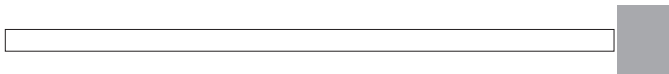




理 论 篇



第 1 章

数据可视化概述

在大数据时代,如何让决策者快速、准确地理解复杂数据(data)背后的隐藏信息,是数据管理者面临的重要挑战。数据可视化就是利用图形化的方法,让复杂的数据变得更加容易理解、更加便于接收者决策。本章首先介绍数据、信息以及两者之间的关系,引入数据可视化的必要性;然后介绍数据可视化的定义、原则、作用和分类等基本概念;接着介绍数据可视化的一般过程;最后介绍数据可视化的发展历史、面临的挑战、发展方向以及与其他课程的关系。

本章学习目标

- (1) 掌握数据可视化的概念;
- (2) 掌握数据可视化的原则;
- (3) 掌握数据可视化的作用;
- (4) 理解数据可视化的分类;
- (5) 掌握数据可视化的过程;
- (6) 了解数据可视化的发展。

1.1 数据和信息

1.1.1 数据

数据是指对客观事件进行记录并可以鉴别的符号,是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合。数据是事实或观察的结果,是对客观事物的逻辑归纳,是用于表示客观事物的未经加工的原始素材。

数据可以是离散的,也可以是连续的,数据属性可分为离散属性和连续属性。离散属性的取值来自有限或可数的集合,如邮政编码、等级、文档、单词等;连续属性则对应于实数域,如温度、高度和湿度等。属性值可以是表达属性的任意数值或符号,同一类属性可以具有不同的属性值,如长度的度量单位可以是米或公里等;不同的属性也可能具有相同的取值和不同的含义,如年份和年龄都是整数型数值,而年龄通常有取值区间。

1.1.2 信息

信息是现代社会中一个被广泛使用的概念,技术信息、经济信息、金融信息、军事信息等充斥整个世界。作为一个科学术语,信息的定义很多,还不统一。信息论的鼻祖香农

(Shannon)将信息定义为“不确定性减少的一种量度”，并给出了信息量的计算公式。此外，关于信息还有以下的概念，如：

信息是关于客观世界某一方面的知识；

信息对接收者来说是预先不知道的报道；

信息是使不确定因素减少的有用知识；

信息能改变决策中预期结果的概率，对决策过程有价值；

信息是经过加工并对人们行动产生决策影响的数据。

虽然上述定义的角度不同，但对信息特征的认识基本一致，均认为信息有如下特征。

1. 客观性

事物的运动 and 变化是不以人的意志为转移的客观存在，所以反映这种客观存在的信息，同样具有客观性。不仅信息反映的内容是客观的，而且信息本身也具有客观存在性。

2. 时效性

信息是对事物运动状态和变化的历史记录，总是先有事实后有信息，信息的使用价值与信息从信源至信宿经历的时间间隔成反比。时间的延误，会使信息的使用价值衰减甚至消失。

3. 共享性

信息在传递和使用中，允许多次和多方共享，原拥有者只失去信息的原始价值，不会失去信息的使用价值和潜在价值。因此，信息不会因共享而消失。

1.1.3 数据与信息的关系

数据与信息是紧密相关的两个概念。

(1) 数据是信息的载体，是用可鉴别的符号记录下来的客观事物的属性描述。例如，“体重 60 千克”“身高 1.65 米”等。数据可以是数字、文字、声音和图像等形式。

(2) 信息是对数据进行解释以后的结果。比如表示学生的一组数据“女生，年龄 20 岁，体重 50 千克，身高 1.65 米”，这一组数据所反映的信息是该女生形体较为匀称。

图 1-1 说明了数据与信息之间的关系。数据是信息的载体，而信息是经过加工后的数据，它对接接收者的行为产生影响，对接接收者的决策或行为有现实或潜在的价值。

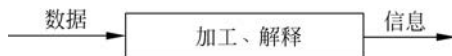


图 1-1 数据与信息

1.2 数据可视化的基本概念

信息对接收者的价值受到很多因素的影响。除了数据本身的完整性、准确性和时效性等之外，数据的展现方式，也将影响信息传递的准确性和接收者的认知，如人们常说的“一图

胜千言”，即图形能够以直观的形式呈现出复杂的数据所要表达的信息，方便接收者对数据进行进一步的分析与应用。特别是面对以高速(velocity)、大量(volume)和多样化(variety)3V为典型特征的大数据，迫切需要利用数据可视化，帮助决策者理解复杂数据背后隐藏的信息。

1.2.1 数据可视化的定义

数据可视化就是使用图形化手段表达数据的变化、联系或者趋势的方法，将数据转换为图形图像显示出来，其目的是清晰有效地传达与沟通信息，让用户更好地理解和使用数据。按过程来讲，数据可视化主要是记录信息、分析推理并进行信息传播与协同的过程。简而言之，数据可视化意味着以可视化图表的形式来显示数据信息，实现发现、分析、预测、监控、决策等目的。

数据可视化是一种关于数据视觉表现形式的方法，这种数据视觉表现形式被定义为一种以某种概要形式抽取出来的信息，包括相应信息单位的各种属性和变量。它是一个处于不断演变之中的概念，其边界在不断地扩大。其主要指的是技术上较为高级的技术方法，而这些技术方法允许利用图形图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性和动画的显示，对数据加以可视化解释。

数据可视化是一个复杂而漫长的过程，首先需要理解模拟信息和数字化数据等基础知识，然后掌握数据获取的技巧和方法，使用多种数据清洗(data cleaning)方法去除“脏”数据，通过数据分析了解数据的整体特征，理解可视化基础并在符合可视化原则的基础上运用可视化工具完成数据可视化作品。

数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。当前，在研究、教学和开发领域，数据可视化乃是一个极为活跃而又关键的方面。“数据可视化”这个术语实现了成熟的科学可视化领域与较年轻的信息可视化领域的统一。

1.2.2 数据可视化的原则

数据可视化并非简单套用图形化方法和工具。数据可视化首先一定要理解数据含义、明确目标，否则非常容易进入的误区就是，拿到一堆数据，还没有理解数据有什么含义，直接就开始套用图形进行展示，把大部分时间用在美化图表上，而完全忽略数据本身传达的意义。数据可视化一般要遵循以下原则。

