第3章

大数据可视化

【导读案例】

南丁格尔“极区图”

弗洛伦斯·南丁格尔(1820年5月12日—1910年8月13日,图3-1)是世界上第一个真正意义上的女护士,被誉为现代护理业之母,5.12国际护士节就是为了纪念她,这一天是南丁格尔的生日。除了在医学和护理界的辉煌成就,实际上,南丁格尔还是一名优秀的统计学家——她是英国皇家统计学会第一位女性会员,也是美国统计学会的会员。据说南丁格尔早期大部分声望都来自其对数据清楚且准确的表达。

图3-1 南丁格尔

南丁格尔生活的时代,各个医院的统计资料非常不精确,也不一致,她认为医学统计资料有助于改进医疗护理的方法和措施。于是,她编著的各类书籍、报告等材料中使用了大量的统计图表,其中最为著名的就是极区图(Polar Area Chart),也叫南丁格尔玫瑰图(图3-2)。

南丁格尔发现,战斗中阵亡的士兵数量少于因为受伤却缺乏治疗的士兵。为了挽救更多的士兵,她画了《东部军队(战士)死亡原因示意图》(1858年)。

这张图描述了1854年4月—1856年3月期间士兵的死亡情况,右图是1854年4月—1855年3月,左图是1855年4月—1856年3月,用蓝、红、黑三种颜色表示三种不同的情况,蓝色代表可预防和可缓解的疾病治疗不及时造成的死亡、红色代表战场阵亡、黑色代表其他死亡原因。图表各扇区的角度相同,用半径及扇区面积来表示死亡人数,可以清晰地看出每个月因各种原因死亡的人数。显然,1854—1855年,因医疗条件而造成的死亡人数远远大于战死沙场的人数,这种情况直到1856年初才得到缓解。南丁格尔的这张图表以及其他图表“生动有力地说明了在战地开展医疗救护和促进伤兵医疗工作的必要性,打动了当局者,增加了战地医院,改善了军队医院的条件,为挽救士兵生命做出了巨大贡献”。

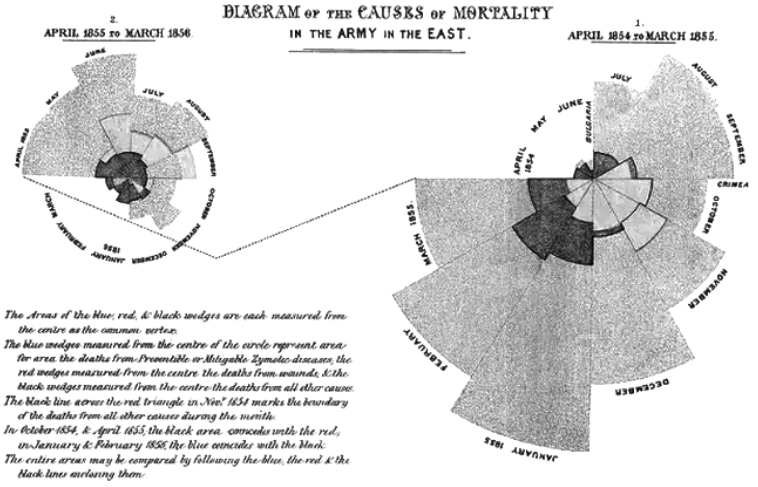

图 3-2 南丁格尔“极区图”

南丁格尔的“极区图”是统计学家对利用图形来展示数据进行的早期探索，南丁格尔的贡献充分说明了数据可视化的价值，特别是在公共领域的价值。

图 3-3 是社交网站(脸书 vs.推特)对比信息图，是一张典型的南丁格尔玫瑰图(极区图)。极区图在数据统计类信息图表中是常见的一类图表形式。

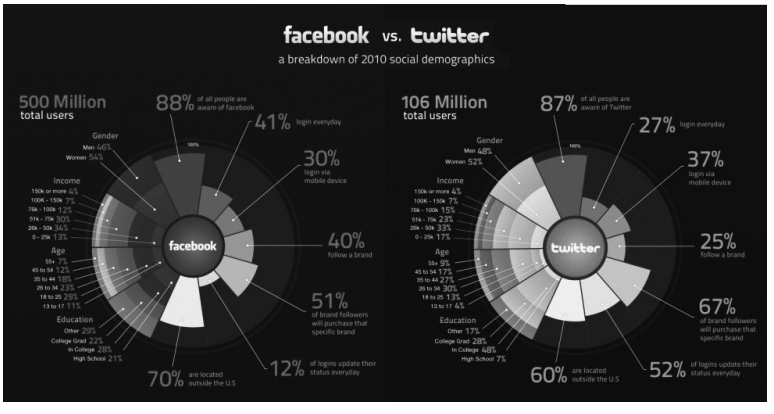

图 3-3 极区图：脸书 vs.推特

阅读上文，请思考、分析并简单记录：

(1) 简述你看到过且印象深刻的数据可视化的案例。

答：_____

_____

_____

_____

(2) 你之前知道南丁格尔吗？南丁格尔玫瑰图还有什么名字？

答：_____

(3) 发展大数据可视化,那么传统的数据或信息的表示方式是否还有意义?请简述你的看法。

答:

(4) 请简单记述你所知道的上一周发生的国际、国内或者身边的大事。

答:

数据与可视化

3.1 数据与可视化

数据是什么?大部分人会含糊地回答说,数据是一种类似电子表格的东西,或者一大堆数字。有点儿技术背景的人会提及数据库或者数据仓库。然而,这些回答只说明了获取数据的格式和存储数据的方式,并未说明数据的本质是什么,以及特定的数据集代表什么。

数据不仅仅是数字,要想把数据可视化,就必须知道它表达的是什么。事实上,数据是现实世界的一个快照,会传递给我们大量的信息。一个数据点可以包含时间、地点、人物、事件、起因等因素,因此,一个数字不再只是沧海一粟。可是,从一个数据点中提取信息并不像一张照片那么简单。你可以猜到照片里发生的事情,但如果对数据心存侥幸,认为它非常精确,并和周围的事物紧密相关,就有可能曲解真实的数据。你需要观察数据产生的来龙去脉,并把数据集作为一个整体来理解。关注全貌,比只注意到局部更容易做出准确的判断。

