

朱廷劭 著

大数据时代的 心理学研究及应用



科学出版社

大数据时代的心理研究及应用

朱廷劭 著

科学出版社

北 京

内 容 简 介

近年来,大数据概念得到了包括心理学在内的各个领域的高度关注。本书以中国科学院心理研究所计算网络心理实验室(CCPL)的一系列研究成果为内容主线,系统介绍了网络心理学的基本概念、研究方法、研究工具及最新研究进展,旨在使读者能够全面地了解这门新型交叉学科的整体概况、清晰地理解大数据对心理科学的研究逻辑和研究方法所产生的深远影响、深刻地领悟利用网络大数据开展心理学研究的非凡科研价值与广阔应用前景,从而为推动后续相关研究课题的开展起到抛砖引玉的作用。

本书可供心理学、医学、教育学、计算机科学工作者和心理学爱好者阅读参考。

图书在版编目(CIP)数据

大数据时代的心理学研究及应用/朱廷劭著. —北京:科学出版社, 2016.3

ISBN 978-7-03-047532-9

I. ①大… II. ①朱… III. ①数据处理—应用—心理学—研究 IV. ①B84-39

中国版本图书馆 CIP 数据核字(2016)第 044348 号

责任编辑: 阚 瑞 / 责任校对: 桂伟利

责任印制: 张 倩 / 封面设计: 迷底书装

科 学 出 版 社 出 版

北京东黄城根北街 16 号

邮政编码: 100717

<http://www.sciencep.com>

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 3 月第 一 版 开本: 720×1 000 1/16

2016 年 3 月第一次印刷 印张: 9 1/4

字数: 180 000

定价: 48.00 元

(如有印装质量问题, 我社负责调换)

序

130年前，威廉·冯特将实验方法引入对人类意识的研究，科学心理学由此诞生。此后心理学的每次重要进步和重大发展，往往离不开对其他学科先进成果的运用与借鉴。无论是在严格控制的实验条件下操纵人的行为，还是在真实生活中无干扰地观察人的行为，有效地利用先进技术都有助于我们更好地理解人类的心理与行为规律。

观察记录人的真实行为，这是心理学研究中的一个重要传统，反映了支撑心理学发展的一种重要价值取向。今天，技术的进步和普及已经使得准确记录和定量分析人类真实行为完全具有了可行性：我们在社交网络上沟通想法，留下了我们的思想踪迹；我们在不同地点用手机签到，留下了我们的行动轨迹……当前这种种真实行为信息都能够以行为大数据的形式存在于由互联网技术构筑的电子空间中，如一座巨大的金矿等待人们来挖掘。

朱廷劭研究员和他的团队以敏锐的眼光和扎实的工作，不仅准确把握住信息技术进步与生活方式变革的潮流，而且有效融合了心理科学和信息科学的研究方法，在网络大数据的金矿上开展心理和行为科学研究，已逐渐显露出强大的生命力与应用前景，并始终保持与国外竞争者同步发展。虽然目前他们的探索性工作距离构建一套完整的理论与实践体系还有不小的距离，但及时总结梳理阶段性成果、付与同仁们交流批评，实为促进学科发展、提升研究水平之正道。

本书所呈现的研究技术对于传统心理学研究者而言也许会感到些许新鲜甚至陌生，但它所反映的研究思路，却又让人觉得似曾相识。那印在教课书上的、写在试卷上的、曾经引发心理学各派别之争的价值判断与思想轨迹，也许会在心理学与其他学科的新一轮交叉融合中焕发新的光彩，成为推动人类认识自我的新力量。

傅小兰

中国科学院心理研究所

2016年3月

前 言

大数据的时代已悄然来临，在得到广泛商业应用的同时，在物理学、天文学、大气学、基因组学、生物学、社会学等众多学科领域也得到了广泛的应用。心理学作为一门研究心理现象及其规律的科学，主要通过研究心理现象的外显表达来间接推测心理现象的变化规律。科技发展为我们获取个体的外显表达提供了坚实的技术基础，能够实现对个体外显表达长时程生态化的记录，在外显行为大数据的基础上可以开展更多心理学研究。如何将大数据技术应用于心理学研究是我们面临的一个重大挑战。