1. 逻辑清晰

数据可视化一定要根据分析目标，确认好内容主线，做到逻辑严密、结构清晰。

2. 表达精准

在数据准确的前提下，根据目的选择正确的图表，表达合适的信息，确保不会产生理解歧义。

3. 设计简洁

数据可视化的重点不是好看，而是突出重点、简洁美观。图表各元素，布局、坐标、单位、

图例交互适中展示,不要过度设计。

1.2.3 数据可视化的作用

数据可视化是借助图形化手段,清晰有效地传达与沟通信息,帮助企业从数据中提取信息、从信息中收获价值。具体而言,数据可视化的作用包括以下几点。

1. 快速理解信息

科学研究表明,人类对图像的处理速度比文本快 6 万倍,同时人类右脑记忆图像的速度比左脑记忆抽象文字快 100 万倍。^① 数据可视化正是利用人类天生技能来增强数据处理和组织效率。通过使用图形能够以清晰、一致的方式查看大量数据,快速理解这些数据隐含的信息。例如,Anscombe 四重奏的例子很好地说明数据可视化对快速理解信息的作用。

1973 年,统计学家 F. J. Anscombe 构造出了四组奇特的数据,如表 1-1 所示。这四组数据中, x 值的均值都是 9.0, y 值的均值都是 7.5; x 值的方差都是 10.0, y 值的方差都是 3.75;它们的相关度都是 0.816,线性回归线都是 $y=3+0.5x$ 。单从这些统计数字来看,四组数据所反映的实际情况非常相近,而事实上,这四组数据有着天壤之别。

表 1-1 Anscombe 的四重奏数据

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

当我们分别绘制四组数据的散点图(scatter graph)后,如图 1-2 所示,即可看出点的分布完全不同。

第一组是线性关系图,数据是大多数人看到上述统计数字的第一反应,是最“正常”的一组数据;第二组是曲线关系图,数据所反映的事实上是一个精确的二次函数关系,只是在错误地应用线性模型后,各项统计数字与第一组数据恰好都相同;第三组是异常值图,数据描述

^① EISENBERG H. Humans process visual data better[EB/OL]. (2014-09-15). <http://www.t-sciences.com/news/humans-process-visual-data-better>.

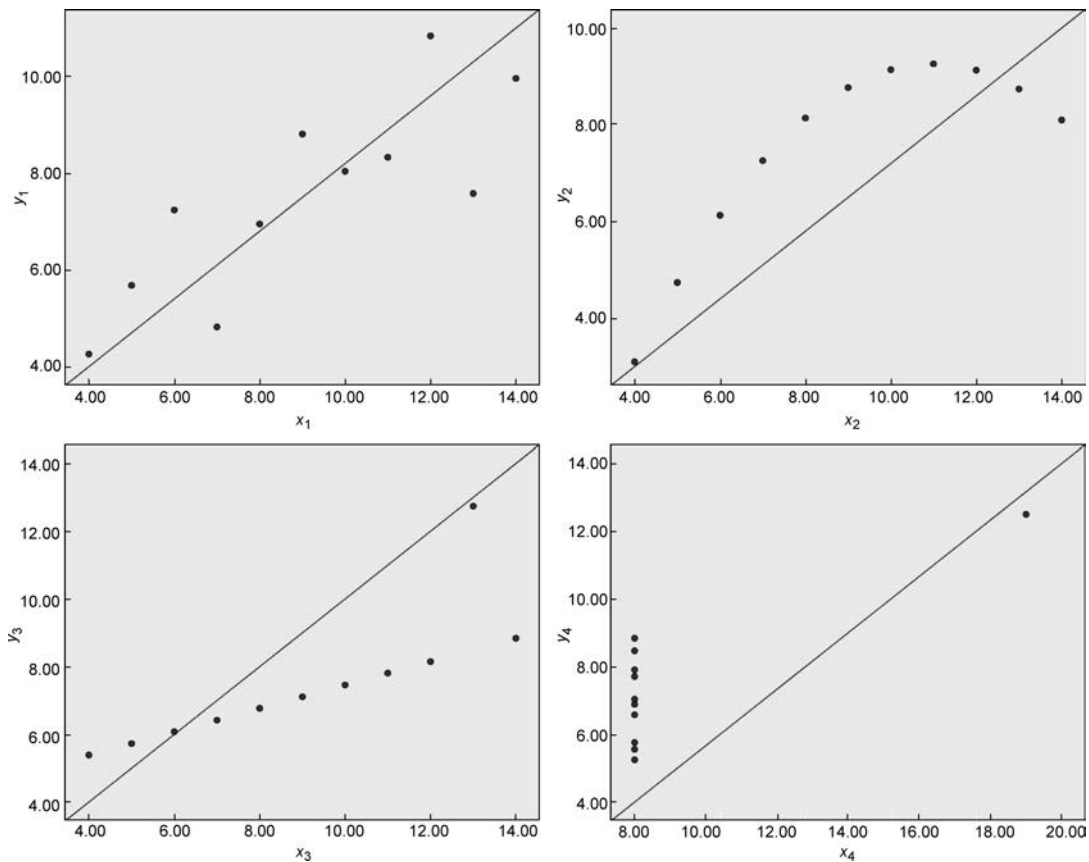


图 1-2 Anscombe 四重奏数据的可视化

的是一个精确的线性关系，只是这里面有一个异常值，它导致上述各个统计数字尤其是相关度值的偏差；第四组为极端异常值图，展示了一个更极端的例子，其异常值导致了均值、方差、相关度、线性回归线等所有统计数字全部的偏差。

2. 识别关系和模式

数据之间以及数据描述的对象之间，往往存在隐含的关系或者难以直接观测的模式，数据可视化以图形方式可以挖掘出这些隐含的关系或者难以观测的模式，这些关系和模式将有助于个人和企业作出科学合理的决策。

案例：不同人群的年龄与脂肪含量

表 1-2 为不同年龄段和不同收入水平人群的脂肪含量，从表 1-2 很难看出年龄收入和脂肪含量之间的关系。

表 1-2 不同年龄段和不同收入水平人群的脂肪含量 %

收入水平	男 性		女 性	
	65 岁以下	65 岁及以上	65 岁以下	65 岁及以上
低收入	25	20	37.5	55
高收入	43	30	55	50

将表 1-2 做个简单的折线图(line chart),结果如图 1-3 所示,很容易发现在男性低收入组、男性高收入组和女性高收入组,随着年龄的增长,脂肪含量会下降;但是在女性低收入组,随着年龄的增长,脂肪含量却会增加。

