

大数据的概念、特征及其应用*

马建光, 姜 巍

(国防科技大学人文与社会科学学院, 湖南 长沙 410074)

【摘 要】 随着互联网的飞速发展, 特别是近年来随着社交网络、物联网、云计算以及多种传感器的广泛应用, 以数量庞大, 种类繁多, 时效性强为特征的非结构化数据不断涌现, 数据的重要性愈发凸显, 传统的数据存储、分析技术难以实时处理大量的非结构化信息, 大数据的概念应运而生。如何获取、聚集、分析大数据成为广泛关注的热点问题。介绍大数据的概念与特点, 分别讨论大数据的典型特征, 分析大数据要解决的相关性分析、实时处理等核心问题, 最后讨论大数据可能要面临的多种挑战。

【关键词】 大数据; 非结构化信息; 解决核心问题; 未来挑战

【中图分类号】 E924.2 **【文献标识码】** A **【文章编号】** 1671-4547 (2013) 02-0010-08

一、引言

自上古时代的结绳记事起, 人类就开始用数据来表征自然和社会, 伴随着科技和社会的发展进步, 数据的数量不断增多, 质量不断提高。工业革命以来, 人类更加注重数据的作用, 不同的行业先后确定了数据标准, 并积累了大量的结构化数据, 计算机和网络的兴起, 大量数据分析、查询、处理技术的出现使得高效的处理大量的传统结构化数据成为可能。而近年来, 随着互联网的快速发展, 音频、文字、图片视频等半结构化、非结构化数据大量涌现, 社交网络、物联网、云计算广泛应用, 使得个人可以更加准确快捷的发布、获取数据。在科学研究、互联网应用、电子商务等诸多应用领域, 数据规模、数据种类正在以极快的速度增长, 大数据时代已悄然降临。

首先, 全球数据量出现爆炸式增长, 数据成了当今社会增长最快的资源之一。根据国际数据公司 IDC 的监测统计^[1], 即使在遭遇金融危机的 2009 年, 全球信息量也比 2008 年增长了 62%, 达到 80 万 PB (1PB 等于 10 亿 GB), 到 2011 年全球数据总量已经达到 1.8ZB (1ZB 等于 1 万亿

GB), 并且以每两年翻一番的速度飞速增长, 预计到 2020 年全球数据量总量将达到 40 ZB, 10 年间增长 20 倍以上, 到 2020 年, 地球上人均数据预计将达 5247GB。在数据规模急剧增长的同时, 数据类型也越来越复杂, 包括结构化数据、半结构化数据、非结构化数据等多种类型, 其中采用传统数据处理手段难以处理的非结构化数据已接近数据总量的 75%。

如此增长迅速、庞大繁杂的数据资源, 给传统的数据分析、处理技术带来了巨大的挑战。为了应对这样的新任务, 与大数据相关的大数据技术、大数据工程、大数据科学和大数据应用等迅速成为信息科学领域的热点问题, 得到了一些国家政府部门、经济领域以及科学领域有关专家的广泛关注。2012 年 3 月 22 日, 奥巴马宣布美国政府五大部门投资 2 亿美元启动“大数据研究和发展计划 (Big Data Research and Development Initiative)”^[2], 欲大力推动大数据相关的收集、储存、保留、管理、分析和共享海量数据技术研究, 以提高美国的科研、教育与国家安全能力。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信息时代的重要资源, 战略地位堪比

