

DOI:10.12154/j.qbzlgz.2021.01.006

# 微博数据研究的伦理反思\*

陈 一<sup>1</sup> 李 斯<sup>2</sup> 孙异凡<sup>1</sup>(<sup>1</sup>武汉大学信息管理学院 湖北 430072; <sup>2</sup>北京大学信息管理系 北京 100871)

**摘要:** [目的/意义] 社交媒体数据已经成为目前学术研究中重要的数据来源。但学者在采集、利用社交媒体数据过程中往往忽略其中存在的伦理风险。本研究从伦理视角剖析社交媒体数据类研究,提醒研究者们更好地理解此类研究面临的研究方法和伦理方面的挑战。[方法/过程] 本文以新浪微博为研究对象,从四个角度对微博数据研究进行伦理反思,研究者缺乏伦理意识;公私数据界定不清;知情同意机制缺失;数据使用规范性有待加强。[结果/结论] 伦理规范即通过有效的规制,保证研究的所有环节都处于伦理的规范空间内,微博蕴含丰富的个人数据,并且微博数据研究的发展演化也对伦理规范有强大需求。本研究认为,应建立数据教育体系,提升“数据伦理”意识;研究者不能假设用户同意;探索科学的知情同意机制;鼓励跨领域合作,平衡创新与风险。

**关键词:** 社交媒体 微博 数据 学术研究 伦理

## Ethical Reflection on Microblog Data Research

Chen Yi<sup>1</sup> Li Si<sup>2</sup> Sun Yifan<sup>1</sup>(<sup>1</sup>School of Information Management, Wuhan University, Hubei, 4430072;<sup>2</sup>Department of Information Management, Peking University, Beijing, 100871)

**Abstract:** [Purpose/significance] Social media data has become an important data source in academic research. However, scholars often ignore the ethical risks in collecting and using social media data. This study analyzes the research on social media data from the ethical perspective, and reminds researchers to better understand the research methods and ethical challenges faced by such research. [Method/process] Taking Sina Microblog as the research object, this paper makes ethical reflection on microblog data research from four perspectives, researchers lack data ethics awareness; unclear definition of public and private data; microblog data research lacks scientific informed consent mechanism; the normative nature needs to be strengthened. [Result/conclusion] Ethical norms, that is, through effective regulation, ensure that all aspects of research are in the ethical space. Microblog contains a lot of personal data, and the development and evolution of microblog data research also has a strong demand for ethical norms. This study suggests that, we should establish the data education system to enhance the awareness of "data ethics", and the researchers cannot assume the consent of users. We should explore a scientific informed consent machine, encourage cross field cooperation, and balance the innovation and risks.

**Keywords:** social media microblog data academic research ethics

## 1 引言

社交媒体的出现和迅猛发展改变了许多人、社区、媒体,渗透率已经处于相当高的水平。而微博因为其

组织的交互方式,使人类使用互联网的方式发生了深刻的变革。微博和微信作为目前我国相对成熟的社交媒体,渗透率已经处于相当高的水平。而微博因为其

\*本文系中国博士后科学基金资助项目“大数据应用中的伦理风险演化及其治理框架构建”(项目编号:2017M622526)的研究成果之一。

公开性,更多的成为表达和衡量用户行为和态度的关键渠道,营销人员和品牌也经常利用微博平台考察用户对产品的实时感受<sup>[1]</sup>。2010年,我国微博出现了井喷式发展,其中新浪微博一直处于领先地位,2018年第四季度,新浪微博月活跃用户4.62亿,微博日活跃用户量达2亿<sup>[2]</sup>。微博已经成为一种极有价值的社会资源<sup>[3]</sup>,成为人类社会中社会关系维系和信息传播的重要渠道和载体<sup>[4]</sup>。海量的用户行为数据使得学术研究发生了深刻的变革。研究人员开发和利用多种技术构建各种在线方法和工具,来探索用户在虚拟世界中的行为<sup>[5]</sup>,数据收集、取样、分析及整个研究过程也发生了巨大的变化。国内外各领域的科研人员开展了大量与微博相关的理论和实践研究工作,利用微博数据从各个角度了解用户行为、情绪,数据用途,对社会及文化产生的影响<sup>[6-7]</sup>。本研究将采集微博上的用户数据或用户所发布的微博作为数据来源的研究称为微博数据研究。