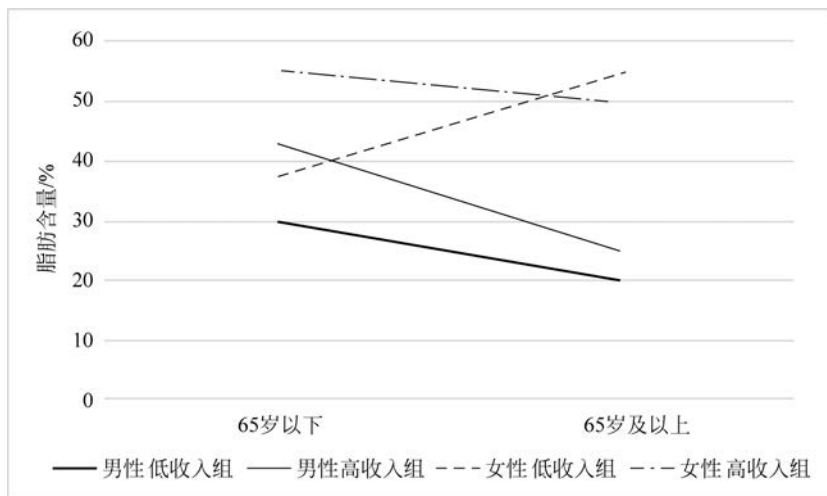


图 1-3 不同年龄段和不同收入水平人群的脂肪含量的折线图

3. 证实假设或者猜想

除了帮助快速理解信息、识别关系和模式之外,数据可视化还能够帮助论证假设或者猜想。常常通过 A/B 测试的模式,来验证假设,对于分析问题过程中的推论进行验证假设,从而发现根本原因。南丁格尔发明的玫瑰图(Nightingale's rose diagram)就是利用数据可视化证实假设的经典案例。

案例：南丁格尔玫瑰图

19 世纪 50 年代,英国、法国、土耳其和俄国进行了克里米亚战争。南丁格尔主动申请,自愿担任战地护士。当时的医院卫生条件极差,伤士死亡率高达 42%,直到 1855 年卫生委员会来到医院改善整体的卫生环境后,伤士死亡率才戏剧性地降至 2.5%。当时的南丁格尔注意到这件事,认为政府应该改善战地医院的条件来拯救更多年轻的生命。

出于对资料统计的结果会不受人重视的忧虑,她发明出一种色彩缤纷的图表形式,让数据能够更加让人印象深刻。如图 1-4 所示,这张图用以表达战地医院季节性的死亡率,从整体来看:这张图是用来说明、比较战地医院伤患因各种原因死亡的人数,每块扇形代表着各个月份的死亡人数,面积越大代表死亡人数越多。

说明:

- (1) 各色块圆饼区均由圆心往外的面积来表现数字。
- (2) 浅灰色区域:死于原本可避免的感染的士兵数。
- (3) 深灰色区域:因受伤过重而死亡的士兵数。
- (4) 黑色区域:死于其他原因的士兵数。
- (5) 1854 年 10 月、1855 年 4 月的深灰色区域和黑色区域恰好相等。

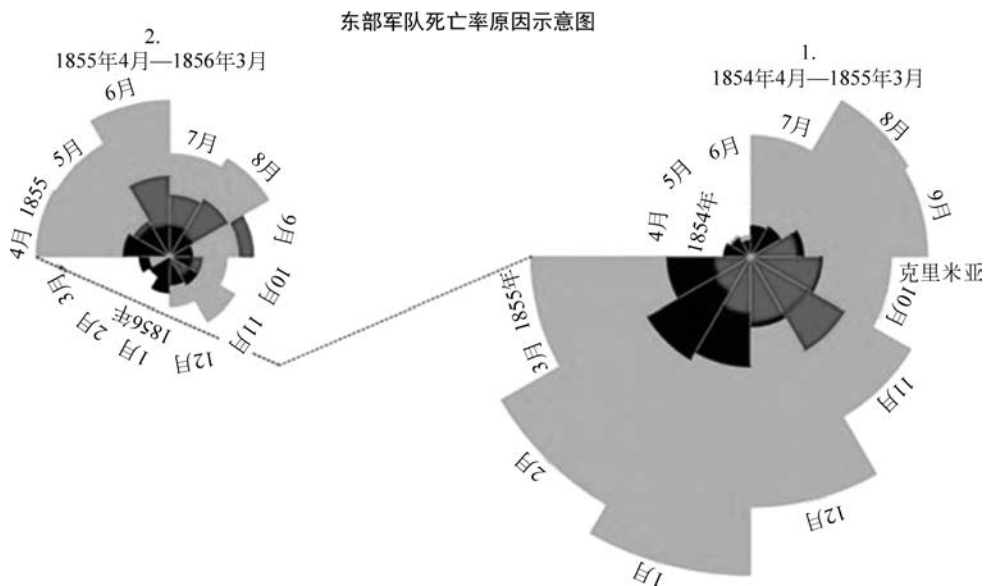


图 1-4 南丁格尔的玫瑰图

(6) 1856 年 1 月与 2 月的浅灰色区域、黑色区域恰好相等。

(7) 1854 年 11 月深灰色区域中的黑线指出该月的黑色区域大小。

由图 1-4 可知,左、右两个玫瑰图被时间点“1855 年 3 月”所隔开。其中,右侧较大的玫瑰图,展现的是 1854 年 4 月至 1855 年 3 月;而左侧的玫瑰图,展现的则是 1855 年 4 月至 1856 年 3 月。通过对两个图大小的对比,可以轻易地得出以下结论。

第一,浅灰色区域的面积明显大于其他颜色的面积。这意味着大多数的伤亡并非直接来自战争,而是来自糟糕医疗环境下的感染。

第二,卫生委员会到达后(1855 年 3 月),死亡人数明显地下降,证实了南丁格尔提出的“改善医院的医疗状况可以显著地降低英军死亡率”的假设。

这幅图让政府相关官员了解到:改善医院的医疗状况可以显著地降低英军的死亡率。南丁格尔的方法打动了当时的高层,包括军方人士和维多利亚女王本人,于是医事改良的提案才得到支持,甚至挽救了千万人的生命。这种新型的图表因为外形很像一朵绽放的玫瑰,被称为“南丁格尔玫瑰图”。

1.2.4 数据可视化的分类

从不同的角度,数据可视化有不同的分类。从数据可视化表现形式的角度,数据可以分为饼图、柱形图、条形图、折线图、圆环图、雷达图、直方图、词云图(word cloud diagram)等;从应用领域的角度,数据可以分为聚类分析中的数据可视化、评价分析中的数据可视化、关联性分析中的数据可视化、状态监控中的数据可视化;从应用行业的角度,数据可以分为制造业的数据可视化、农业的数据可视化、金融业的数据可视化、医学的数据可视化、教育的数据可视化等。

