大数据政治学：新信息时代的政治现象及其探析路径*

孟天广　郭凤林

[内容提要]　伴随着信息化和网络化的深入，数据科学（大数据）正在引发社会科学方法论的变革。大数据政治学至少在方法创新和学科发展两大领域的五个方面具有独特价值：更便捷、更廉价、更大规模的数据采集；数据分析学新方法的引入；定量与定性方法的整合；政治学与计算科学、信息科学等跨学科研究；数据民主化所推进的政治知识平民化的传播和普及。随着大数据政治学方兴未艾，国际学术界已积极将搜索引擎技术、自动文本分析、网络分析、空间分析等应用于政治传播、社会运动与战争、政治文本、投票与选举等政治现象的研究中，并积累了一系列值得借鉴的研究成果。中国学界正积极拥抱大数据政治学，并在理解政府行为、政治传播、互联网政治等方面作出了有益探索。

[关键词]　大数据　政治学　自动文本分析　社会网络分析　研究方法

一、大数据时代的政治现象及其方法论革命

面对全球数据量的指数级增长，《科学》杂志于 2008 年提出用“大数据”来讨论新信息时代（PB 时代）的科学研究。2012 年，《纽约时报》刊文宣告“大数据时代已经到来”。著名信息技术研究机构高德纳（Gartner）认为，大数据是指需要新处理模式才能确保更强的决策力、洞察力和流程优化力的海量高速增长和多样化的信息财富。[①]也有研究认为界定大数据不能简单地以数据规模为准，而要考虑数据管理和分析的复杂程度。除了数据规模，大数据区别于传统数据库的特

* 本文受到清华大学—中国科学院学部科学与社会协同发展研究中心 2014 年先导性项目“基于大数据的网络舆情治理：政府干预在互联网政治传播中的作用”（项目编号：2014A02）的资助。作者感谢香港中文大学博士生苏政和北京大学博士生邵梓捷的研究助理工作。

点还包括：数据形式混合着结构化、非结构化数据；数据存储于不同的数据生产者；对数据挖掘、机器学习和统计分析等数据分析技术的要求较高，等等。

伴随着信息技术和互联网的飞速发展，尤其是 Web 2.0 时代网络数据和社交数据的空前膨胀，传统的数据存储、管理和分析能力已经难以顺应新信息时代的客观要求。大数据应运而生，成为信息科学和计算科学的发展前沿。综合起来，大数据具有如下五大特征：（1）超大规模数据。大数据力图分析全数据，通常指 TB 级别以上的数据量。（2）数据类型多样化。大数据蕴含了文本、图片、视频、音频、邮件、交易信息、社交网络信息等结构化、非结构化数据。（3）数据流动速度快。大数据善于管理和分析动态变化的数据流。（4）大数据蕴含了丰富的时空信息。（5）大数据是贫矿，价值密度低。

大数据时代的来临首先塑造着新信息时代的政治现象。大数据不仅将政治活动场域扩展到虚拟空间，还改造着政府、公民、政党等政治行为主体的行为模式及其关系。保罗·德克尔（Paul T. Decker）将大数据视为“颠覆性创新”，认为它带来了“数据的民主化”，为研究者提供了新机会，有助于推动更高效、更具创新性且更透明的政府建设。[②]为了顺应大数据时代国家治理的客观要求，中国成立了国家安全和信息化领导小组，积极致力于将大数据方法应用于完善国家治理体系和治理能力的建设中，以确保网络繁荣、信息安全和有效治理等战略目标的实现。

国际社会积极将大数据应用于国际发展、政治稳定和公共治理领域。联合国于 2009 年发起全球脉动计划，通过对网络空间海量数据的数据挖掘和统计分析预测各国的失业率、疾病暴发、政治动乱等现象，以此作为国际组织行为的依据。[③]美国政府于 2012 年启动“大数据研究和发展倡议”，通过整合联邦政府各部门的海量数据和大数据分析技术来维护、分析和共享相关成果以服务美国政府的政治利益。[④]

新信息时代的大数据还成为影响现实政治的关键因素。以脸书、推特等为代表的新媒体正成为影响现实政治的重要力量，2010 年底的“阿拉伯之春”政治变革正是源于网络社交平台。脸书和推特等新媒体在“阿拉伯之春”的酝酿、组织、爆发、升级等各个环节均发挥了关键性作用，在新社会运动中扮演着信息传播、动员组织、全球呼应等重要角色。以我国为例，大数据时代给舆情治理带来了严重挑战。互联网时刻进行着信息更新，尤其是自媒体信息，信息量得到了质的增长。互联网的交互性极强，突破了地域、空间、身份的限制，社会各阶层的观点、情绪和诉求在网络空间中迅速集聚、碰撞、流传，信息呈网状传播，速度快且传播范围广，容易引发重大舆情危机，使得网络舆情治理更为困难。

