大数据*

译者：刘宇航
中国科学院计算技术研究所

关键词：大数据

2013年ACM奖获得者
大卫·布雷

2007年ACM奖获得者
达夫尼·考勒

ACM会士
维普·库马尔

2014年ACM图灵奖得主
迈克尔·斯通布雷克

在我们的第四个，也是最后一个论坛，我们邀请了2014年ACM图灵奖得主**迈克尔·斯通布雷克** (Michael Stonebraker)、2013年ACM奖获得者**大卫·布雷** (David Blei)、2007年ACM奖获得者**达夫尼·考勒** (Daphne Koller) 和ACM会士**维普·库马尔** (Vipin Kumar) 讨论大数据的趋势。

高德纳[1]估计，目前约有49亿台连接设备（汽车、家用电器、工业设备等）产生数据。预计2020年这个数字将达到250亿。在你看来，这波数据带来的主要挑战与机遇是什么？

维普·库马尔：我们将看到，其中一个主要挑战是，从这些连接设备和传感器收集的数据与我们的大数据社区过去曾经处理的其他数据集是非常不同的。

我们已经看到，大数据的最大成功是在这些应用上，如互联网搜索、电子商务、投放在线广告、语言翻译、图像处理、自动驾驶。能取得这些成功，在很大程度上，是因为有大量且相对结构化的数据集可用于训练各种机器学习算法。但来自大量互连设备的数据在其原始状态可能是高度分散，在空间和时间上都分离开来，并具有非常高的异质性。分析这些数据，对机器学习与数据挖掘社区来说，将是一个巨大的、新的技术挑战。

---

*本文译自*Communications of the ACM*, “Big Data”, 2017, 60(6):24-25一文。

[1] 高德纳(Gartner)是全球最具权威的IT研究与顾问咨询公司，研究范围覆盖全部IT产业，就IT的研究、发展、评估、应用、市场等领域，为客户提供客观、公正的论证报告及市场调研报告，协助客户进行市场分析、技术选择、项目论证、投资决策。——译者注

大卫·布雷：这里的关键思想是，仅仅从诸如Netflix[2]的观看习惯这样的简单数据，不能推荐一个新电影；只有与来自每个人的所有数据一起才能有助于做出推荐。

这是一个激动人心的世界，因为我们正在通过汇总每个人使用其设备的数据来个性化定制我们与设备的交互。当然，这一切都伴随着对隐私的挑战，当我们使得数据可用时我们做出了什么让步，以及我们可以做出多少让步来换取多大程度的个性化定制的能力作为回报。

另一个机遇是通过许多个人组成的巨大集合以一种前所未有的方式来了解世界。这是一个庞大的数据集，沟通、互动和运动的模式，包括所有其他类型的社会和民众以及世界的宏观层面的描述，现在我们都可以获得。

随着越来越多的数据从一个不断增长的设备池被收集，作为个体的人是否失去了信息隐私权？

迈克尔·斯通布雷克：想象一下这个简单的例子：你出现在医生的办公室，并做了X光检查，你希望医生查询一下还有谁和你一样做了X光检查，他们的诊断是什么，患者的发病率是多少。这就需要从本质上整合该国的整个在线医疗数据库，或许还会扩展到多个国家。因为每家医院的连锁店以不同的格式来存储数据，并对常用术语采用不同的编码格式，这是数据集成的一个艰巨的挑战，可是解决这个问题所获得的社会价值将是巨大的。但这也带来了极为棘手的隐私问题，不仅是一个技术问题。因为如果你寻找一个感兴趣的医疗查询，你不是在寻找普通事件，而是在寻找稀有事件，至少据我所知，没有技术解决方案可以做到允许访问稀有事件的同时不间接透露该事件属于谁。

我认为隐私问题基本上是一个法律问题。我们必须在这方面进行法律上的补救。大量的例子表明，将数据聚集在一起将损害隐私。不幸的是，损害隐私带来的社会影响是巨大的。所以，你可以说技术已经使隐私成为一个悬而未决的问题；或者你可以争辩说，保护隐私是一个法律问题。

随着预测模型日益增加的使用，我们如何在解释和使用数据时避免偏差？

达夫尼·考勒：偏差将永远是一个挑战，并没有一个单一的、神奇的解决方案。更大的问题是：我们如何区分相关性与因果关系？在医学上，黄金标准是随机病例对照。在Web数据的情况下，它被称为AB测试。虽然随机病例对照或AB测试之类的黄金标准不完美，但它是一个很好的工具，因为我们能够用其解决一些混杂因素。不幸的是，这种类型的对照不是在所有情况下都可行。过程必须仔细检查，以查出不同的混杂因素，并寻找引起所看到现象的任何和所有的相关性。这是一个需要大量思考和非常小心的过程，其重要性怎么评价都不为过。

