



大数据技术的应用现状与展望

廖建新

(北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘要:梳理了大数据研究的4项关键技术:“数据的采集、预处理与存储”、“数据的分析与挖掘”、“数据的隐私保护”、“数据中心体系结构”,挖掘和展示了国内外大数据研究的热点,以期对该领域的研究有一个系统而全面的认识。从处理平台、分析产品、标准化3个方面介绍了大数据技术应用的现状,并列举了大数据现有的产品应用及各行业应用案例。最后,从大数据的分析策略、深度学习、隐私保护和数据质量几个方面揭示了大数据研究存在的挑战和机遇,以期为后续大数据技术应用的发展提供参考。

关键词:大数据;数据分析;数据挖掘;非结构化数据;隐私保护;并行计算;社交网络数据

doi: 10.11959/j.issn.1000-0801.2015189

Big Data Technology: Current Applications and Prospects

Liao Jianxin

(State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The research includes computer science, information science, management science, psychology, etc. Four key technologies in big data research, data collection, preprocessing and storage, data analysis and mining, data privacy protection, data center system structure were discussed and the hot research topics from domestic and overseas research groups were demonstrated in order to have comprehensive and systematic understanding of big data. The application status of big data was introduced from three perspectives, application platforms, relevant products and standardization. Meanwhile, description of current business in big data and other field was presented. Eventually, the opportunities and challenges in big data research were revealed through big data analysis strategies, privacy protection and data quality, which provide the guideline for future research.

Key words: big data, data analysis, data mining, unstructured data, privacy protection, parallel computing, social network data

收稿日期:2015-06-03;修回日期:2015-07-14

基金项目:国家重点基础研究发展计划(“973”计划)基金资助项目(No.2013CB329102),国家自然科学基金资助项目(No.61471063, No.61372120, No.61271019, No.61101119, No.61121001),教育部科学技术研究重点(重大)项目(No.MCM20130310),北京高等学校青年英才计划基金资助项目(No.YETP0473),高等学校博士学科点专项科研基金资助课题(No.20100005110008)

Foundation Items: The National Basic Research Program of China (973 Program) (No.2013CB329102), The National Natural Science Foundation of China (No.61471063, No.61372120, No.61271019, No.61101119, No.61121001), The Important Project of Ministry of Education (No.MCM20130310), Beijing Higher Education Young Elite Teacher Project (No.YETP0473), Specialized Research Fund for the Doctoral Program of Higher Education (No.20100005110008)

论文引用格式:廖建新. 大数据技术的应用现状与展望. 电信科学, 2015189

Liao J X. Big data technology: current applications and prospects. Telecommunications Science, 2015189



1 引言

自2012年兴起的“大数据”潮流,让“big data”这个IT圈子里的名词一下风靡了各个行业。虽然大数据的重要性得到了大家的一致认同,但是对大数据的理解却众说纷纭,以至于除了IT以及相关行业以外,各种如餐饮业、房地产业、金融业等行业都声称自己采用了“大数据”技术,并迫不及待地宣布了自己的“大数据”战略。就在这种全民热炒的时代,更需要科技工作者保持冷静的头脑。大数据是一个抽象的概念,除去数据量庞大这一特征,大数据还有一些其他的特征,这些特征决定了大数据与“海量数据”和“非常大的数据”这些概念之间的不同。

高德纳分析员 Doug Laney 曾于2001年在一次演讲中指出,数据增长有3个方向的挑战:数量(volume),即数据多少;速度(velocity),即资料输入、输出的速度;种类(variety),即多样性,这3方面的特征即大数据最先提出的3V模型。2011年,在国际数据公司(IDC)发布的报告中^[1],大数据被定义为:“大数据技术描述了新一代的技术和架构体系,通过高速采集、发现或分析,提取各种各样的大量数据的经济价值。”大数据的特点可以总结为4个V,即volume(体量浩大)、variety(模态繁多)、velocity(生成快速)和value(价值巨大但密度很低)。这种4V定义得到了更广泛的认同,这种定义指出了大数据最为核心的问题,就是如何从规模巨大、种类繁多、生成快速的数据集中挖掘价值。在那以后,业界对大数据的解读越来越全面,相继把大数据的基本特征扩展到了5V、7V、甚至11V特征,扩充了veracity(真实性)、validity(有效性)、variability(易变性)、viability(存活性)、volatility(波动性)、visibility(可见性)、visualization(可视性)等新维度。

除了上述主流的定义,还有人使用3S或者3I描述大数据的特征。3S指的是:大小(size)、速度(speed)和结构(structure);3I指的是:定义不明确的(ill-defined)、令人生畏的(intimidating)、即时的(immediate)。因此,为了保证大数据的可控性,需要缩短数据搜集到获得数据洞察之间的时间,使得大数据成为真正的即时大数据。就大数据究竟该如何定义,工业界和学术界还有不少讨论^[2-4]。但是,大数据的关键并不在于如何定义或如何界定,而是如何提取数据的价值,如何利用数据,如何将“一堆数据”变为“大数据”。

随着数据为王的大数据时代的到来,产业界需求与关

注点发生了重大转变:企业关注的重点转向数据,计算机行业正在转变为真正的信息行业,从追求计算速度转变为关注大数据的处理能力。大数据处理的兴起也改变了云计算的发展方向,使其进入以分析即服务(analytics-as-a-service, AaaS)为主要标志的Cloud 2.0时代。大数据还引起了科技界对科学研究方法论的重新审视,正在引发科学研究思维与方法的一场革命。参考文献[2]对大数据引发的思维变革进行了总结。在大数据时代,通过挖掘和分析处理,大数据可以为人的决策带来参考答案,但是并不能取代人的思考。正是人的思维,促使众多利用大数据的应用出现,因此,大数据更像是人的大脑功能的延伸和扩展,而不是大脑的替代品。如果能有效地组织和使用大数据,将对社会经济和科学研究发展产生巨大的推动作用,同时也孕育着前所未有的机遇。

本文将在系统分析大数据研究关键技术的基础上,梳理国内外大数据现有的技术产品,总结分析企业大数据的策略和商用应用案例。最后,在分析大数据时代面临挑战的基础上,尝试展望大数据研究可能存在的机遇,提出正确应对大数据挑战的观点,以期对该领域的研究有一个系统而全面的认识,为后续大数据研究方向的发展提供参考。

2 大数据的关键技术

大数据技术,从本质上来说就是从类型各异、内容庞大的数据中快速获得有价值信息的技术。目前,随着大数据领域被广泛关注,大量新的技术已经开始涌现出来,而这些技术将成为或者已经成为大数据采集、存储、分析、表现的重要工具。

大数据处理的关键技术主要包括:数据采集、数据预处理(数据清理、数据集成、数据变换等)、海量数据存储、数据分析及挖掘、数据的呈现与应用(数据可视化、数据安全与隐私等)。图1展现了如何将大量的数据最终转化为有价值信息的一般步骤,基本囊括了大数据领域的关键技术。正如图1所示,数据经过一系列的加工和处理,最终以有价值的信息的形式到达用户手中。需要特别注意的是,在数据分析中,云技术与传统方法之间进行联合,使得一些传统的数据分析方法能够成功地运用到大数据的范畴中来。

