农业大数据综述

张浩然 李中良 邹腾飞 魏旭阳 杨国才

(西南大学计算机与信息科学学院 重庆 400715)

摘 要 云计算、物联网、大量社交网络的兴起使我们社会的数据种类和数量都呈井喷式增长,大数据时代已经到来。农业信息化是现代农业建设的重要内容,农业物联网等应用使农业产业发展中的应用日渐深入。在大数据背景下,大数据分析也为农业信息化提供了技术支持。对农业大数据的相关概念进行阐述,介绍了大数据的分析过程,以及对可应用于农业大数据的各项技术进行了介绍。最后简要分析了农业大数据未来发展所要面临的挑战。

关键词 农业大数据,物联网,云计算

中图法分类号 TP311.15 **文献标识码** A

Overview of Agriculture Big Data Research

ZHANG Hao-ran LI Zhong-liang ZOU Teng-fei WEI Xu-yang YANG Guo-cai

(School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract Data type and amount are growing in amazing speed which is caused by the emergency of new services such as cloud computing, internet of things and social network. The era of big data has come. Agricultural informatization is an important part of the construction of modern agriculture. The use of internet of agriculture things makes the applications of agriculture more and more deeply in the big data areas. Big data analytics also provides technical support for the agricultural informatization. This paper elaborated the related concept of agricultural big data, and introduced the progress of big data analytics. The key techniques of agricultural big data were described. Finally some new challenges in the future were summarized.

Keywords Agricultural big data, Internet of things, Cloud computing

1 农业大数据概念

想要得到农业大数据的定义,我们必然要从大数据和农业信息化两个方面来入手。

目前对大数据的准确定义尚有一些争论,这就导致大数据的定义有多种。全球知名的咨询公司麦肯锡认为:大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合[1]。维基百科给出的定义则是:大数据是利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集[2]。另外一种比较有代表性的定义是3V定义[3],即认为大数据需要满足3个特点:规模性(volume),多样性(variety)和高速性(velocity)。也有人在3V的基础上试图增加一个新的特性,即构造4V定义。但是关于第4个特性还没有统一的说法,以国际数据公司IDC为代表的认为大数据应该还具有的是价值性(value)[4],而以IBM为代表的则认为大数据还应具有真实性(veracity)[5]。笔者认为加入价值性较为准确。因为大数据本身存在较大的潜在价值,但由于大数据的数据量过大,其价值往往呈现稀疏性的特点。

农业信息化是一个动态概念,是指利用现代信息技术和信息系统为农业产、供、销及相关的管理和服务提供有效的信息支持,并提高农业的综合生产力,促进农业结构战略性调整和经营管理效率的总称[6]。简单地说,就是在农业领域充分利用信息技术的方法、手段和最新成果的过程。具体地讲,就是在农业生产、流通、消费以及农村经济、社会、技术等各环节全面运用现代信息技术和智能工具,实现农业生产经营、农产品营销、农产品消费的科学化和智能化过程[7]。

王儒敬教授曾提出,我国建成的涉农数据库数量很多,产生的各种数据量非常大,但是数据标准不统一且不规范,如何科学施肥、水肥调控以及品种选择与产业结构布局优化决策,并给出可靠且专业的决策结果的需求十分迫切[8]。解决这个问题就需要用到农业大数据技术。

农业大数据是指以大数据分析为基础,运用大数据的理念、技术及方法来处理农业生产销售整个链条中所产生的大量的数据,从中得到有用信息以指导农业生产经营、农产品流通和消费的过程。山东农业大学校长温孚江教授指出,农业大数据涉及到农业生产销售过程中的方方面面,是跨行业跨专业的数据处理过程[9]。农业大数据的实现过程也是农业信息化很重要的一个组成部分。大数据的应用与农业领域的相关科学研究相结合,可以为农业科研、政府决策、涉农企业发

本文受国家科技支撑计划项目(2012BAD35B08)资助。

张浩然(1990—),女,硕士生,主要研究方向为计算机软件与理论;**杨国才**(1962—),男,博士,教授,主要研究方向为农业信息技术和物联网应用(通信作者)。

展等提供新方法、新思路。王儒敬教授在文献[8]中也指出，建立农业信息化国家大数据中心，努力发展云计算、大数据挖掘等技术，是解决我国农业信息化发展瓶颈的重要手段。

2 农业大数据发展现状

随着物联网、云计算等技术的兴起和以 Facebook 等社交网站为代表的社交网络的发展，数据正以前所未有的速度在不断积累和增长。WinterCorp 的调查表明，目前的数据量正以每两年 3 倍的速率增加[10]，其增长速度远远超过摩尔定律增长速度，大数据的时代已经到来。

