算法技术的盲区与智媒时代的数据伦理困境

王　茜

摘　要　在智能媒体时代，数据正成为新闻内容生产与内容研究的重心，大数据正在成为重要的社会资源。在大数据研究和算法技术的研究中，机器学习手段的出现给传播学以及相关学科的研究者以及新闻业从业者带来新的方法，但同时也给传统的内容分析法带来新的挑战。基于算法分析的计算型数据分析法迎合了大数据时代的挑战，快速便捷且分析量大，然而会丢失很多媒介语境中深层的含义，也无法挖掘语言和文字中的丰富性、复杂性以及内涵的微妙之处，对于媒介内容的研究仅仅依赖大数据计算分析会产生偏差。因此大数据与智能媒体时代依然需要计算机自动分析与传统人工的内容分析法相结合的研究方法。本文以大数据为背景，以媒介内容分析为例，讨论如何根据研究需要，利用数据抓取技术与计算机编程，辅助人工混合交叉进行数据分析，通过一系列案例探索算法技术的不足与盲区，并讨论了智能时代的数据伦理以及困境。

关键词　大数据　智能媒介　算法分析

The Blind-spot of Computational Analysis: The Dispute of Big Data Research Methods in the Era of Intelligent Media

Wang Qian

Abstract　The emergence of big data and new technology has brought new research methods to communication scholars. However, massive data sets of communication are challenging the traditional, human-driven approaches to content

analysis. Computational methods help scholars tackle the big data base more efficiently and accurately, however, the computer is unable to understand the subtleness and richness of human language. The solution is to combine the computational methods with the manual methods. With a detail case study of the news twitter analysis, the research shows the strength of combing two different content research methods in communication research in the era of big data. The ethics in big data analysis is also discussed in the study.

Keyword Big Data, Intelligence Media, Computational Analysis

一 引言

大数据时代的到来，带来了一场关于生活、工作与思维的变革，而关于大数据的分析，也带来了商业、科技、医疗、政府、教育、经济、人文与社会其他各个方面的变化。2012 年联合国发布的《大数据促发展：挑战与机遇》白皮书中，解释了大数据如何帮助政府更好地响应社会和经济指标变化，例如收入、失业、食品价格等，告知各国政府可通过社交网络和手机短信的“情绪分析”来预测失业率等社会数据。[1]在美国的总统竞选中，奥巴马及其团队就将大数据应用到竞选活动中，通过分析挖掘近两年搜集、存储的海量数据，寻找和锁定潜在的选民，运用数字化策略成功定位，拉拢中间派选民及筹集选举资金。在公共健康领域，研究者们利用社交媒介和网络平台收集人们的健康信息，并将这些大数据背后隐藏的信息反馈给医疗人员，借以帮助人们改变自己与健康相关的行为。

大数据是由数量巨大、结构复杂、类型众多的数据构成的集合，是基于云计算的数据处理与应用模式，通过数据的集成共享，交叉复用形成的智力资源和知识服务能力。[2]大数据的特征对使用其的研究者而言有很大的影响，早在 2001 年，时为麦塔集团的分析员道格（Doug Laney）就曾提出关于大数据的“3V”特征，包括数量（Volume）、速度（Velocity）和种类（Variety），还有另一些学者提出 4V 的概念，即还包括价值（value），另有一说认为 4V 代表的是真实性（Veracity）。大数据这几个显著特点，使得它在合理时间内被截取、管理和处理后，成为有用的可被解读的信息。

大数据的出现，使得用户能够在自己的沟通、分享、浏览、网购或聊天等行为中生产出与自身相关的数据。新的技术手段的发展，使研究者们能够追踪、搜集并且分析这些海量的数据。信息领域产生的这些变化，为社会科学包括传播学的研究带来了新路径与机遇，也带来了以大数据为基础的社会科学研究方法的变革。当然，与这些新型研究方法的兴起随之而来的还包括用户的隐私被侵犯以及研究伦理等很多新问题。在新闻传播相关研究领域，国内外很多学者也在尝试着用大数据来研究网络用户在社交网络中的行为特点，比如说对于新浪大 V 用户的微博评论与转贴的分析，以及对于某些社会热点话题的内容分析等，这种内容分析可以涉及健康传播、网络传播等诸多领域。大数据不仅改变了传统意义上的数据搜集方式，也改变了过去我们所熟悉的数据分析的流程。究竟大数据的出现，会给传播学的研究方法带来哪些新的变化呢？本文将以内容分析法为例，并结合具体的研究，深入探讨大数据给传播学领域的研究方法带来哪些机遇和挑战。

