



第一章 数 据

人们总说,我们进入了“数字生活”的时代,社会也依赖于“数字经济”的发展,那么数据到底是什么?本章就从“数据是什么”“数据有什么用”“数据改变了什么”这三个方面,对“数据”这个名词进行全面的讲解。

数据是计算机中的0和1,数据是数学中的数字,数据是一个人每天的生活记录,数据也是一个公司的营业额和运营日常。数据是一个宽泛的概念,我们生活中的一切都与数据有关。数据是需要加工的,从数据清洗到数据分析,看似杂乱的数据会转变为具有价值的信息。最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡,麦肯锡称:“数据已经渗透到当今每一个行业和业务职能领域,成为重要的生产因素。”大数据时代下,人工智能、网购、数字营销等新鲜事物的出现已改变了各个行业的发展模式,带来了新的“数字经济”,本章对典型行业基于数据的改变进行系统的叙述,感受数据给我们带来的变化。

通过本章的学习,我们将发现数据无处不在,将认识到数据的价值和对我们生活的改变。让我们通过学习,拥有一双发现数据的眼睛,拥有挖掘数据价值的能力,在我们的学习或职业中,用好“数据”这个新工具吧。

第一节 数据的概念与特征

数据是什么?本节将介绍关于“数据”的方方面面,包括:最初始的数据是怎样的?数据如何分类?数据有什么特点?以及大数据的利与弊等问题。通过对数据的起源、概念、分类、特征,以及对大数据环境的叙述,我们将更加明确地认识到数据的出现和发展,了解到数据的各种类型,并且对大数据时代有一个更客观的认识。

一、观察事实的结果——数据

在认识数据的开始,我们要了解到数据从哪里来?如何来?在本小节中,我们可以了解到“数据”的概念从最初始的计算机术语到日渐宽泛的定义这一发展过程,以及数据和信息的区别和联系。



（一）计算机的数据

1. 数据的起源

提起数据的起源,我们第一反应就是数据一定和计算机有关。20 世纪被称作第三次科技革命重要标志之一的就是计算机的发明与应用。在计算机系统中,数据是以二进制信息单元 0 和 1 的形式表示的,可以把它当作一个非常微小的开关,用 1 表示“开”,0 表示“关”。我们在计算机上看到的一切数字、文字、图形、图像等信息,都是计算机系统用 0 和 1 两个数的不同编制长度和顺序实现的。

2. 计算机的数据是什么

可以这样说,只要是可输入计算机,并被程序处理的符号、介质、模拟量等,都是计算机中的数据。计算机中的数据是一个广义的概念,包括各种各样的字母、数字符号的组合、语音、图形、图像等。数据也是组成地理信息系统的基本要素,种类很多。计算机对于数据的记录,慢慢取代了人们口口相传、书本文字等一系列记录的方法,成为目前便捷的、大容量的载体,也是较简单、较长久的记录方式。

3. 计算机与人类

有人会问,计算机会代替人类吗? 计算机会反过来控制人类吗? 这仿佛是计算机智能出现以来,我们不断在发问和畅想的一个问题。对于计算机代替了人脑记忆,总有不同的声音。有人认为计算机的出现将使人类变得懒惰、愚钝,对于计算机的依赖,会让我们记忆越发退化。有人认为,计算机是人类的工具,可以准确地记住繁杂的、海量的数据,这本身就简单的人脑和书本所能企及的。计算机记录数据在一定程度上补充了人脑记录数据的缺陷,达到了更好的客观性与永久性。

这个争论没有结果,也没有阻挡计算机科学的飞速发展。随着互联网的飞速发展,计算机将记录数据,互联网将共享数据,计算机中的海量数据成就了这个信息大爆炸的时代,也迎来了大数据的时代。

（二）记录事实就是记录数据

1. 数据是客观的记录者

一切主观的表达和记录总是充满了个人色彩,每个人眼中都只会看到自己眼中的事实。而数据总扮演着最理智的记录者,所谓事实胜于雄辩,而记录数据就是在记录事实、表达事实。数据给了许多词语一个否定,例如,“好像”“大概”“很严重”“差不多”,真实的数据是最真实、客观的记录者。

