

第5章



数据生产

近年来,随着云计算、互联网、物联网等新兴信息技术的发展,社会的数据规模在急速增长,人类社会进入了“大数据时代”。据国际数据公司(IDC)预测,到2025年,全球数据圈(数字化存在,即地球在任何一年时间内创建、捕捉和复制的全部数据总和)将扩展至163ZB(1ZB等于1万亿GB),相当于2016年所产生16.1ZB数据的10倍,这些数据将带来独特的用户体验和众多全新的商业机会,成为数据生产工作面临的机遇与挑战。现如今,数据生产的现象、场景越来越多,主动生产数据的重要性越来越高,构成诸多组织转型升级的基础。可以断言,任何组织都有数据生产的必要性和能力,成功与否则取决于其数据思维的养成。

本章将探讨数据生产的概念、数据生产的特点、数据生产的目标、数据生产的阶段、数据的生产源等内容,以使读者对数据生产有深刻的理解和认识。

5.1 数据生产的概念

简单来说,数据生产就是创造新数据或者以数据材料、原始数据为基础加工成为新的数据或数据产品的过程。

对于数据采集,读者想必并不陌生,而对于数据生产,却是一个新鲜概念。数据采集与数据生产的不同在于,生产数据是从无到有的过程,而数据采集则是从有到获取的过程。

那么,该如何具体理解数据生产呢?可以从数据生产的形式来理解其含义(见图5-1)。数据生产有不同的形式。

第一种,用户生成内容形式(User Generated Content, UGC)。例如,淘宝的用户评论,就是一个用户数据生产的过程。淘宝提供了一个数据生产的平台,设定好了生产的模式,用户在淘宝平台上评论,从无到有地生成新的数据。类似的数据生产还发生在不同的场景下,例如人们因为支付行为创造的支付记录数据,搜索信息时生成的浏览记录,微信微博等社交媒体自发的评论数据等,这些都是用户参与数据生产的形式。可以说,UGC是典型的数据生产模式。

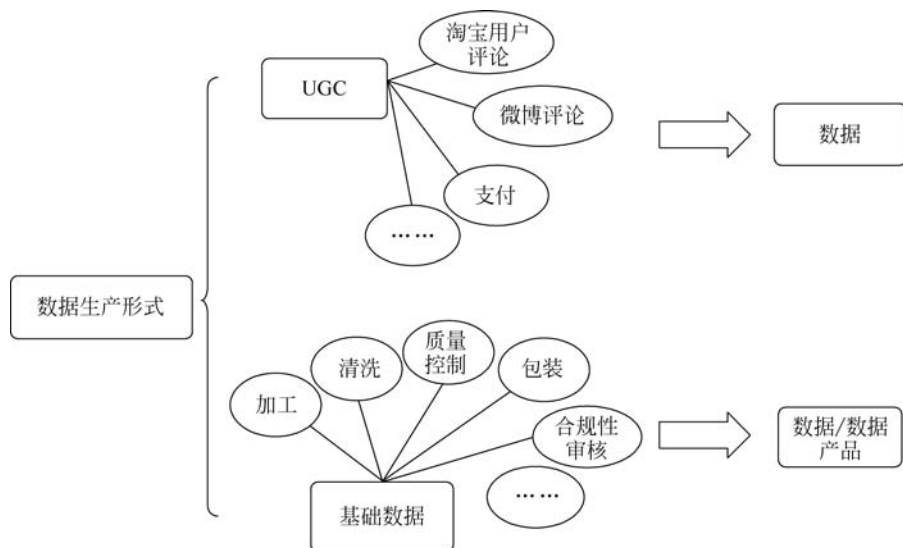


图 5-1 数据生产的形式

第二种,数据产品生产,是将数据材料或者原始数据加工成为数据或数据产品的过程。在数据的分析过程中,常常需要一些数据的基础产品,如清洗过的数据集。从杂乱的数据到形成清洗过的数据集,也可以认为是一种数据生产的过程。这样就大大丰富了数据加工的内涵。人们既可以对数据进行加工,也可以对它清洗,还可以对它包装、质量控制、合规性审核等,从而形成新的数据产品。事实上,这个过程可以进行的工作很多,用生产来代替采集,将“采集”的内涵更多地挖掘出来,让人们更清楚采集的过程中有哪些创新性的工作,进一步改变和改进生产模式,让数据价值的创造更加有深度。

政府开放数据是数据生产的一个典范。政府部门作为最大的数据生产者和收集者,掌握大量的社会(或公共)数据。随着大数据时代的到来,拥有丰富数据的各国政府开始从信息公开走向数据开放。美国是政府数据开放的领导者,2013年,美国数据开放门户网站Data.gov上线,美国联邦政府在网站上公开了包括农业、经济、医疗等来自美国联邦政府各个机构的全面数据。美国联邦政府收集数据的类型主要有三种:一是业务数据,如公民的户口身份信息、企业注册信息等,这部分数据在政府的业务管理中被动产生;二是民意、人口等统计数据,以投入人力、财力、物力的方式主动收集产生;三是气象等环境数据,以传感器等设备主动收集。这三类数据形成了美国政府开放数据的原始数据,联邦政府将各个部门的数据,经过数据脱敏、解密等技术,整合加工形成一个个机器可读、标准化、高价值的开放数据集。

