大数据时代下中国社会调查的科学新观

顾佳峰
北京大学中国社会科学调查中心，北京 100871

摘要
大数据已经成为这个时代的显著特征，大数据的发展为入户调查数据带来了极大的冲击和挑战。在这种情况下，社会调查需要有新的基于中国古老智慧的管理理论，并且把大数据和云计算等都纳入社会调查系统，使其成为社会调查运作系统的有机构成部分。利用大数据分析技术，对社会调查过程中的行为数据进行分析和利用，可以大大提高社会调查的精准度，有效实施社会关系的精准管理。最后，对于大数据和调查数据的未来发展提出了几点看法。

关键词
大数据；社会调查；大智慧；行为数据

中图分类号：C915 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016016

Social science research in China social surveys under the big data revolution

GU Jiafeng
Institute of Social Sciences Survey, Peking University, Beijing 100871, China

Abstract
Big data has become a significant feature of this age. The development of big data brings great impact and challenge to social surveys. To face the challenge, a new management theory based on China's traditional wisdom of social surveys is needed. Big data and cloud computing should become the constituent parts of total survey management system. The big data analytics can give insights of paradata, which can improve the accuracy of social surveys significantly and implementation of precise management of social relations. Finally, some views on the future development of big data and survey data were proposed.

Key words
big data, social survey, big wisdom, paradata

1 大数据时代的基本特点

1.1 大数据的基本特点

大数据和传统意义上的数据有何区别？这是所有关心大数据的人必须回答的问题。关于大数据的定义，有两种说法：其一，大数据就是数据；其二，大数据不是一般的数据。这种界定，有点辩证哲学的味道。事实上，上述说法都对，但不解决问题。大数据具有典型的特征，可以体现在“4V”之上。大数据具有体量上的特征，就是数据量大（volume），大到连“海量”、“巨量”都无法来形容。大数据一般都不是静止不动的，而是时时刻刻都在变化的，而且变化速度很快（velocity）。比如互联网上的数据以及人体生物信息，时时刻刻、分分秒秒都在变化，这么高速变化的数据要求新的分析方法。大数据的变化性更大（variety），要在动态变化的情境下捕捉到大数据背后的规律，传统的数据分析方法就会显得力不从心了。大数据中的内容是与真实世界中的发生息息相关的（veracity），因此，对于大数据的分析，本质上就是要透过数据迷雾，看到现实世界的客观发展规律和本质。唯有如此，大数据分析才有意义。

1.2 大数据时代的挑战

大数据的兴起，对传统意义上的“小数据”形成了很大的冲击。过去，社会问题的诊断和公共决策大多依赖于通过调查收集上来的数据和信息。由于受到调查样本量的限制，这类调查数据的量是有限的。大数据兴起后，这类调查数据首当其冲，受到了很大的冲击。2015年秋季，Meyer B D等人在《经济展望杂志》上发表了一篇《危机中的入户调查》，引起了社会调查界的高度关注[1]。在这篇论文中，提出了一个很重要的观点，就是通过入户调查来收集数据的方式已经遭遇到了前所未有的挑战，入户调查的无响应率（nonresponse rate）甚至高达30%~40%。在这种情况下，入户调查的成本会显著上升，使得入户调查越来越成为一种不经济的数据收集方法。于是，调查机构纷纷通过创新和转型来获得在大数据时代下的生存权。在这种趋势下，调查数据和大数据相结合的混合数据收集模式应运而生，成为了一股新的力量。

2 大数据时代的社会变革与研究

2.1 社会变迁的数据测量

大数据时代的到来，从深层次影响着社会的发展与转型。中国人越来越离不开智能手机、互联网等，几乎生活的每个环节都与大数据或“小数据”有关联。这种社会变革和转型，对社会科学的研究提出了更高的要求，也提供了绝好的研究机会。美国科学院院士、普林斯顿大学著名的社会学国际权威谢宇教授，曾为笔者的专著《调查机构管理：理论与实践》一书作序，他写道：中国正在经历一场急剧、大规模且不可逆转的社会变革，这场变革给社会科学研究提供了前所未有的良好机遇[2]。北京大学召集了包括社会学、人口学、经济学、公共卫生学等近20个社会学科的海内外专家，在2006年成立了北京大学中国社会科学调查中心（Institute of Social Science Survey, ISSS），通过实施全国性的中国家庭追踪调查（China family panel studies, CFPS），系统性地收集旨在刻画中国社会变

