

“Sci-Culturomics” 与科学文化研究

李大光

(中国科学院大学, 北京100049)

[摘要] 中国是一个历史悠久的国家, 具有深厚的传统文化。在科学技术迅速发展的过程中, 中国已经进入到科学发达国家行列。在大数据时代, 应该用大数据的理论和观察在科学文化中中国文化的演变。这个演变过程不仅包括了公众对于科学技术的理解, 同时也包括技术的应用对传统文化的冲击和影响, 以及这种冲击带来的理解偏移和态度变化。本文在欧美学者的“Culturomics”理论与研究方法的启发下, 提出“Sci-Culturomics”概念和研究模式。同时, 在调查研究的基础上提出在中国现阶段的大数据条件下的研究可能性与价值。

[关键词] 科学文化组学 科学文化 大数据 模型 可能性

[中图分类号] N02 **[文献标识码]** A **[文章编号]** 1673-8357 (2015) 04-0021-08

“Sci-Culturomics” and Studies on Science Culture

Li Daguang

(University of Chinese Academy of Sciences, Beijing 100049)

Abstract: China is a country with a long history and profound traditional culture. China has taken faster steps in the advancement of science and technology. The evolution of the culture should be observed with the research model of Big Data in the Internet and computation age. This culture change dose not only include an understanding of science and technology but also the change of attitudes towards technology application, and its impact on traditional culture. This article brings up a concept “Sci-Culturomics” inspired by the concept “Culturomics”. The author also talks about the possibility and value of the research model under the current context of big data in China based on survey.

Keywords: Sci-Culturomics; scientific culture; big data; model; possibility

CLC Numbers: N02 **Document Code:** A **Article ID:** 1673-8357 (2015) 04-0021-08

国际科学素养调查显示出公众对科学技术的了解受到宗教和传统文化的影响。在传统文化的冲击下, 公众对科学技术的理解和态度产

生偏移, 甚至对抗。在大数据时代, 科学文化的研究成为可能。中国在现有条件下, 可以实现对科学文化的研究的尝试和实验。“Sci-Cultur-

收稿日期: 2015-02-15

作者简介: 李大光, 中国科学院大学人文学院教授, Email: ldaguang@vip.sina.com。

omics”在研究科学文化中可以进行尝试，中国的文化特点也需要进行尝试性的实验。任何新方法都需要有人提出和作出初步设计。本文对“科学文化组学”的概念的形成和研究模式进行了初步探讨。

1 大数据的研究应用

大数据时代的开始时间难以确定。很多人认为是2010年。这个看法可能是大数据的概念变得清晰化，同时开始有意识地在各个领域用大数据技术开始研究的时代。关于大数据（Big Data）概念的起源大概可以追溯到2000年斯隆数据调查公司（The Sloan Digital Sky Survey）在位于美国新墨西哥州太空望远镜对宇宙观察的累计数据进行长达几周的分析，整理出天文学研究的全部数据后，将所有的观测数据转化为可以计算的“粒子”，即最小单位，进行储存。大约10年后，其数据库信息量增大到140兆兆节（terabytes）。而其继任者，2016年即将到来的智利流而诞生的“大口径综合巡天望远镜”（The Large Synoptic Survey Telescope），每五天就将获得同样大的数据量。

利用大数据获利的主要是工业和商业。客户的信息记录不仅使得商家获取了巨大的经济利益，而且将带动资本的投资，引导经济的新趋势。世界最大零售商沃尔玛公司在其全世界8860多个超市中积累的数据已经达到2.5拍字节（petabytes），相当于世界最大图书馆美国国会图书馆藏书总量的167倍。关于客户的所有信息，包括出生日期到喜好，甚至嗜好都了如指掌。可以说，通过数据了解的客户超过客户本人对自己的了解。但是与此同时，商业对客户信息的无限量的索取和使用引发了隐私权的争论，从而进一步带动了与商业有关的法律的改革。

