大数据时代的思维特点研究

洪 菁
(佐治亚理工大学计算机学院 亚特兰大 GA 30309)

摘 要 近年来,“大数据”已经成为科技界和企业界关注的热点,所谓 big data(大数据),是指靠专有平台实现价值提炼以帮助使用者决策分析的海量数据集产品。目前,拥有数据的规模大小和运用数据的能力将成为一个国家综合国力的重要组成,一个国家对数据的占有、控制将成为国家间和企业间新的争夺焦点。重点研究分析了大数据的 5 个特性,指出了企业的经济效益是推动大数据发展的主要动力,当前的大数据处理技术使人们从事的工作变得更加智能。通过对大数据时代的思维特点及思维方式进行研究,可以得出大数据时代最大的转变就是研究方式将进入数据密集型科学的范围中进行。

关键词 大数据,思维特点,数据密集型科学
中图法分类号 TP391 **文献标识码** A

Research on Characteristics of Thinking in Era of Big Data

HONG Jing
(School of Computer Science, Georgia Institute of Technology, Atlanta GA 30309, USA)

Abstract In recent years, big data, a massive data set that needs the proprietary platform to realize value abstraction to help decision analysis, has become a focus of the scientific community and the business community. The scale of the data that a country has and its ability to use the data will become the significant components of its comprehensive national strength, and the possession and control of data will become the focal points among countries and enterprises. This paper focuses on the research and analysis of five characteristics, pointing out that the economic benefit of enterprises is the main driving force of the development of big data, and current processing technology of big data makes the work which people engaged more intelligent. Through researching the characteristics and modes of thinking in the era of big data, we concluded that the biggest change in the big data era is that the researches will be conducted in data-intensive science.

Keywords Big data, Thinking characteristics, Data intensive science

1 引言

近年来,大数据(big data)已经成为全球各界关注的热点。2012 年 3 月,美国政府宣布投资 2 亿美元启动“大数据研究和发展计划”[1],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部署[1-3]。

目前,拥有数据的规模大小和运用数据的能力将成为一个国家综合国力的重要组成。一个国家对数据的占有、控制将成为国家间和企业间新的争夺焦点。《纽约时报》2012 年 2 月的一篇专栏指,出当前“大数据”时代已来临,在经济等领域中,将逐步使用数据、分析作出重大决策而并非基于经验和直觉[1],在 IT 产业中尤其如此。

在 21 世纪的今天,24 小时内,全球互联网产生的所有数据能录入近两亿张 4.7G 盘;人类发出的电邮与美国两年的纸质信件数量接近,达 3 亿封,全球论坛发出的各种帖子已达 200 万个,相当于《TIME》杂志 770 年的文字量[3]。当前的数据量已经从 TB(1024GB=1TB)级别跃升到 PB(1024TB=1PB)、EB(1024PB=1EB)乃至 ZB(1024EB=1ZB)级别,可见数据的增长量之大。IBM 公司的研究成果表明整个人类文明所获得的所有数据高达 90%是过去几年产生的[3]。

2 大数据的几大特性

经过近年的研究表明,大数据具有应用价值高、类型多、容量大、存取速度快且真实而准确的 5 个特性,特别是 Value(价值特性),随着研究的深入,技术人员将重新定义并让数据价值体现在应用领域的各个方面。

Volume-(数据量)极大,已经由 TB 增长到 PB 级别。当今全球生产的各种印刷材料的数据量是 200PB,全人类自文明史后说过的所有的话的数据量大约是 5EB[3]。

Velocity-(处理速度)极快,IDC 的“数字宇宙”的报告,经过计算,预计到 2020 年,各个国家数据使用量之和将达到 35.2ZB。在如此海量的数据面前,数据的处理效率就是企业的生命。

Value-(价值密度)高。通常价值密度的高低与数据总量

洪 菁(1995-),女,主要研究方向为数据分析,E-mail:jingh1122@126.com。

的大小成反比。研究通过强大的机器算法快速地完成数据的价值提纯，成为当今大数据背景下急需解决的难题[3]。

Variety-(数据类型)繁多。网络上大量的视频、图片、地理位置信息等类型的多样性也让数据被分为结构化数据和非结构化数据。所有这些各种类型的数据将对数据的处理能力要求更高[3]。