下面将对大数据领域中的核心技术进行简要的介绍与阐述。

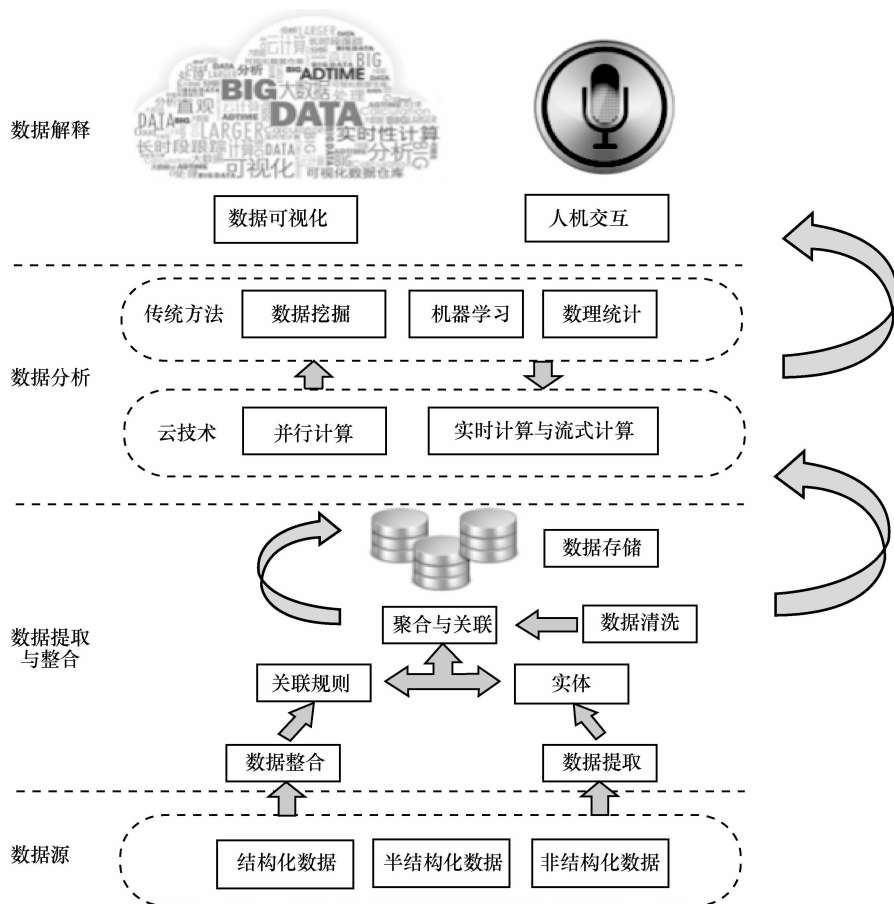


图1 大数据领域关键技术

2.1 数据的采集、预处理与存储

大数据技术的第一步也是关键一步便是数据的采集。从传感器网络、社交媒体等数据源中获取结构化、半结构化和非结构化数据,并将数据主体进行预处理与存储是大数据环境下处理与分析数据的基础。大量数据在这一环节中被清理、集成和变换,成为“合格”的数据,从而被进一步分析与利用。

2.1.1 数据的采集技术

在大数据这一范畴下,数据是指通过传感器网络、无线射频数据、社交网络数据及移动互联网数据等方式获得的结构化、半结构化(或称为弱结构化)及非结构化的海量数据,是大数据分析挖掘的根本。

大数据采集体系如图2所示。由图2可知,一般可将大数据采集体系分为智能感知层和基础支撑层。智能感知层主要包括数据传感网络、无线射频网络、智能识别网络(二维码)及资源接入系统,实现对非结构化、半结构化、结构化的海量数据的智能化识别、定位、接入、传输、监控、初步处理和管理等。而基础支撑层主要提供大数据服务平台

所需的物理介质,如数据库资源、物理传输资源、物联网资源等。

2.1.2 数据预处理技术

数据预处理技术主要包括数据清理、数据集成以及数据变换^[1]。数据清理可以去掉噪声数据以及异常数据,纠正数据中的不一致。而数据集成可以将来自不同数据源的数据合并成一致的数据存储,如数据仓库。数据变换可以改进涉及距离度量的挖掘算法的精度和有效性,将不同度量下的数据归一化^[2],使得数据的应用比较有意义。这些数据预处理技术在数据分析之前使用,可以大大提高数据分析的质量,提高分析的速度与准确性。

2.1.3 海量数据存储技术

在大数据的环境下,为保证高可用、高可靠和经济性,往往采用分布式存储的方式来存储数据,采用冗余存储的方式来保证存储数据的可靠性,即为同一份数据存储多个副本。海量存储的关键技术包括并行存储体系架构、高性能对象存储技术、并行I/O访问技术、海量存储系统高可用技术、嵌入式64 bit存储操作系统、数据保护与安全体

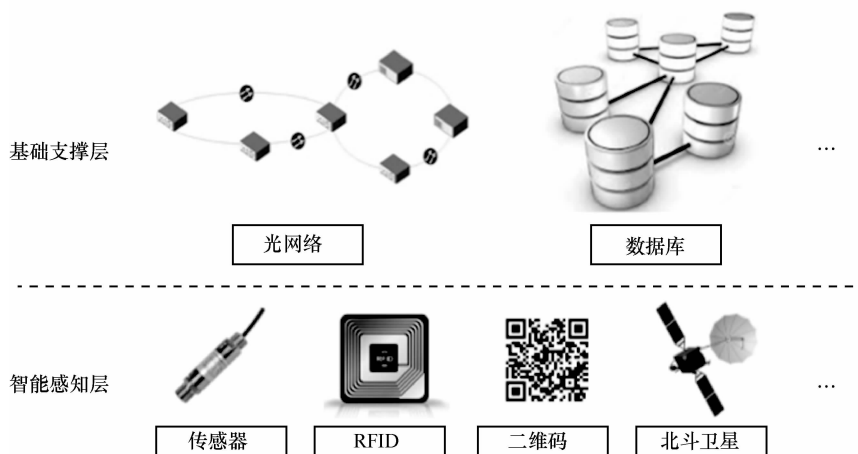


图2 大数据采集体系

系、绿色存储等。

广泛适用的分布式文件存储系统的设计思想不同于传统的文件系统,这类系统往往是针对大规模数据处理而特殊设计的^[13,14]。它们虽然运行在一些非常廉价的普通硬件之上,但是却可以提供容错的功能,从而给用户总体性能较高的服务。一个分布式集群一般由一个主服务器和大量的块服务器构成,许多用户可以同时访问^[15]。主服务器包含了所有的元数据,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。主服务器还控制系统活动范围,定期通过心跳消息与每一个块服务器通信,并收集它们的状态信息。