数据可视化的处理对象是数据,因此数据可视化与所处理的数据类型紧密相关。本书

按照数据类型,将数据可视化分为以下六类。

1. 结构化数据可视化

结构化数据(structured data)主要借助二维表的形式表达数据逻辑,通常数据结构以行和列的形式存在。其中,行数据代表样本或观测,列数据代表变量或属性。结构化数据主要使用关系数据库管理系统进行存储和管理,而且广泛存在于信息系统中。结构化数据是最常见的数据类型。常见的结构化数据可视化视角主要包括比较与排序、局部与整体、分布、相关性、网络关系、时间趋势等,详细内容请见第5章。

2. 关系型数据可视化

关系型数据(relational data)可视化主要是为了揭示实体间的联系或关联,以便分析数据结构、规律和特征。针对不同的关系型数据可视化常用的图表有散点图、气泡图、韦恩图、树状图、旭日图、桑基图、和弦图、漏斗图、节点关系图等,详细内容请见第6章。

3. 文本数据可视化

文字是传递信息最常用的载体。文本数据可视化的需求一般可分为三级:词汇级(lexical level)、语法级(syntactic level)和语义级(semantic level)。针对不同级别的理解需要,其分析内容和可视化技术也是不同的。对于词汇级的理解,主要是通过各类分词算法,基于词频(term frequency, TF)的可视化技术,包括标签云、词云(word cloud)等;对于语法级的理解,需要采用一些句法分析算法,如 Cocke-Kasami-Younger(CKY)算法、Earley 算法和 Chart Parsing 算法等;而对语义级的理解则主要采用主题抽取等算法,详细内容请见第7章。

4. 多媒体数据可视化

多媒体数据一般包括声音(sound)、图像和视频等多种媒体形式。声音是能触发听觉的生理信号,声音属性包括音乐频率(音调)、音量、速度、空间位置等;图像是日常生活中最常见、最容易创造的媒体,图像属性包括明暗变化、场景复杂、轮廓色彩丰富等;视频分析涉及视频结构和关键帧的抽取、视频语义的理解,以及视频特征和语义的可视化与分析。这些多媒体数据的可视化旨在从原始数据集(dataset)中提取有意义的信息,来揭示其内在的结构和模式,并采用适当的视觉表达形式传达给用户,详细内容详见第8章。

5. 时变型数据可视化

随时间变化、带有时间属性的数据称为时变型数据(time-varying data 或者 temporal data),时变型数据可视化旨在揭示数据随时间变化的规律。时变型数据可视化主要从表达维度、比例维度和布局维度进行设计,常见的可视化图表有折线图、日历图、面积图、主题河流图、K线图,具体内容详见第9章。

6. 空间型数据可视化

地理空间型图表主要展示数据中的精确位置和地理分布规律。空间型数据是带有地理

位置信息的数据,它所具有的数据属性跟地理区域有关。空间数据可视化已有诸多成果,从点、线、面的角度出发可分为点数据可视化、线数据可视化和区域数据可视化。区域数据可视化的目的是表现区域的属性,它可以比点数据和线数据表达更多信息,具体内容详见第10章。

1.3 数据可视化的一般过程

数据可视化是以数据流向为主线的一个完整流程,主要包括数据获取、数据处理、可视化映射等环节。一个完整的可视化过程,可以看成数据流经过一系列处理模块并得到转化的过程,用户通过可视化结果获取信息,从而作出决策。数据可视化的一般过程如图1-5所示。

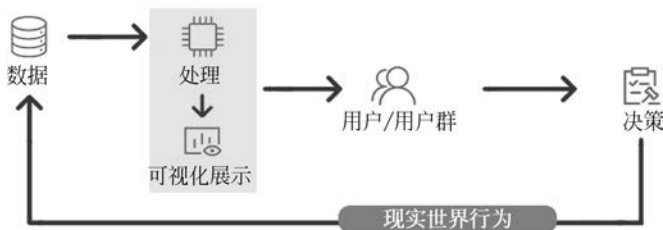


图 1-5 数据可视化的一般过程

1.3.1 数据获取

大数据时代的特点之一是数据开始变得廉价,即获取数据的途径多种多样,成本相对较低。通常来说,数据获取的手段有实验测量、计算机仿真与网络数据传输等。传统的数据获取方式以文件输入/输出为主。在移动互联网时代,基于网络的多源数据交换占据主流。数据获取的挑战主要有数据格式变换和异构异质数据的获取协议两部分。数据的多样性导致不同的数据语义表述,这些差异来自不同的安全要求、不同的用户类型、不同的数据格式、不同的数据来源。

数据获取协议(Data Access Protocol,DAP)作为一种通用的数据获取标准,在科研领域应用比较广泛。该协议通过定义基于网络的数据获取句法,以完善数据交换机制,维护、发展和提升数据获取效率。DAP4 提供了更多的数据类型和传输功能,以适用于更广泛的环境,直接满足用户要求。OPeNDAP(网络数据访问协议的开源项目)是一个研发数据获取协议的组织,它提供了一个同名的科学数据联网的简要框架,允许以本地数据格式快速地获取任意格式远程数据的机制。协议中相关的系统要素包括客户端、浏览器界面、数据集成、服务器等。

除此之外,互联网上存在大量免费的数据资源,这些资源通常由网站进行维护,并开放专门的API(应用程序编程接口)使用户得以访问。Google 作为全世界最大的互联网公司之一,提供了许多用于免费数据获取的API,如用于获取高级定制搜索结果的Google Custom Search,以及用于获取地理坐标信息的Google Geocoding API等。豆瓣和Twitter等社交

网站也开放了数据获取 API,用于获取社交网络相关信息。在法律法规及行业规定允许的条件下,可采用爬虫技术针对网络数据进行爬取,以满足数据分析、可视化的研究需求。

1.3.2 数据处理

1. 数据清洗

数据质量是数据获取后所需考虑的第一个问题。对于海量数据来说,未经处理的原始数据包含大量的无效数据,这些数据在到达存储过程之前就应该被过滤掉。在原始数据中,常见的数据质量问题包括噪声和离群值、数值缺失、数值重复等。解决这些问题的方法称为数据清洗。

(1) 噪声和离群值。噪声是指对真实数据的修改;离群值是指与大多数数据偏离较大的数据。

(2) 数值缺失。数值缺失的主要原因包括:信息未被记录;某些属性不适用于所有实例。处理数据缺失的方法有:删除数据对象;插值计算缺失值;在分析时忽略缺失值;用概率模型估算缺失值等。非结构化数据(unstructured data)通常存在低质量数据项(如从网页和传感器网络获取的数据),构成了数据清洗和数据可视化的新挑战。

(3) 数值重复。数值重复的主要来源是异构数据源的合并,可采用数据清洗方法消除。

处理数据丢失和重复记录仅是数据清洗的一部分,其他操作还包括:运用汇总统计删除、分辨或者修订错误或不精确的数据;调整数据格式和测量单位;数据标准化与归一化等。另外,实际采集的数据经常包含错误和自相矛盾的内容,而且实验、模拟和信息分析过程不可避免地存在误差,从而对分析结果产生很大的影响。