大数据一出现即挑战着传统的科学研究方法论。图灵奖得主詹姆斯·格雷（James Gray）认为大数据时代将形成数据密集型科学研究的“第四范式”（the fourth paradigm）。[⑤]大数据时代的科学研究将不再需要模型和假设，而是利用超级计算直接分析海量数据，发现相关关系，从而获得新知识。《自然》、《科学》等杂志分别组织专刊讨论了大数据对自然科学和社会科学研究模式的挑战和创新价值。

大数据正在引发政治学、经济学等社

会科学的一场方法论革命。2009 年，戴维·雷泽尔（David Lazer）等在《科学》发文提出“计算社会科学”（computational social science）的构想。他们认为，计算社会科学正在兴起，人们将在前所未有的深度和广度上采集和利用数据为社会科学研究服务。[⑥]瑞·M. 张（Ray M. Chang）探讨了大数据带来的社会科学范式的转换，认为大数据带来了更便捷的数据收集技术，社会科学与计算科学、网络科学相结合，正在向“计算社会科学”和“网络社会科学”（E - Social Science）的方向转变。[⑦]

菲利普·郎克尔（Philip J. Runkel）等人提出了社会科学研究的困境，即“普遍性”（generality）、“可控性”（control）和“现实性”（reality）三大目标难以同时实现。而大数据的数据可获得性、低廉的成本和设计上的便利，使得一些过去不能做的研究成为可能，研究者过去所注重的控制变量选择变得更为多元，实验设计可以设定更多条件，能够在很大程度上解决上述困境。在大数据推动社会科学范式转换的过程中，技术进步、学科间融合、新数据分析技术的应用、新的商业和组织环境都会加速这种范式转换。这种转换涉及诸多方面：在研究视角上要实现不同学科间研究方法、理论及测量上的整合；在研究方法上，研究者不再需要构建精巧的研究设计来模拟现实，而是可以直接获取人类行为和互动的基本信息，田野研究和实验研究间的界线会逐渐模糊；在样本选择上，大数据可以突破传统抽样调查的样本限制，观察性研究也能够大幅度提高数据的采集频率。需要注意的是，尽管大数据会对研究方法产生重大影响，但理论的作用并不会因为大数据时代的到来而减弱，仍然在科学研究中占据核心位置。[⑧]

二、大数据政治学的研究主题

国内外学术界将大数据方法应用于政治学已经初见端倪，并在涉及公共政策、政治传播、选举与投票行为和社会运动的广泛主题上取得了一系列丰硕的研究成果。本部分将系统梳理政治学领域应用大数据方法开展的研究主题及其成果。

公共政策

大数据在公共政策领域的应用充满希望，托马斯·库克（Thomas D. Cook）热情展望了大数据在公共政策领域的应用前景。大数据在提高政策描述和强化政策预测能力方面具有强大潜力：借助大数据技术，个体、城市、国家层面以及群体数据，尤其是大规模时间序列数据的实时获取成为可能，会使研究者对公共政策的描述和评估在时间和空间上变得更为丰富。此外，在数量更多、质量更好的数据基础上，公共政策分析的基础工具——成本收益分析将更为适用。利用警察局犯罪数据对稀缺警力进行更有效的配置就是一个可以直接运用大数据的公共政策问题。[⑨]贾斯汀·基恩（Justin Keen）着重探讨了卫生服务信息公开在英国卫生服务领域的前景。卫生服务领域已经具备了大量、完整的信息，这些信息向第三方开放将会带来巨大收益。[⑩]

在许多政策领域，单一数据来源已经不足以应付复杂的公共政策问题，有效的公共治理需要平行使用多个大型数据库。以美国联邦政府为例，“9·11 事件”之后，美国各大部门开始建立数据库，并逐步实现数据库之间的共享和实时连接。比如，美国海关总署要求航空公司提供乘客

所有信息（包括地址、电话、犯罪记录、身份证号以及驾照号码等），交通部则建立了将航空预定系统与私人、政府数据库相连接的智能网络来对乘客进行定位，而地方警局可以与这些数据库进行实时信息交流。在医疗领域，卫生服务效用数据库新近被开发出来，用于推进公共支付过程的改革，其存储的实时支付数据在评估服务绩效和描述质量波动时非常有用。在法律实施和公共安全方面，纽约市将其警务责任系统提升到预警层次，对紧急援助、自行车道管理、林木规划等进行数字化管理，通过在市长办公室视频上滚动显示不同指标的实时结果，使官员和民众能够实时掌握各个区域的情况。[11]

