大数据在心理学研究中的应用

朱廷劭
中国科学院心理研究所

关键词：大数据　心理学　机器学习　研究方法

心理学及其研究范式

心理学是研究心理现象及其规律的科学，是一门既古老又年轻的学科。在19世纪中叶以前，心理学的研究方法都是思辨式的，带有经验描述性质，因此尚不能称之为科学。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专门的心理学实验室，由此开创了科学心理学。

科学心理学的研究立足于反映心理活动的外部表现的客观材料，建立在客观数据的基础之上。然而由于条件所限，长期以来，心理学研究的样本规模都十分有限。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。目前，心理学研究对实验条件的控制通常会营造出不同于真实生活中的行为情境，并在实验过程中始终伴随着实验者效应、要求特征等干扰因素。对于大规模的问卷(questionnaire)和测验研究，传统方法一方面需要用户自陈，数据收集过程比较缓慢，往往需要等待被试反馈足够的信息；另一方面，在数据处理阶段需要大量人工计算整理，效率受到很大限制。传统心理学研究方法所收集的信息要么是回溯性的，容易受到遗忘等因素影响而产生误差；要么是截取单个或有限的几个时间节点，却得出推广到整个时空的研究结论。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部表现数据的实时采集，弥补传统研究方法时效性不足的缺点。

大数据理论与技术的出现，特别是当下数据采集技术的飞速发展和应用范围的极大拓展，使得我们有可能针对极大规模的用户开展研究，进行全时全程的跟踪记录，并实现数据颗粒度的灵活变化，从而使得心理学研究的数据基础更加全面和坚实。基于此，我们利用网络大数据开展多个层面的研究。首先是利用情绪关键词的词频变化，考察群体的情绪变化规律；进而利用机器学习模型，根据用户网络行为进行心理特征的识别。利用预测模型，使得我们通过分析用户生态化的行为数据开展心理学的相关研究成为可能，这种无侵扰的心理学研究使得结果更具有可推广性。

大数据揭示情绪变化规律

社交媒体的出现为情绪变化研究带来了新的机会。人们越来越习惯于将自己的行为和感受展示在网络上。2011年，戈尔德(Golder)等人研究了84个国家的人在推特(Twitter)上不同时间段（以天、周和季节为单位）的情绪变化，发现情绪变化与人的生理节律和日常作息有密切的关系。社交媒体产生的数据体量巨大，例如，新浪微博每天的微博量多达2.5亿条。如果能够利用如此庞大的数据提取有效的情绪信息，就能更生态化地实现对个体乃至群体情绪的实时观察与分析，甚至能够展开回溯和追踪研究。

从2011年9月至2012年6月，我们每隔3个月收集一次数据，每次收集持续一周的时间。鉴于

转发微博并非用户的自我表达，我们只保留原创内容用于分析。原创内容包括原创微博和转发微博中的原创部分。我们利用“中文心理分析系统”（TextMind，简称“文心”）处理微博中的情绪内容，使用“文心”对原创微博中的积极情绪 (Positive Affect, PA) 和消极情绪 (Negative Affect, NA) 进行分析计算，获得不同时间段的积极情绪和消极情绪的数据。积极情绪和消极情绪的计算，分别通过统计“文心系统”字典中的积极情绪词类和消极情绪词类在文本中出现的频率实现。我们计算了情绪在一天内随时间的变化情况，结果见图 1。