但是大数据这一概念引起足够重视并进一步发展则是在 2008 年以后。2008 年《Nature》杂志出版的专刊“Big Data: Science in the Petabyte Era”[11]从互联网技术、网络经济学、超级计算、环境科学、生物医药等多个方面介绍了大数据带来的挑战。《Science》杂志也在 2011 年推出专刊“Dealing with Data”[12]，围绕着科学研究中大数据的问题展开了讨论，说明了大数据对于科学研究的重要性。2011 年 6 月，麦肯锡研究院(MGI)发布研究报告《Big data: The next frontier for innovation, competition, and productivity》[13]，文中详尽地分析了大数据的影响、关键技术和应用领域等方面，指出大数据将带动未来生产力的发展和创新，并且能够拉动消费需求增长。2012 年欧洲信息学与数学研究协会会刊 ERCIM News 出版专刊“Big Data”[14]，讨论了在大数据背景下的数据管理、数据密集型研究的创新技术等问题，并简要描述了欧洲科研机构的部分研究工作和所取得的创新性进展。Gartner 在一年一度的技术成熟度曲线(见图 1)报告中指出，大数据已经进入发展高峰期。可以看出，未来信息技术的发展离不开大数据。IBM、ORACLE、谷歌、微软、亚马逊、Facebook、EMC、惠普等跨国巨头为了提高自身的竞争力也都在努力发展大数据技术[15]。中国科学院院士李国杰教授指出，在“数据大爆炸”时代，我们要做的是把数据“由厚变薄”，把数据去冗分类、去粗存精[16]。2012 年 6 月，中国计算机学会常务理事通过决议成立大数据专家委员会。2012 年 10 月，首个专门研究大数据应用和发展的学术咨询组织——中国通信学会大数据专家委员会成立，推动了我国大数据的研究与发展。2012 年 11 月，在北京举行了以“大数据共享与开放技术”为主题的“Hadoop 与大数据技术大会”，会上总结了多个当下的热点问题。2013 年 12 月，国内召开了中国大数据技术大会，会议主要讨论了大数据应用的技术手段和商业价值。大数据科学作为一个横跨信息科学、社会科学、网络科学、系统科学等诸多领域的新兴交叉学科，已经成为科技界的研究热点。

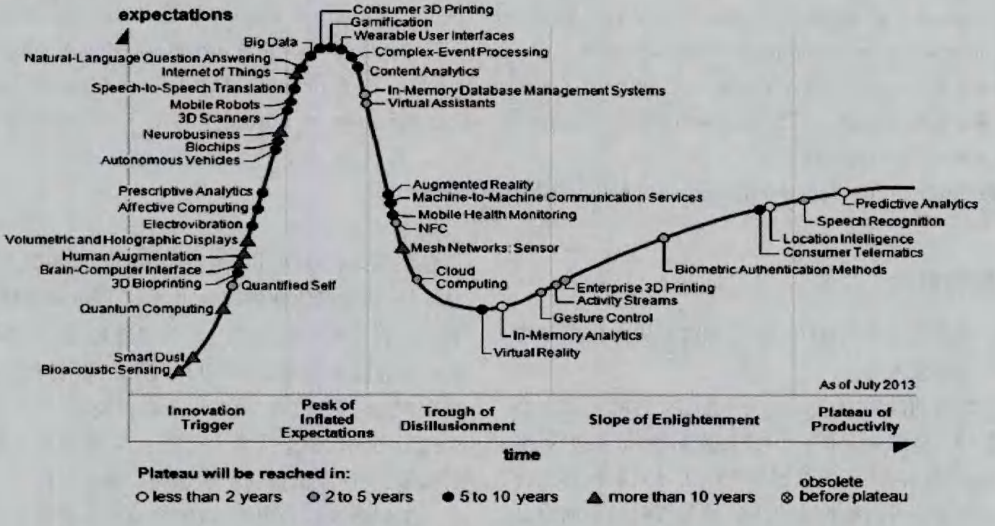

图 1 2013 年 Gartner 技术成熟度曲线[17]

不仅计算机技术的专业人员意识到了这一点，很多国家政府也已经意识到了利用大数据分析得到有用结果的重要性。

美国奥巴马政府发起的“大数据研究和发展倡议”[18]，被认为是继“信息高速公路计划”之后在信息科学领域的又一重要举措；英国政府预计在大数据和节能计算研究上投资 1.89 亿英镑；法国政府宣布投入 1150 万欧元，用于 7 个大数据市场研发项目；日本在新一轮 IT 振兴计划中，将发展大数据作为国家战略层面提出，重点关注大数据应用技术，如社会化媒体、新医疗、交通拥堵治理等公共领域的应用。中国的基础研究“大数据服务平台应用示范项目”正在启动，有关部门正在积极研究相关发展目标、发展原则、关键技术等方面的顶层设计。