* 【收稿日期】 2012-04-12

【作者简介】 马建光, 男, 教授, 国防科技大学;
姜 巍, 男, 博士研究生

工业时代的石油,其影响除了体现在科技、经济方面,同时也对政治、文化等方面产生深远的影响。在商业方面,2013年,Gartner发布了将在未来三年对企业产生重大影响的十大战略技术中,大数据名列其中,提出大数据技术将影响企业的长期计划、规划和行动方案,同时,IBM、Intel、EMC、Walmart、Teradata、Oracle、Microsoft、Google、Facebook等发源于美国的跨国巨头也积极提出自己的应对大数据挑战的发展策略,他们成了发展大数据处理技术的主要推动者。在科技领域,庞大的数据正在改变着人类发现问题、解决问题的基本方式,采用最简单的统计分析算法,将大量数据不经过模型和假设直接交给高性能计算机处理,就可以发现某些传统科学方法难以得到的规律和结论。图灵奖得主吉姆·格雷提出的数据密集型科研第四范式^[3],不同于传统的实验、理论和计算三种范式,第四种范式不需要考虑因果关系,以数据为中心,分析数据的相关性,打破了千百年来从结果出发探究原因的科研模式,大规模的复杂数据使得新的科研模式成为可能。

虽然大数据日益升温,但与大多数信息学领域的问题一样,大数据的基本概念及特点,大数据要解决核心问题,目前尚无统一的认识,大数据的获取、存储、处理、分析等诸多方面仍存在一定的争议,大数据概念有过度炒作的嫌疑。欧洲的一些企业甚至认为大数据就是海量数据存储,仅将大数据视作是可以获取更多信息的平台。本文分析当前流行的几种大数据的概念,讨论其异同,从大数据据有的典型特征角度描述大数据的概念和特点,从整体上分析大数据要解决的相关性分析、实时处理等核心问题,在此基础上,最后讨论大数据可能要面临的多种挑战。

二、国内外开展的相关工作

近年来,大数据成为新兴的热点问题,在科技、商业领域得到了日益广泛的关注和研究,有一些相关的研究成果。早在1980年,阿尔文·托夫勒^[4]等人就前瞻性地指出过大数据时代即将到来。此后经过几十年的发展,特别是移动互联网络和云计算的出现,人们逐渐认识到大数据的重大意义,国际顶级学术刊物相继出版大数据方

面的专刊,讨论大数据的特征、技术与应用,2008年Nature出版专刊“Big Data”^[5],分析了大量快速涌现数据给数据分析处理带来的巨大挑战,大数据的影响遍及互联网技术、电子商务、超级计算、环境科学、生物医药等多个领域。2011年Science推出关于数据处理的专刊“Dealing with data”^[6],讨论了数据洪流(Data Deluge)所带来的挑战,提出了对大数据进行有效的分析、组织、利用可以对社会发展起到巨大推动作用。在大数据领域,国内学者也有大量的相关工作,李国杰等人^[7]阐述了大数据的研究现状与意义,介绍了大数据应用与研究所面临的问题与挑战并对大数据发展战略提出了建议。文献^[8-10]主要关注大数据分析、查询方面的理论、技术,对大数据基本概念进行了剖析,列举了大数据分析平台需要具备的几个重要特性,阐述了大数据处理的基本框架,并对当前的主流实现平台进行了分析归纳。随着大数据理念逐渐被大众了解,出现了一些阐述大数据基本概念与思想的专著,舍恩伯格等在大数据时代^[11]一书中用三个部分讲述了大数据时代的思维变革、商业变革和管理变革。近年来,大数据对经济的推动作用被广泛接受,出现了探讨大数据在商业领域的应用的文章和专著,Martin Klubeck等人在量化:大数据时代的企业管理^[12]一书中提到,进入大数据时代,数据发挥着关键的作用,探讨了如何从空前膨胀的海量数据中挖掘出有用的指标和信息。朱志军等人所著的《转型时代丛书:大数据·大价值、大机遇、大变革》^[13]中介绍了大数据产生的背景、特征和发展趋势,从实证的角度探讨了它对社会和商业智能的影响,并认为大数据正影响着商业模式的转变,并将带来新的商业机会。