本书作者团队从 2008 年开展相关的研究，从不同颗粒度的网络行为，实现对用户心理特征的识别，包括人格、心理健康等，并且从对个体的心理特征识别拓展到对群体的社会心态的分析，同时结合不同的实际应用场景开展应用推广的尝试。在此基础上，也开展移动可穿戴以及体感技术方面的研究，根据对用户行为的全方面记录实现对心理特征及时准确的预测。利用行为大数据实现对心理特征的预测是开展大数据与心理学结合的第一步，我们希望能够将大数据与心理学研究深度融合，真正将大数据应用于心理学研究。

本书内容是基于我们以往在大数据以及心理学方面的研究成果为主，通过课题成果介绍大数据技术在不同心理学研究领域的实例。本书是在课题组同学毕业论文的基础上形成的，在与他们相处的日子里，我们共同面对课题中遇到的难题，有过沮丧，有过喜悦！感谢宁悦、李琳、张帆、李一琳、聂栋、高锐、李昂、关增达、白朔天、郝碧波、管理、焦冬冬、高玉松和卢婷婷。感谢父母和家人对我的帮助和支持！

作 者

2015 年 12 月

目 录

序

前言

第 1 章 绪论	1
1.1 大数据的概念及特点	1
1.2 心理学的历史与概况	2
1.2.1 发展历史	2
1.2.2 研究概况	2
1.3 大数据在心理学中的应用	4
1.3.1 改善了传统的心理学研究方法	5
1.3.2 催生出全新的心理学研究领域	8
1.4 大数据与心理学的结合原则	12
1.4.1 学科交叉	12
1.4.2 数据驱动	13
1.4.3 目标导向	13
参考文献	14
第 2 章 大数据时代的计算网络心理	15
2.1 网络心理学的研究范畴	15
2.2 计算网络心理学的提出	17
2.3 开展心理学研究的理论依据	18
2.4 网络行为与心理特征	19
2.4.1 网络行为与人格的相关性	19
2.4.2 网络行为与心理健康的相关性	21
2.5 网络挖掘技术	24
2.5.1 结构挖掘	24
2.5.2 日志挖掘	25
2.5.3 内容挖掘/内容分析	25

2.6 小结	26
参考文献	26
第3章 计算网络心理学的研究基础	33
3.1 构建网络行为特征指标体系	33
3.2 网络行为测量方法	38
3.3 网络内容分析	40
3.3.1 词袋模型方法	40
3.3.2 同义词字典方法	41
3.3.3 心理语义字典方法	42
3.4 机器学习方法的应用	44
3.4.1 主动学习	45
3.4.2 半监督学习	47
3.4.3 迁移学习	47
3.4.4 多任务学习	50
3.4.5 总结	52
参考文献	52
第4章 个体心理特征计算	55
4.1 人格计算	55
4.1.1 人格研究的理论依据	55
4.1.2 基于微博数据分析的人格预测	57
4.2 心理健康计算	61
4.2.1 心理健康研究的理论依据	62
4.2.2 基于微博数据分析的心理健康预测	63
4.3 主观幸福感计算	73
4.3.1 主观幸福感研究的理论依据	73
4.3.2 基于微博分析的主观幸福感预测	74
参考文献	78
第5章 群体心理特征计算	80
5.1 群体社会态度计算	80
5.1.1 开展群体社会态度研究的必要性	80