通常,在实施记录时,由于成本太高或者缺少人力,或二者皆有,人们不大可能记录下一切,而只能获取零碎的信息,然后寻找其中的模式和关联,凭经验猜测数据所表达的含义,数据是对现实世界的简化和抽象表达。当人们可视化数据的时候,其实是在将对现实世界的抽象表达可视化,或至少是将它的一些细微方面可视化。可视化是对数据的一种抽象表达,所以,最后你得到的是一个抽象的抽象。这并不是说可视化模糊了你的视角。恰恰相反,可视化能帮助你从一个个独立的数据点中解脱出来,换一个不同的角度去探索它们。

数据和它所代表的事物之间的关联既是把数据可视化的关键,也是全面分析数据的关键,同样还是深层次理解数据的关键。计算机可以把数字批量转换成不同的形状和颜色,但是人类必须建立起数据和现实世界的联系,以便使用图表的人能够从中得到有价值

的信息。数据会因其可变性和不确定性变得复杂，但放入一个合适的背景信息中，就变得容易理解了。

3.1.1 数据的可变性

本节以美国国家公路交通安全管理局发布的公路交通事故数据为例，来了解数据的可变性。

2001—2010 年，美国国家公路交通安全管理局发布的数据显示，全美共发生 363 839 起致命的公路交通事故。这个总数代表着逝去的生命(图 3-4)。

然而，除了安全驾驶之外，从这个数据中你还了解到了什么？美国国家公路交通安全管理局提供的数据具体到了每一起事故及其发生的时间和地点，我们可以从中了解更多的信息。

如果在地图中画出 2001—2010 年间全美发生的每一起致命的交通事故，用一个点代表一起事故，就可以看到事故多集中发生在大城市和高速公路主干道上，而人烟稀少的地方和道路几乎没有事故发生过。这样，这幅图除了告诉我们对交通事故不能掉以轻心之外，还告诉了我们关于美国公路网络的情况。

观察这些年里发生的交通事故，人们会把关注焦点切换到这些具体的事故上。图 3-5 显示了每年发生的交通事故数，所表达的内容与简单告诉你一个总数完全不同。虽然每年仍会发生成千上万起交通事故，但通过观察可以看到，2006—2010 年间事故呈显著下降趋势。

图 3-4 2001—2010 年全美公路致命交通事故总数

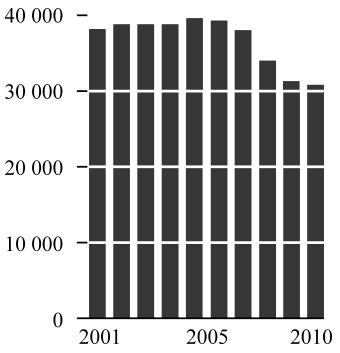

图 3-5 每年的致命交通事故数

从图 3-6 中可以看出，交通事故发生的季节性周期很明显。夏季是事故多发期，因为此时外出旅游的人较多。而在冬季，开车出门旅行的人相对较少，事故就会少很多。每年都是如此。同时，还可以看到 2006—2010 年事故呈下降趋势。

如果比较那些年的具体月份，还有一些变化。例如，在 2001 年，8 月份的事故最多，9 月份相对回落。2002—2004 年每年都是这样。然而，2005—2007 年，每年 7 月份的事故最多。2008—2010 年又变成了 8 月份。另一方面，因为每年 2 月份的天数最少，事故数也就最少，只有 2008 年例外。因此，这里存在着不同季节的变化和季节内的变化情况。

我们还可以更加详细地观察每日的交通事故数，例如看出高峰和低谷模式，可以看出

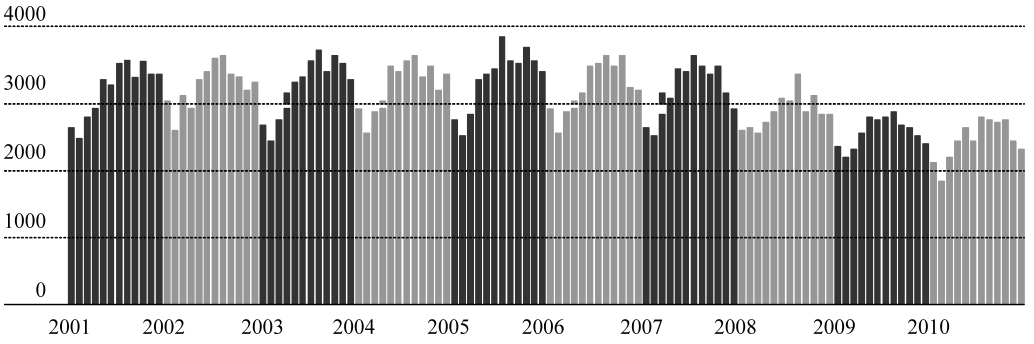

图 3-6　月度致命交通事故数

周循环周期,就是周末比周中事故多,每周的高峰日在周五、周六和周日间的波动。可以继续增加数据的粒度,即观察每小时的数据。

重要的是,查看这些数据比查看平均数、中位数和总数更有价值,那些测量值只是告诉了你一小部分信息。大多数时候,总数或数值只是告诉了你分布的中间在哪里,而未能显示出你做决定或讲述时应该关注的细节。

一个独立的离群值可能是需要修正或特别注意的。也许在你的体系中,随着时间推移发生的变化预示有好事(或坏事)将要发生。周期性或规律性的事件可以帮助你为将来做好准备,但面对那么多的变化,它往往就失效了,这时应该退回到整体和分布的粒度来进行观察。

3.1.2　数据的不确定性

数据具有不确定性。通常,大部分数据都是估算的,并不精确。分析师会研究一个样本,并据此猜测整体的情况。每天你都在做这样的事情,你会基于自己的知识和见闻来猜测,尽管大多数的时候你确定猜测是正确的,但仍然存在着不确定性。

例如,笔记本电脑上的电池寿命估计会按小时增量跳动;地铁预告说下一班车将会在10min内到达,但实际上是11min,预计在周一送达的一份快件往往周三才到。

如果数据是一系列平均数和中位数,或者是基于一个样本群体的一些估算,就应该时时考虑其存在的不确定性。当人们基于类似全国人口或世界人口的预测数做影响广泛的重大决定时,这一点尤为重要,因为一个很小的误差也可能会导致巨大的差异。

