大数据研究

严霄凤,张德馨
(工业和信息化部 计算机与微电子发展研究中心(中国软件评测中心),北京 100048)

摘 要: 大数据是继云计算、物联网之后 IT 产业又一次颠覆性的技术革命。大数据挖掘和应用可创造出超万亿美元的价值,将是未来 IT 领域最大的市场机遇之一。大数据利用对数据处理的实时性、有效性提出了更高要求,需要根据大数据特点对传统的常规数据处理技术进行技术变革,形成适用于大数据收集、存储、管理、处理、分析、共享和可视化的技术。文中介绍了大数据的概念及其关键技术,描述了大数据带来的机遇和挑战,概述了美国政府的"大数据研究和发展倡议"。

关键词: 大数据; 云计算; 信息安全
中图分类号: TP393 **文献标识码:** A **文章编号:** 1673-629X(2013)04-0168-05
doi: 10.3969/j.issn.1673-629X.2013.04.041

Big Data Research

YAN Xiao-feng, ZHANG De-xin
(Research Center for Computer and Microelectronics Industry Development(China Software Testing Center), MIIT, Beijing 100048, China)

Abstract: Big Data is a disruptive technology revolution of IT industry following the cloud computing and Internet of Things. Big Data mining and applications can create significant value for the world economy, will be one of the biggest market opportunities of IT field in the future. Use of Big Data puts forward higher requirements for the real-time and effectiveness of data processing, demands improving conventional data processing technology according to the characteristics of Big Data, forms the techniques appropriate to Big Data collection, storage, management, processing, analysis, sharing and visualization. Introduce the concept of Big Data and its key technologies, describe the opportunities and challenges brought about by the Big Data, outline the "Big Data Research and Development Initiative" of U.S. federal government.

Key words: Big Data; cloud computing; information security

0 引 言

当人们还在津津乐道云计算、物联网等主题时,一个崭新的概念——大数据横空出世。大数据是继云计算、物联网之后 IT 产业又一次颠覆性的技术革命[1],对国家治理模式、企业决策、组织和业务流程,以及个人生活方式等都将产生巨大的影响。大数据的挖掘和应用可创造出超万亿美元的价值,将是未来 IT 领域最大的市场机遇之一,其作用堪称又一次工业革命。

收稿日期: 2012-07-31; 修回日期: 2012-10-31
基金项目: 国家科技支撑计划项目(2009BAH39B00, 2009BAH39B05); 国家发展改革委信息安全专项项目(发改办高技【2010】3044号)
作者简介: 严霄凤(1964-),女,山西太原人,硕士,主要从事信息安全、电子认证相关标准、规范和测评方法研究,长期从事信息系统测试、信息安全测评和政府信息系统安全、电子认证、电子签名相关课题等的研究实施工作。

1 大数据

麦肯锡将大数据定义为: 无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合[2]。

大数据不是一种新技术,也不是一种新产品,而是一种新现象,是近来研究的一个技术热点。大数据具有以下 4 个特点,即 4 个"V"[3]:

(1) 数据体量(Volumes) 巨大。大型数据集,从 TB 级别,跃升到 PB 级别。

(2) 数据类别(Variety) 繁多。数据来自多种数据源,数据种类和格式冲破了以前所限定的结构化数据范畴,囊括了半结构化和非结构化数据。

(3) 价值(Value) 密度低。以视频为例,连续不间断监控过程中,可能有用的数据仅仅一两秒钟。

(4) 处理速度(Velocity) 快。包含大量在线或实时数据分析处理的需求,1 秒定律。

随着互联网技术的不断发展,数据本身就是资产。云计算为数据资产提供了保管、访问的场所和渠道,但

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

如何盘活数据资产,使其为国家治理、企业决策乃至个人生活服务,是大数据的核心议题,也是云计算的灵魂和必然的升级方向。

大数据已经出现。IDC 多年的研究结果告诉我们: 全球数据量大约每两年翻一番,每年产生的数据量按指数增长,数据增速基本符合摩尔定律。全球有 46 亿移动电话用户,有 20 亿人访问互联网,人们以比以往任何时候都高得多的热情在与数据或信息交互。思科公司预计,到 2013 年,在互联网上流动的数据量将达到每年 667 艾字节[4]。