目前，我国大数据也已经运用到医疗业、制造业、交通业、金融业等不同行业。随着农业信息化的发展及物联网在农业中的应用，农业大数据必将成为大数据应用的又一重点。农业大数据是大数据理念、技术和方法在农业领域的实践。农业大数据涉及耕地、育种、播种、施肥、收获、储运、农产品加工、销售等各个环节，是跨行业和跨专业的数据分析和挖掘。2013 年 6 月，国内第一个农业大数据产业技术创新战略联盟在山东农业大学成立。山东农业大学校长温孚江指出，目前在国内，大数据研究虽然刚刚起步，但“农业大数据”研究已经十分领先[9]。

众多企业也都瞄准了农业大数据的机遇。据纽约时报报道，2012 年土壤抽样分析服务商 Solum 获得 Andreessen Horowitz 等公司投入的 1700 万美元的资金。该公司致力于使用数据分析技术来确定化肥的投入量问题，通过对农业大数据的分析来帮助农民提高产出、降低成本。据纽约时报报道，2013 年 10 月跨国农业生物技术公司 Monsanto 以 9.3 亿美元巨资收购意外天气保险公司 Climate Corporation[19]，这

个公司通过分析自己掌握的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。2013 年 5 月召开的一次关于农业数据开放问题的国际论坛上，八国集团(G8)领导人基本讨论了取消农业数据限制的最佳途径，这也为我们带来了大量的原始研究数据。

农业大数据虽然还是一个比较新鲜的词汇，但是农业生产销售的过程中产生的绝不止过去的一些小数据了，其产生已经完全不受时间地点的限制，其产生方式发生了巨大的变化。随着智慧农业的提出，传感器的广泛应用大大促进了农业物联网技术的发展。传感器的职能便是搜集数据，各种声光压力温度传感器就好比人类的五官和触觉，使物联网系统自动产生大量半结构化和非结构化的数据。以物联网为代表产生的如此海量的数据就需要人们使用大数据技术去利用。而云计算的异军突起更是为这些海量数据提供了一个可靠的处理方式。

现阶段，虽然对农业大数据的研究很少，大数据的研究也刚刚起步，但是大量的科研人员已经意识到农业大数据的研究价值，也已经投入到农业大数据的分析、处理过程的优化中。如果能够将农业大数据利用好，不仅是人类农业历史上的一次伟大创举，也将是人类历史上的一次创举。

3 农业大数据的关键技术

3.1 农业大数据的采集

采集是农业大数据价值挖掘最重要的一环，其后的集成、分析、管理都构建于采集的基础上。大数据的来源方式主要有 RFID 射频数据、传感器数据、社交网络交互数据及移动互联网数据等方式。农业物联网的应用使得农业大数据的来源方式更侧重于 RFID 射频数据和传感器数据。

农业大数据采集主要包括农业数据传感体系、网络通信体系、传感适配体系、智能识别体系及软硬件资源接入系统，实现对结构化、半结构化、非结构化的海量数据的智能化识别、定位、跟踪、接入、传输、信号转换、监控、初步处理和管理等。农业大数据的采集主要与数据采集技术、传感器技术、信号处理技术等几个方面有关。

3.2 农业大数据的集成技术

上文提到大数据的来源方式十分广泛，这就导致数据类型极为复杂，农业大数据的数据类型不再是关系型数据库时期的结构化数据，而是转变为结构化数据、半结构化数据和非结构化数据的集合。要想处理农业大数据，就要首先对大数据采集阶段所得到的数据进行预处理，从中提取出关系和实体，经过关联和聚合，采用统一结构来存储这些数据。比如，在农业物联网中的各个传感器终端每天会产生大量数据，这样得到的数据格式可能不一致，或者其格式不是下一步数据分析所需要的，因此就要改变其格式，使其具有统一的结构以方便存储和分析处理。对于大数据，并不是全部有价值，有些数据并不是我们关心的内容，而且有一些数据可能是完全错误的干扰项，因此要对数据通过过滤“去噪”来提取有效数据。

在分布式数据集成中，如何屏蔽数据的分布性和异构性，实现数据高效、安全的交换和传输，并保持局部系统的自治性和目标系统的数据完整性，是需要考虑的主要问题[20]。

数据集成技术在传统数据库领域已有了比较成熟的研究。联邦数据库技术、分布式数据库技术、数据仓库技术都为数据集成应用提出了解决的办法。随着新数据源的出现，数据集成方法也在不断发展之中，相继出现了基于 XML 技术的数据集成[21,22]、基于 CORBA 的数据集成、基于 P2P 技术的数据集成[23]、基于 Web Services 技术的数据集成[24]等。总的来说，从数据集成模型来看，现有的数据集成方式可基本上分为 4 种类型[25]：基于物化或 ETL 方法的引擎(materialization or ETL engine)、基于联邦数据库或中间件方法的引擎(federation engine or mediator)、基于数据流方法的引擎(stream engine)及基于搜索引擎(search engine)的方法。

