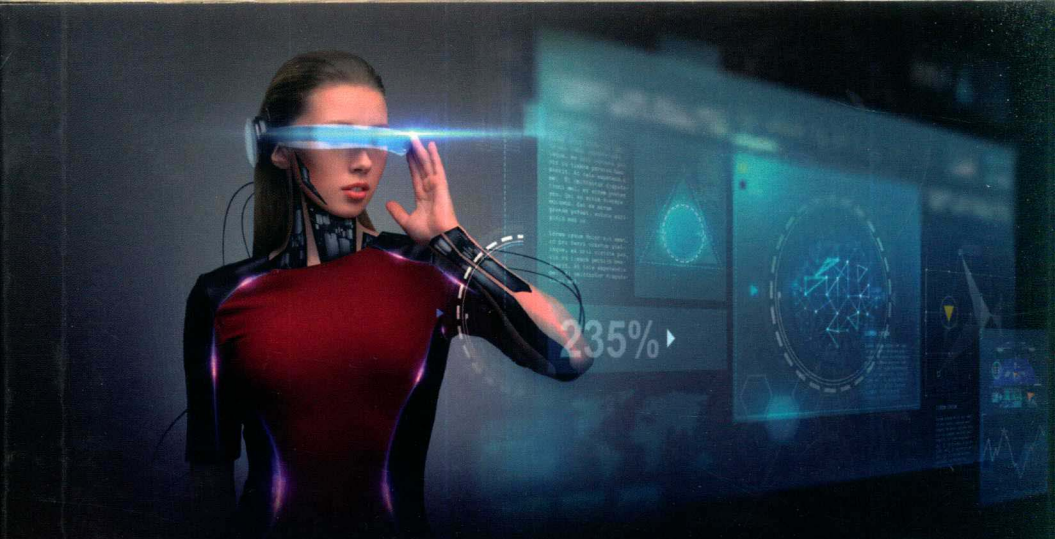

个人大数据

可穿戴计算改变人类生活

LifeLogging: Personal Big Data

[爱尔兰] 卡哈尔·古林 (Cathal Gurrin)
艾 伦·斯密顿 (Alan F. Smeaton)
艾 登·多赫提 (Aiden R. Doherty) 著

王 鹏 秦永强 译著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

个人大数据
可穿戴计算改变人类生活

卡哈尔·古林 (Cathal Gurrin)
[爱尔兰] 艾　伦·斯密顿 (Alan F. Smeaton) 著
艾　登·多赫提 (Aiden R. Doherty)

王　鹏　秦永强　译著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

不难发现，近期大量新技术的涌现使生活记录（lifelogging）逐渐成为一个备受关注的研究方向。计算机存储资源成本的降低，以及电子设备传感能力的增强，都可以帮助人们对个人行为、位置、外界环境进行有效的感知和记录。例如，越来越多的用户参与到量化自我（quantified self）的运动中，并采用可穿戴式传感器对生活行为进行连续跟踪，以便更好地理解人类在执行各项工作或任务中的表现。本书对生活记录领域从多个角度进行了全面总结，其中包括生活记录的研究进程、当前的技术，以及主要应用等。迄今为止，大多数生活记录的研究都聚焦于视觉生活记录（visual lifelogging），即通过视觉媒体记录生活中行为的细节。因此，视觉生活记录也是本书的核心研究内容。除此之外，本书也涉及在生活记录研究过程中遇到的关于信息检索领域的诸多挑战。

版权贸易合同登记号　图字：01-2017-6982

图书在版编目（CIP）数据

个人大数据：可穿戴计算改变人类生活/（爱尔兰）卡哈尔·古林（Cathal Gurrin），（爱尔兰）艾伦·斯密顿（Alan F. Smeaton），（爱尔兰）艾登·多赫提（Aiden R. Doherty）著；王鹏，秦永强译著. —北京：电子工业出版社，2018.7
书名原文：LifeLogging: Personal Big Data
ISBN 978-7-121-33959-2

Ⅰ. ①个…　Ⅱ. ①卡…　②艾…　③艾…　④王…　⑤秦…　Ⅲ. ①便携式计算机　Ⅳ. ①TP368.33

中国版本图书馆 CIP 数据核字（2018）第 064795 号

策划编辑：缪晓红
责任编辑：董亚峰
印　　刷：北京虎彩文化传播有限公司
装　　订：北京虎彩文化传播有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：880×1 230　1/32　印张：5　字数：118 千字
版　　次：2018 年 7 月第 1 版
印　　次：2018 年 7 月第 1 次印刷
定　　价：46.00 元