三、大数据的概念与特点

大数据是一个较为抽象的概念,正如信息学领域大多数新兴概念,大数据至今尚无确切、统一的定义。在维基百科中关于大数据的定义为^[14]:大数据是指利用常用软件工具来获取、管理和处理数据所耗时间超过可容忍时间的数据集。笔者认为,这并不是一个精确的定义,因为无法确定常用软件工具的范围,可容忍时间也是个概略的描述。IDC在对大数据作出的定义

为^[15]: 大数据一般会涉及 2 种或 2 种以上数据形式。它要收集超过 100TB 的数据, 并且是高速、实时数据流; 或者是从小数据开始, 但数据每年会增长 60% 以上。这个定义给出了量化标准, 但只强调数据量大, 种类多, 增长快等数据本身的特征。研究机构 Gartner 给出了这样的定义^[16]: 大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。这也是一个描述性的定义, 在对数据描述的基础上加入了处理此类数据的一些特征, 用这些特征来描述大数据。当前, 较为统一的认识是大数据有四个基本特征: 数据规模大 (Volume), 数据种类多 (Variety), 数据要求处理速度快 (Velocity), 数据价值密度低 (Value), 即所谓的四 V 特性。这些特性使得大数据区别于传统的数据概念。大数据的概念与“海量数据”不同, 后者只强调数据的量, 而大数据不仅用来描述大量的数据, 还更进一步指出数据的复杂形式、数据的快速时间特性以及对数据的分析、处理等专业化处理, 最终获得有价值信息的能力。

(一) 数据量大

大数据聚合在一起的数据量是非常大的, 根据 IDC 的定义至少要有超过 100TB 的可供分析的数据, 数据量大是大数据的基本属性。导致数据规模激增的原因有很多, 首先是随着互联网的广泛应用, 使用网络的人、企业、机构增多, 数据获取、分享变得相对容易, 以前, 只有少量的机构可以通过调查、取样的方法获取数据, 同时发布数据的机构也很有限, 人们难以短期内获取大量的数据, 而现在用户可以通过网络非常方便的获取数据, 同时用户在有意的分享和无意的点击、浏览都可以快速的提供大量数据; 其次是随着各种传感器数据获取能力的大幅提高, 使得人们获取的数据越来越接近原始事物本身, 描述同一事物的数据量激增。早期的单位化数据, 对原始事物进行了一定程度的抽象, 数据维度低, 数据类型简单, 多采用表格的形式来收集、存储、整理, 数据的单位、量纲和意义基本统一, 存储、处理的只是数值而已, 因此数据量有限, 增长速度慢而随着应用的发展, 数据维度越来越高, 描述相同事物所需的数据量越来越大。以当

前最为普遍的网络数据为例, 早期网络上的数据以文本和一维的音频为主, 维度低, 单位数据量小。近年来, 图像、视频等二维数据大规模涌现, 而随着三维扫描设备以及 Kinect 等动作捕捉设备的普及, 数据越来越接近真实的世界, 数据的描述能力不断增强, 而数据量本身必将以几何级数增长。此外, 数据量大还体现在人们处理数据的方法和理念发生了根本的改变。早期, 人们对事物的认知受限于获取、分析数据的能力, 一直利用采样的方法, 以少量的数据来近似的描述事物的全貌, 样本的数量可以根据数据获取、处理能力来设定。不管事物多么复杂, 通过采样得到部分样本, 数据规模变小, 就可以利用当时的技术手段来进行数据管理和分析, 如何通过正确的采样方法以最小的数据量尽可能分析整体属性成了当时的重要问题。随着技术的发展, 样本数目逐渐逼近原始的总体数据, 且在某些特定的应用领域, 采样数据可能远不能描述整个事物, 可能丢掉大量重要细节, 甚至可能得到完全相反的结论, 因此, 当今有直接处理所有数据而不是只考虑采样数据的趋势。使用所有的数据可以带来更高的精确性, 从更多的细节来解释事物属性, 同时必然使得要处理数据量显著增多。