从原始数据到开放的数据集,形成了政府的数据产品。这些开放数据公开供用户进一步开发使用,例如,美国的GPS数据开放后,带动了汽车导航、精准农业、通信等一连串的生产和生活服务创新;美国首都华盛顿,利用公安部犯罪记录和交通部门开放共享的数据,开发出了提示公众避免进入犯罪高发区域和提高警惕的手机短信应用,从而降低犯罪发生概率,维护社会治安。

大数据研究专家托维克托·迈尔·舍恩伯格曾说:“世界的本质是数据,千千万万不起眼的小数据组成了大数据”。随着大数据浪潮来袭,“人无我有”“人有我优”的数据源成为打通大数据应用落地过程中的关键点,拥有原始数据也就具备了数据生产的基础。以百度、阿

里巴巴、腾讯(简称 BAT)为代表的中国互联网企业,在数据领域各有千秋,百度的搜索数据、阿里巴巴的电商数据、腾讯的社交数据,即便放到世界范围来看,我国数亿网络用户产生的数据规模都不容小觑。这些互联网企业利用自己的原始数据生产出各自的数字产品,如百度公司以百度海量网民行为数据为基础,打造了百度指数这一数据分享平台,它能够告诉用户某个关键词在百度的搜索规模有多大,一段时间内的关注度涨跌态势、相关的新闻舆论变化、关注这些词的用户的人群画像,包括其地域分布、年龄、性别、兴趣分布,以及关键词的相关需求图谱,帮助个人用户和企业用户优化数字营销活动方案等,百度指数自发布之日便成为众多企业营销决策的重要依据。阿里巴巴利用网站每日运营的基本数据,包括每日网站浏览量、每日浏览人次、每日新增供求产品数、新增公司数和产品数这 5 项指标,开发出阿里指数数据分析平台,帮助人们观察并预测电子商务平台市场动向。百度指数、阿里指数搜索界面如图 5-2 所示。



图 5-2 百度指数、阿里指数搜索界面

“数据”作为企业和公共组织越来越重要的资产,就像当年“知识产权”对于企业资产形态的突破以及由此带来的企业进步发展一样,将历史性地改变着企业资产的理念和进步发

展的历程。如何利用原始数据生产出新的数据或数据产品,在获得经济利益的同时,让数据价值得到充分发挥,已经成为大数据时代重要的能力之一。

5.2 数据生产的特点

大数据时代数据的产生与生产模式都发生了变化。数据产生的时间、空间、场景等要素更加细密,呈现全时段、大空间与多场景的特征。

1. 数据生产的实时性和移动性

大数据时代的数据呈现出移动性和实时性的特点。随着各种信息技术的发展,联网设备每时每刻都在运作、在线的人们随时随地都可以生产数据。以流文件数据、传感器数据和移动设备数据为代表的实时数据快速流动,速度成为区分大数据与传统数据的重要特征。从电网、供水系统到医院、公共交通及道路网络,实时数据的增长在数量和重要性上都显得极为引人注目。数据已成为消费者、政府和企业日常各方面顺利运作的关键要素。无论何时何地需要数据,人们都希望数据即时可用,这一趋势正变得越来越明显。在实时的海量数据面前,更要求实时分析。

据 IDC 报道,未来几年,移动数据和实时数据都将呈现出强劲的增长势头。移动数据将继续保持自身在数据创建中的占比,实时数据在总体数据创建中的占比则将增长 1.5 倍。实时数据的使用可能(但并非必然)涉及移动设备。例如,生产车间的自动化机器虽然固定不动,但也有赖于实时数据以实施工艺控制和改进。实际上,绝大多数实时数据的使用将由物联网设备驱动(见图 5-3),到 2025 年,在全球数据圈创建的数据中,超过四分之一的数据在本质上都是实时数据,而物联网实时数据将占这部分数据的 95% 以上。

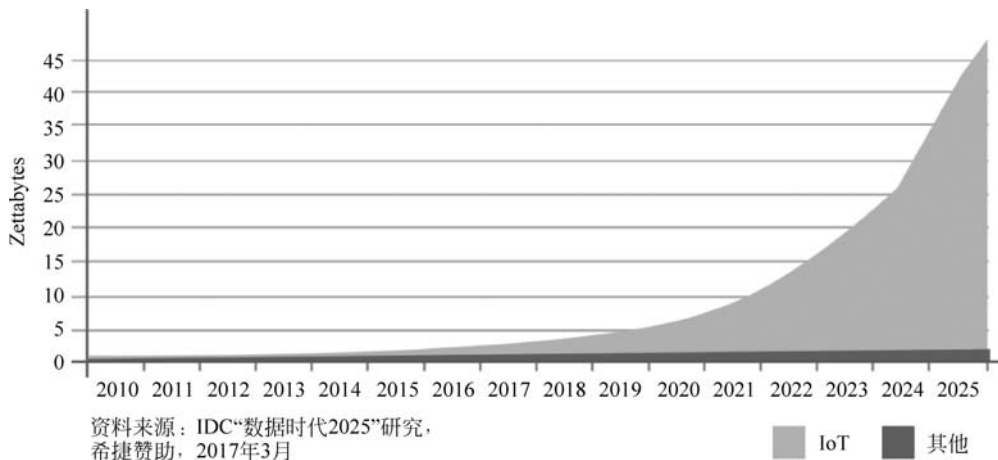


图 5-3 实时数据的种类

2. 数据生产的空间跨度大

大数据时代产生了大量描述在不同时空下不同个体行为的空间大数据,例如手机数据、出租车数据、社交媒体数据等。这些数据为人们进一步定量理解社会经济环境提供了一种新的手段。借助于各类空间大数据,可以研究人类的时间、空间行为特征,进而形成解释社

