



大数据研究 :未来科技 及经济社会发展的重大战略领域 大数据的研究现状与科学思考^{*}

文 / 李国杰 程学旗
中国科学院计算技术研究所 北京 100190

【摘要】 近来,大数据引起了产业界、科技界和政府部门的高度关注。本文简要阐述了大数据的研究现状与重大意义,探讨了大数据的科学问题,介绍了大数据应用与研究所面临的问题与挑战。最后,对大数据发展战略提出了几点建议。

【关键词】 大数据,数据科学,数据工程,第四范式

DOI 10.3969/j.issn.1000-3045.2012.06.001



中国科学院

近年来,大数据引起了产业界、科技界和政府部门的高度关注。2012年3月22日,奥巴马宣布美国政府投资2亿美元启动大数据研究和发展计划(Big Data Research and Development Initiative)。这是继1993年美国宣布信息高速公路计划后的又一次重大科技发展部署。美国政府认为,大数据是未来的新石油,并将对大数据的研究上升为国家意志,这对未来的科技与经济发展必将带来深远影响。

1 何谓大数据

人、机、物三元世界的高度融合引发了数据规模的爆炸式增长和数据模式的高度复杂化,世界已进入网络化的大数据(Big Data)时代^[1,16]。以数据为中心的传统学科(如基因组学、蛋白组学、天体物理学和脑科

学等)的研究产生了越来越多的数据。例如,用电子显微镜重建大脑中的突触网络,1立方毫米大脑的图像数据就超过1PB。但近年来大数据的飙升主要还是来自日常生活,特别是互联网公司的服务。据著名咨询公司IDC的统计,2011年全球被创建和复制的数据总量为1.8ZB(10的21次方),其中75%来自于个人(主要是图片、视频和音乐),远远超过人类有史以来所有印刷材料的数据总量(200PB)^[11]。Google公司通过大规模集群和MapReduce软件,每月处理的数据量超过400PB;百度每天大约要处理几十PB数据;Facebook注册用户超过10亿,每月上传的照片超过10亿张,每天生成300TB以上的日志数据;淘宝网会员超过3.7亿,在线商品超过8.8亿,每天交易数千万笔,产生约20TB数据。传感网和物联网的蓬勃发展是大数据的又一推动力,各个城

^{*} 修改稿收到日期 2012年11月12日

市的视频监控每时每刻都在采集巨量的流媒体数据。工业设备的监控也是大数据的重要来源。例如,劳斯莱斯公司对全世界数以万计的飞机引擎进行实时监控,每年传送PB数量级的数据。

一般意义上,大数据是指无法在可容忍的时间内用传统IT技术和硬件工具对其进行感知、获取、管理、处理和服务的数据集合。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)。首先,数据集合的规模不断扩大,已从GB到TB再到PB级,甚至开始以EB和ZB来计数。IDC的研究报告称,未来10年全球大数据将增加50倍,管理数据仓库的服务器数量将增加10倍^[11]。其次,大数据类型繁多,包括结构化数据、半结构化数据和非结构化数据。现代互联网应用呈现出非结构化数据大幅增长的特点,至2012年末,非结构化数据占有比例将达到整个数据量的75%以上。同时,由于数据显性或隐性的网络化存在,使得数据之间的复杂关联无所不在。再次,大数据往往以数据流的形式动态、快速地产生,具有很强的时效性,用户只有把握好对数据流的掌控才能有效利用这些数据。另外,数据自身的状态与价值也往往随时空变化而发生演变,数据的涌现特征明显。最后,虽然数据的价值巨大,但是基于传统思维与技术,人们在实际环境中往往面临信息泛滥而知识匮乏的窘态,大数据的价值利用密度低。

2 大数据已引起高度关注

毫无疑问,大数据隐含着巨大的社会、经济、科研价值,已引起了各行各业的高度重视^[14,15,17]。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。著名的O'Reilly公司断言:数据是下一个Intel Inside,未来属于将数据转换成产品的公司和人们。

IBM、Oracle、Microsoft、Google、Amazon、

Facebook等跨国巨头是发展大数据处理技术的主要推动者。自2005年以来,IBM投资160亿美元进行了30次与大数据有关的收购,促使其业绩稳定高速增长。2012年,IBM股价突破200美元大关,3年之内股价翻了3倍。华尔街早就开始招聘精通数据分析的天文学家和理论数学家来设计金融产品。IBM现在是全球数学博士的最大雇主,数学家正在将其数据分析的才能应用于石油勘探、医疗健康等各个领域。eBay通过数据挖掘可精确计算出广告中的每一个关键字为公司带来的回报。通过对广告投放的优化,2007年以来eBay产品销售的广告费降低了99%,而顶级卖家占总销售额的百分比却上升至32%。目前推动大数据研究的动力主要是企业经济效益,巨大的经济利益驱使大企业不断扩大数据处理规模^[14,15,17]。