2　大数据关键技术

随着互联网、云计算和物联网的迅猛发展,无所不在的移动设备、RFID、无线传感器每分每秒都在产生数据,数以亿计用户的互联网服务时时刻刻在产生巨量的交互……。要处理的数据量越来越大、而且还将更加快速地增长,同时业务需求和竞争压力对数据处理的实时性、有效性也提出了更高要求,传统的常规数据处理技术已无法应付,大数据带来了很多现实的难题。为了解决这些难题需要突破传统技术,根据大数据的特点进行新的技术变革。大数据技术是一系列收集、存储、管理、处理、分析、共享和可视化技术的集合。适用于大数据的关键技术[5~8]包括:

遗传算法。借鉴生物界的进化规律(适者生存,优胜劣汰遗传机制) 演化而来的随机化搜索方法。采用概率化的寻优方法,自动获取和指导优化的搜索空间,不需要确定的规则,自适应地调整搜索方向。已被人们广泛应用于组合优化、机器学习、信号处理、自适应控制和人工生命等领域。是现代有关智能计算中的关键技术。应用实例包括制造业改善作业调度,以及优化投资回报率等。

神经网络。受生物神经网络结构和运作的启发,模拟动物神经网络行为特征,进行分布式并行信息处理的算法数学模型。应用实例包括识别高价值客户离开特定公司的风险,以及识别欺诈性的保险理赔行为等。

数据挖掘。结合统计数据和机器学习、使用数据库管理技术从大型数据集中提取有用信息和知识的技术。根据其它属性的值预测特定(目标) 属性的值,如回归、分类、异常检测等,或寻找概括数据中潜在联系的模式,如关联分析、演化分析、聚类分析、序列模式挖掘等。

回归分析。确定当一个或多个独立变量值被修改时相关变量如何变化的统计方法。通常用于预测或预报。应用实例如基于不同的市场和经济变量,或通过确定何种制造业参数对客户满意度影响最大来预测销售量等。用于数据挖掘。

分类分析。在训练集包含的数据点已经被归类的基础上,确定新的数据点所属类别的方法。典型应用是在明确假设或客观结果前提下,预测部分特定客户行为(例如,购买决策、流失率、消费率等) 。因为使用训练集,属于监督学习,是无监督学习类型聚类分析的反面。用于数据挖掘。

聚类分析。一种多元化群体的分类统计方法。在事先不知道的前提下,将一个集合分成较小的对象组,组内对象具有相似特点。聚类分析的典型例子是将消费者分割成具有自相似性的群体做针对性营销。因为不使用训练数据,属于无监督学习类型,是监督学习类型分类分析的反面。用于数据挖掘。

关联规则学习。在大数据集变量中发现感兴趣关系(即“关联规则”) 的方法,包括多种生成和测试可能规则的算法。典型应用是市场购物篮分析,其中零售商可以决定哪些产品经常一起购买和如何使用这种营销信息。用于数据挖掘。

数据融合与集成。集成和分析来自多个源的数据的方法。典型应用如,使用来自互联网的传感器数据综合分析如炼油厂这样的复杂分布式系统的性能。使用社会媒体数据,经过自然语言处理分析,并结合实时销售数据,确定营销活动如何影响顾客的情绪和购买行为等。

机器学习。研究计算机怎样模拟或实现人类的学习行为,获取新的知识或技能,重新组织已有的知识结构并不断改善自身的性能,是人工智能的核心,是使计算机具有智能的根本途径。自然语言处理是机器学习的一个例子。

自然语言处理。研究实现人与计算机之间用自然语言进行有效通信的理论和方法。典型应用是使用社交媒体的情感分析来确定潜在客户对品牌活动的反应等。

情感分析。从源文字材料中确定和提取主观信息的自然语言处理和分析方法的应用。分析的主要内容包括识别表达情感的特征、态势或作品。应用实例是分析社会化媒体(如博客、微博客或社交网络) 确定不同客户群和利益相关者对其产品和行为的反应。

