

第 5 章



大数据可视化技术

本章学习目标

- 了解数据可视化的概念
- 了解数据可视化的作用、分类、基本原则、图表
- 熟练掌握数据可视化分析
- 了解大数据可视化
- 熟练掌握大数据可视化工具

大数据可视化是当前大数据应用研究领域及信息安全领域研究的热点,大数据可视化数据信息可快速和直观地洞见大数据的价值。针对海量的半结构化、结构化和非结构化的数据,数据可视化的实现流程较为复杂,需要先对海量数据进行采集、分析、管理和挖掘等处理,然后再进行表现形式(如立体的、动态的、实时的、交互的等)的设计。大数据可视化是指海量数据的可视化,一般包括结构化、非结构化和半结构化的数据,需要先对数据进行采集、分析、管理和挖掘等处理之后再对数据的表现形式进行设计,实现数据动态化、立体化和直观化地呈现。

大数据可视化与数据可视化有着本质的区别,大数据可视化的数据类型主要是针对结构化数据、半结构化数据和非结构化数据,而数据可视化的数据类型主要针对的是结构化的数据;大数据可视化的表现形式有多种,而数据可视化的表现形式主要为统计图表;大数据可视化的结果主要为发现数据中蕴含的规律特征,而数据可视化的结果为注重数据及其结构关系。大数据可视化的过程主要有数据的可视化、指标的可视化、数据关系的可视化、背景数据的可视化、转换成便于接受的形式、聚焦、集中或汇总展示、扫尾的处理和完美的风格化等 9 个方面。

对数据进行可视化是将不可见的现象转换为可见的图形图像符号,可洞悉数据中隐藏的规律、趋势和价值。数据可视化目前主要应用在医学、金融、能源、交通、公安、文化传媒、航空、气象预报以及工程等领域。

5.1 数据可视化概述

数据通常来说是枯燥乏味的,相对来说,图形和图像以及颜色等更富有生动性和极好的表现力。数据可视化是关于数据视觉表现形式的科学技术研究,利用可视化可将枯燥乏味的数据变为丰富生动的视觉效果,对分析处理后的数据以更加直观的图形化形式展现给用户。在

计算机视觉领域,数据可视化是对数据的一种形象直观的解释,实现从不同维度观察数据,从而得到更有价值的信息。

5.1.1 数据可视化的概念

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,它利用计算机对抽象信息进行直观的表示,以利于快速检索信息和增强知识能力。数据可视化是关于数据视觉表现形式的科学技术研究,它让大数据更有意义,更贴近大多数人,因此,大数据可视化是艺术与技术的结合。数据可视化是一个相对的概念,它通过将数据转换为标识从而为人们提供帮助与指导,并最终成为通过数据分析传递信息的一种重要工具。数据可视化与传统的立体建模之类的特殊技术方法相比,其所涵盖的技术方法要更加广泛。数据可视化是利用计算机图形学及图像处理技术,将数据转换为图形或图像形式显示到屏幕上,并进行交互处理的理论、方法和技术。维基百科将数据可视化定义为:它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。

数据可视化主要是指将数据中隐含的价值通过图形图像形式进行表示,实现简洁高效地传达信息的过程。数据可视化一般在时空数据、地理空间数据、高维非空间数据、层次和网络数据、跨媒体数据等数据方面进行可视化的应用较多。时空数据可视化方面,如气象遥感观测数据、流体力学模拟仿真得到的矢量场和张量场数据以及三维医学图像数据等;地理空间数据可视化方面,如百度地图、GPS 导航、出租车轨迹查询、手机信息跟踪等;高维非空间数据可视化方面,如手机通信日志数据、全国人口调查数据等;层次和网络数据可视化方面,如超大规模集成电路设计、软件过程可视化等方面;跨媒体数据可视化方面,如跨媒体数据中的文本数据、社交网络数据以及日志数据等。数据可视化是综合运用计算机图形学、图像、人机交互等技术,将采集或模拟的数据映射为可识别的图形图像、视频或动画,并允许用户对数据进行交互分析的理论、方法和技术。数据可视化是用户理解复杂分析结果的手段,同时,也是数据分析过程中比较重要的且容易被忽视的步骤。数据可视化的目的是对数据进行可视化处理,以便更加直接地为用户提供数据背后隐藏的价值信息。

数据可视化的标准主要有实用性、完整性、真实性、艺术性和交互性。实用性方面,衡量数据实用性的主要参照是要满足使用者的需求,需要清楚地知道这些数据是不是人们想要了解的与他们切身相关的信息。完整性方面,衡量数据完整性的主要指标是该可视化的数据应当能够帮助使用者全面而完整地理解数据的信息。真实性方面,可视化的真实性考量的是信息的准确度和是否有据可依。艺术性方面,主要是指数据的可视化呈现应当具有艺术性,符合审美规则。交互性方面,主要是指实现用户与数据的交互,方便用户控制数据。