凡购买电子工业出版社的图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系。联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254760 或 mxh@phei.com.cn。

目　录

CONTENTS

1 简　介

生活记录（lifelogging）代表了一种崭新的现象，即人们能够以不同的细节程度对他们的日常生活进行数字化的记录，从而达到各种各样的应用目的。在某种程度上，我们也可以将生活记录理解为对人类生活行为无所不包的“黑盒子”，它为我们提供了进行生活挖掘和知识推理的无限可能。如同所有新兴技术都有早期的采纳者一样，一些生活记录的使用者（lifelogger，简称生活记录者）竭尽所能地记录更多的生活片段，并将它们存储到生活记录这一“黑盒子”当中。同样，也有很多的使用者不希望存储如此精细粒度的生活记录。这些早期的采纳者们以及他们所开发出来的配套技术往往对生活记录有着更广泛的诉求，例如，有人希望针对特定的应用获得某种缩小版的生活记录，有人则需要持续多年的全量生活行为记录。

生活记录同样有助于基于内容的信息检索、行为上下文检索、浏览、查询、链接、摘要生成，以及用户交互等多方面的研究。然而，进行生活记录的传感器产生的多模态信息受传感器校准或失效等因素的影响往往具有高噪声、易出错和连续性中断的问题。因此，在对这些信息流进行管理、分析、索引和内容访问

过程中面临着诸多的挑战。生活记录之所以能够提供更多前所未有的应用机会，是因为通过这种方式记录的内容黑盒子是我们日常生活的一个集合，通过这个集合可以获得丰富的上下文信息，而这些信息是发现新的有用信息的阿喀琉斯之踵（Achilles' heel）。如果我们知道用户的详细上下文（context），例如用户是谁、他或她在哪里及最近去过哪里、现在在做什么或已经从事了什么活动、和谁在一起等，我们就有可能对这些上下文信息进行综合，以开发出更加有效的信息获取工具。关于更多上下文信息检索的细节，读者可以参阅信息检索基础和趋势系列（Foundations and Trends in Information Retrieval，FnTIR）的《上下文信息检索》一书（*Contextual Information Retrieval*，*Melucci*，2012）。迄今为止，这些由生活记录提供的有价值的上下文信息在信息检索领域仍然缺乏广泛和深入的研究。

在展开本书的论述之前，我们将首先对生活记录的含义进行定义，并讨论谁是生活记录的执行人员，以及为什么要做这样的记录，然后，介绍一些典型生活记录的应用和该领域的核心主题。

1.1 术语、定义和记忆

目前，关于生活记录还没有公认的定义，但有很多相关工作都被称为生活记录，并产生了不同形式的生活日志（lifelog）

数据存档。其中一些非常流行的关于生活记录的行为还包括量化自我（quantified-self）的分析[1]、生活博客（lifeblogs）、图形化生活博客（lifeglogs）、个人数字记忆（personal digital memory）、终生存储（lifetime store）和人类黑盒子（human black box）等。

为了做出合适的定义，我们引用了文献 Dodge and Kitchin（2007）中对生活记录的定义，即生活记录是“普适计算的一种形式，它对单个个体经历的全部进行统一的数字化记录，采用数字传感器进行多模态的捕捉并以个人多媒体存档的形式得到永久存储”。其中，这种统一的数字记录建立在多模态数据采集的基础之上，对这些数据进行整合、存储并处理成具有语义内含和可供检索的信息，以及通过交互界面提供访问等，均能够极大限度地支持各种各样的应用场景。我们将在本书中对这些内容进行更加详细的介绍。

该定义的一个关键方面是生活记录应该尽量记录个体经历的全部。目前，由于传感器硬件的限制，我们很难做到对个体经历进行更加详尽的记录。然而，我们在本书中继续采纳了这个定义中的理念，并且假设生活记录试图捕捉个人行为的详细轨迹。因此，本书中对生活记录的讨论大部分是从多模态感知的角度进行的，其中包括推动了第一代生活记录研究的可穿戴照相设备（wearable cameras）。

[1] http://quantifiedself.com.