网络分析。在图或网络中描述离散节点之间特征关系的方法。在社会网络分析中,分析个人在社会或组织之间的联系,如信息如何传播或谁拥有了其中的大部分影响。应用实例包括确定营销目标的关键意见负责人,以及确定企业信息流的瓶颈等。

空间分析。分析数据集拓扑、几何或地理编码性能技术的统计方法。数据通常来源于采集地址或纬度/经度坐标等位置的地理信息系统。应用实例包括

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

空间数据的空间回归(例如,消费者是否愿意购买与位置相关的产品) 或模拟(例如,如何将制造业的供应链网络分布到不同的地点) 。

时间序列分析。分析数据点序列表示连续时间值,从数据中提取有意义特征的统计学和信号处理方法。一般通过曲线拟合和参数估计来建立数学模型。应用实例包括销售数字预测、气象预报、水文预报,或将诊断为传染性疾病人数的预测等。

分布式文件系统。最典型的是 Google 的 GFS,部分源自于 Hadoop 的灵感。Hadoop 是一个处理分布式系统问题中庞大数据集的软件框架,具备低廉的硬件成本、开源的软件体系、较强的灵活性、允许用户自己修改代码等特点,同时能支持海量数据存储和计算任务。MapReduce 是谷歌推出的,处理庞大数据集分布式系统的软件框架。

分布式缓存。缓存在 Web 开发中运用越来越广泛,Memcached 是一个高性能的分布式内存对象缓存系统,用于动态 Web 应用以减轻数据库负载。通过在内存中缓存数据和对象来减少读取数据库的次数,从而提供动态、数据库驱动网站的速度,提升性能。MemcacheDB 是一个分布式、Key - Value 形式的持久存储系统,是一个基于对象存取、可靠、快速的持久存储引擎。协议与 Memcached 一致(不完整) ,所以很多 Memcached 客户端都可以跟它连接。MemcacheDB 采用 Berkeley DB 作为持久存储组件,支持很多 Berkeley DB 的特性。类似这样的产品还有很多,如淘宝的 Tair。

分布式数据库。Greenplum 数据引擎软件专为新一代数据仓库所需的大规模数据和复杂查询功能所设计,基于大规模并行处理和完全无共享架构、开源软件和 x86 商用硬件设计,性价比更高。Hive 是一个基于 Hadoop 的数据仓库平台,将转化为相应的 MapReduce 程序,基于 Hadoop 执行。通过 Hive 开发人员可以方便地进行数据提取、转换和加载开发。Big Table 是建立在谷歌文件系统上的专用分布式数据库系统,来源于 HBase 的启发。Cassandra 是一个开源数据库管理系统,处理分布式系统上的大量数据。

非关系型数据库系统。HBase 是一个仿照谷歌 Big Table 的开源分布式非关系型数据库。是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统,利用 HBase 技术可在廉价 PC Server 上搭建大规模结构化存储集群。HBase 是 Big Table 的开源实现,使用 HDFS 作为其文件存储系统。利用 MapReduce 来处理 HBase 中的海量数据。Dynamo 是由亚马逊开发的专用分布式数据存储系统。

可视化技术。可视化是支持大数据蓬勃发展的重要领域。可视化技术通过创建图片、图表或动画等,方便对大数据分析结果的沟通与理解。标签云即加权视觉列表,将其中出现频繁的词以更大的文本呈现,不经常出现的词用较小的文本呈现,帮助读者迅速感知大文本中最突出的概念; Clustergram 是一种聚类分析可视化技术,用于显示随着集群数量的增加,数据集的个别成员如何被分配到集群。使分析师能够更好地了解为何不同的集群数量产生不同的聚类结果; 历史流用图形化的方法表示多个作者编辑文件的历史,在图中很容易发现不同的见解。空间信息流在视图中通过不同亮度、颜色等显示统计分析参数。如利用视图显示纽约和世界各地城市之间 IP 数据流的大小,在图中特定城市所在位置以不同亮度反映该城市和纽约之间的不同 IP 流量,可以快速确定哪些城市与纽约的通信量大。

