiPalantino 等人认为众包是一种基于技能的全支付模式的拍卖,并且发现奖励可以提高用户参与的频率^[20]. Harris 发现在找到合适的人来完成任务的情况下,增加报酬可以提高任务最终的完成质量^[21].目前,像猪八戒这样的众包平台就是通过奖金吸引众多人参与的. Kaggle, Innocentive 等网站的高额奖金也是吸引全世界的科学家参与的重要原因.

但是,很多众包平台,比如 youtube 以及国内的一些字幕组都是没有任何报酬的,吸引他们参与的更多是一些内在需求. Huberman 等人发现 youtube 上没有关注的用户会倾向于停止上传,而持续增加的关注度则会对用户的分享上传行为产生积极影响^[22]. Howe 指出,大众很多都是因为兴趣和爱好参与众包项目^[23]. 夏恩君等人认为,个体参与众包活动的动因既包括内部动机也包括外部动机,完成任务带来的成就感是大众持续参与众包的主要内在动机^[24]. Kaggle 吸引人的另一个原因是它的竞争排名机制已经成为全球顶尖数据科学家证明自己实力的重要参考,参加竞赛增加的经验以及带来的名誉等收获也是激励各方人才积极参加的原因.

在公民科研项目中,参与动机则主要为内在激励. Raddick 研究了志愿者参与这类公民科研的动机.研究发现大部分志愿者参与此类活动的主要动机是参与到科学研究中、为科学研究做贡献.另外,社交互动体验、成就感等也会影响志愿者参与的意愿^[25].此外, Greenhill 等人以 Zooniverse 为例,研究了游戏化行为对于用户的激励作用,研究发现游戏化可以激发用户参与任务、贡献自己的时间和精力,并且有助于提高用户对于平台的忠诚度^[26].

1.3 众包质量控制

质量是任务结果满足任务发布者需求的程度. Allahbakhsh 等人认为质量取决于 2 方面,即发布者的任务设计和任务完成者的资质.任务完成者的资质包括声誉和能力;任务设计是任务发布者发布的信息,包括任务定义、用户界面、任务复杂程度和奖励政策^[27].同时,作者还总结了在任务设计阶段和运行阶段的质量管理措施.

众包面向的用户是不确定的,众包独立匿名的

特点使得对于众包用户的资格审查难度增加,这使得众包的质量具有较大的不确定性.科学大数据的众包对于数据结果质量的要求更高,因此质量控制是设计科学大数据众包平台的重要内容.对众包质量控制的研究集中在 2 方面:1)结果质量控制;2)众包任务设计.

结果质量评估主要是根据用户提交的任务结果的历史数据剔除不合格的参与者.一些众包网站,如 Amazon Mechanical Turk 采取冗余信息标识正确答案,即让大量的人重复做同一任务,然后用投票的方法来决定正确答案.但是,大量的冗余是昂贵的,极大地提高了众包网站的成本.针对此,Dawid 和 Skene 提出了基于期望最大化算法的解决方案.他们的算法首先通过用户提交的结果来估计每个任务的正确答案,然后通过将用户提交的答案与估计的正确答案对比来评估用户质量,最终使用用户的总误差之和来给每个用户的质量评分,从而剔除恶意操作者,提升众包平台的效率^[28].Ipeirotis 等人认为用任务结果对用户进行评价受个人偏好的影响,存在偏差,因此在此基础上提出了消除用户个人偏好、恢复故有误差率的方法,以获得更可靠的质量评估^[29].Tong 等人提出了一个利用众包的方法对复杂版本数据进行检测和修复的模型,用户只需通过对数据进行简单识别.在模型中,作者加入了一个信任模型,根据参与者的历史任务数据对参与者的信用评级,以便筛选出信誉最好的参与者以提高众包质量^[25].

众包的任务设计对于最终的质量也有一定影响.Sorokin 等人发现,酬金的高低会影响不同类型众包参与者的参与意愿,进而影响到任务质量.酬金较低时,对任务感兴趣的人会明显变少;而酬金较高时,则会吸引很多恶意参与者,即想利用规则漏洞赢得酬金而不认真完成任务的用户^[30].Kittur 等人指出众包任务应当保证通过欺骗的手法骗取酬金和认真完成所花的时间基本相同,以此来减少恶意用户行为^[31].此外,一些众包平台还通过只接受发达国家的人承接任务来减少恶意用户,Eickhoff 等人在研究中已经证实了这种做法的有效性,但是也指出,采取人员过滤的做法会使得任务的完成周期变长.他们进一步指出,相比人员过滤的方法,通过任务来过滤用户更加高效,他们发现恶意用户通常较少参与需要较高创造性的任务,因此通过增加任务的复杂程度可以有效减少恶意用户,从而提高任务质量^[32].