近几年,Nature和Science等国际顶级学术刊物相继出版专刊来专门探讨对大数据的研究^[6-9]。2008年Nature出版专刊Big Data^[6],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海量数据带来的挑战。2011年Science推出关于数据处理的专刊Dealing with data^[7],讨论了数据洪流(Data Deluge)所带来的挑战,特别指出,倘若能够更有效地组织和使用这些数据,人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用。2012年4月欧洲信息学与数学研究协会会刊ERCIM News出版专刊Big Data^[9],讨论了大数据时代的数据管理、数据密集型研究的创新技术等问题,并介绍了欧洲科研机构开展的研究活动和取得的创新性进展。在这样的大背景下,2012年5月,香山科学会议组织了以“大数据科学与工程——一门新兴的交叉学科?”为主题的第424次学术讨论会,来自国内外35个单位横跨IT、经济、管理、社会、生物等多个不同学科领域的43位专家代表参会,并就大数据的理论与工程技术研究、应用方向以及大数据研究的组织方式与资源支持形式等重要问题进行了深入讨论。6月,中国计算机学会青年计算机科技论坛

(CCF YOCSEF)举办了 大数据时代 ,智谋未来 学术报告会 ,就大数据时代的数据挖掘、体系架构理论、大数据安全、大数据平台开发与大数据现实案例进行了全面的讨论。总体而言 ,大数据技术及相应的基础研究已经成为科技界的研究热点 ,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、心理学、经济学等诸多领域的新兴交叉学科方向正在逐步形成。

大数据同时也引起了包括美国在内的许多国家政府的极大关注。如前所述 ,2012年3月 ,美国公布了 大数据研发计划^[13]。该计划旨在提高和改进人们从海量和复杂的数据中获取知识的能力 ,进而加速美国在科学与工程领域发明的步伐 ,增强国家安全。根据该计划 ,美国国家科学基金会(NSF)、国立卫生研究院(NIH)、国防部(DOD)、能源部(DOE)、国防部高级研究计划局(DARPA)、地质勘探局(USGS)6个联邦部门和机构共同提高收集、储存、保留、管理、分析和共享海量数据所需的核心技术 ,扩大大数据技术开发和应用所需人才的供给。该计划还强调 ,大数据技术事关美国国家安全、科学和研究的步伐 ,将引发教育和学习的变革。欧盟方面也有类似的举措。过去几年欧盟已对科学数据基础设施投资1亿多欧元 ,并将数据信息化基础设施作为Horizon 2020计划的优先领域之一。2012年1月截止的预算为5 000万欧元的FP7 Call 8专门征集针对大数据的研究项目 ,仍以基础设施为先导^[9]。纵观国际形势 ,对大数据的研究与应用已引起各国政府的高度重视 ,并已成为重要的战略布局方向。

3 大数据研究的重大意义

大数据是与自然资源、人力资源一样重要的战略资源 ,是一个国家数字主权的体

现。大数据时代 ,国家层面的竞争力将部分体现为一国拥有大数据的规模、活性以及对数据的解释、运用的能力。一个国家在网络空间的数据主权将是继海、陆、空、天之后另一个大国博弈的空间。在大数据领域的落后 ,意味着失守产业战略制高点 ,意味着数字主权无险可守 ,意味着国家安全将出现漏洞。大数据将直接影响国家和社会稳定 ,是关系国家安全的战略性问题。因此 ,我国应尽快研究并制定我们国家的大数据战略。

大数据是现有产业升级与新产业诞生的重要推动力量。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处理能力 ,软件也将从编程为主转变为以数据为中心。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(AaaS)为主要标志的Cloud 2.0时代。采用大数据处理方法 ,生物制药、新材料研制生产的流程会发生革命性的变化 ,可以通过数据处理能力极高的计算机并行处理 ,同时进行大批量的仿真比较和筛选 ,大大提高科研和生产效率 ,甚至使整个行业迈入数字化与信息化的新阶段。数据已成为与矿物和化学元素一样的原始材料 ,未来可能形成数据服务、数据探矿、数据化学、数据材料、数据制药等一系列战略性的新兴产业。

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。最早的科学研究只有实验科学 ,随后出现了以研究各种定律和定理为特征的理论科学。由于理论分析方法在许多问题上过于复杂 ,难以解决实际问题 ,人们开始寻求模拟的方法 ,导致计算科学的