二　智能前时代的数据获取与分析

20 世纪 50 年代，美国学者贝雷尔森（Berelson）发表《传播研究的内容分析》一书，确立了内容分析法的地位。[3] 对于内容分析法的定义，是“一种客观、系统、和定量的方式来描述传播的显性内容的研究手段”。内容分析法是一种对文献内容进行客观系统的定量分析的专门方法，其目的是弄清或测验文献中本质性的事实和趋势，揭示文献所含有的隐性信息内容，并对事物发展作出内容预测。[4] 它是一种半定量型的研究方法，其基本做法是把媒介上的文字、非量化的有交流价值的信息转化为定量的数据，建立起有意义的类目分解交流内容，并以此来分析媒介内容的某些特征。

内容分析法包括两方面的工作：一是如何对内容资料进行分析以取得量化的结果；二是如何根据研究需要，设计选择系统化分析的模式，将各种内容分析的量化结果加以比较并定量地说明与解释研究的结果。在传统内容分析法的研究中，研究者一般需要七个步骤：①设定研究问题或研究假设；②选择内容研究的样本，确保总体的完整性和它的特殊性；③设定

类目与分析单元，依据研究假设的需要，制定分类的编码表；④训练编码人员。按预先制定的类目表格，按分析单元进行系统判断，记录各类目所表现的客观事实；⑤进行内容编码分析，按照预先制定的类目表格，系统地判断并记录各类目出现的客观事实和频数；⑥测定信度，确保两个以上参与内容分析的人员对相同类目判断的一致性，信度会直接影响内容分析的结果；⑦编码内容进行分析与解读，对研究结论进行数据比较与分析。这些是传统社会科学时期研究媒介内容的主要研究方法，在数据和算法为主导的时代，这一情形正在发生变化。

因此，在此基础上，提出的研究问题包括：①大数据和人工智能带来什么研究方法与研究路径的新发展？②算法技术为媒介相关的数据获取和数据分析带来了哪些新机遇？③大数据与智能媒介对现有的社会科学和媒介研究带来什么样的研究盲区与伦理困境？

三　大数据带来的机遇与挑战

网络媒体的蓬勃发展带来了新的媒介内容的体系架构，也带来了与新闻传播学相关的新的研究框架。与此同时，越来越多的学者开始转向对网络文本的内容分析。例如，关于微博中的标签研究，及与此相关的社会文化的文本分析等。这样的发展机遇冲击了新闻传播领域内容分析研究法的方方面面。比如有学者总结了内容分析法在网络传播中的样本采集和抽样的问题，另一些研究者对不同类型的微博内容进行分析，包括对健康类微博、科普类微博、政务类微博、纸媒的微博以及学术类微博的内容分析，但大多数研究都使用传统的人工内容分析法。[5]而国外的相关研究则更加立足于社会文化的根基中，例如，有学者曾经对新闻记者发布的 twitter 微博进行内容分析，[6][7]或对新型组织的微博进行分析，[8][9][10]或对外国记者的微博内容进行分析，[11][12]或对公益组织的微博进行内容分析，[13]甚至对无家可归者的微博内容进行分析。[14]在这些研究中，研究者都使用了内容分析法对社交媒介的结构特点及与之相关的社会文化文本进行对比分析。例如，有研究者分析了记者如何使用微博的转发与回复功能来完成新闻生产中“把关人”的作用。[15]

