

国内外农业大数据应用研究分析^①

黎玲萍^{②*} 毛克彪^{③**} 付秀丽^{***} 马莹^{**} 王芳^{****} 刘勃^{**}

(* 北京化工大学信息科学与技术学院 北京 100029)

(** 中国农业科学院农业资源与农业区划研究所, 呼伦贝尔草原生态系统国家野外科学观测研究站 北京 100081)

(*** 北京石油化工学院信息工程学院 北京 102617)

(**** 湖南科技大学地理空间信息技术国家地方联合工程实验室 湘潭 411201)

摘 要 针对农业领域数据规模大、数据结构复杂、空间数据挖掘能力不足等问题,研究了大数据开源技术在农业领域的数据分析体系中的应用。借鉴国内外学者在农业大数据的研究成果,基于农业数据时空属性的特征,结合农业数据的特点分析了 Hadoop、Storm 和 Spark 开源大数据挖掘技术,归纳性阐述了如何开发适合农业需求的大数据系统。最后,简要分析了农业大数据技术所面临的挑战和研究难题,指出需要加大力度进一步深入理论和应用研究,从而推动和实现基于数据的科学决策,为国家粮食提供安全保障。

关键词 大数据, 农业大数据, Hadoop, Storm, Spark

0 引 言

大数据是结构复杂、内容多样的海量数据,具有数据规模大、数据种类多、处理速度要求高、数据价值密度低等特征^[1],对其处理远远超出了现有传统的计算技术和信息系统的能力,因此,寻求有效的大数据处理技术、方法和手段已经成为当前的迫切需求。大数据技术是以新数据处理技术为手段,将不同格式、不同领域的大数据整合成标准统一的数据源,经研究、分析、挖掘后,为各行业提供有用的数据和知识^[2,3]。在 2012 年 3 月,美国政府正式启动“大数据发展计划”,旨在利用大数据技术在科学研究、环境、生物医学等领域进行技术突破^[4]。2013 年, Mike Gautlieri 提出了存储过程访问(SPA)框架,指出大数据是可用于支撑行业的高效运营、决策选择、风险规避和服务用户,具备高性能存储、解析和使用所有数据的前沿领域技术^[5]。 Aaron McKenna

应用大数据分析技术,利用 MapReduce 架构嵌入基因分析工具集 GATK 来处理和分析 DNA 序列海量数据,取得了较好成效^[6]。2013 年末,中国计算机学会(CCF)发布了《2013 年中国大数据技术与产业发展白皮书》^[7],介绍了大数据应用现状并探讨了大数据技术研究面临的科学问题,针对大数据技术产业的发展提供了政策指导和研究建议。

随着大数据相关技术的不断突破、发展以及公众对大数据的认识加深,大数据技术分析和处理现已成为各个领域的研究热点,诸如商业、医疗行业、金融行业等,都取得了相当的研究成果,大数据的应用也在不断的延伸。农业大数据技术已应用于农业信息领域,使用多粒度、多层次、多渠道的分析模型对庞大的数据总体进行挖掘分析,为农业信息技术带来了革命性进展,促进了农业产业的整体进步。我国在农业信息化发展上需要统一规范数据标准,迫切需要解决科学施肥、水肥调控以及品种选择、优化产业结构布局,并给出可靠专业的决策结果^[8]。

① 国家自然科学基金(41571427),中央级公益性科研院所专项资金(IARRP-2015-26)和农业部农业信息预警专项(614-1)资助项目。

② 女,1991 年生,硕士;研究方向:地层剖面数据建模;E-mail: 823947116@qq.com

③ 通讯作者,E-mail: maokebiao@126.com

(收稿日期:2015-12-23)

然而,农业生产周期长,影响因子复杂,而且在农业生产过程中涉及到育种、种植、培育、管理、收获、储藏、运输以及农产品加工和销售等各环节,其中蕴藏着大量农业信息动向及市场变化、农业科技发展等重要数据。采用大数据技术的手段与方法,可对采集到的影响农业生产过程的温度、湿度、光照、水质状况、气象状况^[9]、市场动态等信息进行挖掘、分析、应用,也包括跨行业、跨专业的数据分析与挖掘^[10],从而有望缩短农业研究的周期,加速科研成果转化推广的进程,同时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^[11]。例如,2012年,土壤抽样分析服务商 Solum 通过使用大数据分析技术来确定化肥的投入量问题,帮助农民提高生产、降低成本。跨国农业生物技术公司 Monsanto 通过分析海量天气数据来预测未来可能对农业生产造成破坏的各种天气,农民可以根据这种预测来选择相应的农业保险,以降低恶劣天气对农业生产造成的影响^[12]。