Veracity-(真实而准确)，只有如此，数据的使用才能发挥作用。

因此当今大数据是多维的且极具复杂性，显然它对于每个人来说都是机遇和挑战并存的。

3 大数据成智能手段和致富捷径，推动大数据的动力主要为企业经济效益

Gartner(顾能公司)在2010年提出：信息将成为“21世纪的石油”；而《经济学家》刊文指出，大数据会带来巨大的商机，具有潜在价值。企业大数据研究的动力主要是经济效益的驱动，以巨大的经济利益驱使大企业不断扩大数据处理规模。在制定市场策略政策时，最佳的支撑平台是握有大量的有效数据。这类效益对于企业来说是潜在商机，对于个体则是实现自身价值的高效方法。谁握有数据，利用这些数据交易就能产生好的效益。可以认为尖端的数据处理技术为企业带来了更智能、更富有的益处。当然通过数据挖掘会诞生出许多商业模式，可见开发这类数据在未来有希望成为最大的企业交易品，到那时各个公司的角色有所不同，有数据提供方、管理者，还有监管者等，这样就形成了一大新兴产业：大数据。未来A数据集和B数据集也许可能有机地结合到一起形成交集，会创造出新的信息和知识点，实现企业的大幅增值。今后数据也是战略资源，将与自然资源、人力资源一样重要，已引起人们的高度关注。

美国O'Reilly Media，Inc.(奥莱利公司)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”。在数据为王的时代，企业的战略需求也发生了转变：企业关注重点将转向数据。多数公司将成为真正的IT业，其业务将从追求计算速度转变为高速处理大数据的能力上，软件也将从编程为主转变为以数据为中心为主[4]，此时数据已成为矿物一样的原材料供各类企业使用。到那时将形成诸如数据探矿、数据化学等新学科和新工艺模式。

4 大数据时代的思维特点研究

现代经济活动领域，通过收集整理的业务大数据能预测出每一个用户的真实消费倾向，可以预判其想要什么产品、喜欢哪类特点、单个消费群体的需求区别有哪些，这些都可以被集合到一起，以便进行分类。通常这类大数据在数量上剧增，经过分析研究，供给侧的提供商会实现从量变到质变的升级。

要想实现产品从量变到质变的升级，需要更好地采用thinking of big data(大数据思维)。著名的科学家维克托·迈尔-舍恩伯格的看法较为合理：即首先需要全部数据样本而不是抽样；其次关注效率而不是精确度；最终关注相关性而不是因果关系[4]。Alibaba的王坚对于大数据也有独特的见解，如：“今天的数据不是大，真正有意思的是数据变得在线了，这个恰恰是互联网的特点[5]。”“非互联网时期的产品，功能一定是它的价值，今天互联网的产品，数据一定是它的价值[3]。”真正实现大数据的意义在于能填补无数个之前未实现的空白并创造新产业[4,6]。

今后，各类企业把采收有效的数据看作是发掘富矿煤田。通常按照煤矿的性质有诸如肥煤[3]、贫煤、焦煤、无烟煤等类别，采集露天煤矿、深山煤矿的挖掘成本很不一样；同理，big data并不在它的“big”，而在于它的“Value”，big data的价值含量和挖掘成本比Volume更重要，这是毋庸置疑的。在投资者眼中，big data就是企业的价值，IT公司Facebook路演时，咨询公司评定的公司有效资产中，Facebook网上的各种数据占据绝大部分。

未来大数据产业中快速提高对各类数据加工的能力将成为实现盈利的关键点。企业必须高效加工大数据才可实现产值的剧增。对于这类体量巨大的各种数据质量来说，为避免混乱急需找到数据之间的相互关联性。

典型案例可以证明数据收集和再处理的重要性：二十多年前，Amazon刚成立，杰夫·贝索斯让几十个书评员来销售书，当时他就知道不仅可以请人写书评，也可用数据技术来提供各类图书的推荐。最初使用类推荐时，用户的体验并不好，通过调查分析后，公司决定不对用户进行分类，而是对用户的需求进行分类，该方法取得了巨大成功。今天，这种推荐系统为Amazon带去了30%的营收。显然该案例证明了：今后，通过研究和分析数据并找出其中的一个关联物并进行监控就能预测未来。