中国科学院

兴起。海量数据的出现催生了一种新的科研模式,即面对海量数据,科研人员只需从数据中直接查找或挖掘所需要的信息、知识和智慧,甚至无需直接接触需研究的对象。2007年,已故的图灵奖得主吉姆·格雷(Jim Gray)在他最后一次演讲中描绘了数据密集型科学研究的第四范式(The Fourth Paradigm)^[5],把数据密集型科学从计算科学中单独区分开来。格雷认为,要解决我们面临的某些最棘手的全球性挑战,第四范式可能是唯一具有系统性的方法。其实,第四范式不仅是科研方式的转变,也是人们思维方式的大变化。

4 对大数据研究的科学思考

4.1 数据科学研究的对象是什么?

计算机科学是关于算法的科学,数据科学是关于数据的科学。从事数据科学研究的学者更关注数据的科学价值,试图把数据当成一个自然体(Data nature)来研究,提出所谓数据界(Data universe)的概念,颇有把计算机科学划归为自然科学的倾向。但脱离各个领域的物理世界,作为客观事物间接存在形式的数据界究竟有什么共性问题还不清楚。物理世界在网络空间中有其数据映像,目前一些学者认为,数据界的规律其本质可能是物理世界的规律(还需要在物理世界中测试验证)。除去各个领域的规律,作为映像的数据界还有其独特的共同规律吗?这是一个值得深思的问题。

任何领域的研究,若要成为一门科学,一定是研究共性的问题。针对非常狭窄领域的某个具体问题,主要依靠该问题涉及的特殊条件和专门知识做数据挖掘,不大可能使大数据成为一门科学。数据研究能成为一门科学的前提是,在一个领域发现的数据相互关系和规律具有可推广到其他领域的普适性。抽象出一个领域的共性科学问题往往需要较长的时间,提炼数据界的共性科学问题还需要一段时间的实践积累。至少未来5~10年内计算机界的学者还需多花精力协助其

他领域的学者解决大数据带来的技术挑战问题。通过分层次的不抽象,大数据的共性科学问题才会逐步清晰明朗。

当前数据科学的目标还不很明确,但与其他学科一样,科学研究的道路常常是先做白盒研究,知识积累多了就有可能抽象出通用性较强的黑盒模型和普适规律。数据库理论是一个很好的例子。在经历了层次数据库、网状数据库多年实践后,Codd^[18]发现了数据库应用的共性规律,建立了有坚实理论基础的关系模型。在这之前人们也一直在问数据库可不可能有共性的理论。现在大数据研究要做的事就是提出像关系数据库这样的理论来指导海量非结构化数据的处理。

信息技术的发展使我们逐步进入人-机-物融合的三元世界,未来的世界可以做到机中有人,人中有机,物中有机,机中有物。所谓机就是联系人类社会(包括个人身体与大脑)与物理世界的网络空间,其最基本的构成元素是不同于原子和神经元的bit。物理空间和人类社会(包括人的大脑)都有共性的科学问题和规律,与这两者有密切联系的网络空间会不会有不同的共性科学问题?从人-机-物三元世界的角度来探讨大数据科学的共性问题,也许是一个可以尝试的突破口。

4.2 数据背后的共性问题 关系网络

观察各种复杂系统得到的大数据,直接反映的往往是一个个孤立的数据和分散的链接,但这些反映相互关系的链接整合起来就是一个网络。例如,基因数据构成基因网络,脑科学实验数据形成神经网络,Web数据反映出社会网络。数据的共性、网络的整体特征隐藏在数据网络中,大数据往往以复杂关联的数据网络这样一种独特的形式存在,因此要理解大数据就要对大数据后面的网络进行深入分析。网络有不少参数和性质,如平均路径长度、度分布、聚集系数、核数、介数等,这些性质和参数也许能刻画大数据背后网络的共性。因此,大数据面临的科学问题本质上可能就是网络科学问题,复杂网络分析应该是数据科学

的重要基石。

目前,研究 Web 数据的学者以复杂网络上的数据(信息)传播机理、搜索、聚类、同步和控制作为主要研究方向。最新的研究成果表明^[4] 随机的 Scale-free 网络不是一般的小世界,而是超小世界(Ultrasmall world),规模为 N 的网络的最短路径的平均长度不是一般小世界的 $\ln N$ 而是 $\ln \ln N$ 。网络数据研究应发现网络数据产生、传播以及网络信息涌现的内在机制,还要研究隐藏在数据背后的社会学、心理学、经济学的机理,同时利用这些机理研究互联网对政治、经济、文化、教育、科研的影响。基于大数据对复杂系统内在机理进行整体性的研究,也许将为研究复杂系统提供新的途径。从这种意义上看,数据科学是从整体上研究复杂系统的一门科学。