(二) 数据类型多样

数据类型繁多, 复杂多变是大数据的重要特性。以往的数据尽管数量庞大, 但通常是事先定义好的结构化数据。结构化数据是将事物向便于人类和计算机存储、处理、查询的方向抽象的结果, 结构化在抽象的过程中, 忽略一些在特定的应用下可以不考虑的细节, 抽取了有用的信息。处理此类结构化数据, 只需事先分析好数据的意义以数据间的相关属性, 构造表结构来表示数据的属性, 数据都以表格的形式保存在数据库中, 数据格式统一, 以后不管再产生多少数据, 只需根据其属性, 将数据存储合适的位置, 就可以方便的处理、查询, 一般不需要为新增的数据显著的更改数据聚集、处理、查询方法, 限制数据处理能力的只是运算速度和存储空间。这种关注结构化信息, 强调大众化、标准化的属性使得处理传统数据的复杂程度一般呈线性增长, 新增的数据可以通过常规的技术手段处理。而随着互联网与传感器的飞速发展, 非结构化数据大量涌

现,非结构化数据没有统一的结构属性,难以用表结构来表示,在记录数据数值的同时还需要存储数据的结构,增加了数据存储、处理的难度。而时下在网络上流动着的数据大部分是非结构化数据,人们上网不只是看看新闻,发送电子邮件,还会上传下载照片、视频、发送微博等非结构化数据,同时,遍及工作、生活中各个角落的传感器也时刻不断的产生各种半结构化、非结构化数据,这些结构复杂,种类多样,同时规模又很大的半结构化、非结构化数据逐渐成为主流数据。如上所述,非结构化数据量已占到数据总量的75%以上,且非结构化数据的增长速度比结构化数据快10倍到50倍。在数据激增的同时,新的数据类型层出不穷,已经很难用一种或几种规定的模式来表征日趋复杂、多样的数据形式,这样的数据已经不能用传统的数据库表格来整齐的排列、表示。大数据正是在这样的背景下产生的,大数据与传统数据处理最大的不同就是重点关注非结构化信息,大数据关注包含大量细节信息的非结构化数据,强调小众化,体验化的特性使得传统的数据处理方式面临巨大的挑战。

(三) 数据处理速度快

要求数据的快速处理,是大数据区别于传统海量数据处理的重要特性之一。随着各种传感器和互联网络等信息获取、传播技术的飞速发展普及,数据的产生、发布越来越容易,产生数据的途径增多,个人甚至成为了数据产生的主体之一,数据呈爆炸的形式快速增长,新数据不断涌现,快速增长的数据量要求数据处理的速度也要相应的提升,才能使得大量的数据得到有效的利用,否则不断激增的数据不但不能为解决问题带来优势,反而成了快速解决问题的负担。同时,数据不是静止不动的,而是在互联网络中不断流动,且通常这样的数据的价值是随着时间的推移而迅速降低的,如果数据尚未得到有效的处理,就失去了价值,大量的数据就没有意义。此外,在许多应用中要求能够实时处理新增的大量数据,比如有大量在线交互的电子商务应用,就具有很强的时效性,大数据以数据流的形式产生、快速流动、迅速消失,且数据流量通常不是平稳的,会在某些特定的时段突然激增,数据的涌现特征明显,而用户对于数据的响应时间通常非常

敏感,心理学实验证实,从用户体验的角度,瞬间(moment,3秒钟)是可以容忍的最大极限,对于大数据应用而言,很多情况下都必须要在1秒钟或者瞬间内形成结果,否则处理结果就是过时和无效的,这种情况下,大数据要求快速、持续的实时处理。对不断激增的海量数据数据的实时处理要求,是大数据与传统海量数据处理技术的关键差别之一。

