DOI:10.16182/j.cnki.joss.2013.s1.074

大数据研究综述

陶雪娇，胡晓峰，刘洋
（国防大学信息作战与指挥训练教研部，北京 100091）

摘要：2010 年，全球数据量跨入了 ZB 时代，据 IDC 预测，至 2020 年全球将拥有 35ZB 的数据量，大量数据实时地影响我们工作、生活，甚至国家经济、社会发展，大数据时代已经到来。大数据具有数据量巨大、数据类型多样、流动速度快和价值密度低的特点，大数据技术为我们分析问题和解决问题提供了新的思路和方法，其研究渐渐成为热点。阐述了大数据的相关概念、特点、*大数据技术特别是在数据挖掘方面国内外发展状况以及我们在大数据时代面临的挑战*。通过综述，对大数据有一个全面的认识，为下一步研究打下基础。

关键词：大数据；大数据技术；数据挖掘；挑战
中图分类号：TP301　　文献标识码：A　　文章编号：1004-731X (2013) S-0142-05

Overview of Big Data Research

TAO Xue-jiao, HU Xiao-feng, LIU Yang
(The Department of Information Operation & Command Training of NDU, Beijing 100091, China)

Abstract: In 2010, the quantity of data reached ZB level. According to IDC, there will be at least 35 zettabytes of stored data in 2020. Massive data are affecting our life, even the economy and the development of the society. The Big Data era has already come. There are four defining characteristics of Big Data: volume, variety, velocity and value. It is often referred to them as “the 4Vs”. The Big Data technology will offer new ideas and methods, which is becoming popular. *Introductions to Big Data and Big Data technology with particular emphasis on Data Mining were* given. There will be a comprehensive understanding of Big Data and lay a foundation for further study.

Key words: big data; big data technology; data mining; challenge

引言

当我们对“云计算”、“物联网”等概念还感觉模糊的时候，“大数据”横空出世且其发展呈燎原之势。为了减少火车脱轨造成的伤亡，交通系统变得更加智能。火车上安装了各种传感器来收集各个部位运行情况的数据，以此来检测存在安全隐患的器件，当然，这些还远不能称为“智能”，要对铁轨乃至整个交通系统都能够进行实时的数据采集，甚至对影响交通的天气情况都要考虑在内，现在把这些信息加入到火车的承载量、出发以及到达等数据里，一个大数据问题就出现了[1]。

我们身处数据的海洋，几乎所有事物都与数据有关，环境、金融、医疗……我们每天都在产生数据，打电话、发短信、进地铁站安检、进办公楼刷卡、在 QQ 上聊天、上淘宝网购物……大量数据实时地影响我们的工作、生活乃至社会发展。数据成为与自然资源、人力资源同样重要的战略资源，引起了科技界和企业界的高度重视。

根据国际数据资讯(IDC)公司监测，全球数据量大约每两年翻一番，预计到 2020 年，全球将拥有 35ZB 的数据量(如图 1 所示)，并且 85%以上的数据以非结构化或半结构化的形式存在。IT 专业人员预见数据处理面临的挑战，用“Big Data(大数据)”来形容这个问题。其实，“大数据”这个名词并不新鲜，早在上个世纪 80 年代就有美国人提出来[2]。2008 年 9 月，《科学》杂志发表文章“Big Data: Science in the Petabyte Era”，“大数据”这个词开始广泛传播。

收稿日期：2013-04-15　　修回日期：2013-06-06
基金项目：国家自然科学基金(61174156, 61174035, 61273189)
作者简介：陶雪娇(1986-)，女，山东人，硕士，助工，研究方向为计算机战争模拟；胡晓峰(1957-)，男，山东人，教授，博导，研究方向为军事运筹与军事系统工程、战争仿真系统工程、多媒体及虚拟现实等；刘洋(1975-)，女，江西人，博士，副教，研究方向为作战模拟、数据分析与挖掘、复杂网络等。