5.1.2 基于微博数据分析的社会态度预测	81
5.2 北京市微博用户社会态度研究	88
5.2.1 数据获取	88
5.2.2 北京市微博用户的月度情绪分析	90
5.2.3 北京市微博用户的季度社会态度分析	91
参考文献	93
第 6 章 自杀风险预测	95
6.1 自杀风险预测研究的理论依据	95
6.2 基于微博数据分析的自杀风险预测关键技术研究	97
6.2.1 基于社会媒体的自杀风险预测研究	97
6.2.2 基于微博数据分析的自杀风险预测的关键问题研究	98
6.2.3 网络识别自杀死亡用户与无自杀意念用户的新浪微博特征差异研究	99
6.2.4 不同自杀可能性水平新浪微博用户行为和-content特征差异研究	104
6.2.5 通过新浪微博特征预测用户自杀风险水平——分类预测模型研究	108
6.2.6 总结	115
参考文献	118
第 7 章 智能移动设备带来的新机遇	122
7.1 智能移动设备的主要类型	122
7.1.1 以智能手机为代表的移动电话	122
7.1.2 以传感器为核心部件的智能手环、手表等设备	123
7.1.3 Kinect 3D 摄像头	123
7.2 基于智能移动设备的心理预测：以智能手机为例	124
7.2.1 相关研究概况	124
7.2.2 手机使用行为与心理特征相关研究	124
7.2.3 相关实验案例	129
7.3 未来研究趋势	132
参考文献	133

第 1 章 绪 论

大数据作为信息时代的一个新概念，得到了包括心理学在内的各领域的高度关注。本章首先介绍大数据的概念及特点，然后从心理学研究的视角探讨大数据在心理学研究中的应用，最后总结大数据对心理学的研究逻辑和研究方法产生的影响。

1.1 大数据的概念及特点

大数据 (big data)，又称海量数据，是指所涉及的数据规模庞大到无法通过人工方式在合理时间内达到截取、管理、处理并整理成人类所能解读的信息的目的^[1]。李国杰和程学旗认为大数据是一种无法在一定时间内用常规机器和硬件工具对其进行感知、获取、管理、处理和服务的数据集合^[2]。大数据这一名称及其常用定义会使人误解数据规模是判断其是否属于大数据的唯一依据。事实上，数据的大小并非是大数据的唯一衡量标准。Laney 认为大数据具备数据体量巨大 (volume)、数据类型繁多 (variety) 和数据处理速度快 (velocity) 的 3V 特征^[1]。这意味着，只有当数据同时符合容量巨大、处理高速和类型丰富三个条件才能够称之为大数据。也正是因为同时兼具这三个方面的特性，大数据才被认为具备挑战系统性能的特性。

当前，大数据时代已悄然来临。大数据中包含着丰富多样的数据类型，并且不同的数据类型之间具有一定程度的关联性，这为挖掘数据中的隐含知识提供了可能。更重要的是，大数据充分展示了“整体大于部分之和”的优势，即在总数据规模相同的情况下，与依次分析独立的小型数据集（类似传统科学研究中的样本）相比，将各个小型数据集合并（类似于总体）后进行分析可以析取出更多的知识。大数据可以被用来挖掘总体规律性的知识，例如，判定消费者喜好、预测选举结果、监测实时交通路况以及监控疾病疫情（如谷歌公司利用大数据成功开发了预测流感传播趋势的在线应用）等。正是由于上述优

点,大数据在物理学、天文学、大气学、基因组学、生物学、社会学等众多学科领域得到了广泛的应用^[3]。

1.2 心理学的历史与概况

1.2.1 发展历史

目前,大数据的应用涵盖越来越多的学科领域,包括心理学。心理学是一门研究心理现象及其规律的科学。人类对心理现象的探究有着漫长的历史,其前身可以追溯到几千年前对灵魂活动的阐释。例如,两千年前,我国先秦诸子的著作中,就已经出现了对人类精神活动现象的思考,而古希腊哲学家亚里士多德的《论灵魂》则被认为是人类历史上最早关于心理现象的专著。这些早期的心理思想都带有不同程度的主观思辨或经验描述性质,尚不能称之为科学。19世纪中叶,德国心理学家威廉·冯特把实验法引进心理学,并于1879年在德国莱比锡大学创建了世界上第一个心理学实验室,由此开创了科学心理学的先河,正式确立了心理学的科学地位。因此,作为一门科学,心理学又是一门年轻的学科。