可以说，生活记录是一个刚刚得到快速发展的研究领域[2]，关于这个领域的术语尚未得到充分的考虑和定义。因此，为了便于后续的讨论，我们认为生活记录的处理具有如下三方面的核心要素：

- 生活记录（lifelogging）是对由各种传感器被动（passively）收集的生活经历数据进行整合、处理及深入分析的过程，这个过程由单个个体，即生活记录者（lifelogger）完成。这些生活经历数据大多数由可穿戴式传感器对人的活动进行直接感知获得，有时也可能会整合环境传感器数据或其他传感器的输入。
- 生活日志（lifelog）[3]指的是实际收集的数据，这些数据可能存储在个人硬盘中，也可能放在云端或一些移动存储设备上。这些日志可以是一些简单的照片集合，也可以是大量复杂的、长时期积累的可穿戴传感数据集，例如，GPS位置记录、加速度行为轨迹等。
- 记忆代理（surrogate memory）类似于数字图书馆，但是它由生活日志数据和用于组织管理这些数据的配套软件组成。对于记忆代理来说，开发新一代的检索技术以应用于庞大的新型数据集是信息检索领域的一项巨大挑战。在这里，记忆代理并不代表任何形式的认知过程，这一术语仅仅描述了生活日志数据的数字图书馆表示形式。在此之

[2] 尽管生活记录的理念已经出现几十年了，它在近期才在大众中得到广泛接受。

[3] 为便于后续讨论，我们在本书中分别将 lifelogging 和 lifelog 译为生活记录和生活日志。lifelogging 代表这一研究过程，并在这个过程中产生了生活日志——lifelog。

前，记忆代理更多地致力于以事件列表或生活片段的形式对生活日志进行维护。

需要强调的一点是，生活记录往往以环境的（ambiently）或被动的（passively）方式进行，在这个过程中不需要记录者做任何操作。虽然有一些非常专注（dedicated）的个体希望主动地去记录他们生活的全部，但这毕竟是极少数情况。例如，在文献 Fuller et al.（2008）中介绍，Richard Buckminster Fuller 以手工的方式记录了 1920—1983 年每 15 分钟的行为。更近一点，文献 Bell and Gemmell（2007）中，Gordon Bell 的 MyLifeBits 项目将主动和被动记录相结合，采用了可穿戴相机对现实生活中的信息进行采集。另一个主动记录的例子是 Nick Feltron 的 Reporter 应用，该应用允许用户以他们希望的精细度记录他们想要存储的任何行为。Reporter 应用会定期提醒用户“报告”他们当前所从事的活动。

尽管这种更加专注的生活记录形式并非主流，事实上，我们大多数人也经常会以不同的方式主动记录自己的生活，例如，在社交场合拍照等。在这些情况下，我们有意识地做出了拍照的决定，为此摆出造型并面带微笑。然而，生活记录却截然不同，它以被动的形式执行，这就意味着除非对其进行人为的关闭，否则，默认情况下生活记录将总是在执行。因此，生活记录实施过程中将产生大量的数据，其中也有很多重复的数据。生活日志的内容不仅包含生日聚会上故意摆出造型所拍出的照片，还包含用户全天所做任何事情的记录，其中包括生活中的平凡小事和习惯性的

事务。

将生活记录与最近比较流行的量化自我分析领域进行比较可以发现，量化自我更多的是一种将新技术与人的日常生活各方面数据获取进行结合的运动（movement），这反映在输入（例如，食物消耗和环境空气的质量等）、状态（例如，心情、激励和血氧的水平等），以及表现（包括精神上和身体上）等方面。量化自我和生活记录在很多方面不存在本质区别，因此二者具有一定程度上的含混和交叉。本书中认为，生活记录和量化自我分析最大的不同是：量化自我在数据记录的过程中更多地关注特定的领域，例如，运动水平和健康指标等，并且具有清晰的分析目标；生活记录则是对全量的生活经历进行不加选择的记录，并且在生活记录开始的时候，最终的用户目的和分析目标常常是不明确或不能预先知道的。