2. 用数据体现新闻

以新闻领域为例,新闻讲事实而不讲观点,而直接的数据就是事实的搬运工和新闻的表达者。例如,第四届中国数据新闻大赛中南京师范大学团队的作品《从“超生”到“低生”,数说中国人口困境》。对于人口这一话题,最显而易见的表达形式就是基于数值数据的总结分类和对比。在作品中,对于中国人口老龄化问题,用数值数据构建的簇状柱形图—折线图表现出近 28 年来老年人口的数据变化(见图 1-1),直接明了地展现出老龄化问题的发展趋势



及现状。用数据表现出了老年人数量多了多少,老年人人口占比高了多少,整体表现出社会中确实出现了老龄化问题,这要比用嘴去说“我们国家的老龄化越来越严重了”更加具体,也更加具有可信度和准确性。

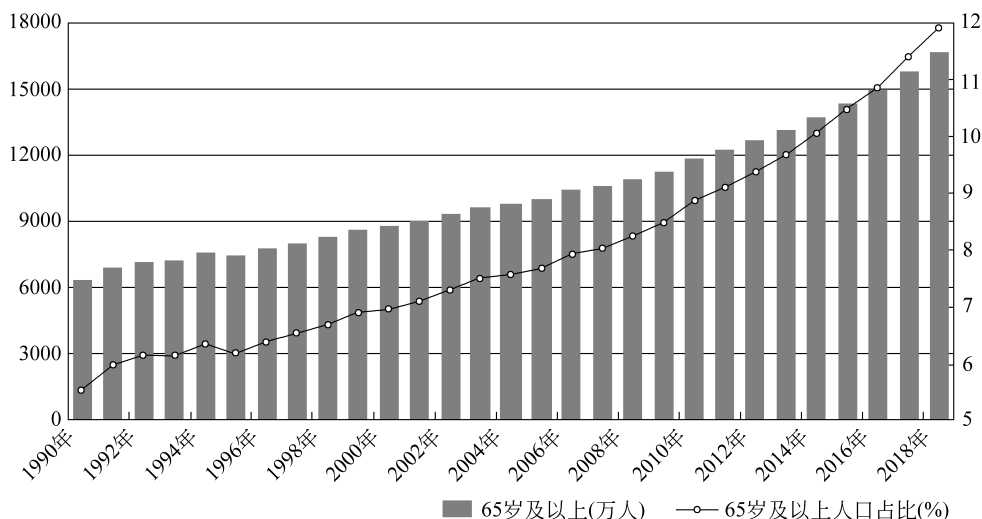


图 1-1 1990—2018 年中国老年人口数及占比情况

3. 数据带来的深层思考

用数据分析原因,例如,下面的作品用数据分析曾经要抑制生育的“超生”人口大国如何在几十年间进入“低生”困境?影响社会生育的主要原因是什么?作品指出“低生”的原因是房价压力,那为什么是房价压力呢?如何证明房价压力是主要原因呢?图 1-2 中各大城市的房价与收入的对比排名就清晰地表现出了生育的主力军——年轻人的城市生存压力,进而影响生育意愿,造成了社会“低生”这一人口现象。

2017年城市房价排名及工资水平

	平均房价 (元/m²)	平均工资 (元/月)
北京	67951	9942
上海	48384	9802
深圳	46879	8892
厦门	44235	7452
天津	26687	6733
广州	26453	7996
南京	26127	7342
石家庄	22934	7608
杭州	22934	7608
福州	21979	7073
苏州	15728	7548
武汉	15716	6769
青岛	15405	6651
济南	14800	6487
合肥	14495	6684
郑州	13574	6692
宁波	13259	7680
东莞	12242	7862
大连	10263	6669
南昌	9824	6631



北京 房价工资比**6.83**

不吃不喝**20**年



深圳 房价工资比**5.27**

不吃不喝**15**年



南京 房价工资比**3.56**

不吃不喝**10**年



大连 房价工资比**1.55**

不吃不喝**4.5**年

图 1-2 2017 年城市房价排名及工资水平



记录数据就是记录事实,新闻是记录者,是对事实的搬运和整理,它不是评论者,也不是解决者。这则数据新闻就很好地做到了用数据记录事实,用数据表现事实这一点。不只是新闻领域,在任何时间、地点,用数据说话都是表达事实的最好方法。