(四) 数据价值密度低

数据价值密度低是大数据关注的非结构化数据的重要属性。传统的结构化数据,依据特定的应用,对事物进行了相应的抽象,每一条数据都包含该应用需要考量的信息,而大数据为了获取事物的全部细节,不对事物进行抽象、归纳等处理,直接采用原始的数据,保留了数据的原貌,且通常不对数据进行采样,直接采用全体数据,由于减少了采样和抽象,呈现所有数据和全部细节信息,可以分析更多的信息,但也引入了大量没有意义的信息,甚至是错误的信息,因此相对于特定的应用,大数据关注的非结构化数据的价值密度偏低,以当前广泛应用的监控视频为例,在连续不间断监控过程中,大量的视频数据被存储下来,许多数据可能是无用,对于某一特定的应用,比如获取犯罪嫌疑人的体貌特征,有效的视频数据可能仅仅有一两秒,大量不相关的视频信息增加了获取这有效的一两秒数据的难度。但是大数据的数据密度低是指相对于特定的应用,有效的信息相对于数据整体是偏少的,信息有效与否也是相对的,对于某些应用是无效的信息对于另外一些应用则成为最关键的信息,数据的价值也是相对的,有时一条微不足道的细节数据可能造成巨大的影响,比如网络中的一条几十个字符的微博,就可能通过转发而快速扩散,导致相关的信息大量涌现,其价值不可估量。因此为了保证对于新产生的应用有足够的有效信息,通常必须保存所有数据,这样就使得一方面是数据的绝对数量激增,一方面是数据包含有效信息量的比例不断减少,数据价值密度偏低。

四、大数据要解决的核心问题

与传统海量数据的处理流程相类似,大数据的处理也包括获取与特定的应用相关的有用数

据,并将数据聚合成便于存储、分析、查询的形式;分析数据的相关性,得出相关属性;采用合适的方式将数据分析的结果展示出来等过程。大数据要解决的核心问题与相应的这些步骤相关。

(一) 获取有用数据

通常认为,数据是大数据要处理的对象,大数据技术流程应该从对数据的分析开始,实际上,规模巨大,种类繁多,包含大量信息的数据是大数据的基础,数据本身的优劣对分析结果有很大的影响,有一种观点认为,数据量大了可以不必强调数据的质量,允许错误的数据进入系统,参与分析。大量的数据中包含少量的错误数据影响不大,事实上如果不加约束,大量错误数据涌入就可能得到完全错误的结果。正是数据获取技术的进步促成了大数据的兴起,大数据理应重视数据的获取,如果通过简单的算法处理大量的数据就可以得出相关的结果,则解决问题的困难就转到了如何获取有效的数据。文献^[8]中指出数据的产生技术经历了被动、主动和自动三个阶段,早期的数据是人们为基于分析特定问题的需要,通过采样、抽象等方法记录产生的数据;随着互联网特别是社交网络的发展,越来越多的人在网络上传递发布信息,主动产生数据;而传感器技术的广泛应用使得利用传感器网络可以不用控制全天候的自动获取数据。其中自动、主动数据的大量涌现,构成了大数据的主要来源。对于实际应用来说,并不是数据越多越好,获取大量数据的目的是尽可能正确、详尽的描述事物的属性,对于特定的应用数据必须包含有用的信息,拥有包含足够信息的有效数据才是大数据的关键。有了原始数据,要从数据中抽取有效的信息,将这些数据以某种形式聚集起来,对于结构化数据,此类工作相对简单。而大数据通常处理的是非结构化数据,数据种类繁多,构成复杂,需要根据特定应用的需求,从数据中抽取相关的有效数据,同时尽量摒除可能影响判断的错误数据和无关数据。

(二) 数据分析

数据分析是大数据处理的关键,大量的数据本身并没有实际意义,只有针对特定的应用分析这些数据,使之转化成有用的结果,海量的数据才能发挥作用。数据是广泛可用的,所缺乏的是