社会媒体的大数据研究由于各个国家的电子媒体受到管理体制限制、电子数据的开发和用户的文化习惯而有所不同，但是，大数据时

代的技术，主要是电子技术仍然在遵循摩尔定律（Moore's Law）发展。估计到2020年，“微处理器的单位价格每18个月就下降一倍，带宽和存储器的价格甚至下降得更快”的定律将仍然有效。美国的“Facebook”已经能够储存400亿张图片。电子技术的迅猛发展使得含有30个碱基对的人类基因组解码时间在“人类基因组工程”开始到2003年用了整整10年，而现在只需要一周。

在大数据时代，整个世界都处于数据的笼罩之下。数据在云计算、云储存和对数据粒子的分析、计算以及机器学习和预测中引导科学研究、城市管理、流行病控制、犯罪控制以及人类了解自己的研究——社会学。

约翰·霍普金斯大学教授亚历克斯·绍洛伊（Alex Szalay）认为，大数据时代呈现出的巨量增长使得对数据的使用、分析和预测变得愈加困难。“如何理解这些数据？这些数据的价值利用不仅仅是科学家的事情，而且是政府管理人员和工业界人士的基本能力。人们应该对培养下一代应用大数据而进行思考。”^①

目前，大数据研究主要在自然科学、工业和商业领域，而社会学领域相对落后。处于落后的原因除了利益原因外，更重要的是方向不够具体化。也就是说，其价值与方向的一致性不确定。但是，大数据时代是未来的金矿和石油宝藏，对于社会学来说，同样具有重要的意义。在现代电子技术的催生下，人类日常生活中从电话和信用卡、电视和电脑、城市基础设施、带有传感器的建筑物、火车、汽车、飞机、桥梁以及工厂等所有传播信息的渠道中每时每刻带来的数据流在不断积累，在过去的两年时间内，积累了惊人的10万亿亿字节（zettabyte），超过了人类文明史上记载的最高纪录。美国计量社会科学研究所以信息技术研究中心（Information Technology, Institute for Quantitative Social Science）教授盖瑞·金（Gary King）认为：“这是一个大数据革命。大数据革命意味着，我们现在

^① “Data, Data Everywhere”. The Economist. 25 February 2010. Retrieved 9 December 2012. <http://www.economist.com/node/15557443> (2015, 3)。

可以通过数据做任何事情。”^②

但是,到目前为止,北美两个国家,即美国 and 加拿大的大数据量为世界半数,欧洲的大数据量约为这两个国家的半数。处于第三位的是日本,而中国则落后于这些国家^[1]。中国如果想在未来成为强国,必须开展大数据的研究和应用。

科学技术,尤其是技术应用在特定的文化和发展阶段中呈现出传统文化、传统伦理、信源公信力以及意识形态,甚至信息管理体制等方面的影响下,而呈现出复杂的状态。因此,仅仅对科学素养进行研究,或者仅仅对科学知识在网络时代的传播状况进行研究,由于其社会学指标的单一,其成果被认为缺乏一定的精确性。但是,在模型设计和社会学调查和信息技术研究中并不十分困难。分析公众与科学技术信息之间的相互影响,从而及时了解公众对科学技术的态度,不仅仅对社会稳定和科学技术的民主化和透明化具有重要的意义,同时,对于公众理解科学的发展趋势进行有效预测,从而对制定科学技术政策具有重要价值。

2 “Culturomic”与文化研究

欧盟的建立意味着欧洲的全面一体化。欧洲一体化不仅仅是经济和货币一体化,同时也进行科学技术研究体制的改革。欧盟成员国打破国家领土界限,形成研究区(Research Area)。欧洲核子研究委员会(European Organization for Nuclear Research, CERN)耗资 54 亿 6 000 万美元在苏黎世地下 100 米深处修建了长度将近达到 28 公里(27.98 公里)的环形强子对撞机(Large Hidron Collider, LHC)。这个世界上最大的机器具有 1 亿 5 000 万个传感器,每秒产生 6 000 亿个数据,超过世界所有数据总量的 60 倍。质子束在接近 99.9999% 的光速对撞,产生粒子浓汤,从中分离出“上帝粒子”,即希格斯·波色子(Higgs Boson)。这个物理学的重大发现不仅意味着理论物理中心从美国转移到欧洲,并且获得了诺贝尔物理学奖。