2 科学数据众包处理体系

科学数据众包处理体系主要包括众包人才筛选、众包任务处理模式与方法以及众包结果质量评估 3 方面.相比于传统的众包任务,科学数据处理专业性较强,需由专业人士或专业基础较好的业余爱好者来完成,这对人才筛选和管理提出挑战;而且科学数据处理通常为数据密集型和计算密集型,往往需要可扩展的计算和存储资源,如何为用户提供一个私有的计算机存储、处理环境并汇聚相应数据资源成为关键;同时科学数据处理任务对数据质量要求较高,质量控制方法复杂.因此本文主要针对科学数据众包处理体系中的人才筛选机制、任务处理模式和结果评估策略 3 方面展开研究.

2.1 众包人才筛选机制

2.1.1 众包人才评价与管理

研究表明,任务领取人的能力对于任务完成的结果质量影响很大^[31,33],因此,在进行任务分配时,人才选取成为获取高质量数据结果的关键.本文采用人才分级管理机制,通过多种方式的审核,对参与科学数据处理的人才进行评价和分级,管理流程图如图 1 所示:

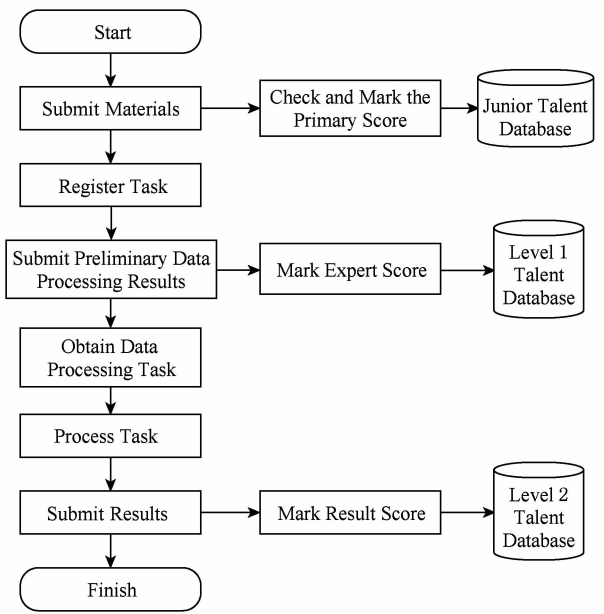


Fig. 1 Flow chart of talents assessment mechanism
图 1 人才评价机制流程图

将所有参与科学数据处理的人才分为 3 类:初级人才、1 级人才和 2 级人才.其中初级人才需在科学数据处理众包平台提交相关材料,经平台审核后,

给予其初级评分 s_0 , 并存储到初级人才数据库. 报名参与科学数据处理众包任务的人需提交数据预处理结果和任务解决方案等补充材料, 经专家评价以及考核后, 给予其专家评分 s_1 , 并存储到 1 级人才库. 对于专家评分高的人才将成为众包任务领取人, 参与科学数据处理. 最终平台根据众包任务领取人提交的数据结果质量, 再次对任务领取人进行结果评分 s_2 , 并存储到 2 级人才库. 具体评分方法如下:

一项众包任务 T 对任务领取人的能力要求集合为 $\{C_1, C_2, \dots, C_z\}$. 对于不同层次的人才评分, 所依据的能力要求不同. 人才的初级评分根据其信息完整程度、数据处理经验和数据处理工具熟练程度等方面来设定; 1 级人才的专家评分的依据是任务申请人时间上是否能够保证完成任务, 以及申请人的数据预处理能力、数据处理能力、编程能力、挖掘算法熟悉程度、数据处理态度等; 2 级人才的结果评分则根据对其提交的数据处理结果的质量评价来确定. 众包候选人对任务数据处理能力 P_x 的满足程度 S_x 可通过模糊性语言变量子集 {不满足, 基本满足, 比较满足, 非常满足} 来表示, 并分别以数值 {0, 0.6, 0.8, 1.0} 来量化表示.

各个层次的人才评分描述为

$$s_i = \sum_{x=1}^z w_x \times S_x, \tag{1}$$

其中 w_x 为众包任务对各方面能力要求的权重, $\sum_{x=1}^z w_x = 1$, z 为完成任务所需的能力的数量, S_x 为各层次人才对数据处理任务所要求的能力的满足程度.

2.1.2 众包人才筛选与任务分配

针对众包任务 T_i , 设定任务能力要求阈值, 阈值需根据子任务的数据处理工作量、重要性等因素确定. 搜索满足能力要求的众包任务候选人, 并进行任务分配, 人才筛选与任务分配算法如算法 1 所示.

算法 1. 人才筛选流程算法.

输入: 任务集 T , 初级人才库 P_0 , 1 级人才库 P_1 , 2 级人才库 P_2 ;

输出: 任务分配结果.