图 3-7　彩虹糖

换个角度,想象一下你有一罐彩虹糖(图 3-7),没法看清罐子里的情况,你想猜猜每种颜色的彩虹糖各有多少颗。如果你把一罐彩虹糖统统倒在桌子上,一颗颗数过去,就不用估算了,你已经得到了总数。但是你只能抓一把,然后基于手里的彩虹糖推测整罐的情况。这一把越大,估计值就越接近整罐的情况,也就越容易猜测。相反,如

果只能拿出一颗彩虹糖，那几乎就无法推测罐子里的情况。

只拿一颗彩虹糖，误差会很大。而拿一大把彩虹糖，误差会小很多。如果把整罐都数一遍，误差就是零。当有数百万个彩虹糖装在上千个大小不同的罐子里时，分布各不相同，每一把的大小也不一样，估算就会变得更复杂了。接下来，把彩虹糖换成人，把罐子换成城、镇和县，把那一把彩虹糖换成随机分布的调查，误差的含义就有分量多了。

如果不考虑数据的真实含义，很容易产生误解。要始终考虑到不确定性和可变性。这也就到了背景信息发挥作用的时候了。

3.1.3 数据所依存的背景信息

仰望夜空，满天繁星看上去就像平面上的一个个点。你感觉不到视觉深度，会觉得星星都离你一样远，很容易就能把星空直接搬到纸面上，于是星座也就不难想象了，把一个个点连接起来即可。然而，实际上不同的星星与你的距离可能相差许多光年。假如你能飞得比星星还远，星座看起来又会是什么样子呢？

如果切换到显示实际距离的模式，星星的位置转移了，原先容易辨别的星座几乎认不出了。从新的视角出发，数据看起来就不同了。这就是背景信息的作用。背景信息可以完全改变你对某一个数据集的看法，帮助你确定数据代表着什么以及如何解释。在确切了解了数据的含义之后，你的理解会帮你找出有趣的信息，从而带来有价值的可视化效果。

使用数据而不了解除数值本身之外的任何信息，就好比拿断章取义的片段作为文章的主要论点引用一样。这样做或许没有问题，但却可能完全误解说话人的意思。你必须首先了解何人、如何、何事、何时、何地以及何因，即元数据，或者说关于数据的数据，然后才能了解数据的本质是什么。

何人(who)："谁收集了数据"和"数据是关于谁的"同样重要。

如何(how)：大致了解怎样获取你感兴趣的数据。如果数据是你收集的，那一切都好，但如果数据只是从网上获取的，就不需要知道每种数据集背后精确的统计模型但要小心小样本，样本小，误差率就高，也要小心不合适的假设，比如包含不一致或不相关信息的指数或排名等。

何事(what)：还要知道自己的数据是关于什么的，应该知道围绕在数字周围的信息是什么。可以跟学科专家交流，阅读论文及相关文件。

何时(when)：数据大都以某种方式与时间关联。数据可能是一个时间序列，或者是特定时期的一组快照。不论是哪一种，你都必须清楚地知道数据是什么时候采集的。由于只能得到旧数据，于是很多人便把旧数据当成现在的对付一下，这是一种常见的错误。事在变，人在变，地点也在变，数据自然也会变。

何地(where)：正如事情会随着时间变化，它们也会随着城市、地区和国家的不同而变化：例如，不要将来自少数几个国家的数据推及整个世界。同样的道理也适用于数字定位。来自推特或脸书之类网站的数据能够概括网站用户的行为，但未必适用于物理世界。

为何(why)：最后，必须了解收集数据的原因，通常这是为了检查一下数据是否存在

偏颇。有时人们收集甚至捏造数据只是为了应付某项议程,应当警惕这种情况。

首要任务是竭尽所能地了解自己的数据,这样,数据分析和可视化会因此而增色。可视化通常被认为是一种图形设计或破解计算机科学问题的练习,但是最好的作品往往来源于数据。要可视化数据,必须理解数据是什么,它代表了现实世界中的什么,以及应该在什么样的背景信息中解释它。

在不同的粒度上,数据会呈现出不同的形状和大小,并带有不确定性,这意味着总数、平均数和中位数只是数据点的一小部分。数据是曲折的、旋转的,也是波动的、个性化的,甚至是富有诗意的。因此,你可以看到多种形式的可视化数据。

3.1.4 打造最好的可视化效果

当然,存在计算机不需要人为干涉就能单独处理数据的例子。例如,当要处理数十亿条搜索查询的时候,要想人为地找出与查询结果相匹配的文本广告是根本不可能的。同样,计算机系统非常善于自动定价,并在百万多个交易中快速判断出哪些具有欺骗性。

但是,人类可以根据数据做出更好的决策。事实上,拥有的数据越多,从数据中提取出具有实践意义的见解就显得越发重要。可视化和数据是相伴而生的,将这些数据可视化,可能是指导我们行动的最强大的机制之一。

可视化可以将事实融入数据,并引起情感反应,它可以将大量数据压缩成便于使用的知识。因此,可视化不仅是一种传递大量信息的有效途径,它还和大脑直接联系在一起,并能触动情感,引起化学反应。可视化可能是传递数据信息最有效的方法之一。研究表明,不仅可视化本身很重要,何时、何地、以何种形式呈现对可视化来说也至关重要。

通过设置正确的场景,选择恰当的颜色甚至选择一天中合适的时间,可视化可以更有效地传达隐藏在大量数据中的真知灼见。科学证据证明了在传递信息时环境和传输的重要性。

3.2 数据与图形

假设你是第一次来到华盛顿特区——美利坚合众国的首都,你很兴奋,激动地想参观白宫和所有的纪念碑、博物馆。从一个地方赶到另一个地方,为此,你需要利用当地的交通系统——地铁(图 3-8)。这看上去挺简单,但问题是:你如果没有地图,不知道怎么走,那么,即使遇上个好心人热情指点,要弄清楚搭哪条线路,在哪个站上车、下车,简直就是一场噩梦。不过,幸运的是,华盛顿地铁图(图 3-9)可以传达这些数据信息。