2.2 数据的分析与挖掘

为了实现抽取大数据中价值的信息,必须对经过预处理之后的数据进行分析与挖掘。目前针对海量数据的分析方法主要有并行计算、实时计算与流式计算。

2.2.1 并行计算

并行计算是指同时使用多个计算资源完成运算。其基本思想是将问题进行分解,由若干个独立的处理器完成各自的任务,以达到协同处理的目的^[16]。在大数据时代下,串行的处理方式难以满足人们的需求,现在主要采用并行计算的方式。目前在大数据环境下所提出的并行计算往往是指任务级别的并行计算,而指令或进程级别的并行计算模型往往具有更强大的处理大规模数据的能力,只是现有的GPU(graphics processing unit, 图像处理器)编程模型还不够完善。大量挖掘算法开始针对并行架构进行调整,使其能够利用并行的优势,对大规模的数据进行更好的处理。Luo D等提出的非平凡策略,可以将一系列的数据挖掘问题进行并行处理,其中包括支持向量机(support vector

machine, SVM)、非负最小二乘问题等。由此得到的算法,可以通过目前流行的并行处理结构进行实现,数据分析效果获得了显著的提升^[17]。Gao F等提出了一种新的近似算法,将基于核函数的数据挖掘算法的应用范围推广到了海量数据集上。该算法大幅度降低了计算核矩阵时的计算开销和内存使用,但却没有对结果的精确度产生很大的影响。此外,并行结构下该算法的实现也被提出,真实数据集上的测试结果显示,提出的算法可以大量节省时间和空间上的消耗^[18]。Shim K在并行结构的框架下,讨论了如何将传统的数理统计方法和并行结构相结合,以便进行大数据分析^[19]。

2.2.2 实时计算与流式计算

传统的数据流处理^[20,21]受到数据采集速度和内存容量等因素的限制,往往只能处理小规模的数据流。但随着数据采集技术和数据传输技术的进一步发展,使得短时间内积累大量的历史数据成为可能。与此同时,目前大数据环境下对数据流处理的要求不断提升,使得历史数据规模的增加也成为必然。

目前关于数据流的处理研究主要可以分为两类:集中式和分布式。在集中式环境下,数据流的计算受到存储资源,特别是内存容量的限制,主要通过概要数据、准入控制和QoS降价等方法,以牺牲服务质量为代价最终实现伸缩性。而在分布式环境下,针对由多个算子组成的数据流处理网络,主要是通过平衡在多个节点上的算子分布来最终实现伸缩性。

实时计算同样为实时数据的挖掘奠定了基础,此时数据挖掘有了特定的对象——数据流。为了有效地处理数据流,需要建立新的数据结构,并将新的数据结构与传统的挖掘算法相结合^[22]。因为并不存在无限大的空间存储数据

流,所以需要在正确性和存储空间之间进行平衡。常用的数据结构和挖掘技术有滑动窗口、多分辨率方法、梗概、随机算法等。

2.2.3 深度学习技术

深度学习概念的出现,源自于复杂数据结构处理以及复杂特征提取任务之中遇到的与人工智能相关的问题,这些问题普遍需要高阶抽象概念进行表述,具有非线性、语意性等特征。深度学习网络结构通常由多层非线性运算网络组成,每一层的输出作为下一层的输入,能够从海量数据中提取并学习到有效的复杂特征,进而用于数据的检索、分类、回归等问题之中^[24]。

深度学习概念起源于人工神经网络的研究,而人工神经网络结构在复杂高维数据处理问题上有许多不足,并且有着泛化能力差、训练速度慢等问题。2006 年 Hinton 等^[25]提出的用于训练深信度网络(deep belief network, DBN)的无监督学习算法,将之前用于训练网络的全局优化算法拆分成若干个子任务逐次执行,在保证算法性能的前提下极大地提高了网络训练效率。在此之后,大量研究工作被引入,深度学习理论被不断丰富,将深度学习技术应用于信号处理、图像识别、语音处理等领域的尝试与研究也日益增加。

深度学习结构试图找到数据内部的结构特征,发觉数据间的真实关联规则。在处理实际任务的过程中,数据的表现形式、关系模式是多种多样的;与之对应,深度学习结构也发展出多种结构,以应对不同场景下的数据处理需求。目前有卷积神经网络(convolutional neural network, CNN)^[26]和 DBN 两种主流的深度学习结构。在自然语言处理和信息检索领域,已经有了大量的 DBN 应用案例^[27]。

2.3 数据的隐私保护

在大数据时代,云端可以每时每刻对用户的信息进行采集,使每一个用户成为“透明人”,因此当前亟需针对大数据面临的用户隐私保护、数据内容可信验证、访问控制等安全挑战,提出相应的解决方案。

2.3.1 数据发布匿名保护技术

对于大数据中的结构化数据(或称关系数据)而言,数据发布匿名保护是实现其隐私保护的核心关键技术 with 基本手段。以典型的 k 匿名方案为例,通过元组泛化、抑制等数据处理,将准标识符分组。每个分组中的准标识符相同且至少包含 k 个元组,因而每个元组至少与 $k-1$ 个其他元组不可区分。实现方法包括基于裁剪算法的方案以及基于数据置换的方案等。在大数据场景中,数据发布匿名保护

问题较之更为复杂:攻击者可以从多种渠道获得数据,而不仅仅是同一发布源。

2.3.2 社交网络匿名保护技术

由于社交网络具有图结构特征,其匿名保护技术与结构化数据有很大不同。

社交网络中的典型匿名保护需求为:用户标识匿名与属性匿名(又称点匿名),在数据发布时隐藏了用户的标识与属性信息;用户间关系匿名(又称边匿名),在数据发布时隐藏用户间的关系。目前的边匿名方案有基于边的随机增删方案、基于节点聚集的匿名方案、基于基因算法的实现方案、基于模拟退火算法的实现方案以及先填充再分割超级节点的方案等。

2.3.3 数据溯源技术

数据溯源(data provenance)也被译成“数据世系”。其基本出发点是帮助人们确定数据仓库中各项数据的来源,例如,了解它们是由哪些表中的哪些数据项运算而成的,据此可以方便地验算结果的正确性,或者以极小的代价进行数据更新。数据溯源的基本方法是标记法,如通过对数据进行标记来记录数据在数据仓库中的查询与传播历史。后来概念进一步被细化为 why 和 where 两类,分别侧重数据的计算方法和出处。除数据库以外,数据溯源还包括 XML 数据、流数据与不确定数据的溯源技术。数据溯源技术也可用于文件的溯源与恢复。例如,通过扩展 Linux 内核与文件系统,创建一个数据起源存储原型系统,可以自动搜集起源数据。未来数据溯源技术将在信息安全领域发挥重要作用。