有效组织庞大生活日志数据存档的时候，我们认为应该以类似于大脑存储记忆的形式进行结构化。对人类记忆模型的讨论显然超出了本书的范围，我们选择了 Cohen and Conway（2008）的人类记忆建模方法，这是由于很多研究记忆方面的科学家在涉足生活记录的应用过程中都采用了这个模型，例如，文献 Doherty et al.（2012）、Pauly-Takacs et al.（2011）和 Silva et al.（2013）。Cohen 和 Conway 的模型认为特定事件和经历的记忆应该被称为情节记忆（episodic memory）。这种记忆形式是自传体的（autobiographical）和个人的（personal），并且可以用于对日期、时间、地点、人物、情感和其

他上下文事实进行回忆。我们的语义记忆则是另一种记忆形式，即大脑对知识、真实世界中的事实、随时间获得的含义和概念等的存储和记录。可以说，我们的情节记忆是个人的，而我们的语义记忆是与别人共享的，并且与我们各自的阅历和情感无关，这些记忆的内容可以独立存在并且是抽象的。文献 Cohen and Conway （2008）表明，我们的语义记忆通常源自情节记忆，这个过程可以认为是从我们自身经历中学习新的事实和知识的过程。对于生活记录来说，迄今为止的大量研究都聚焦在支持和生成情节记忆的数字替代品上。

基于这样的模型，我们可以将典型的一天分割为一系列包含不同时长的事件。图 1.1 展示了一天中事件的时间线形式（timeline），其中事件由图像和各种元数据进行表示。例如，穿衣服及自我清洁，准备食物，吃东西，乘公交车，看电视，听音乐，使用电脑，参加会议，听报告，整理花园，去健身，等等，这些都是日常事件的典型例子，其中有些事件是经常性的事件，因此会重复出现。例如，我们很多人每天都差不多在相同的时间和地点吃相同的或类似的早饭。尽管看电影或参加派对可能不会经常发生，但是可能会以周或月为周期。不可否认，当前对于人类记忆的形式化问题仍然存在争论，本书中关于这方面的观点则认为，生活记录者创建生活日志的过程与 Cohen 和 Conway 建模情节记忆的方法类似（Cohen and Conway，2008）。生活日志捕捉的是我们生活片段中的“事实”部分，而并非它们的情感解释（emotional interpretation）。

图 1.1　事件的时间线表示：生活日志中关键图片和元数据的关联

生活日志通常并不能捕捉和存储语义记忆，因此，当我们想知道阿塞拜疆（Azerbaijan）的首都巴库（Baku），或者 2000 年足总杯（The Football Association Challenge Cup，FA Cup）的冠军切尔西（Chelsea）等，我们并不会向生活日志寻求帮助，而去请教维基百科（Wikipedia）或者进行网络搜索。截至现在，我们并不会参考生活日志来获得语义事实。在这一点上，生活记录存在一个现实的挑战，即如何从生活日志中查询相关信息。这一挑战产生的原因是，前几十年逐步发展起来的信息检索技术针对的是语义记忆而不是情节记忆，我们将在本书后面继续探讨这一观点。

其他的关于生活记录的应用场合也多种多样，例如，采用类似于量化自我的分析方式，可以对我们日常生活进行深入的

探测和挖掘，我们将在后面对应用案例做进一步的详细讨论。如同其他任何技术一样，无论我们在挖掘生活记录过程中采用哪一种应用，我们都应该将这种新技术结合到我们的生活中，并且开发能支持而不是试图改变我们生活的技术。因此，在起始阶段，我们应该问一下自己，形成我们生活的组织形式有什么样的特征和结构，基于这个问题的答案，我们可以建立生活记录的应用。

1.2 发展的推动力

由于数据的获取和低成本的数据存储变得越来越便利，生活记录也逐步被每个人接受。微软的 Gordon Bell 是第一批充分接受生活数字化的先驱，并将他自身生活的数字化作为 MyLifeBits 项目（Gemmell et al. 2002，2006）的一部分，在这个过程中形成了生活记录的基本轮廓。生活记录可以为单个个人生成大量的数据，正如在《经济学人》（*The Economist*，2010）中讨论的那样，这种观点在我们面对大量信息时经常存在，并且在我们各自的生活中也同样存在大量信息。为达到对我们生活日志进行语义内容富化（semantically enrich）的目的，通常需要分析如何将我们的个人生活日志和“外部数据”进行互联。而作为一项信息管理的任务，如何最大限度地挖掘这部分工作的潜力变得非常具有挑战性。生活记录并非是一项新的想法，也不是一项新的实践，然而，撇开对 MyLifeBits 等项目的媒体报道，生活记录之所以在最近变得如