数据清洗最终需要达到的目标,包括有效性、准确性、可信性、一致性、完整性和时效性六个方面。在数据清洗步骤完成后,可以依据这 6 个指标对已清洗的数据进行数据质量评估,以满足可视化需求。

2. 数据精简

由高维性带来的维度灾难、数据的稀疏性和特征的多尺度性是大数据时代数据所特有的性质。直接对海量高维的数据集进行可视化通常会产生杂乱无章的结果,这种现象被称为视觉混乱。为了能够在有限的显示空间内表达比显示空间尺寸大得多的数据,需要进行数据精简。

在数据存储、分析层面进行的数据精简能降低数据复杂度,减少数据点数目并同时保留数据中的内涵特征,从而减少查询和处理时的资源开销,提升查询的响应性能。经典的数据精简包括统计分析、采样、直方图、聚类 and 降维,也可采用各类数据特征抽取方法,如奇异值分解、局部微分算子、离散小波变换等。

面向大数据的交互可视化对数据组织和管理提出了更高的要求。实施计算机图形学发展的一些理念经常被应用在交互式数据可视化应用上,如可伸缩的数据结构和算法、层次化数据管理和多尺度表达等。在选择恰当的数据精简方法时,使用者必须对时机、对象、使用策略和视觉质量评估等因素进行综合考察,这些考察项目不仅针对数据管理、数据可视化等

学科,还往往涉及认知心理学、用户测试、视觉设计等相关学科。以是否使用可视化为标准,数据精简方法可分为两类:①使用质量指标优化非视觉因素,如时间、空间等;②使用质量指标优化数据可视化,称为可视数据精简。

3. 其他数据预处理过程

在解决质量问题后,通常需要对数据集进行进一步的处理操作,以符合后续数据分析步骤要求。这一类操作通常被归为数据预处理步骤。常用的预处理操作有以下几种。

(1) 合并。合并是将两个以上的属性或对象合并为一个属性或对象。合并操作的效用包括:有效简化数据;改变数据尺度(例如,从乡村起逐级合并,形成城镇、地区、州、国家等);减小数据的方差。

(2) 采样。采样是统计学的基本方法,也是对数据进行选择的主要手段,在对数据的初步探索和最后的数据分析环节经常被采用。统计学家实施采样操作的根本原因是获取或处理全部数据集的代价太高,或者时间开销无法接受。如果采样结果大致具备原始数据的特征,那么这个采样是具有代表性的。最简单的随机采样可以按某种分布随机从数据集中等概率地选择数据项。当某个数据项被选中后,它可以继续保留在采样对象中,也可以在后继采样过程中被剔除。在前一种模式中,同一个数据项可能被多次选中。采样也可分层次进行:先将数据全集分为多份,然后在每份中随机采样。

(3) 降维。维度越高,数据集在高维空间的分布越稀疏,从而减小了数据集的密度和距离的定义对数据聚类与离群值检测等操作的影响。将数据属性的维度降低,有助于解决维度灾难,减少数据处理的时间和内存消耗;可以更为有效地可视化数据;降低噪声或消除无关特征等。降维是数据挖掘的核心研究内容,常规的做法有主成分分析、奇异值分解、局部结构保持的 LLP(局部线性投影)、ISOMAP(等度量映射)等方法。

(4) 特征子集选择。从数据集中选择部分数据属性值可以消除冗余、与任务无关的特征。特征子集选择可达到降维的效果,但不破坏原始的数据属性结构。特征子集选择的方法包括暴力枚举法、特征重要性选择、压缩感知理论的稀疏表达方法等。

(5) 特征生成。特征生成可以在原始数据集基础上构建新的能反映数据集重要信息的属性,常用的方法有特征抽取、将数据应用到新空间、基于特征融合与特征变换的特征构造三种。

(6) 离散化与二值化。将数据集根据其分布划分为若干个子类,形成对数据集的离散表达,称为离散化。将数据值映射为二值区间,是数据处理中的常见做法。将数据区间映射到 $[0, 1]$ 区间的方法称为二值化。

(7) 属性变换。将某个属性的所有可能值一一映射到另一个空间的做法称为属性变换,如指数变换、取绝对值等。标准化与归一化是两类特殊的属性变换,其中,标准化将数据区间变换到某个统一的区间范围,归一化则变换到 $[0, 1]$ 区间。

1.3.3 可视化映射

对数据进行清洗、去噪、精简,并按照业务目的进行数据处理之后,接下来就到了可视化映射环节。可视化映射是整个数据可视化流程的核心,是指将处理后的数据信息映射成可

可视化元素的过程。可视化元素由可视化空间、标记、视觉通道三部分组成。

1. 可视化空间

数据可视化的显示空间,通常是二维。三维(3D)物体的可视化,通过图形绘制技术,解决了在二维平面显示的问题,如 3D 环形图、3D 地图等。

2. 标记

标记是数据属性到可视化几何图形元素的映射,用来代表数据属性的归类。根据空间自由度的差别,标记可以分为点、线、面、体,分别具有零自由度、一维自由度、二维自由度、三维自由度。如我们常见的散点图、折线图、矩形树图、三维柱状图,分别采用了点、线、面、体这四种不同类型的标记。

3. 视觉通道

数据属性的值到标记的视觉呈现参数的映射,叫作视觉通道,通常用于展示数据属性的定量信息。常用的视觉通道包括标记的位置、大小(长度、面积、体积……)、形状(三角形、圆、立方体……)、方向、颜色(色调、饱和度、亮度、透明度……)等。

标记和视觉通道是可视化编码(visual encoding)元素的两个方面,两者的结合,可以完整地将数据信息进行可视化表达,从而完成可视化映射这一过程。

1.3.4 用户交互

1. 人机交互

可视化的目的,是反映数据的数值、特征和模式,以更加直观、易于理解的方式,将数据背后的信息呈现给目标用户,辅助其作出正确的决策。但是通常,我们面对的数据是复杂的,数据所蕴含的信息是丰富的。如果在可视化图形中,将所有信息不经过组织和筛选,全部机械地摆放出来,不仅会让整个页面显得特别臃肿和混乱、缺乏美感,而且模糊了重点,分散用户的注意力,削弱用户单位时间获取信息的能力。

常见的人机交互方式包括以下几种。

(1) 滚动和缩放:当数据在当前分辨率的设备上无法完整展示时,滚动和缩放是一种非常有效的交互方式,如地图、折线图的信息细节等。但是,滚动与缩放的具体效果,除了与页面布局有关外,还与具体的显示设备有关。