政治传播

很多学者利用大数据技术对互联网空间的政治传播进行研究。作为一种虚拟公共空间，互联网空间存在门户网站、网络论坛、社交平台等公共空间，充斥着文本、视频、关系等结构化和非结构化信息，为大数据政治学的发展提供了前所未有的试验场。现有研究利用大数据方法探讨了网络政治传播的方式、影响及其与传统政治传播方式的联系。米歇尔·詹森（Michael J. Jensen）利用推特数据预测了2011年美国共和党总统提名。他搜集了盖洛普民意调查数据和2011年实际投票结果，并利用推特应用程序接口搜集了爱荷华州党内提名会议前的有关竞选人姓名的推特留言，得到了195737位推特用户的697065条推特留言，随后将每位竞选人的推特提到率、民意调查支持率与实际提名结果进行比较，发现尽管推特提到率与最终投票结果不完全一致，但推特传播中存在着一些里程碑式的转折点，对于竞选者有较大影响。[12]

卡琳娜·拿翁（Karine Nahon）考察了政治竞选活动中视频博客传播的模式。作者从网络视频中选择了“政治”、“选举”、“大众”三大主题中排名前100的视频，再由一名教授和三名博士生对300个视频的内容进行分类，最终获得了120个与选举相关的视频样本，随后利用谷歌博客搜索技术寻找与120个视频链接的博客，在清除了重复信息后共获得9765个博主发布这些视频的13173篇博客。作者根据每个博客日浏览量的不同，将博主区分为精英、政治领袖、一般领袖和普通博主四种类型。最后，作者通过多元时间序列模型发现，网络信息传播并不是单向度的（monolithic），精英和政治领袖是信息传播的发起者，决定着信息传播的内容和时间；一般领袖和普通博主是信息接收者和跟随者，维持着信息传播的链条。[13]

社会运动

大数据使得获取个体层面数据变得更为可行，不少学者利用大数据方法来预测社会运动。海伦·玛格斯（Helen Margetts）指出，由于大数据能够提供个体行为和意愿数据，使得研究者能够探讨一些过去无法研究的问题，比如由网络社交平台引发的“阿拉伯之春”。[14]托马斯·沙德福（Thomas Chadefaux）利用文本分析方法来预测大规模冲突事件，他分析了1990—2013年间166个国家的报纸文章，来验证同期超过200次的战争冲突。运用这些信息，他能够在85%的置信水平下推断下一年是否会发生战争，成功预测了超过70%的大规模战争。[15]

现有研究非常关心网络社交平台对社会运动的影响。奥丽莎·科特佐娃（OlessiaKoltsova）研究了俄罗斯网络社交平台——交友网（Live Journal）中大户

(top bloggers) 在舆论引领中的角色。他利用计算机模拟了微博大户的话题结构，发现他们的关注度平均分布在“社会—政治”以及“私人—娱乐”话题上，而俄罗斯 2011 年的街头抗议能够从博客有关政治的内容上得到明显体现。这种消息最初在某个社会话题小组内部显现，很大程度上充当了公共舆情的指示计。[16]与此类似，托马斯·兰德 (Thomas Lansdall) 收集了 4 亿多条社交媒体信息，检验了经济衰退对英国民众情绪的影响，发现了支出削减声明与 2011 年 8 月骚乱之间的强相关关系。[17]

选举与投票

大数据在选举研究中得到了广泛应用。基于谷歌搜索记录数据，莎娜·莱莉 (Shauna Reilly) 探究了 2008 年总统选举前一周谷歌上投票法案名称 (或主题) 的搜索次数与实际投票率的关系。他选用投票流失率 (roll - off) 作为因变量，以谷歌上 153 个投票法案的名称搜索率和主题搜索率作为自变量。相关分析发现名称搜索率和主题搜索率对投票流失率有负向影响，即谷歌上对投票法案的搜索率越高，选民放弃投票的可能性就越低。[18]布鲁斯·宾伯 (Bruce Bimber) 介绍了大数据在奥巴马竞选美国总统时发挥的重要作用。奥巴马团队在竞选宣传时注重获得选民个体化的信息，并基于模型化分析进行更精准和更有技巧性的宣传。他们对人群的分类不再是简单的人口统计学变量如中产阶级妇女或工薪阶层，而是包含了选民的投票记录、慈善捐赠乃至音乐偏好、汽车品牌、杂志订阅、社交网络等信息，对选民进行更为个体化的分析和宣传。[19]