1.2.2 研究概况

心理现象是一种内部的主观精神现象,摸不着,看不见,没有体积、大小和重量,是一种内隐变量,需要借助外显指标予以表达。因此,心理学研究需要首先探讨心理现象的外显指标,在此基础上通过研究心理现象的外显表达来间接推测心理现象变化规律,研究心理世界的“黑盒子”,实现描述、解释、预测、干预心理现象的目的(见图1.1)。

1. 研究逻辑

心理学的研究逻辑与统计学中的假设检验思想有着密切的联系,即证伪而非证实。具体来说,心理学研究通常以提出假设为起点,一般包括备择假设(H_1)

与虚无假设 (H_0)，前者是与研究预期相符合的推测，而后者则是与之对立的推测。心理学研究并非力求直接证实备择假设，而是通过一定的研究方法来获取研究结果，根据研究结果来判断虚无假设是否可以被接受或拒绝，以此反推备择假设的可接受程度。如果虚无假设被接受则表明不能接受备择假设；如果虚无假设被拒绝则表明可以接受备择假设。

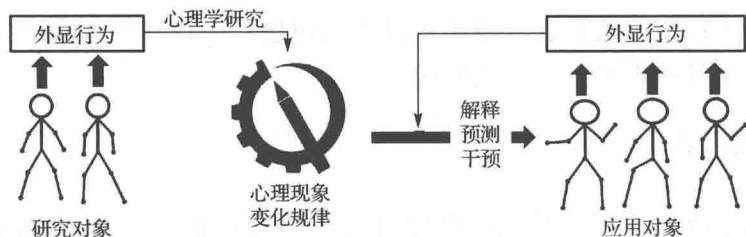


图 1.1 心理学的一般研究流程

2. 研究方法

为了获取真实可靠的研究结果来检验研究假设，需要采用科学的研究方法。心理学的研究方法包括以下几种主要类型。

(1) 观察法

观察法是对心理现象进行系统观察的研究方法。对于一些无法直接操作或难以干预的变量，通常会使用观察法。由于心理变量本身是无法直接观察到的，因此，观察法的观察对象一般能够反映心理现象的外显行为。

(2) 调查法

调查法主要包括问卷法和访谈法。其中，问卷法是心理学中应用最广泛的方法之一，它通过被试评价问卷项目的内容描述与自身实际情况的符合程度，以此判断其心理特征水平。访谈法是通过直接与被试进行交流互动来获取心理信息的方法，常作为编制问卷的前期准备或实验法的有效补充。

(3) 测验法

测验法通过一套标准化的试题来测量所需研究的心理变量。与问卷法相比，测验法使用的工具更具标准化，对数据类型也有更加严格的要求，一般应为等距或等比变量^[4]。测验法主要分为纸笔测验和操作测验两类，其中前者需要用文字回答问题，后者通过手动操作来完成，更适用于受到文化程度或文化背景制约的特殊情境。

(4) 实验法

实验法是心理学中的重要研究方法。它可以用来探究变量之间的因果关系，主要分为实验室实验与自然实验两类，前者要求对实验条件进行严格控制，而后者则仅仅需要对自然情景进行必要的干预。实验法常常会结合问卷法实施。

(5) 个案法

个案法仅对数量有限的案例进行深入研究，通常在一个问题的探索阶段使用。它适用于探索“为什么”或“怎样”的问题，可以详细描述研究对象的特征和解释产生某些现象的深层次原因。

3. 研究工具

在心理学研究中，占重要地位的研究工具是问卷（questionnaire）和量表（scale）。许多问卷和量表是自陈式的，即要求被试根据自身的情况回答问题。而在测量某些心理特征（如人格）时，除了会采用自陈量表外还会采用投射式和情景式量表。投射测验会向被试提供一些刺激情境（通常是图片、句子），让被试自由表达（如看图说故事、补充不完整的句子），根据其反应分析推断其心理特征。情景测验则把被试置于特定的情景中，观察其实际的行为反应并由此推断其心理特征。

