大数据及其“社会学后果”

文／张　旭　唐魁玉

摘要：随着“大数据”一词逐渐被人们所熟知，各学科的研究者们也开始应用大数据进行研究。社会学者已将大数据纳入社会研究的范围，而且有逐渐扩大或蔓延的趋势。大数据对社会学的影响，既体现为正面后果又体现为负面后果。即大数据思维的运用，对社会学研究来说既具有方法论的意义，可以激发社会科学研究的认识论变革，同时也存在着因大数据观念的引入而产生的社会学方法论的局限性。大数据为社会学研究打开了一扇新的大门，但是这些研究方法只能作为传统社会学研究的补充，而不能完全替代传统的小数据研究方法。

关键词：大数据；社会学后果；社会学研究方法；方法论创新

中图分类号：C91—03　**文献标识码**：A　**文章编号**：1006—0138(2016)03—0042—06

近几年来，随着计算机科技的进步，“大数据”一词也逐渐被人们所熟知。这种大数据的变革为社会学研究带来了改变的机遇以及对传统社会学研究方法论的挑战。一直以来，有限的样本量是社会学研究的瓶颈。即使可以收集到大量数据，对这些数据的记录、储存和分析也被当时的技术所限制。在预算范围内，研究者们追求着合理抽样方法和样本量的完美组合。而在大数据时代，云储存和云计算使得对大量数据的记录、储存和分析成为了可能。社会学家们也与时俱进，将大数据纳入到社会学研究中；但是，要想将大数据应用于社会学研究中，无论是研究者的思维还是研究方法，都需要进行一些转变。

一　大数据对传统研究方法的冲击

20 世纪后半叶是实证社会学的黄金年代。20 世纪 50 年代到 90 年代间，实证社会学通过抽样调查和访谈将其他学科远远地抛在了后面。但是近十年来这种优势正在逐渐消失。抽样调查和访谈等传统社会学研究方法已经无法使社会学继续伫立在人文科学的塔尖。

萨维奇（Savage）和布罗斯（Burrows）早在 2007 年就发表了一篇论文用以提醒社会学家们注意这种危机并采取相应措施应对危机。[1] 这篇论文在全球社会学界获得了较大的关注并且得到了广泛引用。萨维奇和布罗斯也

作者简介：张旭，哈尔滨工业大学社会学系博士研究生，哈尔滨市，150001；唐魁玉，哈尔滨工业大学社会学系教授、博士生导师，中国网络社会学会副会长，《哈尔滨工业大学学报》（社会科学版）副主编，哈尔滨市，150001。

是在实际的研究过程中发现了实证社会学的危机。萨维奇在2004年参加了由ESRC提供基金的关于社交网络研究方法的项目。研究者们通过对来自三个机构的成员进行问卷调查，研究三个机构成员之间的私人联系。研究者们耗费了大量的时间来分析问卷数据并对部分受访者进行访谈以了解更多的细节。而项目中的一名非正式研究者则通过非常简单的分析就获得了研究结果，只因为他是一家知名电信公司的员工，并且该公司拥有这些受访者多年来的通话记录。一名社会学外行人仅仅通过大量数据和简单的统计分析就完成了与社会学家们耗费大量时间和资源所完成的同等的研究，甚至获得了比社会学家们更精确的结果，仅仅是因为他拥有大量的数据。布罗斯也是在研究中意识到了实证社会学潜在的危机。2005年，布罗斯也在一次实地的研究中发现将已经存在的公共数据资源（如人口普查等数据）集中在一起，可以迅速地绘制出某一区域的精密的社会—空间地图。如果忽略隐私等道德问题对研究者们的限制，这种社会—空间地图可以呈现一定等级内的细节信息，并在间隔尺寸一定的情况下将地图的范围进一步扩大。