计算任务数 n ;

初始化能力要求阈值 λ_1 和 λ_2 ;

$m_1, m_2 \leftarrow 0$;

for each talent in P_2 do

if $s_2 > \lambda_2$ then m_2++ ;

end if

end for

if $m_2 > n$ then

按分数 s_2 从高到低对人才进行排序;

为 s_2 值最大的 n 个人分配任务;

else

为 m_2 个人分配任务;

for each talent in P_1 do

if $s_1 > \lambda_1$ then m_1++ ;

end if

end for

if $m_1 + m_2 > n$ then

为 s_1 值最大的 $n - m_2$ 个人分配任务;

else

do

根据 s_0 对初级人才进行排序筛选并标记专家评分;

将标记有专家评分的人才加入到 P_1 人才集合中;

while $m_1 + m_2 < n$

分配剩下的 $n - m_1 - m_2$ 个任务;

end if

end if

Return 任务分配结果.

首先将大的科学数据处理任务划分为 n 个子任务; 然后根据数据质量要求, 筛选领取过科学数据处理众包任务的 2 级人才, 如果符合要求的人才数大于或等于 n , 则将任务分配给评分高的前 n 人, 否则先对 2 级人才进行任务分配, 再从 1 级人才库中筛选评分高于设定的能力要求阈值的人才, 并将剩余任务进行分配; 当筛选出的 1 级和 2 级人才数目仍小于 n 时, 则从初级人才库中筛选, 并标记专家评分, 直到任务分配完成.

2.2 众包任务处理模式

针对科学数据处理性质, 众包任务的执行可分为线上和线下 2 种方式. 线下方式由任务领取人下载数据后, 利用个人的计算资源完成科学数据的处理, 最后提交众包结果; 线上方式为任务领取人在线访问和组织数据资源, 并利用在线的计算资源进行数据的处理和存储. 针对线上方式, 平台需为众包领取人提供在线的数据组织和访问, 以及计算资源的管理与使用.

2.2.1 数据资源的组织与访问

众包数据分为公共数据和私人数据. 平台采用分布式文件系统对数据进行存储, 对于不能提供下载的科学数据, 可通过存储在公共数据中, 任务领取

人可进行读取,在线处理,将数据处理结果直接存储到私人数据存储空间中. 数据管理系统如图 2 所示:

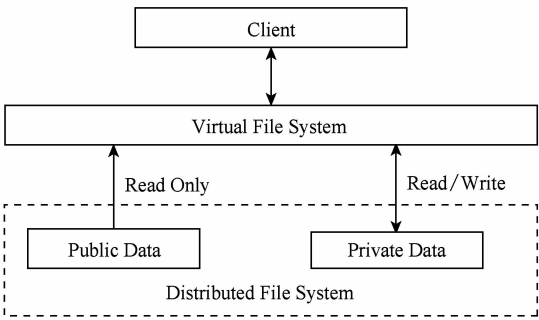


Fig. 2 Data resources management system
图 2 数据资源管理系统

系统采用用户空间文件系统(FUSE 技术)实现符合 POSIX 的文件接口,用户通过虚拟文件系统访问分布式文件系统的数据,公共数据的访问权限

为只读,私有数据则可进行读写操作. 同时任务领取人可通过客户端接口或 FTP 进行数据的组织. 系统允许任务领取人上传本地存储的数据. 基于 jquery/Django 实现的多文件上传模块,前端窗口支持拖拽上传、可对等待上传的文件进行排序管理、实时显示正在上传文件的状态并针对上传失败的情况给出错误提示信息等. 后端针对上传文件的大小采用不同的存储方案:小文件由内存直接写入硬盘;大文件则进行分片存储,接受到的小片文件首先写入临时文件夹,当所有文件全部接受成功后再写入目标文件夹.

2.2.2 计算资源的管理与使用

对于数据量较大的科学数据处理任务,本文采用计算资源隔离容器为任务领取人提供数据快速获取平台和可配置的计算资源,具体架构体系如图 3 所示:

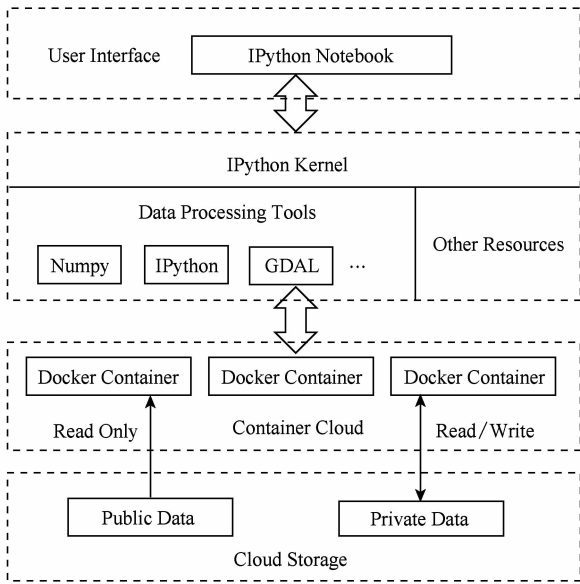


Fig. 3 Architecture of data and computing resources service
图 3 计算资源服务架构图

系统为每个任务领取人提供可扩展的计算资源和云存储环境. 任务领取人可通过 IPython Notebook 编程进行计算资源、数据处理工具和数据资源的调用. IPython Notebook 是一个连接到 IPython Kernel 的基于 Web 的交互式计算环境,允许用户在 Python 环境中灵活定制和实时执行代码,并可交互式地调用数据处理工具包和系统资源. 采用 IPython Notebook 实现,可支持基于网络交互实现复杂的科学计算,如科学绘图、并行计算、Linux 系统 shell 调用等. IPython Kernel 负责管理这些工具资源,并为 IPython Notebook 提供公共接口. IPython Kernel 和数据处

理工具被封装在 Docker 的虚拟文件系统中,每个 Docker 实例执行一个 IPython Kernel,负责接收和处理来自 IPython Notebook 的请求,以及读取云存储中的数据资源.