2.4 数据中心体系结构

网络数据中心作为大数据服务的天然载体,是实现大数据分析的基础设施。传统数据中心的拓扑采用树型分层结构,但性能不佳且可扩展性差。为此,Al-Fares 等提出了一种基于“胖树”的类树型结构拓扑,其主要目的是在网络端节点实现更高的聚合带宽。通过增加一定的布线复杂度来连接形成一个胖树型网络,端交换机用来连接服务器。在全负载最坏的情况下,这个结构仍然可以实现约 87% 的聚合带宽。此外,美国加州大学的 Guo 等人将并行计算的一些思想引入数据中心结构的设计中,提出了 DCell^[28]。DCell 是一种递归构建的数据中心,它使用性能强大的商业级 PC 和低端的交换机,拓展性相对于节点的度具有双倍指数的增长关系。清华大学的李丹提出了一种使用双网卡 PC 机和低端交换机来搭建数据中心的方案 FiConn^[29]。



在 FiConn 这一拓扑中,链路被分为若干级别,通过一种自适应的机制来实现平衡负载和提高网络吞吐量的目的。

3 大数据技术应用现状

大数据已经渗透到每一个行业和业务领域,逐渐成为重要的生产因素,对大数据的分析处理正是业界的流行趋势。

3.1 大数据处理平台

大数据时代产生了许多更为复杂的新数据类型,企业需要更加强大的数据处理平台来分析这些数据。国外的如 Google、Amazon、IBM、Microsoft 等企业,都已经意识到大数据对企业长远发展的重要意义,推出了多种有关大数据的应用平台,非常有名是 Spark 平台和 Oracle 大数据机等知名项目,为企业的产品规划、决策及战略发展提供了重要的数据支撑^[30]。

3.1.1 大数据计算框架

目前,基于分布式处理技术的开源软件框架有 Apache Hadoop、Storm 和 Spark 等。Hadoop 能够利用简单的模型并行处理分布在各个集群计算机中的数据,只是难以满足大数据的实时处理和挖掘的需求。Storm 是基于拓扑的实时流处理计算逻辑,由 Zookeeper 连接 Nimbus 和 Supervisor,并协调分布式环境中的同步、命名等问题。Spark 是一种基于内存的流式实时数据处理平台,如图 3 所示。Spark 处理某个时间段窗口内的事件流,将用户定义的一系列 RDD 转化成 DAG 图,然后 DAG scheduler 把 DAG 再转化成 taskset (作业集),这样 taskset 就可以向集群申请计算资源,集群把这个 taskset 部署到 worker 中进行运算。开发者的任务是定义一些 RDD,并定义相应的转化动作,最后将这一系列的

RDD 投放到 Spark 的集群中运行。

比较典型的分布式流计算框架还有 IBM 的 InfoSphere Streams,它可以根据从多样的海量数据中提取出的自相关数据,实现对来自相关数据的信息和知识的积极分析和管理的。Apache Drill 可以用于解决大数据集的互动分析问题,可以处理分散在上千台服务器中的数据,它的主要优势在于可以在较小的时延内快速响应交互式查询。

3.1.2 大数据存储框架

对于大数据应用, NoSQL 数据库被认为比 SQL 更适合处理和捕捉海量数据。NoSQL 数据库是一种分布式横向扩展技术,它使用了分布式节点集来提供高度弹性扩展功能,让用户可以通过添加节点来动态处理负载。MongoDB 是一个介于关系数据库和非关系数据库之间的产品,支持的数据结构非常松散,因此可以存储比较复杂的数据类型。Neo4j 是一个高性能的 NoSQL 图形数据库,它将结构化数据存储在网络上而不是表中,Neo4j 具有嵌入式、高性能、轻量级等优势,在处理复杂的网络数据时,具有很好的性能表现。Oracle NoSQL Database 是来自 Oracle 的键值 NoSQL 数据库,具备数据备份和分布式键值存储系统。设计时考虑到了高扩展性和高可用性,并可部署于多个互相复制的节点上,以便进行快速故障切换及负载均衡。Apache HBase 源于 Google 的 Bigtable,是一个分布式面向列存储的 NoSQL 数据库。在 Hadoop 和 HDFS 之上提供了像 Bigtable 一样的功能。HBase 快照添加了类似“过程协作”的新功能,用于在线快照、写时备份、恢复和克隆。

除此之外, Greenplum 的统一分析平台结合 Greenplum DB 和 Greenplum Hadoop 为企业构建高效处理结构化、半

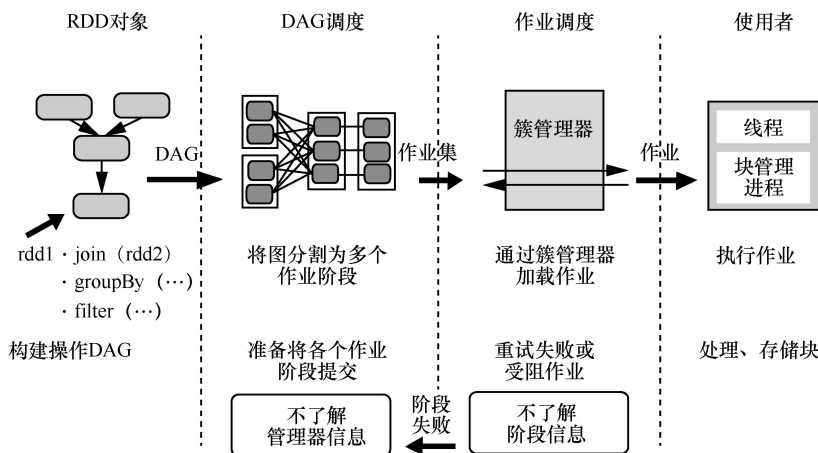


图 3 Spark 框架

结构化、非结构化数据的大数据分析平台。GoodData 提供的是基于云的数据分析服务,向 SaaS 提供商提供技术集成服务,使得这些 SaaS 提供商可以向最终用户提供诸如仪表盘、报表等功能。

3.1.3 大数据云平台

国内外的众多技术公司将大数据处理技术融入云平台^[31]。Amazon EMR (elastic MapReduce) 是亚马逊 (Amazon) 提供的大数据分析云服务,使用 Hadoop 开源框架将数据分布在 Amazon EC2 实例集群中并进行处理,分析用户提交的海量数据分析作业。BigQuery 是谷歌 (Google) 基于 Dremel 和 Hadoop 集群所提供的交互式大数据分析云服务^[32]。BigQuery 允许企业用户上传大数据集到谷歌的存储器中分析数据,同时可以建立应用程序共享服务。国内的引跑科技推出的大数据平台 EngineOne,是大数据处理技术和云平台的结合。EngineOne 云操作系统提供了构建和部署云应用程序所需的全部工具和 API,能让用户在基础设施上弹性部署并运行应用程序。

华存数据、曙光、Oracle 等知名国内外公司纷纷推出了性能卓越的大数据一体机。软件方面应用相应的数据库系统进行高级分析,而硬件方面一般都是云服务器、集群等作为软件的支持。华为公司还推出了名为 FusionCube 的融合一体机,把计算、存储、网络以及虚拟化平台有机融合在统一的架构之下。

3.2 大数据分析产品

面对如此庞大而复杂的数据,需要专门设计的硬件和软件工具进行专业处理。该类型的处理工具分为两种:数据可视化产品可将纷繁的数据呈现在用户面前,并对数据进行统计整理,方便用户了解当前数据分布情况;数据分析类产品则侧重于数据的分析和挖掘以及为企业用户提供比较完整的解决方案。