农业是我国的基础产业,为了实现精准农业^[13]、智慧农业、现代化农业^[14],国内专家将大数据分析技术与农业紧密结合,取得了不错的成效。杨锋等人为解决海量农业数据在传统分布式数据库架构中资源效率不高及存储能力不足的问题,提出基于 Hadoop 的大文件分块存储方法和海量农业数据资源检索方法。该方法可为构建海量农业数据资源管理平台提供支持,实现海量农业数据资源高效的组织和管理^[15]。而且,为了全面、及时地采集到农业数据,量子数聚(北京)科技有限公司搭建了一个农业大数据应用云平台,对获取到的复杂的多源化的农业数据进行处理,该平台可整合国家权威机构发布的农业相关数据、共享数据,汇集政府、企业、社会三方数据,打破信息孤岛,实现资源互联互通,采集的涉农企业数据,可帮助用户准确定位企业以及群体的地理分布。该平台为科研机构、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^[16]。

大数据技术拥有出色的数据分析挖掘能力,在各个领域都有广泛的应用前景。目前能够进行大数据挖掘、分析的工具和软件很多,功能与使用方法也

各有不同。本文分析了国内外大数据应用软件,同时重点分析了大数据技术在农业领域的应用,并对大数据技术在农业领域的应用进行了展望。

1 国内外大数据软件应用研究

1.1 Hadoop

2004 年,Google 发布了关于分布式计算框架 MapReduce 的大规模数据并行处理技术的论文。Doug Cutting 根据 Google 提出的设计思想,用 Java 设计出一套与 Nutch 分布式文件系统(Nutch Distributed File System, NDfs)相结合且支持 Nutch 搜索引擎的并行处理软件系统。该系统从 Nutch 中分离出来并被命名为“Hadoop”。Hadoop 作为 Apache 最大的一个开源项目,是以 Hadoop 分布式文件系统(Hadoop Distributed File System, HDFS)和 MapReduce 为核心的大数据处理平台和生态系统,该系统包含了 HBase、Hive、Zookeeper 等一系列相关子项目。Hadoop 的技术架构如图 1 所示。

分布式文件系统 HDFS 作为 Hadoop 最核心的设计之一,为 Hadoop 分布式计算框架提供高性能、高可靠、高可扩展的存储服务。HDFS 是一个典型的主从架构,其主要包含一个名字节点 NameNode(主节点)和多个数据节点 DataNode(从节点),并提供应用程序访问接口。NameNode 是整个文件系统的管理节点,负责文件系统名字空间的管理与维护,同时负责客户端文件操作的控制以及具体存储任务的管理和分配;DataNode 提供真实文件数据的存储服务,启动 DataNode 线程同时向 NameNode 汇报数据块的情况^[17],系统架构如图 2 所示。

在分布式文件系统 HDFS 中,可能会有多个机架 Rack,每个机架上管理多个 DataNode。为防灾容错,一个数据块 Block(数据块大小默认 64M)的三个副本通常会保存到两个或两个以上的机架中,例如,一个放在机架 Rack1 中的 DataNode1 中,一个放在不同机架 Rack2 的 DataNode1 中,另一个则放在 Rack1 的 DataNode2 中,此分配方法确保系统中的机架掉电或者机架的交换机发生错误时,系统的运行不受到影响,以及数据不会遗失。DataNode 通过