2.3 众包结果评估策略

质量评价的目的是在一定范围内量化数据结果质量或根据数据质量对数据结果进行分类,识别出合格的数据结果. 本文采用迭代回归的质量控制方法,结合多源数据、专家知识以及定量评价算法等,实现高效、准确的众包结果质量评价. 数据质量控制算法如算法 2.

算法 2. 数据质量控制流程.

输入: 任务处理结果集合 R , 当前任务领取人集合 $K, P=\{P_0, P_1, P_2\}$;

输出: 合格的数据集结果.

```
for each result in  $R$  do
    根据质量评价算法计算  $E(R)$ ;
    if  $E(R)=1$  then
        集成数据结果;
    else if  $E(R)=0$  then
        反馈修改意见给任务领取人  $k$ ;
    else if  $E(R)=-1$  then
        Replace the talent  $k$ ;
    end if
    对任务领取人  $k$  进行评分;
end for
```

Return 合格的数据集成结果.

对于每个众包任务的数据结果, 根据数据质量评价算法计算该任务的评价结果 $E(R)$, 数据质量评价算法如下:

1) 确定数据质量评价指标

参考 Wang 等人提出的全面描述数据质量的指标体系, 包括可信性、准确性、完整性、一致性、易理解性和客观性等质量指标^[34], 根据科学数据处理任务的内容和领域特点, 确定数据质量评价指标集合 $\{I_1, I_2, \dots, I_q\}$.

2) 数据质量评价指标的计算

收集并整理相关数据资源, 利用专家知识和经验, 结合多源数据对数据结果进行各个质量指标的计算, 计算公式如下:

$$I_i = f(E, M, A), \tag{2}$$

其中, I_i 为第 i 个指标的评分, 为取值介于 0 到 100 的实数; E 为专家知识; M 为多源数据资源; A 为定量评价算法库, 数据质量评价指标和定量评价算法库采取“边构建, 边使用, 边修改”的方式来补充和完善.

3) 指标权重的确定

将已完成的同类任务的评价数据作为训练样本, 利用机器学习方法, 如逻辑回归, 进行指标权重参数的优化.

4) 得到评价结果

计算数据质量评价结果:

$$E(R) = g\left(\sum_{i=1}^q \theta_i \times I_i\right), \tag{3}$$

其中, θ_i 为第 i 个指标的权重, $E(R)$ 取值范围为

$\{-1, 0, 1\}$. “-1”代表“放弃”, 表示数据结果完全不合格; “0”代表“修改”, 表示为数据基本合格, 但还需进一步修改和完善, 对于“修改”类的数据结果, 根据评价结果反馈修改意见; “1”代表“合格”, 表示数据质量满足处理要求的结果.

采用定性和定量指标相结合的方法, 不仅可避免评价结果的模糊性, 还可充分利用专家知识提出修改意见, 提高数据处理结果的精度, 同时保障了任务领取人的利益.

3 众包案例研究

随着遥感技术的发展, 新一代遥感平台的出现, 遥感影像在空间分辨率、时间分辨率、光谱分辨率和辐射分辨率上都有了很大程度的提高. 遥感影像信息提取 (remote sensing information extracting) 是从遥感影像上获取目标地物信息的过程. 由于遥感具有覆盖面广、及时快速的优势^[35], 通过遥感影像提取地物类型或土地利用类型信息已经成为监测城市扩张、环境监测和灾害评估的有力手段, 对人类的可持续发展具有重要意义^[36].

随着遥感数据集的不断增长, 各类研究对遥感图像的时间、空间跨度要求越来越大. 由于遥感影像成像条件不同, 目前没有一种完美的分类器或算法能够实现大量遥感影像的全自动提取, 因此, 众包模式成为解决这一问题的解决方案. 本文通过一系列的遥感影像信息提取任务对提出的众包体系进行了实践. 所提取的土地利用类型包括水体、建筑用地、耕地、林地、草地、未利用地等, 覆盖区域包括北京、石家庄、上海、天津、青海、武汉、承德、张家口、苏州等省市.

3.1 数据

Landsat 数据是覆盖全球的中等空间分辨率的对地观测数据, 从 1972 年 7 月发射第一颗 Landsat 系列卫星开始, 已进行了 40 多年的连续对地观测, 成为长期陆表状态及其变化监测研究的最有效遥感数据之一. 2008 年由美国 USGS 免费对全球开放使用后, Landsat 系列数据在生态环境监测、能源与资源管理、灾害监测、城市规划等领域得到广泛应用^[37-40]. Landsat 数据可通过地理空间数据云平台 GSCloud (www.gscloud.cn) 免费下载.