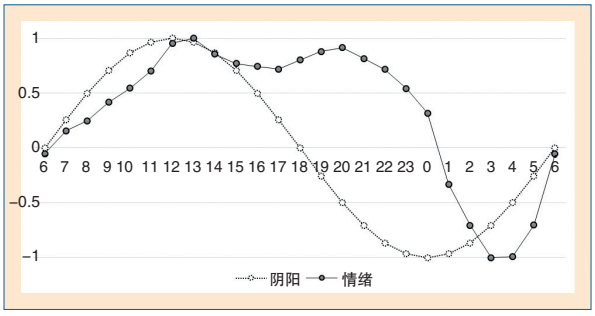

图1 情绪在一天中随时间变化的情况及其与阴阳变化的比较

情绪在一天中的变化，可拆分为前半天和后半天。在前半天，情绪在 0~6 点之间变化最大（见图 1），6 点之后开始升高，直至 13 点达到最高点。在后半天，情绪在 13 点之后有所下降，17 点之后开始恢复，20 点后又开始大幅度下降。从图 1 中可以观察到，从 3 点到 15 点这个时间段，人们的情绪与阴阳变化的趋势是相似的，但从 17 点开始至次日 2 点则不同。阴阳变化规律的总结出自古人对其所属时代的观察，那时的人们保持“日出而作，日落而息”的生活习惯。而如今人们在夜幕降临之后还有很长一段时间的活动，所以旧时的阴阳变化规律（也即生物节律）可以解释当代人在上半日的情绪变化，而余下时间的情绪变化则更多受到当代人的生活节奏和习惯的影响。

在情绪的波形变化方面，相比推特，新浪微博中的积极情绪的统计结果与其昼夜变化（中午最高，稳定至晚上 21 点，以后开始锐减）更为一致。微博的积极情绪的高峰出现在中午 12~13 点和晚上 19~20 点，消极情绪的高峰时间是 0~6 点。在休息时间，还在使用社交网络的人们可能受到睡眠障碍的困扰。研究表明，睡眠质量与消极情绪有显著的负相关。人们在秋天时积极情绪和消极情绪都最低的结果，与中国人自古以来对秋天的复杂情结颇为一致。人们的综合情绪的两个高峰分别是中午和晚上 20 点。前者与古人总结的阴阳变化规律相符，后者则与当代的作息习惯有关。

虽然我们利用大数据的原理能够揭示微博中的情绪在不同时间段的变化情况，但研究的结果是否具备普遍性，或者能否推广还有待验证。原因之一是，本研究的情绪词是从句子中直接提取的，因此筛选出的词汇在原句中的含义与词汇表达的一致性需要进一步验证。另一个原因在于，在微博上随机选取的被试可能存在取样偏差。因为使用微博的人群中，年轻人、城市人居多，且受教育程度偏高，这与真实的人口分布是不一致的。简单统计情绪词频的变化，确实从一个侧面反映了人们的情绪变化，但是某些关键词的使用还不能作为情绪的标记物，这也启发我们能否通过对用户网络行为的分析，实现对其情绪等心理特征的识别。

基于行为数据的人格预测模型

人格 (personality) 是心理科学领域中的重要研究课题，涵盖了个体稳定的行为模式与心理过程，能够解释存在于人际之间的稳定的个性化差异，并且能够与个体、人际、社会等多个研究层面上的结果变量同时保持稳定的预测关系。心理学家已经建立了许多关于人格的理论和模型，具有代表性的理论有卡特尔的人格特质论、艾森克的三因素理论、塔佩斯的大五理论和特里根的七因素理论。其中大五人格模型 (five-factor model 或 big-five) 是目前使用最广泛的人格模型之一，它将人格分为五个因子：开放性、尽责性、外向性、随和性和神经质。传统的人格测量方法主要通过自陈量表的方式来进行。但是，由于自陈量表需要用户人工填写，难以实现

针对大规模用户的实时测量，亟待改善。

近年来，随着社交网络和社会媒体的兴起，有研究者开始尝试利用用户的网络留痕预测其人格，并已经获得了理想的预测效果。2013 年，科辛斯基 (Kosinski)、史迪威 (Stillwell) 和格拉帕尔 (Graepel) 利用脸书 (Facebook) 的“like”（类似于关注点赞）这一属性，抽取用户行为特征矩阵，实现了对用户大五人格指标的自动识别；2015 年，吴（Wu，音译）、科辛斯基和史迪威发现，随着纳入特征矩阵的“like”数目的增多，对用户人格识别的准确度甚至能超过家人对他 / 她的了解程度。