迁的微观数据，为政府决策和社会科学研究提供重要的数据支撑[3]。这个中心刚成立时，只有两位创始者，而笔者很荣幸就是其中之一，参与了这个中心筹建、发展、壮大的全过程，也见证了中国第一个家庭入户跟踪调查项目的酝酿、设计、测试、实施和发展壮大的过程。目前，该数据已全部免费向社会开放，数据使用者通过ISSS官方网站（www.isss.edu.cn）注册后，就可以申请获得数据。调查中心还通过微信公众号（中国民生观察）及时发布数据信息。

2.2 测不准定律与社会调查研究

物理学上有一个测不准定律，不管用人或再怎么精良的仪器测东西，一定会有误差。在CFPS设计过程中，在控制测量误差上下足了功夫。调查设计、抽样、问卷设计、执行、质量监控、数据清理等所有环节，都尽可能减少误差，提高调查数据的精准度。在社会调查理论上，西方有所谓的调查总误差（total survey error）理论。这个理论在传统的非大数据时代中比较适用。但是，当中国同时经历大数据的洗礼和剧烈的社会变革与转型时，继续沿用西方的调查总误差理论，通过社会调查去测量社会变迁就会出现较大的误差。因为大数据时代的到来，令社会信息和社会数据传播方式发生了重大改变。大数据时代的社会测量，需要有对应的调查方法。

大数据有时候会被误解，认为只要数据的量大，就称之为大数据。其实不然。大数据的“大”，主要指的是数据所包含的信息意义重大。所以，有些数据量并不大的“小数据”，其实也是名副其实的大数据。中国古时候有个成语——“微言大义”，说的就是这个意思。数据量很小，但是内涵和意义却非常丰富。这类数据，也是大数据。所以，中西方对于大数据的理解，其实是有细微差别的。西方的大数据，主要从量上来讲，因为数据存储技术的不断升级换代，使得存储和分析海量数据成为可能。中国的大数据，更多地强调数据所蕴含的信息。《韩非子·说林上》云：“圣人见微以知萌，见端以知末，故见象箸而怖，知天下不足也。”这说明，即便是“小数据”，智者也能见微知著，看到微小的苗头，知道其中的规律，预测出可能会发生的显著变化。

其实，今日的大数据思想，早在《易经》中就有体现：“仰观天文，俯察地理，近取诸身，远取诸物，乃作八卦”意思是说，由天文、地理和人文大数据信息汇集在一起，才形成了八卦。所以，对于这类包罗万象的大数据的分析和挖掘，要上观天文，俯观地理，中看人文，这就是古代的大数据挖掘技术。在《黄帝内经》中，已经提出“大数”的概念。当然，中国古代朴素的“大数”与现代的“大数据”在技术和分析方法上是不同的。但是，在基本思想上是相通的，都是试图通过对现象和数据的分析来把握事物发展的客观规律。在这种大数据思想的指导下，笔者根据社会调查的实践，提出了全面调查管理（total survey management，TSM）理论，以期通过社会调查的有效管理，尽可能减少社会调查的测量误差，提高社会监测的精准度。这个社会调查理论把整个社会测量实践分成阴、阳两个层面，如**图1**所示。阳的层面是调查管理的5种核心主题：使命愿景、公共关系、督导关系、访员关系和访问关系。阴的层面是调查管理的5种核心力量：战略力、组织力、凝聚力、执行力和控制力。调查机构通过对阴阳消息的平衡把握，实施社会调查和社会监测项目与活动，确保实现测量误差最小化。