发现 Scale-free 网络的 Albert-László Barabási 教授在 2012 年 1 月的 *Nature Physics* 上发表一篇重要文章 *The network take-over*^[3]。文章认为 20 世纪是量子力学的世纪,从电子学到天文物理学,从核能到量子计算,都离不开量子力学;而到了 21 世纪,网络理论正在成为量子力学的可尊敬的后继,正在构建一个新的理论和算法的框架。

4.3 大数据研究中的关联关系与因果关系

大数据研究不同于传统的逻辑推理研究,而是对数量巨大的数据做统计性的搜索、比较、聚类、分类等分析归纳,因此继承了统计科学的一些特点。统计学关注数据的相关性或称关联性,所谓相关性是指两个或两个以上变量的取值之间存在某种规律性。相关分析的目的是找出数据集里隐藏的相互关系网(关联网),一般用支持度、可信度、兴趣度等参数反映相关性。两个数据 A 和 B 有相关性,只有反映 A 和 B 在取值时相互有影响,并不能告诉我们有 A 就一定

有 B,或者反过来有 B 就一定有 A。严格来讲,统计学无法检验逻辑上的因果关系。如根据统计结果可以说吸烟的人群肺癌发病率会比不吸烟的人群高几倍,但统计结果无法得出吸烟致癌的逻辑结论。统计学的相关性有时可能会产生把结果当成原因的错觉。如统计结果表明:下雨之前常见到燕子低飞,从时间先后看两者的关系可能得出燕子低飞是下雨的原因,而事实上,将要下雨才是燕子低飞的原因。

也许正是因为统计方法不能致力于寻找真正的原因,才促使数据挖掘和大数据技术在商业领域广泛流行。企业的目标是多赚钱,只要从数据挖掘中发现某种措施与增加企业利润有较强的相关性,采取这种措施就是了,不必深究为什么能增加利润,更不必发现其背后的内在规律和模型。一般而言,企业收集和处理大数据,不是按学者们经常描述的从数据到信息再到知识和智慧的研究思路,而是走从数据直接到价值的捷径。Google 广告获得巨额收入经常被引用作为大数据相关分析的成功案例,美国 *Wired* 杂志主编 Chris Anderson 在他的著名文章 *The End of Theory* 的结尾发问:现在是时候问这一句了,科学能从谷歌那儿学到什么?^[2]

因果关系的研究曾引发了科学体系的建立,近代科学体系获得的成就已经证明,科学是研究因果关系最重要的手段。相关性研究是可以替代因果分析的科学新发展,还只是因果分析的补充,不同的学者有完全不同的看法。我们都是从做平面几何证明题开始进入科学大花园的,脑子里固有的逻辑思维模式少不了因果分析,判断是否是真理也习惯看充分必要条件,对于大数据的关联分析蕴含的科学意义往往理解不深。对于简单封闭的系统,基于小数据的因果分析



中国科学院

容易做到。当年开普勒发现行星三大定律,牛顿发现力学三大定律都是基于小数据。但对于开放复杂的巨系统,传统的因果分析难以奏效,因为系统中各个组成部分之间相互有影响,可能互为因果,因果关系隐藏在整个系统之中。现在的因果可能是过去的果,此处的果也可能是别处的因,因果关系本质上是一种相互纠缠的相关性。在物理学的基本粒子理论中,颇受重视的欧几里德量子引力学(霍金所倡导的理论)本身并不包括因果律。因此,对于大数据的关联分析是不是知其然而不知其所以然,其中可能包含深奥的哲理,不能贸然下结论。

4.4 社会科学的大数据研究

根据数据的来源,大数据可以粗略地分成两大类:一类来自物理世界,另一类来自人类社会。前者多半是科学实验数据或传感数据,后者与人的活动有关系,特别是与互联网有关。这两类数据的处理方式和目标差别较大,不能照搬处理科学实验数据的方法来处理Web数据。

科学实验是科技人员设计的,如何采集数据、处理数据事先都已想好了,不管是检索还是模式识别,都有一定的科学规律可循。美国的大数据研究计划中专门列出寻找希格斯粒子(被称为上帝粒子)的大型强子对撞机(LHC)实验。这是一个典型的基于大数据的科学实验,至少要在1万亿个事例中才可能找出1个希格斯粒子。2012年7月4日,CERN宣布发现新的玻色子,标准差为4.9,被认为可能是希格斯玻色子(承认是希格斯玻色子粒子需要5个标准差,即99.99943%的可能性是对的)^[12]。设计这一实验的激动人心之处在于,不论找到还是没有找到希格斯粒子,都是物理学的重大突破。从这一实验可以看出,科学实验的大数据处理是整个实验的一个预定步骤,发现有价值的信息往往在预料之中。