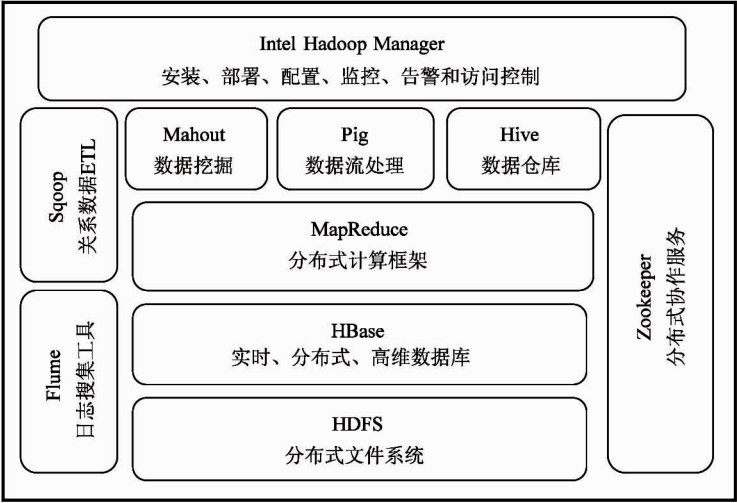


图 1 Hadoop 技术架构

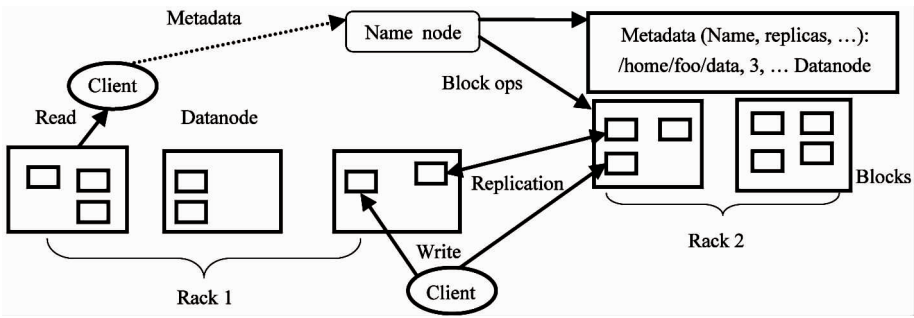


图 2 HDFS 架构图

每 3 秒向 NameNode 发送心跳来保持通信的,一旦心跳停止,将会认为 DataNode 错误,NameNode 会将其上的数据块复制到其他的 DataNode 上。

分布式计算框架 MapReduce 作为 Hadoop 另一最核心的设计,是一个针对大规模群组中的海量数据处理的分布式编程模型。MapReduce 实现两个功能。其一,Map 把一个函数应用于集合中的所有成员,返回一个基于这个处理的结果集。其二,Reduce 把通过多个线程、进程或者独立系统并行执行处理的结果集进行分类和归纳。Map 和 Reduce 两个函数可并行运行。MapReduce 程序的第一步叫做 mapping,将数据元素作为 Mapper 函数的输入数据,每次一个,Mapper 把每次 mapping 得到的结果单独传输到输出数据元素里。Mapper 模型如图 3 所示。

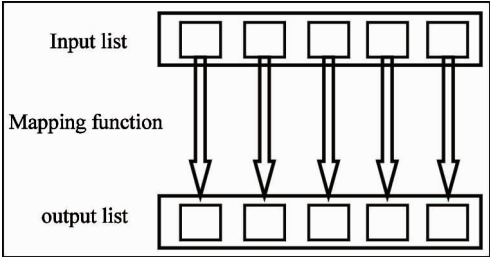


图 3 Mapper 模型

Mapping 对输入数据列表中的每一个元素应用一个函数创建一个新的输出数据列表,即键值对^[17]。Reducing 具有聚集数据功能,Reducer 函数接收来自输入列表的迭代器,将数据聚集在一起,返回一个值。Reducer 模型如图 4 所示。

Reducing 一般用来生成汇总数据,将大规模的数据转变成更小的总结数据。数据通过 Reducer 的分析处理,最终将结果输出。通常作业的输入和输

出都存储在文件系统中,MapReduce 框架和分布式文件系统运行在一组相同的节点上。MapReduce 框架可在已经存好数据的节点上高效运行,使整个集群的网络带宽高效利用。

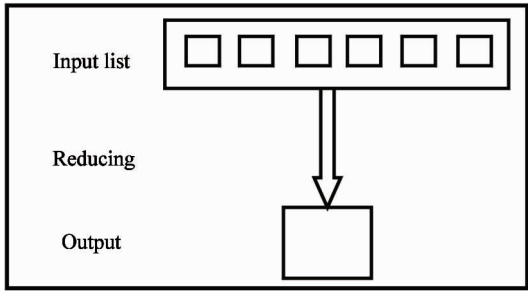


图4 Reducer 模型

Hadoop 具有高可靠性,高扩展性,高容错性,高效性等特点。这些特点符合农业数据处理的要求,可将 Hadoop 用于农业数据的处理中,分布式文件系统 HDFS 可存储农业数据,结合 MapReduce 将数据处理,输出处理结果,为用户提供技术支持。Hadoop 将海量农业数据分割于多个节点,由每一个节点并行计算,将得出的结果归并到输出。运用迭代算法的思想,构建树状结构的分布式计算图,并行、串行结合的计算在分布式集群的资源下高效处理。Hadoop 数据处理方法相对传统数据分析方法不仅快速,而且在处理农业数据时将多元化的农业数据融合分析,分析出数据的相关性。