史提芬·安索雷布哈尔 (Stephen Ansolabehere) 利用大数据探讨了民调中自报投票率超过实际投票率的原因。他通过凯利板 (Catalist) 公司搜集了 50 个州的投票登记数据，包括投票人姓名、住址等信息，以姓名和住址为参照，与 2008 年国会选举调查数据库的受访者信息相匹配，从而得到涵盖选民姓名、地址、性别、年龄、自报投票记录以及实际投票情况等信息的数据库，去除无回答的人数后，形成一个由 26181 名选民组成的样本。同时，作者也采用“美国国家选举研究”1988 年、1984 年和 1980 年的调查数据。在这两大数据库的基础上，作者比较了报告投票率 (report vote rate) 和有效投票率 (validated vote rate) 的时间变化及其差异，并利用性别、年龄、教育、宗教信仰、种族、婚姻状况、流动情况以及党派等因素分别对报告投票率和有效投票率进行回归分析。结果表明，缺失的自报投票者集中在高教育程度、高收入、较活跃的党派成员、经常参加教堂活动和流动性较强的人口中，即流失的自报投票者偏向相对优势群体，这种系统性偏差使得利用民调数据预测投票情况会出现推论偏差。[20]

议会政治

有研究利用大数据方法考察西方国家的议会政治，对议员的政治话语及行为展开了实证研究。贾斯汀·格里默 (Justin Grimmer) 提出了议程表达模型来解释美国参议员的行为逻辑，他利用自动文本分析研究了美国参议员与选民的政治沟通。他利用互联网搜集了美国参议院自 2007 年以来发布的 24000 余份新闻通告，利用无人监督机器学习法，由计算机自动识别单词并进行归类，然后应用贝叶斯分层分析模型来预测特定参议员的议题关注。基于对文本资料的分析，他发现每个参议员

的议题关注与其他参议员的议题关注之间存在着显著相关，重点关注议题的地域分布具有一定的集聚性，议员对参议院拨款法案的关注程度与他们对德敏特—麦凯恩（Demint－McCain）修正案的反对票呈现正相关关系。[21]

乔纳森·布莱特（Jonathan Bright）利用议会文本记录分析了英国议会中议会争论的演变特点。他利用英国议会解析网站提供的议会资料，构建了1936—2011年间英国下议院发布的由7.4亿个单词所构成的数据库，利用自动编码技术对法律、国防、环境、卫生、就业、权利、教育、农业、经济等关键词进行了编码，对这些词汇在这75年间的出现频率进行了描绘，发现这些关键词的出现频率具有一定的稳定性，但也存在很大变化，争论变得更加激烈，环境议题变得更为突出，而农业等问题则逐渐衰落。同时，作者还对不同的文本进行了自动的性别和身份识别，分析了女性以及贵族身份议员在议会争论中的地位和特点，发现前者倾向于较长的发言时间，而后者被打断的频率更高一些。[22]

三、大数据政治学的研究方法

自动文本分析

政治文本分析是探析政治现象的重要途径，是获取政治态度、政治立场以及观测其随时间变化的重要方法。大数据技术出现以前，人工编码数量浩瀚的政治文本非常困难，而自动文本分析技术的出现可以将这项繁琐的工作交由计算机处理，使得这种大规模的文本分析成为可能。

格里默专门探讨了自动文本分析方法的前景和“陷阱”。他认为，文本分析的核心工作是分类。分类有三种方法：字典法（dictionary methods），根据关键词的出现次数来确定文本；有监督学习法（supervised learning methods），先由人工构建编码练习库，然后让机器根据人工编码模式进行自动编码，最后将机器编码与人工编码相比较检验其效度；无监督学习法（unsupervised learning methods），不需要人工事先编码，而是基于模型假设和文本性质来分类并自动将文本分配到各类别。第三种方法比较便捷，但容易混淆重点，可以通过两项技术进行改进：一是通过混合成员模型（mixed membership models），将具体问题结构纳入分析以辅助分类；二是通过计算机辅助分类（computer assisted clustering）来探索众多潜在分类方法。自动文本分析可以确保研究者便捷地实现文本分类和定位，但仍需进一步完善。格里默还总结了自动文本分析的四大基本规律：机器自动识别有很多不准确的地方，但仍然在很多方面给学者提供了研究便利；自动文本分析不能取代学者的阅读和思考；没有一个最完美的自动识别方法；对自动文本分析结果的效度分析非常重要。[23]

斯拉瓦·米哈伊洛夫（Slava Mikhaylov）分析了自动文本分析中编码和分类的效度问题。在文本编码过程中，无论是人工还是机器编码都容易产生效度问题：不同的人对同一文本可能有不同的理解，而不管是有监督、半监督还是无监督自动编码都依赖于参考样本，从而导致编码和分类中误差的存在。作者通过一个编码实验来评估人工输入过程的信度。他利用欧洲比较政党项目（Comparative Manifestos Project，CMP）数据，利用卡帕（Kappa）分析法进行统计检验，发现无论是在单项类别