从数据中提取知识的能力,当前,对非结构化数据的分析仍缺乏快速、高效的手段,一方面是数据不断快速的产生、更新,一方面是大量的非结构化数据难以得到有效的分析,大数据的前途取决于从大量未开发的数据中提取价值,据 IDC 统计^[2]:2012 年,若经过标记和分析,数据总量中 23% 将成为有效数据,大约为 643EB;但实际上只有 3% 的潜在有效数据被标记,大量的有效数据不幸丢失。预计到 2020 年,若经过标记和分析,将有 33% (13000EB) 的数据成为有效数据,具备大数据价值。价值被隐藏起来的数据量和价值被真正挖掘出来的数据量之间的差距巨大,产生了大数据鸿沟,对多种数据类型构成的异构数据集进行交叉分析的技术,是大数据的核心技术之一。此外,大数据的一类重要应用是利用海量的数据,通过运算分析事物的相关性,进而预测事物的发展。与只记录过去,关注状态,简单生成报表的传统数据不同,大数据不是静止不动的,而是不断的更新、流动,不只记录过去,更反映未来发展的趋势。过去,较少的数据量限制了发现问题的能力,而现在,随着数据的不断积累,通过简单的统计学方法就可能找到数据的相关性,找到事物发生的规律,指导人们的决策。

(三) 数据显示

数据显示是将数据经过分析得到的结果以可见或可读形式输出,以方便用户获取相关信息。对于传统的结构化数据,可以采用数据值直接显示、数据表显示、各种统计图形显示等形式来表示数据,而大数据处理的非结构化数据,种类繁多,关系复杂,传统的显示方法通常难以表现,大量的数据表、繁乱的关系图可能使用户感到迷茫,甚至可能误导用户。利用计算机图形学和图像处理的可视计算技术成为大数据显示的重要手段之一,将数据转换成图形或图像,用三维形体来表示复杂的信息,直接对具有形体的信息进行操作,更加直观,方便用户分析结果。若采用立体显示技术,则能够提供符合立体视觉原理的绘制效果,表现力更为丰富。对于传统的数据表示方式,图表、数据通常是二维的,用户与计算机交互容易,而通过三维表现的数据,通常由于数据过于复杂,难以定位而交互困难,可以通过最

近兴起的动作捕捉技术, 获取用户的动作, 将用户与数据融合在一起, 使用户直接与绘制结果交互, 便于用户认识、理解数据。数据显示以准确、方便的向用户传递有效信息为目标, 显示方法可以根据具体应用需要来选择。

(四) 实时处理数据的能力

大数据需要充分、及时地从大量复杂的数据中获取有意义的相关性, 找出规律。数据处理的实时要求是大数据区别于传统数据处理技术的重要差别之一。一般而言, 传统的数据处理应用对时间的要求并不高。运行 1-2 天获得结果依然是可以接受的。而大数据领域相当大的一部分应用需要在 1 秒钟内或瞬间内得到结果, 否则相关的处理结果就是过时的、无效的。先存储后处理的批处理模式通常不能满足需求, 需要对数据进行流处理。由于这些数据的价值会随着时间的推移不断减少, 实时性成了此类数据处理的关键。而数据规模巨大、种类繁多、结构复杂, 使得大数据的实时处理极富挑战性。数据的实时处理要求实时获取数据, 实时分析数据, 实时绘制数据, 任何一个环节慢都会影响系统的实时性。当前, 互联网络以及各种传感器快速普及, 实时获取数据难度不大; 实时分析大规模复杂数据是系统的瓶颈, 也是大数据领域亟待解决的核心问题; 数据的实时绘制是可视计算领域的热点问题, GPU 以及分布式并行计算的飞速发展使得复杂数据的实时绘制成为可能, 同时数据的绘制可以根据实际应用和硬件条件选择合适的绘制方式。