(2) 颜色映射的控制:一些可视化的开源工具,会提供调色板,如 D3 等。用户可以根据自己的喜好,去进行可视化图形颜色的配置。在自助分析等平台型工具中,此功能会相对多一些,但是在一些自研的可视化产品中,一般有专业的设计师来负责这项工作,从而使可视化的视觉传达具有美感。

(3) 数据映射方式的控制:是指用户对数据可视化映射元素的选择,一般一个数据集,是具有多组特征的,提供灵活的数据映射方式给用户,可以方便用户按照自己感兴趣的维度去探索数据背后的信息。其在常用的可视化分析工具中都有提供,如 Tableau、Power BI 等。

(4) 数据细节层次控制：如隐藏数据细节，鼠标悬停(hover)或单击才出现。

2. 用户感知

可视化的结果，只有被用户感知之后，才可以转化为信息和知识。用户在感知过程中，除了被动接受可视化的图形之外，还通过与可视化各模块之间的交互，主动获取信息。如何让用户更好地感知可视化的结果，将结果转化为有价值的信息用来指导决策，涉及的影响因素很多，包括心理学、统计学、人机交互等多个学科的知识。

1.4 数据可视化的发展

1.4.1 数据可视化的发展历史

数据可视化的发展与测量、绘画、文明、科技的发展是密切相关的，数据可视化广泛出现在地图、科学、工程制图、统计图表中。数据可视化的发展大致可以分为九个阶段，如图 1-6 所示。

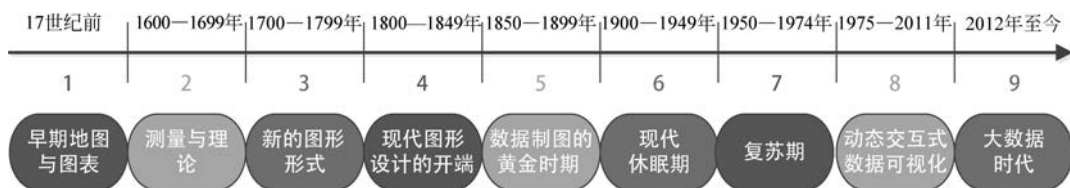


图 1-6 数据可视化的发展历史

1. 17 世纪前：早期地图与图表

在 17 世纪以前，人类研究的领域有限，总体数据量处于较少的阶段，因此几何学通常被视为可视化的起源，数据的表达形式也较为简单。但随着人类知识的增长，活动范围不断扩大，为了能有效探索其他地区，人们开始汇总信息绘制地图。16 世纪用于精确观测和测量物理量以及地理与天体位置的技术和仪器得到了充分发展，尤其在 W. Snell 于 1617 年首创三角测量法后，绘图变得更加精确，形成更加精准的视觉呈现方式。由于宗教等因素，人类对天文学的研究开始较早。一位不知名的天文学家于 10 世纪创作了描绘 7 个主要天体时空变化的多重时间序列图，图中已经存在很多现代统计图形的元素坐标轴、网格图系统、平行坐标和时间序列。

2. 1600—1699 年：测量与理论

大航海时代，欧洲的船队出现在世界各地的海洋上，发展欧洲新生的资本主义，这对于地图制作、距离和空间的测量都产生了极大的促进作用，更为准确的测量方式在 17 世纪得到了更为广泛的使用。同时，伴随着科技的进步以及经济的发展，数据的获取方式主要集中在时间、空间、距离的测量上，对数据的应用集中在制作地图、天文分析（如开普勒的行星运

动定律,1609)上。

与此同时,笛卡儿创立了解析几何和坐标系,在两个维度或者三个维度上进行数据分析,成为数据可视化历史中重要的一步。同时,早期概率论和人口统计学研究开始出现。这些早期的探索,开启了数据可视化的大门,数据的收集、整理和绘制开始了系统性的发展。在此时期,由于科学研究领域的增多,数据总量大大增加,出现了很多新的可视化形式。人们在完善地图精度的同时,不断在新的领域使用可视化方法处理数据。17 世纪末,启动“视觉思维”的必要元素已经准备就绪。

3. 1700—1799 年：新的图形形式

18 世纪是科学史上承上启下的年代,英国工业革命、牛顿对天体的研究、微积分方程等的建立等,都推动着数据向精准化以及量化的阶段发展,统计学研究的需求越发显著,用图形的方式来表示数据的想法也不断成熟。此时,经济学中出现了类似当今柱状图的线图表述方式,英国神学家 Joseph Priestley 也尝试在历史教育上使用图的形式介绍不同国家在各个历史时期的关系。法国人 Marcellin Du Carla 绘制了等高线图,该图用一条曲线表示相同的高程,对于测绘、工程和军事有着重大的意义,成为地图的标准形式之一。

数据可视化发展中的重要人物 William Playfair 在 1765 年创造了第一个时间线图,其中,单个线用于表示人的生命周期,整体可以用于比较多人的生命跨度。这些时间线直接启发他发明了条形图以及其他一些我们至今仍常用的图形,包括饼图、时序图等。他的这一思想可以说是数据可视化发展史上一次新的尝试,用新的形式直观地表达尽可能多的数据。

随着对数据系统性的收集以及科学的分析处理,18 世纪数据可视化的形式已经接近当代科学使用的形式,条形图和时序图等可视化形式的出现,体现了人类数据运用能力的进步。随着数据在经济、地理、数学等领域不同应用场景的应用,数据可视化的形式变得更加丰富,也预示着现代化的信息图形时代的到来。

4. 1800—1849 年：现代信息图形设计的开端

19 世纪上半叶,受到 18 世纪的视觉表达方法创新的影响,统计图形和专题绘图领域出现爆炸式的发展,目前已知的几乎所有形式的统计图形都是在此时被发明的。在此期间,数据的收集整理范围明显扩大,由于政府加强对人口、教育、犯罪、疾病等领域的关注,大量社会管理方面的数据被收集用于分析。1801 年,英国地质学家 William Smith 绘制了第一幅地质图,引领了一场在地图上表现量化信息的潮流,也被称为“改变世界的地图”。

这一时期,数据的收集整理从科学技术和经济领域扩展到社会管理领域,对社会公共领域数据的收集标志着人们开始以科学手段进行社会研究。与此同时,科学研究对数据的需求变得更加精确,研究数据的范围也有明显扩大,人们开始有意识地使用可视化的方式尝试研究、解决更广泛领域的问题。