Hadoop 在农业数据的分析处理中具有很好的优势,但是仍存在不足。第一,Hadoop 的高可用性。HDFS 与 MapReduce 都采用单主机(master)的处理方式,集群中的其他机器都与中心机器进行通信,一旦中心机器损坏,集群停止工作,可用性降低,即单点失败。目前,针对这个已研究出初步解决方案,即对 master 节点进行镜像备份,一旦 master 出现错误,即刻换上镜像备份的机器,提高可用性,但这是基于内置的解决方案,不能从根本上解决问题。第二,Hadoop 不适用迭代次数过多的算法(比如:矩阵的奇异值分解)。MapReduce 每次迭代都需将数据往文件里面读写一遍,浪费大量的时间。第三,不能实现实时性,不能响应毫秒级响应,事物型查询用时

较长。农业数据中含有海量的非结构数据,Hadoop 处理起来十分的复杂,处理速度缓慢,一旦发生单点失败,机器需重启重新处理,会导致处理时间过长。而且 Hadoop 不能实现实时性,导致很多的数据不能及时处理,对于需要实时分析数据输出结果以便作出调整的指标没有作用。

相对于 Hadoop,Storm 主要采用全内存计算,Storm 接收到数据就实时处理并分发,因而 Storm 处理速度快,被广泛应用于实时日志处理、实时统计、实时风控、实时推荐等场景中。

1.2 Storm

2011 年,Hadoop 凭借着高吞吐,方便处理海量数据的优势奠定其在大数据的主要地位。Hadoop 专注于大规模数据存储和处理,但是 Hadoop 的 MapReduce 延迟大,响应缓慢,运维复杂,不适用于实时大数据方面。针对业务场景中对秒级别甚至毫秒级别响应的需求,Twitter 公司推出开源分布式、容错的实时流计算系统 Storm,解决了大规模数据实时处理的问题。Storm 是一个分布式的、容错的实时计算系统,适用于计算机集群中编写与扩展复杂的实时计算。

Storm 集群由一个主节点和多个工作节点组成,Storm 架构图如图 5 所示。主节点运行 Nimbus 守护进程,用于分配代码、布置任务及故障检测。每个工作节点运行 Supervisor 守护进程,开始并终止工作进程。ZooKeeper 用于管理集群中的不同组件,协调 Nimbus 和 Supervisor 两者的工作,Topology 是由计算节点组成的图,节点包括处理的逻辑。

Storm 的工作过程是 Nimbus 针对该拓扑建立本地的目录,根据 Topology 的配置计算作业、分配作业,在 zookeeper 上建立任务节点 assignments 存储作业 task 和 supervisor 机器节点中 worker 的对应关系,以及创建作业节点 taskbeats 来监控作业的心跳,启动 Topology。Supervisor 从 zookeeper 上获取分配的作业 task,启动多个 worker,worker 生成作业 task,一个作业 task 一个线程;根据 Topology 信息初始化建立作业 task 之间的连接;作业和作业之间通过可嵌入的网络通讯库管理,最后整个拓扑运行起来。

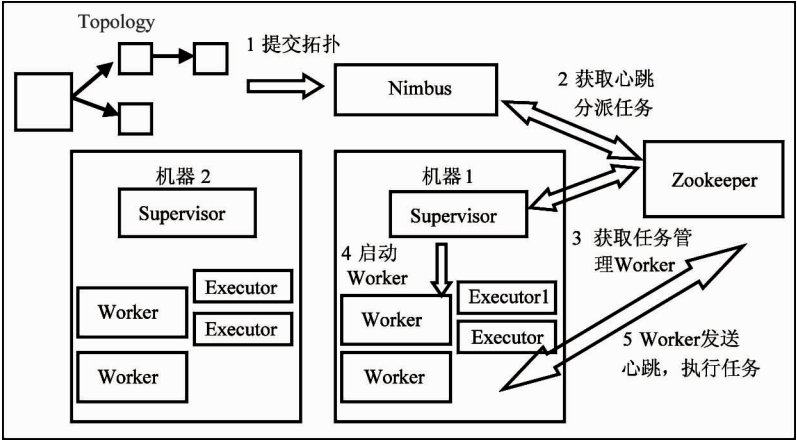


图 5 Storm 架构

Storm 有很多的优势。第一,简单的编程模型。Storm 降低实时处理的复杂性。第二,可以使用各种编程语言。Storm 默认支持 Clojure、Java、Ruby 和 Python。通过实现一个简单的 Storm 通信协议,可增加对其他语言的支持。第三,容错性。Storm 管理工作进程和节点的故障。第四,水平扩展。计算在多个线程、进程和服务端之间并行进行。第五,可靠的消息处理。Storm 保证每个消息至少能得到一次完整处理。第六,快速。系统的设计使用网络通讯库作为其底层消息队列,保证消息能得到快速的处理。第七,本地模式。Storm “本地模式”可在处理过程中完全模拟 Storm 集群,可快速进行开发和单元测试。