我们通过采集“人人网”的用户行为数据来分析用户人格。通过其日志功能招募被试，共有 335 名用户参加了实验，其中 209 名用户（141 位男性，68 位女性，平均年龄 23.8 岁）符合被试要求。

根据“人人网”提供的社交功能，我们总结出 41 个用于人格预测的 5 类行为特征指标：用户基本信息类特征、社交使用情况统计特征、近期社交状况特征组、情感表达情况特征以及近期情感状况，将大五人格的各个维度划分为低分组、中等组和高分组三个类别。我们采用 C4.5 决策树模型，对数据集的分类达到了很好的结果，并使用了 10 倍交叉验证。表 1 是分类的效果评估。

表1　人格分类结果

人格维度	正确率	召回率
随和性	0.725	0.722
尽责性	0.702	0.703
外向性	0.718	0.718
神经质	0.713	0.708
开放性	0.697	0.694

我们还利用新浪微博，通过机器学习建立基于用户微博行为的人格预测模型。在前人对微博特征研究的基础上，我们引入动态行为的概念，提出了两种提取动态行为的时序特征方法，从而挖掘出能够预测人格的复杂行为模式。通过在线的微博用户实验、被试填写人格问卷的形式，我们获取了 547 名用户的人格得分，并利用微博 API 下载了用户的在线微博数据。利用两种特征提取方法，分别提取了 845 和 795 个特征。对于第一个特征集（845 个特征），在大五人格每个维度，我们分别训练了连续预测模型和分类模型，连续模型的相关性系数达到 0.48~0.54，分类模型的精确度 (accuracy) 达到 84%~92%；对于第二个特征集（795 个特征），训练的连续预测模型预测的人格数值与真实用户填写问卷获取的人格得分的相关性系数达到 0.5~0.63。

利用微博行为对人格进行预测的最佳观察周期（出现最优的模型精度的时间段）在 90~120 天之间。而对于不同的人格维度，利用微博行为来预测人格也存在着难度的差异。例如，预测用户的开放性维度相对容易（模型的预测精度随着观察周期的延长快速提高，30 天后达到收敛），而预测用户的随和性维度则困难（模型的预测精度随着观察周期的延长缓慢提高，并且预测精度的变化趋势不稳定）。这与既有研究的结论保持一致。

利用用户的网络行为来预测用户的人格特征是可行的，这为改善人格测量方法提供了新的视角。由于研究所收集的行为均是客观的，同时模型的预测精度较高，因此基于网络行为分析的人格预测方法能够克服传统人格测量方法的不足（例如，数据追踪困难、资源消耗巨大、测验效率低下），从而为人格研究提供了有力的研究工具，并且为其他相关研究领域提供了有益的借鉴。

由于网络数据具有时间可回溯性，我们可以利用心理计算模型获取任意时间点的用户心理特征指标，从而大大扩展了传统的心理学研究范畴，使得开展跨时空的心理学研究成为可能。

心理计算模型应用于家庭暴力的跨时空研究

家庭暴力 (domestic violence) 广泛存在于世界各国的家庭之中，全世界有 35% 的妇女在一生中曾经遭受亲密伴侣的身体暴力和 / 或性暴力，或者非伴侣的性暴力。根据全国妇联 2002 年的调查显示，我国约有 30% 的家庭存在着不同程度的家庭暴力。随

着家庭暴力事件的屡屡曝光，这一社会问题越来越受到人们的关注。对于家庭暴力的探索研究具有十分重要的意义。

家庭暴力对受害者具有恶劣影响。2016年，布莱德恩(Bleidorn)、霍普伍德(Hopwood)和卢卡斯(Lucas)发现，家庭暴力带来的环境变化对人格会造成明显影响；1992年，达顿(Dutton)发现，遭受家庭暴力的女性会经历更多情绪方面的问题。家庭暴力不仅会带来身体损害，更会造成精神伤害。抑郁、自杀意念是家庭暴力受害者经常出现的两大心理症状。