3.3 农业大数据的存储和处理技术

农业大数据的存储和处理涉及许多方面，传统的关系数据管理技术经过了长时间的发展，在扩展性方面遇到了巨大的障碍，无法胜任大数据分析的任务；以 MapReduce 为代表的非关系数据管理和分析技术以其良好的扩展性、容错性和大规模并行处理的优势，得到了很大的发展，并逐渐成为大数据处理和存储的主要技术手段[26]。现在已经有以 Hadoop、MapReduce、HBase 分布式框架为处理平台，构建“农业用户兴趣社交云”的系统[27]。具体来说，农业大数据的主要存储技术有分布式文件系统、分布式数据库等。本文主要介绍分布式文件系统。

3.3.1 分布式文件系统

文件系统是存储系统的重要组成部分，也是支撑大数据处理过程中上层应用的基础。在大数据处理过程中，通常采用分布式文件系统来应对海量数据存储和快速访问。分布式文件系统是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过计算机网络与节点相连。目前，众多处理海量数据的公司都有其自己的分布式文件系统，如 Google 公司的 GFS(Google File Sytem)[28,29]、淘宝网的 Taobao File System(TFS)[30]、IBM 的 GPFS(General Parallel File System)[31]、微软自行开发的 Cosmos[32]和阿里巴巴网站的 Alibaba Distributed File System(ADFS)[33]。另外，也有一些开源的分布式文件系统，包括 HDFS[34]、NFS[35,36]、pNFS[37]、xFS[38]、PIOUS[39]、PVFS[40]、Lustre[41]等。

Google 文件系统 GFS 构建在大量廉价服务器之上，具有较好的容错性和可扩展性，其基本架构[30]如图 2 所示。

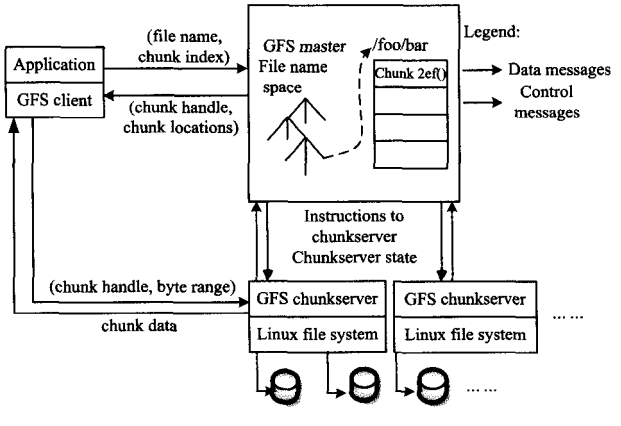

图 2 GFS 基本架构[30]

一个 GFS 系统由一个 master 和大量的 Chunk Server 构成，并被许多客户 Client 访问。Master 服务器维护文件系统所有的元数据，包括名字空间、访问控制信息、从文件到块的映射以及块的当前位置。文件被分成固定大小的块。每个块

由一个不变的、全局唯一的、64 位的 chunk-handle 标识，chunk-handle 是在块创建时由 Master 分配的。Chunk Server 将块存储在本地磁盘并可以读和写由 chunk-handle 和位区间指定的数据。出于可靠性考虑，每一个块被复制到多个 Chunk Server 上。多个并发 Client 可以同时访问 GFS 的文件数据。

TFS(Taobao File System)是一个高可扩展、高可用、高性能、面向互联网服务的分布式文件系统。TFS 为淘宝提供海量小文件存储，通常文件大小不超过 1M，满足了淘宝对小文件存储的需求，被广泛地应用在淘宝各项应用中。它采用了 HA 架构和平滑扩容，保证了整个文件系统的可用性和扩展性。同时扁平化的数据组织结构可将文件名映射到文件的物理地址，简化了文件的访问流程，在一定程度上为 TFS 提供了良好的读写性能。

3.3.2 分布式数据库

分布式数据库是计算机网络技术与传统的集中式数据库技术相结合的产物，除继承了传统的集中式数据库的特点外，还具有其自身的特点。分布式数据库系统是指物理上分布的，但逻辑上却是集中的数据库系统[42]。分布式数据库系统将物理上分散而管理和控制又需要集中的多个独立的数据库系统通过网络连接起来，构成了一个统一的数据库系统[43]。一般根据需要将数据分散地存放在网络中的各个站点上，每一个站点在逻辑上都是独立的，它们都拥有各自的数据库、局部数据库管理系统、CPU、内存、硬盘等其他设备，并通过计算机网络连接成一个逻辑整体，可以进行统一的控制和管理[44]。此外，还有一种分布式数据库系统在物理上和逻辑上都是分布的，也就是所谓的联邦式分布数据库系统。