2009 年,美国生物学家克里格·温特(Craig Venter)用 100 万美金将 DNA 双螺旋结构发现者和诺贝尔生理或医学奖获得者詹姆士·沃森(James Watson, 1926—)的个人基因图谱解读,从隐含在 60 亿个碱基对的密码中分析出其遗传基因,从而判断出其可能寿命和由于基因缺损而导致的疾病。2001 年 5 月,克里格·温特将 4 个碱基对组合为一个基因,用组合基因组成基因组,将人工制造的基因组注入一个被敲除了基因的细菌中,这个细菌被激活,人造生命诞生了。这个震惊世界的发现再次将人类带入到数据世界。

物理研究中的大数据计算和分析方法被广泛地应用于工业数据管理和商业数据管理与预测。而生物学研究中的数据分析和组合分析对社会学研究和文化研究产生了启示作用。

社会学家让-巴蒂斯特·米歇尔(Jean-Baptiste Michel)和其他社会学家和文化研究学者在其《对百万电子书进行文化量化分析》(Quantitative Analysis of Culture Using Millions of Digital Books)中用生物学中“基因组学”(Genomics)对文化量化进行分组定义,创造了“文化分组学”(Culturomics)概念。“文化分组学”的概念为:“文化分组学是采用对数据文本进行量化分析,从而进行人类行为和文化趋势研究的一种计算词汇学(computational lexicology)方法。研究人员通过对电子档案的数据挖掘,对隐含在语言和词汇使用的文化现象进行分析。”^[2]

图书出版总量阶段分析:1500 年代:每年出版仅仅为几本书,字数约几万字;1800 年代:每年 980 万字;1900 年代:每年 18 亿字;2000 年代:每年 110 亿字。按目前每年出版总量,如果一个人不吃不喝不睡,不做任何事情,只读 2000 年出版的图书,以每分钟 200 字的阅读速度,需要阅读 80 年。字母链长度比人类基因组长 100 倍,大约从地球到月球往返 10 次。

米歇尔等人用生物学中对基因组作为分析

② Jonathan Shaw. Why “Big Data” is a Big Deal. Harvard Magazine. <http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal>(2015,4)。

基本单位，从中分离遗传信息，从而发现基因进化与变异过程的方法进行“文化基因”的解读。“文化分组学”的基本分析单位为“1-gram”或者“n-gram”。“1-gram”设定为一个不可分割的单词，比如“banana”；或者一组不可分割的数据，比如“3.1415926”。而“n-gram”则是由“1-gram”组成的信息粒子链，多为短语（phrase），如“stock-market”（2-grams）或者“The United States of America”（5-grams）。在数据分析中，为了防止粒子链的断裂和分离，将“n-gram”的长度规定为不超过5。在粒子链的解读中，“n-gram”出现40次为1链（“n-gram”）。出现频率计算以一年中出现的“n-gram”的次数。

采用这种单位计算方法进行文化基因分组研究的研究人员认为，如果用时序模式研究，最好的方法就是已经数据化的图书。从15世纪中叶，古登堡（Johannes Gensfleisch zur Laden zum Gutenberg，1398—1468）发明活字印刷术以来，共出版了1.5亿种图书。迄今为止，其中的1500万本已经由Google和各个大学出版社进行了数据化，约占出版总数的12%。这些研究人员将其中的500万本可以用“光学字符辨识”方法识别的字符，根据研究内容进行符号粒子设计。文字范围为：英语：361亿；法语：450亿；西班牙语：450亿；德语：370亿；中文：130亿；俄语：350亿；希伯来语：20亿。

数据化的图书中，以最重要的词典中以频率最高为统计指标，英语词典中频繁（common）出现的词汇已经超过10亿个。以1900年、1950年和2000年前10年出现的“1-gram”频率分别为1117997、1102920、1489337。词典观察主要是英文词汇的变化以及词汇的收集编辑入词典的变化。