上述案例中Amazon有大量的交易数据，客户每买一本书就是一个交易数据，公司对这个数据进行研究分析后，可以为用户提供后续相关产品的购买意见。然而用户如今已不再满足于已有的交易数据了，因此也大量开始收集起沟通数据等。然而这类数据产生和收集后，其自身并不能直接产生服务，体现其价值的部分在于当这些数据被收集以后，会被相关公司用于不同的目标，这类巨大的数据由此被重新使用，从而产生新的重要商机，重复使用是大数据的另一突出优点。美国著名的实时交通数据提供商Inrix公司已有近1.5亿个手机用户使用INRIX Traffic APP[8]。该应用软件可以辅助开车并规避堵车严重路段，使用时该软件能用不同的颜色呈现交通道路的热量图(显示为红色时就表明有堵车现象)。如果这个产品只提供数据就无特色可言，特别值得关注的是软件并没有利用交警监控系统，INRIX Traffic APP的手机客户端在使用过程中会给公司的运行服务器发送实时数据(走的速度及位置)，这样就成了数量巨大的道路交通探测器[8]。Inrix公司还有更大的秘密是可以重复利用数据[9]：当它了解到周末哪个路段堵车时，哪里有堵车哪里就有更好的销售商机，公司把数据提供给投资方，买方再根据数据对零售业进行新的投资以方便司机同时也促进了零售业的壮大发展，这类提供商在以前从未出现。

由于大数据的使用而带来的IT革命开启了人类重大的时代转型，它正在对人类的思维、生活和工作产生极大的影响。图灵奖得主、关系数据库的鼻祖Jim Gray描绘了20世纪下半叶，由于有科学计算机的产生而诞生了著名的“计算科学”，并研究出了如“银河”、“深蓝”等超级计算机，它们能对以前无法处理的复杂现象进行模拟动态仿真，也可以推演出越来越多的复杂现象，其典型案例如模拟核试验、天气预报等。

(下转第505页)

ty[J]. Innovations in Systems & Software Engineering, 2010, 6(4): 299-310

[10] Alexandrescu R, Bottle A, Min H J, et al. Mining Software Repositories with iSPARQL and a Software Evolution Ontology [C]// International Workshop on Mining Software Repositories, 2007. ICSE Workshops MSR. 2007: 10-10

[11] Robles G, Herraiz I, German D M, et al. Modification and developer metrics at the function level: Metrics for the study of the evolution of a software project[C]// International Workshop on Emerging Trends in Software Metrics. IEEE, 2012: 49-55

[12] D'Ambros M, Lanza M. A Flexible Framework to Support Collaborative Software Evolution Analysis[C]// Csmr. IEEE Computer Society. 2008: 3-12

[13] Emanuel A W R, Wardoyo R, Istiyanto J E, et al. Modularity Index Metrics for Java-Based Open Source Software Projects[J]. International Journal of Advanced Computer Sciences & Applications, 2013, 2(11): 52-58

[14] Nakamura T, Basili V R. Metrics of Software Architecture Changes Based on Structural Distance[C]// IEEE International Symposium on Software Metrics. IEEE, 2005

[15] Le D M, Behnamghader P, Garcia J, et al. An empirical study of architectural change in open-source software systems[C]// MSR. 2015: 235-245

[16] Lehman M M. Laws of software evolution revisited[C]// European Workshop on Software Process Technology. Springer-Verlag, 1996: 108-124

[17] Kouroshfar E, Mirakhorli M, Bagheri H, et al. A Study on the Role of Software Architecture in the Evolution and Quality of Software[C]// Mining Software Repositories. IEEE, 2015: 246-257

[18] Tzerpos V, Holt R C. MoJo: A Distance Metric for Software Clusterings[C]// Working Conference on Reverse Engineering. IEEE Computer Society, 1999: 187-193

