通往正义的数据

林 衍

中国青年报社,北京 100702

头一回在美国上统计课，在国内政府部门做过近 10 年数据统计工作的涂子沛觉得“眼前一亮”。

这一讲的内容是统计学的意义。大胡子的印度裔教授一本正经地对学生们说:“我们信靠上帝。除了上帝,任何人都必须用数据来说话。”

教授大胆的观点一下子震撼了中国学生涂子沛。在他的印象中,数据往往被当成论证工具,更像是一种“证明领导意图的手段”。

但在这里,“数据”二字似乎有另一重含义。

2007 年底,奥巴马访问了谷歌公司的总部。作为他的支持者,涂子沛在视频网站上观看了这次演讲。

“人民知道得越多,政府官员才可能更加负责任。”这位总统候选人一上来就表达了建设开放政府的决心。

面对谷歌公司的员工,奥巴马继而雄心勃勃地说:“我将把联邦政府的数据用通用的格式推上互联网。我要让公民可以跟踪、查询政府的资金、合同、专门款项和游说人员的信息。”

涂子沛记得,演讲在此时被热烈的掌声所打断。

这是涂子沛第一次将“公民权利”这样的大词与“数据”联系起来。也正是从那时起,这个从卡内基·梅隆大学毕业、并已经在一家美国公司就职的数据库程序员萌生了一个系统的写作计划。

2011 年 12 月,涂子沛将 21 万字的书稿寄给国内几家出版社。

一家出版社的编辑对他说:“不管别人开出什么样的条件,我们都要。”而另一家出版社的主编则给涂子沛回复了一封邮件,写道:“这是一本中国社会需要的书。”

2012 年 7 月,《大数据》一书出版。翻开这本以 0 和 1 的二进制代码图案为封面的新书,扉页上的题记别具一格:“一个真正的信息社会,首先是一个公民社会。”

这几乎相当于 13 亿中国人人手一本 1500 页的书加起来的信息量

涂子沛在美国匹兹堡市一家联邦政府的合同商公司做程序员,每天面对的东西都是数据、代码或大大小小的表格。

但无论从哪个角度观察,他都不是一个单纯的 IT 行业的从业者。涂子沛爱读刘瑜和陈丹青的书,和匹兹堡大学著名史学教授许倬云是好朋友,还会在一个人开车的时候听几段古典诗词的朗诵。他的房间里,大部头的编程书籍和不少从国内带来的人文类图书整齐码放在书架上。书房的窗外,大树的树叶伸手可及,他喜欢对着一片新绿写作。

20 世纪 70 年代初,涂子沛出生在一个法官家庭。因为从小看父亲断案,他很早就开始思考什么是“正义”这样的大问题。20 世纪 90 年代中期,他进入当时的华中理工大学读书。学校里有人文讲堂的讲座活动,这个计算机专业的学生是最忠实的拥趸之一。讲座结束后,他常会在昏暗的路灯下追着讲师们请教问题,直到老师的家门口。

时任中国人民大学副校长的谢韬也来过人文讲堂。当时他给涂子沛签名并留下这样一句赠语:“要做新世纪国家的建设者。”

年已不惑的涂子沛至今仍然保留着这份情怀。他在博客上记录匹兹堡市的学生们如何因为征税问题质询市长,也写下这里的市民对阻止他们游行的匹兹堡警方怎样提起集体诉讼。

2012 年 7 月,在微博上看到“7.21”北京暴雨的新闻后,他连夜写了一篇介绍美国如何发布气象灾害信息的文章。

“没有有效的预警,再准确的数据预报也是一个零。”涂子沛在文中介绍到，美国国家天气服务局不仅开通了推特、脸谱等社交媒体账号,还推出了一种叫做天气收音机的预警产品。一旦气象预警后,平时沉默不语的“收音机”会立刻成为“闹钟”。负责短信预警的部门则专门开发了一个系统，能根据用户手机发出的信号,来判别其是否位于暴风或者恶劣天气覆盖的区域,再决定是否发送信息,以提高准确性并减少信息扰民。