(三) 数据与信息区别与联系

我们已经知道,数字、文字是数据,图表、图像也是数据,这样看来数据好像是毫无意义地大量堆积,显得毫无用处。那要怎样数据才能变得有用、有意义呢?这就好像搭积木,一个一个的小积木块看起来乱七八糟、毫无美感,但当有人有目的地去排列和搭建它,一个个积木块组合在一起,它就有成为城堡的可能。数据也是这样,不是我们记录下来一堆堆的数据,这个工作就结束了。我们还需要有目的、有逻辑地去清洗和加工它,去分析、去组合、去创造、去探索出数据可以表达的问题和意义。

当数据具有了意义,它就有了一个新的名字——信息。我们总说,今天是一个信息大爆炸的时代,我们刷微博,看推送信息,乐此不疲地接收着大量与自己有关的或无关的信息。我们也总说,是互联网缔造了这个信息大爆炸的时代。确实,是互联网把信息传递到千家万户,让我们可以足不出户而知天下事,它开辟了信息共享的新纪元。但是,本质上是数据成就了这个信息大爆炸的时代。人们把大量的数据加工成信息,通过互联网这个载体,传播到世界各地。互联网是传播数据、传播信息的载体,但制造信息的本质是数据。

二、数值数据与多样化数据

在了解到“数据”的发展变化后,我们认识到生活中的一切都与数据有关。那么如何把数据进行简单分类?不同的数据类型有什么区别?大量的数据就是“大数据”吗?以数据为基础的大数据时代究竟如何?本小节将就“数据的分类”和“大数据时代”两个问题进行探讨,学习对数据简单的分类,从而对数据影响有一个客观认识。

数据有着多种多样的形式。最直接的数据形式当然就是数据数值,它是一个具体的数。除此之外,模拟量也是数据,语音也是数据,图形、表格也是数据。下面认识一下数据多种多样的形式。

(一) 数值数据

数值数据是表示数量、可以进行数值运算的一种数据类型。在计算机编程语言中,数值型数据由数字、小数点、正负号和表示乘幂的字母 E 组成,数值精度达 16 位。按存储、表示形式与取值范围不同,数值型数据又分为多种不同类型,如数值型、浮点型(单精度型、双精度型)和整型等。

1. 数值数据的固定性

数值数据是简单明了的一个数字定量。在计算机中,数值数据用二进制表示。数值数据在生活中也是非常常见的一种数据。例如,一个青少年的身高体重,某人一次考试的成绩等,这些都是基本的一个数值数据。而当搜集了一组青少年的身高体重,某人一个月考试的成绩,我们就得到了一组数值数据。当同一主题的数值数据有了一定的量,我们就可以进行数据的分析。根据一组身高的数值数据,可以推算出青少年的身高水平。根据某人一个月的成绩,可以得出某人近一个月的学习水平。这就是找到了一定量数值数据的规律和变化,从而实现数值数据从一个简单的数值单位到有意义的数据的转换。最终可以达到产出有意

义的信息的目的。

2. 国家数据

不限于日常小数据信息的处理,如图 1-3 所示国家数据网站关于全国全年粮食产量的数值数据统计(国家数据是一个国家各方面数据信息公开的平台)。通过对粮食产量数值数据的记录分析,可以在一定程度上得出一个国家第一生产力的水平和一年的变化。国家数据包括了各行业、产业月度、季度和年度的数据记录,也涵盖了各个省份、地区的各种数据。通过对数据检索,可以简单宏观地得出一个产业或地区的发展现状。可见,数值数据对于一个国家、一个社会的发展和数据分析是至关重要的。



指标	地区	数据时间	数值	所属栏目	相关报表
①粮食产量(万吨)	全国	2019年	66384.34	年度数据	相关报表
①粮食产量(万吨)	全国	2018年	65789.22	年度数据	相关报表
②人均粮食产量(公斤/人)	全国	2019年		年度数据	相关报表
②人均粮食产量(公斤/人)	全国	2018年		年度数据	相关报表
③秋粮产量(万吨)	全国	2019年	49597.00	年度数据	相关报表
③秋粮产量(万吨)	全国	2018年	49049.18	年度数据	相关报表
④粮食单位面积产量(公斤/公顷)	全国	2019年		年度数据	相关报表
④粮食单位面积产量(公斤/公顷)	全国	2018年	5621.17	年度数据	相关报表