五、大数据面临的挑战

当今社会, 互联网络和传感器技术飞速发展, 大规模非结构化数据快速积累, 适应时代发展的大数据理论和技术其前瞻性是显而易见的, 但同时, 大数据的概念也有过分炒作的可能。大数据这种新的理念一出现, 就出现了大数据当立, 传统方案当下的论调, 似乎大数据是万能的, 传统的数据分析、处理方法可以淘汰了, 以数据为中心, 当数据多到一定程度时, 用最简单的算法就可以得到结果, 不需要关注算法的优劣, 只需关注数据的质量, 大数据带来的巨大运算量可以由计算优势来应对。实际上, 大数据是

一种新兴的理论, 大数据的概念、技术、方法还远不成熟, 在其发展的过程中还将面临多种挑战, 不应过分夸大其先进性。

(一) 不能完全代替传统数据

当前大数据尚不能完全取代传统结构化数据, 尽管大数据关注的非结构化数据的绝对数据量占总数据量的 75%, 但由于非结构化数据的价值偏低, 有效的非结构化数据与结构化数据相比并不占绝对优势, 对于某些特定的应用, 结构化数据仍然占据主导地位。对于互联网、社交网络、传感器网络等应用, 利用大数据分析可以更好的分析相关的非结构化海量数据, 因此前面所述的 EMC、Google、Facebook 等面临数据爆炸的商业巨头积极推动大数据技术发展。而对于传统的结构化数据密集型的应用, 相关研究已经持续了几十年, 传统数据处理方法可以很好的处理这些结构化数据, 对于这些应用则没有必要应用大数据相关技术, 没有必要盲目的追逐潮流。此外, 商业上一些所谓大数据应用, 甚至就是对原来技术进行新的包装, 并没有革命性的突破。大数据当立, 传统方案当下的论调当前并不准确, 非结构化数据完全替代传统数据尚需时日, 用户需要根据实际应用需要选择合适的数据处理方式。

(二) 数据保护

大数据时代, 互联网络的发展使得获取数据十分便利, 给信息安全带来了巨大的挑战。当前, 数据安全形势不容乐观, 需要保护的数据量增长已超过了数据总量的增长。据 IDC 统计^[15]: 2010 年仅有不到 1/3 的数据需要保护, 到 2020 年这一比例将超过 2/5; 2012 年的统计显示, 虽然有 35% 的信息需要保护, 但实际得到保护的不到 20%。在亚洲、南美等新兴市场, 数据保护的缺失更加严重。首先个人隐私更容易通过网络泄露, 随着电子商务、社交网络的兴起, 人们通过网络联系的日益紧密, 将个人的相关数据足迹聚集起来分析, 可以很容易获取个人的相关信息, 隐私数据就可能暴露, 而数据在网络上的发布机制使得这种暴露似乎防不胜防; 在国家层面, 大数据可能给国家安全带来隐患, 如果在大数据处理方面落后, 就可能导致数据的单向透明, 美国发布大数据研发计划, 大力发展大数据技术就

有增强国家安全方面的战略考量。

(三) 相关性预知

大数据时代,人们不再认为数据是静止和陈旧的,而是流动的、不断更新的。大数据是人们获得新的认知,创造新的价值的源泉,通过分析数据的相关性可能预知事物的发展方向。但是从数据来的结论不一定能反映真实,比如随着数据的增多,会带来部分错误的数据,使得数据价值大大降低,影响分析的结果,甚至可能得出错误的结论。此外,大数据获取的统计学上的宏观结论,对于一些微观的问题并没有意义,比如抛硬币,抛的次数越多,得到正反两面的次数越接近,概率越接近 0.5,但不管已经抛了多少次,还是不能分析出下一次得到正面还是反面。因此,不能希望通过大数据可以预知一切。