会经济现象的时空分布、联系及过程的理论和方法。例如,通过社交媒体数据获取人们对某个场所的感受、评价,得到人们对地理环境的情感和认知;再如,基于出租车行驶记录、签到打卡等数据获取海量移动轨迹,得知人们在地理空间中的活动和移动范围;以及基于手机数据获取用户之间的通话联系信息,了解个体之间的社交关系。由于空间大数据包含海量人群的时空间行为信息,使得人们可以基于群体的行为特征揭示空间要素的分布格局。

3. 数据生产的场景多

数据的产生与使用呈现出多场景的特点。以淘宝评论为例,商品交易完成,用户对产品进行评论,一方面,其他用户在购买这一产品时可以查看评论,根据评论的好坏来决定自己购买与否;另一方面,淘宝可以收集所有的评论,将评论数据以一定的价格出售给需要的商家用户,商家对评论进行数据分析,从中挖掘出用户对产品的使用体验,获得该产品的优缺点,以不断改进及时调整自己的产品战略。同样都是评论数据,在不同场景下对不同用户发挥着不同的作用。

5.3 数据生产的目标

5.3.1 采集全量数据

数据是对人类生活和客观世界的测量和记录。过去,因记录、存储和分析数据的工具较为落后,故只能收集少量数据进行分析。数据的采集与生产多以业务为导向,即根据需求去收集所匹配的数据,并且为了让分析变得简单,通常会将数据量缩减到最少。因此,随机抽样一直被公认为是统计时代最有效率的数据分析方法,抽样的目的是用最少的数据代表最准确的信息。大数据时代,随着信息技术的进步,各种传感器和智能设备的普及,能实现数据的实时监测和数采集、传输,人们可以轻易地获得海量数据,选择收集全面而完整的数据进行分析,有助于深入地透析数据,更全面地分析和把握事物的特征和属性,这是传统的随机抽样法无法达到的效果。

1948年,杜鲁门和杜威竞选,盖洛普通过抽样调查预测杜威将当选,结果让所有人都大跌眼镜。其失败的原因在于,抽样调查需要经过问卷设计、信息收集、数据分析等多个步骤,导致其数据滞后于真实情况。在最后两周里,盖洛普不得不停止调查,而杜鲁门恰恰在最后的关头扭转了乾坤。在大数据时代,对谁将当选总统的预测已经出现了新方法:在投票前后,对社交媒体的数据进行挖掘,可以较为准确地预测出谁能当选,如有人通过挖掘Twitter、Facebook等数据,准确预测到奥巴马的当选。这种基于网络数据的挖掘,不需要制定问卷,也不需要逐一调查,数据获取的成本低廉。更重要的是,这种分析是实时的,没有滞后性,所以有越来越多的科学家相信,因为大数据的出现,统计科学将再次发生革命,进入统计2.0时代。事物的诸多真相往往藏匿于细节之中,而随机抽样方法无法捕捉到这些细节。因而要采集全量数据,把数据材料、原始数据中 valuable 的数据尽量全都保留下来。

全量数据采集指尽可能采集所有数据。除了单位内部纵向不同层级、横向不同部门间的数据积累外,还应注重相关外部单位的数据储备,以实现创新应用所需数据全集的流畅协同。实际上,数据只需在纵向上有一定的时间积累,在横向上有细致的记录粒度,再与其他数据整合,就能产生较大价值。以餐饮行业为例,绑定会员卡记录顾客消费行为和消费习

惯,记录顾客点菜和结账时间,记录菜品投诉和退菜情况,形成月度、季度和年度数据,进而判断菜品销量与时间的关系、顾客消费与菜品的关系等,为原材料和营销内容、策略等方面的调整提供决策依据。此外,餐厅还可以与附近加油站建立数据分享协议,主动为就餐时间范围内加油的客户推送餐厅优惠信息,以提升利润空间。当然,将世界上所有产生的数据全部记录下来是不现实的,因而这里的全量数据采集是指在满足应用需求前提下,基于适当的成本,把观测对象在连续时空里的数据全部记录下来。

5.3.2 发现数据的新价值

大数据呈现出了数据的新价值。原先,人们收集数据,是把数据作为资料、档案或者是辅助工作的参考,也就是做一些统计、归纳、检索和归档的工作。数据量相对较小,数据量是分散的、局部的、不成系统的,所以数据所起的作用也是极为有限的。虽然在最近的几十年里,科技发展有了长足的进步,数据处理技术也同步获得巨大的发展,但是整个社会对数据价值的认识和利用还是局限在一个非常狭小的范围内。由于社会对数据价值的漠视,绝大多数的行业对数据价值仍然认识不清。