还是在整体位置的测度上，实验编码结果和 CMP 原始编码结果的一致性都比较低。简言之，编码误差几乎超过了文本形成过程和编码不一致所带来的误差。因此，在利用自动文本分析对文本进行分类时，必须注意到分类过程的信度和效度问题。[24]

社会网络分析

社会网络分析是社会学中常见的对关系型数据的分析方法。在大数据时代，随着数据抓取能力的增强和处理复杂网络之分析软件的出现，社会网络分析在研究领域、研究方法上得到了长足的发展。不少学者尝试利用该方法对政治选举、集体行动、政治传播等问题进行研究。结合大数据强大的结构性和非结构性数据的获取能力，社会网络构建将变得更为丰富细致，许多过去难以研究的问题会在数据可获得性的基础上得到新的生命力，政治传播、集体行动等研究将会取得新的进展。

罗伯特·邦德（Robert M. Bond）等人在《自然》杂志上发文比较了网络社会网络和面对面社会网络影响政治行为的路径。他们在 2010 年美国国会大选期间对 6100 万脸谱网用户实施了一项发送政治动员消息的随机控制实验，研究发现政治动员消息直接影响着网民的政治自我表达、信息搜寻和现实投票行为。值得注意的是，政治动员消息不仅影响了接受者，还影响了接受者的网友、网友的网友，而这种社会传递效应对投票行为的影响要强于直接效应，而传递效应主要发生在更可能直接接触过的“亲密网友”间，从而凸显出政治行为中强联系的价值。[25]

桑德拉·冈萨雷斯－贝隆（Sandra González－Bailón）等人讨论了线上网络对征兵抗议演变的影响。他们以西班牙动员浪潮中推特网络中的征兵抗议模式为例，试图探讨新媒体如何影响抗议活动的扩散，在识别征兵领导人的网络位置和信息散布者的网络位置后，研究发现消息散布者比征兵领导人更位于网络中心，对征兵抗议过程发挥着更重要的影响。[26]康伟将数据抓取技术与社会网络分析方法相结合，探究了“7·23 动车事故”中网络舆情传播的网络结构、节点位置和关键时点等问题。他对与此次事故相关的个人微博和机构微博信息进行了抓取，获得了主要节点账户间的关注信息，构建了一个社会关系网络，并对其密度、规模、结构等进行了测量，探讨了网络传播在节点以及传播上的一些特点。[27]

可视化和空间分析

可视化是大数据时代社会科学研究的新趋势，是大数据应用最显著的效果之一，更为优化的数据处理技术使得过去的描述性信息可以变得更加直观，增强了对数据信息的发现、跟踪、分析和理解，还能够显著提高表达主题的吸引力和说服力。此外，大数据可视化分析与传统统计分析的区别在于它的动态性，其数据容量、内容及更先进的处理方法都使得动态可视化分析成为可能。

目前不少软件可用于可视化分析，海杜普（Hadoop）即是一个比较成熟的可视化软件，能够对大量数据进行即时处理，淘宝、百度等大型商业网站就利用海杜普来完成每天数以亿计的访问量数据存储、查询统计以及用户行为分析等。美国环境系统研究所（Environmental Systems Research Institute，ESRI）在开源网站基哈伯（GitHub）上共享了“海杜普地理信息系统工具”（GIS Tools for Hadoop），用户可以利用其对上亿条空间数据记录进行过滤和聚合操作，在报告中嵌入大数据

地图进行发布。然而，可视化分析在政治学研究中的应用非常缺乏，因而相关技术和方法普及是至关重要的。

空间分析与可视化密切相关，但具有超出可视化的诸多功能。大数据卓越的数据获取能力及网络化获取方法使得数据获取在很大程度上突破了地理范围的限制，能够同时获取区域乃至全球层面的数据。例如，百度迁徙可实时记录并分析中国人口流动的方向、数量等信息，构建清晰美观的全国人口流动图。俄罗斯工程师鲁斯兰·艾尼基维（Ruslan Enikeev）利用2011年全球196个国家200多万个网站链接将不同国家的网站流量信息构建了一个网络星球（The Internet Map），每个星球的大小根据其网站流量来决定，而星球间的距离则根据链接出现的频率、强度和用户跳转时创建的链接来确定。空间数据的丰富与共享为政治学提供了将空间概念引入政治学分析框架的新机遇，然而，受到数据获取能力和分析能力的限制，政治学研究中空间分析的应用非常缺乏。