图 1-3 全国全年粮食产量(来源:国家数据)

(二) 非数值型数据

除了最典型的数值数据外,我们还可以把其他数据统称为非数值型数据,也就是多样化的数据类型。多样化的数据包括声音、文字、图形、图像等多种表现方式。对数值型数值,我们可直接用算术方法进行汇总和分析,而对其他类型的数值则需特殊方法来处理。

1. Excel 中的文字数据

以最典型的文字数据为例,在 Excel 中的数字数据包含数字、符号、括号、除号、美元符号、百分号、小数点和科学记数符号等。在单元格中可以输入数值、负数、日期、时间等多种数值数据。数值数据可以直接在工作表中进行加、减、乘、除等数值运算。而对于文字数据,Excel 中的文本是字符、数字及特殊符号的组合,不能进行加、减、乘、除等运算,但可以进行连接对齐等数据操作。

2. 多样化数据的应用

除了文字数据,非数值数据还有多种类型。例如图像和声音,语音录音等编码属于声音数据,来源于相机拍摄的外界图像属于图像数据。这些非数值型数据不同于数值型数据的固定性,它们具有多样的灵活性,所以更适合服务于教学、创作或语言等能动性较强的领域。



非数值数据处理属于计算机应用领域,在计算机互联网的发展中,越来越多的智能功能也需要非数值数据的支撑,如模式识别、情报检索、人工智能、数学定理证明、语言翻译、计算机辅助教学等,都需要计算机编制多样化数据达到功能实现。

(三) 科技成就大数据

我们的生活好像被窥探?当你在某个电商平台购买了一个台灯后,你会发现购买成功的那一刻并不是结束。接下来,打开短视频平台,你会看到台灯视频广告;看某篇文章,你会发现台灯软文推荐;用的浏览器会有一条一条的台灯信息被推送。不只是台灯,只要与其相关的各种灯光家装等都会被推送,都会从各个平台进入你的视野。这些都是大数据在“分析”你。

1. 大数据是什么

我们说过,这个信息爆炸的时代也是数据爆炸的时代。到底什么是大数据?大量的数据就是大数据吗?数据的累积就是大数据吗?在互联网趋势下,大数据横空出世,影响着我们的方方面面,让我们不得不去认识一下“大数据”到底是什么?

在《开讲啦》的一期节目中,中国科学院院士梅宏给我们带来了一场演讲“大数据时代,你准备好了吗”,预见了大数据时代。根据 IDC(互联网数据中心)的估算,我们可以了解到近年数据量的极速增长。2003 年,全球的数据量是 500 万 TB。2009 年,达到了 0.8ZB。2012 年,达到了 2.8ZB。预计到 2025 年,将达到 163ZB 的庞大数据量。对于 2030 年,更是预测了一个惊人的数据值——2500ZB。书本文字、个人信息、科学建模都是数据。那么怎样才算是大数据呢?梅宏也谈道:“大数据的本质体现在哪儿?体现在数据挖掘的深度和应用的广度,也就是跨界融合的广度,这个才是所谓的大数据。”我们说到,加工过的数据才是有意义的数据。大数据体现在深度、广度的范围性覆盖上。对于行业的影响力,对于各个领域的波及程度和范围,才是衡量大数据的标准。随着信息化进程的不断发展,大数据引领的“数字经济”“数据时代”已经来临,并且已经成为一个国家、一个社会必不可少的竞争力和发展引擎。

2. 大数据时代的隐私问题

凡事有利必有弊,大数据带来了便利,也带来了困扰。大数据在窥探你的隐私?以电商网购为例,理论上,购买者的个人数据都是个人隐私。但现实是我们不得不交出一些个人隐私,才能换取这样的服务。比如,我们网购一个台灯,就在主动将自己的个人数据上传到电商平台和网络世界。网络平台利用大数据分析我们的个人数据,我们才会收到一系列的台灯信息推送。当我们的姓名、住址、年龄、工作等一系列个人隐私暴露在网络世界后,我们无疑是惶恐不安的。一个新的科学技术是没有错的,重要的是如何去约束、去控制、去应用这一新技术,去更好地服务于大众与社会。大数据时代下的网络安全与个人隐私还要依赖于法律制度的保护。不仅是大数据技术,从法律规则上制止技术的滥用对于每一项技术发明都是十分必要的。