图表是表达数据最直观和最强大的方式之一,通过图表展示能够让枯燥的数字吸引人们注意力。在统计图表中,每一种类型的图表都可包含不同的数据可视化图形,如饼状图、柱状图、折线图、直方图、散点图、面积图、雷达图、箱线图、漏斗图、树状图、气泡图、条状图等。图 5-1 是 EVenn 在线可视化平台可视化示例图效果。

数据可视化可以通过视觉元素透视出数据的信息变化,从中发现和捕捉难以察觉的数据中的关系和变化趋势。如图 5-2 所示的是气泡图示例。

数据可视化是将数据以图形、信息图等以直观形式呈现出来的技术,有助于发现数据中蕴含的规律。例如,根据中国互联网络信息中心(CNNIC)发布的第 54 次《中国互联网络发展状

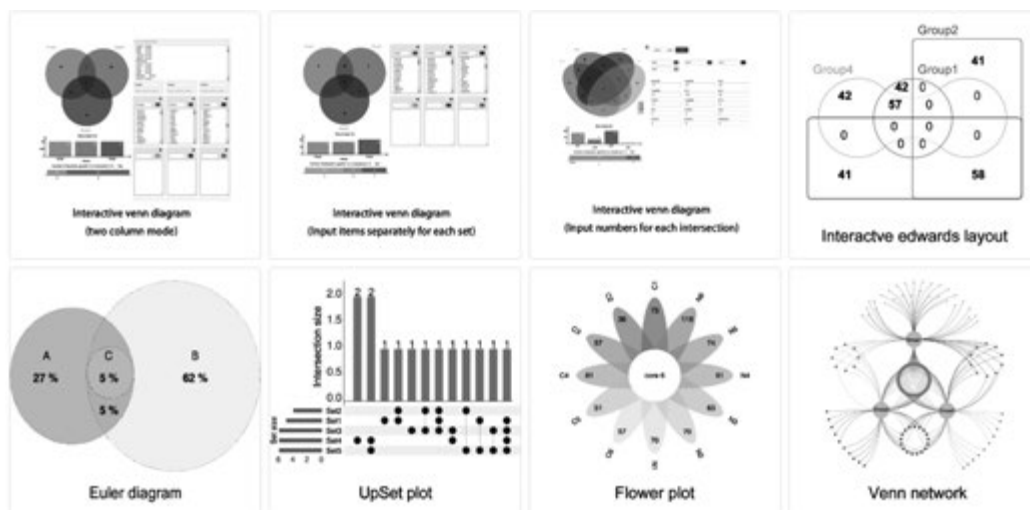


图 5-1 EVenn 示例图

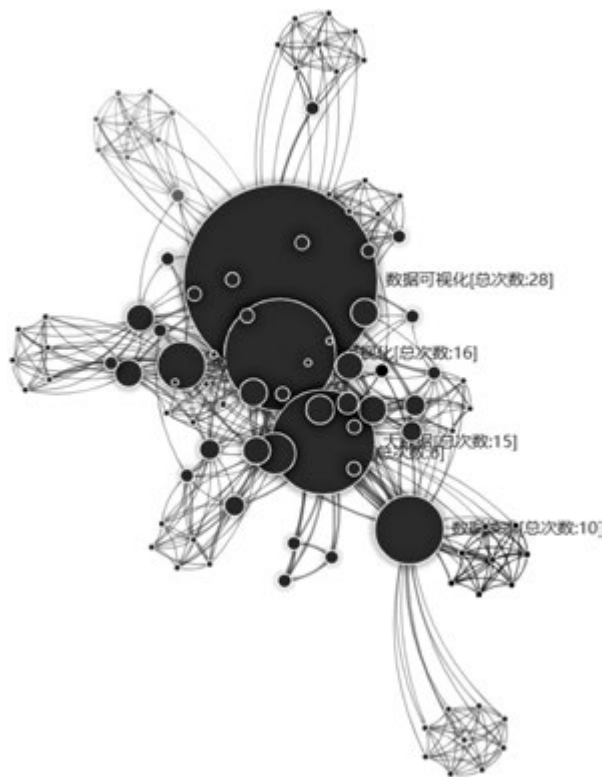


图 5-2 气泡图示例

况统计报告》，2024 年 6 月，手机网民较常使用的 5 类 App 中，即时通信类 App 在 9 点至 20 点间用户使用时段分布较为均匀，占比均在 5% 以上；网络视频类 App 在 12 点、20 点分别出现两次使用小高峰，符合大部分网民闲暇娱乐时间规律；网络购物类、网络支付类 App 的用户使用时段分布趋势较为接近，7 点到 22 点间使用时长合计占比超过 80%；网上外卖类 App 使用时段分布峰值明显，与网民用餐时间关联度很高，分别在 11 点及 18 点出现使用峰值。具体可视化图如图 5-3 所示。

