

■ 文化传媒研究

当大数据遭遇人性：兼论技术的断层与无奈

李炜炜

(北京邮电大学人文学院, 北京 100876)

摘要:在数据科学滥觞之初,“模型推动运营”还是“数据推动运营”就是学术界争论的焦点。时至今日,在“算法统治世界”的话语喧嚣中,我们依然需要清醒。当大数据遭遇人性,我们该如何应对技术的断层与无奈。在越来越多的人文社会科学都宣称进行大数据研究转向的语境下,我们更需要逆流而思,冷静分析大数据方法热下的人文社会科学研究方法的大作为。

关键词: 大数据;人性;技术;研究方法

中图分类号: N02

文献标识码: A

文章编号: 2095-770X(2017)12-0001-04

PDF 获取: <http://sxxqsfxy.ijournal.cn/ch/index.aspx>

doi: 10.11995/j.issn.2095-770X.2017.12.001

When Big Data Encounters Humanity: On the Fault and Helplessness of the Technology

Li Wei-wei

(School of Humanities, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The beginning of the data science has witnessed the heated argument between model-driven operation and data-driven operation in business. Up to now, among the noise of discourse of algorithm domination, we should still be sober towards the trend of big data. When big data encounters humanity, what can people reply to the helplessness of the technology? In a period when an increasing number of humanity and social sciences proclaim to turn to the big data research paradigm, humanity studies need to forge a level head and analyze the big data method in an objective way.

Key words: big data; humanity; technology; research; method

随着信息技术的深入,加上各种媒体报道议程设置的推波助澜,越来越多的人知道了“数据”和“大数据”的概念。就连数学能力较为薄弱的很多人文学者现在也是言必称“大数据”,一时之间,大数据成了产业运营和学术研究的标准配置,大数据好像成了万能灵药,更有媒体不遗余力地鼓吹“21世纪最重要的战略资源是数据”。作为一篇研究方法的论文,本文无意对大数据技术进行刻意吹捧或诋毁,只想从更加客观和理性的角度出发来审视大数据方法,尤其基于互联网的人文社会科学研究中的大数据方法的不足以及改进之道。

一、“大数据”概念的滥觞与发展

“大数据”的概念首先源自于未来学家阿尔文·

托夫勒^[1](Alvin Toffler)《第三次浪潮》(1980)一书,但大数据发展的物质基础则是始于21世纪信息量的爆炸性增长。计算机公司首先推进了数据计算能力,但是大数据真正勃兴还是得益于营销公司对于市场与用户孜孜不倦的渴求。在“数据推动运营”模式的感召下,不仅互联网公司痴迷于数据采集与数据分析,就连房地产企业和汽车销售公司等也纷纷加入数据家族,优质低价地将顾客的信息转卖给任何感兴趣的个人或实体。笔者日前在买房时,房地产中介就强行安装某一客户端应用(不安装无法交易),结果就是不仅姓名等信息清晰可见,就连住址、单位、位置、银行卡号与交易密码等个人可识别信息(PII, Personal Identifiable Information)也得拱手托出,宛如“透明人”。

收稿日期:2017-06-12;修回日期:2017-07-13

作者简介:李炜炜,男,安徽桐城人,北京邮电大学人文学院讲师,主要研究方向:理论传播学。

大数据的应用外延并不仅限于商业领域,早在2012年的时候,美国前总统奥巴马就宣布启动“大数据研究和发展计划(Big Data Research and Development Initiative)”^[1],以提高美国的科研、教育与国家安全能力,这是继1993年美国宣布“信息高速公路”计划之后的又一次重大科技发展部署。美国政府认为大数据是未来信息时代的重要资源,战略地位堪比工业时代的石油。