在对家庭暴力的影响研究，尤其是对受害者心理影响的研究时，常用的方法包括量表法、个案法，或两者相结合的方式。当研究目的是测量受害者心理特征被影响的程度时，常采用问卷法。例如，在家庭暴力对于人格、抑郁与自杀的关系研究中，研究者采用汉密顿抑郁量表(HAMD)、社会支持评定量表(SSRS)、特质应对方式问卷(TCSQ)、艾森克人格问卷(EPQ)等作为测量工具。

传统测量方式存在一定的不足。首先，传统测量方法的测量结果依赖于样本，若选取不同的样本，则测得的结果可能不同，其结果代表性较差；其次，对于问卷、结构式访谈等传统方法，若多次重复测量，容易导致被试疲劳，效度降低，因此无法对受害者的心理特质进行实时检测；第三，传统测量方式只能测量被试在填写问卷时刻的心理状态，很难获取被试以往时刻的心理状态，难以记录个体心理状态实时的变化趋势；第四，传统的研究方法需要大量的人力、物力，且测量周期较长，时效性较差。因此，若想要更加简单高效地测量个体心理特征的变化，则需要寻找更为直接的测量方法。

为了研究家庭暴力对受害者抑郁程度的影响，我们主要分析家暴受害者初次受到家庭暴力前后抑郁程度的变化。取受害者受到初次家庭暴力的时间点，将受害者与对照组在此时间点前后一个月的微博文字与行为数据代入心理计算模型，比较受害者与对照组在此时间点前后抑郁程度的变化，以印证家庭暴力对于受害者抑郁的影响程度。

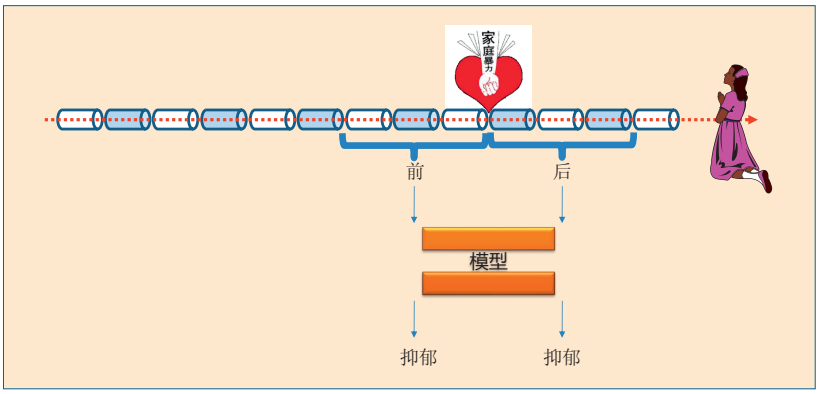

图2 利用网络数据根据抑郁预测模型获取家庭暴力前后的抑郁程度

经过两轮筛选，我们找出233名家庭暴力受害者（简称家暴组）组成一组，并匹配233名受害者组成另外一组。根据抑郁预测模型，计算出每个组别首次遭受家庭暴力前后的抑郁程度（如图2所示），并计算平均值与标准误差。K-S正态性检验结果表明，每个组别的抑郁程度预测值满足正态分布($P<0.05$)，对每组家暴前后数据进行配对样本t检验，结果发现，家庭暴力受害者在首次经历家庭暴力之后，抑郁程度显著升高。此外，对照组在家庭暴力前后抑郁程度没有显著变化。因而可以排除地域特殊事件以及时间因素等系统变量对于结果的干扰。对比家暴组与对照组在家庭暴力前的抑郁程度，并进行独立样本t检验，结果发现，二者的抑郁程度并无显著性差异。

以施暴方式分类进行分析发现，身体暴力与精神暴力均会导致受害者短时间抑郁程度的增加。以受害对象分类进行统计分析发现，夫妻间家庭暴力受害者、受虐儿童在家庭暴力过后的一月内，抑郁程度均显著增加。而目睹亲人家庭暴力的受害者，在家暴发生一个月内抑郁程度没有显著变化。