事实上,正是从看到奥巴马 2007 年在谷歌公司的演讲开始,涂子沛才意识到,原来自己一直关心的“公平正义”,竟然与每天朝夕相处的数据有着如此紧密的联系。

在此以前,和大部分人一样,涂子沛更愿意从技术层面去关注什么才是 “大数据”——这是对信息爆炸时代的崭新描述。它的基本单位是“太”(TB),而 1000 个太则等于一“拍”(PB)。打个直观的比方,美国国会图书馆是世界上最大的图书馆之一,它所有印刷品的信息量加起来只有 15 太。而全美

本文转载自《中国青年报》2012 年 8 月 22 日 12 版冰点特稿。

国仅在 2010 年一年的新增数据量就足足有 3500 拍,这比 13 亿中国人人手一本 1500 页的书加起来的信息量还要大。

麦肯锡咨询顾问公司曾做出估测,未来数据仍然会以每年 50%的速度增长,美国还需要 14 万—19 万名拥有"深度分析数据"专长的工作者。

涂子沛便是其中之一。他会在自己的专栏文章中记录那些数据改变商业的故事:比如,沃尔玛的研究人员通过数据挖掘,发现 4 成左右的年轻爸爸在购买婴儿尿布时会顺手买点啤酒犒劳自己,便对这两种商品进行了捆绑销售,结果销售量双双增加。更夸张的事例是,一个高中女孩某天突然收到了超市寄来的婴儿服广告,父亲大为光火,但就在超市公开道歉几天后,这位父亲发现自己的女儿真的怀孕了。原来,超市已经可以通过顾客的食品消费数据做出趋势判断。

事实上,数据挖掘已经在美国形成了一条完整的产业链,不少大学还设立了相关的硕士学位。

但当涂子沛开始写作《大数据》一书时,商业已经不再是他头脑中最重要的东西。

他想在书里讲一个数据与正义的故事。

所有的政府机关都应该利用最新的技术推进信息公开,这种公开应该是及时的

2009 年 1 月 17 日,新任美国总统奥巴马主持内阁的宣誓仪式并发表讲话。

"很长时间以来,我们已经习惯了为华盛顿这个城市蒙上一层神秘的面纱,但从今天起,我们将不再沿承旧例。"

"为了引领一个开放政府的新时代,面对信息,政府机关的第一反应必须是公开。这意味着我们必须坚定地公开信息,而不是等待公众查询。所有的政府机关都应该利用最新的技术推进信息公开,这种公开,应该是及时的。"

这一天,奥巴马伏案用他标志性的左手姿势签署了他的首份总统备忘案《透明和开放的政府》。

120 天后,一个叫做 Data.gov 的网站正式上线发布。这是一个数据开放的门户网站,旨在全面开放美国联邦政府拥有的数据。

项目的负责人是维伟克·昆德拉,美国政府历史上的第一位首席信息官。"这是一场数据民主化的运动,我们正在把信息的力量放到美国人民手中。"这个不到 35 岁的印度裔帅小伙同样雄心勃勃。

事实上,作为雇用了约 200 万名工作人员的全美最大雇主,联邦政府共拥有 848 拍字节的数据总量。全世界最大的零售巨头沃尔玛,其数据库大小还赶不上商务部下属的美国普查局。

昆德拉的任务就是把它们统统摆上互联网,供全世界任何人访问和下载。

起初,这个数据开放网站有点名不副实,即使包括地理数据在内,这个新生网站上也仅仅只有 47 组数据和 27 个数据分析工具。

要知道,年轻的首席信息官要面对的,是美国政府 15 位内阁部长,以及 70 多个独立机构的局长、主任和主席。

"这层层关卡的背后,是一种沿袭了上百年的行政文化。面对庞大的公共信息,政府首脑的第一反应往往是安全第一,保密为上。"涂子沛在《大数据》中写道。

信息之于民主,就如货币之于经济

尽管早在建国之初,《独立宣言》的起草人、美国第 3 任总统托马斯·杰斐逊就曾说:"信息之于民主,就如货币之于经济。"但事实上,在美国 200 多年的历史中,开放政府信息的雄心并不多见。

1789 年,美国国会制定了《管家法》,规定了行政机关必须在统一的出版物上公开政务信息,但对于公开的内容,行政长官有自由裁定权。进入 20 世纪后,政府规模不断膨胀,国会先后通过了《联邦登记法》和《行政程序法》,规定公众可以向政府提出信息公开的要求,但如果危及公共利益,政府有权拒绝。