Web上的信息(譬如微博)是千千万万的人随机产生的,从事社会科学研究的学者要从这些看似杂乱无章的数据中寻找有价值的蛛丝马迹。网

络大数据有许多不同于自然科学数据的特点,包括多源异构、交互性、时效性、社会性、突发性和高噪声等,不但非结构化数据多,而且数据的实时性强,大量数据都是随机动态产生。科学数据的采集一般代价较高,LHC实验设备花了几十亿美元,因此对采集什么数据要做精心安排。而网络数据的采集相对成本较低,网上许多数据是重复的或者没有价值,价值密度很低。一般而言,社会科学的大数据分析,特别是根据Web数据做经济形势、安全形势、社会群体事件的预测,比科学实验的数据分析更困难。

未来的任务主要不是获取越来越多的数据,而是数据的去冗分类、去粗取精,从数据中挖掘知识。几百年来,科学研究一直在做从薄到厚的事情,把小数据变成大数据,现在要做的事情是从厚到薄,要把大数据变成小数据。要在不明显增加采集成本的前提下尽可能提高数据的质量。要研究如何科学合理地抽样采集数据,减少不必要的数据采集。两岁的小孩学习识别动物和汽车等,往往几十张样本图片就足够了,研究清楚人类为什么具有小数据学习能力,对开展大数据分析研究具有深刻的指导意义。

近10年来增长最快的数据是网络上传播的各种非结构化或半结构化的数据。网络数据的背后是相互联系的各种人群,网络大数据的处理能力直接关系到国家的信息空间安全和社会稳定^[10]。从心理学、经济学、信息科学等不同学科领域共同探讨网络数据的产生、扩散、涌现的基本规律,是建立安全和谐的网络环境的重大战略需求,是促使国家长治久安的大事。我国拥有世界上最多的网民和最大的访问量,在网络大数据分析方面已有较强的基础,有望做出世界领先的原始创新成果,应加大网络大数据分析方面的研究力度。

4.5 数据处理的复杂性研究

计算复杂性是计算机科学的基本问题,科学计算主要考虑时间复杂性和空间复杂性。对于大数据处理,除了时间和空间复杂性外,可能还需要

考虑解决一个问题需要多大的数据量,暂且称为“数据量复杂性”。数据量复杂性和空间复杂性不是一个概念,空间复杂性要考虑计算过程中产生的空间需求。

设想有人采集完全随机地抛掷硬币的正反面数据,得到极长的01数字序列,通过统计可计算出出现正面的比例。可以肯定,收集的数据越多,其结果与0.5的误差越小,这是一个无限渐进的过程。基于唯象假设的数据处理常出现这类增量式进步,数据多一点,结果就好一点。这类问题的数据科学价值可能不大。反过来,可能有些问题的数据处理像个无底洞,无论多少数据都不可能解决问题。这种问题有些类似NP问题。我们需要建立一种理论,对求解一个问题达到某种满意程度(对判定问题是有多大把握说是或否,优化问题是接近最优解的程度)需要多大规模的数据量给出理论上的判断。当然,目前还有很多问题没有定义清楚,比如,对于网络搜索之类的问题,如何定义问题规模和数据规模等。

对从事大数据研究的学者而言,最有意思的问题应该是,解决一个问题的数据规模有一个阈值。数据少于这个阈值,问题解决不了;达到这个阈值,就可以解决以前解决不了的大问题;而数据规模超过这个阈值,对解决问题也没有更多的帮助。我们把这类问题称为“预言性数据分析问题”,即在做大数据处理之前,我们可以预言,当数据量到达多大规模时,该问题的解可以达到何种满意程度。

与社会科学有关的大数据问题,例如舆情分析、情感分析等,许多理论问题过去没有考虑过,才刚刚开始研究。迫切需要计算机学者与社会科学领域的学者密切合作,共同开拓新的疆域。借助大数据的推力,社会科学将脱下“准科学”的外衣,真正迈进科学

的殿堂。

4.6 科研第四范式是思维方式的大变化

已故图灵奖得主吉姆·格雷提出的数据密集型科研第四范式(the fourth paradigm),将大数据科研从第三范式(计算科学)中分离出来单独作为一种科研范式,是因为其研究方式不同于基于数学模型的传统研究方式^[5]。Google公司的研究部主任Peter Norvig的一句名言可以概括两者的区别:所有的模型都是错误的,进一步说,没有模型你也可以成功(All models are wrong, and increasingly you can succeed without them)^[2]。PB级数据使我们可以做到没有模型和假设就可以分析数据。将数据丢进巨大的计算机机群中,只要有相互关系的数据,统计分析算法可以发现过去的科学方法发现不了的新模式、新知识甚至新规律。实际上,Google的广告优化配置、战胜人类的IBM沃森问答系统都是这么实现的,这就是第四范式的魅力!