网络时代的蓬勃发展给内容分析法带来很多新的挑战，很多学者开始

对用传统的内容分析法来研究浩瀚的网络媒介内容持怀疑态度。麦克米伦（McMillan）等在研究中就曾经指出使用传统的内容分析法分析网络内容可能带来的一系列问题。这些问题包括网络信息的宏大性使得研究者很难获得具有代表性的研究样本；以及由于网络内容的流动性与变化性会导致在信度的测定中，编码人员很难使用完全相同的内容进行分析，进而影响到内容分析法的信度测量。[16]

大数据时代的到来为社会科学的研究分析方法的重生与发展带来了机遇，研究者可以在更丰富与浩大的内容层面进行数据的抓取与分析。很多社会科学家已经在利用大数据来研究互联网用户的人际互动，这些网络互动中潜藏了大量有用数据；而现在完成这些工作再也不用去问卷或者电话调查成千上万的用户，只需要一些代码就可以解决。

即便是大数据描绘了一幅关于未来数据分析的美好图景，但在实际操作层面，研究者对大数据的截获也存在一定困难。例如，很多新闻传播学的研究者都在试图找到截获关于微博数据、网页新闻页面、网络流量等相关数据的方式，当然研究者可以获得一些部门或企业发布的相关公共数据。例如，美国政府就将很多相关的公共事业相关的大数据公布出来供学者们作为二手数据进行研究；但是对绝大多数的社会科学的研究者来说，得到这些庞大的且具有研究和商业背景的数据并非易事。一方面，很多公司对自己的数据不完全开放，他们不会提供应用数据接口或者 API；另一方面，大数据中的隐私问题仍然没有得到解决。虽然许多学者正在试图开始研究社交媒体，由于用户隐私问题的存在，互联网公司越来越不愿意和研究者或公司分享这些数据。受制于互联网公司的服务条款，很多研究人员即使拿到了数据，他们也无法使用。所有这些现状都给大数据时代的内容分析法带来了新的挑战。

四　算法型数据分析法的优势与局限

如果说大数据为传播学者们提供了海量的可供分析的数据，那么大数据同时也为传播学的研究者们带来了一系列新的处理和分析数据的工具与方法。虽然可以不断听到各方提出大数据研究中如何保障被使用者的隐私不被泄露这样的质疑，大数据却已经给社会科学研究带来了新的技术层面

的创新。正如有学者所言，“我们再也不需要在数据的量与数据的深度之间做出取舍”。[17]这种全新的视野给我们带来了在大数据时代传播学研究的新思路。

在过去的传播学研究中，研究者总是会面临数据过量或者取样过少这样的困扰。例如，在对报纸的内容分析中发现需要分析的新闻篇数过多，对电视进行内容分析时发现电视中有难以计数的节目或广告。而通常来说，研究者采取的对应策略无外乎使用分层抽样或者随机抽样的方式来减少内容分析的文本数量；或者找更多的人员参与到研究的编码与分析工作中。利用计算机直接进行内容分析可以让这一过程变得简单。理论上来说，计算机的分析与处理方式，会帮助研究者解决传统内容分析法中这些关于样本数量或者编码局限性的问题。例如，使用算法分析与系统采集数据的方式，就可以将任意一家新闻机构中全年所有的新闻内容进行分析与归类。这些在传统的内容分析法中以人工的方式是难以想象且不能完成的。

算法技术还可以帮助研究者将庞大的数据分割成我们所需要的精确数据或类别数据，比如研究者需要某个特定时期对某个话题的所有博客讨论贴，或者某类帖子中出现最多的关键词。针对这样一个内容分析研究，在编码过程中，我们就可以利用计算机搜索整个数据库，并使用语句分析与概念图绘制工具等方式来确定最常出现的关键词，最后将这些关键词的频次与关系用视觉化的方式表达出来。例如，在布伦斯（Bruns）等人的研究中，研究者就将网上关于某一政治议题讨论的博客帖子用计算机进行数据处理，并依据结果绘制出一张网络图。[18]