3.3.3 云计算技术

云计算[45]是一种可以调用的虚拟化的资源池，这些资源池可以根据负载动态重新配置，以达到最优化使用的目的。用户和服务提供商事先约定服务等级协议，用户以用时付费模式使用服务。云计算从用户的角度来说，就是一台机器很难满足任务繁重的工作，我们通过网络把世界各个地方的计算机联合起来，将繁重的工作分解，通过各种技术轻松地解决问题。

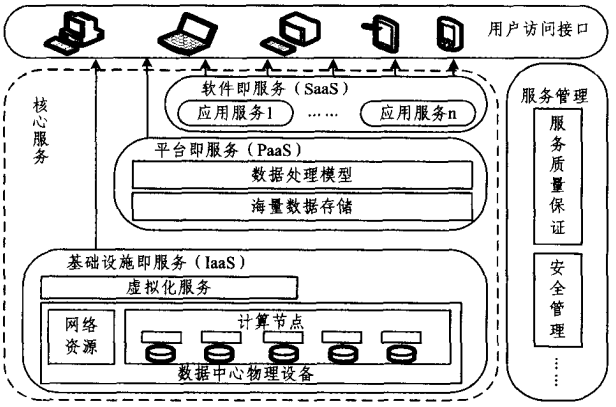

图 3 云计算体系结构[46]

云计算的体系结构可以分为 3 个层次：核心服务、服务管理、用户访问接口。核心服务层将硬件基础设施、应用程序、软件运行环境抽象成服务；服务管理层为核心服务提供支持，确保核心服务的可靠、可用与安全；用户访问接口层实现端到云的访问[46]。其结构如图 3 所示。

云计算的核心服务通常包括基础设施即服务(IaaS)、平台即服务(PaaS)和软件即服务(SaaS)这 3 种类型。我们通过表 1 来介绍这 3 种服务类型。

表 1 3 种服务类型比较

	IaaS	PaaS	SaaS
服务内容	提供基础设施部署服务	提供应用程序部署与管理服务	提供基于互联网的应用程序服务
服务对象	需要硬件资源的用户	程序开发者	企业和需要软件应用的用户
使用方式	使用者上传数据、程序代码和环境配置	使用者上传数据和程序代码	使用者上传数据
关键技术	数据中心管理技术、虚拟化技术等	海量数据处理技术、资源管理与调度技术等	Web 服务技术、互联网应用开发技术等
系统实例	Amazon EC2、Eucalyptus 等	Google App Engine 、MicrosoftAzure、Hadoop 等	Google Apps、Salesforce CRM 等

IaaS 提供硬件基础设施部署服务，简单地说，就是将实体或虚拟的计算、网络和存储等资源按需提供给用户。PaaS 提供应用程序部署与管理服务，是云计算应用程序的运行环境。SaaS 是指基于云计算基础平台所开发的应用程序，它通过 Internet 提供软件，用户不需要购买软件，而是向供应商租用基于 Web 的软件，来经营管理企业活动。

4 面临的挑战及解决方案

大数据技术目前处于起步阶段，主要面临数据的广泛异构性、数据的不完备性、数据处理的实时性、缺乏先验知识、隐私问题等挑战。农业大数据所面临的问题基本与大数据技术一致，但是农业大数据技术对数据的安全性或隐私问题不是那么敏感，挖掘农业大数据本身就是为了让农业生产销售过程中的人利用挖掘的信息来指导其行为，挖掘的对象是农业信息。因此，农业大数据主要面临的以下几个问题。

4.1 农业大数据的异构性

农业大数据的来源不同，有的来自物联网中的射频设备，有的来自农业信息化的网站，有的来自各种先进的移动终端等，这就导致数据类型从以结构化数据为主转向结构化、半结构化、非结构化三者的融合。怎样将这些异构的数据进行统一的存储，怎样运用分析手段来统一分析这些异构的数据，将是值得深究的问题。

4.2 农业大数据的实时性

随着时间的流逝，数据中所蕴含的知识价值往往也在衰减，因此实时性也是农业大数据分析过程中必须考虑的问题。尤其是在与天气、环境状况相关的数据分析的方面，大数据的分析不及时可能会导致农业生产灾害的发生。由于数据量的密集，能否在所能忍受的时间内完成指定工作也成为衡量大数据分析的方法。解决这个问题的一个途径是采用流处理与批处理相结合的数据处理模式。