3. 大数据的未来发展

大数据仍在发展的道路中,目前它的推送还不够智能,隐私保护也不够完善,但是一系列的问题并没有阻止它的崛起。越来越多的行业依赖大数据,越来越多的学者研究大数据,越来越多的技术人员完善大数据,它影响着我们生活的柴米油盐,也影响着一个国家社会的



发展前途。在技术日新月异的同时,我们的法律制度也需要快速跟进,技术和制度的共同发展才能更好地把科技应用到社会中。

三、数据的规律性与变异性

每一个事物都有它自己的独特之处。那么,数据有什么特点呢?本节介绍数据的两大特点:规律性和变异性。这看起来是自相矛盾的两种特性,在下面的学习中,我们就会看到这两种特性在数据中的并存和发展。

数据之所以有意义,是因为我们可以找出数据的规律特点进而对事物进行分析预测。反之,数据之所以需要二次加工,是因为数据本身是杂乱无章的、充满不确定性的。了解数据的规律性,也接受数据的复杂性和不确定性,是我们利用好数据的根本。

(一) 杂乱无章的数据

我们在网络上直接搜索得到的数据,往往已经经过了分组加工等行为。仅在记录数据这一步,得到的最原始数据往往是杂乱无章的。例如,通过生活数据预测用户的餐饮喜好。可是,一个人生活的方方面面都会或多或少地影响到其饮食,如他的睡眠时间影响到是否吃早饭,体重水平影响到是否在刻意改变饮食习惯,工资收入、出行路线、身体健康、家乡所在地等因素,我们会搜集到一大堆物理数据。一眼看去,这些数据好像互相没有联系又互相影响,也许还会出现结论冲突。这些数据就是杂乱无章的数据。

那么,这些数据就是无用的吗?显然不是。饮食是我们调查的主题,围绕它展开的一系列数据搜集当然是有目的的。可是为什么我们最终得到了杂乱无章的一堆数据?这时候就需要进行数据分析。当我们从这一大堆数据中,删去无用数据,找到重点参考值,计算出平均影响度,进行一系列的数据加工。在这之后,我们就可以推断主要原因,进而进行合理预测了。

(二) 数据变异的不确定性

我们总说,数据是有规律的,所以才有意义。其实,这种规律只是从整体的角度看得出来的宏观的规律。一组数据是来源于调查和记录的,所以它是不会完全按照非常准确的规律排列的。当我们在分析数据时,不但要分析多数数据的规律性,也应该注意到数据变异的不确定性。

1. 个别数据的变异

个别数据是有特殊性与变异性的。就像上面提到的通过生活数据预测用户的餐饮喜好这一例子,总会搜集到特殊的、变异的数据。也许今天天气不好,也许生病了,也许家里有意外情况,现实中总有各种在我们预判之外的事件发生,从而影响到我们对其饮食的调查分析。但只要这种特殊数据没有对整体数据有剧烈的影响,它就不会影响我们对整体数据的分析。

2. 整体数据的变异

如果将个别数据的变异称为特殊情况,忽略不计,那么对于整体数据的变异,就需要格外被关注了,数据的变化规律背后往往是真正值得关注的特殊信息。例如,我们调查一个超市的营业额,一开始的数据是规律且稳定的,后期却出现了不小的涨幅或下降,数据发生明显的整体变异。这个时候,我们就可以利用发散思维去分析和调查,为什么出现了变异?



是政策支持引起的涨幅?还是特殊疫情等公共卫生引起的下降?虽然我们有目的、有预想地去调查,但是现实往往是瞬息万变的,采集到的数据也许与我们预想的背道而驰。但也正是因为调查数据的这种未知性与不确定性,数据分析才更加值得我们去探索。

(三) 大量数据的内在规律

科学的数据分析往往需要大量数据的支撑。当数据达到一定的量,才更有说服力,预想结论也更加科学合理。大量的数据不仅可以分析粮食产量、营业额等这些数值型的领域,也完全可以对人文社科领域进行分析和预判,这也就是所谓的人工智能。