网络数据研究带来的新型定量和统计分析改变了传统的社会研究,已经成为当下最普遍的社会科学研究形式之一<sup>[8]</sup>,数据革命也已经成为全球发展议程的重要组成部分<sup>[9]</sup>。这类数据研究所特有的优势很容易使研究者追求便捷和廉价而忽视在采集、分析,甚至使用数据进行预测时所伴随的伦理问题。在2018年4月10日,Facebook创始人扎克伯克就Facebook泄露用户隐私数据事件接受了美国国会44名参议员的尖锐拷问。在此之后,Facebook已经宣布限制第三方收集数据,其中也包括应用程序开发者可访问的大量信息,研究人员收集在线数据也将面临更严格的审查<sup>[10]</sup>。没有约束的科技是危险的<sup>[11]</sup>,而这类利用网络数据开展的研究所固有的缺陷也使研究的科学性和客观性受到质疑<sup>[12]</sup>。在利用微博数据进行学术探索的同时,学者们也应该关注到其中潜在的风险。

## 2 微博数据研究现状

笔者在CNKI数据中检索“篇名”或“关键词”中含有“微博”的文章,2008年,有学者开始关注到微博的发展,2009年随着国内微博的大规模兴起,研究也逐渐多样化,2011年开始,有关“微博”的研究陡增,达到3456篇,2012—2014年均保持在5000篇以上,近两年研究势头逐步平缓,但也均保持在2000篇左右。可见,微博及微博数据成为数据驱动的科学研究中重要数据来源,给研究者们提供了全新的研究思路,学者们利用社交媒体数据在各个领域展开探索,微博数据被用于事件跟踪、预测<sup>[13-14]</sup>、疾病预防<sup>[15]</sup>、企业产品推广<sup>[16-17]</sup>等。

国内外也有很多学者从宏观角度梳理了“微博研究”的研究进展。Ngai等<sup>[18]</sup>梳理了2002—2011年46篇有关社交媒体的研究,归纳了先前的研究维度和变量,以及所提出的概念模型和框架,在此基础上建立了因果链框架说明所采用的研究结构之间的相互关系。Williams等<sup>[19]</sup>对2007—2011年之间的Twitter研究进行了系统的识别和分类,研究发现大多数研究是关于信息传播以及用户研究。国内也有学者对微博研究中的技术方法和关键问题进行了总结,刘晓娟等<sup>[20]</sup>研究发现科研人员通常通过API获取数据,刘超等<sup>[21]</sup>从长期保存视角,归纳了微博信息采集的三个关键问题,即采集范围的确定、采集权利的获取和采集方法的选择。也有一些学者从宏观角度梳理了微博领域研究进展,刘志明<sup>[22]</sup>总结了使用微博数据对突发事件领域、经济领域、政治领域、文化领域进行预测的情况;周金元等<sup>[23]</sup>对我国微博研究的研究作者、主题进行了可视化分析,认为未来微博研究的趋势是技术突破、学科合作、实践导向和自律。严威等<sup>[24]</sup>从理论基础、研究方法、研究主题和研究层面四个方面对微博研究进行了归纳。

国外有一些学者意识到利用社交媒体数据进行研究存在风险,Moreno等<sup>[25]</sup>社交媒体网站成为越来越流行的研究工具,这在给研究人员提供新的研究机会的同时,也给审查这些研究方案的机构审查委员会(IRB)提出了新的挑战。而目前,IRB针对社交媒体研究几乎没有指导意见。作者认为,机密性是社交媒体研究的关键,保护参与者身份至关重要。Fiesler和Proferes<sup>[26]</sup>也表示公共社交数据的可用性会带来道德挑战,研究者应该从用户出发探索用户在接受研究时的认知感受。研究发现,很少有用户知道他们的公开推文可以被研究人员使用,而且大多数用户认为研究人员不应该在未经同意的情况下使用推文。Swirsky等<sup>[27]</sup>表示使用社交媒体数据可能会带来意想不到的伦理争议,特别是我们在医疗研究中使用社交媒体时,必须深思熟虑,以确保我们处在伦理正轨上。

而目前我国学者的相关研究中,未见到有学者关注到社交媒体数据研究存在伦理隐忧。高质量期刊论文能够真实反映学术发展,为进一步验证,本研究组以CSSCI中收录的图书馆、情报与文献学期刊上的文献为数据来源,根据当年CSSCI来源期刊目录,在数据库中检索2009—2018年标题或关键词中带有“微博”的所有文献,共计726条。通过逐篇阅读,剔除文中没有利用微博数据进行研究的文献,如微博介绍性研究、单纯的

理论策略型研究,最终获得470篇基于微博数据的研究文献。其中,仅有一篇在文中有直接关于伦理的相关说明,“引入基于多主体建模的计算机仿真方法建立模型,从而避免实验可能带来的道德问题”<sup>[28]</sup>。目前我国研究中更多的是关注微博数据所带来的商业和学术价值,几乎没有学者意识到此类研究中潜在的伦理问题。基于此,本文拟从伦理视角,重新审视微博数据类研究。