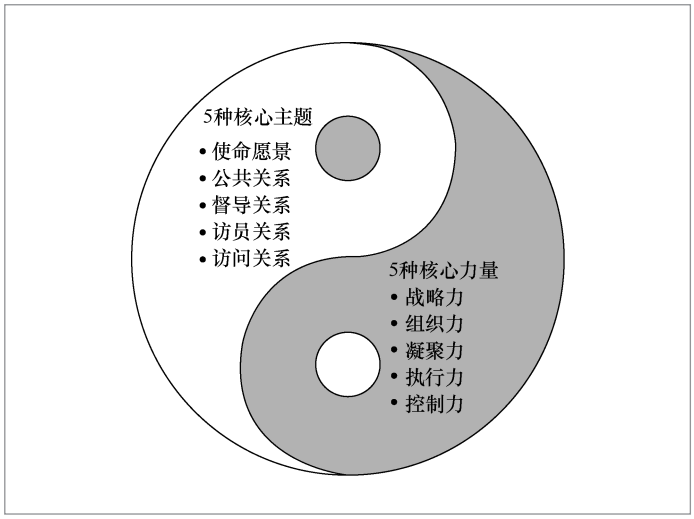

图 1 全面调查管理的基本结构 [2]

3 大数据技术在社会调查中的应用

3.1 行为大数据及其应用

根据TSM理论，任何数据都包含阴阳消息。因此，在社会调查的设计和执行过程中，要同时对两方面的数据进行管理。社会调查的问卷数据是阳层面上的数据，也是社会调查所需要收集的目标数据。但是，要降低这些数据收集的误差，就需要同时收集另一部分数据，就是行为数据（paradata）。调查过程中的行为数据一般都是隐秘不公开的，仅仅作为内部管理和质量监控之用，所以可以归于阴层面上的数据[4]。在CFPS项目的执行过程中，采用的是计算机辅助面访（computer assisted personal interviewing，CAPI）系统。当访员入户打开调查专用的笔记本电脑进行调查时，收集行为数据的软件就开始启动。访员在用笔记本进行调查的每个动作数据，都被同步纪录了下来。**图2**为调查数据和行为数据的同步收集系统。

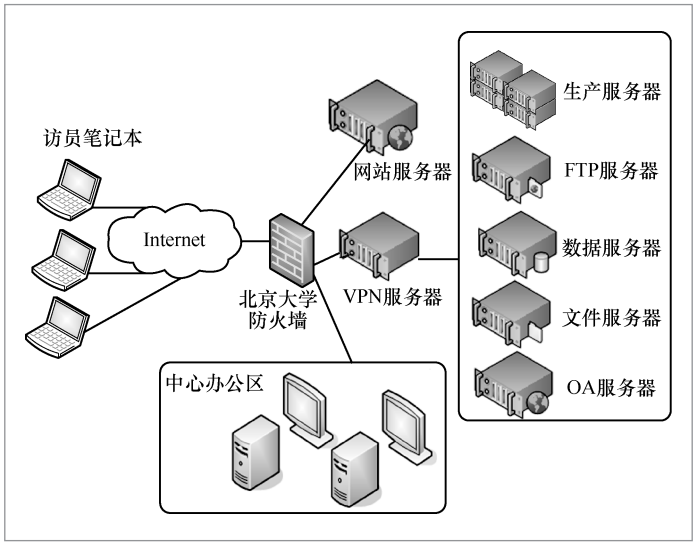

图 2 调查数据和行为数据的同步收集系统

根据**图2**的架构，整个社会调查的数据流都是整合在一起的，在信息系统中进行及时传输和共享。调查数据和行为数据经过传输后，进入不同的数据库进行存储，并用于不同的用途。调查数据收集上来后，就是层层数据质量的查核。行为数据收集上来后，主要用于访问管理。社会调查过程的行为数据，包括了方方面面的信息，比如访员的地理位置信息以及移动的空间路线、访员敲击笔记本电脑键盘的信息、每道题所问的时间长短信息、中间停顿时间信息等[5]。所有这些行为数据都客观如实地记录了身处在调查现场的访员的一举一动，让访员的行为可控，进而确保把访问误差控制到最低程度。

3.2 云计算与访员行为管理

一旦行为数据采集进来，基于云计算的大数据分析就自动启动和运转。例如，当每个访员的键盘行为数据采集起来后，系统的云计算就可以通过分析每个访员的敲击键盘的特征，识别出每个访员的用指习惯，从而自动识别出是否为指定访员在通过笔记本进行入户调查。因为每个人用手指敲打键盘的方式是不同的，体现在