图 3-8 华盛顿地铁

地图上每条线路的所有站点都按照顺序,用不同颜色标记出来的,还可以在上面看到线路交叉的站点。这样,要知道在哪里换乘就很容易了。突然之间,弄清楚如何搭乘地铁变成

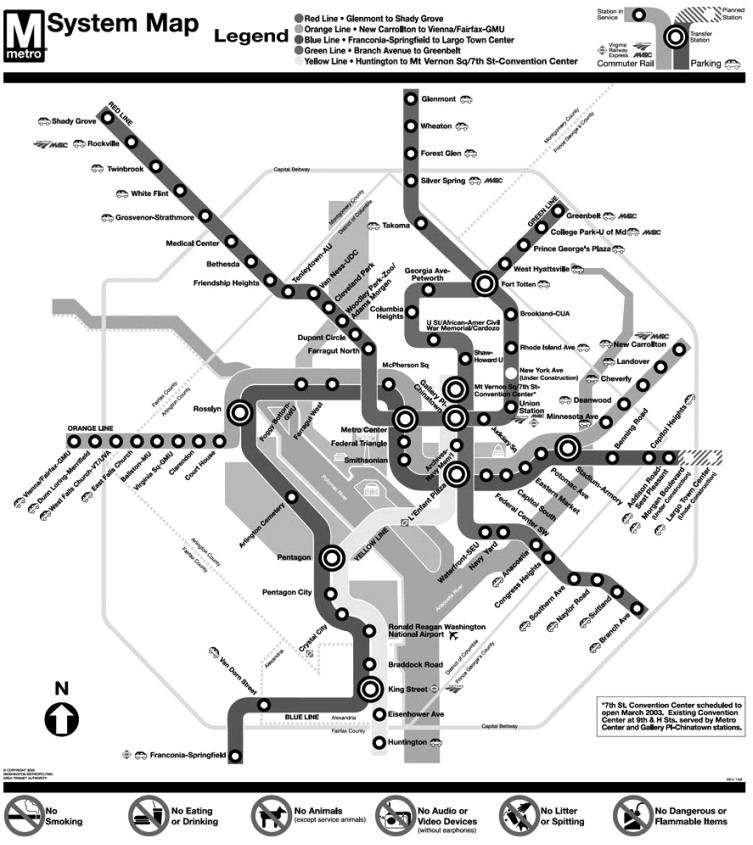

图 3-9 华盛顿地铁图

了轻而易举的事情。地铁图呈给你的不仅是数据信息,更是清晰的认知。

你不仅知道了该搭乘哪条线路,还大概知道了到达目的地需要花多长时间。无须多想,你就能知道到达目的地有 8 个站,每个站之间大概需要几分钟,因而可以计算出从你所在的位置到“航空航天博物馆”要花上 20 多分钟。除此之外,地铁图上的路线不仅标注了名字或终点站,还用了不用的颜色——红、黄、蓝、绿、橙来帮助你辨认。每条线路用的是不同的颜色,如此一来,不管是在地图上还是地铁外的墙壁上,只要你想查找地铁线路,都能通过颜色快速辨别。

将信息可视化能有效地抓住人们的注意力。有的信息,如果通过单纯的数字和文字来传达,可能需要花费数分钟甚至几小时,甚至可能无法传达;但是通过颜色、布局、标记和其他元素的融合,图形却能够在几秒钟之内就把这些信息传达给我们。

通过仔细阅读华盛顿地铁图,理清了头绪,你发现其实华盛顿特区只有 86 个地铁站。日本东京地铁系统包括东京地铁公司(Tokyo Metro)和都营地铁公司(the Toei)两大地铁运营系统,一共有 274 个站。算上东京更大片区的所有铁路系统,东京一共有 882 个车站(图 3-10)。如果没有地图,人们将很难了解这么多的站台信息。

3.2.1 数据与走势

在使用电子表格软件处理数据时会发现,要从填满数字的单元格中发现走势是困难的。这就是诸如微软电子表格软件(Microsoft Excel)和苹果电子表格软件(Apple Numbers)这类程序内置图表生成功能的原因之一。一般来说,人们在看一个折线图(图 3-11)、饼状图或条形图的时候,更容易发现事物的变化走势。