Storm 也存在一些亟待解决的问题。第一,目前的开源版本中只是单节点 Nimbus,故障只能自动重启,可考虑实现一个双 Nimbus 的布局。第二,Clojure 是一个在 Java 虚拟机 (Java Virtual Machine, JVM) 平台运行的动态函数式编程语言,优势在于流程计算,Storm 的部分核心内容由 Clojure 编写,性能提高的同时也提升维护成本。因此,Storm 适合用于需要实时分析的农业数据,如农业大棚中的温度、湿度监测等,可以迅速地分析出结果。

1.3 Spark

Spark 是起源于美国加州大学伯克利分校 AMPLab 的大数据计算平台,在 2010 年开源,目前是 Apache 软件基金会的顶级项目。2014 年 11 月,Spark 在 Daytona Gray Sort 100TB Benchmark 竞赛中利用 1/10 的节点数,把 100TB 数据的排序时间从 72min

提高到了 23min,打破了由 Hadoop MapReduce 保持的排序记录。随着 Spark 在大数据计算领域的崭露头角,越来越多的企业开始关注和使用。

弹性分布式数据集 (resilient distributed dataset, RDD) 是 Spark 的最基本抽象,是对分布式内存的抽象使用,实现以操作本地集合的方式来操作分布式数据集的抽象实现。RDD 是 Spark 最核心的东西,其表示已被分区、不可变的、可序列化的、能被并行操作的数据集合,不同的数据集格式对应不同的 RDD 实现。RDD 缓存到内存中,RDD 数据集的操作基于内存,节省 MapReduce 大量的磁盘 IO 操作。提升迭代机器学习算法、交互式数据挖掘的效率。

Spark 提供多种数据集操作类型,如 map, filter, flatMap, sample, groupByKey, reduceByKey, union, join, cogroup, mapValues, sort, partitionBy 等操作类型,将这些操作类型称为 Transformations。同时还提供 Count, collect, reduce, lookup, save 等多种 Actions。Transformation 操作是通过转换从一个或多个 RDD 生成新的 RDD。Action 操作是从 RDD 生成最后的计算结果。Spark 编程模型相对灵活,用户可以命名、物化、控制中间结果的分区。

Spark 架构采用 Master-Slave 模型。Master 对应集群中的含有 Master 进程的节点,Slave 是集群中含有 Worker 进程的节点。Master 作为整个集群的控制器,负责整个集群的正常运行;Worker 是计算节点,接收主节点命令与进行状态汇报;Executor 负责任务执行;Client 作为用户的客户端负责提交应用,Driver 负责控制一个应用的执行,如图 6 所示。

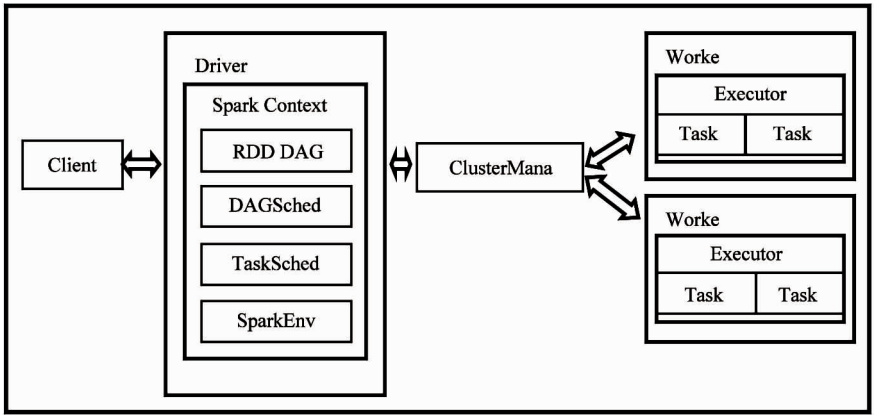


图 6 Spark 架构

Spark 集群部署后,主节点和从节点分别启动 Master 进程和 Worker 进程,对整个集群进行控制。在 Spark 应用的执行过程中,Driver 程序是应用逻辑执行的起点,负责作业的调度,即 Task 任务的分发,Worker 管理计算节点和创建 Executor 并行处理任务,SparkContext 控制应用的生命周期。DAG Scheduler 是根据作业(Job)构建基于 Stage 的 DAG,并提交 Stage 给 TaskScheduler。TaskScheduler 将任务(Task)分发给 Executor 执行。SparkEnv 是线程级别的上下文,创建并包含一些重要组件的引用。ClusterManager 控制整个集群,监控 Worker。在执行阶段,Executor 为执行器,是在 worker node 上执行任务的组件,用于启动线程运行任务。Driver 将 Task、file 和 jar 序列化后传递给对应的 Worker 机器,同时 Executor 对相应数据分区的任务进行处理。