### 3 微博数据研究的伦理反思

#### 3.1 研究者缺乏伦理意识

网络数据研究带来的新型定量和统计分析改变了传统的社会研究,已经成为当下最普遍的社会科学研究形式之一<sup>[8]</sup>,而这些数据的来源——社交媒体用户,可能永远不知道自己已经成为研究对象了,更不用说对研究的知情同意了。即便对数据进行清洗,大数据关联分析还是存在被重新识别的风险<sup>[29]</sup>。国外很多学者认为这类研究不够尊重社交媒体用户<sup>[30-31]</sup>,而也有许多研究者没有意识到这一点。Zimmer等<sup>[3]</sup>对380篇利用Twitter的数据的学术论文进行统计,发现其中仅有4%提到了与研究设计和数据收集方法有关的伦理问题或者相关因素。其中有5篇文献承认了在研究中存在伦理风险,并阐述了他们在研究中为确保用户的匿名性而采取的举措。

上文中也已提到,在470篇基于微博数据的研究中,仅有一篇文献在文中有直接关于伦理的相关说明。在调查中发现,有不少研究直接截取了用户的微博信息,包括头像、个人简介、用户名及微博内容;有研究直接截取了新浪爱问医生中用户医疗咨询的内容;也有文章提到会对博文截图保存。但也有文章在实际分析过程中对样本进行了匿名化处理,用代码表示相关博文;对抓取后的数据进行“虚构和改编”;也有研究同时采集了机构类微博账户和个人微博账户,直接使用了机构类微博的截图,而对个人账户的个人信息进行了马赛克处理。

研究者的科研伦理意识主要包括研究者们对研究活动本身伦理问题的认识以及以此为基础形成的社会责任感以及价值取向<sup>[32]</sup>。可以看出,目前我国学者已经有少数学者在实际研究过程中采取了相应隐私保护措施,但对于利用网络数据研究这一研究模式中存在的伦理问题还没有形成明确的认知。

#### 3.2 公私数据界定不清

数据在什么情况下属于公共或是私有领域通常是存在争议的问题,一些学者认为这是造成网络数据研究中存在伦理争议的核心问题<sup>[33]</sup>。从严格的法律角度

来看,只要《著作权法》不保护的文件都应该归类于“公共领域”,很多学者也认为网络属于公共区域,在这一区域中从事研究是公平的<sup>[34]</sup>。那么微博数据是否受到《著作权法》保护?有人认为微博信息达不到“作品”所应具有的科学性、艺术性高度,而有专家认为微博具有独创性应受到著作权保护,这主要是因为:(1)字数与作品没有直接关系;(2)内容的独创性才是判断作品的标准;(3)《著作权法》第五条规定时事新闻不适用于本法,故微博中时事新闻类信息不是作品,不受《著作权法》保护。简言之,除时事新闻类,具有独创性内容的微博应作为作品对待,相应的著作权人拥有包括复制权在内的所有权利<sup>[35]</sup>。也有学者表示“一个清晰的、分门别类的、量化的独创性认定指引是不存在的”,不能简单否定“日常对话式”博文的独创性<sup>[36]</sup>。我国《著作权法》中虽然规定了一些著作权例外情况,但有明确的限制规范,其中第二十二条说明“为学校课堂教学或者科学研究,翻译或少量复制已经发表的作品,供教学或者科研人员使用”属于例外情况,但同时也强调了“不得出版发行”。因此,无限制的采集微博数据用于学术研究并没有明确的法律支撑,更不能简单认为微博数据属于公共领域而随意采集。

#### 3.3 知情同意机制缺失

新浪微博开放自己的接口给第三方应用程序公司,使其可以调取平台数据和用户信息,进一步为用户提供更多有偿或者无偿服务,但同时新浪微博对于开发者有以下规定,“开发者应用或服务需要收集用户数据的,必须事先获得用户的同意……开发者应当告知用户相关数据收集的目的、范围及使用方式,以保障用户的知情权”“开发者应当就应用向微博平台用户提供隐私保护政策,告知微博平台用户该应用收集哪些用户数据,如何使用,是否会将用户数据传播或提交给他人等,以保障用户的知情权及选择权”<sup>[37]</sup>。这类授权是基于“开发的应用接入”“向用户提供各种服务”这一目的,而对于学术研究中个人或组织对于微博数据的采集行为,并没有明确的规定。在所调研的文献中,没有文献表示在采集数据时有任何的知情同意程序,研究者多运用成熟的爬虫工具或自行设计软件对数据进行爬取,很多文献采集了微博用户的账户信息。而用户的教育信息、职业信息以及手机号信息等是在高级权限下才能调取的信息,研究者的数据采集行为是否合规目前还没有明确的判定。