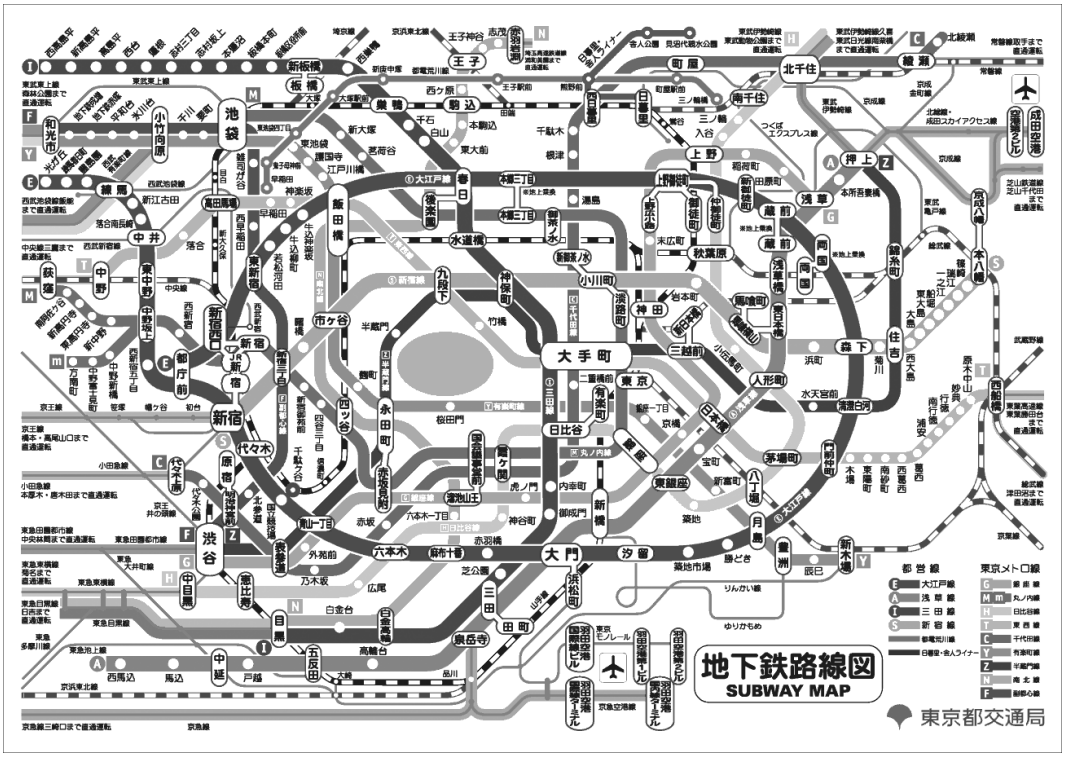

图 3-10 东京地铁图

制订决策时,了解事物的变化走势至关重要。不管是讨论销售数据还是健康数据,一个简单的数据点通常不足以呈现事情的整个变化走势。

投资者常常要试着评估一个公司的业绩,一种方法就是及时查看公司在某一特定时刻的数据。比如,管理团队在评估某一特定季度的销售业绩和利润时,若没有将之前几个季度的情况考虑进去,他们可能会总结说公司运营状况良好。但实际上,投资者没有从数据中看出公司每个季度的业绩增幅都在减少。表面上看,公司的销售业绩和利润似乎还不错,而事实上,如果不想办法来增加销量,公司甚至很快就会走向破产。

管理者或投资者在了解公司业务发展趋势的时候,内部环境信息是重要指标之一。管理者和投资者同时也需要了解外部环境,因为外部环境能让他们了解自己的公司相对于其他公司的运营情况如何。

在不了解公司外部运营环境时,如果某个季度销售业绩下滑,管理者就有可能错误地认为公司的运营情况不好。可事实上,销售业绩下滑的原因可能是由大的行业问题引起

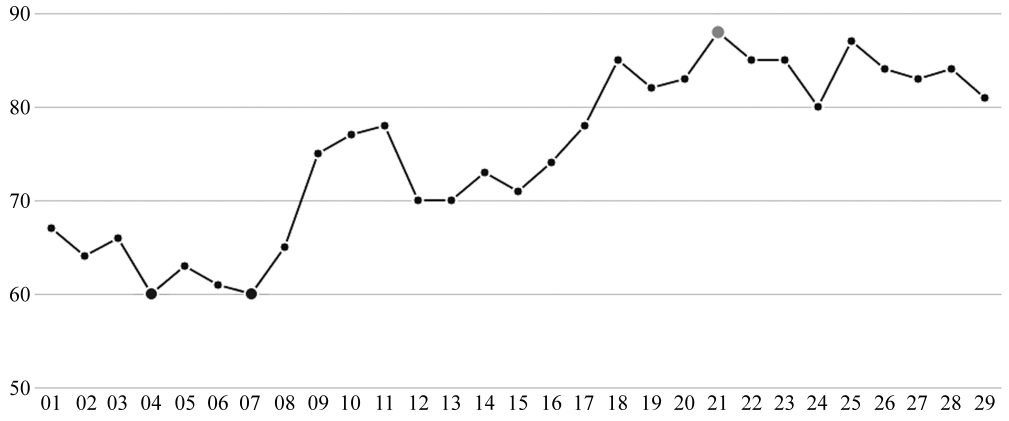

图 3-11 折线图示意图

的，例如，房地产行业受房屋修建量减少的影响，航空业受出行减少的影响等。

外部环境是指同行业的其他公司在同一段时间内的运营情况。不了解外部环境，管理者就很难洞悉究竟是什么导致了公司的业务受损。即使管理者了解了内部环境和外部环境，但要想仅通过抽象的数字来看出端倪还是很困难的，而图形可以帮助他们解决这一问题。

大卫·麦克坎德莱斯说："可视化是压缩知识的一种方式"。减少数据量是一种压缩方式，如采用速记、简写的方式来表示一个词或者一组词。但是，数据经过压缩之后，虽然更容易存储，却让人难以理解。然而，图片不仅可以容纳大量信息，还是一种便于理解的表现方式。在大数据里，这样的图片就叫作"可视化"。

地铁图、饼状图和条形图都是可视化的表现方式。乍一看，可视化似乎很简单。但由于种种原因，要理解起来并不容易。

首先，它很难满足人们希望将所有数据相互衔接并出现在同一个地方的愿望。其次，内部环境和外部环境的数据信息可能存储在两个不同的地方。行业数据可能存储在市场调查报告之中，而公司的具体销售数据则存储在公司的数据库中。而且，这两种数据的存储模式也有细微的差别。公司的销售数据可能是按天更新存储的，而可用的行业数据可能只有季度数据。

最后，数据信息不统一的表达方式也使我们难以理解数据真正想传达的信息。但是，通过获取所有这些数据信息，并将之绘制成图表，数据就不再是简单的数据了，它变成了知识。可视化是一种压缩知识的形式，因为看似简单的图片却包含了大量结构化或非结构化的数据信息。它用不同的线条、颜色压缩这些信息，然后快速、有效地传达出数据表示的含义。

3.2.2 视觉信息的科学解释

在数据可视化领域，爱德华·塔夫特被誉为"数据界的列奥纳多·达·芬奇"。他的一大贡献就是：聚焦于将每一个数据都做成图示物——无一例外。塔夫特的信息图形不仅能传达信息，甚至被很多人看作是艺术品。塔夫特指出，可视化不仅能作为商业工具发

挥作用,还能以一种视觉上引人入胜的方式传达数据信息。

在通常情况下,人们的视觉能吸纳多少信息呢?根据美国宾夕法尼亚大学医学院的研究人员估计,人类视网膜“视觉输入(信息)的速度可以和以太网的传输速度相媲美”。在研究中,研究者将一只取自豚鼠的完好视网膜和一台叫作“多电极阵列”的设备连接起来,该设备可以测量神经节细胞中的电脉冲峰值。神经节细胞将信息从视网膜传达到大脑。基于这一研究,科学家们能够估算出所有神经节细胞传递信息的速度。其中一只豚鼠的视网膜含有大概 1 00 000 个神经节细胞,相应地,科学家们就能够计算出人类视网膜中的细胞每秒能传递多少数据。人类视网膜中大约包含 1 000 000 个神经节细胞,算上所有的细胞,人类视网膜能以大约每秒 10MB 的速度传达信息。