简言之,信息是否公开,最终还要政府说了算。到了 1951 年,在当时的冷战格局下,美国历史上首次把军事机构的保密制度引入普通的行政部门。一时间"保密"之风在政府各个部门蔓延。

同一时期,哥伦比亚大学的新闻学教授克劳斯对当时的信息管理情况进行了调研,并出版了《人民的知情权》一书。他在书中明确提出:只有拥有信息自由,人民才能真正拥有对公共事务的发言权。但美国的"知情权",并没有明确的法律保障。所有行政部门的档案盒记录,实际上都处于"准机密"的状态。因为是否属于"机密",完全取决于当权领导一时的态度。

后来这本书被誉为美国信息自由运动的"圣经"。书出版那一年,一个叫约翰·摩斯的新任国会议员走进了首都华盛顿。不久后,他提出了《信息自由法》的草案。

在摩斯任职国会议员的 25 年里,经历了数任美国总统。然而即使是以开明著称的肯尼迪,也对这个草案避犹不及。约翰逊在听到这个法案后的第一反应则是:"摩斯想干什么?他想搞砸我这届政府吗?"

直到 1964 年,摩斯的提案才进入辩论阶段。那时候的联邦政府有 27 个部门,无一例外地在听证会上大声反对。白宫新闻秘书莫耶斯甚至在法案的封面留下了这样的字迹:"对!我们必须反对这个法案!"后来,莫耶斯曾回忆过这段历史,他谈到了总统约翰逊对于这个法案的厌恶和无奈,并承认当时的联邦政府确实以国家安全的名义,掩盖了太多的秘密。

1966 年 7 月 4 日,由于参众两院对于该法案的支持率都高于 2/3,无法动用否决权的总统约翰逊在家中签署了这份法案。信息自由胜利的这一天,正是美国的独立日。

12 年后,六朝元老摩斯的国会生涯画上了句号。这位《信

息自由法》之父回顾说:“我们所做的,只是一个开始。那最好的仗,我已经打过。”

如今,仅联邦政府每年就会收到 50 多万宗信息公开的申请,如果政府以保密的要求拒绝,公民可以提起司法诉讼,法院拥有是否公开的最终裁判权。

Data.gov 的出现则为信息公开的发展注入了新的活力。昆德拉在演讲中表示,政府数据作为一项公共资源,应该像天气预报、体育赛事和股票信息一样实时公开。

充满了挑战性的新事物同样引发了联邦政府各部门头头脑脑们的忧虑。有人担心,万一民间机构根据原始数据分析出的结论与政府不一样,是不是等于“搬起石头砸自己的脚”;还有人认为,这场面对全世界的数据开放运动,会在不知不觉间让国家利益受到损害。

但改革并没有止步。截止到 2009 年底,这个网站收到了社会各界约 900 项开放数据的申请。联邦政府最后回复:16%的数据立即开放、26%将在短期内开放、36%将计划开放,还有 22%因为国家安全、个人隐私以及技术方面的限制无法开放。与此同时,行政管理预算局发布了《开放政府的指令》,命令各个联邦部门必须在 45 天之内,至少再开放 3 项高价值的数据。

在 Data.gov 上线发布一周年的时候,联邦政府开放数据的总数已经达到了 27 万项。

因为这些创新型的应用,数据的能量将层层放大

在涂子沛看来,创建 Data.gov 的价值并不仅仅在于满足民众的知情权。

2009 年 1 月,美国联邦政府跨部门工作组曾做出一份报告提供给总统科学技术委员会,该报告这样写道:一组数据,可能会得到数据收集人难以想象的应用,也可能会在另一个看起来毫不相关的领域得到应用,而因为这些创新型的应用,数据的能量将层层放大。

某种程度上,Data.gov 的确掀起了一个社会创新的浪潮。截止 2011 年 12 月,在这个政府主导的数据开放网站上,汇集了 1140 个应用程序和软件工具、85 个手机插件。其中有近 300 个应用程序由民间的程序员、公益组织自发开发。

就在 Data.gov 上线不到一个月的时候,民间的一位程序员便利用美国交通部开放的数据开发了一个航班延误的免费查询系统。

在《大数据》里,涂子沛用可视化图表的形式呈现了 2010 年波士顿至纽约的航线情况:

这一年总共有 6735 次航班,其中 62%准点或提前到达,14%有 20 分钟以内的延误,20%有 20 分钟以上的延误,还有 4%的航班最终取消。

如果你打算把航班延误的事情搞清楚,必须使自己能够承受更多的数据冲击。在这条航线上,天气良好的情况下,多数航班会提前 5 分钟到达;但下雪天的平均延误时间为 7 分钟;下雨天的平均延误时间则为 4 分钟。

就“打飞的”而言,出行的时间也至关重要。通常星期六将有 78%的最高准点率,而不幸赶上星期一坐飞机的人则要承受 30%的最高延误率。而在一天之内,每天早晨 5 点到 6 点起飞的航班有 83%的最高准点率;而晚上 9 点到 10 点的准点率则低达 53%。

好消息是,感恩节、圣诞节当天的飞机一般都会准时到达。坏消息则是节日前后的延误情况颇为严重,比如圣诞节次日的平均延误时间就有 34 分钟,最长需要等待 80 分钟,更糟的事情则是航班居然有 41%的可能性被取消。

当然,这个应用程序最有力度的数据还是一份所有航班的延误大排名,“Delta1807” 便以平均延误 26 分钟成了最不靠谱的航班。

这样的数据挖掘直接给那些落后的航空公司敲响了警钟。其实早在 Data.gov 之前,国家交通安全局曾经提供过一个“航班等待时间计算器”。后来发现这个民间免费工具如此强大后,交通安全局便知趣地关闭了那个计算器。

联邦政府首席信息官昆德拉在接受媒体采访时,也曾高度赞扬这个由民间开发的第三方程序,认为这正是奥巴马政府希望促成民众共同参与解决社会问题的最好例证。

类似的故事并不鲜见。不少人习惯于在购物时,先到 Data.gov 上查询商品是否存在质量问题召回的记录。而在 2010 年发生的一起矿难案件中,网民们通过 Data.gov 上的煤老板捐款数据记录,发现接收款项最多的个人居然是地方上诉法庭的法官。后来,最高法院依此认定这名法官存在“重大偏袒”的可能。

随着 Data.gov 在美国大获好评,从 2009 年 5 月起,已经先后有加州、纽约州等 31 个州和芝加哥、亚特兰大等 13 个大城市推出了各自的 Data.gov 门户网站。

“数据权”是信息时代每一个公民都拥有的一项基本权利

推动数据开放运动的国家并不只有美国。

2006 年 3 月,英国的《卫报》刊登了一篇名为《把皇冠上的明珠还给我们》的专栏文章,这被视为英国数据开放运动的序幕。

设计了全世界第一个网站的伯纳斯·李曾被评选为最伟大的英国人。2009 年 2 月,他受邀在 TED 大会上发表演讲。

“你想象不出政府会找出多少个借口来拒绝开放数据。”这一天,一向以“内向”闻名于英国新闻界的伯纳斯·李一改往日作风,甚至在演讲台上带领观众们一句一句地喊起口号:

“原始!”

“数据!”

“现在!”

“原始数据,现在就要!”

此后不久，他和时任英国首相戈登·布朗一同出席一次颁奖典礼。布朗问他，英国政府应该如何利用互联网。伯纳斯·李立即回答说：把政府的数据推上互联网。

2010 年 1 月，英国政府的 Data.gov.uk 正式上线发布，第一天就公布了 3000 多项民生数据。而这个时候，已经经营半年多的美国 Data.gov 还仅仅有 1000 多项民生数据。

“为什么别人一出手，数据量就是我们的 3 倍?”美国的报纸发出惊呼。而在卡梅伦出任英国首相后，更是率先提出了“数据权”的概念，并将其视为信息时代每一个公民都拥有的一项基本权利。后来，两国间的竞争被伯纳斯·李称为“美丽的竞赛”。

2011 年，这场竞赛被扩展到全世界。这一年的 9 月 20 日，由 8 个国家发起的“开放政府联盟”在纽约成立。几个月后，开放政府联盟又迅速收到了加拿大、意大利、韩国等 42 个国家或地区的加盟申请。其中，有 31 个国家或地区都建立了公共数据的开放网站。