在“新浪微博起诉脉脉抓取使用微博用户信息案”中,脉脉公司抓取微博数据用于商业化,法院判定其构成侵权行为。法院在判决中明确将用户的权益保护作

为案件裁判的一个考虑因素,并提出用户授权+平台授权+用户授权的三重授权原则<sup>[38]</sup>,即开放平台直接收集、使用用户数据需获得用户授权、第三方开发者通过开放平台 Open API 接口间接获得用户数据,需获得用户授权和平台方授权,三重授权需同时满足,缺少任何一方都违反“三重授权原则”。基于学术目的的数据采集并没有建立起成熟的知情同意机制,而且对于个人研究者来说,获取所有用户的同意通常也是不切实际的<sup>[39]</sup>,急需确立适用于学术研究数据采集的知情同意机制。

### 3.4 数据使用规范性有待加强

基于数据的研究不同于传统的逻辑推理研究,它是对一定数量的数据做统计性的搜索、比较、聚类、分析,对数据获取方式、数量及分析方式的描述不但是研究的重要组成部分,更确保了研究的可回溯性和可验证,也有助于之后研究在此基础上的科研创新。规范都是来自过去经验的总结与提升,研究形式的系统规范性能在一定程度上保障研究的伦理规范。

但根据研究组对 470 篇文献的统计发现,有 28 篇文献没有说明微博数据或账户采集数量,有 46 篇文献没有数据采集方式的相关说明,有 140 篇文献有明确的研究方法说明,有 58 篇文献有单独一章描述数据,有 116 篇文献在实证部分单独说明了数据获取方式及数量,其余三分之一的文献仅在正文中简单提及或没有相关描述。我国也有学者曾表示,从公开发表的论文看,一些作者仍然缺乏对研究方法进行选择、说明和使用的意识<sup>[40]</sup>。Miles 和 Huberman<sup>[41]</sup>强调,定性数据分析过程应包括数据提炼、数据显示和数据推导,过程应被记录,也要说明数据收集和分析所运用的程序,这样其他人原则上遵循同样的程序可以得到相同的结论,或者更有可能追随下去并评价内在的逻辑<sup>[40]</sup>。而目前我国数据类研究在数据使用规范性上还亟待加强。

## 4 讨论及建议

社交媒体账户中包含着大量的个人信息。在 Twitter 中,40%~50%的推文中包含有关作者的信息<sup>[42-43]</sup>,其中还可能包含其联系人的相关数据,如个人身份信息、位置数据、健康信息以及图像等<sup>[44]</sup>。而我国新浪微博用户基本注册信息真实程度最高为 100%,最低为 17.3%。信息发布是微博用户信息泄露的关键路径,将一段时间内某用户所发布的图片、信息进行整合,掌握该用户较全面的个人基本信息甚至个人私密信息并不困难。图像信息是微博用户信息泄露的重要方式,图像的直观真实可靠将用户的真实世界呈现在虚拟空间

中<sup>[45]</sup>。数据价值多来自对数据集的二次利用所产生出的多样化的预测<sup>[46]</sup>,研究者们随意大量采集用户数据,运用数据及一些机器分析方法使得海量相关关系过度分析可能会大面积披露隐私数据。

从微博数据研究的发展来看,通过对 470 篇文章逐篇阅读统计发现,近年来,微博数据研究的研究方式、研究主题、研究层次都在不断演进变化,研究者们数据采集、分析能力也在不断提升。在图书情报领域,早期的微博数据研究多采用人工网页检索、浏览获取数据的研究较多,“图书馆利用微博”及“政务微博”研究占有比例较高,微博被当作一种新的工作方式和服务手段,研究人员多运用一些评价指标来评价图书馆微博及政务微博,评价指标较为类似且内容缺乏深度,多为单纯的计数统计。而近年来,影响力研究、预测、相关性研究及网络结构研究所占比例越来越大,研究者更多的利用爬虫等工具大规模获取数据,也不再单纯描述微博内容,进行数量统计,而是更多的利用算法,分析数据关联,挖掘数据内涵。数据分析工具与开发环境的不断革新,带来更大学术和应用价值的同时,也可能带来更严峻的伦理问题。