丹麦的著名科学作家陶·诺瑞钱德证明了人们通过视觉接收的信息比其他任何一种感官都多。如果人们通过视觉接收信息的速度和计算机网络相当,那么通过触觉接受信息的速度就只有它的 1/10。人们的嗅觉和听觉接收信息的速度更慢,大约是触觉接收速度的 1/10。同样,我们通过味蕾接收信息的速度也很慢。

换句话说,人们通过视觉接收信息的速度比其他感官接收信息的速度快了 10～100 倍。因此,可视化能传达庞大的信息量也就容易理解了。如果包含大量数据的信息被压缩成了充满知识的图片,那人们接收这些信息的速度会更快。但这并不是可视化数据表示法如此强大的唯一原因。另一个原因是人们喜欢分享,尤其喜欢分享图片。

3.2.3 图片和分享的力量

人们喜欢照片(图片)的主要原因之一,是现在拍照很容易。数码相机、智能手机和便宜的存储设备使人们可以拍摄多得数不清的数码照片。现在,几乎每部智能手机都有内置摄像头。这就意味着不但可以随意拍照,还可以轻松地上传或分享这些照片。这种轻松、自在的拍摄和分享图片的过程充满了乐趣和价值,人们自然想要分享它们。

和照片一样,如今制作信息图也要比以前容易得多。公司制作这类信息图的动机也多了。公司的营销人员发现,一个拥有有限信息资源的营销人员该做些什么来让搜索更加吸引人呢?答案是制作一张信息图。信息图可以吸纳广泛的数据资源,使这些数据相互吻合,甚至编造一个引人入胜的故事。博主和记者们想方设法地在自己的文章中加进类似的图片,因为读者喜欢看图片,同时也乐于分享这些图片。

最有效的信息图还是被不断重复分享的图片。其中有一些图片在网上疯传,它们在社交网站如脸书、推特、领英、微信以及传统但实用的邮件里被分享了数千次甚至上百万次。由于信息图制作需求的增加,帮助制作这类图形的公司和服务也随之增多。

3.3 实时可视化

很多信息图提供的信息,从本质上看是静态的。通常制作信息图需要花费很长的时间和精力:它需要数据,需要展示有趣的故事,还需要以图标将数据以一种吸引人的方式呈现出来。但是,工作到这里还没结束。图表只有经过发布、加工、分享和查看之后才具有真正的价值。当然,到那时,数据已经成了几周或几个月前的旧数据了。那么,在展示

可视化数据时，要怎样在吸引人的同时又保证其时效性呢？

数据要具有实时性价值，必须满足以下三个条件。

(1) 数据本身必须要有价值。

(2) 必须有足够的存储空间和计算机处理能力来存储和分析数据。

(3) 必须要有一种巧妙的方法，及时将数据可视化，而不用花费几天或几周的时间。

想了解数百万人是如何看待实时性事件的，并将他们的想法以可视化的形式展示出来，看似遥不可及，但其实很容易达成。

在过去的几十年里，美国总统选举过程中的投票民意测试，需要测试者打电话或亲自询问每个选民的意见。通过将少数选民的投票和统计抽样方法结合起来，民意测试者就能预测选举的结果，并总结出人们对重要政治事件的看法。但今天，大数据正改变着我们的调查方法。

捕捉和存储数据只是像推特这样的公司所面临的大数据挑战中的一部分。为了分析这些数据，公司开发了推特数据流，即支持每秒发送 5000 条或更多推文的功能。在特殊时期，如总统选举辩论期间，用户发送的推文更多，大约每秒 2 万条。然后公司又要分析这些推文所使用的语言，找出通用词汇，最后将所有的数据以可视化的形式呈现出来。

要处理数量庞大且具有时效性的数据很困难，但并不是不可能。推特为大家熟知的数据流人口配备了编程接口。像推特一样，Gnip 公司也开始提供类似的渠道。其他公司(如 BrightContext)提供实时情感分析工具。在 2012 年总统选举辩论期间，《华盛顿邮报》在观众观看辩论的时候使用 BrightContext 的实时情感模式来调查和绘制情感图表。实时调查公司 Topsy 将大约 2000 亿条推文编入索引，为推特的政治索引提供了被称为 Twindex 的技术支持。Vizzuality 公司专门绘制地理空间数据，并为《华尔街日报》选举图提供技术支持。

与电话投票耗时长且每场面谈通常要花费大约 20 美元相比，上述公司所采用的实时调查只需花费几个计算周期，并且没有规模限制。另外，它还可以将收集到的数据及时进行可视化处理。

但信息实时可视化并不只是在网上不停地展示实时信息而已。“谷歌眼镜”(图 3-12)被《时代周刊》称为 2012 年最好的发明。“它被制成一副眼镜的形状，增强了现实感，使之成为人们日常生活的一部分。”将来，人们不仅可以在计算机和手机上看可视化呈现的数据，还能边四处走动边设想或理解这个物质世界。

图 3-12 谷歌眼镜

3.4 数据可视化的运用

人类对图形的理解能力非常独到,往往能够从图形当中发现数据的一些规律,而这些规律用常规的方法是很难发现的。在大数据时代,数据量变得非常大,而且非常烦琐,要想发现数据中包含的信息或者知识,可视化是最有效的途径之一(图 3-13)。