3.2.1 可视化产品

这类软件可以将复杂的大规模数据可视化,允许用户通过操作界面方便地与数据进行互动,并对其执行进一步的高级分析,以便获得更深入的业务洞察,更好地帮助用户进行企业决策。借助这类软件,用户可以从原始信息出发,最终得到隐含在结构化与非结构化数据中的重要信息。

IBM 的 Cognos Insight 商业智能解决方案可以向业务线使用人员提供丰富多变的可视化界面。它通过所有的系统和资料资源,提供了无缝密合的报表、分析、记分卡、仪表盘等解决方案。Google 等诸多公司也推出了适用于其他

行业的数据可视化工具。

3.2.2 分析类产品

Google Analytics (谷歌分析) 是 Google 公司推出的网站分析工具,用户把网站信息提交给 Google Analytics, Google Analytics 自动给出分析结果及丰富的图表式报告。Google Analytics 也可分析社交广告、移动广告、搜索广告的投放效果,使广告的作用极大地发挥。百度统计比 Google Analytics 更加本土化,更加适合中国用户。百度统计提供独立的 IP 地址数,分析报告的时效性更强,而且将百度指数和热门搜索词功能整合其中,把相关的信息串在一起,帮助分析流量原因。

对于多数企业来说,仅仅依靠网站分析是远远不能满足需求的,需要功能更强的大数据分析工具来辅助企业决策。阿里云开放数据处理服务 (open data processing service, ODPS),就是构建在大规模分布式计算系统上的,针对海量数据的处理工具,适用于海量数据统计、数据模型、数据挖掘、数据商业智能等诸多互联网应用。RadiumOne 通过分析 Facebook 中用户的社交网络行为,预测用户会对哪类产品感兴趣。而 Mintigo 推出的搜索引擎用来挖掘潜在用户,它首先对营销需求、目标行业以及用户人群进行分析。Connotate 的数据采集和分析服务适合企业使用,以指导他们更好地做出决策。Luminoso 利用 Connotate 在社交网站中发现消费者对新产品的喜好,以帮助他们做出更好的决策和提高。

3.3 大数据的标准化

在大数据标准化方面,目前最为实质性的是国际标准化组织 (International Organization Standardization, ISO)/国际电工委员会 (International Electrotechnical Commission, IEC) 第一联合技术委员会 (Joint Technical Committee 1, JTC1) 成立了大数据研究组,由美国国家标准技术研究所 (National Institute of Standards and Technology, NIST) 牵头, NIST 系统地开展了大数据架构、数据、安全需求等方面的研究。由于大数据作为一项新兴技术,国内外相关标准的研制还处于起步阶段^[33],简要介绍如下。

(1) ISO/IEC JTC1 SC32

持续致力于研制信息系统环境内及之间的数据管理和交换标准,为跨行业领域协调数据管理能力提供技术支持。SC32 下设 4 个工作组和几个研究组: WG1, 电子业务; WG2, 元数据; WG3, 数据库语言; WG4, SQL 多媒体和应用包。JTC1/SC32 现有的标准制定和研究工作为大数据



的发展提供了良好基础。

(2)ISO/IEC JTC1 SG2

负责大数据国际标准化的大数据研究组,有利于统筹开展大数据的标准化工作。调研各个标准化组织在大数据领域的关键技术、参考模型以及用例等标准基础;确定大数据领域应用需要的术语与定义;评估分析当前大数据标准的具体需求,提出 ISO/IEC JTC1 大数据标准的优先顺序;向 2014 年 ISO/IEC JTC1 全会提交大数据建议的技术报告和其他研究成果。

(3)NIST

力争形成达成共识的定义、术语、安全参考体系结构和技术路线图,提出数据分析技术应满足的互操作性、可移植性、可用性和扩展性需求,安全有效地支持大数据应用的技术基础设施。

(4)全国信息技术标准化技术委员会(以下简称全国信标委)

全国信标委(TC28)负责开展国内数据标准化工作,在元数据、数据库、数据建模、数据交换与管理等领域推动相关标准的研制与应用,为提升跨行业领域数据管理能力提供标准化支持。

3.4 大数据产品应用

互联网行业大数据应用的起步较早,其数据类型丰富、范围广,与各垂直行业不断进行着深度的融合^[34]。目前,互联网行业的大数据主要应用在社交网络、B2C 业务、精准营销、广告监测等方面,经典案例有百度、新浪微博、阿里巴巴等。

百度以传统的搜索引擎为基础,进一步推出了特色鲜明的百度大数据引擎。该引擎包括开放云、数据工厂和百度大脑 3 个部分。此前这些技术被应用在语音、图像和文本识别以及自然语言和语义理解方面,现在这些技术被用来对大数据进行智能化的处理、分析和学习,甚至组建了拥有 2×10^{11} 个参数的深度神经网络。基于图片识别技术,百度还新推出了拍照搜索功能,命名为百度识图。在使用手机拍摄照片后,搜索引擎能自动识别出用户拍摄的对象,并返回该对象的相关信息。对于搜索未知事物时经常遇到的词不达意的困惑,拍照的方式将更加直观、全面、易用。

新浪微博通过各种渠道聚集了海量数据信息,已沉积了将近 5×10^{11} 个人与人之间的联系图、人与物之间的联系图谱。经过研讨这些数据可以更精准地为每个用户推送更准确的广告^[35]。“带着微博去旅行”就是新浪微博推出的一

项微博旅游品牌活动,那么针对到达这些地区的旅行用户,可以及时推送当地的旅行信息、抢手景点以及电话费用等信息。移动端产品“Page 页面”是聚合了用户兴趣爱好社交关系数据的综合展示页面,连接人、物和兴趣,将微博的弱社交网络转换为真正的社会网络。

为了使淘宝商家能够实现数据化运营,阿里巴巴的数据魔方应运而生。根据用户的互联网痕迹进行渠道营销效果优化,从而调整营销资源在各个渠道的投放。例如,数据魔方可以根据不同时段在网站访问情况,分析顾客的需求和地理位置,实现按时按地点地精准地投放广告。还可以结合历史数据,给商家提供比较合理的价格参考。

3.5 大数据行业应用

大数据在现今的公共管理、零售、互联网、电信、金融等众多行业快速渗透^[35],影响力迅速扩大,已经成为重要的生产因素。目前各行业纷纷加快大数据的应用规模,重构未来的核心竞争力。对于很多行业而言,如何有效利用这些海量数据正成为赢得竞争的至关重要的因素之一。