当然对于在内容分析中采用计算编码方式进行分析的方式，也存有很多质疑与讨论。这种研究方法看似快捷，但存在很多弊端，比如说电脑的内容分析一般只能停留在文字的浅表层次，这样对研究者而言，很多文本中深层的含义就会被牺牲掉。而对于使用计算机进行内容分析的方式来说，其最大的不足之处莫过于计算机无法理解人类语言和文字的丰富性、复杂性以及内涵的微妙之处，而这些人工编码的内容分析法却可以做到的。[19]

另一个值得警惕的现象是，在浩瀚的网络中我们所获得的大数据集不是客观的，智能媒介时代，大数据都来源于有关于人类本身以及人类行

为，展现的是人类如何生活的一些合集，然而，这不代表整体。这些数据本身也不是中立的，使用者所依赖的算法其实也不能解释它们。在社会科学领域，我们经常看到一些大数据背后解释的人类行为与故事，通常它们是不完整的，不确定的，甚至不具备科学性的。算法和我们通过网络爬取的大数据其实并不如我们所想象的客观，相反带有很强的主观性。

五　人工智能与大数据分析中的盲区

大数据时代需要更多的学科交叉方法，例如，新的内容分析研究方法就立足于计算机科学与社会科学的交叉融合中。因而传播学者需要应用算法技巧与分析方法来研究宏大的网络媒介内容的数据库，这种方法对社交媒体的内容分析而言更不可或缺。我们看到，一方面，传统的人工编码与计数的内容分析方法已经完全不能应付与处理大数据时代的宏大媒介数据；另一方面，如果仅仅对于媒介内容做计算分析的话，会让研究者难以发现研究内容中人类语言中的隐含意义与微妙之处。[20]因此，在大数据时代的内容分析研究中，我们需要结合传统的人工计数与计算机数据处理的双重方式，来帮助我们获得更多的信息分析数据，得到更准确的分析结果。[21]

在传播学研究中，一些学者也意识到这一问题的存在，并试图在研究中结合这两种不同类型的内容分析法，进而做到扬长避短。例如，有学者在研究挪威广播公司全年的网络新闻内容时，就曾经将计算机数据搜集法与人工内容分析法相结合。[22]在该研究中研究者使用 Python 脚本语言，截取了 74430 篇以文本为主的新闻，并将这些新闻下载到本地的服务器上。在第一阶段的研究中，研究者用计算机内容分析的方式，对网页中的超链接、读者评论、视频等网页内容特征进行自动计数；在第二阶段的研究中，研究者对数据库进行分析，从网页上选取了约 2000 篇文章，采用量化分析与人工编码的方式，对其新闻类别、主题链接、工具条内容等进行了内容分析。这些人工分析的方式，在计算机上都是无法识别和完成的。

通过这个研究不难看出，在内容分析中，计算方法的优势在于确保了数据的质量、准确性与范围，但是采用人工的内容分析法，可以确保我们对主题种类鉴别的准确性。[23]在分析媒介内容的内在含义这一层面，大数

据时代使用传统的人工内容分析法依然保有自己的优越性。因此，在大数据时代对网络媒介内容进行内容分析，如果研究对象是新媒体的结构特征，我们可以用算法内容分析的方式来完成；然而，如果要研究这些媒介的内容结构特征，或者研究分析那些与社会文化语境相关的媒介内容，则依然需要使用人工的内容分析法。这给我们的启示是，传播学研究者可以使用计算机与人工混合研究的方式，以避免丢失那些与社会文化语境相关的文本内容。

六　数据获取与分析的伦理困境

另一个与伦理相关的主题是，在获得大数据进行科研或者新闻报道的同时，应该如何避免对网络的使用带来隐私泄露与伤害。Facebook 中的数据泄露事件已经给了人们足够的教训与启发，这对于社交媒体以及大数据时代带来的数据伦理与算法道德困境有着重要的启示意义。这个技术景观中最大的变化特征是成本，随着数据库变得越来越大，算法技术不断更新，技术侵入隐私的成本已经大大减少。而与此同时，在数据获得过程中，从主体获得知情同意变得越来越难。例如在社会科学研究伦理中，一个很重要的原则是，如果研究主体是未成年人，父母必须被告知并获得知情同意权，但是大数据数据库中没有知情同意书。在智能媒体时代如何获取未成年人在智能终端或社交媒介上的行动痕迹，而不担心存在侵犯和暴露儿童隐私的风险，是一个具有争议而且亟待解决的伦理问题。