3 大数据带来的机遇和挑战

大数据瓦解了传统信息体系架构,将数据仓库转化为具有流动、连接和信息共享的数据池。大数据技术使人们可以利用以前不能有效利用的多种数据类型,抓住被忽略的机遇,使企业组织更加智能和高效。大数据技术也将推动新兴信息安全技术与产品的形成[2,3] 。

3.1 大数据带来的机遇

(1) 大数据的挖掘和应用成为核心,将从多个方面创造价值。

大数据的重心将从存储和传输,过渡到数据的挖掘和应用,这将深刻影响企业的商业模式。据麦肯锡测算,大数据的应用每年潜在可为美国医疗健康业和欧洲政府分别节省 3000 亿美元和 1000 亿欧元,利用个人位置信息潜在可创造出 6000 亿美元价值,因此大数据应用具有远超万亿美元的大市场。

(2) 大数据利用中安全更加重要,为信息安全带来发展契机。

随着移动互联网、物联网等新兴 IT 技术逐渐步入主流,大数据使得数据价值极大提高,无处不在的数据,对信息安全提出了更高要求。同时,大数据领域出现的许多新兴技术与产品将为安全分析提供新的可能性; 信息安全和云计算贯穿于大数据产业链的各个环节,云安全等关键技术将更安全地保护数据。大数据对信息安全的要求和促进将推动信息安全产业的大发展。

(3) 大数据时代来临,使商业智能、信息安全和云计算具有更大潜力。

大数据产业链按产品形态分为硬件、基础软件和应用软件三大领域,商业智能、信息安全和云计算主题

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

横跨三大领域,将构成产业链中快速发展的三驾马车。就国内而言,商业智能市场已步入成长期,预计未来 3 年复合年均增长率(CAGR) 为 35%, "十二五"期间潜在产值将超 300 亿元; 信息安全预计未来 3 年 CAGR 有望保持 35% ~40% 的快速增长, "十二五"期间潜在产值将超 4000 亿元; 云计算刚进入成长期,预计未来 5 年 CAGR 将超 50% ,2015 年产业规模预计将达 1 万亿元。

3.2　大数据带来的挑战

大数据在带来机遇的同时也在人才、技术、信息安全等方面带来了很大的挑战。

(1) 大数据需要专业化的技术和管理人才。

大数据解决方案的设计和实施,需要专业化分析复杂数据集的工具和技术,包括统计学、机器学习、自然语言处理和建模,以及可视化技术,例如标签云、集群、历史流、动画和信息图表等。

大数据时代,企业、组织需要大量既精通业务又能进行大数据分析的人才,美国目前面临 14 万至 19 万分析和管理人才,以及 150 万具备理解和基于大数据研究做出决策的经理和分析师人才的缺口,我国目前 IT 人员本身配备不足的现状与大数据需要 IT 人员增加的矛盾更加突出,大数据对我国人才的培养模式以及现有人才的储备提出了严峻的挑战。

(2) 大数据的有效应用需要解决大容量、多类别和高时效数据处理的问题。

传统数据库的管理能力无法应付大数据体量的数据。传统数据库处理不了数 TB 级别的数据,也不能很好支持高级别的数据分析,大数据急速膨胀的数据体量已经超越了传统数据库的管理能力。

大数据中不同格式的数据需要复杂的处理方法。大数据包括了越来越多的数据格式,囊括了半结构化和非结构化数据,非结构化数据的多样性和海量性,决定了大数据技术的复杂性,这些数据的处理超出了目前常规数据软件工具所能承受的极限。

大数据处理需要满足极高的时效性。在当今快速变化的社会经济形势面前,把握数据的时效性,是立于不败之地的关键。数据量大意味着计算开销大,数据多样性意味着算法可扩展性要强,二者制约了大数据处理技术的时效性,大数据的实时处理给大数据技术带来了更大的挑战。