[19] 杨芙清. 软件工程技术发展思索[J]. 软件学报, 2005, 16(1): 1-7

[20] 张路, 谢冰, 梅宏, 等. 基于构件的软件配置管理技术研究[J]. 电子学报, 2001, 29(2): 266-268

[21] 钟林辉, 谢冰, 邵维忠. 扩充 CDL 支持基于构件的系统组装与演化[J]. 计算机研究与发展, 2002, 39(10): 1361-1365

[22] 钟林辉, 侯长源, 宗洪雁, 等. 构件化软件演化信息及演化相似性度量技术研究[J]. 计算机应用研究, 2015, 32(5): 1399-1402, 1416

(上接第 473 页)

Jim Gray 认为当今以及未来科学的发展趋势是随着数据量的高速增长，计算机将不仅仅能做模拟仿真，还能进行分析总结得到理论[6,10]。也就是说，过去由牛顿、爱因斯坦等科学家从事的工作，未来可以由计算机来做。这种科学研究的方式，即为今后会高速发展的数据密集型科学。通俗来说，大数据的核心就是预测。实际上，谷歌的广告优化配置、战胜国际围棋大师李世石的谷歌机器人 AlphaGo 也是极其成功的案例，这就是数据密集型科学的魅力所在。

显然，处在这个时代，随着大数据技术的快速发展，人类对科学研究思维方式的转变将会逐步放弃对以往各种传统因果关系的探求并开始重视相关关系的研究。换言之，这将颠覆以往人类的思维惯例，人们将不需要像以往那样知道为什么而只需要知道是什么就行，这对世界交流的方式是一种挑战，对各国人民的认知也是一种挑战。Chris Anderson(克里斯·安德森)2008 年曾发出惊人的断言：The data deluge makes the scientific method obsolete(数据洪流使传统科学方法变得过时)[11]。人类在获得海量数据(即 big data)后，运用各种大数据分析工具，如 Hadoop 等来处理这些数据，将为后人理解世界提供的一条完整的新途径。当前各国投入巨大的人力物力研究开发大数据处理技术，将会为人类创造出不容置疑的变革，这类可量化的维度是前所未有的。大数据及其处理技术即将成为全球人类创造新发明和新服务的思维方式的源泉，越来越多的改革正蓄势待发，世界将由大数据带来巨变。

结束语　大数据的研究成果让人类能够完成没有建立完整的模型和假设，也可以对收集到的数据进行分析。如果将收集的数据录入计算机集群，只要有相互关系的数据，通过专业的大数据分析工具就可以发现过去的科学方法发现不了的新模式、知识和规律。数据量巨大的大数据让我们体会到：Correlation is enough。常规的模型的探索可以停顿，因为因果关系已经被相互关系取代。不管对“数据密集型科学”的理解有多深，必须得承认：数据密集型科学不仅是科研方式的转变，也是人们思维方式的大变化。

参 考 文 献

[1] 范平. 大数据时代，你不得不懂中关村在线[EB/OL][2013-02-18]. http://chuansong.me/, http://blog.sina.com, http://oatos.diandia

[2] 大数据. 维基百科[EB/OL]. [2012-10-5]. http://zh.wikipedia.org/wiki

[3] 王雄. 大数据究竟是什么？一篇文章让你认识并读懂大数据[EB/OL]. [2015-06-18]. http://blog.sina.com.cn/u/2262489275

[4] 维克托迈尔·舍恩伯格, 肯尼思·库克耶著. 大数据时代[M]. 杭州: 浙江人民出版社, 2013

[5] 黄宜华. 深入理解大数据-大数据处理与编程实践[M]. 北京: 机械工业出版社, 2014

[6] 孙晓立. 大数据：让“云”落地成“雨”[J]. 中国科技投资, 2012, (Z2): 43-45

[7] 周傲英. 数据密集型计算-数据管理技术面临的挑战[J]. 中国计算机学会通讯, 2009, 5(7): 50-53

[8] Big data -Wikipedia. the free encyclopedia[EB/OL]. [2013-09-23]. http://en.wikipedia.org/wiki/Big_data

[9] http://v.qq.com/boke/page/y/0/e/y0115diwl0e

[10] 刘念真. 利用 Oracle 信息模型驾驭大数据[EB/OL]. [2014-06-22] http://wenku.baidu.com/view/abfb3a1552d380eb62946d9d.html

[11] 克里斯·安德森. 长尾理论[M]. 北京: 中信出版社, 2006