5. 1850—1899 年：数据制图的黄金时期

19 世纪上半叶末,数据可视化领域开始快速发展,随着数字信息对社会、工业、商业和交通规划的影响不断增大,欧洲开始着力发展数据分析技术。高斯和拉普拉斯发起的统计理论给出了更多种数据的意义,数据可视化迎来了它历史上的第一个黄金时代。

统计学理论的建立是推动可视化发展的重要一步,此时数据的来源也变得更加规范化,由政府机构进行采集。随着社会统计学的影响力越来越大,在1857年维也纳的统计学国际会议上,学者就已经开始对可视化图形的分类和标准化进行讨论。不同数据图形开始出现在书籍、报刊、研究报告和政府报告等正式场合之中。这一时期,法国工程师 Charles Joseph Minard 绘制了多幅有意义的可视化作品,被称为“法国的 Playfair”,他最著名的作品是用二维的表达方式,展现六种类型的数据,用于描述拿破仑战争时期军队损失。

1879年,Luigi Perozzo 绘制了一张1750—1875年瑞典人口普查数据图,以金字塔形式表现了人口变化的三维立体图,此图与之前所看到的可视化形式有一个明显的区别:开始使用三维的形式,并使用彩色表示了数据值之间的区别,提高了视觉感知。

在对这一时期数据可视化的探究中发现,数据来源的官方化以及对数据价值的认同,成为数据可视化快速发展的决定性因素,如今几乎所有的常见可视化元素都已经出现,并且出现了三维的数据表达方式,这种创造性的成果对后来的研究有十分重要的作用。

6. 1900—1949年：现代休眠期

20世纪上半叶,随着数理统计这一新数学分支的诞生,追求数理统计严格的数学基础并扩展统计的疆域,成为这个时期统计学家的核心任务。数据可视化成果在这一时期得到了推广和普及,并开始被用于尝试解决天文学、物理学、生物学的理论新成果。如 Hertzsprung-Russell 绘制的温度与恒星亮度图成为近代天体物理学的奠基之一;伦敦地铁线路图的绘制形式如今一直在沿用;E. W. Maunder 的“蝴蝶图”用于研究太阳黑子随时间的变化。

然而,这一时期人类收集、展现数据的方式并没有得到根本上的创新,统计学在这一时期也没有大的发展,所以整个20世纪上半叶都是休眠期。但正是这一时期的蛰伏与统计学者潜心的研究,才让数据可视化在20世纪后期迎来了复苏与更快速的发展。

7. 1950—1974年：复苏期

从20世纪上半叶末到1974年这一时期被称为数据可视化的复苏期,在这一时期引起复苏的最重要的因素就是计算机的发明,计算机的出现让人类处理数据的能力有了跨越式提升。在现代统计学与计算机计算能力的共同推动下,数据可视化开始复苏,统计学家 John W. Tukey 和制图师 Jacques Bertin 成为可视化复苏期的领军人物。

John W. Tukey 在第二次世界大战期间对火力控制进行的长期研究中意识到了统计学在实际研究中的价值,发表了具有划时代意义的 *The Future of Data Analysis* 论文,让科学界将探索性数据分析(exploratory data analysis, EDA)视为不同于数学统计的另一独立学科,并在20世纪后期首次采用了茎叶图、盒形图等新的可视化图形形式,成为可视化新时代的开启性人物。Jacques Bertin 发表了他里程碑式的著作 *Sémiologie Graphique*。这部书根据数据的联系和特征,来组织图形的视觉元素,为信息的可视化提供了一个坚实的理论基础。

随着计算机的普及,20世纪60年代末,各研究机构就逐渐开始使用计算机程序取代手绘图形。由于计算机的数据处理精度和速度具有强大的优势,高精度分析图形已不能用手绘制。在这一时期,数据缩减图、多维标度法、聚类图、树形图等更为新颖复杂的数据可视化形式开始出现。人们开始尝试在一张图上表达多种类型数据,或用新的形式表现数据之间的复杂关联,这也成为现今数据处理应用的主流方向。数据和计算机的结合让数据可视化

迎来了新的发展阶段。

8. 1975—2011 年：动态交互式数据可视化

在这一阶段,计算机成为数据处理必要的成分,数据可视化进入新的黄金时代,随着应用领域的增加和数据规模的扩大,更多新的数据可视化需求逐渐出现。20 世纪 70 年代到 80 年代,人们主要尝试使用多维定量数据的静态图来表现静态数据。80 年代中期,动态统计图开始出现,最终在 20 世纪末两种方式开始合并,试图实现动态、可交互的数据可视化,于是动态交互式的数据可视化方式成为新的发展主题。

数据可视化在这一时期的最大潜力来自动态图形方法的发展,其允许对图形对象和相关统计特性进行即时和直接的操纵。早期就已经出现为了实时地与概率图进行交互的系统,通过调整控制来选择参考分布的形状参数和功率变换。这可以看作动态交互式可视化发展的起源,推动了这一时期数据可视化的发展。

9. 2012 年至今：大数据时代

在 2003 年全世界创造了 5 EB 的数据量时,人们就逐渐开始对大数据的处理进行重点关注。到 2011 年,全球每天的新增数据量已经开始以指数级增长,用户对于数据的使用效率在不断提升,数据的服务商也就开始需要从多个维度向用户提供服务,大数据时代正式到来。

2012 年,进入数据驱动的时代。掌握数据就能掌握发展方向,因此人们对数据可视化技术的依赖程度也不断加深。大数据时代的到来对数据可视化的发展有着冲击性的影响,试图继续以传统展现形式来表达庞大的数据量中的信息是不可能的,大规模的动态化数据要依靠更有效的处理算法和表达形式才能够传达出有价值的信息,因此大数据可视化的研究成为新的时代命题。

在应对大数据时,不但要考虑快速增加的数据量,还需要考虑到数据类型的变化,这种数据扩展性的问题需要更深入的研究才能解决;互联网的加入增加了数据更新的频率和获取的渠道,并且实时数据的巨大价值只有通过有效的可视化处理才可以体现,于是在上一历史时期就受到关注的动态交互的技术已经向交互式实时数据可视化发展,这也是如今大数据可视化的研究重点之一。综上,如何建立一种有效的、可交互式的大数据可视化方案来表达大规模、不同类型的实时数据,成为数据可视化这一学科主要的研究方向。

1.4.2 数据可视化面临的挑战

伴随着大数据时代的到来,数据可视化日益受到关注,数据可视化在人们日常生活中扮演着重要的角色,可视化技术也日益成熟,越来越多的可视化画面出现在人们生活中,成为人们逐渐关注和进行研究的一门课题。随着社会发展的需求,数据可视化仍存在许多问题,且面临巨大的挑战和各种各样的难题,数据可视化发展中面临的困境主要有以下几方面。

