

大数据的“大”问题

本报记者 计红梅

连日来，2012 年美国总统大选成为信息技术领域大数据研究者们最爱探讨的话题。

与此前的历届选举不同，此次大选伊始，美国总统的竞选团队们就纷纷利用数据分析方法来寻找和锁定潜在的己方选民，并使用最前沿的数字化策略定位并拉拢那些中间派选民。整个过程中，大数据应用的威力可谓发挥得淋漓尽致。

可以想见的是，无论四年一度的美国总统大选投票日（11 月 6 日）之后结果如何，本已位居 2012 信息技术时髦词汇榜首的大数据，热度还将继续攀升。

不过，与外行们的看热闹不同，大数据的研究者们更关心的，则是这一事件体现出的大数据的“大”问题。

“大”数据

每天要处理 25 亿条消息、500 多 TB 的数据、上传 3 亿张照片、每半个小时扫描的数据大约为 105TB……一说到大数据，人们首先想到的便是以社交网络 Facebook 为代表所产生的大量数据。

在近日由中国计算机学会（CCF）主办、大连大学承办的 2012 中国计算机大会上，中国工程院院士李国杰在谈及大数据的定义时，也开明宗义地说到了上述数据所体现出的四个“V”，即体量巨大（Volume）、数据类型多样（Variety）、价值大但密度低（Value）以及处理速度快（Velocity）等特点。

在当天的大会论坛上，加拿大西安大略大学教授凌晓峰表示，信息技术正在以一种难以置信的速度发展，其表现之一就是数据的成倍增长。

如何在信息技术发展的大背景下理解大数据因何为“大”？李国杰告诉《中国科学报》记者，信息技术的发展使得信息采集的成本大大降低，这是数据快速增长的主要原因。根据数据的来源，大数据大略可以分为两类：一类来自物理世界，如天文、生物等研究领域；另一类则来自人类社会，特别是与互联网有关。以脑科学为例，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了1PB。

“有了大量的数据，自然就要挖掘其价值。大数据的出现可以说是自然而然、水到渠成的。” 李国杰说。

然而，在信息技术领域，数据分析的历史远远要比大数据长。以上世纪90年代中期兴起的数据挖掘技术为例，该技术的宗旨就是发现数据中有用的模式，并以之解释当前的行为或预测未来的结果，提供有用的决策信息。

数据挖掘是如何演变为大数据的呢？香港中文大学常务副校长华云生认为，这主要是由于随着信息技术的发展，挖掘数据价值的速度大大慢于数据产生的速度。在数据量越来越大、数据变化又很快的情况下，就诞生了大数据，即在固定的时间内找到所需信息的价值。

“大数据的‘大’是相对的”。美国罗格斯-新泽西州立大学教授熊辉也认为。他举例说，10MB的数据量并不大，但要在1毫秒之

内对 10MB 数据完成复杂的数据挖掘分析，可能超越目前常用设备的数据处理能力。因此，大数据的“大”只是相对的概念，不只是量大，而且对处理的速度也提出了苛刻的要求。

“大” 应用

“目前推动大数据研究的动力主要来自企业的经济效益。”李国杰告诉记者，巨大的经济利益驱使大企业不断扩大数据处理规模，IBM、甲骨文、微软、谷歌、亚马逊、Facebook等跨国巨头是发展大数据处理技术的主要推动者。

而在金蝶国际软件集团有限公司首席科学家、高级副总裁张良杰博士看来，大数据在企业中的应用主要有三个模式。一类是最早的沃尔玛和Visa模式，例如沃尔玛之所以将某些商品放置在一起就是基于数据分析的结果。其次是互联网企业对用户消费行为进行分析，进而更精准地投放广告。数据显示，eBay通过数据分析技术可以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%。

第三种则是把大数据的分析能力放在“云”里，通过数据共享的方式，实现众包服务。对此，张良杰举例说，6000家公司为了解决同一问题，每家公司都需要提供6个分析师。而一家128人的小公司利用云存储和云计算，就可以通过众包的模式接受来自这6000家公司的外包服务，并解决其问题。

“从公司内部的数据挖掘，到互联网公司的消费行为分析，再到

分布式众包模式，大数据的应用处于不断演进的过程中。”张良杰说。

熊辉自1999年开始就介入了对数据挖掘应用的研究。近年来，他与七八家企业在不同领域进行了大数据方面的合作。他的心得是，“大数据应用已如水银泻地，无孔不入。”

“大”平台

正因为大数据的应用已渗透到各个行业，因此，业内专家们已达成共识，大数据的研究一定要依托于某一特定领域和特定行业才能做出贡献。

张良杰举例说，一家美国公司把气象数据放在亚马逊的“云”平台上进行处理，以及时提供气象预报的方式保证当地的农牧业主在特定的地区以及特定的季节不会遭受天气灾害，并由此获得了丰厚的利润。

华云生也认为，解决的问题和目的都不太明确是大数据最困难的地方。以美国总统大选为例，要预测哪些问题，如何预测，都需要与实际的用户相结合。

因此，李国杰呼吁，大数据研究需要倡导“大平台，大联合，大合作”。“搞计算机研究的学者们要放下身段，甘当配角，老老实实地为具体领域的从业者们帮忙。”

李国杰指出，企业对大数据的关注只是着眼于对相关性的研究，可以说是只看现象，不看其背后更深层次的规律。作为一名科学家，则不能止步于此，而是要从各领域具体的技术问题出发，进而发现其背后的科学问题。

2007年，已故的图灵奖得主吉姆·格雷在他最后一次演讲中描绘了数据密集型科研“第四范式”的愿景。他之所以将大数据科研从第三范式（计算机模拟）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。李国杰认为，科研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。

正是看到了大数据对于科学研究的重要意义，作为国内计算机及相关领域专业学术团体CCF的名誉理事长，李国杰担任了CCF大数据专家委员会的主任。

10月20日，聚合了学界、工业界和海外专家的CCF大数据专家委员会在中国计算机大会上举行了正式的成立仪式。李国杰表示，该委员会成立的主要目的，就是为了发现大数据的核心问题，推动大数据的学科发展，并打造产学研用的平台，促进大数据产业的良性发展。