对于定性研究方法，如深度访谈，最初并不为社会学家所应用，更多的则是被社工和心理学家们所应用。直到后来，实证社会学兴起，社会学家们发现一些有影响力的人可以代表一个广大群体的看法，而对他们进行访谈则可以有效率地得知这部分群众的看法。毫无疑问，这种方法在技术不发达的年代，可以有效率地收集数据。并且，通过更大覆盖面的问题，可以获得一些小范围的特质概括用以作为将来大范围定量研究的假设。但是在现在，每天产生的大量基于网络平台的交互性数据完全可以收集到比访谈更加丰富的数据，只要技术手段可以达到，对这些数据的定性分析完全可以达到数倍于访谈的效果。[2] 而对于各种人文类学科一直以来都在应用的史料分析方法，应用计算机技术等手段，无疑可以一次分析更大规模的资料。而且，一些大公司（如Google）这些年致力于将纸质资料数字化，更为这种大规模的史料分析奠定了基础。

涂子沛在其风靡全国的专著《大数据》的封面上写到："除了上帝，任何人都必须用数据说话。"[3] 在数据如此丰富的现在，除了一些坚持传统社会学研究方法的小部分社会学家，大部分社会学家已经开始将大数据纳入到研究范围中，并开始尝试一些革新的研究方法以适应新的数据。同时，这些变革也正在为社会学重新回归"社会事实"奠定方法论基础，而这也是社会学重新回到领先地位的绝好机会。

二　大数据的正面后果：研究方法的变革与创新

大数据现在被引申为关于某个特殊平台或某个特殊领域的全部数据。对于社会学研究来说，一些特殊的平台，例如Facebook和Twitter，以及和他们具有相似功能的我国的人人网和微博，具有极大的意义。这些数据是动态的，体现着实时的社会活动，并且这些数据记录了人们在自然环境下所说的话和所做的事，而不是像常规问卷调查和访谈中获得的那些僵化了的信息。[4] 同时，这些实时性的信息可以提供有关网络信息传播的速度以及方式和方向。区别于这种来自于某个特殊平台的数据，来自于一些特殊领域的数据则包括更大范围的信息，例如Google曾经应用往年搜索结果建立数学模型用来预测流感疫情，以及奥伦·艾奇奥尼（Oren Etzioni）应用以往的机票价格预计机票价格的涨跌。大数据的兴起使社会学研究向更广泛的方向发展，大量的网络数据也随之被应用。据统计，1995至2008年间，随着互联网的广泛传播，基于文字的网页增长了6600万，并且还在持续增长，最近已经增长了超过1万亿。[5] 这些网页内的信息无疑可以作为社会学研究的数据，但是社会学研究并不仅仅只能应用这类数据。陈云松应用Google图书的最新语料库进行关键词的词频统计，用以阐释19世纪中期以来社会学各方面的发展。[6] 龙瀛及同事使用北京1周间产生的855万个公交IC卡的数据结合市民出行情况及城市地图及土地利用信息，分析了市民的职住关系和通勤行为。[7] 这类已经被收集完成的资料可以成为社会学家们的研究对象并进行分析和再利用，同时，另一部分研

究者们选择亲自收集资料用于研究和分析。

（一）收集数据

区别于以往社会学研究中的抽样方法，在大数据背景下的数据收集需要有所变革才能应对收集全部数据的要求。针对来自网络社交平台这类特殊平台的数据，可以根据他们本身提供的功能进行收集，同时，一些平台提供专用软件用来收集数据。以Twitter为例，它向用户提供“发表”“转发”“回复”这些针对微博客的功能，“关注”“取消关注”“提起”这种针对用户的功能，以及有助用户发表相关话题的“标签”功能。由于Twitter的完全开放性，研究者们可以获得某一用户发布的所有微博客，[8]在相关话题标签下的全部微博客，[9]以及通过搜索功能搜集所有包含关键字的微博客。[10]研究者们也可以通过应用程序接口（Application Programming Interface，简称API）进行数据收集。API可以实现几个方面的功能：(1）通过搜索关键词和话题标签的微博客收集；(2）在所有微博客中抽取10%作为随机样本；(3）收集所有已发布的微博客。自2008年，Twitter获得了研究者们的广泛关注，也有很多论文发表，但是只有极少数发表在了主流期刊上。[11]我国的微博（新浪微博、腾讯微博等）也具有极其相似的功能，唯一与之区别的是在应用程序的接口上并未完全开放，可能无法达到以上描述的全部功能。