贯穿数据采集、存储、处理、检索、分析和展现的全生命周期,大数据将挑战企业的存储架构、数据中心的基础设施等,还将引发数据仓库、数据挖掘、商业智能、云计算等应用的连锁反应。

(3) 大数据利用对信息安全提出了更高要求。

大数据时代,数据价值越来越大,面对海量数据的收集、存储、管理、分析和共享,信息安全问题成为重中之重。

防止数据被窃取或篡改。大数据的海量数据,通常采用云端存储,数据管理比较分散,对用户进行数据处理的场所无法控制,很难区分合法与非法用户,容易导致非法用户入侵,窃取或篡改重要数据信息。如何保证大数据的安全以及分析结果的可靠是信息安全领域需要解决的新课题。

防止个人信息泄漏。大数据中包含了大量的个人隐私,以及各种行为的细节记录。如何做到既深入挖掘其中给人类带来利益的智慧部分,又充分保护个人隐私不被滥用,在大数据的利用中找到个人信息开放和保护的平衡点,是大数据提出的又一巨大难题。

4　美国政府的"大数据研究和发展倡议"

今年 3 月 29 日,美国政府宣布"大数据研究和发展倡议"(以下简称"倡议") [10,11] ,提出通过增强收集海量数据、分析萃取信息的能力,加快美国在科学与工程领域发明的步伐,增强国家安全,转变现有的教学和学习方式,同时希望与 IT 行业和相关机构携手迎接大数据的机遇和挑战。

"倡议"发布的同时,第一轮扶持大数据的六个美国联邦部门和机构,包括美国国家科学基金会(NSF) 、美国国家卫生研究院(NIH) 、美国能源部(DOE) 、美国国防部(DOD) 、美国国防部高级研究计划局(DARPA) 和美国地质勘探局(USGS) ,承诺将提供两亿多美元来推动大数据相关工具及技术的研发。

"倡议"还透露了联邦政府各部门正在进行的多项大数据计划。在解决新一代网络以及云计算环境中的信息安全问题方面设置了相关项目,包括 DARPA 的多尺度异常检测(ADAMS) 、网络内部威胁(CINDER) 、Insight、面向任务的弹性云、加密数据编程计算(PROCEED) 和国家安全局的预警网络等。在大数据人才培养和基础技术研究方面设置了相关项目,包括 NSF 的 21 世纪科学与工程网络基础设施框架(CIF21) 、CIF21 对 IGERT 的跟踪、创意实验室、开放科学网格(OSG) 、推进大数据科学与工程的核心技术(BIGDATA) 、数据引用、开放科学数据和软件保护(DASPOS) 、数据挖掘挑战、计算探索、随机网络模型、信息集成和信息学、计算与数据处理科学工程(CDS&E) 等。

"倡议"重要意义在于:

(1) 数据主权继边防、海防、空防之后,将成为另一个大国博弈的空间,占有和控制数据主权是维护美国新信息时代霸主地位的重要措施。

(2) 通过全球战略下的"新军事战略"和"反恐战

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

略”,将军方纳入“倡议”,并为美国整合强化国家情报信息网络体系和提高军事情报信息处理能力提供有效的技术手段和工具。

(3) 研发新的技术、方法、工具,提高国家安全等领域利用大数据能力,增强国家安全力量,巩固美国信息安全保障体系。

(4) 将大数据技术从商业行为上升为国家意志,推动大数据相关产业链发展,使大数据产业迎来快速发展的机会,通过高科技寻求经济增长和复苏的新途径。

5 结束语

大数据时代已经来临,各国将在这一新的领域展开新一轮的竞争,我国应当与时俱进、及时转型,适应大数据时代的到来。可以借鉴美国“倡议”的做法,抓住大数据时代的关键点,从国家战略制定、人才培养、基础技术研究、信息安全保障体系建设等方面展开相应的工作。

参考文献:

[1] 赛迪智库: 大数据时代需要加快布局 [EB/OL]. 2012-05-17. http://www.cio360.net/index.php? m=content&c=index&a=show&catid=201&id=53375.