在数据图书的研究中，用于实验研究的文化组分为以下几类：

1. 名人跟踪

(1) 法国画家马克·夏卡尔（Marc Chagall，1887—1985）；

(2) 列夫·托洛斯基（Leon Trotsky，1879—1940）；

(3) 格里高利·佩雷尔曼（Grigory Perelman，俄罗斯著名数学家，庞加莱猜想解读者，菲尔兹奖获得者）；

(4) 列夫·鲍里索维奇·加米涅夫（Lev Kamenev，1883—1936，俄罗斯政治家）。

2. 好莱坞明星跟踪

3. 纳粹迫害艺术家与作家、政治家、文学家、历史学家、哲学家和宗教人士

4. 传染病

5. 女权主义：英文和法文

6. 科学家：伽利略、达尔文、弗洛伊德、爱因斯坦（见图1）

7. 科学词汇：演化论、细胞、细菌、DNA

8. 食物：牛排、香肠、冰激凌、披萨、意大利面

9. 上帝

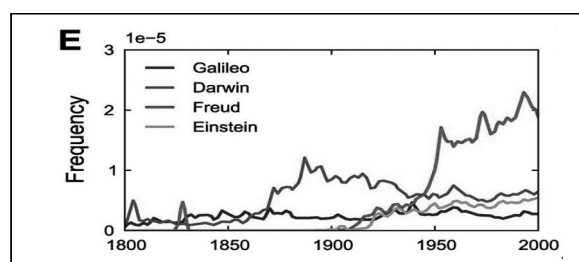


图1 数据图书实践研究的文化组：科学家^③

以米歇尔等人创立的“Culturomic”为基本文化分析单位的文化分组研究方法产生一定的学术影响，开创了研究的思路，同时也引发了文化领域的讨论。

3 “Sci-Culturomic”与科学文化研究

关于科学文化理论的阐述和研究在中国不是很多。在西方国家不仅有定义和理论的探索，而且已经有很多社会学、传播学、语言学以及公众对科学和技术的理解和态度方面的研究。

③图片来源：Jean-Baptiste Michel et al. “Quantitative Analysis of Culture Using Millions of Digital Books”, Science, 14 January, 2011, Vol. 331。

加拿大科学传播学者伯纳德·席勒 (Bernard Schiele) 主持的一次关于科学文化的讨论, 文章结集成书《当科学成为文化》(When Science Becomes Culture)。书中关于科学文化的观点认为: “今天, 特别是从 20 世纪 70 年代初以来, ‘科学和技能文化’ (scientific and technical culture), 以及最近更多谈论的 ‘科学和技术文化’ (scientific and technological culture) 已经开始取代长期以来一直占统治地位的科学普及 (scientific popularization)。”^④ 书中关于科学文化的意义认为: “科学与技术文化是知识、技能和态度的综合概念。知识需要通过学校提供的传播知识的活动获得。这是在工作中、家庭生活中和日常生活中需要的技能, 也是我们成长和创新所须臾不可离开的。科学技术文化同时也是在面临发展时我们的观点阐述和采取关键行动时的态度, 也是我们在采取行动和作出选择时的态度宣示。”^⑤

还有学者认为科学文化是: “个人和社会体味和鉴赏科学技术的多方位的表达方式。”^⑥ 可见, 科学文化是在多维文化体系中存在的一种以科学技术的信息为主要传播内容的一种文化。其存在的形式与影响随着科学技术的发展与传播而日益变得多元化。而多元化形式随着大数据时代变得愈加复杂。这种复杂性已经远远超过在传统研究阶段那种仅仅依靠一种指标和监测维度就能够得到清晰的解释, 得出基本规律, 同时做出预测。

科学文化的量化研究在大数据时代前基本依赖的是社会学方法和文献学方法。中国科学情报专家梁立明和武夷山等人早在 2006 年就用科学计量学方法进行了西方国家科学情报和中国科学情报的研究。他们对瑞典情报学家、图书馆学和信息科学教授奥尔·佩尔森 (Olle Persson) 用可视化映射方法展示出的 1986—1996 年 10 年间的世界情报学研究领域最具影响力的 57 位科学家的共引关系。佩尔