除此之外，心理学研究中还有大量针对某一问题专门设计的实验工具，如研究操作性行为的斯金纳箱、研究深度知觉的视崖、研究认知地图的触棒迷津等。近二十年来，随着认知神经科学的兴起，EEG、ERPs、fMRI 等神经生理记录技术与设备也被大量运用在心理学研究中。

1.3 大数据在心理学中的应用

2014 年在美国开展的一项关于情绪传染的研究备受关注，该研究对近 70 万 Facebook 用户的动态信息做了设置，使一组用户接收到以反映积极情感为主的信息，另一组则接收到以反映消极情感为主的信息^[5]。结果显示，用户的情绪会受到这些动态信息所包含的情感的影响，即主要接受积极情感信息的用户的情绪会变得更加积极，而主要接受消极情感信息的用户的情绪会变得更加消极。在传统的心理学研究中，收集处理近 70 万被试样本将耗费巨大的人力与时

间资源，根本无法想象，而利用大数据却可以轻松地解决这个难题。大数据的兴起为心理学研究带来了极大的机遇，对心理学的研究逻辑、研究方法以及研究工具产生深远的影响。

1.3.1 改善了传统的心理学研究方法

1. 研究逻辑

在心理学研究中，假设检验的研究逻辑是先验的，是在得出具体研究结果之前就已经作出的推断。不同于心理学的传统研究逻辑，大数据是根据数据分析得出结论，其研究逻辑是后验的。福尔摩斯曾经说过“*It is a capital mistake to theorise before one has data*”（先于数据的理论是推理的首要错误）。这种思想被许多学科（例如，法医学、刑侦学）奉为研究原则，研究不应使事实符合理论（即不应该先得出理论再去寻找支持理论的证据的研究过程），而应让理论符合事实（即应该先得出事实再从事实中总结出理论）。传统的心理学研究之所以会采用先验逻辑，归根结底是对传统研究方法在数据采集、处理等方面所存在的局限性的妥协。大数据为心理学研究能够采纳后验逻辑提供了一种可能性，进而完善了心理学研究的科学性，同时也符合从实践到理论再到实践的科学研究思路。

根据假设检验的研究逻辑，拒绝虚无假设只代表排除了一个不正确的推断，而由于采样带来的样本集的不同，对于证实备择假设还有一定的距离，仍然需要进一步的检验。大数据则不会受到类似的困扰，这是因为大数据的数据量更接近真实的总体规模，不再局限于采样规模，因此可以通过直接分析大数据来归纳出结论，即数据驱动，而无需从样本上推论总体情况，这大大提高了最终获取正确推断的效率与可能性。

2. 研究方法

大数据不仅为改进心理学的研究逻辑带来了新契机，对心理学的研究方法也将产生深刻的影响。

（1）代表性

传统的心理学研究方法没有能力直接对总体进行研究，这导致需要采取迂

回的方式研究样本,样本规模通常都十分有限,因此传统研究多采用抽样的方式从总体中抽取样本,再把样本的研究结果推广到总体上,这就使研究结论的有效性不可避免地受到样本代表性的影响。大数据研究可以不再通过样本间接推论总体的情况,而是能够做到直接对接总体的全部数据进行分析处理,从根本上消除了样本代表性的问题。

(2) 生态性

传统心理学研究的一个重要缺陷在于对实验条件的过分控制可能会营造一个不同于真实生活的行为情境,从而对被试的行为产生干扰。大数据方法可以在不直接接触被试的前提下收集他们在日常生活中的真实行为数据,从而避免非自然的实验场景对数据质量的负面影响,使得收集到的数据具有更好的生态效度。

(3) 高效性

基于传统研究方法的数据收集、分析过程通常会比较缓慢,结果反馈的时间滞后性严重且需要耗费大量的人力、物力。大数据技术具备对海量数据的高效存取能力,并能够通过对既存的、原始的数据记录进行高速处理,提取出有效特征,使心理学研究在数据处理的规模与效率方面实现质的飞跃。