Spark 运行的整体流程为:Client 提交应用,Master 找到一个 Worker 启动 Driver,Driver 向 Master 或者资源管理器申请资源,将应用转化为 RDD Graph,由 DAGS cheduler 将 RDD Graph 转化为 Stage 的有向无环图提交给 TaskScheduler,由 TaskScheduler 提交任务给 Executor 执行。在任务执行的过程中,其他组件协同工作,确保整个应用顺利执行。

Spark 采用了内存计算。从多迭代批处理出发,允许将数据载入内存作反复查询,此外还融合数据仓库、流处理和图形计算等多种计算方式。Spark 构建在 HDFS 上,能与 Hadoop 紧密结合。Spark 可用一个技术堆栈解决云计算大数据中流处理、图技术、机器学习、交互式查询、误差查询等所有的问题,可

以快速地处理农业非结构数据,处理速度比 Hadoop 要快。Spark 正在逐渐走向成熟,并在农业大数据分析这个领域扮演更加重要的角色。

2 农业大数据应用研究分析

大数据在 IT 等行业已经有了相当的研究成果,大数据与农业等传统行业结合是未来的研究热点。农业大数据作为大数据的重要分支,包括气象数据、生物信息数据、环境数据、农业生产数据、管理数据、市场数据和统计数据等,涉及种植、养殖、加工等方面。面对复杂、多源化、大量的农业数据,传统的数据分析方法已经不能满足现代农业的要求,存在着数据分析不到位、数据存储空间不足、数据分析不及时等问题。随着大数据与农业的深度融合发展,依靠传统方法不能解决的诸多问题也会迎刃而解。通过农业大数据技术,对农业领域的粮食安全、生态环境、病虫害预防、动植物育种、农业结构调整、农产品市场预测等诸多问题进行预测和提升。用大数据技术整合分析农业领域数据,改变传统农业生产缺少量化数据支撑的问题,针对农业高度分散、生产规模小、时空变异大、规模化程度差、稳定性和可控程度低等行业性弱点,通过对历史数据和现实数据进行挖掘,很有可能发现和解决新的农业科学问题,拓展人类探索农业科学的广度和深度,提升农业创新水平;另一方面挖掘农业大数据技术的发展前景,丰富了大数据的应用功能,增强了农业与其他领域之间的联系。

通过农业大数据技术与互联网络、计算机技术的结合,农业信息化已经融入农业产业,对于加快转变农业发展方式,建设现代农业具有重要的推动作用。目前,大力发展农业大数据技术,加快推进信息化发展,促进信息化和现代化融合,已经成为各国发展农业的重要趋势。在英国,提出将大数据技术应用于农业领域,农业向“精准农业”迈进,与数字技术、传感技术和空间地理技术结合,更为精准地进行种植和养殖作业;另一方面,通过农业大数据技术数据搜集和分析处理平台,提升农业生产部门和市场需求的对接,加强其对于市场的理解。在 2013 年,英国政府正式启动“农业技术战略”,提出高度重视利用“大数据”和信息技术提升农业生产效率。在美国,为推进信息化以支撑农业发展,美国政府提出利用农业大数据技术,采取政府投入与资本市场运营相结合的投资模式,从农业信息技术应用、农业信息网络建设和农业信息资源开发利用等方面全方位推进和完善农业信息化建设,构建规模和影响力较大的涉农信息数据中心,有力促进农业整体水平的提高。在法国,农业部门完备农业信息数据库,结合农业大数据技术,分析整合各项农业数据,通过互联网络,定期公布农业生产信息,管控农业生产销售环节的正常秩序,向农民提供更为详尽与专业的农业信息资讯,更好地服务于农业生产。目前,德国正致力于发展更高水平的“数字农业”。通过大数据和云技术的应用,将天气、土壤、降水、温度、地理位置等数据上传到云端进行处理,结果数据发送到智能化的大型农业机械上,指挥机器进行精细作业。但是,德国农业数字化建设面临农村地区宽带覆盖率不够高,以及数据安全等问题^[18-20]。在农业大数据的应用方面,我国已经初步构建基于物联网技术的数据采集,基于互联网、移动互联网及 GPRS 的信息传输,基于数据挖掘技术的分析及可视化技术的数据展示技术体系,其应用正向着产前、产中、产后的整个农业生产过程延伸^[21]。由此可知,发展农业大数据技术有利于构建面向农业的综合信息服务体系,为农业生产提供综合、高效、便捷的信息服务。加强农业大数据技术建设,完善农业生产数据采集、传输、共享基础设施,建立农业生产数据采集、运算、