森教授的数据取自美国费城科学情报研究所, 他从 11 种图书馆与情报学科学期刊检索出 7 001 篇论文, 计量出这些文献被引用情况, 并以被引用 30 次以上为标准, 遴选出 57 位最有影响力的情报学家。采用的研究方法为文献计量学和数学模型方法、科学评价以及信息检索等。这是中国比较早的用量化分析方法进行科学文献和情报分析的成果^⑦。

大数据的发展主要是因为无线传感器的快速普及。2001 年, 著名的高德纳咨询公司 (Gartner) 在研究报告中指出, 数据的爆炸是“三维”的: 1. 同一类型的数据量在快速增大; 2. 数据增长的速度在加快; 3. 数据的多样性, 即数据的来源和新的数据的种类在不断增加。新型的数据化使得科学技术的信息呈现出多样化。人们对于科学技术的理解受到来自各种文化形态的影响。这些影响来自国际的、国内科学技术界、各种社会媒体、电子数据渠道以及各种文化立场的声音的影响。公众对于科学技术的理解受到多种因素的影响, 他们对于科学发现和新的技术应用的态度在不断变化中^⑧。

强子对撞机是从数以百亿计的数据中, 经过几年的分析分离出希格斯·波色子。生物研究中的遗传密码是从 60 多亿碱基对中隐藏密码中通过计算机分析出来的。科学文化因素的分析必须从基本的文化基因中分析。就像迈克尔等人运用生物学概念而创立的“Culturomics” (文化组学), 同时运用这个基本的数据粒子进行纵向文化分析, 在科学文化研究中, 可以用“Sci-Culturomics” (科学文化组学) 进行科学文化研究。

科学文化分组研究应该通过纵向和横向的数据中提取与研究目标一致的科学文化基本单位进行观测, 从而分析出科学文化的影响。

科学文化分组应该具备以下基本条件:

(1) 能够准确代表某一科学领域概念的字、词或者词组, 这些字、词或者词组可以是中文

④ “科学普及”一词究其根源产生于中国民国时期, 也就是 20 世纪 30 年代的“科学社会化”和“社会科学化”讨论阶段。英文中对应的表述国内学者一直采用“scientific popularization”或者“science popularization”, “popularization of science”。鉴于中外文化表述中的差异, 本文中的英文注示是为了解释翻译的原词。

⑤ “3D Data Management: Controlling Data Volume, Velocity and Variety”, Laney, Douglas, Feb 2001. <http://wenku.baidu.com/view/c4ddd5400b4c2e3f5627633d.html>.

的，也可以是外来语翻译的，还可以是外来语惯用的。比如“核能”、“霾”、“DNA”、“PM2.5”、“GDP”、“PX”（对二甲苯），等等^⑥。

（2）这些字、词或者词组是不可分的。即分离后意义发生变化。

（3）选择字、词或者词组应该进行误读或者误解实验，比如“干细胞”中的“干”不能理解为与“湿”对应的“干”，“转基因”中的“转”不能理解为“转圈”的“转”，“纳米”并不是一种新的粮食的种类，等等^⑦。

（4）对于某些容易与传统文化混淆的科学项目命名应该恰当地配以其他相关字词，增强其意义的完整性。比如，中国航天“嫦娥”与“嫦娥奔月”传说；中国航天“神舟X号”中的“神舟”与中国有关神话故事等。

用科学文化组学方法研究科学文化的演化需要通过确定的科学文化“粒子”在出版物中出现的频率而观察其变化。文化分组研究的是文化“基因”在历史演化中的趋势，研究的历史阶段可以以出版历史阶段进行研究，米歇尔等人的研究是从1800—2010年为历史阶段进行研究和观察的。这是因为虽然印刷术早在15世纪就出现并影响了欧洲的文明进程，但是，当时主要的书是《圣经》。科学文化研究应该与特定国家的文化发展特征紧密连接，才能制定观察时间。