(4) 瞬时性

人类的心理活动会随时间流逝不断变化,而传统研究方法所收集的数据信息或是带有回溯性质,容易受到遗忘等因素的干扰而产生误差;或是仅通过截取、分析单个或有限几个时间节点,就得出了推广到整个时空的研究结论。如果能够采集到瞬时性信息并对此加以及时分析,那么对心理活动的研究将会更加准确。利用大数据信息采集与处理技术,可以实现对个体和群体外显行为数据的瞬时采集、分析,从而弥补传统研究方法的不足。

3. 研究工具

伴随着大数据概念的提出与信息技术的发展,研究工具也随之得到了较大的改善,从而使得我们可以全时程地跟踪、记录个体的海量行为数据。无论是在现实社会还是虚拟社会中,我们都可以借此获取全时程的、无缝衔接的、多维度的电子数据记录。

(1) 可穿戴设备

可穿戴设备泛指内嵌在服装中,或以饰品、随身佩带物品形态存在的电子

设备。它们以传感器作为收集数据的源装置,不仅可以采集使用者的行为、生理与生化指标(如步态、心率、血氧饱和度等),还具有全面的环境感知能力(通过GPS、摄像头、麦克风等)。由于体积小巧并且与使用者实时接触,能够较为完整地记录个体的外部表现,为开展心理学研究提供丰富而全面的数据来源。

(2) 智能移动终端

智能移动终端是指具备开放的操作系统平台、个人电脑级别的处理能力、高速接入能力和丰富的人机交互界面的移动终端,包括智能手机和平板电脑,其中又以智能手机的使用最为普遍。智能手机和平板电脑,不仅为人们的生活提供了极大便利,而且能够记录用户使用过程中的行为数据(如APP使用日志,短信,电话录音等),其内嵌的多种传感器也可记录用户在使用中的环境数据(如重力加速度等),这些都为我们开展心理学研究提供了更具生态效度的详实数据。

(3) 网络行为记录

网络是网络与网络之间所串连成的逻辑上的单一巨大国际网络,是一个能够相互交流沟通、相互参与的互动平台。不管人们承认与否、喜欢与否,网络上的所有一切内容和使用行为都在被跟踪和记录。这一方面固然会带来对隐私的担忧,但是从研究的角度考虑,这种无时不在的跟踪记录能够为心理学研究提供最真实、最全面的行为数据源。

(4) 社会活动行为记录

当今技术的发展可以详实地记录个体社会活动的行为轨迹。人们外出活动的行程,包括交通工具使用以及旅馆入住信息等都在相应的系统中留下记录,而散布各地的监控探头更能够记录人们在现实社会中的行动轨迹,例如,英国的CCTV(closed circuit television)系统就实现了全国范围内的视频监控。

(5) 大数据存储管理和云计算(cloud computing)

大数据对存储介质容量和数据检索速度都提出了更高的要求,传统的数据库难以胜任这一角色。为应对这一问题,新型的大数据存储技术被不断提出,其中NoSQL数据库逐渐脱颖而出,承担起高效存取大数据的任务。云计算是利用分布式部署的计算机集群,将计算任务分布在由大量计算机构成的资源池上,使用户能够按需获取计算力、存储空间和信息服务,赋予用户前所未有的计算能力,适应大数据计算分析的需要。

(6) 数据挖掘(data mining)

数据挖掘又称数据库中的知识发现,是从大量的数据中挖掘有趣模式和知识的过程,包括多模态数据分析和融合、数据清理、特征提取和选取、数据建

模、模式评估以及知识应用等步骤，其中的核心环节数据建模是利用统计分析以及机器学习等方法对数据进行分析建模。数据挖掘经过多年的发展，目前主要的挖掘建模算法比较成熟，并且在不同领域得到了很好的应用。