3.2 任务处理方法

根据时空规则将整个数据处理任务划分为多个子任务, 对于大规模遥感影像信息提取任务, 通常采用时间和空间相结合的方法进行任务划分, 每个子

任务由领取人通过人机协作方式使用面向对象方法提取地物信息,最后进行数据质量的控制,并合并各子任务合格的数据处理结果. 基于众包的遥感影像

信息提取的整体流程如图 4 所示. 任务领取人可通过 GSCloud 对 Landsat 影像进行在线处理,如植被指数计算、条带修复等.

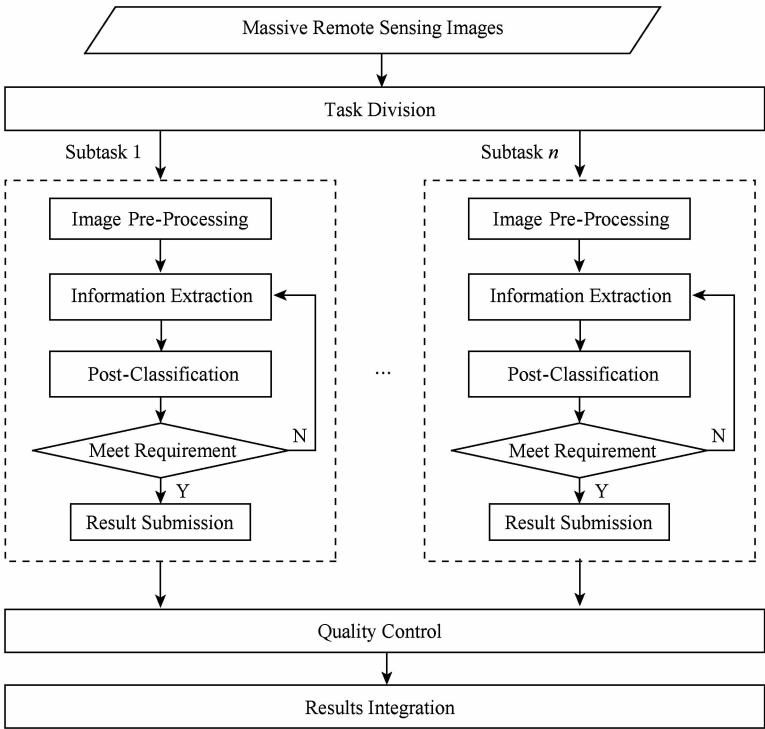


Fig. 4 Flow chart of crowdsourcing based remote sensing image information extraction

图 4 基于众包的遥感影像信息提取流程图

3. 2. 1 影像信息提取方法

为保证信息提取结果的质量一致性,我们对数据处理方法做了规定. 遥感影像信息提取方法可分为面向像元的方法和面向对象的方法. 面向对象方法的基本思想是首先将遥感影像中空间相邻的像元分割成一个个同质性的对象,然后将这些对象作为最小的分类单元利用光谱、纹理、几何和上下文信息对影像进行分类,完成地物信息的提取^[41]. 和面向像元的方法相比,面向对象方法通过图像分割将目标从影像中分离出来,将原始图像转化为更抽象更紧凑的形式,便于特征提取和参数测量,因此使得更高层次的分析和理解成为可能^[42].

在使用面向对象方法进行影像信息提取时,由于遥感影像的成像条件不同,空间变异性高,复杂多样,且遥感影像的分析与理解需要从不同的尺度着手,所以不存在一种完美的分类器可自动准确地提取遥感影像的信息. 实验任务规定采用半自动的遥感影像信息提取方法,即利用人脑对遥感影像的综合理解和分析,将人的知识与经验融入到面向对象信息提取中来,具体方法如图 5 所示.

半自动化遥感影像信息提取方法首先选择并提

取特征进行影像分割,根据目视判读优化分割参数;其次,根据影像分割结果,构建特征向量;再次,选择分类器对影像进行分类;再次,对分类结果进行后处理及质量评价,对于精度较差的结果,重新优化影像分割参数或调整分类器参数. 通过将人机交互和自动计算相结合,不仅降低地物信息识别和提取的错误,还可提高信息提取效率.