微博数据自身所蕴含的丰富的个人数据,以及微博数据研究的发展演化使其对伦理有着强大的需求。伦理规范即通过有效的规制,保证研究的所有环节都处于伦理的规范空间内,这也保障了研究能够持续向纵深推进。

### 4.1 建立数字教育体系,提升“数据伦理”意识

人是技术使用的主体,也是引发伦理问题的关键<sup>[47]</sup>,有专家表示:大多数违规都是人为的,而不是技术<sup>[48]</sup>,人才建设是关键。目前,很多国家都非常重视对数字教育的投入,2018年,英国额外投入 4.06 亿英镑,用于数学、数字化和技术教育<sup>[49]</sup>。而我国的教育体制缺少数据使用规范性的系统训练,导致我国科研人员对研究伦理的认知和知识储备严重不足,这极有可能造成研究收益归个人,而风险留给社会,这将带来极大的科技风险<sup>[50]</sup>。我国应借鉴有益经验,加大对数据教育的投入,针对不同的对象设计不同的课程体系。义务教育阶段,以科普性教育为主,使下一代掌握数据使用的基本常识,树立正确的“数据伦理”观念。大学阶段,设置系统的课程体系和培养计划,广泛开设数据利用课程,并培养专业型数据人才。社会教育也应与学校教育联动,发挥图书馆的社会教育职能,举办专题讲座普及数据利用常识,指导、规范数据传播、利用行为,同时也要提高社会成员对数字社会的接纳程度,提升其数据素养。

## 4.2 研究者不能假设用户“同意”

《贝尔蒙特报告》<sup>[51]</sup>提出了人体受试者研究的三大伦理准则,其中“尊重”排在首位。用户在网络平台上发表观点与在报纸上或公开会议上发表意见之间存在重要的“心理差异”,不能总被假设是为了寻求“公众可见”,有些用户认为一些对话、评论、问答只是存在于私人之间<sup>[52]</sup>。实际上,很少有互联网用户阅读网站的隐私条款,也鲜有能够完全理解<sup>[26]</sup>。本研究组在另一项研究中调查了用户对于微博数据类研究的态度,有超过81%的受访者认为“研究者们不可以未经许可使用微博数据”。有学者未经用户允许利用Facebook用户数据进行“情绪感染”实验,虽然Facebook的研究涉及直接实验而不是收集现有数据,但是也引发了公众的强烈抗议<sup>[53-54]</sup>。除此之外,有学者认为用户同意公开并不意味着是允许它们被持续系统地跟踪、获取、挖掘和存档,可以设想一下,如果用户删除了自己账户中的数据,那研究者应该如何处理之前采集的数据?<sup>[55]</sup>国外一项研究显示,有32.5%的被调查者对于自己删除的社交媒体数据被用于研究感到不舒服(somewhat uncomfortable),有31.3%的用户感到非常不舒服(very uncomfortable)<sup>[26]</sup>。

学术研究应该是“充分自由、完全知情以及持续同意”<sup>[56]</sup>,研究者在采集、应用用户数据过程中,应不断征求研究参与者的参与意愿,以确保该参与行为是一种有意义、完全知情且持续性的自主选择。Fiesler等<sup>[26]</sup>研究表示,“知情”和“同意”可被理解成两种含义,对很多用户来说,“知情”对他们来说就足够了。对于用户来说,研究者征求用户的“同意”或“许可”体现了对用户的基本尊重,大多用户真正介意的是自己未被告知而数据被采集。总之,研究者不能忽视用户的伦理需求,研究者要尊重数据自治,维护个人和社区的隐私和尊严<sup>[57]</sup>。

## 4.3 探索科学的知情同意机制

目前,网络数据研究并没有建立起成熟的知情同意机制,甚至连数据采集之前是否有必要获取参与者的知情同意都存在争议。一方面是因为数据权属不清晰,另一方面是知情同意程序工作量巨大,对于个人研究者实现起来有困难。但是,客观上,微博数据包含了大量个人数据,通过关联分析也存在大面积披露隐私的风险;主观上,用户也有极强的伦理需求。因此,建立科学的知情同意机制规范数据采集过程,并且实现真正的匿名,规避其中存在的伦理风险很有必要。有学者建议,如果研究人员没有事先获得许可,可以选择事后通知相关用户<sup>[26]</sup>,用户也可持续性地追踪研究进展,同时享有可在任意环节选择退出的权利。

本研究还倡导建立数据使用者的持续性审核责任机制。除了在数据申请阶段对数据使用者的资质、使用目的等进行详细的审核外,在进程中,也需定期就数据利用情况进行说明,特别是较申请时数据利用模式有变更的,需要提交申请并详细说明。要明确权责关系,数据使用者必须对授权数据的合理使用和存储负责,一旦发生违法事件,需承担全部责任,并禁止再次利用数据。