近年来,随着数据量的急剧膨胀,高新技术的研发和运用越来越依靠数据。社会生活中,数据扮演的角色也越来越重要,一些“大数据先行者”所获得的极大成功,使“数据为王”的观点被越来越多的人接受。大数据应用真正要实现的是“用数据说话”,而不是依靠直觉或经验。大数据应用价值体现在以下三个方面:

一是,发现过去未被挖掘出的价值。在大数据应用的背景下,一些企业开始关注过去其不重视、丢弃或者无能力处理的数据,从中分析潜在的信息和知识,用于客户拓展、市场营销等。例如,企业在进行新客户开拓、新订单交易和新产品研发的过程中,产生了很多用户浏览的日志、呼叫中心的投诉和反馈,这些数据过去一直被企业所忽视。通过大数据的分析和利用,这些数据能够为企业的客户关怀、产品创新和市场策略提供非常有价值的信息。

二是,通过不同数据集的整合创造出新的数据价值。在互联网和移动互联网时代,企业收集了来自网站、电子商务、客户积分卡、移动应用呼叫中心、企业微博等不同渠道的客户访问、交易和反馈数据,把这些数据整合起来,形成关于客户的全方位属性,构建完整的用户画像,将有助于企业给客户提供有针对性、更贴心的产品和服务。

三是,把在一个领域已经发挥过价值的数据再次应用在其他领域创造出新价值。数据是企业的宝贵资源,特别是客户数据、行业数据等。当企业把这些数据从一个业务领域向另一个业务领域拓展进行再利用,这就以低成本的复制发挥了数据的增值价值。很多成功的互联网企业就是基于原始用户群的数据再利用,不断进行业务创新,以在新的领域挖出更高的价值。

5.3.3 考虑外部用户的需求

许多组织内部的部门拥有大量数据,其能产生的作用对本部门来说可能微乎其微,但对于组织外部用户来说,却能产生极大的效益。如视频监控的例子,银行、地铁、电力设施等一些敏感部门或者地点,摄像头都是24小时运转,产生了非常丰富的数据。通常情况下,大部

分视频拍摄的是正常行为,没有长期保留的必要;只有小部分内容包含比较重要的事件,将来或许需要调用;在交通摄像机生成的数据中,地方交通管理部门重视的是交通违法或交通异常的视频,而其会在创建适当的元数据之后丢弃大部分正常的、合法的交通记录;对于娱乐场所的视频监控系统来说,运营商仅重视和保留含有可疑行为的视频,其余部分在创建好元数据和经过一段时间后也会被丢弃;也许保存了一年的视频数据,只有一帧是有用的,但是在研究人类行为的社会学家眼中,这些视频可能就是难得的第一手资料,可以借此窥探和解释人类的某些行为特征。

5.4 数据生产的阶段

大量数据的产生是计算机和信息通信技术(ICT)广泛应用的必然结果,特别是互联网、云计算、移动互联网、物联网、社交网络等新一代信息技术的发展,起到了催化剂的作用。

由于数据是可被计算机读取的信息抽象,ICT 是使得信息可读并且产生或捕获数据的主要驱动力。因此跟随 ICT 的发展与应用历程,阐述数据爆炸式生成与增长的演变过程。概括而言,人类数据的产生大致经历了三个阶段:数据运营阶段、用户生成内容阶段和感知生产阶段,如图 5-4 所示。

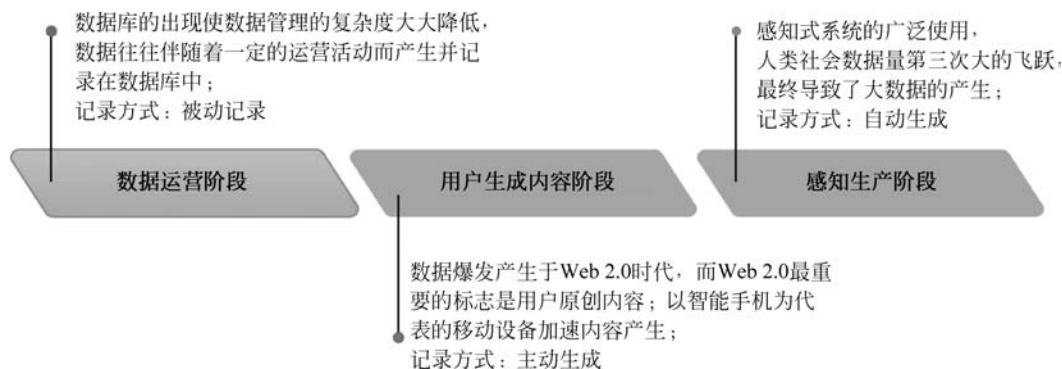


图 5-4 数据产生方式的演变

阶段一：数据运营阶段。该阶段大致开始于 20 世纪 90 年代。随着数字技术和数据库系统的广泛使用,许多企业以及组织的管理系统存储了大量的数据,如大型零售超市销售系统、银行交易系统、股市交易系统、医院医疗系统、企业客户管理系统等大量运营式系统,都是建立在数据库基础之上的,数据库中保存了大量结构化的信息,用来满足企业各种业务需求。在这个阶段,数据的产生方式是被动的。只有当实际的业务发生时,才会产生新的记录并存入库。例如,对于股市交易系统而言,只有当发生一笔股票交易时,才会有相关记录生成。