例如，有时偏见反映在从数据得出的结论中。在某些网站上进行搜索时，把“Steph”自动补全为“Stephen”而不是“Stephanie”，因为Stephen是一个更常见的搜索术语。有人会说这是性别偏见，应该被消除。作为一个从事技术工作的女性，我当然与那个观点有关且可以理解。或许也有人会说，数据就是数据本身，如果Stephen是比Stephanie更常见的搜索词，那么我们是否真的想使算法做有助于提高用户效率之外的事情？这是一个真正的困境，无论哪种方式都有合理的论据。

迈克尔·斯通布雷克：预测模型的问题在于它们是由人建立的，人天生容易产生偏见。如果我们看看最近的总统选举，我们看到了现有投票模型的一个惊人的失败。事后观察表明，没有人认为特朗普[3]真的能赢，事实上，更可能是投票模型微妙

[2] Netflix是全球十大视频网站中唯一收费的网站。——译者注

[3] 唐纳德·特朗普(Donald Trump)，现任美国总统。2016年11月9日，美国大选计票结果显示：共和党候选人特朗普已获得了276张选举人票，超过270张选举人票的获胜标准，锁定美国总统宝座。——译者注

地对他怀有偏见。

因此，预测模型的问题在于模型本身。如果预测模型中包括欺诈、偏见等，他们可以产生非常坏的答案。人们必须采取半信半疑的态度对待预测模型。我们过去过于相信预测模型。

大数据和机器学习在帮助科学家理解数据（例如，人类基因组计划中的数据）中发挥什么作用，并在健康和医学方面带来哪些潜在的现实机遇？

达夫尼·考勒：我回到医疗保健领域的一个主要原因是，我认为这里有巨大的机遇。随着成本的下降，对新基因组进行排序的能力急剧增长。不只是对基因组，还有转录组和蛋白质组以及其他数据形式。当我们将这些与可以看到表型效果的可穿戴设备结合在一起时，我们可以访问的数据就惊人地爆炸了。其中一个有益的原因是它会提高我们确定引起某些疾病的遗传因素的能力。虽然我们之前也能做，但那时面对数以千万计的基因组的变化，只有几百个例子可被使用，除了最强烈的信号外，就真的很难提取其他的信号了。

是否有几乎肯定会发生的潜在技术突破可以在不久的将来再次改变这个领域？

大卫·布雷：我认为我们处在一个由一些想法激发起来的机器学习和统计学的变革时代的中期。强化学习是一个大的想法。这种想法是，我们可以学习如何面对不确定的环境来行动，而我们行动的后果也是不确定的。强化学习促成了我们看到机器学习和人工智能正在产生许多惊人的结果。而深度学习是另外一个想法：一类非常灵活的学习器(learner)，当给定了大规模数据集，可以识别高维数据中复杂和组合的结构。第三个想法有60年的历史了，就是优化：我有某种函数，我希望得到该函数的最大值，我该如何做到这一点？这被称之为一个优化过程。优化告诉我们如何使用大规模数据集有效地求出函数的最大值。

维普·库马尔：新型传感器和通信技术将具有相当强的革新性。我们今天看到的各种传感器，在几十年前都想象不到。在过去的十年中，移动医疗传感器如Fitbit[4]和苹果手表，能够前所未有地详细记录我们的生理参数。基于电子学、纳米技术和生物医学的进步的新型传感器，已经使得我们能够部署小型廉价的卫星以从前难以企及的空间和时间分辨率监测地球及其环境。如果没有如无线射频识别(RFID)那样的技术，将很难想象可以走进一个商店，通过看或接近一件商品而完成购买。而这在一家位于西雅图、没有收银台的杂货店Amazon Go就可以做到。基于量子技术的新型传感器可能会开辟全新的、甚至今天我们没有考虑到的应用。

还有什么最后要说的想法吗？

迈克尔·斯通布雷克：我们希望从大数据中获取的社会效益取决于对大数据的无缝集成。解决如何提高数据集成度这个问题，将是从所有被创造的数据中获取最大利益的关键。 ■

译　者：刘宇航（作者简介见本刊P50）

[4] Fitbit是美国旧金山的一家新兴公司，以记录器产品闻名于世，通过追踪全天活动、锻炼、睡眠和体重，帮助人们过上健康、平衡的生活。——译者注