应用、服务体系,推动农业信息化快速发展。统筹国内国际农业数据资源,强化农业资源要素数据的集聚利用,提升预测预警能力。整合构建国家涉农大数据中心,推进各地区、各行业、各领域涉农数据资源的共享开放,加强数据资源发掘运用。加快农业大数据关键技术研发,加大示范力度,提升生产智能化、经营网络化、管理高效化、服务便捷化能力和水平。

农业科学中因果关系较为复杂,大数据技术可以突破传统数据分析方法的局限,对农业领域中的相关关系进行分析。动态流数据^[22]和非结构化数据是大数据技术分析的重点,Hadoop、Storm、Spark 等作为开源的大数据处理平台,在大数据领域起到至关重要的作用。Hadoop^[23]作为一种开源的架构适合在廉价机器上对各种资源数据进行分布式存储和分布式管理,具有可伸缩性和高容错性。在 Hadoop 分布式文件系统 HDFS 的基础上,对目前传统的云存储架构进行适当的改进,可以实现对海量农业数据的存储管理。Storm 适用于在线实时数据流的处理,在农业领域可用于分析大棚湿度、温度、二氧化碳含量等实时数据的处理,及时快速地将大棚内的状况反馈给用户,方便用户调节大棚内的各项指标。Spark 是离线的大数据处理技术,可以在离线情况下快速的处理农业数据。Spark 是基于内存计算的开源项目,其处理速度相比 Hadoop 更快,将 Hadoop 和 Spark 结合,即将数据存储在分布式文件系统 HDFS 中,采用 Spark 内存计算的方式,提高数据的处理速度,及时地处理数据。

然而,想要农业大数据挖掘工具在农业中发挥实际的功效,我们还有许多具体的内容需要研究和实践,主要包括数据存储方面和数据挖掘分析方面。在数据存储方面,由于农业数据具有时空属性,如土壤类型众多,农作物病虫害次数频繁且症状变化,受气候、肥水等相互之间关系和影响较大^[24],因而农业数据具有多维、动态、非线性等特征,导致在数据存储方面传统关系数据库难以满足农业大数据的存储需求,难以对非结构的农业大数据进行加工和管理。在数据挖掘分析方面,由于农业基础数据存在数据资源薄弱、数据结构不合理、数据粒度不够、数

据标准化与规范化程度低等问题,数据分析过程中会要求考虑时间因素、经济学因素等要求,为保证数据分析的精度,大数据分析技术不仅要具备对非结构数据高吞吐量,而且还需多角度分析数据。因而要加快我国现代化农业的建设步伐,完成海量农业数据存储、管理和分析,建立适合于农业体系的大数据系统。

3 结 论

大数据的产生给海量信息处理技术带来新的挑战,大数据技术的方法、手段的运用,促进了金融、证券、保险等领域的信息化发展。同时,在科学研究领域,将大数据技术架构到传统的研究方法中,进一步推进了实验研究的准确性和快速性。大数据技术在我国农业领域的应用研究刚刚起步,拥有市场和发展的潜力,农业朝着信息化和智能化的方向发展。农业大数据技术从复杂的海量农业数据中,分析获得有价值的信息,对提高农户决策水平和提高农业科技含量、农业生产效率等都有着重大意义。文中分析了 Hadoop、Storm、Spark 开源的大数据挖掘技术,将这些开源的大数据技术应用到农业中,通过采集数据、分析数据、发布有价值信息,从而为农业服务。但由于大数据挖掘技术软件在农业领域的应用尚不成熟,数据分析过程不稳定,可能会导致分析结果达不到预期的目标,因而需要加大力度改造大数据技术软件在农业方面的应用。农业大数据技术以数据流引领农业信息化发展方向,深刻影响农业分工的组织模式,促进农业生产组织方式的集约和创新。农业大数据技术推动农业生产要素的网络化共享、集约化整合、协作化开发和高效化利用,改变了传统农业的生产方式,可显著提升农业生产运行水平和效率。在农业现代化建设中,夯实农业大数据在智慧农业中的应用,加强农业大数据采集分析、共享开放和开发利用,提升以农业大数据为支撑的农业信息化,开拓智慧农业新局面,实现农业供应链的改造、农产品流通体系的再造以及价值体系的再造,实现与农民的交互、满足农业研究的专业化和个性化需求,促进农业信息化的发展。