学术界对于大数据的探索则主要集中在自然科学领域,2008年Nature出版了专刊“Big Data”,2011年Science推出关于数据处理的专刊“Dealing with Data”。与自然科学研究领域关注大数据对超级计算、互联网技术、生物制药的钟爱不同,人文学科更多是在介绍大数据方法对于人文社会科学研究范式的影响,如彭兰^[2]、喻国明^[3]、曾凡斌^[4]、黄升民^[5]等。

在众说纷纭的大数据解读中,产业和学界、自然科学和社会科学都比较公认的认知是大数据的“4V”特征,即,数据规模大(Volume)、数据种类多(Variety)、数据要求处理速度快(Velocity)、数据价值密度低(Value)。和传统的数据处理流程相似,大数据也是从“数据获取—数据分析—数据显示—数据处理”的流程来进行数据处理的,比较特殊的是大数据除了包括传统的数字和结构化数据以外,还包括了很多从文本或评论采集而来的文字和其它非结构化数据,以及通过传感器从物体上采集回来的物联网数据。大数据是一个介于云计算和人工智能中间的产物,因为大数据必然要求传统算法向云计算算法演进,而随着数据量的增加,计算机的深度学习(Deep Learning)能力也越来越强,从而在人工智能的路上又迈进了一大步。

二、大数据方法的不足

(一)大数据技术本身的不成熟

尽管大数据的呼声不绝于耳,但是毫无疑问,绝对的技术至上主义和计算神学^[6]是偏激的。既然大数据是一种技术,我们不妨首先就从技术角度来分析大数据技术本身的不成熟。

以大数据的第一个公认的标志性参数“数据规模大(Volume)”来看,目前很多宣称使用大数据的研究只能被称之为“较大的数据研究”,其规模远远不够。机器学习的原理是通过海量数据来从事深度学习的。举个例子来说,AlphaGo之所以打败了围棋高手,是因为在之前的屡战屡败中,AlphaGo积累了大量与高手对弈的数据,在这过程中其实现了深度智能学习。所以,如果不想AlphaGo胜利,方法很简单,不要请高手与其对弈,不要为其积累大量博弈数据就好了。能够被称之为大数据的数据量至

少需要有超过100TB的数据量,且需要包括非结构化数据等多维数据,目前我们很多的大数据应用的数据规模远未到“大数据”的量级。

其次,人工智能技术自20世纪70年代提出之后,沉寂了将近30年,直到近些年大数据算法的优化才又使其火爆了起来。原因很简单,技术本身并不成熟。计算机可以进行每秒钟数十亿次的计算和自动驾驶飞机,却学不会简单的骑自行车。因为驾驶飞机的技术是程式化的,而骑自行车却需要瞬间的反应以及相应的平衡力。在人工智能界,比较公认的结论是包括AlphaGo在内的机器人其实只是实现了弱人工智能,实现强人工智能的数据量还远远不足。

此外,人们现在已知晓大数据推送的不足,但从技术上尚无法克服,信息茧房效应就是不足之一。作为对现代社会信息过载的一种自我保护,多数用户顺从地接受了可以节省他们信息获取成本的个性化信息推送算法。“今日头条”、“一点资讯”、“天天快报”等新闻客户端主打的个性化新闻噱头的弊端已经日益显现。在企鹅智酷面向用户的调查中,认为个性推荐能完全满足获取资讯的用户为15.2%,另有70.3%的用户认为满足程度为一般。认为个性推荐的内容太少(32.6%)和认为它会让视野变狭窄的用户(32.3%)比例相当,而认为推荐内容不准(30.7%)和推荐内容低俗(29.4%)的比例也相当^[7]。

而且,很多用户都有这样的经验,我们可能是出于对标题党的好奇而点击的一条资讯可能在客户端引发后续的无限推送(美其名曰“猜你喜欢”),并且机器抓取的同类资讯常常是过时的信息,这不仅违反了新闻是“新近发生事实的报道”的定义,更与“阅读是为了发现未知世界”的初衷背道而驰,使得新闻价值无法体现,也使得我们在被大数据建构的碎片化、肤浅化和娱乐化的“媒介景观”中作茧自缚,无法提升媒介素养,信息辨识能力不强,失去批判的本能。