这个联盟里，最引人瞩目的国家并非美英，而是来自非洲大陆的肯尼亚。

2010 年 8 月，肯尼亚通过了新的宪法，其中第 35 条规定：“每一个公民都有权获得政府拥有的信息……每一个公民都有权修改、删除(政府保存的)不真实、有误导倾向的错误信息。”

一年后，总统齐贝吉宣布，肯尼亚人拥有了属于他们的公共数据开放网站，它的名字叫做 openData.go.ke。

《1984》电影海报中的“老大哥”直接被换成了布什的头像

美国联邦政府在推动数据开放运动中功不可没，但在日常生活中，这个庞大的官僚机构却不得不在数据问题上面对种种“憋屈”。

美国人通常将政府的信息收集工作视为对公民和社会造成的负担，为此，联邦政府特意成立了一个叫做“信息和管制办公室”的机构，其主要任务就是编制“年度信息收集预算”。这份预算与钱无关，它要计算的是政府机构的信息收集计划会给全社会带来多大的负担。1995 年的时候，这个数字是 65 亿小时，这相当于 320 万人整整一年的工作量。从那一年开始，信息和管制办公室不得不制订出更审慎的“减负计划”，力争每年的“信息扰民”时间减少 10%。

随着大数据时代的到来，2009 年联邦政府数据的信息收集负担不减反增，居然达到了 99 亿小时。这个办公室的同志们十分“搓火”，不得不隆重推出了 5 项减负措施，其中包括了尽量使用电子签名以减少信息传送、投递时间这样的招数。

更令人意想不到的事情则是，在美国这个以知识产权保护为荣的国家，联邦政府所拥有的数据竟然没有版权可言。原来，美国的《版权法》早在 1976 年就已经规定：对于联邦政府的工作和作品，无论是文字、图像、软件，还是信息和数据，只要是美国联邦政府工作人员为了完成本职工作而取得的成果，都不能申请版权。

没有版权，就意味着政府发布的数据在法理上失去了收费的理由和根据。不过，“便宜没好货”这样的老话在这里却万万行不通，他们还要面对另一道紧箍咒《数据质量法》。这条法案规定，联邦政府发布的数据，其获得的方式、产生的方法必须是透明公开的，而且别人通过相同的方法，应该能够产生、复制相同的数据。

事实上，由于数据的质量问题，联邦政府没少被较真的企业或个人告上法庭。

数据帝国里，“吃力不讨好”的故事不止于此。

从 1965 年开始，当时的预算局就建议，成立一个统一的“数据中心”，最终的目标是为全国每一个人建立一个数据档案，档案里将包括每个人从摇篮到坟墓的所有数据记录。

预算局将这个大型数据库称为“中央数据银行”。但很快，这一创举就被民意推翻，人们担心一旦全部数据经过彼此印证与互相揭示，个人隐私会无所遁形。

而在“9·11”爆发一年后，中央数据银行计划再次浮出水面，并更名为万维信息触角计划。

没想到，这个为了推动反恐工作而酝酿的计划竟然再次搁浅。《纽约时报》的专栏作家萨菲尔将该计划称为“超级侦探的终极梦想”，美国公民自由联盟更是直言不讳地警告：“美国人民将生活在奥威尔《1984》中所描绘的监控当中，唯一不同的是，监控我们的不是电幕，而是数据库！”

普通老百姓则用行动表达了情绪，他们直接把《1984》电影海报“老大哥在看着你”中的“老大哥”，换成了时任总统布什的头像。

一个真正的信息社会是一个信息自由流动而不受操纵的社会

《大数据》一书在国内出版后，涂子沛在一家图书网站上看到了这样一条推荐语：这里有中国的问题，这里有中国的财富，这里有中国的乡愁。

“当时就掉眼泪了。”曾经在边防部队生活过 8 年的涂子沛压低声音说，他平时车里放的歌就是罗大佑的《乡愁四韵》。在 8 个月的成书过程中，白天他从事技术工作，晚上常常写作直到半夜，次日开车上班时再继续构思。好几次下班，都浑然不觉地把车开过了家门口，绕个大弯才掉头回来。