3. 2. 2 质量评价方法

针对任务领取人提交的影像解译结果,采用定性和定量相结合的方式对数据质量评价. 首先由 GSCloud 平台专家利用同一区域多时期解译结果以及 OpenStreetMap 数据进行综合评价,评价指标选择准确性、完整性和一致性. OpenStreetMap 为由公众编辑的全世界地图,数据类型包括点、线、面数据. 其中,点数据为感兴趣区;线数据包括道路、水系、铁路等;面数据包括土地利用数据、自然地物、居民区等,其数据精度与专业数据相当^[43]. 通过叠加遥感影像、多时期数据结果以及 OpenStreetMap 数据检查数据的拓扑关系、位置和属性信息,并统计数据质量问题 and 误差信息. 定量评价步骤为:

1) 采用规则采样与选择性采样相结合的方式

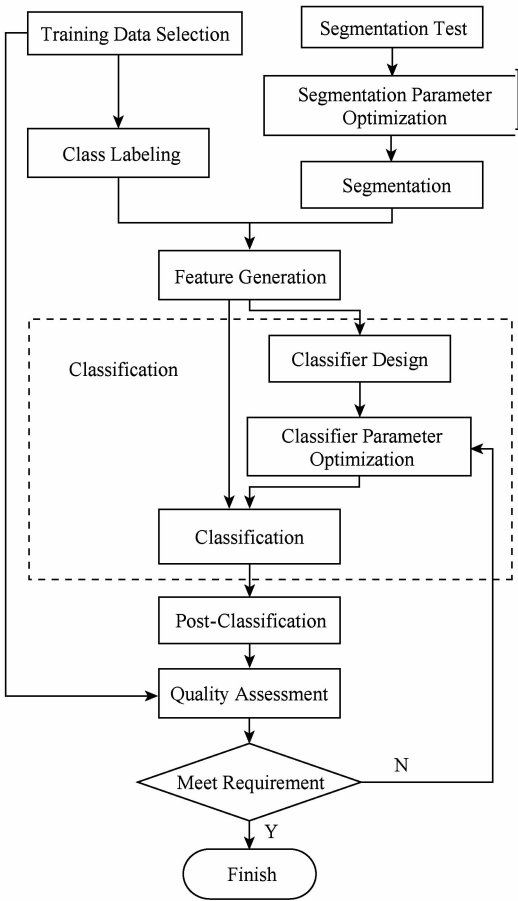


Fig. 5 Semi-automatic remote sensing image information extraction framework
图5 半自动化遥感影像信息提取技术路线图

对研究区进行验证点的混合采样,即在规则采样的基础上再根据定性评价的统计信息对研究区易出现提取信息误差的区域增加采样点;

2) 人工标记采样点真实地物信息作为参考数据,并基于混淆矩阵对分类结果进行精度评价.混淆矩阵是一个用于表示分为某一土地利用类型的像元个数与地面检验为该类别数的比较阵列.矩阵中的列代表参考数据,行代表数据分类结果的土地利用类型,如表1所示.

3) 根据混淆矩阵计算总体分类精度.总体精度(overall accuracy)为被正确分类的像元总和除以总像元数,计算方法如下:

$$OA = \frac{\sum_{i=1}^k N_{ii}}{N}, \quad (4)$$

其中, k 为类别数, N_{ii} 为第*i*类土地利用信息提取正确的像元数; N 为影像的像元总数.规定总体精度大于85%的结果为合格结果.

Table 1 The Confusion Matrix

表1 混淆矩阵

Task Results	Reference Data				
	A	B	C	D	N_{i+}
A	N_{11}	N_{12}	...	N_{1k}	N_{1+}
B	N_{21}	N_{22}	...	N_{2k}	N_{2+}
C	$\vdots$	$\vdots$		$\vdots$	$\vdots$
D	N_{k1}	N_{k2}	...	N_{kk}	N_{k+}
N_{+j}	N_{+1}	N_{+2}	...	N_{+k}	N

3.3 结果

从2015年5月至今,基于地理空间数据云平台累计拟定并发布基于Landsat影像的遥感信息提取任务共36个.GSCloud是一个基于云计算技术的海量地学数据资源查询、下载、在线处理和可视化的服务平台.经过对遥感数据的长期整理、存储与处理,GSCloud已积累了成规模的遥感数据资源,包括LANDSAT,MODIS,Sentinel,EO-1,DEM,NCAR,NOAA及LUCC数据集等,并集聚了全国14万专业用户.用户主要为科研院所的研究人员和学生,以及科技公司专业技术人员,均可利用相应的技能和业余时间参与科学数据处理工作.通过该任务系列的发布,目前GSCloud初级人才库共有专业人员1106人,1级人才库即任务申请人共697人,2级人才库即最后参与数据处理共64人.其中典型的大规模遥感影像信息提取任务——青藏高原5期湖泊提取任务——报名人数为239人,经过初级评分、专家评分以及人才筛选后,共23人参与数据处理,进入2级人才库.至此地理空间数据云已初步形成科学数据采集、数据众包处理与人才管理的生态系统.部分众包任务的数据处理结果已在GSCloud(<http://www.gscloud.cn/help/cases/>)发布.

4 结论

利用众包模式将大量具有一定专业技能的公众的时间和精力有效地聚集在一起来处理计算机难以自动化处理的大量科学数据,一直是一项复杂的挑战.本文针对基于众包的科学数据处理所面临的人才评价与任务分配、数据与计算资源服务以及数据质量评价与控制3个关键问题进行研究,提出了人才分级管理机制、隔离容器提供数据和计算资源、以及迭代回归的质量控制方法.并借助理空间数据云平台在地学领域海量数据积累、在线计算服务、

以及 14 万专业用户的优势,开展了一系列基于众包的遥感影像信息提取实验.结果表明,在有良好的流程机制和平台支持下,公众可以参与到科学大数据处理中来,并产生合格的数据结果.