参考文献

- [1] 建光,姜巍. 大数据的概念、特征及其应用. 国防科技, 2013,34(2):10-17
- [2] 蔡书凯. 大数据与农业:现实挑战与对策. 电子商务, 2014,1:3-4
- [3] 官建文,刘振兴,刘扬. 国内外主要互联网公司大数据布局与应用比较研究. 中国传媒科技, 2012,17:45-49
- [4] 孟小峰,慈祥. 大数据管理:概念、技术与挑战. 计算机研究与发展, 2013,50(1):146-169
- [5] Mike G. Big data in 2013:What to expect. *Information & Management*, 2013,3/4:20-25
- [6] McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 2010,20(9):1297-1303
- [7] CCF 大数据专家委员会. 中国大数据技术与产业发展白皮书. <http://www.ccf.org.cn/sites/ccf/ccfziliao.jsp?contentId=2774793649105>; 中国计算机学会, 2013
- [8] 王儒敬. 我国农业信息化发展的瓶颈与应对策略思考. 中国科学院院刊, 2013,28(3):337-343
- [9] 覃志豪,唐华俊,李文娟等. 气候变化对农业和粮食生产影响的研究进展与发展方向. 中国农业资源与区划, 2013,34(5):1-7
- [10] 温学江. 农业大数据研究的战略意义与协同机制. 高等农业教育, 2013,11:3-6
- [11] 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. 安徽农业科学, 2014,42(27):9642-9645
- [12] Monsanto Company. Monsanto Acquires The Climate Corporation. <http://www.monsanto.com/features/pages/Monsanto-acquires-the-climate-corporation.aspx>; Monsanto Company, 2002
- [13] 徐可英. 国内外精确农业发展现状与对策. 中国农业资源与区划, 2000,21(2):53-56
- [14] Li D. Internet of things and smart agriculture. *Agricultural Engineering*, 2012,2(1):1-7
- [15] 杨锋,吴华璃,朱华吉等. 基于 Hadoop 的海量农业数据资源管理平台. 计算机工程, 2013,37(12):242-244
- [16] 量子数聚(北京)科技有限公司. 农业大数据应用云平台. <http://www.dataagri.com/agriculture/index.action>; 量子数聚(北京)科技有限公司, 2013

[17] 刘刚. Hadoop 应用开发技术详解. 北京:机械工业出版社,2014. 40-126

[18] Chute C, Ullman-Cullere M, Wood G . Some experiences and opportunities for big data in translational research. *GENETICS in Medicie* , 2013 ,15(10) :802-809

[19] Gu M, Li X P, Cao Y Y. Optical storage arrays: a perspective for future big data storage . *Light: Science & Applications* ,2014 ,58 :1-11

[20] Haluk D, Dursun D. Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. *Decision Support Systems* ,2013, 55 :412-421

[21] 宋长青,高明秀,周虎等. 农业大数据与现代农业发展探析. <http://www.cnki.net/kcms/detail/11.4922.S.20150104.1322.002.html>;中国农业信息. 2015

[22] 孙忠富,杜克明,尹首一. 物联网发展趋势与农业应用展望. 农业网络信息, 2010, (5) :5-8

[23] White T. Hadoop: The Definitive Guide. O’ Reilly-Media, 2009. 200-345

[24] 邢平平,施鹏飞,熊范纶. 数据挖掘技术在农业数据中的有效应用. 计算机工程与应用, 2001 ,37(2) :4-6

Analysis of the research on agricultural big data applications at home and abroad

Li Lingping^{* **} , Mao Kebiao^{**} , Fu Xiuli^{***} , Ma Ying^{**} , Wang Fang^{****} , Liu Qing^{**}

(^{*} College of Information Science & Technology, Beijing University of Chemical Technology, Beijing 100029)

(^{**} National Hulunber Grassland Ecosystem Observation and Research Station, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081)

(^{***} Information Engineering Institute, Beijing Institute of Petrochemical Technology, Beijing 102617)

(^{****} National-Local Joint Engineering Laboratory of Geo-spatial Information Technology, Hunan University of Science and Technology, Xiangtan 411201)

Abstract

In view of the problems of large scale, complicated structure and scarce data mining of agricultural data, the application of big data open source technologies to the agricultural data analysis system is investigated. Based on the characteristics of the agricultural data’s spatio-temporal attribution, the classical big data mining technologies of Hadoop, Storm and Spark are analyzed under the consideration of the features of agricultural data. Then, how to develop a big data system for agriculture is described summarily. Finally, the challenges and research problems to agricultural big data are concisely analyzed and summarized. It is pointed that it is necessary to do further research on the related theory and applications to boost and realize the scientific decision based on data to provide the secure guarantee for national food.

Key words: big data, agricultural big data, Hadoop, Storm, Spark