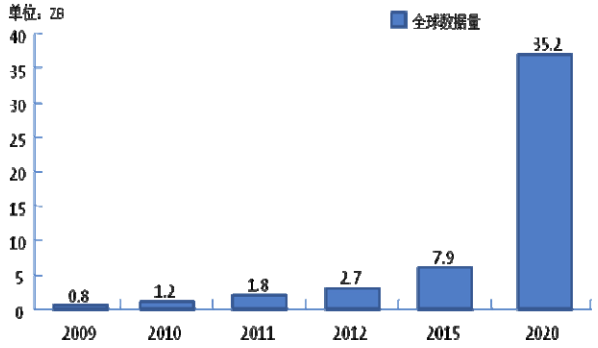

图 1　IDC 全球数据使用量预测

2011 年 6 月，IDC 研究报告《从混沌中提取价值》中三个基本论断构成了大数据的理论基础[3]，人们对大数据的关注程度日益上升，据统计，Google“大数据”搜索量自 2011 年 6 月起呈直线上升趋势，大数据时代的到来毋庸置疑。

1　定义

研究机构 Gartner 的定义：大数据是指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。

维基百科的定义：大数据指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策目的的资讯。

麦肯锡的定义：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。

无论哪种定义，我们可以看出，大数据并不是一种新的产品也不是一种新的技术，就如同本世纪初提出的“海量数据”概念一样，大数据只是数字化时代出现的一种现象。那么海量数据与大数据的差别何在[4]？从翻译的角度看，“大数据”和“海量数据”均来自英文，“big data”翻译为“大数据”，而“large-scale data”或者“vast data”则翻译为“海量数据”。从组成的角度看，海量数据包括结构化和半结构化的交易数据，而大数据除此以外还包括非结构化数据和交互数据。Informatica 大中国区首席产品顾问但彬进一步指出，大数据意味着包括交易和交互数据集在内的所有数据集，其规模或复杂程度超出了常用技术，按照合理的成本和时限捕捉、管理及处理这些数据集的能力。可见，大数据由海量交易数据、海量交互数据和海量数据处理三大主要的技术趋势汇聚而成。

上个世纪 60 年代，数据一般存储在文件中，由应用程序直接管理；70 年代构建了关系数据模型，数据库技术为数据存储提供了新的手段；80 年代中期，数据仓库由于具有面向主题、集成性、时变性和非易失性特点，成为数据分析和联机分析的重要平台；随着网络的普及和 web 2.0 网站的兴起，基于 Web 的数据库和非关系型数据库等技术应运而生……目前，智能手机和社交网络的广泛使用，使得各种类型的数据呈指数增长，渐渐超出了传统关系型数据库的处理能力，数据中存在的关系和规则难以被发现，而大数据技术很好的解决了这个难题，它能够在成本可承受的条件下，在较短的时间内，将数据采集到数据仓库中，用分布式技术框架对非关系型数据进行异质性处理，通过数据挖掘与分析，从大量化、多类别的数据中提取价值，大数据技术将是 IT 领域新一代的技术与架构。

2　特征

2.1　数据体量巨大(Volume)

大数据通常指 10TB(1TB=1024GB)规模以上的数据量。之所以产生如此巨大的数据量，一是由于各种仪器的使用，使我们能够感知到更多的事物，这些事物的部分甚至全部数据就可以被存储；二是由于通信工具的使用，使人们能够全时段的联系，机器-机器(M2M)方式的出现，使得交流的数据量成倍增长；三是由于集成电路价格降低，使很多东西都有了智能的成分。

2.2　数据种类繁多(Variety)

随着传感器种类的增多以及智能设备、社交网络等的流行，数据类型也变得更加复杂，不仅包括传统的关系数据类型，也包括以网页、视频、音频、e-mail、文档等形式存在的未加工的、半结构化的和非结构化的数据。

2.3　流动速度快(Velocity)

我们通常理解的是数据的获取、存储以及挖掘有效信息的速度，但我们现在处理的数据是 PB 级代替了 TB 级，考虑到“超大规模数据”和“海量数据”也有规模大的特点，强调数据是快速动态变化的，形成流式数据是大数据的重要特征，数据流动的速度快到难以用传统的系统去处理。