键盘上，就可以清晰发现在键盘敲打的力度、持续的时间等方面，每个人都会有一种独特的模式。基于云计算的大数据分析，能够通过键盘敲打的行为数据，从中找出个性化的用指模式，进而可以精准识别出是给定的访员在用笔记本做调查，还是冒充访员的人在用笔记本做调查。由于这些行为信息的采集是在访员并不觉察的情况下进行的，因此，这些行为数据的可靠性极强。即便访员意识到有键盘使用行为采集系统在收集信息，想要刻意去制造噪音，以混淆键盘使用信息，但是实际上这很难做到。因为每个人的用指习惯是很难改变的。

图3显示的是基于云计算的数据链管理系统。这是一套实时联动的无缝大数据系统。比如，当在调查现场的访员使用键盘时，基于云计算的数据分析系统发现该访员的键盘使用与过去一贯的模式不同，大数据分析系统就会给出警示，建议督导及时查核这名访员，确定使用该笔记本进行入户调查的人的真实身份，避免他人冒充访员进行调查的情形出现。这套大数据系统不仅能够识别笔记本电脑的访员身份，而且还能精准测量访员的调查访问状态。访员的个人情绪，往往会影响调查访问的数据质量。为了提高调查数据的精确度，减少访问过程中的人为误差，都需要访员按照规定的调查行为标准开展入户调查，尽可能减少访员的个人因素的干预和影响。例如，在访问过程中，访员的情绪大幅度波动，往往会影响调查数据的质量。因此，一般都要求访员在访问过程中保持情绪平稳，心平气和地完成调查。基于云计算的键盘使用模式分析系统能够对所有访员的键盘使用大数据进行分析，提炼出若干典型的情绪模式，比如激动、愤怒、压力、害怕等[6]。一旦某个访员在键盘使用上出现这些负面情绪特征，相关的督导就需要予以注意，及时和该访员进行电话沟通，第一时间安抚访员的情绪，并鼓励其继续按照预定计划完成调查目标[7]。

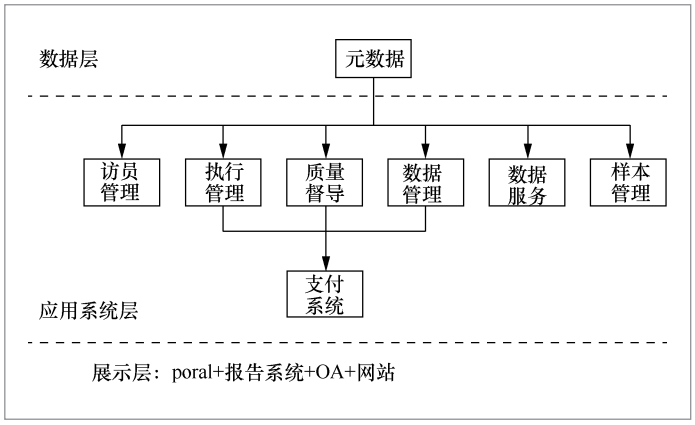

图 3 基于云计算的数据链系统

3.3 大数据与社会跟踪调查

社会调查分成两种类型：截面调查和跟踪调查。前者就是在特定时间和地点进行抽样调查，每次重复调查时，都需要进行再次抽样。后者是在调查之前确定样本后，就跟踪这些样本进行反复调查，因此，基础样本基本上是不变的。跟踪调查的优点是能积累信息丰富的面板数据（panel data），具有历时效应，能够观察特定样本随着时间的发展演变趋势，便于更好地预测未来[8]。CFPS就是典型的跟踪调查，基础样本是16 000户，每两年做一次跟踪调查。但是，跟踪调查有个劣势，就是样本跟踪难度大、成本高。尤其在中国，当前正值城市化不断深入、社会急剧转型的阶段，人口迁徙范围广、变动大。在这种环境下，CFPS样本中不少家庭在第二次进行入户跟踪调查时，就已经迁移到别的地方，有的已经找不到联系方式。若无法找到这些迁徙样本，那么CFPS样本就会出现严重流失。样本一旦出现误差，缺乏代表性，通过