四、大数据政治学视角下的中国政治

大数据方法的出现和运用在一定程度上可以穿透政治现象的复杂性和特殊性，为中国政治的研究者带来深刻而丰富的洞见，并为其理论提供更强大的说服力。国内外学者已经利用大数据方法在政治传播、互联网政治、网络舆情治理和分析方法创新等方面进行了有益的尝试。

政治传播会通过影响人们对于特定事件的认识和态度、塑造人们的价值观而进一步影响人们的政治参与行为，而报刊、广播电视以及电子邮件、手机、网站、博客等新媒体都是政治传播的重要载体。在一项针对报刊审查机制的研究中，作者通过追踪《广州日报》和《南方周末》从2002年12月至2003年6月间的全部报道来分析政府干预对于“非典事件”曝光的影响。他们发现，通过宣传部门委任报刊主编、在各个层级发布指令和通告、传播领导人在特定场合的直接指示是影响“非典”曝光率的三种主要机制。[29]另一项研究通过收集几十万条新浪博客和校内网的帖子，比较了两种网上社交网络的传播特点。[29]此外，有研究者对2008年“汶川地震”发生后天涯论坛一周内的2266个主题帖进行了分类，并分析了论坛在信息、观点、行动、情感和社区建设等方面的作用。[30]

网络政治关注的一个核心问题是网络参与对实际政治行为的影响，它既可能成为消除潜在社会不安的“解压阀”，也有可能成为酝酿激进行为的“高压锅”。而造成二者区别的关键在于网络讨论的时点、议题选择和参与者本身的意图。乔纳森·哈西德（Jonathan Hassid）对2198个博客从2010年8月30日至11月7日的发帖内容进行文本分析后指出，在涉及腐败、环保、领土争端等由主流媒体发起的议题时，政府对于参与者的评论、批评和正式行动会表现得较为宽容，网络参与起到了一种“安全阀”效应；而当议题超前或涉及敏感领域，如城乡差异、宗教问题时，过多讨论则会加剧社会紧张和不安，发帖者也更可能遭到严格的审查。[31]另一项研究在分析了2003年、2005年和2007年收集的调查数据后发现，互联网使用与线上观点表达存在正向关系，而即使存在政府审查，互联网的网络效应也可能会给中国社会带来增量的变化。[32]

在大数据时代，网络舆情成为影响国

家治理的重要因素，因而网络空间的政府干预变得不可或缺。加里·金（Gary King）首创性地使用自动文本分析技术，对 2011 年上半年 1400 多个网站的上百万个帖子进行了内容分析，并将其归入不同的议题领域。研究发现，相比于其他议题，审查机构对批评政府、领导人和政策的帖子的删帖率较低；而无论内容为何，有可能导致集体行动或强化社会动员的帖子成为政府审查的主要对象，即防止潜在的集体行动是政府审查的主要动机。[33]

五、大数据时代的政治学研究：机遇与挑战

综上所述，大数据方法不仅为深入探析选举政治、社会运动等传统政治现象提供了创新性工具箱，更为挖掘信息时代的信息政治、互联网政治等新生政治现象创造了方法和理论视角。大数据方法对政治学研究的核心贡献体现在研究方法创新和学科发展两个领域。

大数据方法空前催化了政治学研究方法的开拓创新，这反映在以下三个具体方面：（1）大数据方法革新了政治学研究中数据获取与管理的既有模式。大数据方法使得廉价便捷地获取总体数据而不是抽样数据成为可能，更进一步拓宽了传统政治学对数据的界定，历史文本、社交媒体、多媒体等结构化、非结构化、关系型的数据都成为研究对象。（2）机器学习、数据挖掘等数据分析学（data analytics）的发展空前催化了政治学研究方法的创新，诸如自动文本分析、主题模型、情感分析等前沿方法被及时应用于政治学研究。（3）大数据方法强化了定量方法与定性方法的对话。传统政治学研究中长期存在的定量和定性方法分野有望在大数据时代合流，大数据方法可以有效利用定量技术分析大规模的定性资料，同时运用定性方法来呈现和阐释定量分析结果。

此外，将大数据方法应用于政治学研究还极大地拓宽了政治学的学科界限和社会价值。首先，大数据方法大规模地拓宽了政治学的学科界限，将互联网、社交网络、信息流和语义等纳入政治学研究范畴，促使政治学与计算科学、信息科学、传播学、语言学等相关学科的跨学科研究落到实处。其次，借助大数据方法与互联网和可视化的无缝对接，大数据政治学的研究成果得以实时、直观、平民化地传播和普及，这不仅保证了政治知识的大众启蒙和社会积累，更强化了政治学研究对现实政治的直接影响。