1. 视觉噪声

在数据大规模集中环节,大多数数据在一定程度上具有较强的关联性,不可以将其分离

作为一种单独的对象显示。在数据的处理和运行过程中会使用相关的设备,从而出现不同类型的噪声现象,也是影响数据可视化进一步发展的重要因素。

2. 信息丢失现象严重

在进行数据的可视化转化过程中,会导致相关的关键信息流失,也是数据可视化进程中的困境之一。

3. 大型图像感知方面

数据可视化的实现受到设备方面的长度比和设备分辨率、现实世界的感受方面的制约。在数据可视化的发展与延伸方面受到制约,同时受限于现实,影响大型图像感知的实际效果,也是影响数据中心可视化进程的因素之一。

4. 高速图像变换

用户虽然能够观察数据,却不能对数据强度变化作出反应。用户在数据的分析方面存在局限性,缺少对相关数据深入的认识和研究,不能直接反映图像高速度变换的发展。

5. 高性能要求

数据可视化在静态方面画面的要求极低,因为静态可视化速度方面要求较低,性能标准不高,然而动态可视化对性能要求会比较高。数据处理及转换成静态和动态的画面可视化效果截然相反,需要对两者进行有效的区分和针对性的利用。

6. 数据可视化过程中的表现方式

目前,在数据简约可视化研究中,高清晰显示、大屏幕显示,高可扩展数据投影、维度降解等技术都试着从不同角度解决这个难题。可感知、交互的扩展性是大数据可视化面临的挑战之一。从规模较大的数据库进行查询工作量极大、工作复杂程度高,影响数据交互转换的效率。在大数据实际操作程序中,数据的复杂性和多维度特点会加大数据可视化的难度。不同的数据可视化需要不同的数据中心的建立和完善,推动相关的数据可视化研究和技术的发展。

7. 查看大量数据

数据可视化的一般挑战是处理大量的数据。很多创新的原型仅能处理几千个条目,或者当处理数量更大的条目时难以保持实时交互性。显示数百万条目的动态可视化证明,数据可视化尚未达到人类视觉能力的极限,用户控制的聚合机制将进一步突破性能极限。较大的显示器能够有帮助,因为额外的像素使用户看到更多的细节,同时保持合理的概览。

8. 集成数据挖掘

数据可视化和数据挖掘起源于两条独立的研究路线。数据可视化的研究人员相信让用户的视觉系统引导他们形成假设的重要性,而数据挖掘的研究人员则相信能够依赖传统计算方法和机器学习来发现有趣的模式。使用适当的数据可视化方式,消费者的购买习惯,如

不同种类商品选择之间的相关关系就可以凸显出来。然而,统计实验有助于发现在购买产品的顾客需求或人口统计的连接方面的更微妙趋势。研究人员正逐渐把这两种方法结合在一起。

9. 集成分析推理技术

为了支持评估、计划和决策,视觉分析领域强调数据可视化与分析推理工具的集成。业务与智能分析师使用来自搜索和可视化的数据以及洞察力作为支持或否认有竞争性的假设的证据。他们还需要工具来快速产生他们分析的概要和与决策者交流他们的推理,决策者可能需要追溯证据的起源。

10. 与他人协同

发现是一个复杂的过程,它依赖于知道要寻找什么、通过与他人协同来验证假设、注意异常和使其他人相信发现的意义。因为对社交过程的支持对数据可视化是至关重要的,所以让用户通过数据可视化软件交互协同,会比发给用户、发布到网站等效果更佳。

11. 实现普遍可用性

可视化工具作为面向大众推广的产品,因每个使用者的生活、工作、教育和技术背景参差不齐,实现可视化工具的普遍可用性是设计人员面临的巨大挑战之一。

1.4.3 数据可视化的发展方向

在学科融合发展背景下,数据可视化的发展方向有以下三个方面。

(1) **数据可视化与数据挖掘将联系更紧密。**数据可视化可以帮助人类洞察数据背后隐藏的潜在规律,进而提高数据挖掘的效率。因此,与数据挖掘紧密结合是数据可视化研究的一个重要方向。

(2) **数据可视化与人机交互将联系更紧密。**更好地实现人机交互是人类一直追求的目标,而用户与数据的友好交互,能方便用户控制数据。因此,与人机交互结合是数据可视化的一个重要发展方向。

(3) **数据可视化与大规模、高维度、非结构化数据将联系更紧密。**目前,我们正处在大数据时代,大规模、高维度、非结构化数据层出不穷,要将这些数据以可视化形式完美地展示出来,并非易事。因此,与大规模、高维度、非结构化数据的结合是数据可视化的一个重要发展方向。

1.4.4 数据可视化与其他课程的关系

以大数据管理与应用专业为例,其专业核心课程一般包括数据科学导论、数据库原理、数据结构与算法、数据挖掘与商务智能、大数据分析技术、机器学习、文本分析与文本挖掘、社交网络分析、数据可视化、数据治理、商业模式分析等。图 1-7 利用三层结构表示各门课程之间的关系,其中:基础层的课程包括数据科学导论、数据库原理、数据结构等;应用层的

课程包括数据挖掘与商务智能、数据可视化、文本分析与文本挖掘等；管理层的课程包括数据治理、大数据决策、商业模式分析等。数据可视化作为大数据与应用专业的核心课程，属于应用层的课程之一，该课程主要应用基础层课程的知识，进行数据的处理、分析和展示，为管理层课程提供决策支持。

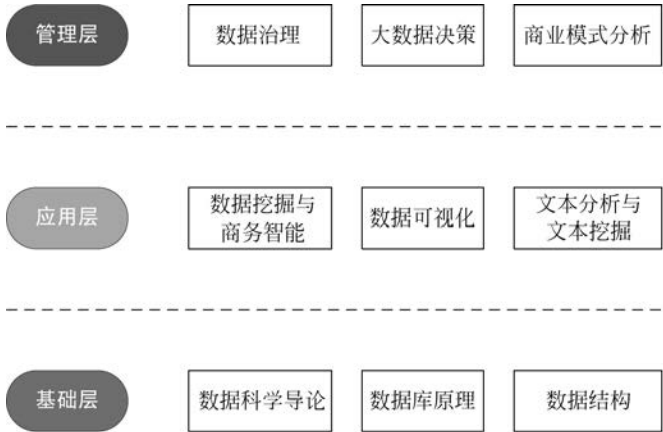


图 1-7 数据可视化与其他课程之间的关系

即测即练



思考题

1. 什么是数据可视化？
2. 数据可视化的原则有哪些？
3. 数据可视化的作用有哪些？
4. 数据可视化的一般过程包括哪些步骤？
5. 简要阐述数据可视化的发展历史。