此外，越来越多的人在研究社交媒体社区时，采用了参与式观察，网络民族志，或者数字人类学田野调查的新型研究方法。然而，这里的伦理困境是被很多人忽视的，例如，当社会科学家或新闻工作者在网上观察和研究社区以及社群互动时，他们进行观察以及与他人交谈的事实可能会破坏社区和安全感。观察网络社区的研究人员可能在无意中造成社区中的纠纷和对相应群体的伤害，例如，如果有研究者借助数据分析微信群中多次的抢红包行为，并将个体行为，时间节点与红包金额进行数据可视化呈现，那些只抢红包不发红包的人就自然成为研究中的有趣焦点，作为被研究对象的个体的抢红包行为将会在可视化数据中被匿名呈现出来，他们或许会在有意无意之间感受到自己的隐私被侵犯，以及被贬损与曝光的不良

感受，这些在数据收集、分析与呈现中产生的伦理问题也并未被前人研究提及。

当社会科学与新闻结合在一起时，大数据的结果呈现也会形成研究伦理面临的重要挑战。大数据研究对少数族裔人口和社会边缘人群可能产生的影响是不可忽视的，研究者可以使用大数据分析产生最好的误导以及最坏的统计数据。例如，有一些研究者试图从司法判决的材料中收集一些关于犯罪分子的统计资料，他们便可以获得一些关于种族、地域和宗教偏见的重要数据分析结论并形成看似很有逻辑的支持，例如，会得出更多来自某个族裔的人是犯罪分子，或者来自某个地区的人更加具有伤害性，以及某个籍贯的人不可信任等这些不良的结论。

在研究中另一种伦理挑战是在研究者使用他人收集智能媒介数据中，今天手机应用、移动媒介和各种智能媒介都在暴露个人的数据，那么对于使用其他人收集的次级数据库中相关人类行为数据的社会科学家和新闻记者而言，他们是否必须获得第二次知情同意才能进行研究，这个问题仍然是悬而未决的伦理困境。

七　如何弥补数据获取与分析缺陷

网络与数字技术的优势使社交媒体能够吸引足够多学者的参与，并把社交媒介作为其进行公共传播的第二平台。作为新闻传播的重要平台与渠道，微博俨然是全世界记者们搜集信息、发现线人、与不同信源进行快速便捷沟通的一个必备武器。对于记者们而言，他们不用走出办公室甚至不接听电话就能够完成一篇不错的新闻报道。以微博为代表的社交网络已经成为很多记者搜集新闻信息与新闻热点的一个重要来源。作为新闻热线，微博可以给记者们提供实时更新的一手新闻材料；作为新闻编辑的公共空间，它也可以帮助记者核实数据，搜索相关信息等编辑事务。与主流媒体新闻编辑中的新闻“把关人”不同，社交媒体例如微博中对于媒介内容的“把关”与选择在新闻发表之前就已经完成了，并且这种对于新闻内容的选择是由微博信息的发布者通过转发的方式选择完成的。

新闻的信源不仅会影响很多记者的能否获得新闻，同时也为新闻信息的呈现与解读提供相应的框架。[24]一般来说，新闻记者喜欢援引具有话语

权的社会个人、专家或者权威机构发布的信息，以增加其新闻的权威性与可信度，在主流媒体中社会边缘化的声音往往会被忽视。比如曼宁（Manning）2001 年在对“绿色和平”和“地球之友”组织的微博研究中发现，那些被边缘化的信源可能会在记者的信源等级中缓慢上升，并慢慢被认定为权威。[25]

微博的特殊性在于它可以保证一种非主流报道的存在，带来社会中的另一种声音。比如在对 2011 年埃及革命中社交媒体的影响力的研究中，研究者发现由于社交媒介上大量转发的内容而形成了社会的另一种声音。[26]因此对于微博的研究，更需要用内容分析的方式进行，比如通过点击量最高的微博或者转发量最多的微博来评判这些新媒体内容的相关性，而不是传统的内容分析中词频的方式来进行判断。[27]