[2] 计算机行业-大数据(Big Data) 专题报告 [R]. 上海: 光大证券股份有限公司研究所,2011.

[3] 大数据分析技术的发展 [EB/OL]. 2012-05-16. http://tech.ccidnet.com/art/32963/20120516/3859799_1.html.

[4] 大数据时代来临何为大数据? [EB/OL]. 2012-05-12. http://datacenter.ctocio.com.cn/464/12331964.shtml.

[5] Big data: The next frontier for innovation, competition, and productivity [R]. USA: The McKinsey Global Institute,2011.

[6] 刘 俊. 基于大数据流的 Multi-Agent 系统模型研究 [J]. 计算机技术与发展,2007,17(5):166-169.

[7] 张 辉,赵郁亮,徐 江,等. 基于 Oracle 数据库海量数据的查询优化研究 [J]. 计算机技术与发展,2012,22(2):165-167.

[8] 林 昕,李心科. 一种 OLAP 海量数据载入技术的研究 [J]. 计算机技术与发展,2008,18(2):51-54.

[9] 赵国栋. 大数据专题: 大数据时代的三大发展趋势及投资方向 [R]. 上海: 国金证券股份有限公司,2012.

[10] Fact Sheet: Big Data Across the Federal Government [R]. USA: Executive Office of the President,2012.

[11] Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments [R]. USA: Executive Office of the President,2012.

(上接第 150 页)

据传输的可靠性具有明显效果。

4 结束语

文中通过对 WSN 节点性能的分析,构造出基本的对等设备之间数据传输模型。在此基础上分别对传统方式与网络编码情况下数据可靠性进行分析。通过分析与仿真表明,采用网络编码可以提高数据传输的可靠性。在高斯白噪声的信道中,信噪比越小,网络编码的优势愈加明显。

参考文献:

[1] 孙利民,李建中,陈 渝,等. 无线传感器网络 [M]. 北京: 清华大学出版社,2005.

[2] Ahlswede R, Cai N, Li S Y R, et al. Network Information Flow [J]. IEEE Trans. on Inform. Theory, 2000, 46(1): 1204-1216.

[3] Chen Yu-Hsun, Chen Gen-Huey, Wu E H. Multiple Trees with Network Coding for Efficient and Reliable Multicast in MANETs [C]//2010 39th International Conference on Parallel Processing Workshops. [s.l.]: [s.n.], 2010: 581-585.

[4] Chi Kaikai, Jiang Xiaohong, Ye Baoliu, et al. Efficient network coding-based loss recovery for reliable multicast in wireless network [J]. IEICE Trans. on Communication, 2010, E93B(4): 971-981.

[5] 孙 敏. 无线自组织网络中基于网络编码的可靠中继多播方案 [J]. 计算机系统应用,2011,20(5):60-63.

[6] 许胤龙,詹 成,罗 文,等. Ad hoc 网络中基于网络编码的可靠组播 [J]. 中国科学技术大学学报,2008,38(7):860-869.

[7] 李盼盼,洪佩琳. 基于流的无线网络编码 [J]. 通信技术,2009,42(9):148-153.

[8] 唐文胜,王 威,罗 娟,等. WSN 中基于网络编码的可靠传输算法 [J]. 湖南师范大学自然科学学报,2008,31(1):59-64.

[9] Li S Y, Yeung R W, Cai N. Linear network coding [J]. IEEE Trans. on Information Theory, 2003, 49(2): 371-381.

[10] Nguyen D, Nguyen T, Bose B. Wireless broadcasting using network coding [R]. Oregon: Oregon State University, 2006.

[11] Ghaderi M, Towsley D, Kurose J. Reliability Gain of Network Coding in Lossy Wireless Networks [C]//IEEE 27th Conference on Computer Communications. [s.l.]: [s.n.], 2008: 2171-2179.

[12] 陈 敏. OPNET 网络仿真 [M]. 北京: 清华大学出版社, 2004.

[13] 樊昌信,曹丽娜. 通信原理 [M]. 北京: 国防工业出版社, 2008.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net