对运营商来说,可利用数据与网络资源优势,挖掘数据的商业智能应用,为运营商创造更大的价值,这对于运营商转型升级具有重大的战略意义^[36]。为了应对大数据时代的挑战,中国移动率先提出了大数据超细分微营销服务的理念,将助力其全面提升精细化运营水平,改善用户体验。借助大数据分析识别用户特征,细分用户,进而指导公司制定合适的营销策略。其中,手机阅读业务需要数据分析与挖掘支持的解决方案,为精细化运营提供科学依据。杭州东信北邮信息技术有限公司为此研究、设计和开发了一整套大数据系统框架,具体架构如图 4 所示。特别是在 ETL、分布式计算以及智能推荐引擎方面分别设计和开发了一套完整的产品,有效地完成了新用户的挖掘和各种用户群体的分析,并针对不同情况给出了详细的营销方案,实现了精细化的营销策略。

4 大数据存在的挑战与机遇

目前对大数据的探讨还处于起步阶段,一些有价值的研究开始尝试全局性地整理大数据生态环境之中各部分、各模块之间的关系^[37,38]。大数据带来的描述与存储的挑战、分析与理解的挑战、挖掘与预测的挑战都有别于传统数据,除去开发新型数据库形式、提高数据挖掘算法性能、增强并行数据处理能力的研究与尝试,还需要对大数据有全局性的认识,通过分析大数据相关技术

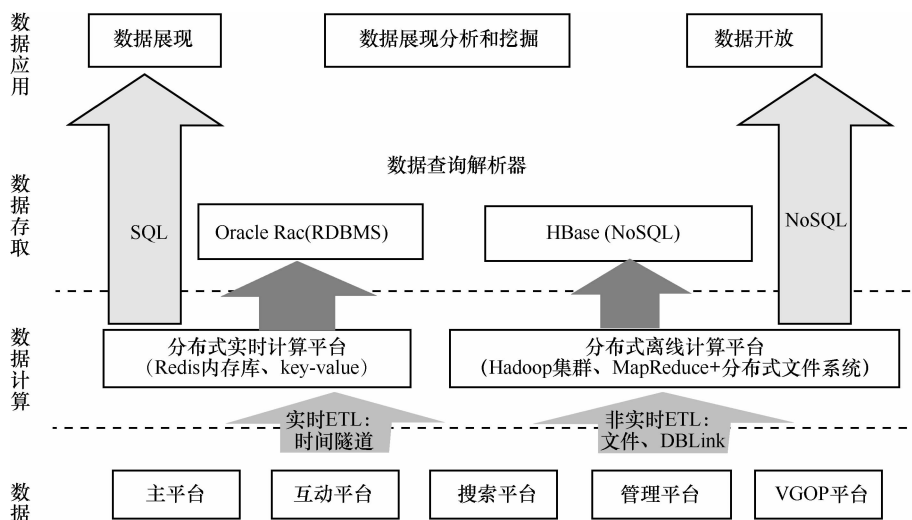


图4 中国移动手机阅读业务的大数据系统架构

的发展情况与不足,找到大数据发展过程中的痛点,以应对大数据发展中的机遇和挑战。结合大数据关键技术,从如下几个角度对大数据技术存在的机遇和挑战进行分析和预测。

4.1 大数据的新型分析策略

大数据时代的数据分析需求,不仅体现在对海量数据的处理能力上,同时也体现在对新的知识类型的兴趣上。进入大数据时代以来,数据查找、筛选甚至回归、关联规则分析等传统数据分析目标已经渐渐无法令人们满足,从而有了新的数据分析的目的,这就必然促使有针对性地开发设计新的数据分析策略。

数据是对现实事物的采样而非完整描述,这是大数据时代的一个崭新视角。在这个思想下,大数据处理就变成了大规模的常规数据处理,而除了依赖针对大数据的规模性数据处理算法之外,又多了依据数据采样进行数据分析的一条新路。以伯克利大学的 BlinkDB^[9]数据库系统为例,其运算速度如图5所示。在执行大规模数据查找的任务时,系统将数据库视为一个大样本集,逐次从数据库中以不同规模进行采样,使得通过采样结果得到的查询信息不断逼近实际结果,这一策略使得其在保证相同误差水平的条件下取得比 Hive 快 200 倍的查找速度。在社交网络应用方面,采样的策略同样有非常广泛的应用空间。往往不需要了解一个人的全部操作和发言,而只需要随机地从这个人的全部操作之中取规模很小的一个样本集,就可以得到与原有策略准确度不相上下的用户行为分析结果。

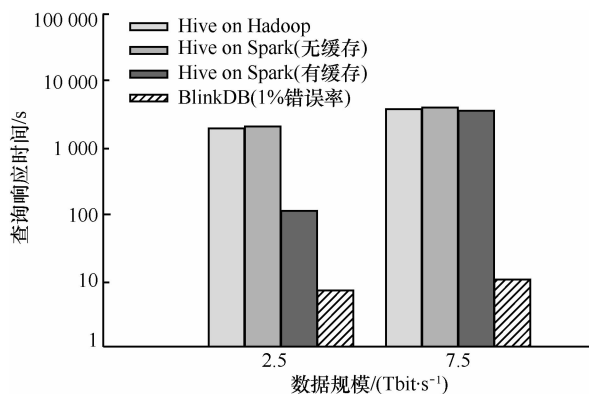


图5 BlinkDB 的运算速度

4.2 大数据的深度学习问题

在近几十年的研究中,研究人员普遍发现,基于人工神经网络结构的深度学习技术具有很好的学习能力、应用空间。然而目前深度学习领域的研究仍然存在一些有待解决的问题,这些问题又集中体现在如下3个方面。

(1) 理论分析

深度学习技术的成型依赖于对网络训练方法理论的创新,而现有深度网络所存在的诸多问题,也需要理论方面的突破性进展才能得到真正的解决。深度模型之所以存在训练方面的困难,一个主要的原因在于深度网络使用了非线性神经元结构,而非线性系统虽然具有更好的拟合目标函数的能力,却较难应用于基于梯度的算法进行优化训练。如果能够找到更好的 RBM 求解策略,或是发现新的可以替代 RBM 的结构,都将对深度学习技术的研究与应用产生举足轻重的影响。



(2) 数据特征提取

现阶段的一些结合 RBM 与 HMM、CRF 等传统特征提取算法的资料还基本处于将二者机械叠加的水平。由于 DBN 结构本身就是多层简单结构的叠加,这种叠加虽然具有形式简明、训练策略明朗的优势,却也牺牲了很多扩展性与针对性。发掘深度结构的潜在优势,找到更好的方法,建立适合不同数据特点的深度神经网络,将成为下一阶段深度学习领域研究的重点。

(3) 训练与优化求解技术

基于梯度的训练算法对随机初始化的深度神经网络结构收敛较差,容易陷入局部最优而导致性能下降,这一现象引出了一系列半监督训练策略、回火策略等算法及变形,依然需要在训练算法、并行性优化等方面大幅度提升现有技术水平。