阶段二：用户生成内容阶段。互联网的出现,使得数据传播变得更加快捷,不需要借助于磁盘、磁带等物理存储介质传播数据。网页的出现进一步加速了网络内容的产生,从而使人类数据量开始呈现“指数级”增长。但是,真正的数据爆发产生于以“用户生成内容”为特征的 Web 2.0 时代。Web 1.0 时代主要以门户网站为代表,强调内容的组织与提供,大量上网用户本身并不参与内容的产生。而 Web 2.0 技术以 Wiki、博客、微博、微信等

自服务模式为主,强调用户参与,大量上网用户成为内容的生成者。这个阶段数据的产生方式是自动生成的,尤其是随着移动互联网和智能手机终端的普及,人们更是可以随时随地使用手机发微博、传照片,数据量开始急剧增加。

阶段三：感知生产阶段。主要是物联网的发展,导致了人类社会数据量的第三次跃升。物联网中包含大量传感器,如温度传感器、湿度传感器、压力传感器、位移传感器、光电传感器等。此外,视频监控摄像头也是物联网的重要组成部分。物联网中的这些设备,每时每刻都在自动产生大量数据,与 Web 2.0 时代的人工数据产生方式相比,物联网中的自动数据产生方式,将在短时间内根据所有者需要生成合规、全量、密集的数据,使得人类社会迅速步入“大数据时代”。

5.5 数据的生产源

自从发明文字以来,各种数据就被记录在不同的载体上。早期数据保存的介质一般是纸张,汇总困难且无法直观地加以分析加工。随着现代信息技术与存储设备的发展以及万物互联的过程,数据爆发的趋势势不可挡。普通大众也时时刻刻身处数字信息的环境之中,在互联网上查阅信息,每次用数码相机拍照,通过电子邮件、社交软件把信息、照片、文件发送给朋友和家人等,人类的种种行为都可能产生数字信息。

一般来说,数据的生产源根据数据生产场景和设备的不同可以分为以下几个部分。

5.5.1 互联网数据

世界互联网统计中心(Internet World Stats)的数据显示,截止到 2019 年 6 月 30 日,全球互联网用户数量已达 44.22 亿,亚洲、欧洲、非洲是人口密集区,同时也是互联网用户的集中区域(见图 5-5)。以中国为例,中国互联网络信息中心(China Internet Network Information Center,CNNIC)发布的第 49 次《中国互联网络发展状况统计报告》显示,截至 2019 年 6 月,中国网民规模已达到 8.54 亿,互联网普及率达到 61.2%。规模庞大的网民数量,在网络上留下多种多样的“使用痕迹”,生产出大量的互联网数据。

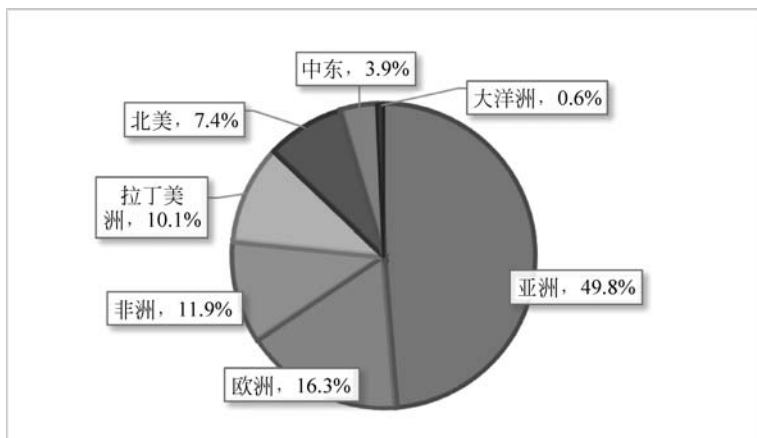


图 5-5 世界互联网用户集中区域

互联网数据由搜索引擎记录、互联网论坛社区动态、聊天记录、社交评论等组成,具有高价值、低密度等相似特征,网络应用如搜索引擎、社交网络平台、网站和点击流是典型的大数据源,这些数据总价值高,数据的价值分散在数据源的各个部分。随着社交网络的发展,互联网进入了 Web 2.0 时代,每个人既是数据的使用者又是数据的生产者,数据规模迅速扩大,数据在每分每秒中被大量创造出来,用户的每一个网络状态,阅读的每一篇文章,上传和分享的每一张照片,都在创造一个数字踪迹,讲述一个故事,而用户可能本身并没有察觉,这其实就是数据生产的过程。

5.5.2 移动网络数据

“生命在于运动,通信在于移动”。移动通信技术的介入已经将世界带到互联网的下一站——移动互联网。得益于更快、更好的连接,移动互联网将成为真正个性化和移动化的网络,它势必会改变人们使用互联网的方式,甚至改变互联网的本质。移动通信和互联网的融合,正在改变全世界人民的生活。

第 44 次《中国互联网络发展状况统计报告》显示,截至 2019 年 6 月,我国手机网民规模达 8.47 亿,较 2018 年年底增长 2984 万,网民使用手机上网的比例达 99.1%,较 2018 年年底提升 0.5 个百分点。在移动互联网的加持下,人们与网络的连接更加密切,更多的数据被生产出来。