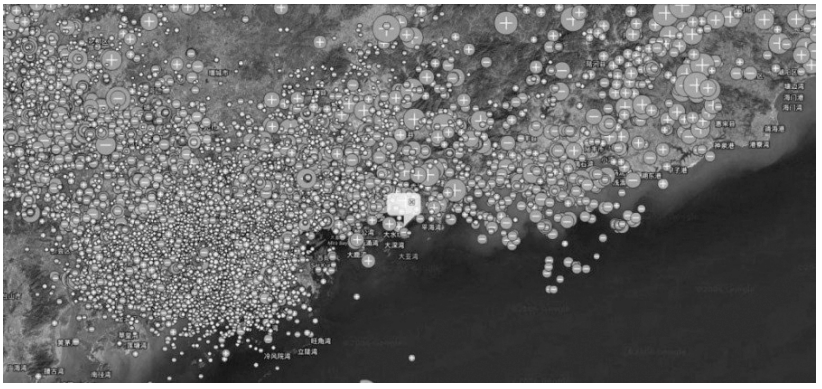

图 3-13 深圳受大面积雷电影响,某时间段共记录到 9119 次闪电

数据可视化要根据数据的特性,如时间信息和空间信息等,找到合适的可视化方式,如图表、图和地图等,将数据用直观地展现出来,以帮助人们理解数据,同时找出包含在海量数据中的规律或者信息。数据可视化是大数据生命周期管理的最后一步,也是最重要的一步。

数据可视化起源于图形学、计算机图形学、人工智能、科学可视化以及用户界面等领域的相互促进和发展,是当前计算机科学的一个重要研究方向。它是利用计算机对抽象信息进行直观表示,以利于快速检索信息,增强认知能力。

图 3-14 CLARITY 成像技术

数据可视化系统并不是为了展示已知数据之间的规律,而是为了帮助用户通过认知数据有新的发现,发现这些数据所反映的实质。如图 3-14 所示,CLARITY 成像技术使科学家们不需要切片就能够看穿整个大脑。

斯坦福大学生物工程和精神病学负责人卡尔·戴瑟罗斯说:“以分子水平和全局范围观察整个大脑系统,曾经一直都是生物学领域一个无法实现的重大目标”。也就是说,用户在使用信息可视化系统之前往往没有明确的目标。信息可视化系统在探索性任务(例如包含大数据量信息)中有突出的表现,它可以帮助用户从大量的数据空间中找到关注的信息来进行详细的分析。因此,数据可视化主要应用于下面几种情况。

(1) 当存在相似的底层结构、相似的数据,可以进行归类时。

(2) 当用户处理自己不熟悉的数据内容时。

(3) 当用户对系统的认知有限,并且喜欢用扩展性的认知方法时。

(4) 当用户难以了解底层信息时。

(5) 当数据更适合感知时。

【作　　业】

1. 数据是现实世界的一个快照,会传递给人们大量的信息。一个数据点可以包含(　　)等因素,因此,一个数字不再只是沧海一粟。

A. 时间　　B. 地点　　C. 任务　　D. A、B、C

2. 数据是对现实世界的(　　)。可视化能帮助你从一个个独立的数据点中解脱出来,换一个不同的角度去探索它们。

A. 简化和抽象表达　　B. 复杂化和抽象表达

C. 简化和分解表达　　D. 复杂化和分解表达

3. (　　)数据和它所代表的事物之间的关联,既是把数据可视化的关键,也是全面分析数据的关键,同样还是深层次理解数据的关键。

A. 数据之间的关联　　B. 数据和它所代表的事物之间的关联

C. 事物之间的关联　　D. 事物和它所代表的数据之间的关联

4. 一个(　　)独立的离群值可能是需要修正或特别注意的。也许在你的体系中,随着时间推移发生的变化预示有好事(或坏事)将要发生。

A. 总计值　　B. 平均值

C. 独立的离群值　　D. 普遍的连续值

5. 通常,大部分数据都是估算的,并不精确。分析师会研究一个样本,并据此猜测整体的情况。你会基于自己的知识和见闻来猜测,即使大多数时候你确定猜测是正确的,但仍然存在着(　　)。

A. 确定性　　B. 不确定性　　C. 唯一性　　D. 稳定性

6. (　　)可以完全改变你对某一个数据集的看法,它能帮助你确定数据代表着什么以及如何解释。确切了解了数据的含义之后,你的理解会帮你找出有趣的信息,从而带来有价值的可视化效果。

A. 前景信息　　B. 合计信息　　C. 背景信息　　D. 独特信息

7. 使用数据而不了解除数值本身之外的任何信息,就好比拿断章取义的片段作为文章的主要论点引用一样。你必须首先了解(　　)、何时、何地以及何因,即元数据,或者说关于数据的数据,然后才能了解数据的本质是什么。

A. 何人　　B. 如何　　C. 何事　　D. A、B、C

8. (　　)不仅是一种传递大量信息的有效途径,它还和大脑直接联系在一起,并能触动情感,引起化学反应,可能是传递数据信息最有效的方法之一。

A. 可视化　　B. 个性化　　C. 现代化　　D. 集中化

9. 通过(　　)颜色、布局、标记和其他元素的融合,图形能够在几秒钟之内就把这些信息传达给人们。将信息可视化能有效地抓住人们的注意力。

A. 颜色　　B. 布局　　C. 标记　　D. A、B、C

10. 人们在使用电子表格软件处理数据时发现,要从填满数字的单元格中发现走势是困难的。一般来说,人们在看一个(　　)的时候,更容易发现事物的变化走势。人们在制订决策的时候,了解事物的变化走势至关重要。

A. 折线图　　B. 饼状图　　C. 条形图　　D. B、C、D

11. (　　)不仅可以容纳大量信息,还是一种便于理解的表现方式。在大数据里,这样的东西就叫作“可视化”。可视化是压缩知识的一种方式。

A. 文字　　B. 数字　　C. 图片　　D. 表格

12. 通常,内部环境和外部环境的数据信息存储在(　　)地方。而且,这两种数据的存储模式也有细微的差别。

A. 同一个　　B. 两个不同的　　C. 隐蔽的　　D. 突出的

13. 根据美国宾夕法尼亚大学医学院的研究人员估计,通常情况下,人类视网膜“视觉输入(信息)的速度(　　)”。

A. 可以和以太网的传输速度相媲美　　B. 远远落后于以太网的传输速度

C. 远快于以太网的传输速度　　D. 很慢但很精确

14. 丹麦科学作家陶·诺瑞钱德证明了人们通过(　　)视觉接收的信息比其他任何一种感官都多。

A. 嗅觉　　B. 触觉　　C. 听觉　　D. 视觉

15. 数据要具有实时性价值,但(　　)不是实时性必须满足的条件。

A. 数据本身必须要有价值

B. 必须有足够的存储空间和计算机处理能力来存储和分析数据

C. 数据必须纯粹由数字和字符组成

D. 必须要有一种巧妙的方法及时将数据可视化,而不用花费几天或几周的时间

16. 信息实时可视化并不只是在网上不停地展示实时信息而已。将来人们不仅可以在计算机和手机上看可视化呈现的数据,还能(　　),在移动中设想或理解这个物质世界。

A. 手持 PAD 设备　　B. 身着可穿戴设备

C. 带着 U 盘　　D. 使用移动光盘

17. 数据可视化要根据数据的特性找到合适的可视化方式,将数据直观地展现出来,以帮助人们理解数据,同时找出包含在海量数据中的规律或信息。(　　)不属于这样的可视化元素。