一种新的新闻产生的方式正在社交网络空间中形成，例如，记者对外界更加多样化与丰富的信源的依赖以及与信源之间的互动。如果研究者想研究以微博为代表的社交网络如何为新闻的生产与重组提供了平台，那他们可以去了解这种互动是基于转发微博，还是在微博上被@ 以后所获得。[28]在微博上，记者寻找信息来源的过程变得公开且有迹可寻。当记者在转发的微博中标明@ 的新闻来源，或者援引某个个人或机构发布的信息时，研究者都可以去捕捉新闻记者与信源对象在微博上的互动；而在传统的新闻采集过程中，研究者无法获知这种记者与信源之间的交流过程。

微博平台的特殊性使得研究者可以追踪网络新闻生产的过程，并通过内容分析法去研究记者如何利用微博发现新闻线索并编辑新闻的社会现象。研究者可以去了解网络空间中的新闻信息来源，比如记者的信源，记者援引或转发了哪些个体或者机构的信息。但是，由于网络信息的容量大、变化速度快等特征会导致抓取这些网络内容有一定困难。大数据赋予了新闻传播学的研究者很多新的可操作的研究方法。

我们以 2010 年底在北非和西亚的阿拉伯国家和相关地区发生的一系列以“民主”和“经济”等为主题的反政府运动“阿拉伯之春”事件为例，在这次运动中，现代移动通讯技术尤其 Facebook、Twitter、Youtube 等社交媒体发挥了重要作用。刘易斯（Lewis）等学者对“阿拉伯之春”事件期间美国国家电视台记者的微博进行了内容分析的研究，并利用大数据对新闻记者的 Twitter 微博进行内容分析。这一事件不仅让人们看到传统媒体和

新媒体之间的平衡正逐渐向社交媒体和公民记者倾斜；也让很多人看到公民记者能够提供有价值的报道，而越来越多该地区新闻专业人员也开始使用社交媒体与读者进行互动。这是一项针对新闻来源如何在媒介对社会现实建构中发挥作用的研究，同时也是一项在大数据时代如何获取、分解，解释数据以及对数据做计算与人工混合型内容分析的研究范例。[29]

这样的研究为大数据时代传播学的内容分析法提供了一种新的范式与样板，作为研究者，也可以看到完全利用抓取软件进行内容分析天然存在着缺陷与不足。如果完全依赖抓取软件去做内容分析，有可能会丢失掉一些重要的媒介信息或者代码，而如何将人工的内容分析步骤通过软件筛选、编程语言等创新方式融入大数据计算为基础的内容分析研究中，以弥补数据计算型内容分析法的缺陷，并为内容分析法的变革带来新的研究方法上的升华与创新，才是传播学研究者面临的新课题与挑战。

八　结语

综上所述，将算法为基础与人工为基础的内容分析法相结合并运用在新闻与传播领域仍然存在很多挑战，依然有很多值得学者们去探索的空间。从上述讨论中可以看出，采用计算的方式进行内容分析是对传统人工方式的效率提升与范围扩大。这两种方式的结合不仅可以帮助研究者解决在编码过程中可能出现的输入错误，而且可以去处理一个非常庞大的数据量。保留传统人工的方式可以帮助研究者保持对研究的文本内容细微处的敏感度，提高研究的深度。因此，这种两者结合的研究方式，可以最大化大数据中的数据体量，以及发挥计算效率的优势；当然，我们还可以通过计算的方式自动选找关键词，再将搜索到的代表性类目提供给人工的编码者进行研究，这时，人工编码的研究人员可以依据文本的实际解读方式选择接受或者拒绝。另一种具有挑战的可能性是，如果我们能将这种研究方式与编程语言相结合，会有更加强有力的分析力量。

即便在拥有强大的计算工具的前提下，研究者依然不能忽视用传统的人工方法进行内容分析的重要性。即便在对大数据的研究中，也不能忽略对文本内容的隐含意义的敏感度。虽然大数据时代的研究者可能更看重数据中重复出现的规律与漂亮的形式，但这些毕竟只停留在数据的表浅层