即使在大数据时代，研究者们主张收集全部数据，而不再完全依靠随机样本，收集所有数据再筛选出需要的信息也是极大的工程，因此，即使是收集全部信息也需要一些相应的方法。例如有研究者将滚雪球抽样方法与计算机技术结合形成了一种适用于大数据时代的数据收集方法。[12]研究者首先输入一个起始网页，并规定关键字或者对搜索目标更细节的描述，而后该程序将访问每一个与起始网页相关的网页，如遇到与关键字相关的网页将提取出文字资料，并在当前网页重复之前过程。如果程序一直运行下去，将得到一种类似蜘蛛网的扩散结构。但是由于计算机的硬件限制，这种过程无法一直持续下去。而且由于网页的互相关联性，在几轮之后，可能出现相关网页在之前已经被搜索过的情况。因此，用户可以通过规定多个关键词以及限定搜索的轮数，得到一个接近于整体的样本。

（二）整理数据

由于在大数据时代，数据量极大，收集到的数据无法保证一致性。而我们如果追求大量的数据，就必须接受随之而来的数据的驳杂性。在分析数据之前，研究者们不可避免地需要整理这些数据，减少驳杂性，使它们具有可分析性。因为通过互联网收集的数据大多是文字性的，无论是定性还是定量的研究，都需要将这些信息加以整理。如果是定量的研究，数字性的信息可以结合某些计算机技术手段筛选后使用，而文字信息则可以通过对信息进行量化使用。陈云松关于社会学发展的研究，就是通过统计Google最新语料库中的某些社会学关键字的词频来阐述社会学自19世纪中期以来的发展。[13]

由于计算机科学与社会学的学科合作并没有非常深入，所以现在大部分的对于文字信息进行定量研究都只是简单的统计频率。如果社会学科可以和计算机学科进一步合作，研究者们可以获得更加丰富的数据，并在一定程度上取代一些样本量较大的问卷调查。如美国的综合社会调查（General Social Survey，简称GSS），通过在全部成年美国公民内抽取3000人左右的样本，进行登门问卷调查。自互联网兴起之后，一些研究者们也在网络上发布问卷调查，但是由于网民群体并不是特别具有代表性的公民样本，因此结果也无法具有特别高的代表性。如果研究者们收集来自社交平台的数据，筛选出关于某些话题的信息，再应用计算机技术分析情绪并将之量化，可以得到民众关于某些话题的看法。但同时，这种方法也被样本代表性限制着，只是由于数据量较大，且这些社交平台的用户数量在持续增长，可以弥补部分代表性的缺点。

如果是定性的研究，则需要先根据关键字对信息进行分类。对于这种极大量的数据，全部依靠人工分类显然不可能。依靠计算机手段也有诸多的缺陷。比如有些词有许多同义词或者类似含义的表述，而研究者们很难将这些表

述收集完全，因此在未阅读这些信息前，研究者们很难选出可以足够合适并完全的关键字，并且由于这种方式忽略了句子中的大部分成分，很容易造成语义理解错误。基于这种困境，研究者们采取了一种新的方式，通过“机器学习”进行主题建模。这种方式产生于社会学家、语言学家以及计算机学家的合作。这种方式通过对主题的描述寻找几个词汇同时出现的概率，进而进行分类。[14] 即使这样，对信息进行分类依然会产生错误。比如，这种方法假设顺序无关紧要，包括词汇的顺序和在极大样本中文字篇章的顺序。而且，这种方法对分类完全采取单一结果的方式，一段文字信息只能对应一个关键字，而忽略了一些关键字之间的联系。为了克服这些限制,有研究者提出了在“人工前导”下的主题建模。研究者们先从广泛的数据中随机抽选出一些篇章进行人工分类，并将结果作为机器学习的训练样本。霍普金斯（Hopkins）和金（King）采用了这种方法进行了研究。他们通过在 2008 年美国总统选举期间的几千篇相关博客分析群众们对候选人的看法。经过学习了一些训练样本后，计算机的分类结果比人工分类更加精准。[15] 但是对于这种观点，并没有更进一步的验证。