美国Wired杂志主编Chris Anderson 2008年曾发出“理论已终结”的惊人断言:“数据洪流使(传统)科学方法变得过时(The Data Deluge Makes the Scientific Method Obsolete)^[2]。他指出,获得海量数据和处理这些数据的统计工具的可能性提供了理解世界的一条完整的新途径。Petabytes让我们说:相互关系已经足够(Correlation is enough)。我们可以停止寻找模型,相互关系取代了因果关系,没有具有一致性的模型、统一的理论和任何机械式的说明,科学也可以进步。

Chris Anderson的极端看法并没有得到科学界的普遍认同,数据量的增加能否引起科研方法本质性的改变仍然是一个值得探讨的问题。对研究领域的深刻理解(如空气动力学方程用于风洞实验)和数据量的积累



中国科学院

应是一个迭代累进的过程。没有科学假设和模型就能发现新知识究竟有多大的普适性也需要实践来检验,我们需要思考:这类问题有多大的普遍性?这种优势是数据量特别大带来的还是问题本身有这种特性?所谓从数据中获取知识要不要人的参与,人在机器自动学习和运行中应该扮演什么角色?也许有些领域可以先用第四范式,等领域知识逐步丰富了再过渡到第三范式。

5 面临的主要问题与挑战

现有的数据中心技术很难满足大数据的需求,需要考虑对整个IT架构进行革命性的重构。而存储能力的增长远远赶不上数据的增长,因此设计最合理的分层存储架构已成为IT系统的关键。数据的移动已成为IT系统最大的开销,目前传送大数据最高效也最实用的方式是通过飞机或地面交通工具运送磁盘而不是网络通信。在大数据时代,IT系统需要从数据围着处理器转改变为处理能力围着数据转,将计算推送给数据,而不是将数据推送给计算。大数据也导致高可扩展性成为对IT系统最本质的需求,并发执行(同时执行的线程)的规模要从现在的千万量级提高到10亿级以上。

在应对处理大数据的各种技术挑战中,以下几个问题值得高度重视:

(1)大数据的去冗降噪技术。大数据一般都来自多个不同的源头,而且往往以动态数据流的形式产生。因此,大数据中常常包含有不同形态的噪声数据。另外,数据采样算法缺陷与设备故障也可能导致大数据的噪声。大数据的冗余则通常来自两个方面:一方面,大数据的多源性导致了不同源头的数据中存在有相同的数据,从而造成数据的绝对冗余;另一方面,就具体的应用需求而言,大数据可能会提供超量特别是超精度的数据,这又形成数据的相对冗余。降低噪声、消除冗余是提高数据质量、降低数据存储成本的基础;

(2)大数据的新型表示方法。目前表示数据

的方法,不一定能直观地展现出大数据本身的意义。要想有效利用数据并挖掘其中的信息或知识,必须找到最合适的数据表示方法。在一种不合适的数据表示中寻找大数据的固定模式、因果关系和关联关系时,可能会落入固有的偏见之中。数据表示方法和最初的数据产生者有着密切关系。如果原始数据有必要的标识,就会大大减轻事后数据识别和分类的困难。但标识数据会给用户增添麻烦,所以往往得不到用户认可。研究既有效又简易的数据表示方法是处理网络大数据必须解决的技术难题之一;

(3)高效率低成本的大数据存储。大数据的存储方式不仅影响其后的数据分析处理效率也影响数据存储的成本。因此,就需要研究高效率低成本的数据存储方式。具体则需要研究多源多模态数据高质量获取与整合的理论和技術、流式数据的高速索引创建与存储、错误自动检测与修复的理论和技術、低质量数据上的近似计算的理论和算法等;

(4)大数据的有效融合。数据不整合就发挥不出大数据的大价值。大数据的泛滥与数据格式太多有关。大数据面临的一个重要问题是个人、企业和政府机构的各种数据和信息能否方便地融合。如同人类有许多种自然语言一样,作为网络空间中唯一客观存在的数据难免有多种格式。但为了扫清网络大数据处理的障碍,应研究推广不与平台绑定的数据格式。大数据已成为联系人类社会、物理世界和网络空间的纽带,需要通过统一的数据格式构建融合人、机、物三元世界的统一信息系统;