面。有一些研究者依然停留在数据的表面，或许已经为能够利用端口抓取大数据这样的简易操作程序而欢呼。作为研究者，我们要对数字信息的生产过程中的细微之处保持敏锐的观察。对研究者而言，当大数据被处理简化成一个数据模型以后，其中隐藏的语境意义会更加难以保存。当然，传统的人工方式的内容分析法也存在自身的不足。在两种方法相结合的过程中，研究者需要寻找到这种方法间的平衡，通过计算分析的方式将大数据简化，转换成不需要太多的编码人员就可以进行分析的数据，帮助研究者更加有效与准确地分析出微博内容背后的社会语境含义，并进行人工方式的分析。这也可以看作是大数据时代传统内容分析法的一种重生。

大数据时代，数据的获取变得更加容易，研究者有机会获得几乎所有的数据而非少量的抽样，然而这也带来更多的伦理挑战。大数据中研究的对象是数据的总体而非部分，展示的更多是数据的相关性而非因果关系；研究者也不再关心如何花费高昂的代价去消除数据里面的不确定性，而是考虑如何从海量的数据中获益。[30] 基于这些特点，大数据所呈现的客观性与准确性可能对研究者的方向有一定的误导。一方面，用户或许不能代表所用的被研究对象全体；另一方面，离开了语境，大数据将失去其内在含义。[31]

对传播学者和新闻记者来说，使用大数据进行研究依然需要更多的思考，也带来更多的挑战。在智能媒体时代，数据科学变得越来越重要，传播学研究方法不断革新的过程中，我们要保持批判的心态；在人们为大数据带来的研究方法的划时代变革而欢呼时，研究者却应该保持冷静，并具有一定的大数据媒体素养与伦理意识。大数据时代如何处理社会语境的解析问题，如何避免伦理困境，并将数据分析法进行与时俱进的发展，对于研究者来说依然是一个具有挑战性的议题。

（作者单位：上海交通大学媒体与传播学院）

［本文为上海市哲学社会科学青年课题（项目编号：2017EXW002）“健康传播框架下上海青少年控烟教育与行为改变的实证研究”的阶段性成果，以及教育部青年基金项目（项目编号：13YJCZH205）的阶段性研究成果］

注释

[1] United Nations, "Big data for development: Challenges and Opportunities", White Papers, 2012.

[2] Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., & Gutmann, M., "Life in the network: The coming age of computational social science," Science, Vol. 323, 2009, pp. 721 – 723.

[3] Berelson, B., Content Analysis in Communication Research, New York: Free Press, 1952.

[4] Holsti, O. R, Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley Publishing Company, 1969.

[5] 徐硕. 传播学学者在微博中的知识生产现状分析——基于传播学学者新浪微博内容分析 [J]. 东南传播, 2012 (6): 69 – 71.

[6] Bruns, A, "Journalists and Twitter: How Australian news organisations adapt to a new medium," Media International Australia incorporating Culture and Policy, Vol. 144, 2012, pp. 97 – 107.

[7] Herrera, S. & Requejo, J. L, "10 good practices for news organizations using Twitter," Journal of Applied Journalism & Media Studies, Vol. 1, 2012, 79 – 95.

[8] Blasingame, D, "Gatejumping: Twitter, TV news and the delivery of breaking news," ISOJ Journal: The Official Research Journal of the International Symposium on Online Journalism, Vol. 2, 2011.

[9] Greer, C. F., & Ferguson, D. A., "Using Twitter for promotion and branding: A content analysis of local television Twitter sites," Journal of Broadcasting & Electronic Media, Vol. 2, 2011, pp. 198 – 214.

[10] Messner, M., Linke, M., & Eford, A, "Shoveling tweets: An analysis of the microblogging engagement of traditional news organizations," ISOJ Journal: The Official Research. Journal of the International Symposium on Online Journalism, Vol. 1, 2012.