六、结语

随着社交网络、物联网、云计算的飞速发展,大量非结构化数据呈指数级快速增长,数据样式高度复杂,为人类认识世界、改造世界提供了重要的资源,企业和个人通过网络可以大规模的收集和分析数据,也可以产生、发布数据,个体在互联的网络中既是数据的消费者又是数据的生产者,大规模生产、分享、应用数据的大数据时代已经来临。与此同时,数量巨大、种类繁多的数据给传统的数据获取、分析、处理、存储、检索技术带来了挑战,大数据成为广泛关注且亟待解决的热点问题,并已经开始影响社会的发展与人们的日常生活。然而大数据的概念和相关技术还远未成熟,尚存在着一定的争议,面临着诸多挑战,甚至有人认为大数据有过分炒作的可能。本文从几种常见的描述大数据的概念出发,分析大数据的典型特征,依据这些特征来讨论大数据技术可能的要解决的核心问题,最后讨论了大数据可能要面临的多种挑战。

大数据的概念来源于、发展于美国,并向全球扩展,必将给我国未来的科技与经济发展带来深远影响。根据 IDC 统计,目前数据量在全球比例为:美国 32%、西欧 19%、中国 13%,预计

到 2020 年中国将产生全球 21% 的数据,我国是仅次于美国的数据大国,而我国大数据方面的研究尚处在起步阶段,如何开发、利用保护好大数据这一重要的战略资源,是我国当前亟待解决的问题。

[参考文献]

- [1] Obama Administration Unveils “Big Data” Initiative: Announces \$200 Million in New R&D Investments. Office of Science and Technology Policy Executive Office of the President. 2012. 5.
- [2] David Reinsel, John Gantz. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. 2012. 12.
- [3] Tony Hey, Stewart Tansley. The Fourth Paradigm: Data - Intensive Scientific Discovery. Microsoft, 2009. 10.
- [4] 托夫勒. 第三次浪潮[M]. 北京: 中信出版社 2006.
- [5] Big Data. Nature, 2008, 455(7209): 1 - 136.
- [6] Dealing with data. Science, 2011, 331(6018): 639 - 806.
- [7] 李国杰, 程学旗. 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[J]. 中国科学院院刊. 2012(06): 647 - 657.
- [8] 孟小峰, 慈祥. 大数据管理概念技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146 - 169.
- [9] 覃雄派, 王会举, 杜小勇, 王珊. 大数据分析——RDBMS 与 MapReduce 的竞争与共生[J]. 软件学报, 2012, 23(1): 32 - 45.
- [10] 王珊, 王会举, 覃雄派, 周炯. 架构大数据: 挑战、现状与展望[J]. 计算机学报, 2012, 34(10): 1741 - 1752.
- [11] 维克托·迈尔-舍恩伯格. 大数据时代[M]. 上海: 浙江人民出版社, 2012.
- [12] Martin Klubeck. 量化: 大数据时代的企业管理[M]. 北京: 人民邮电出版, 2013.
- [13] 朱志军, 闫蕾. 转型时代丛书: 大数据·大价值、大机遇、大变革[M]. 北京: 电子工业出版社, 2012.
- [14] Big data, http://en.wikipedia.org/wiki/Big_data.
- [15] Benjamin Woo World wide Big Data Technology and Services 2012 - 2015 Forecast. 2012. 5 [16] Big data <http://www.gartner.com/it-glossary/big-data>.

The Concept , Characteristics and Application of Big Data

Ma Jian – guang , JIANG Wei

(1. School of Humanities and Social Sciences , National University of Defense Technology , Changsha , Hunan 410073 , China)

Abstract: With the rapid development of the Internet , especially the wide application of social networking , the Internet of Things , cloud computing as well as a variety of sensors in recent years , unstructured data , which have large numbers , varieties and also timeliness , continue to emerge. The importance of the data becomes more prominent. It is difficult to use the traditional data storage and analysis technology to handle large volumes of unstructured information in a real – time manner , and that's how the concept of big data came into being. How to obtain , aggregate and analyze big data becomes a hot issue. This paper introduces the concept and characteristics of big data , analyzes the core issues , such as the correlation analysis , real – time processing , etc. , and finally discusses many challenges large data may face.

Key words: big data , unstructured information , resolve of the core issues , future challenges