为了让更多的公众参与到科学数据处理中来,如何进行领域知识的封装,向公众提供简单可用的处理工具或任务完成方法,让无专业知识的公众可以参与其中将成为下一步要解决的问题.同时,建立一个公众可以高效协作的平台还可以加快科学数据的处理速度.可以预见,众包在科学数据处理领域的强大优势将逐渐显现并最终带来科研模式的改变.

参 考 文 献

- [1] Saxton G D, Oh O, Kishore R. Rules of crowdsourcing: Models, issues, and systems of control [J]. *Information Systems Management*, 2013, 30(1): 2-20
- [2] Koulouzis S, Vasyunin D, Cushing R, et al. Cloud data federation for scientific applications [G] //LNCS 8374. Berlin: Springer, 2014: 13-22
- [3] Liu Shaowei, Kong Lingmei, Ren Kaijun, et al. A two-step data placement and task scheduling strategy for optimizing scientific workflow performance on cloud computing platform [J]. *Chinese Journal of Computers*, 2011, 34(11): 2121-2130 (in chinese)
(刘少伟, 孔令梅, 任开军, 等. 云环境下优化科学工作流程执行性能的两阶段数据放置与任务调度策略[J]. *计算机学报*, 2011, 34(11): 2121-2130)
- [4] Juve G, Rynge M, Deelman E, et al. Comparing futuregrid, Amazon EC2, and open science grid for scientific workflows [J]. *Computing in Science and Engineering*, 2013, 15(4): 20-29
- [5] Berriman B, Deelman E, Juve G, et al. High-performance compute infrastructure in astronomy: 2020 is only months away [J]. *Astronomical Data Analysis Software and Systems XXI*, 2012, 461: 91-94
- [6] Kulkarni A P, Can M, Hartmann B. Turkomatic: Automatic recursive task and workflow design for mechanical turk [C] //Proc of CHI'11 Extended Abstracts on Human Factors in Computing Systems. New York: ACM, 2011: 2053-2058
- [7] Liu J, Pacitti E, Valduriez P, et al. A survey of data-intensive scientific workflow management [J]. *Journal of Grid Computing*, 2015, 13(4): 457-493
- [8] Bernstein A, Klein M, Malone T W. Programming the global brain [J]. *Communications of the ACM*, 2012, 55(5): 41-43
- [9] Howe J. The rise of crowdsourcing [J]. *Wired Magazine*, 2006, 14(6): 1-5
- [10] Schenk E, Guittard C. Towards a characterization of crowdsourcing practices [J]. *Journal of Innovation Economics & Management*, 2011, 7(1): 93-107
- [11] Wei Tongqi, Jiang Tao, Tao Siyu, et al. Science sourcing—A new model of scientific cooperration [J]. *Scientific Management Research*, 2015, 33(2): 16-19 (in chinese)
(卫桐圻, 姜涛, 陶斯宇, 等. 科研众包——科研合作的新模式[J]. *科学管理研究*, 2015, 33(2): 16-19)
- [12] See L, Comber A, Salk C, et al. Comparing the quality of crowdsourced data contributed by expert and non-experts [J]. *PLoS ONE*, 2013, 8(7): 1-11
- [13] Source J P C. Citizen science: Can volunteers do real research? [J]. *BioScience*, 2008, 58(3): 192-197
- [14] Smith A M, Lynn S, Lintott C J. An introduction to the zooniverse [C] //Proc of the 1st AAAI Conf on Human Computation and Crowdsourcing. Palo Alto, CA: AAAI, 2013
- [15] Arganda-Carreras I, Turaga S C, Berger D R, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics [J]. *Frontiers in Neuroanatomy*, 2015, 9: 1-13
- [16] Kittur A, Smus B, Khamkar S, et al. Crowdforge: Crowdsourcing complex work [C] //Proc of the 24th Annual ACM Symp on User Interface Software and Technology. New York: ACM, 2011: 43-52
- [17] Simperl E. How to use crowdsourcing effectively: Guidelines and examples [J]. *LIBER Quarterly*, 2015, 25(1): 18-39
- [18] Brabham D C. Crowdsourcing as a model for problem solving an introduction and cases [J]. *Convergence: The International Journal of Research into New Media Technologies*, 2008, 14(1): 75-90
- [19] Boudreau K, Lakhani K. Using the crowd as an innovation partner [J]. *Harvard Business Review*, 2013, 91(4): 60-69
- [20] DiPalantino D, Vojnovic M. Crowdsourcing and all-pay auctions [C] //Proc of the 10th ACM Conf on Electronic Commerce. New York: ACM, 2009: 119-128
- [21] Harris C. You're hired! An examination of crowdsourcing incentive models in human resource tasks [C] //Proc of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the 4th ACM Int Conf on Web Search and Data Mining (WSDM). New York: ACM, 2011: 15-18
- [22] Huberman B A, Subrahmanyam A, Romero D M, et al. Crowdsourcing, attention and productivity [J]. *Journal of Information Science*, 2009, 35(6): 758-765
- [23] Howe J. Crowdsourcing: Why the power of the crowd is driving the future of business [M]. Danvers, MA: Crown Business, 2009
- [24] Xia Enjun, Zhao Xuanwei, Li Sen. Current situation and trend of overseas crowdsourcing research [J]. *Technology & Economy*, 2015, 34(1): 28-36 (in chinese)
(夏恩君, 赵轩维, 李森. 国外众包研究现状和趋势[J]. *技术经济*, 2015, 34(1): 28-36)
- [25] Reed J, Raddick M J, Lardner A, et al. An exploratory factor analysis of motivations for participating in Zooniverse, a collection of virtual citizen science projects [C] //Proc of the 46th Hawaii International Conf on System Sciences (HICSS). Piscataway, NJ: IEEE, 2013: 610-619