[11] Bruno, N., "Tweet first, verify later: How real-time information is changing the coverage of worldwide crisis events," Reuters Institute for the Study of Journalism, http://reutersinstitute. politics. ox. ac. uk/fileadmin/documents/Publications/fellows_papers/2010 – 2011/TWEET_FIRST_VERIFY_LATER. pdf, 2011.

[12] Cozma, R., & Chen, K.-J., "What's in a tweet? Foreign correspondents' use of social media," Journalism Practice. Advance online publication, 2012.

[13] Waters, R. D., & Jamal, J. Y., "Tweet, tweet, tweet: A content analysis of non-profit organizations' Twitter updates," Public Relations Review, Vol. 3, 2011, pp. 321 – 324.

[14] Koepfler, J. A., & Fleischmann, K. R., "Studying the values of hard-to-reach populations: Content analysis of tweets by the 21st century homeless," Proceedings of the 2012iConference, iConference '12, New York: ACM, 2012, pp. 48 – 55,

[15] Lasorsa, D. L., Lewis, S. C., & Holton, A., "Normalizing Twitter: Journalism Practice in an Emerging Communication Space," Journalism Studies, Vol. 1, 2012, pp. 19 – 36.

[16] McMillan, S. J., "The microscope and the moving target: The challenge of applying content analysis to the World Wide Web," Journalism & Mass Communication Quarterly, Vol. 1, 2000, pp. 80 – 98.

[17] Manovich, L., "Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.)," Debates in the Digital Humanities, Minneapolis, MN: University of Minnesota Press, 2012, pp. 460 – 475.

[18] Bruns, A., Burgess, J., Highfield, T., Kirchhoff, L., & Nicolai, T., "Mapping the Australian Networked Public Sphere," Social Science Computer Review, Vol. 3, 2011, 277 – 287.

[19] Simon, A. F., "A unified method for analyzing media framing. In R. P. Hart & D. R. Shaw (Eds.)," Communication in U. S. elections: New agendas, Lanham, MD: Rowman and Littlefield, 2001, pp. 75 – 89.

[20] Conway, M., "The subjective precision of computers: A methodological comparison with human coding in content analysis," Journalism & Mass Communication Quarterly, Vol. 1, 2006, pp. 186 – 200.

[21] Lewis, S. C., "The tension between professional control and open participation: Journalism and its boundaries," Information, Communication & Society, Vol. 6, 2012, pp. 836 – 866.

[22] Karlsson, M., & Strömbäck, J., "Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news," Journalism Studies, Vol. 1, 2010, pp. 2 – 19.

[23] Sjøvaag, H., Moe, H., & Stavelin, "E. Public service news on the web: A large-scale content analysis of the Norwegian Broadcasting Corporation's online news,"

Journalism Studies, Vol. 1, 2012, pp. 90 – 106.

[24] Sigal, L. V., Reporters and Officials: The organization and politics of newsmaking. Lexington, MA: D. C. Heath, 1973.

[25] Manning, P., News and News Sources: A Critical Introduction. London: Sage, 2001.

[26] Papacharissi, Z., & de Fatima Oliveira, M., "Affective news and networked publics: The rhythms of news storytelling on Egypt," Journal of Communication, Vol. 2, 2012, pp. 266 – 282.

[27] Poell, T., & Borra, E., "Twitter, YouTube, and Flickr as platforms of alternative journalism: The Social Media Account of the 2010 Toronto G20 Protests," Journalism, Vol. 6, 2012, pp. 695 – 713.

[28] Silverman, C., "Is this the world's best Twitter account?" Columbia Journalism Review, http://www.cjr.org/behind_the_news/is_this_the_worlds_best_twitter_account.php, 2011.

[29] Lewis, S. C., Zamith, R., & Hermida, A., "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods," Journal of Broadcasting & Electronic Media, Vol. 1, 2013, pp. 34 – 52.

[30]〔英〕维克托·迈尔 – 舍恩伯格、肯尼斯·库克耶. 大数据：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。

[31] Boyd, d. & Crawford, K., "Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon," Information, Communication & Society, Vol. 5, 2012, pp. 662 – 679.