（三）分析数据

在数据被整理之后，研究者们将使用这些数据进行分析。对于定量研究来说，整理后的数据已经可以通过相应统计软件进行直接使用，对于定性研究来说,数据依然需要继续处理。“主题建模”等方式依然可以继续使用，用于将信息进一步细化并摘取出有意义的片段。比如在霍普金斯和金的研究中，在将博客按候选人进行分类后，还要对人们对候选人的态度进行分析，并摘录出关键语句。[16] 现在的社会学，计算机学和语言学的合作还只能将这些有关于态度的内容进行简单分类，而对于更深层次的应用则需要各学科之间更加深入的合作。而对于定量研究来说，分析数据也包括将现有数据制作成图表，以便于更加方便地阐述研究结果。比如缇娜提（Tinati）及同事通过统计在学费抗议期间 Twitter 上面微博客转发数分析网络数据的传播规律。在数据整理之后通过软件生成了散射状的信息流动图，[17] 这种通过统计极大数据而生成的图表，如果不使用相关软件基本无法实现。

三　大数据的负面后果：更新研究方法的局限性

对于大数据的研究虽然将社会学推上了一个新的高度，但是却依然有其局限性，并不能完全取代传统的实证社会学研究方法。

首先，以定量分析方法抽样调查为例，在一些案例中，抽样调查更加适用于那些有“遗失”的数据和代表性的样本。比如，一些没有被警察发现的犯罪记录。为了保护自己，人们一般不会在社交网络平台上袒露自己的犯罪记录，尤其是那些警方没有掌握证据的犯罪记录。这些记录就是数据库中大量缺失的数据。而且，如果真的有人存在犯罪未被发现的情况，他们一般倾向于不在公共场合和平台上坦白自己。在这种情况下，社会学家们可以通过统计方法估计这部分缺失的数据，从而预测整体的行为特点，而并非依赖于对全部所收集到的数据进行分析。另外，社会学家们倾向于使用来自社交平台的大数据，然而由于不同社交网络平台的用户群体在族群背景、教育、收入等方面都有所差异，在某一平台收集的数据并非如众多社会学家所想象的那样，可以代表某一概念下的整体。[18]

其次，虽然当前的技术水平已经足够研究人员们储存和分析如此大量的数据，但是对于普通的研究者来讲，如此大规模的运算还是比较困难的。大数据的收集是一个费时费力的工作，需要大量的资金支持。除了谷歌和微软这样的大型 IT 公司，只有那些像沃尔玛这样的大型商业公司才有这样的实力。也曾经有社会学家收集过一些来自于社交网络平台的数据，但是由于技术和资金限制，这些数据的规模无论是纵向还是横向都远远无法和那些大公司的数据库相比。[19] 社会学家也可以应用政府、机构、和企业已经收集整理好的数据库。但是，除了政府的开放数据和一些机构的免费数据，大部分数据都需要研究者们向数据的所有者购买。对于研究经费有限的社会学家们，这可能

是一笔不菲的支出。而且，通信领域和社会媒体领域的企业通常拒绝或者限制向研究者们分享数据。[20] 即使企业统一分享数据，由于机构和企业收集数据时并没有针对社会学研究的需要，所以一些数据库可能并不能完全适用于社会学研究。因此，社会学家研究大数据的最理想选择还是自己收集数据或者使用其他社会学家或者社会学研究机构已经收集好的数据。这无疑需要与其他学科，尤其是计算机学科更大程度的合作与交流。而对于整理和分析这些数据，对传统的社会学方法也具有很大的挑战。因为数据的巨大规模和驳杂性，用人工去整理和分析这些数据几乎是不可能的事情。于是社会学家们同样也需要更加先进的方法去处理这些数据。