移动互联时代,数以百亿计的机器、企业、个人随时随地都会获取和产生新的数据。即便是在“摩尔定律”的支撑下,硬件性能进化的速度也早已赶不上数据增长的速度,并且差距越来越大。随着传统互联网向移动互联网发展,全球范围内,除了个人计算机、平板、智能手机、游戏主机等常见的计算终端外,更广阔的、泛在互联的智能设备,例如智能汽车、智能家居、工业设备和可移动手持设备等都联接到网络之中。基于社会化网络平台和应用,让数以百亿计的机器、企业、个人随时随地都可获取和产生新的数据。

联网的数字设备逐渐取代了独立的模拟设备,产生了大量的数据,而这些数据体验反过来又让用户得到改良和改进系统、流程和用户体验的机会。大数据和元数据(数据的数据)最终将触及人们生活中几乎每一个方面,并带来深远的影响。预计到 2025 年,全球平均每人每天与联网设备互动的次数将达到近 4800 次,基本上每隔 18s 一次。

5.5.3 物联网数据

除了互联网,数据的产生源仍需要说到能量更大、影响更深的物联网世界。1999 年,麻省理工学院自动识别中心创始人凯文·艾什顿(Kevin Ashton,见图 5-6)首次提出“物联网”的概念。物联网是新一代信息技术的重要组成部分,也是“信息化”时代的重要发展阶段。其英文名称是“Internet of Thing”,即 IoT。顾名思义,物联网就是物物相连的互联网。它是通过射频识别、红外感应器、全球定位系统、激光扫描器等信息传感设备,按约定的协议把需要联网的物品与网络连接起来,进行信息交换和通信,以实现智能化识别、定位、跟踪、监控和管理的一种网络。这有两层含义:其一,物联网的核心和基础仍然是互联网,是在互联网基础上延伸和扩展的网络;其二,其用户端延伸和扩展到了任何物品与物品之间,进行信息交换和通信,也就是物物相联。物联网通过智能感知、识别技术与普适计算等通信感知技术,广泛应用于网络的融合中,也因此被称为继计算机、互联网之后世界信息产业发展的第三次浪潮。



图 5-6 物联网之父——凯文·艾什顿

根据物联网中数据采集和传输的过程,其网络架构可分为三个层,即感知层、网络层和应用层(见图 5-7)。感知层由各种传感器以及传感器网关技术架构而成,包括各种传感器(如二氧化碳浓度传感器、温度传感器、湿度传感器等)、二维码标签、RFID 标签和读写器、摄像头、GPS 等感知终端。感知层是物联网识别物体、采集信息的来源,其主要功能是识别物体,采集信息。网络层由各种私有网络、互联网、有线和无线通信网、网络管理系统和云计算平台等组成,相当于人的神经中枢和大脑,负责传递和处理感知层获取的信息。应用层是物联网和用户(包括人、组织和其他系统)的接口,它与行业需求结合,实现物联网的智能应用。

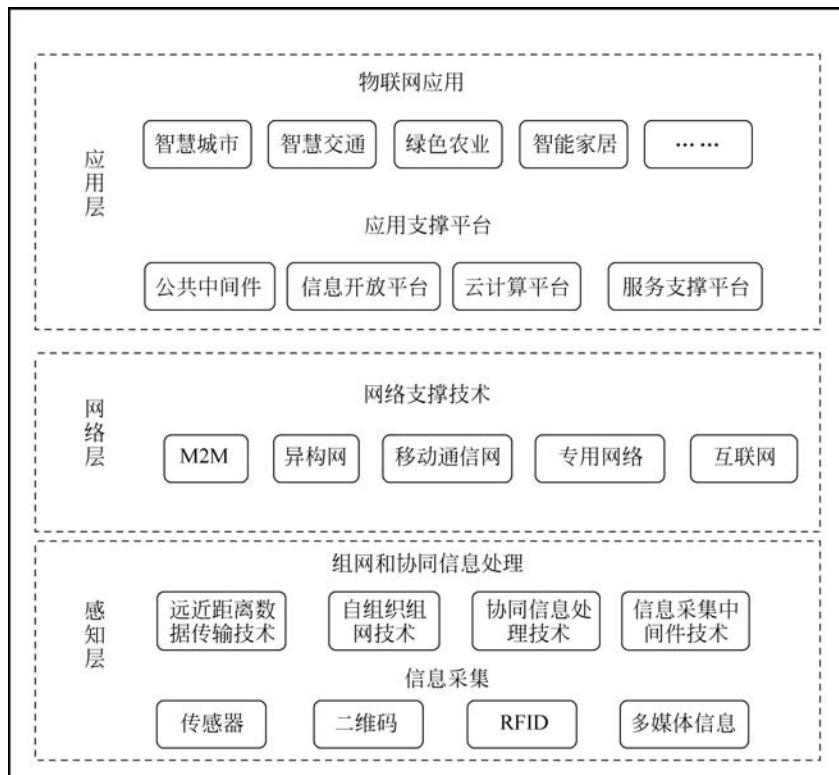


图 5-7 一般物联网架构

物联网的行业特性主要体现在其应用领域内,绿色农业、公共安全、城市管理、远程医疗、智能家居、智能交通和环境监测等各个行业均有物联网应用的尝试,某些行业已经积累了一些成功的案例,真正实现了物联网的功能。