中国的近代科学文化研究大约应该从明晚期传教士的活动开始，明末清初一清晚期（主要是江南制造局翻译馆）—1897年商务印书馆的出版物—中国学者早期科学刊物（如1890年的《亚泉杂志》）—晚清时期—国民时期（1915年的中国科学社的书籍以及《科学》杂志）和“中国科学化运动”出版物—新中国成立—改革开放前—改革开放后，直至目前阶段。科学文化的“基因”在中国文化历史中引进和演化可以为我们观察中国科学文化的趋势描绘出精确的时序进程。

中国在改革开放以后，以科学技术为主要驱动力的经济发展显示出与历史任何阶段都

截然不同的科学文化模式。教育的发展和科学技术教育的正规教育与非正规教育的发展，以及国际科学文化的交融，使得中国的传统文化不同于历史任何阶段的科学，尤其是技术应用飞速发展，对公众的生活和工作带来了巨大的，甚至是突然的变化。中国的传统文化与科学文化不可避免地产生冲突。尤其是技术应用对环境、能源、伦理、道德以及科学家行为和利益的冲突，通过发达的电子媒介的迅速传播，公众文化和科学文化之间的矛盾导致了公众舆论对科学技术发展的影响。在现阶段，对科学文化的影响的研究可以使得数据精确化，同时，社会学在大数据时代的研究会产生突破进展，脱离简单和传统的方式（问卷和质化调

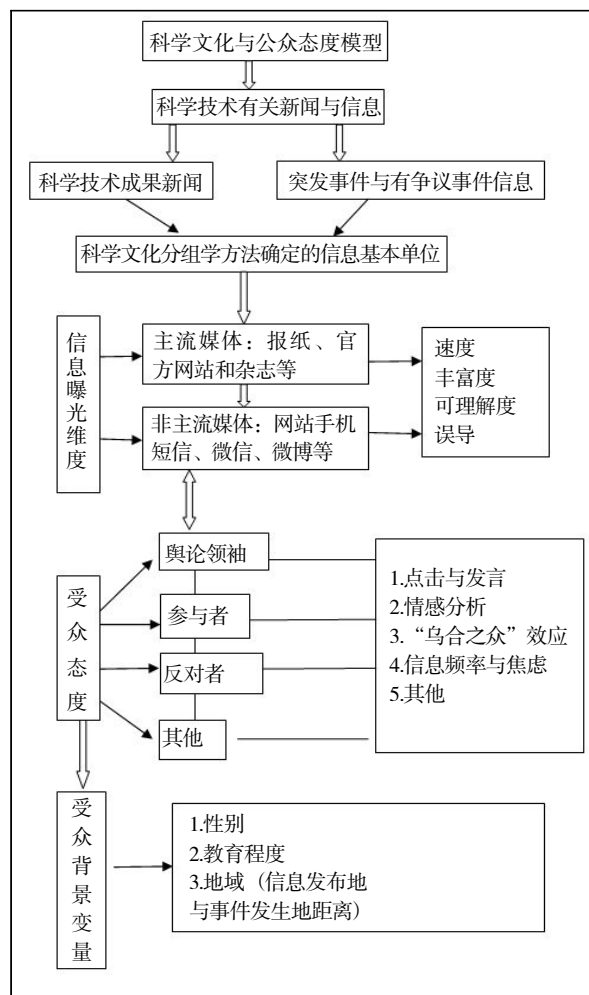


图2 大数据科学文化与公众态度研究模型

⑥诸多的外来语已经融入中国文化，成为媒介和人际传播的惯用语。

⑦在以前质化调查中已经发现这种误读和错误理解。

查),通过电子数据的观察,可以最大限度地避免系统误差,不仅能够观察到实时状况,而且大数据之间的相关性,可以为我们描绘出公众对科学技术的理解所受到的影响因素(详见图2)。