4.3 农业大数据的挖掘能力

农业大数据的异构性导致农业数据的类型多种多样，并且由于数据集过大，使得传统的数据挖掘、机器学习等算法不再适用于对农业大数据的挖掘。因为现有算法往往是用于常驻的小数据集，用于大数据集可能会导致效率过差甚至根本无法使用。一方面，云计算是农业大数据处理的主要工具，传统的算法并不能直接适用于农业大数据的处理平台；另一方面，农业大数据的应用有一个很重要的特点就是实时性，算法的准确率不再是主要指标。我们需要在处理的实时性和准确

率之间取得平衡。为了解决这个问题，我们可以对处理小数据的算法进行改进，使其既可以适用于大数据的处理平台，又可以兼顾准确率和实时性，这也将会是未来一段时间内科学研究的热点。

结束语 农业大数据是目前研究的热点，也是大数据应用的一个重要方面。正确认识农业大数据的有关内容具有重要意义。本文首先由大数据和农业信息化的相关概念引出了农业大数据的概念；其次，介绍了目前农业大数据的发展现状，其中既有国内现状，也有国外现状；再次，讨论了农业大数据技术的关键技术，并进行了详细的介绍；最后，提出了农业大数据所面临的挑战和研究难题。

参 考 文 献

[1] James M, Michael C, Brad B. Big data: The next frontier for innovation, competition, and productivity[J]. The McKinsey Global Institute, 2011(5)

[2] Big data[EB/OL]. [2014-03-19]. http://en.wikipedia.org/wiki/Big_data

[3] Grobelnik, Marko. Big Data Tutorial[EB/OL]. [2014-03-19]. http://videolectures.net/eswc2012_grobelnik_big-data

[4] Barwick H. The "four Vs" of Big Data. Implementing Information Infrastructure Symposium[EB/OL]. [2014-03-19]. http://www.Computerworld.com.au/article/396198/iiis_four_vs_big_data

[5] IBM. What is big data? [EB/OL]. [2014-03-19]. http://www901.ibm.com/software/data/bigdata

[6] 曾晓娟，丁超英，文红霞，等. 探讨实现农业信息化的有效途径[J]. 农业图书情报学刊，2004(2)：34-37

[7] 钱学军. 中国农业现代化进程中的农业信息化研究[D]. 北京：中国农业大学，2005

[8] 王儒敬. 我国农业信息化发展的瓶颈与应对策略思考[J]. 中国科学院院刊，2013，28(3)：337-343

[9] 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，2013，11：2

[10] WinterCorp. The Large Scale Data Management Experts[EB/OL]. http://www.wintercorp.com

[11] Nature. Big Data[EB/OL]. [2014-03-19]. http://www.nature.com/news/specials/bigdata/index.html

[12] Science. Special Online Collection: Dealing with Data[EB/OL]. [2014-03-19]. http://www.sciencemag.org/site/special/data

[13] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity[R]. McKinsey Global Institute, 2011：1-137

[14] ERCIM News. Big Data[EB/OL]. [2014-03-19]. http://ercim-news.ercim.eu/en89

[15] 李国杰. 大数据研究的科学价值[J]. 中国计算机学会通信，2012，8(9)：8-15

[16] 甘晓，李国杰. 大数据成为信息科技新关注点[J]. 中国科学报，2012

[17] Jackie Fenn, Hung LeHong. Emerging Technologies Hype Cycle for 2013: Redefining the Relationship[EB/OL]. [2014-03-19]. http://my.gartner.com/portal/server.pt? open=512&objID=202&mode=2&PageID=5553&showOriginalFeature=y&resId=2546719&fnl=search&srcId=1-3478922244

[18] Big Data Across the Federal Government[EB/OL]. [2014-03-19]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheetfmall.pdf

[19] Monsanto Acquires The Climate Corporation[EB/OL]. [2014-03-19]. http://www.monsanto.com/features/pages/monsanto-acquires-the-climate-corporation.aspx

[20] 靳强勇，李冠宇，张俊. 异构数据集成技术的发展和现状[J]. 计算机工程与应用，2002，11(7)：112-114

[21] Lehti P, Fankhauser P. XML data integration with OWL: Experiences and challenges[C]// Proceedings 2004 International Symposium on Applications and the Internet. IEEE, 2004：160-167

[22] Passi K, Lane L, Madria S, et al. A model for XML Schema integration[M]// E-Commerce and Web Technologies. Springer Berlin Heidelberg, 2002：193-202

[23] Ng W S, Ooi B C, Tan K L, et al. PeerDB: A P2P-based system for distributed data sharing[C]// Proceedings 19th International Conference on Data Engineering, 2003. IEEE, 2003：633-644