总之，大数据方法在政治学等社会科学中具有广阔的应用前景和开发潜力。清华大学、北京大学等科研机构已经启动了利用大数据方法开展政府质量、政治传播和互联网政治研究的项目，并取得了初步成果。然而，将现有大数据方法应用于政治学等社会科学研究也面临若干重要挑战，明确这些挑战有助于我们深刻理解大数据政治学的本质及其发展趋势。

首先，大数据方法的数据测量面临严重的信度和效度问题。雷泽尔在《科学》杂志上撰文指出，以谷歌流感趋势为代表的大数据预测技术尽管有其价值，但仍然存在不可忽视的预测误差，作者将其称为“大数据分析的陷阱”。大数据分析的陷阱主要源于所谓的“大数据傲慢”，即研究者假定大数据是传统数据采集和分析方法的替代而不是补充。然而，大数据并不意味着人们可以忽视信度、效度和数据相依等基本测量问题，大数据的核心挑战在

于广受关注的数据信息缺乏科学研究的效度和信度。[34]

第二，大数据强调相关性而非因果性的研究取向限制了其探究因果关系的能力。在著名的《大数据时代》一书中，维克托·迈尔－舍恩伯格（Viktor Mayer－Schönberger）认为大数据时代相关关系优于因果关系，相关性可以让我们在分析某些现象的时候不用了解其内部运作机制即可预测未来。然而，因果推论是科学研究的最终目标，即利用我们已知的知识来了解我们未知的世界，而抽离因果关系是这一过程的核心环节。[35]大数据缺乏发现因果关系的优势，应该将其与实验设计和观察研究相结合来获取有价值的知识。

第三，缺乏透明性和开放性极大地限制着大数据方法的应用。商业机构和公共机构掌握的大数据不仅涉及个人和商业机构的隐私，还涉及利益分配等问题，数据开放的前景尚不明朗。此外，出于经济和政治利益的考虑，大数据提供者或使用者经常性地调整数据算法（algorithm dynamics），导致研究者不仅无法获得稳定且可比的测量数据，更缺乏对数据生成过程的基本知识。因而，很多学者倡导大数据提供者应该确保基本的数据透明性。

第四，技术壁垒也限制着大数据在社会科学中的广泛应用。应用大数据方法不仅需要强大的数据采集和存储技术，而且需要开发数据分析学、预测分析学（predictive analytics）等数据分析和计算技术。毫无疑问，熟练掌握和应用以上技术对于社会科学研究者而言是不小的挑战，因而，强化社会科学与计算科学、信息科学的跨学科合作，培育社会科学领域的大数据分析人才将不可或缺。■

注 释

① 参见维基百科“Gartner”词条，http://en.wikipedia.org/wiki/Gartner。

② Paul T. Decker, “Presidential Address: False Choices, Policy Framing, and the Promise of ‘Big Data’”, *Journal of Policy Analysis and Management*, Vol. 33, No. 2, 2014, pp. 252－262.

③ 参见联合国全球脉动（global pulse）计划，http://www.unglobalpulse.org/research。

④ Tom Kalil, “Big Data is a Big Deal”, 2012, http://www.whitehouse.gov/blog/2012/03/29/big－data－big－deal.

⑤ J Gray, Tony Hey, “The Fourth Paradigm－Data－Intensive Scientific Discovery”, *E－Science and Information Management*, 2012, p. 1.

⑥ David Lazer, et al, “Life in the network: the Coming Age of Computational Social Science”, *Science* (New York, NY), Vol. 323, No. 5915, 2009, p. 721.

⑦ Ray M. Chang, Robert J. Kauffman, YoungOk Kwon, “Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data”, *Decision Support Systems*, 63, 2013, pp. 67－88.

⑧ Philip Runkel, Joseph Edward MacGrath, *Research on Human Behavior*, Holt, Rinehart & Winston, 1972.

⑨ Thomas D. Cook, “‘Big Data’ in Research on Social Policy”, *Journal of Policy Analysis and Management*, Vol. 33, No. 2, 2014, pp. 544－547.

⑩ Keen, Justin, et al, “Big Data + Politics = Open Data: The Case of Health Care Data in England”, *Policy & Internet*, Vol. 5, No. 2, 2013, pp. 228－243.

⑪ Michael E. Milakovich, “Anticipatory Government: Integrating Big Data for Smaller Government”, *Internet, Politics, Policy*, 2012.

⑫ Michael J. Jensen, Nick Anstead, “Psychological Investigations: Tweets, Votes, and Unknown Unknowns in the Republican Nomination Process”, *Policy & Internet*, Vol. 5, No. 2, 2013, pp. 161－182.

⑬ Nahon Karine, et al, “Fifteen Minutes of Fame: The Power of Blogs in the Lifecycleof Viral Political Information”, *Policy & Internet*, Vol. 3, No. 1, 2011, pp. 1－28.