入户所收集上来的数据质量就会出现严重问题。所以，所有迁徙的样本都必须确认其新地址，并且获得其联系方式，继续进行跟踪调查。

为了做到更精准地识别和确定迁徙样本的新地址，大数据挖掘技术发挥了强大的威力。2010年，CFPS做完基线调查之后，2012年开始做跟踪调查。在这轮跟踪调查做好后，迁徙样本家庭就出现了。通过互联网大数据挖掘技术，结合线下的人员打探，基本上能够再次联系上这些迁徙样本[9]。在这个过程中，采用大数据和大地图相结合的分析方法，在地图上精准画出每个迁徙样本的迁徙空间路线。根据这些迁徙空间制图的数据，再加上大数据建模和挖掘技术，就能模拟出样本家庭空间迁徙的情况，预测出2014年样本家庭迁徙的路线和区域，提前予以核实信息和联络，确保迁徙家庭主动提供迁徙后的新联系方式。同样的道理，在2012年和2014年数据的基础上，可以刻意预测2016年的迁徙情况。如此循环，大数据加上地图的分析，让很困难的样本追踪成为了相对比较容易的事情。

4 大数据与精准关系管理

4.1 大数据需要大智慧

从哲学上讲，数据无论多大，都是客体，是被认知的对象。要从数据中找出对于指导人们行为有用的信息，就需要发挥主体的主观能动性。如此，大数据才能转化为大智慧。但是，人类社会世事无常，一切都在变化着。如何用大数据来刻画转瞬即逝的社会关系，就成为大数据时代普遍的挑战。谷歌公司的流感预测这两年失灵，对于原因的剖析，可谓是仁者见仁、智者见智。哈佛大学政治学金加里（Gary King）教授等人认为，造成谷歌流感趋势预测结果偏差的重要原因是大数据傲慢（big data hubris）和算法变化（algorithm dynamics）[10]。2015年5月份，笔者专门到金加里教授的办公室和他讨论这个问题。笔者的观点是由于大数据模型无法捕捉住瞬息变化着的社会关系，导致预测失效的后果，其失效的原理如同中国古代成语“刻舟求剑”所揭示的那样，当环境发生变化了，依然沿用过去的模型去挖掘规律，往往是失效的。中国古代智慧强调的是“阴阳消息，五行转移”，强调的是用动态大数据去分析动态的社会变迁，方能在变化无常的社会关系中把握住发展的规律。调查机构在进行数据收集的过程中，会遇到方方面面的关系，需要协调和处理这些时刻都在变化着的关系。于是，在长期的调查实践摸索与总结的基础上，基于大数据的精准关系管理就产生了。

通过大数据来把握复杂多变的社会关系，从而能够精准地协调和处理好方方面面的关系，需要从内部和外部两个方面同时进行大数据收集和分析，就是内部修炼和外部整合。笔者与管理大师、《第五项修炼》的作者彼得·圣吉曾经专门讨论过大数据和组织修炼之间的关系，认为在大数据时代，组织修炼能力包括了大数据收集、分享、提炼等能力。除了内部修炼之外，大数据驱动的组织还需要能够整合外部数据和信息，具有强有力的外部大数据吸附、消化能力。在大数据时代，有了这两项基本能力，调查机构就可以对变化多端的社会关系进行精准分析和把握，使得大数据上升为大智慧，通过实施精准公关来协调好内外部关系，进而成功实施社会调查项目。基于大数据的精准公关管理如**图4**所示。

4.2 文本数据挖掘与精准关系管理

当访员入户开展调查活动时，调查现场各种情况都可能发生。访员要成功实施入户调查，就需要第一时间协调好相关的关系，获得当地社区和受访户的支持。在CFPS开始以来，通过OA系统收集所有访员在现场遭遇情况的文本信息。任何一个访员在现场遇到任何情况，都鼓励其通过OA系统记录下来。如此，随着CFPS实施的推进，逐渐积累起越来越多的关于现场突发情况和遭遇问题的文本信息。通过文本数据挖掘技术，根据广大访员的经验，对现场的情况加入分门别类的问题识别，并予以最优化应对[12]。在具体技术上，采用了文本特征提取技术[13]、文本检索技术、文本自动分类技术、文本自动聚类技术、话题检测跟踪技术、文本过滤技术、关联分析技术以及智能问答技术等。有了这套系统，访员到调查现场遇到任何问题，可以通过文本输入的方式描述问题的情况，可以第一时间获得解决问题的对策和协助，帮助访员协调好现场关系，顺利完成每一个入户调查。这是文本大数据挖掘在精准关系管理上的应用之一，效果显著。