(二)大数据技术对于人性的不可知

相对于技术本身的掣肘,大数据技术对于人性的不可知更是让人唏嘘。我们人类的大脑尽管在数学计算速度上无法和机器相提并论,但是大脑懂得人性。人们擅长反射彼此的情绪状态,擅长侦测出不合时宜的行为,擅长用情绪为事物赋予价值。在可预见的未来,这恐怕都是大数据无法企及的高度。

在营销学中,人们津津乐道的案例是“即食通心粉”。按照大数据分析,消费者在烹饪即食通心粉的时候会加上一点洋葱,于是体贴的通心粉厂家就发明了一个新产品,在即食调料包里为消费者加上一些洋葱。但在实际销售时,没有加洋葱的通心粉却

比加了洋葱的新产品卖得更好。尽管大数据分析人们的行为数据,也得出了即食通心粉与洋葱之间的相关性关系,但是大数据却忽视了一种人性:家庭主妇在给家人烹制即食通心粉的时候,有一种没有尽到家庭主妇职责的内疚感,为了消除这种内疚感,她们会选择在烹饪通心粉时,加入一点自己准备的洋葱,表明这顿饭是自己精心准备的,自己不是一个偷懒的、不称职的家庭主妇,所以她们选择购买没有添加洋葱的即食通心粉。对消费者的深刻洞察不是来自于量化的研究数据和书面的研究报告,而是来自于与消费者的直接、深度接触中,比如街头暗访、消费行为的观察、与目标人群的谈话等等更接地气、更原生态的研究方法。与量化研究不能取代定性研究一样,大数据方法也不能取代假设。

计算机数据分析擅长的是测量社会交往的“量”而非“质”。数据科学家可以测量出你在微信上的互动数量,但是他们不可能捕捉到你心底在读朋友圈里分享的圈文时的情感。因此,在社交关系的决策中,我们不能一味地相信机器。

美国社会学家库利认为,人的行为很大程度上取决于对自我的认识,而这种认识主要是通过与其他人的社会互动形成的,他人对自己的评价、态度等等,是反映自我的一面“镜子”,个人通过这面“镜子”认识和把握自己^[8]。因此,人的自我是通过与其他人的相互作用形成的。从这个意义上说,基于大数据技术的用户画像画出来的可能是“镜中我”,而非“本我”。美国著名社会学家戈夫曼也认为人生就是一出戏,社会是一个大舞台,社会成员作为表演者都渴望自己能够在观众面前塑造能被人接受的形象^[9]。如果微博和微信等媒介中呈现出来的大都是“更好的”的伪装的自己,那么这肯定不是真实的社会状态。当媒介呈现的景观与真实的社会反差巨大的时候,对于缺乏媒介素养的青少年绝非利好。

此外,就算算法是机器的力量,但决定算法的还是人。最典型的例子就是使用不同搜索引擎搜出来的排序结果常常并不一致。算法的常见陷阱是它会带来偏见与歧视,而我们却常常对此一无所知。所以,在大数据算法时代,人还是需要保持自身对现实世界的洞察力与判断力。

(三)大数据技术对于隐私的侵害

相比起大数据对于人性洞察的无力,大数据技术对于隐私边界的侵蚀则是公认的事实。如今国际主流研究将信息隐私权看作一项社会权利。它的本质是“信任”,核心在于分享个人信息后依然能够保留某些控制。但是个性化信息服务与隐私通常是相矛盾的。