以农牧业为例,1990年以来,全球各地陆续爆发动物疫情。2003年12月,美国发现了第一宗疯牛病病例,2004年起联邦政府农业部启动了“全国动物身份识别系统”的项目,为全国新生牲畜建档立户、安装射频识别耳标。通过这个医用传感器,对牲畜进行连续跟踪,一旦家畜疫情爆发,就能通过数据库追踪溯源,快速确定传染源和传播范围。目前,美国已经装备射频识别耳标的家畜总数无从得知,但可以肯定这个数据库也是海量级的。

一架波音787飞机,每一次飞行所产生的物联网数据量大约有500GB。在物联网领域,有超过3000万的物联网传感器工作在运输、汽车、工业、家电、公用事业和零售部门并产生数据,而这些传感器每年仍将以超过30%的速率增长。据著名咨询公司Gartner预计,到2020年全球将有250亿台的设备通过物联网连接,这些联网设备中存在着各式各样的音频采集器、视频采集器、多样的虚拟感官系统(视觉、听觉、嗅觉)等。未来几年内,传感和移动设备将更深入延伸至人们的日常生活,导致数据爆发。另根据相关研究统计,物联网中产生的来自传感器的数据逐步超越互联网的数据量,如果算上工业企业自动化生产线及设备上的运行数据,特别是随着工业4.0推进而带来的数据爆炸,物联网数据的量更是呈现几何级数增长。可以说,未来人们谈到或研究“大数据”,无疑物联网将是主要的数据来源。如此的物联网世界,其数据产生速度不可避免地会大于人的互联网世界。

除了按照上述数据来源架构和设备的不同对数据生产源进行分类,数据生产源还可以按照用途和目的大致分为不同行业数据和科学研究数据。

行业数据生产是指不同的行业如电信、银行、金融、医药、教育、电力等行业在每天的运转过程中产生的数据,如医疗行业产生的数据集中在患者的数据,通过对患者数据的分析,可以更精确地预测病理情况,从而对患者采取恰当的措施;再如银行业产生的数据包括用户存款交易流水、利率市场投放信息、业务信息等。此外,这些行业的信息化系统如自动化办公系统(OA)、企业资源计划(ERP)、客户关系管理系统(CRM)等,每一天都会有大量的数据产生并沉淀下来。例如,OA系统中各种办公流程所产生的人事、财务、业务、项目等方面的数据,以及后台的日志数据;ERP系统中关于企业人、财、物、时间、空间等资源与企业供应链方面的数据。

科学研究数据是指科学家为了获取准确的数据通过观测、监督、实验、记录、计算等科学研究行为而产生的专业数据。例如,在计算生物学领域,自人类基因组计划启动以来,以新一代测序技术和质谱技术为代表的各类组学技术的飞速发展,推动了基因组、转录组、表观遗传组、蛋白质组、代谢组等海量生命科学组学数据的指数级的增长。同时,机器学习和人工智能技术极大提升了医学影像和分子影像技术的分析能力,正在改变这些专业的科学研究数据的应用方式。



延伸阅读

物联网无处不在

20世纪90年代,一群卡耐基·梅隆大学的程序员去楼下自动售货机买可乐时,经常会碰上缺货或可乐不是很凉的情况,这群懒家伙灵机一动,写了个程序来监控可乐的状态:是

否有货,是否够凉,并把这台自动售货机连进网络——这台自动售货机,大概算是物联网的鼻祖之一。

可能,目前你的家用电器还不够智能;可能,你的信息数据还不能完全被收录;可能,你觉得万物互联的时代还要等到 2049 年才有眉目;但其实,你早已成为物联网中的一员,并且受益良久。

你网购过没有?当你下单后,产品安排出库,通过 RFID 技术便拥有了射频标签。这些标签上传至网络,经读取应用后就以物流信息形式出现在你的眼前。近年来,有了 GPS 的加入,你便可以清晰地看见货物随着卡车跑到了何地,将由哪位送货小哥亲自送到你的府上,而商家也知道何时会收到你的货款。如果从技术的角度出发,物联网就是把所有的物品通过射频等信息传感设备与互联网连接起来,实现智能化识别和管理。因此你的这次消费就是物联网中一次成功的操作。

你打过滴滴没有?那些成天被你呼来唤去的滴滴用车也是物联网的一个例证。当你发布信息的一刻,手机就是镶嵌在滴滴网络上的一个传感器。在数据中心历经 1000 次运算后,你的信息就会推送给当前合适的司机。他对你的目的地了如指掌,你对他的行踪也尽在眼里。你不担心他放你鸽子,他也不怕你不付车费。最终,你们相视一笑,一辆别人的车就为你这件特殊的“货物”开启了在物联网信道里的行程。

你骑过共享单车没有?只要打开手机对着车上的二维码“扫一扫”,即可解锁骑车,开启一次动感单车之旅。共享单车的智能锁采用的正是物联网的“手机端-云端-单车端”的架构。你在手机端使用着查看单车、预约开锁等功能;云端控制整个共享单车系统,与所有的单车进行数据通信,收集信息指令,响应你的需求和管理员的操作;单车端的智能锁内部集成了 GPS 和物联网卡 SIM,将车辆的所在位置和电子锁的状态数据传输给云端,物联网卡通过装置在单车上的二维码,经过接口和无线网络连接,实现人与车、车与车之间的沟通和对话。