5 未来发展的若干思考

5.1 生物社会调查（biosocial survey）的深化

人的行为不仅仅受到社会的影响，还与其生理、心理等因素有关。笔者曾经与美国科学院院士、美国人文与科学院院士、麻省理工学院资深教授哈维·罗德士

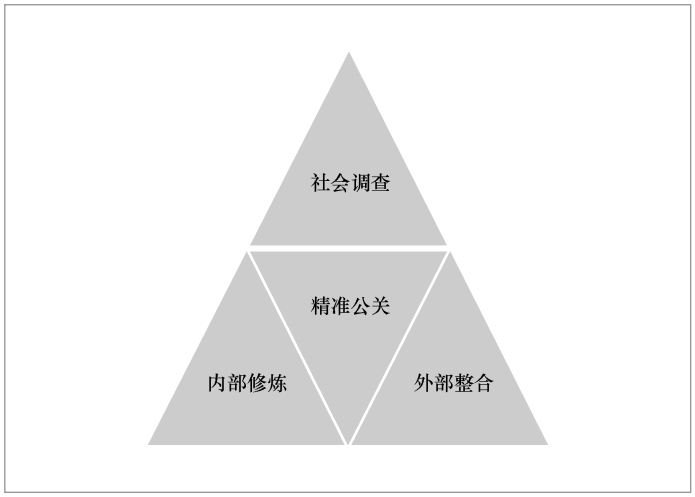

图 4 基于大数据的精准公关管理 [11]

（Harvey Lodish）深入探讨过这个问题，认为非常有必要同时调查人的生理指标[14]。随着可穿戴技术的迅速发展，已经可以突破社会调查在时间和空间上的限制，实现时时刻刻的数据收集工作。所以，未来的社会调查，问卷调查技术势必会结合可穿戴技术，整合共同收集样本的社会特征和生理指标数据，实现对样本的立体调查和监测，找出个体社会行为与其生理特征之间的关联关系。目前，已经有一些社会调查项目开始抽取血样等，为下一步开展DNA等生化因素与社会行为关系研究提供基础性数据。

5.2 大数据的精准化

数据不是越大越好，而是能推断出精准的信息便于更好地解决问题才好。当前的硬件存储技术以及数据收集能力都大为提高，收集大量的数据根本不是问题。问题在于，收集到如此海量的数据之后，能获得哪些有用的信息？这就要强调数据挖掘的精细化和精准性。如同沙里淘金，不是沙子越多越好，而是最后能掏出多少金子才是关键的核心所在。无论是大数据，还是

“小数据”，数据本身都不是问题，而是数据能提供的信息是否能帮助人们解决其所面临的问题，这才是关键。所以，问题的核心不在于量，而在于精[15]。

5.3 数据挖掘思想的拓展

在《庄子·秋水》中记录过一个故事。庄子和惠子一道在濠水的桥上游玩，庄子说：“儵鱼出游从容，是鱼之乐也。”惠子曰：“子非鱼，安知鱼之乐？”再来看当今的大数据专家，动不动就搬出大数据来说事，指点江山，预测未来。惠子式的质疑就出现了：你不是大数据，怎么知道大数据的规律呢？这是一个普遍性的问题，就是主体如何去认知客体[16]。当人们从大数据中推演出所谓的趋势、规律时，一定要谨慎，因为这些规律很可能是表象。这就要求大数据专家在数据挖掘上，要反复推敲和检验，而且不能就数据论数据，需要在分析方法上突破数据的限制和束缚，阴阳消息，五行转移，从综合和动态的角度把握大数据背后的真正规律。