社会学在大数据时代的研究具有生物学研究相近的属性。这是因为社会是由人构成的。这种差异与工业化的数据和商业化的数据有不同之处。理查德·道金斯(Richard Dawkins)^⑧在其《自私的基因》(*The Selfish Gene*)^⑨一书中谈到:“在自然界中,最明显的方式是由基因关系——亲属而集合。大多数动物喜欢同自己的兄弟姐妹与表亲们,而不是种族中其他成员居住在一起。这并不一定是出于选择,而是自动跟随种群中的‘粘性’。这里的‘粘性’指的是任何使个体持续居住于出生地的趋势。”^⑩一个民族的种群关系形成了其对发生的突发事件的影响,甚至是重大影响。人际传播即使在今日电子时代仍然在中国的文化群体中具有十分重要的影响。人们的手机、微信,甚至电子信件的传播首先是在熟人或者朋友群体中传播,然后再通过微博和博客或者网络信息扩散到网友群体中。有研究者认为,社交网络会加剧“物以类聚,人以群分”效应,从而加剧了社交圈的“孤岛”效应。但是,目前比较有影响的一项研究通过对 Facebook 的人群采用随机跟踪的方法跟踪特定的一组网页地址,发现别人分享这个地址给我们,我们看到后转发的可能性(pfeed)比我们自己看到这个地址直接分享的可能性(pno feed)大5倍以上。强联系的放大效应为6,而弱联系的放大效应是9^⑪。

在不断加剧的信息扩散中,与科学技术有关的信息也逐步进入网络传播中。传播的走向与公众的生活密切相关的利益有关,同时,

也与商业的广告有关。与此同时,科学带来的益处与技术带来的环境问题以及与公众的生活密切相关的事件也成为公众最关注的信息。其中包括环境污染、核能利用、转基因食品、科学家行为、PX化工厂^⑫、传染病、药物、医院救治、地震等突发事件都与科学技术有关。这些事件在传播过程中携带的各种信息不仅快速传播,更重要的是在传播中变形,改变了公众对科学技术的看法,导致公众对科学技术的理解偏离,从而引发态度和行为的转变。

在传播中,大约由几种人组成:最接近媒体信息,善于归纳甚至加入自己的看法的舆论领袖(或者意见领袖);参与者,即对科学技术感兴趣,却又并不十分了解,但是愿意参与讨论,同时传播信息者;反对者,即在信息传播中积极表达自己的反对意见;而更多的是随声附和者。他们受到自己的专业知识、认识方法、情感、切身体验或者间接听说,在接到第一信息后,传播行为会有所不同。舆论领袖一般速度会最快,发表意见或者议论最多。而更多的是在“乌合之众”效应中非理性的态度表达。无论如何,在海量的信息中,公众在对真相未揭示之前,必定会产生怀疑、疑惑、反对,甚至更为极端的言论和行动。大数据的分析会使研究者基本看到发展趋势以及相互之间的关联性影响。

大数据时代的公众对科学技术的理解和态度的研究与传统的社会学的随机样本调查不同。虽然大数据研究也是采用样本,但却是大数据的样本在大数据海洋中的样本,因此,全数据模式:样本=总体。传统社会学随机抽样从主观上会遗漏可能具有重大价值的样本群,而导致数据结果的偏移(bias)。比如,中国是一个有55个少数民族的国家,而这些少数民

⑧英国皇家学会会员。在2005年英国《前景》杂志同美国《外交》杂志网络调查中,他被评为在世的全世界100名最有影响的公共知识分子之一。

⑨在英国《新科学家》(*New Scientist*)杂志于2012年进行的10本“迄今为止最有影响的大众科学书”公众评选中,《自私的基因》名列其中。<http://www.newscientist.com/blogs/culturelab/2012/10/top-10-most-influential-popular-science-books.html>。

⑩<http://www.scribd.com/doc/78445521/Role-of-Social-Networks-in-Information-Diffusion>。