2016年,谷歌人工智能“阿尔法围棋”(AlphaGo)战胜围棋冠军李世石,引发了全世界对于“人机大战”“人工智能”的高度关注。阿尔法围棋为什么可以赢?就是因为它有大量的数据作为支撑。围棋冠军也是在一场场的比赛中进行经验获取与技术提升,这可以看作人脑对于数据的储存分析。阿尔法围棋也是一样,它把围棋对战中的所有行为都看作一个数据并进行记录和分析,通过训练形成一个策略网络。在对战时,将棋盘上的局势作为输入信息,并对所有可行的落子位置生成一个概率分布,从而分析出最优结果。人类会受比赛状态的影响,但人工智能机器人不会,它只是在一次又一次地把数据录入并计算分析。

寻找大量数据的内在规律,就是像阿尔法围棋一样,以量取胜。把所有的数据规律和概率分布都考虑到,怎么会不赢呢?

第二节 数据的商业价值

本节重点讲述数据的加工和利用。从记录数据、分析数据、产出价值,到建设庞大的主题性数据库、打造数据运营和服务,再到用数据分析用户画像和商业前景。通过本节的学习,我们将对数据的作用有更深层次的认识,并且掌握一定加工数据的能力。

一、数据的搜集与价值转换

搜集数据和分析数据是发挥数据价值的前提。对数据从收集到价值转换这一过程,进行一个简单的认识和学习,是我们利用数据的第一步。

数据在哪里?吃饭、看书、听音乐都是数据。各大商业公司谷歌、亚马逊、Facebook、百度、阿里巴巴等均陷在其中而不能自拔。

中国领先的云计算方案和服务供应商浪潮集团在京发布基于全新技术架构的大数据处理平台——云海大数据一体机,面向产业界正式发布大数据整体战略,渗透到当今社会互联网、金融、电信等多个行业。大数据将带来“数字经济”的时代。IDC 报告指出,截至 2011 年底,全球新增的数据量已达到了 1.8 万亿吉字节(GB),未来十年还将增长 50 倍,迅速积累的海量数据蕴含着重大的商业价值和社会价值,通过挖掘海量数据,公司的决策、运行会建立在更加科学的基础上,失误更少,效率更高。

第四届中国数据大赛的作品之一——《2022 北京冬奥会是笔好“投资”吗?》就很好地利用了数据,分析利益价值的可行性。我们将以此为例,阐述如何从数据中提取价值。

(一) 搜集大量的数据

要用数据,首先就要先有数据。数据从哪里来?如何搜集到大量的数据?商业的目的

是利润,所以最重要的数据就是收入与支出。从各个方面记录影响收入与支出的数据,就是最有用的商业数据。所以,对于这则关于冬奥会利益价值的新闻,最必不可少的数据就是收入与支出数据。申办奥运会对于主办城市是不是一项“赔本的买卖”?我们需要看的是历年奥运会的收支账。

从局部看一场赛事的利益分析,就是赛事准备投入与回报的对比。所以对于奥运会赛事,我们需要搜集的数据有场馆建设、人力成本、服务支出等赛事投入数据与赞助商、门票等直接收入数据(见图 1-4)。



图 1-4 国家奥组委支出收入图

从长远看奥运会的效益分析,不应该只包括赛事本身的收支账。奥运村建设、环境治理、城市基础设施完善,这些要为奥运准备的环境与服务,也是奥运的支出账单。旅游知名度、场馆和基础设施后期利用、国内品牌的广告度,这也是奥运带来的隐形财富。

所以对于奥运会数据的搜集,其涵盖的部分不应只局限于一场赛事的经济投入。仅仅以目前收支来判断一个赛事或项目的利益可行性是不可取的,其带来的文化影响力等隐形收入的数据也具有很重要的参考价值。那么我们必须找到交通、环保、各大国货等社会每个方面的数据吗?当然不是。奥运会对一个国家社会的影响是方方面面的,可以说在多个领域的的数据量都会受到奥运的影响。从时间纵向看,对于奥运经济,我们首先可以考虑的就是奥运场馆的后期利用性和基础设施的长久价值;从横向覆盖范围看,一场奥运会对于城市的交通、环保、服务业等很多方面或多或少都有影响,我们只要找到典型领域进行着重调查即可。多方面、全方位地去搜集数据,才可以达到一定的数据量和范围广度,利于对数据的分析与预测。