从技术层面来看,目前的技术是无法保证绝对的隐私安全的。互联网的底层技术支撑是通信网,

在量子通信没有普及以前,基于光纤的通信技术无法保证绝对的网络安全和信息安全。

现在大数据通常和移动互联网交织在一起,因此在互联网时代并不凸显的个人位置信息数据在移动互联网的环境下变得异常活跃。无处不在的无线网络(SSID)与有线通信网络一起编织起成一张巨大的“泛在网”,在这张网里,人们没有隐私,近乎“裸奔”。更加让人担心的是,人们似乎并不满足于人机互联,人们通过传感器接入了物体、通过移动的虚拟现实(VR)和增强现实(AR)设备构建场景,我们在实现万物互联的理想之时,也将自己置身于隐私被暴露和售卖的时代桎梏之中。

对于大数据时代对隐私的侵蚀引起的伦理问题,现在国际社会已经有了较多的技术手段和立法手段来进行规避和处理^[10]。但在笔者看来,切实提升个人的媒介素养以及养成对个人隐私的保护意识才是此类问题得以解决的王道。正如联合国教科文组织在2013年12月发布的“媒体和信息素养(Media and Information Literacy)”评估框架中所提到的那样,“每一个居民都需要和理解媒体和信息供应者的规则,以发挥他们在社会中的功能,了解更多来自虚拟世界的机会和威胁,学会管理资源”^[11]。

鉴于大数据时代个人隐私保护的困难程度,已经有人提出“遗忘”的必要性,甚至有观点称“被遗忘是人的一项基本权利”。尽管实施起来还是困难重重,但是我们欣喜地看到越来越多的用户逐渐了解了隐私保护的意义,像Snapchat这样主打“阅后即焚”的保护隐私的产品风靡也说明了这一趋势。有时候,笔者甚至认为现在技术整合复杂多维数据尚有难度,政府和企业一些数据由于不开放而导致的“信息孤岛”现象常被人认为是不足需要攻克,但从隐私保护的角度来看,这又何尝不是一种另类之“幸”呢?

三、大数据技术在人文社会科学研究中的应用和挑战

传统的人文科学研究方法主要是定性研究和量化研究。量化研究秉承实证主义的传统,主要使用基于统计学和概率论的方法来进行实验和调查,各种问卷与量表是量化研究中必不可少的分析手段。定性研究则更多地是从人文主义和理论建构的基石出发来收集人们的感受、见解与经历等资料,故而研究手段常常是参与观察与深度访谈。与量化研究主要使用演绎推理不同,定性研究常使用归纳推理法,带有浓郁的思辨色彩。

鉴于学科的分野以及研究方法的差异,传统的人文学科,如文学多使用定性研究,而社会学科,如社会学则多诉诸于量化研究。但是,进入到“数据推

动学术”之后,人文学科和社会学科皆出现了不同程度的“数据转向”热。跟自然科学走得比较近的量化研究方法对大数据方法的狂热程度更强,“数据主宰一切”、“算法统治世界”之类的声音不绝于耳。从研究范式到具体方法,从人才培养到产学研联合,大数据对人文社会科学研究产生了深远的影响。

从具体操作层面来说,与传统人文科学领域中的基于假设或模型的小网络小抽样不一样,大数据研究方法使用的是面向复杂网络的全样本分析。由于集成了自然语言处理技术、分词抓取技术、语料库语言学技术,现在的大数据研究可以更加快捷地分析人们的情感语义特征,从而使得用户画像更为准确,也就使得个性化的推介变得可能。

以新闻学为例,大数据在新闻上的应用涵盖了根据业务板块的需求定制发稿模板、数据自动抓取与采集、稿件自动生成等,那么传统意义上基于采编模式的新闻研究与新闻教育就需要更新。早在2011年,用 Narrative Science 软件写作的新闻稿件已出现在一些美国媒体上。2016年里约奥运会期间,机器写作的新闻体育报道更是大行其道,占据了数字化媒体的版面。除了体育新闻之外,财经新闻也是数据新闻和机器写作的先驱者,国内领先的互联网巨头腾讯网的财经新闻机器写作已经颇具规模。国内也已经有高校开设了“数据新闻”专业。相比于火爆的新闻业务方向的研究,对传统新闻学研究,尤其是新闻史论的研究则更加式微。