4.3 大数据的隐私保护问题

大数据时代的隐私问题与以往常规数据时代的数据隐私问题相比,具有很多新的特征,其中较为主要的包括数据收集过程、数据分析过程中带来的隐私问题。

随着个人电子终端的传感器化、用户行为记录的细粒度化、信息收集的隐秘化等现象的出现,数据收集过程逐渐呈现出一种以往完全无法想象的状态,即由用户主动提供数据向数据收集者主动收集数据变化、用户知情向用户不知情变化、数据实际使用方式模糊化^[40]。用户往往在毫不知情的前提下,个人数据、操作记录等信息就已经被隐秘地收集了,而实际上用户有权力知晓这些数据是否被共享、误用、恶意传播、销毁。数据收集过程中的隐私保护问题有别于一般的数据安全问题,较难通过技术方式进行解决,主要依赖相关政策法规的约束。

现有的大数据隐私研究除去继承常规数据的隐私处理方法之外,还使用了诸如分级访问控制策略^[41]、使用添加数据机制防范背景知识分析攻击^[42]等策略。基于“知情与同意”策略的位置隐私保护策略在大数据时代已经不足以满足隐私保护的实际需求,基于启发式隐私度量、概率推测和隐私信息检索的大数据隐私保护机制的相关研究,仍有大量的工作需要进一步进行。

4.4 大数据的数据质量与可用性问题

在解决大数据存储、传输、处理的相关问题,建立数据中心基础架构、实现大数据标准化规范的过程中,数据质量与可用性始终是无法忽视的问题^[43,44]。大数据之中蕴藏着巨大的潜在价值,要想确保能够从大数据中取得正确的

信息,进而以之辅助决策、实现相关应用,关键的前提是保证数据质量。然而在实际情况中,能够取得的数据往往是有许多冗余、噪点、误差的质量不高的数据,这对大数据的知识发现、应用都有负面的影响。确保数据可用性同样是一项十分困难的任务,大数据在一致性、精确性、完整性、时效性、实体统一性这 5 个数据可用性维度上都存在着一定的可用性难题。

大数据可用性面临的主要挑战包括如下几个方面。

- 探索从物理信息系统等多数据源有效地获取高质量大数据的理论和算法,研究高效数据过滤方法,建立多模态大数据融合计算的理论和算法,实现高质量数据的获取和精准整合,继而发现数据演变规律。
- 大数据可用性面临很多数据恢复、定量评估、质量融合管理、数据演化方面的新问题,应当结合大数据处理的新方法、新技术,有针对性地对大数据可用性进行定性定量的分析,设计并建立大数据可用性的理论模型、形式化系统和评估算法。
- 弱可用数据指的是无法彻底修复的数据,在大数据环境下,弱可用数据的规模较以往大幅度增加。除进一步研究数据错误自动检测与修复的理论和算法之外,对近似计算理论和技术的研究、对弱可用数据的知识校验和纠偏、对弱可用数据的知识演变机理方面的研究,也是解决弱可用数据问题的重要思路。

5 结束语

在大数据时代,数据的收集、获取和分析都更加快捷,这些海量的数据将对人们的思考方式产生深远的影响。纵观人类社会的发展史,人的需求及意愿始终是推动科技进步的源动力。随着物联网的兴起,移动感知技术的发展,数据采集技术的进步,人不仅是大数据的使用者和消费者,还是生产者和参与者。基于大数据的社会关系感知、众包、社交网络大数据分析等与人的活动密切相关的应用,大数据将在各行业产生重大的社会价值,也必将引起社会活动的巨大发展和变革。未来还需要进一步加强大数据标准化顶层设计,推动开放数据集建设,建立健全的大数据标准体系,重点突破一批涉及大数据发展的基础性、方法性、公共性标准的研制,为大数据发展和应用夯实基础。

参考文献

- 1 Gantz J, Reinsel D. Extracting Value from Chaos. IDC iView Report, 2011
- 2 Schnberger V M, Cukier K. 大数据时代: 生活、工作与思维的大变革. 盛杨燕, 周涛译. 杭州: 浙江人民出版社, 2013
Schnberger V M, Cukier K. Big Data: A Revolution that Will Transform How We Live, Work, and Think. Translated by Sheng Y Y, Zhou T. Hangzhou: Zhejiang People's Publishing House, 2013
- 3 Team O R. Big Data Now: Current Perspectives from O'Reilly Radar. Sebastopol: O'Reilly Media, 2011
- 4 Grobelnik M. Big data tutorial. [http://videlectures.net/eswc2012grobelnik big data/](http://videlectures.net/eswc2012grobelnik%20big%20data/), 2012
- 5 张引, 陈敏, 廖小飞. 大数据应用的现状与展望. 计算机研究与发展, 2013, 50(S2)
Zhang Y, Chen M, Liao X F. Big data applications: a survey. Journal of Computer Research and Development, 2013, 50(S2)
- 6 Binzenhöfer A, Tutschku K, Graben B A D, *et al.* A P2P-based framework for distributed network management. Lecture Notes in Computer Science, 2006(3883): 198~210
- 7 Tutschku K, Chevul S, Binzenhfer A, Schmid M, *et al.* A self-organizing concept for distributed end-to-end quality monitoring. University of Wurzburg Institute, Wurzburg, Germany, 2006
- 8 李强, 王宏, 王乐春. 基于 P2P 的分布式网络管理模型研究. 计算机工程, 2006, 32(13): 150~152
Li Q, Wang H, Wang L C. Research of P2P based distributed network management. Computer Engineering, 2006, 32(13): 150~152
- 9 Karagiannis T, Papagiannaki K, Faloutsos M. Blinc: multilevel traffic classification in the dark. Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, Pennsylvania, USA, 2005
- 10 Karagiannis T, Roido A, Aloutsos M, *et al.* Transport layer identification of P2P traffic. Proceedings of the 2004 ACM SIGCOMM Internet Measurement Conference, Taormina, Italy, 2004
- 11 鲍静, 范生万. 基于数据挖掘的图书数据预处理. 大学图书馆学报, 2008, 26(2): 31~33
Bao J, Fan S W. Pre-processing library data based on data milling. Journal of Academic Library and Information Science, 2008, 26(2): 31~33
- 12 刘云霞. 数据归约的统计方法研究及应用 (博士学位论文). 厦门大学, 2007
Liu Y X. Research and application of the statistical method of data reduction (doctor dissertation). Xiamen University, 2007
- 13 刘鹏. 云计算技术基础. 北京: 电子工业出版社, 2011
Liu P. Fundamental Cloud Computing Technology. Beijing: Publishing House of Electronics Industry, 2011
- 14 王庆波, 金津, 何乐. 虚拟化与云计算. 北京: 电子工业出版社, 2010
Wang Q B, Jin X, He L. Virtualization and Cloud Computing. Beijing: Publishing House of Electronics Industry, 2010
- 15 王鹏. 云计算的关键技术与应用实例. 北京: 人民邮电出版社, 2010
Wang P. Key Technologies and Applications of Cloud Computing. Beijing: Posts and Telecom Press, 2010
- 16 Agrawal D, Bernstein P, Bertino E, *et al.* Challenges and opportunities with big data. Challenges and Opportunities with Big Data-ResearchGate, 2012, 6(12): 2032~2033
- 17 Luo D, Ding C, Huang H. Parallelization with multiplicative algorithms for big data mining. Proceedings of IEEE 12th International Conference on Data Mining, Brussels, Belgium, 2012: 489~498
- 18 Gao F, Abd-Elmageed W, Hefeeda M. Distributed approximate spectral clustering for large-scale datasets. Proceedings of the 21st International ACM Symposium on High-Performance Parallel and Distributed Computing, Delft, the Netherlands, 2012: 223~234
- 19 Shim K. MapReduce algorithms for big data analysis, and storage of big data. Proceedings of the VLDB Endowment, Istanbul, Turkey, 2012: 2016~2017
- 20 Abadi D J, Ahmad Y, Balazinska M, *et al.* The design of the borealis stream processing engine. Proceedings of 2nd Biennial Conference on Innovative Data Systems Research (CIDR'05), Asilomar, CA, USA, 2005
- 21 Motwani R, Widom J, Arasu A, *et al.* Query processing, resource management, and approximation in a data stream management system. Proceedings of the 2003 CIDR Conference, Asilomar, CA, USA, 2003
- 22 Han J W, Kamber M, Pei J, *et al.* 数据挖掘: 概念与技术. 范明, 孟小峰译. 北京: 机械工业出版社, 2005
Han J, Kamber M, Pei J, *et al.* Data Mining: Concepts and Techniques. Translated by Fan M, Meng X F. Beijing: China Machine Press, 2005
- 23 Al-Fhres M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture. Proceedings of ACM SIGCOMM, Seattle, WA, USA, 2008
- 24 Mitchell B, Sheppard J. Deep structure learning: beyond connectionist approaches. Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA), Boca Raton, Florida, USA, 2012: 162~167
- 25 Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527~1554