(二) 学会清洗数据

当完成了搜集记录这一部分,我们会得到与主题相关的一系列大量的数据。接下来,就要对原始数据进行加工,让数据变得“有用”。数据分析首先要进行数据清洗,主要的过程就是去除无用数据,进行数据分组、连接数据和填补遗漏数据。

第一,去除无用数据。由于在搜集我们希望数据量多且涵盖广,所以在进行数据整理时,往往会发现许多无用的或者说参考价值较弱的的数据。这些数据可以在一开始的数据分析中被摒弃掉。比如对于“奥运会隐形财富”这一举例,可能前期我们收集了奥运会对于旅



游业、房地产、基础工程等多个领域的影响数据,但房地产远远没有旅游业和基础设施这些领域受奥运会影响大,或者不够成为“奥运会遗产”的代表性支撑,数据充足的情况下,我们就可以对其进行删减。

第二,进行数据分组。数据被记录之后,仍然是相对分散的。我们可以把这些分散的数据进行简单的分组,把相同类型或相同单位的数据放在一起,形成一个个的数据小组。分组的标准依情况而定,时间、地区、三大产业都可以作为分组的基本依据。根据自己的调查需求把数据分组,也是初步完成了数据从零散到整体的整理。

第三,连接数据。在分组后,我们把相同类型或者关联性强的数据放在了一起,得到了一组一组的数据。我们可能简单地把总产值放在一组,把时间线放在一组,把地区分布放在一组。但是它们依然是简单分散的,并没有成为一个整体,这时候我们就可以进行合理的数据连接。比如我们把总产值、时间线和地区分布放在一起,进行数据连接,就可以得出季度、地区、占比等多种数据信息。把数据连接起来,就可以直观地看出事物之间的许多联系和有价值的信息,是数据分析很重要的一步。

第四,填补遗漏数据。在进行数据的加工分析中,我们可以找到无用的、参考价值不大的数据,同时可以发现遗漏了的需要数据支撑的部分。在搜集中,我们有目的地寻找记录里的数据,基于的是我们的预判和推测。到了数据加工时,有可能发生一定的偏差,一些必要的数据可能并没有被采集。所以,填补遗漏的数据,在加工中进行必要数据的再次收集,可以让我们的数据分析更加全面完整。

(三) 寻找数据规律

如何从大量的数据中得出规律? 首先,看数据的整体和变化趋势。从整体多角度地看数据,往往可以得出一个整体现状或规律变化。仍以“奥运会经济”为例,可以看图 1-5,图中是国家奥组委收入与支出的对比数据,用支出作为单位 1。2000 年奥组委收入支出的对比值为 1:1.16。2002 年的对比值为 1:1.02。同样,后面的六个年份中,也是收入略高于支出。所以从这组数据来看,可以发现收入一直是略高于支出这个规律。

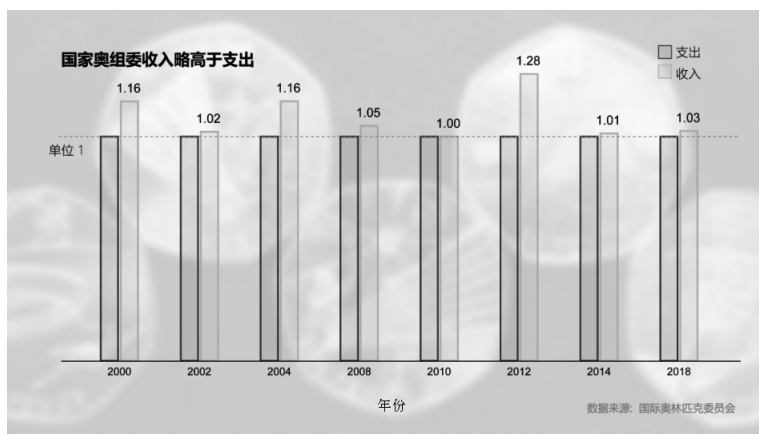


图 1-5 国际奥委会收入与支出比

其次,看数据特殊的地方。我们可以分析最大值和最小值,或者变量大、具有特殊性、具有典型性的数据量。如图 1-6 所示,图中是各城市申办奥运会的预算数据,每格代表 30 亿