小结

近年来,随着云计算、互联网、物联网等信息技术的发展,数据的规模在快速增长,人类已经进入了“大数据时代”。大量数据的产生能够让各行业更好地了解客户需求,提供个性化的服务,并催生出数据生产这一概念。

数据的产生和收集本身并没有直接产生价值,最具价值的部分在于:当这些数据在收集以后,会被用于不同的目的,数据被重新再次使用,从而创造新数据或者以数据材料、原始数据为基础加工成为数据产品,从而为人们提供全新的服务。

大数据时代数据的产生与生产模式也发生了变化,数据产生的时间、空间、场景等要素更加细密,呈现实时性、大空间与多场景的特征。以物联网数据的实时数据在大规模的增长,给采集和处理实时数据的系统提出新的要求;具有时空标记、能够描述个体行为的空间大数据使得人们可以研究人类时空行为特征,进而形成解释社会经济现象的时空分布、联系及过程的理论和方法;同一种数据也可能在不同场景发挥着不同的用处。

大量数据的产生是计算机和网络通信技术广泛应用的必然结果,受各种信息技术的影响,数据经历了从数据运营阶段数据的被动产生,到用户生成内容阶段的数据的主动产生,再到数据感知生产阶段的数据自动产生,每一阶段都呈现出不同的特点。

讨论与实践

1. 结合自己的思考,谈谈对数据生产的理解。
2. 阅读相关文献,理解大数据时代数据生产的特点。
3. 结合自己的理解与思考,对数据生产源进行进一步的探究。

参考文献

- [1] 霍雨佳,周若平,钱晖中.大数据科学[M].成都:电子科技大学出版社,2017.
- [2] 李学龙,龚海刚.大数据系统综述[J].中国科学:信息科学,2015,45(1):1-44.
- [3] 郑英豪.大数据与企业大数据的来源[J].上海经济,2015(Z1):98-100.
- [4] 庄红韬.2年生成人类史上9成数据,迅猛增长的“大数据”将改变未来[EB/OL].(2012-12-20)[2018-11-30].<http://finance.people.com.cn/n/2012/1220/c348883-19959503.html>.
- [5] 佚名.物联网[EB/OL].(2018-06-16)[2018-12-01].<https://baike.so.com/doc/53278345563006.html#5327834-5563006-3>.
- [6] 曹洋,王建平.物联网架构及其产业链研究[J].技术经济与管理研究,2013,(2):98-101.
- [7] 王伟军,刘蕤,周光有.大数据分析[M].重庆:重庆大学出版社,2017.
- [8] 张国庆,李亦学,王泽峰,等.生物医学大数据发展的新挑战与趋势[J].中国科学院院刊,2018,33(8):853-860.
- [9] 宁康,陈挺.生物医学大数据的现状与展望[J].科学通报,2015,60(Z1):534-546.
- [10] 郎为民.漫话大数据[M].北京:人民邮电出版社,2014.
- [11] Min Chen, Shiwen Mao, Yin Zhang, et al. Big Data Related Technologies, Challenges and Future Prospects[M]. Springer Cham Heidelberg New York Dordrecht London, 2014.
- [12] Jose' Marl'a Cavanillas, Edward Curry, Wolfgang Wahlster. New Horizons for a Data-Driven Economy[M]. SpringerLink.com, 2016.
- [13] 中国互联网络信息中心.第44次《中国互联网络发展状况统计报告》[EB/OL].http://www.cnnic.net.cn/hlwfzyj/hlwxbzg/hlwjbg/201908/t20190830_70800.htm, 2019-08-30.
- [14] 苏商会.凯文·艾什顿:物联网的一个世纪也是中国的一个世纪,它给中国带来了什么?[EB/OL].(2017-12-12)[2018-12-01].http://www.sohu.com/a/210048786_465557.
- [15] 佚名.阿里指数[EB/OL].(2018-12-30)[2019-01-05].<https://baike.so.com/doc/559846-592727.html>.
- [16] 波金金.百度指数[EB/OL].(2018-11-20)[2019-01-05].<https://baike.so.com/doc/5347819-5583266.html>.
- [17] 张铭睿,谢安.美国政府开放数据的实践及启示[J].中国统计,2015,(5):24-26.
- [18] 21世纪经济报道.BAT大数据野心:数据生产全链条浮现[EB/OL].(2016-07-04)[2019-01-06].<https://finance.sina.cn/usstock/hlwgs/2016-07-04/tech-ifxtrwtu9775904.d.html?from=wap>.
- [19] 涂子沛.数据之巅:大数据革命,历史、现实和未来[M].北京:中信出版社,2014.
- [20] 王桂玲,王强,赵卓峰,等.物联网大数据处理技术与实践[M].北京:电子工业出版社,2017.
- [21] 赵国栋,易欢欢,糜万军,等.大数据时代的历史机遇:产业变革与数据科学[M].北京:清华大学出版社,2013.
- [22] 娄岩.大数据技术应用导论[M].沈阳:辽宁科学技术出版社,2017.
- [23] 赵刚.大数据技术与应用实践指南[M].北京:电子工业出版社,2013.
- [24] 刁生富,姚志颖.论大数据思维的局限性及其超越[J].自然辩证法研究,2017,33(5):87-91,97.