2.4　价值密度低(Value)

数据量呈指数增长的同时，隐藏在海量数据的有用信息却没有相应比例增长，反而使我们获取有用信息的难度加大。以视频为例，连续的监控过程，可能有用的数据仅有一两秒。

大数据的“4V”特征表明其不仅仅是数据海量，对于大数据的分析将更加复杂、更追求速度、更注重实效。

3　国内外发展情况

1989 年在美国底特律召开的第 11 届国际人工智能联

合会议专题讨论会上，首次提出了“数据库中的知识发现（KDD）”的概念。1995 年召开了第一届知识发现与数据挖掘国际学术会议，随着与会人员的增多，KDD 国际会议发展为年会。1998 年在美国纽约举行了第四届知识发现与数据挖掘国际学术会议，不仅进行了学术讨论，而且 30 多家软件公司展示了自己的产品，比如，IBM 公司研制的 Intelligent Miner，用来提供数据挖掘的解决方案；SPSS 股份公司开发了基于决策树的数据挖掘软件 Clementine；Oracle 公司开发的 Darwin 数据挖掘套件，另外还有 SAS 公司的 Enterprise 和 SGI 公司的 Mine Set 等。

经济利益成为主要的推动力，IBM、ORACLE、微软、谷歌、亚马逊、Facebook、Teradata、EMC、惠普等跨国巨头也因大数据技术的发展而更加具有竞争力[5]。仅 2009 年一年，谷歌公司通过大数据业务对美国经济贡献 540 亿美元；2005 年以来，IBM 投资 160 亿美元进行 30 多次与大数据相关的收购，使业绩稳定高速增长，2012 年，IBM 股价每股突破 200 美元大关，3 年内翻了 3 番；eBay 通过数据挖掘精确计算出广告中每个关键字带来的回报，2007 年以来，广告费降低了 99%，同时顶级卖家占总销售额的百分比上升至 32%；2011 年，Facebook 首次公开新数据处理分析平台 PUMA，通过对数据多处理环节区分优化，相比之前单纯采用 Hadoop 和 Hive 进行处理的技术，数据分析周期从 2 天降到 10 秒以内，效率提高数万倍。

2012 年 3 月，奥巴马政府公布“大数据研发计划”，旨在提高和改进人们从海量、复杂的数据中获取知识的能力，发展收集、储存、保留、管理、分析和共享海量数据所需要的核心技术，大数据成为继集成电路和互联网之后信息科技关注的重点。

与国外相比，国内起步稍晚，还未形成整体力量，企业使用数据挖掘技术尚不普遍,但近几年出现了蓬勃发展的态势。

我国国家自然科学基金于 1993 年首次支持对数据挖掘领域的研究项目。1999 年，在北京召开第三届亚太地区知识发现与数据挖掘国际会议(PAKDD)，收到论文 158 篇[6]。2011 年，第十五届 PAKDD 在深圳举办，会议就数据挖掘、知识发现、人工智能、机器学习等相关领域的主题进行交流讨论，反响热烈。2012 年 6 月 9 日，中国计算机学会常务理事会决定成立大数据专家委员会。2012 年 10 月，成立了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, 推动了我国大数据的科研与发展。2012 年 11 月，“Hadoop 与大数据技术大会”以“大数据共享与开放技术”为主题，总结了八个热点问题：数据科学与大数据的学科边界、数据计算的基本模式与范式、大数据的作用力和变换反对、大数据特性与数据态、大数据安全和隐私问题、大数据对 IT 技术架构的挑战、大数据的生态环境问题以及大数据的应用及产业链。大会还成立了“大数据共享联盟”，旨在搜集大数据、展示大数据、促进大数据的研究与开发。

目前，国内相关技术主要集中于数据挖掘相关算法、实际应用及有关理论方面的研究，涉及行业比较广泛，包括金融业、电信业、网络相关行业、零售业、制造业、医疗保健、制药业及科学领域，单位集中在部分高等院校、研究所和公司，特别是在 IT 等新兴领域，华为、阿里巴巴、百度等对技术进步起到了很大的推动作用。