(5)非结构化和半结构化数据的高效处理。据统计,目前采集到的数据85%以上是非结构化和半结构化数据,而传统的关系数据库技术无法胜任这些数据的处理,因为关系数据库系统的出发点是追求高度的数据一致性和容错性。根据CAP (Consistency, Availability, tolerance to network Partitions) 理论,在分布式系统中,一致性、可

用性、分区容错性三者不可兼得,因而并行关系数据库必然无法获得较强的扩展性和良好的系统可用性。系统的高扩展性是大数据分析最重要的需求,必须寻找高扩展性的数据分析技术。以 MapReduce 和 Hadoop 为代表的非关系数据分析技术,以其适合非结构数据处理、大规模并行处理、简单易用等突出优势,在互联网信息搜索和其他大数据分析领域取得了重大进展,已成为大数据分析的主流技术。MapReduce 和 Hadoop 在应用性能等方面还存在不少问题,还需要研究开发更有效、更实用的大数据分析和管理工作技术;

(6)适合不同行业的大数据挖掘分析工具和开发环境。不同行业需要不同的大数据分析工具和开发环境,应鼓励计算机算法研究人员与各领域的科研人员密切合作,在分析工具和开发环境上创新。当前跨领域跨行业的数据共享仍存在大量壁垒,海量数据的收集,特别是关联领域的同时收集还存在很大挑战。只有跨领域的数据分析才更有可能形成真正的知识和智能,产生更大的价值;

(7)大幅度降低数据处理、存储和通信能耗的新技术。大数据的获取、通信、存储、管理与分析处理都需要消耗大量的能源。在能源问题日益突出的今天,研究创新的数据处理和传送的节能方法与技术是重要的研究方向。

6 建议和举措

尽管大数据意味着大机遇,但同时也意味着工程技术、管理政策、人才培养等方面的大挑战。只有解决了这些基础性的挑战问题,才能充分利用这个大机遇,得到大数据的大价值。因此,我国亟需在国家层面对大数据给予高度重视,特别需要从政策制

定、资源投入、人才培养等方面给予强有力的支持;另一方面,建立良性的大数据生态环境是有效应对大数据挑战的唯一出路,需要科技界、工业界以及政府部门在国家政策的引导下共同努力,通过消除壁垒、成立联盟、建立专业组织等途径,建立和谐的大数据生态系统。

就大数据研究计划与措施,我们有如下的建议:

6.1 优先支持网络大数据研究

大数据涉及物理、生物、脑科学、医疗、环保、经济、文化、安全等众多领域。网络空间中的数据是大数据的重要组成部分,这类大数据与人的活动密切相关,因此也与社会科学密切相关。而网络数据科学和工程是信息科学技术与社会科学等多个不同领域高度交叉的新型学科方向,对国家的稳定与发展有独特的作用,因此应特别重视与支持网络大数据的研究。大数据涉及应用领域很广,当前大数据的研究应与国计民生密切相关的科学决策、环境与社会管理、金融工程、应急管理(如疾病防治、灾害预测与控制、食品安全与群体事件)以及知识经济为主要应用领域。

6.2 大数据科学的基础研究

无论是国外政府的大数据研究计划,还是国内外大公司的大数据研发,当前最重视的都是大数据分析算法和大数据系统的效率。因此,当工业界把主要精力放在应对大数据的工程技术挑战的时候,科技界应开始着手关注大数据的基础理论研究。大数据科学作为一个新兴的交叉学科方向,其共性理论基础将来自多个不同的学科领域,包括计算机科学、统计学、人工智能、社会科学等。因此,大数据的基础研究离不开对相关学科的领域知识与研究方法论的借鉴。在大数据的基础研究方面,建议研究大数据的



中国科学院

内在机理,包括大数据的生命周期、演化与传播规律,数据科学与社会学、经济学等之间的互动机制,以及大数据的结构与效能的规律性(如社会效应、经济效应等)。在大数据计算方面,研究大数据表示、数据复杂性以及大数据计算模型。在大数据应用基础理论方面,研究大数据与知识发现(学习方法、语义解释),大数据环境下的实验与验证方法,以及大数据的安全与隐私等。