1.3.2 催生出全新的心理学研究领域

大数据不仅改善了传统的心理学研究方法，而且也催生出一些全新的心理学研究领域。结合近期国内外相关研究的进展情况，我们提出如下的可能研究创新方向。

1. 网络大数据环境下的心理过程与机制

(1) 网络使用与心理因素的关系

目前网络逐渐深入人们的日常生活，各种各样的网络应用与服务使得人们的生活更便捷，同时每天数十亿用户在网络上留下的痕迹会产生海量数据，将这些数据记录保存下来可以用于探究用户在网络使用中的相关心理因素，如个体的网络行为习惯、这些习惯与心理特征之间的关系等。

(2) 网络使用对认知能力的影响

随着大数据概念的提出，人们在日常生活中可以接触到的知识量与日俱增，但对这些知识的理解可能只是浮于表面，未能深入。大数据时代的信息碎片化一方面扩展了用户接收信息的范围，另一方面却又分散了用户的注意力、消耗了用户的认知资源。研究大数据时代信息碎片化给人们带来的影响，有助于科学认识网络对人类认知的影响，从而在信息加工和利用过程中扬长避短。

(3) 网络使用对社交行为的影响

社交网络指一类可以提供社交用途的网站，其注册用户数量通常可以达到百万量级，如此海量的用户在网络平台上进行互动从根本上改变了传统的交流方式。通过分析用户在使用社交网络时产生的大数据，可以解读这一全新的社交组织方式对人类心理与行为的深刻影响。例如，社交网络是提供了社会支持还是带来了孤独感、情绪如何在社交网络进行传播和传染等。

(4) 虚拟社会与现实社会的关系

虚拟世界之所以能够称之为虚拟社会，是因为在虚拟网络环境中，个体的表现以及人与人之间的关系是人类现实社会的映射。个体在现实社会与虚拟社会中的行为，呈现出复杂的相互作用。个体在现实社会与虚拟社会的一致性与

相关性,构成了目前我们对虚拟社会绝大部分认识的前提。然而,人们在虚拟社会中的行为表现可能受到自身有意识和无意识操纵的影响,因此开展虚拟社会与现实社会的一致性研究,探究两者之间相互作用的机制,对于充分认识虚拟世界、拓展相关心理学理论、防范相关风险都具有重要的意义。

2. 个体/群体多模态大数据分析处理

(1) 多模态数据整理

行为是受心理因素支配所表现出的外在活动,是人类心理现象的外在体现。现代信息技术使我们能够事无巨细地对个体和群体的行为做跟踪记录,但是这些原始数据无法直接用来进行心理学意义上的分析处理,需要首先对这些原始数据进行整理、清洗等预处理。尤其是当前各种数据采集工具已经能够帮助我们获得多模态的用户数据时(包括生理、生化、行为数据等),如何对这些多模态大数据进行有效整理,是有效利用大数据的关键。

(2) 特征设计和提取

在对大数据整理和清洗之后,需要从数据中提取出有意义的特征。如何进行有效的特征设计和提取极具挑战性。传统研究多采用启发式方法,依赖研究人员的专家知识与创造性工作。计算机科学与信息技术的发展可能有助于解决这一问题。例如,目前出现的深度学习(deep learning)方法可以帮助我们实现在大数据之上进行特征学习,找到纷繁复杂的行为表现背后的最基础的特征表达;此外,在充分利用心理学既有研究成果的基础上,建立一个能够完备描述外部表现的特征指标体系也是实现有效特征提取的关键之一。我们针对新浪微博数据建立的网络行为特征体系^[6],作为微博用户行为特征提取的指导方案框架,取得了较好的效果。

(3) 文本处理与分析

网络用户会通过发表网络文本的方式来表达其自身的观点、想法与情绪等心理特征,因此对网络文本内容进行心理语义分析具有重要意义。利用自然语言处理技术,结合心理学研究成果,可以建立起心理语义分析词典。基于心理语义分析词典,我们可以高效地分析用户的网络文本大数据,探究其心理特征。我们在LIWC(Linguistic Inquiry and Word Count)的基础上,成功开发了“文心”(Textmind)(<http://ccpl.psych.ac.cn/textmind/>)中文心理语义分析系统,为相关研究提供了中文网络文本心理语义分析的有力工具。