第三，在大数据时代，道德也成为大数据社会学研究的限制。在传统社会学研究中，研究者必须先得到受访者的“知情同意”后才能进行数据的收集，即数据收集者必须告诉受访者，有哪些数据将要被收集，这些数据将用来做什么，在受访者对研究过程充分了解的基础上方可进行。虽然这并非是数据收集的唯一方式，但已经成为了基于隐私政策的共识性基础。然而在大数据时代，许多数据在收集的时候并无意用作其他用途，而最终却产生了许多创新性的用途。许多研究也验证了，大数据的价值不再单纯地来自于其基本用途，更多源于对它的二次甚至于多次利用。经常被作为数据收集对象的各种网络社交平台，虽然在用户开始使用服务之前通过一些使用许可，但这些简陋的许可并没有规定这些数据的具体用途，同时由于冗长的篇幅，很少有用户将这些许可全部阅读。这些都造成了大数据时代的社会学研究缺乏对受访者隐私的严格保护。而由于IP地址的唯一性，用户很容易通过IP地址被追溯，而这显然不利于研究者们对受访者的保护。另外，收集好的数据集可以作为资源出售，这也使得某些平台在利益的驱动下非法获得用户的隐私信息并用以进行商业销售，而购买者对这些数据的应用并不被出售者和数据相关者所掌控，从而对用户的正常生活产生不良的影响。[21]

第四，使用大数据之后也存在着对公平性的影响。以行车保险业为例，很多保险公司在用户的车辆中装载行车记录设备，这些设备记录了用户的行车路线、驾驶习惯等等。保险公司将这些记录与用户的违章记录合并，对用户的交通事故概率进行预测，并以此确定用户应当缴纳的保险费用。交通事故概率的预测值越高，则用户需要缴纳的保险费用就更高。乍看之下，这并没有什么问题，但是，在这其中，弱势群体的劣势被进一步累计。具有更高经济等级的人可以选择在工作地点附近居住，或者选择在更好交通条件的地点居住，他们的上下班时间也更加方便他们的驾驶，从而他们拥有更高的驾驶安全系数。而那些低经济等级的人，可能住在距离上班地点比较远的地方，所行驶的道路和上下班的时间也不利于安全驾驶，因而只有较低的驾驶安全系数。由此，低经济等级的民众需要缴纳更高的保险费用，而高经济等级的人反而需要缴纳较少的费用。这从另一个方面进一步拉大了贫富差距，导致了更加严重的社会不公平。

不仅如此，大数据研究的强大预测性也可能导致更加严重的问题。如果警察应用大数据的预测来预防犯罪，一些素行不良的人可能会因为尚未发生的犯罪得到惩罚，这无疑是不公平的。而社会学研究所应用的大规模交互性数据可能会包含某些人有犯罪意图的信息，是否将这些信息提交给警方也将是社会学研究道德的一部分。总之，大数据为社会学研究建立了一个全新的国度，而这个国度的道德规范还没有建立完全。

四 结 论

大数据对社会学的影响，体现为两种社会学后果：正面后果和负面后果。它改变了社会学研究方式，开创了社会学研究的新时代，但是这并不代表以抽样调查和访谈作为根本的传统社会学研究方式从此退出了历史舞台。由于研究基金和技术等方面的限制，在很长一段时间内，大部分社会学家们还将继续使用传统社会学研究方法。也许今后计算机科学技术不断发展，同时社会学与计算机学科和语言学进一步合作，基于大数据的社会学研究可以进一步

增大范围。但是由于社会学以社会为研究对象，即使在研究基金充足、技术也达到标准的情况下，依然有研究要依赖传统的抽样调查与访谈的方法。因而，大数据社会研究方法并非是传统社会学研究方法的替代，而是补充。

注释：

[1]M.Savage, R.Burrows, "The Coming Crisis of Empirical Sociology", *Sociology*, vol.41 (October 2007), pp.885—899.