4 相关技术

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”。之前，我们手中的数据量相对不足，对数据的研究是“由薄变厚”，把“小”数据变“大”，而在“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[7]。大数据时代，“数据丰富、信息匮乏”的现象越来越突显，使人们产生对数据分析工具的强烈需求，数据挖掘的地位越来越重要。

4.1 数据挖掘

数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的，但又是潜在有用信息和知识的过程。

目前广为接受的一种处理模型是Fayyad等人设计的多处理阶段模型（如图 2 所示）。

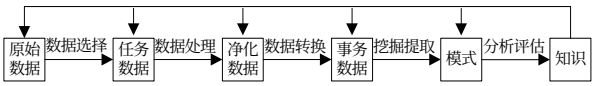

图 2 多处理阶段模型

数据挖掘的焦点集中在寻求数据挖掘过程中的可视化方法，使知识发现过程能够被用户理解，便于在知识发现过程中的人机交互；研究在网络环境下的数据挖掘技术，特别是在 Internet 上建立数据挖掘和知识发现(DMKD)服务器，与数据库服务器配合，实现数据挖掘；加强对各种非结构化或半结构化数据的挖掘，如多媒体数据、文本数据和图像数据等。

目前，大数据的研究主要是将其作为一种研究方法或一种发现新知识的工具，而不是把数据本身当成研究目标[5]，它与传统数据挖掘方法有密切联系又有根本不同。

4.2 对比分析[8]

4.2.1 分析对象

传统数据分析主要针对已知的数据范围中易处理的数据进行，大多数据仓库都有一个完善的 ETL 流程和数据库限制，这意味着加载进数据仓库的数据是容易理解的、洗清过的并符合业务的元数据。而大数据分析针对传统手段捕捉到的数据之外的非结构化数据，意味着不能保证输入的数据是完整的、清洗过和没有错误的。这使它更有挑战性，但同时提供了在数据中获得更多洞察力的范围。

4.2.2 分析基础

传统分析是建立在关系数据模型之上的，主题之间的关系在系统内就已经被创立，而分析也在此基础上进行。而在典型的世界里，很难在所有的信息间以一种正式的方式建立关系，因此非结构化以图片、视频、移动产生的信息、无线射频识别等的形式存在，被考虑进大数据分析，绝大多数的分析基于纵列数据库之外。

4.2.3 分析效率

传统分析是定向的批处理，而且我们在获得所需的洞察力之前需要等待 ETL 等工作的完成。而大数据分析是利用对数据有意义的软件的支持，对数据进行实时分析。

4.2.4 硬件要求

在一个传统的分析系统中，平行是通过昂贵的硬件，如大规模并行处理系统或对称多处理系统来实现的。而大数据分析的应用系统，可以通过通用的硬件和新一代的分析软件来实现，加之成本的考虑，由高端的服务器向中低端硬件构建的大规模机群平台发展[9]。

4.3 国内外技术发展

在相关技术中，比较具有代表性的是 Apache 软件基金会开发的 Hadoop，以 MapReduce 和 Hadoop 为代表的非关系数据分析技术，凭借其适合非结构处理、大规模并行处理和简单易用等优势，在互联网搜索和其他大数据分析技术领域取得重大进展，成为主流技术。

MapReduce 是 2004 年谷歌公司提出的用来进行并行处理和生成大数据的模型，是一种线性的、可伸缩的编程模型[10]。其可扩展性得益于 shared-nothing 结构、各节点间的松耦合性和较强的软件级容错能力。MapReduce 被设计在处理时间内解释数据，所以对非结构化、半结构化的数据处理非常有效。针对 MapReduce 并行编程模型的易用性，产生了多种大数据处理高级查询语言，如 Facebook 的 Hive、雅虎的 Pig、谷歌的 Sawzall 等。但 MapReduce 作为典型的离线计算框架，无法满足在线实时计算需求，目前在线计算主要基于两种模式：一是基于关系型数据库，通过提高其扩展性，增加查询通量来满足大规模数据处理需求，但通常不能对上层提供原存储引擎的全部查询功能；二是基于 NoSQL 数据库，通过提高其查询能力、丰富查询功能来满足需求，典型的 NoSQL 有 Oracle NoSQL DB、MySQL Cluster、MyFox 等，典型应用是谷歌的 BigTable 及其一系列扩展系统。