其实，他的书中与中国有关的内容并不多，却大多富有意味。

2011 年，麦肯锡公司曾以 2010 年度各国新增的存储器为基准，对全世界大数据的分布做了一个研究，结果发现中国这一年新增数据量约为 250 拍，不及日本的 400 拍、欧洲的 2000 拍，和美国的 3500 拍相比，则连 1/10 都不到。与此相对应的另一个数据是，中国拥有 4.8 亿互联网用户，几乎是美国

的两倍;拥有近 9 亿部手机,是美国的 3 倍,而互联网和手机正是产生数据的重要来源。

在涂子沛看来,这意味着中国并不缺乏可供收集的数据,而是缺乏收集数据的意识。

他回忆起,还在卡内基·梅隆大学读书时,有一回师门聚会,大家相约每个人都要贡献一个拿手菜。一位来自中国的博士生以一道卤牛肉赢得满堂彩,但当他公布自己的烹饪配方时,却令一位美国教授不知所措。原来,这位教授不太能理解“盐少许”、“酒若干”、“醋一勺”这样的表述到底是什么意思。

涂子沛记得,在国内做程序员的时候,要是做一个数据系统供本单位使用,那么上级和下级单位一般都无法登录这个系统。还有人告诉他,国内有些城市会把环境监测点刻意设在人工湖畔的柳树林中,或湖中心的小亭子里。

这些都让他想起美国《数据质量法》中的严苛规定:任何联邦政府部门收集的数据,必须无偿与其他部门共享。而在发布数据时,必须同时发布一系列的文档,说明数据的来源、产生的方法,以及用户复制过程当中可能出现的问题和错误。

涂子沛把他的乡愁写进了书的尾声,题为:《挑战中国,摘下“差不多先生”的标签》。他在文中提到胡适对于中国人“凡事差不多、凡事只讲大致如此”的判断,也引用了史学家黄仁宇认为中国在历史上缺乏"数目字管理"这种现代治国手段的观点。

“数据不是任人打扮的小姑娘,漠视精确就是不尊重事实。”回国时,他还以“用数据说话,而不是用数据说谎”为主题进行了演讲。

前不久,国内一位学者针对一项社会调查,得出了“科学主义一定会导致严重的偏颇,其具体弊端就是迷信数据”的观点。

涂子沛第一时间完成了一篇专栏文章。他反驳道:中国社会治理领域的问题恰恰不是数据迷信。相反,现实情况往往是,决策者没有合理使用数据,同时又受制于错综复杂的理念和利益之争,导致数据意识形态化,在中国缺乏公信力。

他的观点是,收集数据,使用数据,开放数据,都是大数据时代我们中国人需要一一面对的挑战。

“如果前两者是文化和习惯,那后者则是一种态度。”涂子沛强调,一个真正的信息社会是一个信息自由流动而不受操纵的社会,这种开放意味着信息与每一个公民之间都是等距的,当然,也意味着公平与正义。

令涂子沛欣慰的是,《大数据》一书出版后,有几位读者给他发了邮件,他们的愿望是建设一个非政府组织,推动国内的数据开放运动。

更棒的事情是,2012 年 5 月,复旦大学的硕士研究生吴恒建立了一个叫做“掷出窗外”的网站,上面收录了全国 2000 余条有毒食品数据供社会查询。这让涂子沛想起,2004 年的时候,普林斯顿大学的本科生乔舒亚·陶伯拉曾建立了一个叫做 TrackGov.us 的网站,专门收集和公开国会的法案信息。《纽约时报》根据这个网站所提供的数据,揭露了奥巴马、希拉里与麦凯恩上百次缺席国会投票的事实。

“太棒了,他们真的很像。”涂子沛兴奋地笑出了声。

·学术动态·

“第一届全国软土工程学术会议”征文

“第一届全国软土工程学术会议”将于 2013 年 11 月 15—17 日在上海召开。此次大会是由中国土木工程学会土力学及岩土工程分会与软土工程专业委员会共同主办。

征稿范围:软土强度与变形特性;软土特性测试技术;软土基础工程理论设计与施工;软土地下工程理论设计与施工;软土地区重大工程实践;软土工程的其他问题。

全文截稿日期:2013 年 4 月 30 日。

通信地址:上海市四平路 1239 号同济大学 地下建筑与工程系。

电子信箱:softsoilcom@163.com。

大会网站:http://geonjut.42137.east-icp.cn/show.asp?id=706。