[24] 张实. 基于 Web Services 的异构数据源共享[D]. 长沙：国防科技大学，2003：12-17

[25] Haas L. Integrating Extremely Large Data is Extremely Challenging[C/OL]//[2014-03-19]. Proc of XLDB Asia 2012. http://idke.ruc.edu.cn/xldb/www.xldb-asia.org/program.html

[26] 覃雄派，王会举，杜小勇，等. 大数据分析——RDBMS 与 MapReduce 的竞争与共生[J]. 软件学报，2012，23(1)：32-45

[27] 郭平，刘波，沈岳. 农业云大数据自组织推送关键技术综述[J]. 软件，2013，34(3)：1-6

[28] Ghemawat S, Gobioff H, Leung S. The Google file system[C]// the ACM Symposium on Operating Systems Principles. 2003：29-43

[29] McKusick M K, Quinlan S. GFS: Evolution on Fast-forward[J]. ACM Queue, 2009，7(7)：10-20

[30] Chucai. TFS Introduction[EB/OL]. [2014-03-19]. http://rdc.taobao.com/blog/cs

[31] Jones T, Koniges A, Yates R K. Performance of the IBM General Parallel File System[C]// the 14th International Symposium on Parallel and Distributed Processing, 2000：673-681

[32] Yu W, Vetter J S, Canon R S. Exploiting Lustre File Joining for Effective Collective IO[C]// the Seventh IEEE International Symposium on Cluster Computing and the Grid(CCGrid2007). 2007：267-274

[33] Ye W, et. al. The Software Revolution during Internet Ear: The Design of SaaS Infrastructure[M]. Beijing: Publishing House of Electronics Industry, 2010

[34] Konstantin S, Hairong K, Sanjay R, et al. The Hadoop Distributed File System[C]// the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies(MSST). 2010：1-10

[35] Osadzinski A. Network File System (NFS) [J]. Computer Standards and Interfaces, 1988，8(1)：45-48

[36] Anderson T E, Dahlin M D, Neefe J M, et al. Serverless Network File Systems[J]. ACM Transaction on Computer Systems, 1996，14(1)：41-79

[37] The Internet Engineering Steering Group. pNFS File System [EB/OL]. [2014-03-19]. http://www.pnfscom

[38] Randolph Y W, Thomas E A. xFS: A Wide Area Mass Storage File System[C]// the 1933 Workshop on Workstation Operating Systems. 1933：71-78

[39] Steven A M, Sunderam V S. PIOUS: A Parallel File Sytstem for Distributed Computing Environments[C]// the Scalable High-Performance Computing Conference. 1994：71-78

[40] Carns P H, Ligon III W B, Ross R B, et al. PVFS: A Parallel

Virtual File System for Linux Clusters[C]// the 4th Annual Linux Showcase and Conference. 2000:317-327
[41] Sun Microsystems, Inc.. Lustre™ 1.6 Operations Manual[EB/OL]. [2014-03-19]. http://wiki.lustre.org/images/alae/820-3681.pdf
[42] 邵佩英. 分布式数据库系统及其应用[M]. 北京:科学出版社,2005,7
[43] 李川. 分布式数据库查询策略优化的研究[D]. 西安:西安电子科技大学,2012
[44] 萨师煊,王珊. 数据库系统概论(第三版)[M]. 北京:高等教育出版社,2000
[45] Vaquero L, Rodero-Marino L, Caceres J, et al. A break in the clouds: towards a cloud definition[J]. SIGCOMM Computer Communication Review,2009,39(1):50-55
[46] 罗军舟,金嘉晖,宋爱波. 云计算体系架构与关键技术[J]. 通信学报,2011,32(7)

(上接第 367 页)

nal functions),内部函数和内嵌函数的区别在于前者没有链接,所以只能被 GCC 内部调用,而不能被用户手工调用。

3. SIMD 循环的向量化

SIMD 循环在向量化阶段的处理主要涉及如下 3 个函数:

1)入口函数 vectorize_loops()

该函数中对循环的向量化可能性进行分析前,先判断 cfun 所带循环数目。若数目不超过 1,函数将返回,不执行向量化。但在返回之前,若发现 cfun 有 SIMD 循环,则调用 adjust_simduid_builtins(),该函数为 IFN_GOMP_SIMD_VF, IFN_GOMP_SIMD_LANE 和 IFN_GOMP_SIMD_LAST_LANE,3 个内部调用函数分别创建一个整型树节点,替换 cfun 中调用该内部函数的语句。

而若 cfun 的循环个数超过 1,则调用 note_simd_array_uses(),构建从 simd 数组到 simduid 的映射。

在分析循环阶段,全部分析完后,执行代码的向量化转换。在循环已被向量化后,为了使得该循环能被顺利展开,新建 simduid_to_vf 类型的变量 simduid_to_vf_data,并给其成员变量赋值,如成员 simduid 的值来自已向量化循环的 simduid,成员 vf 的值来自已向量化循环的向量化因子。