Hadoop 起源于 Apache Nutch，2006 年 2 月，NDFS 和 MapReduce 从 Nutch 转移出来，成为一个独立的 Lucene 子项目，称为 Hadoop。Hadoop 是开放源码并行运算编程工具和分散式档案系统，是 MapReduce 的开源实现，凭借其开源和易用的特性，成为大数据处理的首选。2008 年 2 月，雅虎宣布其搜索引擎产品部署在一个拥有一万个内核的 Hadoop 集群上，此外，Hadoop 还被 Last.fm、Facebook 和《纽约时报》等公司应用，其核心功能是提供一个稳定的共享存储和分析系统，存储由 HDFS 实现，分析由 MapReduce 实现。Hadoop 架构支持在公有云端存储 EB 量级数据的应用，许多互联网公司，包括 Facebook、谷歌、eBay 和雅虎等，都已开发了基于 Hadoop 的 EB 量级超大规模数据应用。据统计，云计算与大数据的深度融合位列 2013 年大数据发展趋势的第三名，大数据为云计算大规模和分布式的计算能力提供了广阔的应用空间，云计算正在进入以“分析即服务(AaaS)”为主要标志的 Cloud 2.0 时代[5]。数据的爆发对 IT 架构提出了大的挑战，今天的 Hadoop 实际上是应对大数据及大数据处理的相关的架构。

国内，大数据技术的发展呈现良好势头。华为提供了基于 x86 服务器的 SmartVision 大数据处理解决方案[11]，催生数据基础架构的革新。SmartVision 方案引入了流处理机制、提供统一的存储处理平台、提供基于虚拟机的弹性服务方案，是一个系统性的工程，在存储、计算、网络等硬件方面拥有完整的多层次的产品线。在“2012 华为云计算大会”上，推出了 OceanStor MVX 大数据存储解决方案[12]，存储系统是融合了 Scale-out NAS、Scale-out Database 和 Scale-out Backup，实现存储、分析、备份多位一体，面向大数据存储的集群存储系统。

可见，大数据技术不是一款简单的数据分析软件，要从大体量、多类别的数据中快速提取价值，几乎需要重构整个数据库技术体系。

5 面临挑战

5.1 数据量的成倍增长挑战数据存储能力

大数据及其潜在的商业价值要求使用专门的数据库技

术和专用的数据存储设备，传统的数据库追求高度的数据一致性和容错性，缺乏较强的扩展性和较好的系统可用性，不能有效存储视频、音频等非结构化和半结构化的数据。目前，数据存储能力的增长远远赶不上数据的增长，设计最合理的分层存储架构成为信息系统的关键[5]。

5.2 数据类型的多样性挑战数据挖掘能力

数据类型的多样化，对传统的数据分析平台发出了挑战。从数据库的观点看，挖掘算法的有效性和可伸缩性是实现数据挖掘的关键，而现有的算法往往适合常驻内存的小数据集，大型数据库中的数据可能无法同时导入内存，随着数据规模的不断增大，算法的效率逐渐成为数据分析流程的瓶颈。要想彻底改变被动局面，需要对现有架构、组织体系、资源配置和权力结构进行重组。

5.3 对大数据的处理速度挑战数据处理的时效性

随着数据规模的不断增大，分析处理的时间相应地越来越长，而大数据条件下对信息处理的时效性要求越来越高。传统的数据挖掘技术在数据维度和规模增大时，需要的资源呈指数增长，面对 PB 级以上的海量数据，NlogN 甚至线性复杂度的算法都难以接受，处理大数据需要简单有效的人工智能算法和新的问题求解方法。