6 结束语

当人们讨论大数据时，很容易把大数据与小数据对立起来。这种先入为主的判断，不利于真正认识大数据的本来面目。大数据兴起之后，的确对包括社会调查数据在内的所谓小数据产生了强大的冲击。不少市场调查公司也纷纷调整经营方向，从通过市场调查获得数据转向通过大数据技术来获得数据。但是，大数据和小数据之间并非是天然对立的，而是对立统一的，是可以互为补充的。本文以北京大学中国社会科学调查中心的中国家庭追踪调查为例，阐述了在大数据环境下的微观入户调查如何整合大数据技术而获得发展，进而收集到精准的数据。大数据的兴起引起了社会治理方式的变化。在这种背景下，中国家庭追踪调查通过无缝整合大数据技术来精准获得关于中国社会变迁测量的微观数据，进而为社会治理提供基础性决策支持数据，提供社会决策和社会治理的能力与效率。事实上，中国家庭追踪调查所获得的数据已经为政府的重大决策提供了关键性决策依据，比如政府开放单独二胎政策，就是以这套数据的模拟结果作为决策依据的。因此，在大数据盛行的今日，大数据与小数据互相融合与互相补充而形成的决策信息，将会是社会治理的重要决策依据。

参考文献:

[1] MEYER B D, MOK C, SULLIVAN J X. Household surveys in crisis[J]. Journal of Economic Perspectives, 2015, 29(4): 199-226.

[2] 顾佳峰. 调查机构管理：理论与实践[M]. 北京：人民出版社，2013.
GU J F. Survey Organization Management: Theory and Practice[M]. Beijing: People's Publishing House, 2013.

[3] XIE Y, LU P. The sampling design of the China family panel studies (CFPS) [J]. Chinese Journal of Sociology, 2015, 1(4): 471-484.

[4] STIEGER S, REIPS U. What are participants doing while filling in an online questionnaire: a paradata collection tool and an empirical study[J]. Computers in Human Behavior, 2010, 26(6): 1488-1495.

[5] KREUTER F. Improving Surveys with Paradata: Analytic Uses of Process Information[M]. New York: John Wiley & Sons, 2013.

[6] MAEHR W. eMotion: estimation of user's

emotional state by mouse motions[J]. Elsevier, 2008, 15(3): 15-17.

[7] GALESIC M, TOURANGEAU R, COUPER M P, et al. Eye-tracking data: new insights on response order effects and other cognitive shortcuts in survey responding[J]. Public Opinion Quarterly, 2008, 72(5): 892-913.

[8] TACH L, CORNWELL B. Social networks and social capital: new directions for a household panel survey[J]. Journal of Economic and Social Measurement, 2015, 40(1-4): 249-281.

[9] PAKIZE S, GANDOMI A. Comparative study of classification algorithms based on MapReduce model[J]. International Journal of Innovative Research in Advanced Engineering, 2014, 1(7): 251-254.

[10] LAZER D, KENNEDY R, KING G, et al. The parable of Google flu: traps in big data analysis[J]. Science, 2014, 343(6176): 1203-1205.

[11] 顾佳峰. 调查机构公共关系经营与管理[M]. 北京: 经济日报出版社, 2014.
GU J F. Public Relations Management for Survey Institute[M]. Beijing: The Publishing House of the Economic Daily, 2014.

[12] WILLIAMS T, GONG J. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers[J]. Journal of Automation in Construction, 2014, 43(7): 23-29.

[13] HEARST M. TextTiling: segmenting text into multi-paragraph subtopic passages[J]. Computational Linguistics, 1997, 23(1): 33-64.

[14] CHAWLA N V, DAVIS D A. Bringing big data to personalized healthcare: a patient-centered framework[J]. Journal of General Internal Medicine, 2013, 28(3): 660-665.

[15] COLLINS F S, VARMUS H. A new initiative on precision medicine[J]. The New England Journal of Medicine, 2015, 372(9): 1-3.

[16] DANAH B, KATE C. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon[J]. Information, Communication & Society, 2012, 15(5): 662-679.

作者简介

顾佳峰(1975-),男,博士,北京大学中国社会科学调查中心研究发展部主任,北京大学创新研究院副院长,美国加州伯克利大学、哈佛大学高级访问学者,OECD组织“产业监管调查项目”中国地区项目首席科学家。

收稿日期: 2016-01-20
* 本文为2015中国大数据技术大会(BDTC)演讲约稿