## 4.4 鼓励跨领域合作,平衡创新与风险

虽然微博数据研究存在一定的伦理争议,但是我们并不是建议平台禁止研究人员收集数据来解决这一困境。完全不允许在研究中使用公共数据和没有任何伦理考量地使用公共数据可能同样会带来糟糕的后果。

李国杰院士<sup>[58]</sup>表示,“复杂网络分析应该是数据科学的重要基石”。网络大数据(如微博数据)一般多源异构交互性、时效性、社会性、突发性和高噪声等,非结构化数据多,而且数据的时效性强,大量数据都是随机动态产生,网络上许多数据都是重复的或没有价值的,根据网络数据做经济形势、安全形势和社会群体事件的预测相较于实验数据则更加困难。我国学者数据采集、分析能力不断提升,但较之国际水平还有一定差距。Zimmer等<sup>[13]</sup>研究发现,利用Twitter数据进行研究的文献绝大多数的推文采集量在10万条以上。而在本研究组统计的470篇文献中,微博数据采集量在10万条以上的仅有63篇,而且也有研究认为我国早期微博研究存在技术局限<sup>[25]</sup>。

不止在图书情报领域,我国的数据研究应鼓励计算机算法研究人员与各个领域的科研人员密切合作,在分析工具和开发环境上创新,实现跨领域的数据分析,揭示数据背后的复杂关系网络,加速形成真正的知识和智能,产生更大的学术和应用价值<sup>[59]</sup>。提升科学采集、处理、分析数据的能力,积极联合确立研究规范,平衡创新与风险。

## 5 结语

社交媒体的发展不但改变了在线沟通的形式,同时也影响了学术领域数据收集、分析、取样的研究过程。越来越多的学者利用数据分析技术挖掘数据内涵,如上文统计,有更多的学者利用微博数据进行预测和关联性分析,但也有学者认为微博是虚假信息扩散平台<sup>[60]</sup>,甚至利用数据预测存在“结果预判挑战自由、隐私披露挑战尊严、信息垄断挑战公平、固化标签挑战争议”<sup>[61]</sup>。本从伦理视角对此类研究进行反思。作为研

究群体,我们应更好的理解网络数据类研究面临的研究方法和伦理方面的挑战。随着网络数据类研究的持续推进,数据采集规模进一步扩大,分析手段也更加多样化、智能化,数据研究则更应该关注数据的科学性,有效性和规范性,重视网络数据研究中的伦理规范。