A. 大字符集　　B. 图表　　C. 图　　D. 地图

18. 数据可视化起源于图形学、计算机图形学等领域的相互促进和发展,是当前计算机科学的一个重要研究方向。但(　　)不属于相关的起源领域。

A. 人工智能　　B. 科学可视化　　C. 二进制算法　　D. 用户界面

19. 数据可视化系统并不是为了(　　)。

A. 帮助用户认知数据

B. 帮助用户通过认知数据发现这些数据所反映的实质

C. 帮助用户通过认知数据有新的发现

D. 展示用户的已知数据之间的规律

20. 信息可视化系统可以帮助用户从大量的数据空间中找到关注的信息来进行详细的分析。但(　　)不是数据可视化的主要应用情况。

A. 当存在相似的底层结构,相似的数据可以进行归类时

B. 当用户处理自己非常熟悉的数据内容时

C. 当用户对系统的认知有限时,并且喜欢用扩展性的认知方法时

D. 当用户难以了解底层信息时

【实验与思考】 绘制南丁格尔极区图

1. 实验目的

(1) 熟悉大数据可视化的基本概念和主要内容。

(2) 通过绘制南丁格尔极区图,尝试了解大数据可视化的设计与表现技术。

2. 工具/准备工作

在开始本实验之前,请认真阅读课程的相关内容。

需要准备一台带有浏览器,能够访问因特网的计算机。

3. 实验内容与步骤

(1) 请结合查阅相关文献资料,简述什么是数据可视化,数据可视化系统的主要目的是什么。

答:_____

_____

_____

_____

(2) 南丁格尔“极区图”是数据统计类信息图表中常见的一类图表形式,下面来了解这类图表的一般绘制方法。

① 设计分析

设计的最终效果图如图 3-15 所示。

图表中包括性别、年龄、教育、收入等 11 个分类的对比信息指标,每个指标占用的圆周的角度相同,即任一指标的扇区角度为 360°/11=32.723°。在 CorelDraw 中,其表现为角度相同,半径不等的扇区图。

在 Gender、Income、Age、Education 四个指标中,又分别划成几个不同的区段。在 CorelDraw 中,同一扇区图中不同的区段由角度相同,半径不等的扇区图依次叠加而成。

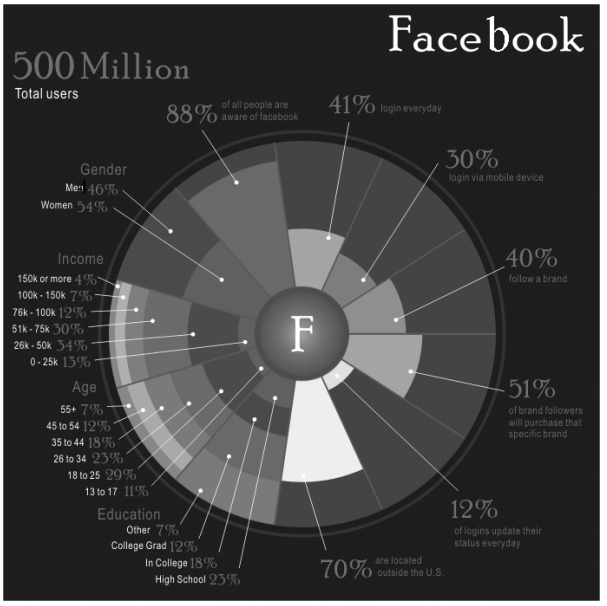

图 3-15　脸书极区图

② 绘图步骤

绘制此信息图,主要应用 CorelDraw 软件中的“旋转”和“分层叠加”两个功能。脸书极区信息图在 CorelDraw 中的具体绘制步骤如下:

步骤 1:绘制定位圆环和背景圆,以及 11 等分扇形。

步骤 2～3:依次绘制 11 个指标对应的不同长度的扇区图。

步骤 4～6:依次绘制 4 个指标中不同区段的扇区图(图 3-16)。

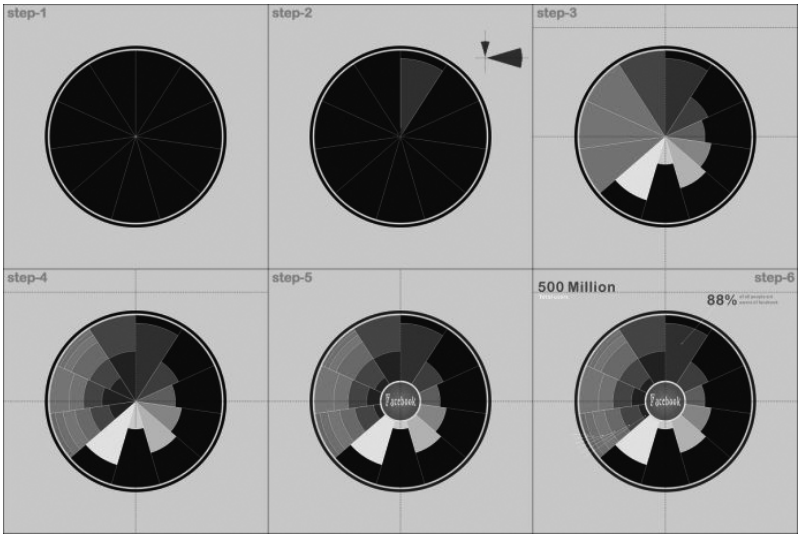

图 3-16　绘制极区图的步骤 1～6

读者也可尝试用自己熟悉的其他作图软件工具绘制此图。

4. 实验总结

5. 实验评价(教师)