⑪最有影响的是在2009到2011年间的厦门PX事件引发国际关注。最近漳州PX工厂起火事件引发公众再一次对化工企业的担忧。

族占据总人口的9%的比例^⑫，更重要的是，少数民族分布范围很广，但是，随着信息化的发展，大多数少数民族中的“一神论”宗教信仰对科学技术的看法可能与人口占绝对总量的汉族有很大不同。在大数据中，这些在社会学调查中最容易在抽样中遗漏的样本，在大数据的研究中，都会在数据总量中得到统计和分析。与此同时，地理位置也具有重要的意义。地理位置主要体现在发布信息者距离科学技术事件中心的距离对理解和态度具有十分重要的影响。“大数据是指不用随机分析法这样的捷径，而采用所有数据的方法。”^⑬所有数据的分析使得公众对科学技术的态度的信息表述都在观察范围之内。其结果增加了判断的准确性。

4 可能性

大数据时代的研究主要依据的是大数据的巨量、飞速与数据多样性。在2015年的“政府工作报告”中，出现了“互联网+”的概念：“推动移动互联网、云计算、大数据、物联网等与现代制造业相结合，促进电子商务、工业互联网和互联网金融健康发展。”^⑭2014年，中国的网民达到6.49亿，占人口总数的47.8%；手机网民数量达到5.57亿，达到85.8%。互联网已经将金融、汽车、教育、泛娱乐以及智能软件融合为一体。大数据与云计算成为连接一切和智能化的根基^⑮。

新一代的信息技术发展催生了创新2.0，而创新2.0又反过来促进了新一代信息技术的新形态的形成与发展，这种新形态的主要特征就是物联网、云计算、社会计算和大数据，从而进一步推动了“互联网+”全新概念以及打造“互联网+”生态链的诞生。

目前，在科学文化大数据研究中，大数据的丰富度还有待于提升。我国目前还没有像Google这样巨型的网络将大量的图书数据化。

目前，腾讯、百度和阿里巴巴是中国最大的网络集团，还没有实现图书数据化。图书数据化仍然处于分散的状态。我国大型的省市级图书馆和部分大学图书馆实现了图书数据化，但是，数据化的量仍然不是很大。与此同时，经济利益导致数据相对封闭。

但是，我国目前的手机信息和微信数据量巨大。微博与博客数据量以及网络数据量也十分庞大。这些为大数据时代公众对科学技术的理解和态度的研究提供了可能性。

我国目前已经有一批电子数据公司，具备了很强的实力。这些公司完全能够承担起这种科学文化数据研究。

因此，目前在中国进行科学文化（甚至其他文化）研究基本具备了条件，尽管这个条件仍然有待于提升。在有限的数据量中进行有条件的研究，至少能够为今后的研究创造条件。

参考文献

- [1] 李大光. 大数据时代的公众理解科学[J]. 科普研究, 2015 (2): 5-11.
- [2] Jean-Baptiste Michel et al. Quantitative analysis of culture using millions of digital books [J]. Science, 14 January, 2011, Vol. 331.
- [3] Bernard Schiele et al. "Introduction" When science becomes culture[M]. University of Ottawa Press, 2003.
- [4] Godin and Gingras. What is scientific and technological culture and how is it measured? A Multidimensional Model[J]. Public Understanding of Science, January 2000 (9): 43-58.
- [5] 梁立明, 武夷山. 科学计量学: 理论探索与案例分析[M]. 北京: 科学出版社, 2006.
- [6] 理查德·道金斯. 自私的基因[M]. 卢允中, 张岱云, 译. 北京: 中信出版社, 2012.
- [7] 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[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.

(编辑 张南茜)

⑫ 《中国统计年鉴2013》少数民族分布的主要地区及人口。http://wenku.baidu.com/link?url=xuwNnboGwByEn_rf1LK-eR1cOW DN8ftKDAKwbEbKVGZFgSmptRiWH-gMdhwA_N7qI_UZx1 E CJobsFMf5HdJIgI8mZvFud7Z9LVp18qXoIC。

⑬ 《2015年政府工作报告》(全文实录)。http://lianghui.people.com.cn/2015npc/n/2015/0305/c394298-26642056.html。

⑭ “中国互联网络信息中心”(CNNIC)发布的《第35次中国互联网发展状况统计报告》。www.gsdata.cn。http://www.cnnic.cn/hlwfzyj/hlwfzxx/qwfb/。