[2]R.J.Smith, "Missed Miracles and Mystical Connections: Qualitative Research, Digital Social Science and Big Data", *Studies in Qualitative Methodology*, vol.13 (2014), pp.181—204.

[3]涂子沛:《大数据：正在到来的数据革命，以及它如何改变政府、商业与我们的生活》，桂林:广西师范大学出版社，2015年，第4—12页。

[4]R.Inati, S.Halford, L.Carr, C.Pope, "Big Data: Methodological Challenges and Approaches for Sociological Analysis", *Sociology*, vol.48 (2014), pp. 663—681.

[5]C.Bail, "The Cultural Environment: Measuring Culture with Big Data", *Theory & Society*, vol. 43, no.3/4 (2014), pp.465—482.

[6]陈云松:《大数据中的百年社会学》，《社会学研究》2015年第1期。

[7]龙瀛、张宇、崔承印:《利用公交刷卡数据分析北京职住关系和通勤出行》，《地理学报》2012年第4期。

[8]R.Waters, J.Williams, "Squawking, Tweeting, Cooing, and Hooting: Analyzing the Communication Patterns of Government Agencies", *Journal of Public Affairs,* vol.124, no.4 (2011), pp.353—363.

[9]N.Jackson, D.Lilleker, "Microblogging, Constituency Service and Impression Management: UK MPs and Their Use of Twitter", *Journal of Legislative Studies,* vol.17, no.1 (2011), pp.86—105.

[10]A.Larsson, H.Moe, "Studying Political Micro—Blogging: Twitter Users in the 2012 Swedish Election Campaign", *New Media and Society*, vol.14 (2011), pp.729—747.

[11]D.Murthy, "Towards a Sociological Understanding of Social Media: Theorizing Twitter", *Sociology*, vol.46 (2012), pp.1059—1073.

[12]A.Gong, "An Automated Snowball Census of the Political Web", http://papers.ssrn.com/sol3/paper.cfm?abstract_id=1932024. SSRN eLibrary, 2011.

[13]陈云松:《大数据中的百年社会学》，《社会学研究》2015年第1期。

[14]K.M.Quinn, B.L.Monroe, M.Colaresi, M.H.Crespin, D.R.Radev, "How to Analyze Political Attention with Minimal Assumptions and Costs", *American Journal of Political Science*, vol. 54 (2010), pp.209—228.

[15]D.Hopkins, G.King, "A Method of Automated Nonparametric Content Analysis for Social Science", *American Journal of Political Science*, vol.54 (2010), pp.229—247.

[16]D.Hopkins, G.King, "A Method of Automated Nonparametric Content Analysis for Social Science", *American Journal of Political Science*, vol.54 (2010), pp.229—247.

[17]R.Tinati, S.Halford, L.Carr, C.Pope, "Big Data: Methodological Challenges and Approaches for Sociological Analysis", *Sociology*, vol.48 (2014), pp. 663—681.

[18]"国内外新闻与传播前沿问题跟踪研究"课题组:《大数据时间与研究：批判性反思与研究推动》，《新闻与传播研究》2015年第8期。

[19]A.Edwards, W.Housley, M.Willianms, L.Sloan, "Digital Social Research, Social Media and the Sociological Zmagination: Surrogacy, Augmentation and Re—orientation", *International Journal of Social Research Methodololgy*, Vol.16 (2013), pp.254—260.

[20]孟小峰、李勇、祝建华:《社会计算：大数据时代的机遇与挑战》，《计算机研究与发展》2013年第12期。

[21]沈浩、黄晓兰:《大数据助力社会科学研究:挑战与创新》，《现代传播》2013年第8期。

责任编辑　刘秀秀