### 参考文献

- [1] Jansen B J , Zhang M , Sobel K , et al. Twitter power: Tweets as electronic word of mouth[J]. *Journal of the American Society for Information Science and Technology*, 2010,60(11):2169-2188.
- [2] 2018微博用户发展报告[EB/OL].[2019-05-18].<http://data.weibo.com/report/reportDetail?id=433>.
- [3] Zimmer M, Proferes N J. A topology of Twitter research: disciplines, methods, and ethics[J]. *Aslib Journal of Information Management*, 2014, 66(3):250-261.
- [4] 丁兆云,贾焱,周斌. 微博数据挖掘研究综述[J]. *计算机研究与发展*, 2014, 51(4):691-706.
- [5] James N, Busher H. Ethical issues in online research[J]. *Educational Research & Evaluation*, 2015,21(2):89-94.
- [6] 王清华,朱岩,闻中. 新浪微博用户满意度对使用行为的影响研究[J]. *中国软科学*,2013(7):184-192.
- [7] Choi S, Han W P. Using social media data to explore communication processes within South Korean online innovation communities[J]. *Scientometrics*, 2012, 90(1):43-56.
- [8] Koene A, Perez E, Carter C J, et al. Research ethics and public trust, preconditions for continued growth of Internet mediated research: public confidence in Internet mediate research[C]// *International Conference on Information Systems Security and Privacy*. IEEE, 2016:163-168.
- [9] A World that Counts: Mobilising the Data Revolution for Sustainable Development[EB/OL].[2019-07-20]. <https://repositorio.cepal.org/bitstream/handle/11362/40319/AWorldThatCounts.pdf;jsessionid=760A6B527959A95E9F001701CB210E1A?sequence=1>.
- [10] Nature. Cambridge Analytica Controversy must Spur Researchers to Update Data Ethics[EB/OL].[2018-05-10]. <https://www.nature.com/articles/d41586-018-03856-4>.
- [11] 李侠. 科技伦理:没有约束的科技是危险的[N]. *光明日报*, 2015-07-31(10).
- [12] 陈一. 网络研究中的数据伦理问题研究进展[J]. *信息资源管理学报*,2019(2):58-65.
- [13] 吴小兰,章成志. 基于突发事件特征网络的用户社区发现与社区主题演化研究——以新浪微博H7N9事件为例[J]. *情报理论与实践*, 2017, 40(5):94-98.
- [14] 朱廷劭,白朔天,李昂. 基于社交媒体大数据的群体事件风险预警[C]// *心理学与创新能力提升——第十六届全国心理学学术会议论文集*, 2013.
- [15] 杨桂丽,陈镭,张晓红,等. 微博在疾病预防控制工作中的运用探讨[J]. *中国健康教育*, 2014, 30(9):856-857.
- [16] 周凯,徐理文. 基于5T理论视角下的企业微博营销策略及应用分析——以欧莱雅的微博营销为个案研究[J]. *图书与情报*, 2012(5):120-127.
- [17] 张伟,李晓丹,郭立宏. 不同微博营销渠道对产品销量的影响研究:品牌自有媒体VS第三方媒体的路径对比[J]. *南开管理评论*, 2018, 21(2):43-51.
- [18] Ngai E W T , Tao S S C , Moon K K L . Social media research: theories, constructs, and conceptual frameworks[J]. *International Journal of Information Management*, 2015, 35(1):33-44.
- [19] Williams S A , Terras M M , Warwick C . What do people study when they study Twitter? Classifying Twitter related academic papers[J]. *Journal of Documentation*, 2013, 69(3):384-410.
- [20] 刘晓娟,尤斌,张爱芸. 基于微博数据的应用研究综述[J]. *情报杂志*,2013,32(9):39-45.
- [21] 刘超,郑建程. 长期保存视角下的中文微博信息采集关键问题探讨[J]. *图书情报工作*,2015,59(3):134-139.
- [22] 刘志明. 使用微博数据进行预测的研究综述[J]. *科技管理研究*, 2014(13):165-169.
- [23] 周金元,张莎莎,刘桂锋,等. 国内微博研究综述[J]. *情报杂志*, 2013(9):46-51.
- [24] 严威,黄京华,张瑾. 微博研究回顾:信息、服务和网络[J]. *科研管理*, 2017(4):123-131.
- [25] Moreno M A, Goniú N, Moreno P S, et al. Ethics of social media research: common concerns and practical considerations[J]. *Cyberpsychology, Behavior and Social Networking*,2013,16(9):708-713.
- [26] Fiesler C, Proferes N. "Participant" perceptions of Twitter research ethics[J]. *Social Media Society*, 2018(3):1-14.
- [27] Swirsky E S , Hoop J G , Labott S . Using social media in research: new ethics for a new Meme?[J]. *The American Journal of Bioethics*, 2014, 14(10):60-61.
- [28] 王树义,刁海伦. 基于复杂系统仿真的微博客虚假信息扩散模型研究[J]. *情报科学*, 2014(11):133-138.
- [29] Lease M, Hullman J, Bigham J, et al. Mechanical Turk is Not Anonymous[EB/OL].[2019-08-20].<http://dx.doi.org/10.2139/ssrn.2228728>.
- [30] Facebook Manipulated User News Feeds To Create Emotional Responses[EB/OL].[2019-08-20]. <https://www.forbes.com/sites/gregorymcneal/2014/06/28/facebook-manipulated-user-news-feeds-to-create-emotional-contagion/#14ccbc3539dc>.
- [31] Zimmer M. "But the data is already public": on the ethics of research in Facebook[J]. *Ethics & Information Technology*, 2010, 12(4):313-325.
- [32] 王前. 科技伦理意识养成研究[M]. 北京:人民出版社,2012:1.
- [33] Convery I, Cox D. A review of research ethics in Internet-based research[J]. *Practitioner Research in Higher Education*,