此流行，是出于如下几个原因：

（1）计算机存储无论以云式存储，还是以个人方式存储都变得非常廉价。事实上，我们已经见证了在数字化存储领域较长一段时间内硬盘存储能力的指数级增长。

（2）我们同样可以看到传感器在感知人自身和环境方面的不断进步，在这个过程中传感器成本得到降低，变得更加鲁棒（robust）并且不太会引起人的注意（unobtrusive）。

（3）逐渐增长的社会兴趣也促进了个人感知和记录现象的产生，即所谓量化自我运动的兴起。有时候，它也会受到一些运动或者健康方面应用的推动，而有时我们采用这种方式进行感知仅仅是因为我们能够做这样的尝试。

（4）正如在社交媒体中看到的那样，在我们自身信息的存储和共享方面同样体现了日益增长的开放性。

（5）谷歌眼镜（Google Glass）等新技术（如图 1.2 所示）也对生活记录的发展起到了积极的作用，并将生活记录推上了公共热议的话题之列。

这些贡献因素相互之间各自独立发展，并且其中一些伴随着 2004 年的 CARPE 专题讨论会（Continuous ARchiving of Personal Experiences Workshop）出现（Gemmell et al. 2004）。在这次讨论会上，Steve Mann、Kiyoharu Aizawa、Gordon Bell、Jim Gemmell 等聚在一起，当时，他们很可能会被 Steve Jobs 称为“反叛者”（rebels）及“如方枘圆凿般格格不入的人”（the square pegs in round holes）。2004 年这一届专题研讨会，第一次真正地将这些原本各

自独立进行研究工作的人聚在一起，在研讨会上大家共享了很多研究工具和经验，并且生活记录也作为一项研究领域逐步浮出水面。

图 1.2 典型的可穿戴视觉感知设备。第一行从左至右：谷歌眼镜、百度的 BaiduEye、微软的 SenseCam；第二行：Memoto 微型相机

尽管大多数人对生活记录的兴趣来自我们能够使用的新技术，或者受到生活记录可以应用的新领域的吸引，但是无论是技术还是应用，从它们自身来说，都代表了相当大的挑战。从信息科学的角度来看，生活记录可以提供给我们大量的个人数据档案，然而，这些数据并没有人工标注和语义的描述，而通常是采用原始的传感器数据进行表示（有时还有错误存在）。由此带来的挑战是如何构建对这些数据进行语义理解的工具，以达到对这些数据充分利用的目的。

这个过程类似于早期基于内容的图像检索，但是又与之有所不同，这表现在这种多模态传感器信息构成了生活日志的一部分，并且给大数据分析带来了新的机会。“大数据”是一个经常

被错误使用的术语，由于经常与大容量的信息进行不正确的联系，因此经常采用“大”来对这些数据进行形容。事实上，“大数据”并不仅仅是因为数据量大，它同样关系到数据的真实性（veracity），即数据的准确性（accuracy）和正确性（correctness），这在生活记录中很可能受到诸如数据传感器校准问题等的影响；时效性（velocity），反映了数据随时间的变化模式；多样性（variety），表明了数据获取的多种异质来源。大数据是当前的一个研究问题，这个问题与对不同来源的信息进行挖掘和交叉引用（cross-referencing）紧密相关，并致力于通过这样的挖掘分析发现新的有用知识。生活记录同样带来了新的机遇，即对这种大数据挖掘以个人的方式而不是群体的方式进行。正如最近在文献Melucci（2012）中描述的那样，我们可以认为个人生活记录是一个新的搜索挑战，它可以采用个人提供的新数据源定义新的应用场景和搜索访问方法，并为重新审视上下文信息检索（contextual IR）带来了新的机遇。

1.3 谁，及为什么进行生活记录

正如其他新技术一样，在生活记录领域同样存在本书1.2节提到过的先驱人物，以及将生活记录带入新的应用中的早期采纳者们。他们的这些应用充分展示了对个体生活进行更好的理解所能带来的巨大优势，对这种生活行为的理解远远越过了他们在社交