5.4 数据跨越组织边界传播挑战信息安全

随着技术的发展，大量信息跨越组织边界传播，信息安全问题相伴而生，不仅是没有价值的数据大量出现，保密数据、隐私数据也成倍增长，国家安全、知识产权、个人信息等等都面临着前所未有的安全挑战。大数据时代，犯罪分子获取信息更加容易，人们防范、打击犯罪行为更加困难，这对数据存储的物理安全性以及数据的多副本与容灾机制提出了更高的要求。要想应对瞬息万变的安全问题，最关键的是算法和特征，如何建立相应的强大安全防御体系来发现和识别安全漏洞是保证信息安全的重要环节。

5.5 大数据时代的到来挑战人才资源

从大数据中获取价值至少需要三类关键人才队伍：一是进行大数据分析的资深分析型人才；二是精通如何申请、使用大数据分析的管理者和分析家；三是实现大数据的技术支持人才。此外，由于大数据涵盖内容广泛，所需的高端专业人才不仅包括程序员和数据库工程师，同时也需要天体物理学家、生态学家、数学和统计学家、社会网络学家和社会行为心理学家等。可以预测，在未来几年，资深数据分析人才短缺问题将越来越突显。同时，需要具有前瞻性思维的实干型领导者，能够基于从大数据中获得的见解和分析，制定相应策略并贯彻执行。

6 结论

图灵奖得主吉姆·格雷在最后一次演讲中描绘了数据密集型科研“第四范式(the fourth paradigm)”的愿景，其研究方式不同于基于数学模型的传统研究方式，PB 级数据可以没有模型和假设就分析数据，将海量数据丢进巨大的计算机集群中，只要有相互关系的数据，统计分析计算就可以发现过去科学方法发现不了的新模式、新知识甚至新规律。这也对大数据分析平台提出了很高的要求，需要扩展以下功能：高度可扩展性、高性能、高度容错性、支持异构环境、较低的分析延迟、易用且开放的接口、较低成本和向下兼容性[9]。

“2012Hadoop 与大数据技术大会”发布了今年大数据发展趋势预测调研结果，位列前三的分别是：一、数据资源化；二、大数据隐私问题；三、大数据与云计算等深度融合。《连线》杂志主编克里斯·安德森曾断言“数据洪流使传统科学方法变得过时”，尽管有些极端，但大数据确确实实已经在改变我们的生活、改变我们的思维方式。

参考文献：

[1] Paul C Zikopoulos, Chris Eaton, Dirk de Roos, Thomas Deutsch, George Lapis. Understanding Big Data [M]. USA: The McGraw-Hill Companies, 2012.
[2] 徐子沛. 大数据[M]. 广西: 师范出版社, 2012: 57.
[3] 符健. 解读大数据[Z]. 证券研究报告, 2011.
[4] 涂兰敬. 大数据与海量数据的区别[J]. 网络与信息, 2011, 25(12): 37-38.
[5] 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.
[6] 员巧云, 程刚. 近年来我国数据挖掘研究综述[J]. 情报学报, 2005.
[7] 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.
[8] 人大经济论坛, 传统分析与大数据分析的对比[DB/OL]. (2012) http://bbs.pinggu.org/forum.php?mod=viewthread&tid=2140833&page=1.
[9] 王珊, 王会举, 覃雄派, 等. 架构大数据: 挑战、现状与展望[J]. 计算机学报, 2011, 10: 1472.
[10] 马帅, 李建欣, 胡春明. 大数据科学与工程的挑战与思考[Z].
[11] CT 论坛. 华为 SmartVision 大数据解决方案[Z]. 2012 http://ec.ctiforum.com.
[12] 大数据解决之道: 华为 OceanStor MVX 存储系统技术漫谈[Z]. http://www.cww.net.cn, 2012.
[13] 夏岩, 赵慧英, 贾军帅. 数据挖掘发展综述[J]. 通信与计算技术, 2009.
[14] 郭海涛, 段礼祥, 闫春颖. 数据挖掘方法综述[J]. 计算机科学, 2009.
[15] 彭渊. 大数据机遇与挑战[Z].
[16] 5 联网, 大数据的本质解析[DB/OL]. [2012]. http://www.5lian.cn/html/2012/yunjisuan_0817/33929.html.