对于广大的人文社会科学研究者而言,其数据处理的能力本来就偏弱,更没有必要迷信大数据,鼓吹所谓的“大数据转向”。

具体来说,只有数据其实是没有意义的,不管是学术上的意义,还是商业上的意义。数据的价值无法充分发挥,并不仅仅是数据质量问题或是数据分析能力问题。问题的关键原因之一,还在于一门重要学科——运筹学(Operations Research)在学术界和产业界的稀缺。运筹学不同于数据科学,是一门致力于研究由数据到决策的科学。如果说数据科学旨在理解数据中的规律,运筹学则是将理解的规律为最后的决策服务,从而给决策者带来效益,以体现自身的价值。但在国内,学界和业界的合作尚未启步。作为学术水平要求较高的交叉学科,国内运筹学长期面临着人才匮乏的尴尬。多年来,中国都没有自己的优化求解器(solver),主要靠海外购买或者使用海外的开源求解器。

鉴于此,大数据学和运筹学等一众复杂的技术将会是人文社会科学研究者路上难以逾越的一道巨坎。那么,在大数据时代,人文社会学科,尤其是重实证和调查的社会学科难道就难有作为了吗?答案很明显是否定的!如果将大数据热看成一种“能指

狂欢”,那么原始数据本身可能也只是一种隐喻,数据从来都不可能是“原始”的,数据总是依照某些人的倾向和价值观念而被构建出来的。数据分析的结果看似客观公正,但其实价值选择贯穿了从构建到解读的全过程。人文社会科学的研究者完全可以摒弃自己不擅长的数据采集和分析,选择和自然科学研究者,尤其是运筹学研究者主动对接、深度合作,实现研究方法的跨学科交叉融合与协同创新,自己专攻价值判断,使用深描等方法来阐释意义,将量化研究与定性研究进行深度融合,以便获取更准确、更严谨、更有解释力、预测力和指导价值的研究成果。

四、结语:技术的断层与无奈

历史上,诸如麦克卢汉等学者常过分强调技术的加持,著作中常充盈着技术乐观主义的气质。但是,我们知道,不管在技术的演进还是在历史的更迭中,人才是最重要的常量。没有对人性的尊重,技术一定是处于断层之中,充满无奈。很多人文社会科学研究本来就是基于人性的深刻洞察,因此,人文社会学科在应用大数据方法时,要不忘初心,只有将人和人性至于整体性的技术图谱和生态中进行系统考查,才会有惊人的发现和深刻的研究。

【参考文献】

- [1] 阿尔文·托勒夫. 第三次浪潮[M]. 黄明坚,译. 北京: 中信出版社, 2006.
- [2] John Gantz, David Reinsel, The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, [J]. IDC iview, 2012(12).
- [3] 彭兰. 大数据时代:新闻业面临的新震荡[J]. 编辑之友, 2013(1).
- [4] 喻国明. 传播学研究:大数据时代的新范式[J]. 新闻记者, 2013(6).
- [5] 曾凡斌. 大数据对媒体经营管理的影响及应对分析[J]. 产业论坛, 2013(2).
- [6] 黄升民. 大数据时代,电视如何作为[J]. 南方论坛, 2013(3).
- [7] 彭兰. 机器与算法的流行时代,人该怎么办[J]. 新闻与写作, 2016(12).
- [8] 彭兰. 智媒化:未来媒体浪潮——新媒体发展趋势报告(2016) [J]. 国际新闻界, 2016(11).
- [9] 查尔斯·库利. 人类本性与社会秩序[M]. 包凡一,王媛,译. 北京: 华夏出版社, 2015.
- [10] 尔文·戈夫曼. 日常生活中的自我呈现[M]. 冯钢,译. 北京: 华夏出版社, 2008.
- [11] 陶媛. 联合国教科文组织发布全球媒体和信息素养评估报告[J]. 世界教育信息, 2014(3).