根据预测模型进行家暴前后抑郁程度的对比发现，亲密关系间暴力会导致受害者抑郁程度显著升高。在出现家庭暴力之后，受害者（本研究中85%

以上为女性）在首次报告家庭暴力出现之后抑郁程度增加。而亲密关系受害者在自杀意念与幸福感两项心理特征中没有表现出显著性差异，但预测值略有变化。可能的原因是本次测量的时间较短，而亲密关系暴力对于受害者的影响表现不充分。

由于网络可以记录大量的用户行为数据与文本数据，使得我们得以追踪家庭暴力受害者首次遭遇家庭暴力之前的心理状态，并以此为基线进行家暴前后的心理状态对比。因此，与传统研究方法相比较，利用计算模型的一大优势是可以跨时计算微博用户在任意时刻的心理特征，并且可以快速对其在相关时间内的心理特征进行计算，快速进行追踪研究。与实验室研究相比较，基于网络的心理计算研究拥有另一项优势，即生态性。近年来，心理学实验室研究的生态效度 (ecological validity) 逐渐引起研究者的重视。我们在未加实验操作控制的条件下，利用网络上自然发生、内容丰富、规模庞大的行为数据与文本数据，对用户心理特征进行预测，有效地避免了用户在填写问卷时的掩饰性作答以及猜测效应。

总结

大数据的出现是技术进步的必然产物，而心理学作为以人类心理现象的外显数据为分析对象的学科，可以抓住这样的机遇。大数据对心理学研究的不同层面都能起到提高效率、增强效度的作用。我们可以充分利用现代信息技术，将大数据同心理学问题和研究范式有机结合起来，一则有望开拓心理学研究的领域和思路，同时也能够为大数据和信息科学的研究提出新的科学问题。大数据与心理学的有机结合，可以通过学科交叉，利用数据驱动，以目标为导向，逐渐实现两者的深度融合。

学科交叉 心理学有着深厚的积累，并且顺应时代的发展需求，能够通过借鉴现代科学技术不断创新。科技发展的终极目标之一是为人类服务，心理学“以人为本”应该能够起到一定的引领作用。心理学借力现代信息技术，通过学科交叉融合，能够实现互相促进的良性循环。

数据驱动 现代信息技术的发展，已经使得我们有可能实现几乎相当于研究对象总体的数据采集和处理。通过对大数据的处理，从数据中发现潜在知识，一方面能大大提高研究的效率，另一方面可在验证阶段对新发现的知识进行快速修正。充分利用大数据覆盖全面、处理高效的优势，以数据驱动将归纳法的边界由样本推向总体，将为心理学及其他社会科学的研究注入全新的动力。

目标导向 心理学考察和分析的对象是人的外显表现，并落实在预测和控制人的外部表现。如果我们以预测与控制人类行为的最终目标为导向，不过度追求可解释性，则有望采用大数据方法直接通过对外部表现的计算分析，实现对外部表现的预测和控制。这一全新路径对于实现心理学研究的终极目标，也许会起到更加直接有效的作用。

心理学的持续发展，不仅需要横向拓展更多的研究领域，而且需要纵向挖掘更深层的关系。得益于大数据技术，未来心理学的研究必将越来越稳固地建立在对客观数据的全面准确分析基础之上，并在研究的效率和效果上实现新的飞跃。在信息技术飞速发展的时代，积极开展大数据与心理学的交叉研究，将极大地促进心理学的学科发展和价值产出。 ■

朱廷劭

中国科学院心理研究所研究员。1999年和2005年分别获得中科院计算所和加拿大阿尔伯塔大学博士学位。主要研究方向为数据挖掘、汉语文语转换以及大数据行为心理等。tszhu@psych.ac.cn

参考文献

[1] Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior[J]. *Proceedings of the National Academy of Sciences*, 2013, 110(15): 5802-5805.

[2] Wu Y, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2015, 112(4): 1036-1040.