在结束阶段,完成各数据结构的销毁后,仍调用 adjust_simduid_builtins()来转换内部调用函数语句,之后,对 simd 数组(simd array)遍历,对其中的每个元素取其 simduid,再到哈希表 simduid_to_vf_htab 中去匹配,找到对应元素并取出 vf 值赋给最终的向量化因子。

2)函数 vect_analyze_data_refs()

若存在 SIMD 循环,则可确定存在支持 simd lane 的访问的可能性,然后还需要对数据引用进行相关性分析,确定 simd lane 是否能真的访问。若不能访问,则以向量化分析失败返回。

3)函数 vectorizable_call()

该函数用来向量化调用语句(即调用其他函数的语句)。分析当前语句 stmt 调用的函数发现其如果不能被向量化,则还会检查一种特殊情况,即若 stmt 调用的是 IFN_GOMP_SIMD_LANE 且 stmt 的循环是 SIMD 循环,则仍可向量化。在后面实现向量化的过程中,IFN_GOMP_SIMD_LANE 被处理成$\{0,1,2,\cdots,v_{f-1}\}$向量。

4)vect_estimate_min_profitable_iters()

该函数用于判断向量化收益,在向量收益小于标量收益时,不能进行向量化。此时,若有 SIMD 循环,仅输出信息以说明该 SIMD 循环不可向量化。

结束语 从以上分析可以看出,在当前编译器的实现中,simd 的 3 种结构都能被前端识别,且在 OpenMP 下降和扩展及自动向量化阶段都有处理。GCC 4.9 的新增遍 pass_omp_simd_clone 专门为声明为 declare simd 的函数创建向量化版本。在已有的如 function、loop 等数据结构中新增域来记录 simd 编译指导的相关信息,还会临时创建 simduid_to_vf 等数据结构辅助相关功能的实现。在编译各阶段间,全局变量 cfun 作为主要数据结构来传递 simd 编译指导的信息。但是,当前实现中仍存在较多不足,主要表现在以下两个方面:

1)SIMD 循环的识别受限。当前只能识别带 private、firstprivate 或 reduction 3 种从句的 SIMD 循环;

2)向量化阶段功能实现不完善。主要是指两种 simd 编译指导下的 SIMD 循环在向量化阶段,仅是将内部函数转换成整型树节点、构建一些映射、部分影响向量化条件等,并未真正实现直接指导向量化或转换 SIMD 循环。而且在循环转换成向量化代码的过程中,并未对 SIMD 函数或 SIMD 循环有任何特殊处理。

在下一步工作中,将针对其不完善的部分进行改进,实现用 SIMD 编译指导提升 GCC 自动向量化的能力。

参 考 文 献

[1] Khronos OpenCL Working Group. The OpenCL Speci? cation [R]. [2009]. http://www.khronos.org/registry/cl/
[2] 姚远,赵荣彩. 基于编译指示的向量化方法[J]. Computer Engineering,2012,38(12):272—275
[3] Tian X, Saito H, Preis S V. Compiling C/C++SIMD Extensions for Function and Loop Vectorization on Multicore-SIMD Processors[C]// Multicore and GPU Programming Models, Languages and Compilers Workshop. 2012:2349-2358
[4] Klemm M, et al. Extending OpenMP* with vector constructs for modern multicore SIMD architectures[C]// OpenMP in a Heterogeneous World,2012:59-72
[5] 黄娟娟,李春江,徐颖. GCC 中自动向量化代价模型剖析[C]// 第 17 届计算机工程与工艺年会暨第三届微处理器技术论坛论文集. 长沙:国防科技大学出版社,2013:259-268
[6] OpenMP Architecture Review Board: OpenMP Application Program Interface[M]. Version 4.0(July 2013)
[7] Free Software Foundation Inc. GCC 4.9 Release Series http://gcc.gnu.org/gcc-4.9/
[8] Novillo D. Design and Implementation of Tree SSA[C]// Proceedings of the 2004 GCC Summit. Ottawa, Canada, 2004
[9] 黄娟娟. 多线程多 SIMD 自动向量化技术研究[D]. 长沙:国防科学技术大学,2013
[10] 辛乃军,陈旭灿,等. 基于 GCC 的高性能 DSP Matrix 向量指令集扩展[J]. 计算机工程与科学,2012,34(1):58-63
[11] 徐颖. 编译器自动向量化效能评估与分析[D]. 长沙:国防科学技术大学,2012
[12] 李春江,黄娟娟,徐颖. 典型编译器自动向量化效果评估与分析[J]. 计算机科学,2013(4):41-46