- 26 FuKusHIMA K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980, 36(4): 193~202
- 27 Ranzato M, Susskind J, Mnih V, *et al.* On deep generative models with applications to recognition. *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, 2011: 2857~2864
- 28 Guo C, Wu H, Tan K, *et al.* DCell: a scalable and fault-tolerant network structure for data center. *Proceedings of ACM SIGCOMM*, Seattle, WA, USA, 2008
- 29 Li D, Guo C, Wu H, *et al.* FiConn: using backup port for server interconnection in data centers. *Proceedings of IEEE INFOCOM*, Rio de Janeiro, Brazil, 2009
- 30 胡雄伟, 张宝林, 李抵飞. 大数据研究与应用综述. *标准科学*, 2013(10): 18~21
Hu X W, Zhang B L, Li D F. Overview of big data research and application. *Standard Science*, 2013(10): 18~21
- 31 方巍, 文学志, 潘吴斌等. 云计算: 概念、技术及应用研究综述. *南京信息工程大学学报: 自然科学版*, 2012, 4(4): 351~361
Fang W, Wen X Z, Pan W B, *et al.* Cloud computing: survey on concept, technology and application. *Journal of Nanjing University of Information Science & Technology*, 2012, 4(4): 351~361
- 32 Chandarana P, Vijayalakshmi M. Big data analytics frameworks. *Proceedings of 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, Mumbai, India, 2014: 430~434
- 33 中国电子技术标准化研究院. 大数据标准化白皮书, 2014
China Electronics Standardization Institute. *Big Data Standardization White Paper*, 2014
- 34 宫夏屹, 李伯虎, 柴旭东等. 大数据平台技术综述. *系统仿真学报*, 2014, 26(3): 489~496
Gong X Y, Li B H, Chai X D, *et al.* Survey on big data platform technology. *Journal of System Simulation*, 2014, 26(3): 489~496
- 35 黄哲学, 陈小军, 李俊杰等. 面向服务的大数据分析平台解决方案. *科技促进发展*, 2014, 10(1): 52~59
Huang Z X, Chen X J, Li J J, *et al.* Service oriented big data analytics platform. *Science & Technology for Development*, 2014, 10(1): 52~59
- 36 袁玮. 云计算在电信行业经营分析系统中对大数据的处理探析. 硅谷, 2014(6)
Yuan W. Processing analysis for big data from cloud computing in telecom industry management analysis system. *Silicon Valley*, 2014(6)
- 37 Demchenko Y, De Laat C, Membrey P. Defining architecture components of the big data ecosystem. *Proceedings of 2014 International Conference on Collaboration Technologies and Systems (CTS)*, Minneapolis, MN, USA, 2014: 104~112
- 38 Zeng D, Lusch R. Big data analytics: perspective shifting from transactions to ecosystems. *IEEE Intelligent Systems*, 2013, 28(2): 2~5
- 39 Agarwal S, Mozafari B, Panda A, *et al.* BlinkDB: queries with bounded errors and bounded response times on very large data. *Proceedings of the 8th ACM European Conference on Computer Systems*, Prague, Czech Republic, 2013: 29~42
- 40 孟小峰. 位置大数据隐私保护研究综述. *软件学报*, 2014, 25(4): 693~712
Meng X F. Location privacy preservation in big data era: a survey. *Journal of Software*, 2014, 25(4): 693~712
- 41 Cheng Y, Park J, Sandhu R. Preserving user privacy from third-party applications in online social networks. *Proceedings of the 22nd International Conference on World Wide Web Companion*, New York, USA, 2013: 723~728
- 42 Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms. *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, Bethesda, Maryland, USA, 2009: 351~360
- 43 李默涵, 李建中, 高宏. 数据时效性判定问题的求解算法. *计算机学报*, 2012, 35(11): 2348~2360
Li M H, Li J Z, Gao H. Evaluation of data currency. *Chinese Journal of Computers*, 2012, 35(11): 2348~2360
- 44 刘波, 耿寅融. 数据质量检测规则挖掘方法. *模式识别与人工智能*, 2012, 25(5): 835~844
Liu B, Geng Y R. Mining method for data quality detection rules. *Pattern Recognition and Artificial Intelligence*, 2012, 25(5): 835~844

[作者简介]



廖建新, 男, 北京邮电大学网络与交换技术国家重点实验室网络智能研究中心主任、二级教授、博士生导师, “长江学者”特聘教授, 教育部“创新团队”带头人, 国务院学位委员会博士、硕士学位授权点评议专家, 北京邮电大学学术委员会委员, 《北京邮电大学学报》、《电信科学》编委。主要从事业务网络智能化、下一代互联网和大数据方向的研究工作; 主持国家“973”计划项目、自然科学基金项目、国家高技术产业化信息化装备专项基金课题等国家级、部级科研项目 20 余项; 作为第一完成人获国家科技进步奖二等奖 2 项, 原信息产业部重大技术发明奖 1 项及部级科技进步奖一等奖 3 项; 获国家杰出青年基金资助, 入选新世纪百千万人才工程国家级人选、教育部新世纪优秀人才支持计划, 被评选为全国优秀博士后, 获第九届中国青年科技奖, 信息产业科技创新先进个人。在国际著名期刊发表 SCI 检索论文 30 余篇; 出版专著 3 部, 合作出版国际专著 2 部; 作为第一完成人申请国家发明专利 50 余项, 其中 37 项已授权, 1 项获得中国专利优秀奖。