6.3 大数据研究的组织方式

2012年10月,中国计算机学会和中国通信学会各自成立了大数据专家委员会,从行业学会的层面来组织和推动大数据的相关产学研用活动。但这还不够,建议中科院、科技部、基金委共同推动成立一个组织机构,建立一个大数据科学研究平台,更好地组织大数据的协同创新研究与战略性应用,成立国家级的行业大数据共享联盟,使产业界、科技界以及政府部门都能够参与进来,一方面为学术研究提供基本的数据资源,另一方面为大数据的应用提供理论与技术支持。此外,还需成立国家级的面向大数据研究与应用的开源社区,同时也向国际开源社区的核心团队举荐核心成员,使国际顶级的开源社区能够听到来自中国的声音。

6.4 大数据研究的资源支持

在资源支持方面,建议启动中国大数据科学与工程研究计划,从宏观上对我国的大数据产学研用做出系统全面的短期与长期规划。设立自然科学重大研究计划(基金重大)以及重大基础科学研究项目群(973项目群或863重大项目)等专项资金,有针对性地资助有关大数据的重大科研活动。此外,国家在大数据平台的构建、典型行业的应用以及研发人才的培养等方面应提供相应的财力、物力与人力支持。

主要参考文献

1 李国杰. 大数据研究的科学价值. 中国计算机学会通讯, 2012,

8(9) 8-15.

- 2 Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired, 2008, 16(7).
- 3 Albert-László Barabási. The network takeover. Nature Physics, 2012, 8(1): 14-16.
- 4 Reuven Cohen, Shlomo Havlin. Scale-Free Networks Are Ultrasmall. Physical Review Letters, 2003, 90,(5).
- 5 Tony Hey, Stewart Tansley, Kristin Tolle (Editors). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft, 2009, October 16.
- 6 Big Data. Nature, 2008, 455(7 209): 1-136.
- 7 Dealing with data. Science, 2011, 331(6 018): 639-806.
- 8 Complexity. Nature Physics, 2012, 8(1).
- 9 Big Data. ERCIM News, 2012, (89).
- 10 David Lazer, Alex Pentland, Lada Adamic et al. Computational Social Science. Science, 2009, 323(5 915): 721-723.
- 11 The 2011 Digital Universe Study: Extracting Value from Chaos. International Data Corporation and EMC, June 2011.
- 12 CERN experiments observe particle consistent with long-sought Higgs boson. CERN press release, July 4, 2012.
- 13 Tom Kalil. Big Data is a Big Deal, March 29, 2012. Available at: <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>.
- 14 Divyakant Agrawal, Philip Bernstein, Elisa Bertino et al. Challenges and Opportunities with Big Data, Cyber Center Technical Reports, February 2012. Available at: <http://docs.lib.purdue.edu/cctech/1>.
- 15 James Manyika, Michael Chui, Brad Brown et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.
- 16 Steve Lohr. The Age of Big Data. New York Times, February 11, 2012.
- 17 Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release, July 2011.
- 18 Codd E F. A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, 1970, 13(6): 377-387.

Research Status and Scientific Thinking of Big Data

Li Guojie Cheng Xueqi

(Institute of Computing Technology, CAS 100190 Beijing)

Abstract Recently, big data attract great attention from industries, academia, and governments. This paper briefly describes the research status and significance of big data and discusses the corresponding scientific research issues. It further introduces the problems and challenges faced by big data study. Finally, a few suggestions are presented on the research and applications of big data.

Keywords Big Data ,Data Science ,Data Engineering ,The Fourth Paradigm

李国杰 中国工程院院士 ,世界科学院院士 ,中科院计算所首席科学家、研究员、博士生导师。1985 年获美国普渡大学博士学位。主要从事计算机体系结构、并行算法、人工智能、计算机网络等方面的研究 ,发表论文 100 余篇 ,合著英文专著 4 本。曾任中科院计算技术所所长。现任中国计算机学会名誉理事长、国家信息化专家咨询委员会信息技术与新兴产业专委会副主任、中科院学位委员会副主席、国科大计算机与控制学院院长等职。
E-mail: lig@ict.ac.cn

程学旗 中科院计算所副总工程师、网络数据科学与技术实验室主任、研究员、博士生导师。1994、1996 年分别获东北大学计算机科学与技术专业学士与硕士学位 ,2006 年获中科院计算所计算机系统结构博士学位。在信息网络建模与社区结构分析、Web 搜索与挖掘等领域发表学术论文 130 余篇 ,获授权发明专利 10 项 软件著作权 19 项。现任中国计算机学会大数据专委会秘书长、中文信息学会信息检索与内容安全专委会常务副主任、国家信息安全专项计划管理专家。 E-mail: cxq@ict.ac.cn



中国科学院