[26] Greenhill A, Holmes K, Lintott C, et al. Playing with science; Gamised aspects of gamification found on the online citizen science project-zooniverse [J]. *Game-On*, 2014, (11): 15-24

[27] Allahbakhsh M, Benatallah B, Ignjatovic A, et al. Quality control in crowdsourcing systems; Issues and directions [J]. *IEEE Internet Computing*, 2013, 17(2): 76-81

[28] Dawid A P, Skene A M. Maximum likelihood estimation of observer error-rates using the EM algorithm [J]. *Applied Statistics*, 1979, 28(1): 20-28

[29] Ipeirotis P G, Provost F, Wang J. Quality management on Amazon mechanical turk [C] //Proc of the ACM SIGKDD Workshop on Human Computation. New York: ACM, 2010: 64-67

[30] Sorokin A, Forsyth D. Utility data annotation with Amazon mechanical turk [J]. *Urbana*, 2008, 51(61): 820

[31] Kittur A, Nickerson J V, Bernstein M, et al. The future of crowd work [C] //Proc of the 2013 Conf on Computer Supported Cooperative Work. New York: ACM, 2013: 1301-1318

[32] Eickhoff C, de Vries A. How crowdsourcable is your task [C] //Proc of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the 4th ACM Int Conf on Web Search and Data Mining (WSDM). New York: ACM, 2011: 11-14

[33] Ho C J, Vaughan J W. Online task assignment in crowdsourcing markets [C] // Proc of the 26th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2012: 45-51

[34] Wang R Y, Strong D M. Beyond accuracy: What data quality means to data consumers [J]. *Source Journal of Management Information Systems*, 1996, 12(4): 5-33

[35] Bello O M, Aina Y A. Satellite remote sensing as a tool in disaster management and sustainable development; Towards a synergistic approach [J]. *Procedia-Social and Behavioral Sciences*, 2014, 120: 365-373

[36] Xu Y, Liu Y. Monitoring the near-surface urban heat island in Beijing, China by satellite remote sensing [J]. *Geographical Research*, 2015, 53(1): 16-25

[37] Qin Y, Xiao X, Dong J, et al. Mapping paddy rice planting area in cold temperate climate region through analysis of time series Landsat 8 (OLI), Landsat 7 (ETM+) and MODIS imagery [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015, 105(2016): 220-233

[38] Otukey J R, Blaschke T. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2010, 12 (Suppl): 27-31

[39] Jung H S, Park S W. Multi-sensor fusion of landsat 8 thermal infrared (TIR) and panchromatic (PAN) images [J]. *Sensors*, 2014, 14(12): 24425-24440

[40] Fritz S, See L, McCallum I, et al. Mapping global cropland and field size [J]. *Global Change Biology*, 2015, 21(5): 1980-1992

[41] Blaschke T. Object based image analysis for remote sensing [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2010, 65(1): 2-16

[42] Vieira M A, Formaggio A R, Rennó C D, et al. Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas [J]. *Remote Sensing of Environment*, 2012, 123: 553-562

[43] Haklay M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets [J]. *Environment and Planning B: Planning and Design*, 2010, 37(4): 682-703



Zhao Jianghua, born in 1989. PhD candidate from University of Chinese Academy of Sciences. Her main research interests include massive data processing, data mining and analysis, Web mining.



Mu Shuting, born in 1993. MSc. Her main research interests include big data analysis for social and crowdsourcing.



Wang Xuezhi, born in 1979. PhD, associate professor. His main research interests include massive temporal-spatial data processing and analysis.



Lin Qinghui, born in 1979. PhD, associate professor. Her main research interests include massive data resource aggregation, analysis, and sharing.



Zhang Xi, born in 1982. PhD, professor. His main research interests include big data analysis for social, business and science, crowdsourcing, and data policy and practice.



Zhou Yuanchun, born in 1975. PhD, professor. Senior Member of CCF. His main research interests include data mining, and big data processing.