⑭ Margetts Helen, David Sutcliffe, “Addressing the Policy Challenges and Opportunities of ‘Big data’”, *Policy & Internet*, Vol. 5, No. 2, 2013, pp. 139－146.

⑮ Thomas Chadefaux, "Early Warning Signals for War in the News", *Journal of Peace Research*, Vol. 51, No. 1, 2014, pp. 5 – 18.

⑯ Olessia Koltsova, Sergei Koltcov, "Mapping the Public Agenda with Topic Modeling: The Case of the Russian Livejournal", *Policy & Internet*, Vol. 5, No. 2, 2013, pp. 207 – 227.

⑰ Lansdall – Welfare Thomas, Vasileios Lampos, et al, "Effects of the Recession on Public Mood in the UK", *Proceedings of the 21st international conference Companion on World Wide Web*, ACM, 2012, pp. 1221 – 1226.

⑱ Shauna Reilly, Sean Richey, J. Benjamin Taylor, "Using Google Search Data for State Politics Research an Empirical Validity Test Using Roll – Off Data", *State Politics & Policy Quarterly*, Vol. 12, No. 2, 2012, pp. 146 – 159.

⑲ Bruce Bimber, "Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the Personalized Political Communication Environment", *Journal of Information Technology & Politics*, 2014.

⑳ Stephen Ansolabehere, Eitan Hersh, "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate", *Political Analysis*, Vol. 20, No. 4, 2012, pp. 437 – 459.

㉑ Justin Grimmer, "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases", *Political Analysis*, Vol. 18, No. 1, 2010, pp. 1 – 35.

㉒ Bright Jonathan, "The Dynamics of Parliamentary Discourse in the UK: 1936 – 2011", *Draft Paper for Presentation at "IPP 2012: Big Data, Big Challenges"*, 2012.

㉓ Justin Grimmer, and Brandon M. Stewart, "Text as data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts", *Political Analysis*, Vol. 21, No. 3, 2013, pp. 267 – 297.

㉔ Slava Mikhaylov, Michael Laver, Kenneth R. Benoit, "Coder Reliability and Misclassification in the Human Coding of Party Manifestos", *Political Analysis*, Vol. 20, No. 1, 2012, pp. 78 – 91.

㉕ Robert M. Bond, et al, "A 61 – million – person Experiment in Social Influence and Political Mobilization", *Nature*, Vol. 489, No. 7415, 2012, pp. 295 – 298.

㉖ Sandra González – Bailón, et al, "The Dynamics of Protest Recruitment through an Online Network", *Scientific Reports*, Vol. 1, 2011, pp. 1 – 7.

㉗ 康伟《基于 SNA 的突发事件网络舆情关键节点识别——以"7·23 动车事故"为例》,载《公共管理学报》2012 年第 3 期。

㉘ Ernest Zhang, Kenneth Fleming, "Examination of Characteristics of News Media under Censorship: A Content Analysis of Selected Chinese Newspapers' *SARS* Coverage", *Asian Journal of Communication*, Vol. 15, No. 3, 2005, pp. 319 – 339.

㉙ Feng Fu, Lianghuan Liu, Long Wang, "Empirical Analysis of Online Social Networks in the Age of Web 2.0", *Physica A: Statistical Mechanics and its Applications*, Vol. 387, No. 2, 2008, pp. 675 – 684.

㉚ Yan Qu, Philip Fei Wu, Xiaoqing Wang, "Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake", *Proceedings of the 42nd Hawaii International Conference on System Sciences*, 2009.

㉛ Jonathan Hassid, "Safety Valve or Pressure Cooker? Blogs in Chinese Political Life", *Journal of Communication*, Vol. 62, No. 2, 2012, pp. 212 – 230.

㉜ Fei Shen, et al, "Online Network Size, Efficacy, and Opinion Expression: Assessing the Impacts of Internet Use in China", *International Journal of Public Opinion Research*, Vol. 21, No. 4, 2009, pp. 451 – 476.

㉝ Gary King, Jennifer Pan, Margaret E. Roberts, "How Censorship in China Allows Government Criticism But Silences Collective Expression", *American Political Science Review*, Vol. 107, No. 2, 2013, pp. 326 – 343.

㉞ David M. Lazer, et al, "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, 343 (6176), 2014, pp. 1203 – 1205.

㉟ [美] 维克托·迈尔 – 舍恩伯格、肯尼思·库克耶:《大数据时代:生活、工作与思维的大变革》,浙江人民出版社 2013 年版。

[孟天广:清华大学政治学系;郭凤林:北京大学政府管理学院]

(责任编辑 闫 健)