- 2012, 6(1):50-57.
- [34] Mckee H A. Ethical and legal issues for writing researchers in an age of media convergence[J]. *Computers & Composition*, 2008, 25(1):104-122.
- [35] 张书乐. 微博版权:140个字的烦恼[N]. *中国文化报*, 2011-10-21(5).
- [36] 刘文杰. 微博平台上的著作权[J]. *法学研究*, 2012(6):119-130.
- [37] 微博, 开发者协议 [EB/OL]. [2019-09-18]. <https://open.weibo.com/wiki/%E5%BC%80%E5%8F%91%E8%80%85%E5%8D%8F%E8%AE%AE?sudaref=www.baidu.com&display=0&retcode=6102>.
- [38] 中国日报. 专家认为平台开放数据应坚持“三重授权”[EB/OL]. [2019-07-20]. <http://ex.chinadaily.com.cn/exchange/partners/82/rss/channel/cn/columns/vyuat/stories/WS5cb6c5eea310e7f8b15771ef.html>.
- [39] James N, Busher H. Ethical issues in online research[J]. *Educational Research & Evaluation*, 2015, 21(2):89-94.
- [40] 于英香. 档案学术研究“麦当劳化”置疑——从研究程序规范的视角考察[J]. *档案学通讯*, 2007(5):7-10.
- [41] Miles M B. *Qualitative Data Analysis*[M]. *Biostatistics and Microbiology: A Survival Manual*. Springer New York, 1994.
- [42] Honey C, Herring S C. Beyond microblogging: conversation and collaboration via Twitter[C]. *Hawaii International Conference on System Sciences*. IEEE, 2009.
- [43] Naaman M, Boase J, Lai C H. Is it really about me? Message content in social awareness streams[C]. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010*, Savannah, Georgia, USA, February 6-10, 2010. ACM, 2010.
- [44] Loose Tweets: An Analysis of Privacy Leaks on Twitter[EB/OL]. [2020-10-14]. <https://cs.indiana.edu/~kapadia/papers/loosetweets-wpes11.pdf>
- [45] 常文英, 刘冰. 基于可信度分析的微博用户个人信息泄露实证研究——以新浪微博为例[J]. *情报杂志*, 2015(8):169-176.
- [46] Richards N M, King J. Big data ethics[J]. *Social Science Electronic Publishing*, 2014, 1(2):1-6.
- [47] GOV.UK.AI Sector Deal [EB/OL]. [2019-03-26]. <https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal>.
- [48] The Big Data Dilemma [EB/OL]. [2019-03-25]. <https://publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf>.
- [49] Industrial Strategy: Building a Britain Fit for the Future [EB/OL]. [2019-03-26]. <https://www.gov.uk/government/publications/industrial-strategy-building-a-britain-fit-for-the-future>.
- [50] 李侠. 科技伦理:没有约束的科技是危险的[N]. *光明日报*, 2015-07-31(10).
- [51] Belmont Report [EB/OL]. [2019-11-12]. <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/>.
- [52] Bruckman A. Studying the amateur artist: a perspective on disguising data collected in human subjects research on the Internet[J]. *Ethics & Information Technology*, 2002, 4(3):217-231.
- [53] Facebook Manipulated User News Feeds to Create Emotional Responses[EB/OL]. [2020-10-15]. <https://www.forbes.com/sites/gregorymcneal/2014/06/28/facebook-manipulated-user-news-feeds-to-create-emotional-contagion/#665c181e39dc>
- [54] Kramer A D I, Guillory J E, Hancock J T. Experimental evidence of massive-scale emotional contagion through social networks[J]. *Proceedings of the National Academy of Sciences*, 2014, 111(24):8788-8790.
- [55] Zimmer M. Is It Ethical to Harvest Public Twitter Accounts without Consent[EB/OL]. [2019-07-20]. <http://www.michaelzimmer.org/2010/02/12/is-it-ethical-to-harvest-public-twitter-accounts-without-consent/>.
- [56] TCPS 2 (2014)—the latest edition of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans[EB/OL]. [2020-10-15]. <http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-epc2/Default/>.
- [57] British Psychological Society (2017). *Ethics Guidelines for Internet-mediated Research*[EB/OL]. [2019-09-12]. <https://www.bps.org.uk/news-and-policy/ethics-guidelines-internet-mediated-research-2017>.
- [58] 李国杰. 大数据研究的科学价值[J]. *中国计算机学会通讯*, 2012(8):8-15.
- [59] 李国杰, 程学旗. 大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. *中国科学院院刊*, 2012, 27(6):647-657.
- [60] Doerr B, Fouz M, Friedrich T. Experimental Analysis of Rumor Spreading in Social Networks[M]. Even G, Rawitz D. (eds) *Design and Analysis of Algorithms*. MedAlg 2012. *Lecture Notes in Computer Science*, vol 7659. Springer, Berlin, Heidelberg.
- [61] 蒋洁, 陈芳, 何亮亮. 大数据预测的伦理困境与出路[J]. *图书与情报*, 2014(5):61-64.
- [作者介绍] 陈一, 女, 1990年生, 武汉大学信息管理学院助理研究员。  
李斯, 女, 1991年生, 北京大学信息管理系助理研究员(通讯作者)。  
孙异凡, 女, 1991年生, 武汉大学信息管理学院博士研究生。  
收稿日期:2019-12-11