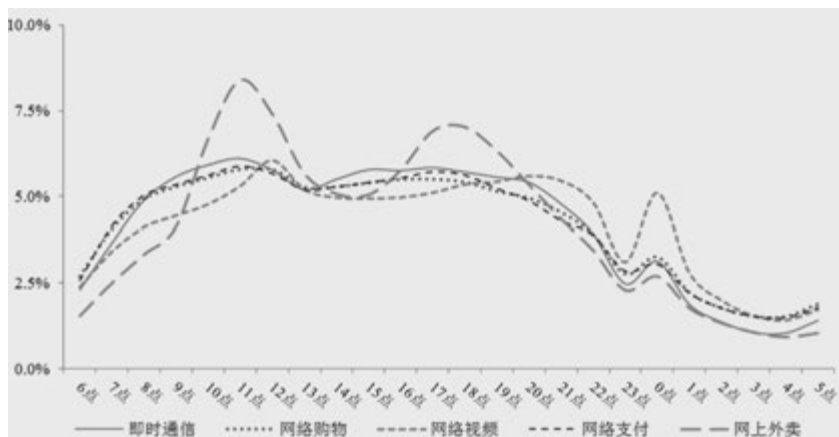


图 5-3 5 类应用使用时段分布可视化图

5.1.2 数据可视化的发展历史

数据可视化的历史与人类现代文明的启蒙以及测量、绘画和科技的发展一脉相承。数据可视化的萌芽源自几何图表和地图的生成,其目的是展示一些重要信息。1700—1799 年,图形符号开始进入数据地图和图表中。到 19 世纪中叶,人类在地图、科学与工程制图和统计图表中,开始有意识地使用可视化理念与技术。在 1853—1856 年克里米亚战争期间,英国女护士南丁格尔通过搜集发现,很多士兵并非战死沙场,而是因为战场外感染了疾病,或是在战场上受伤却没有得到适当的护理。为了解释这个原因,并降低英国士兵的死亡率,她绘制了著名的玫瑰图,如图 5-4 所示。

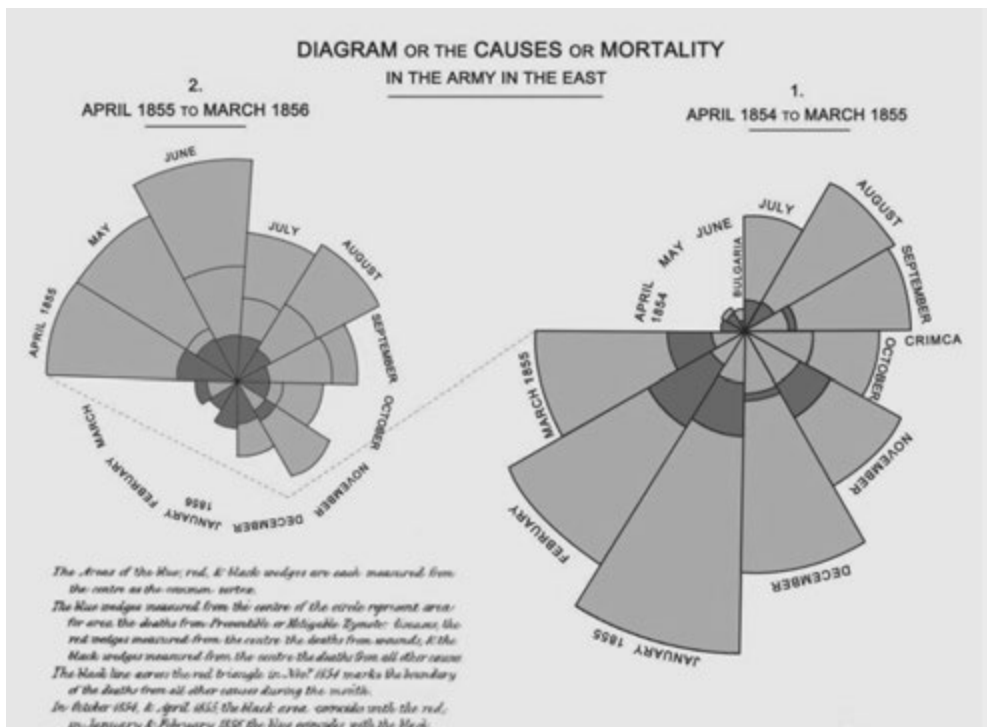


图 5-4 南丁格尔玫瑰图

南丁格尔玫瑰图,是将柱形图转换为更加美观的饼图形式,但区别于饼图,以扇形角度大小及面积表示数据多少,以扇形的半径表示数据的大小,辅助以颜色深浅,美观又清晰地展示出数据的关系。

1879年,Luigi Perozzo 绘制立体图(三维人口金字塔)。此图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,标志着可视化开始进入了三维立体图时代。Luigi Perozzo 绘制的立体可视化图如图 5-5 所示。

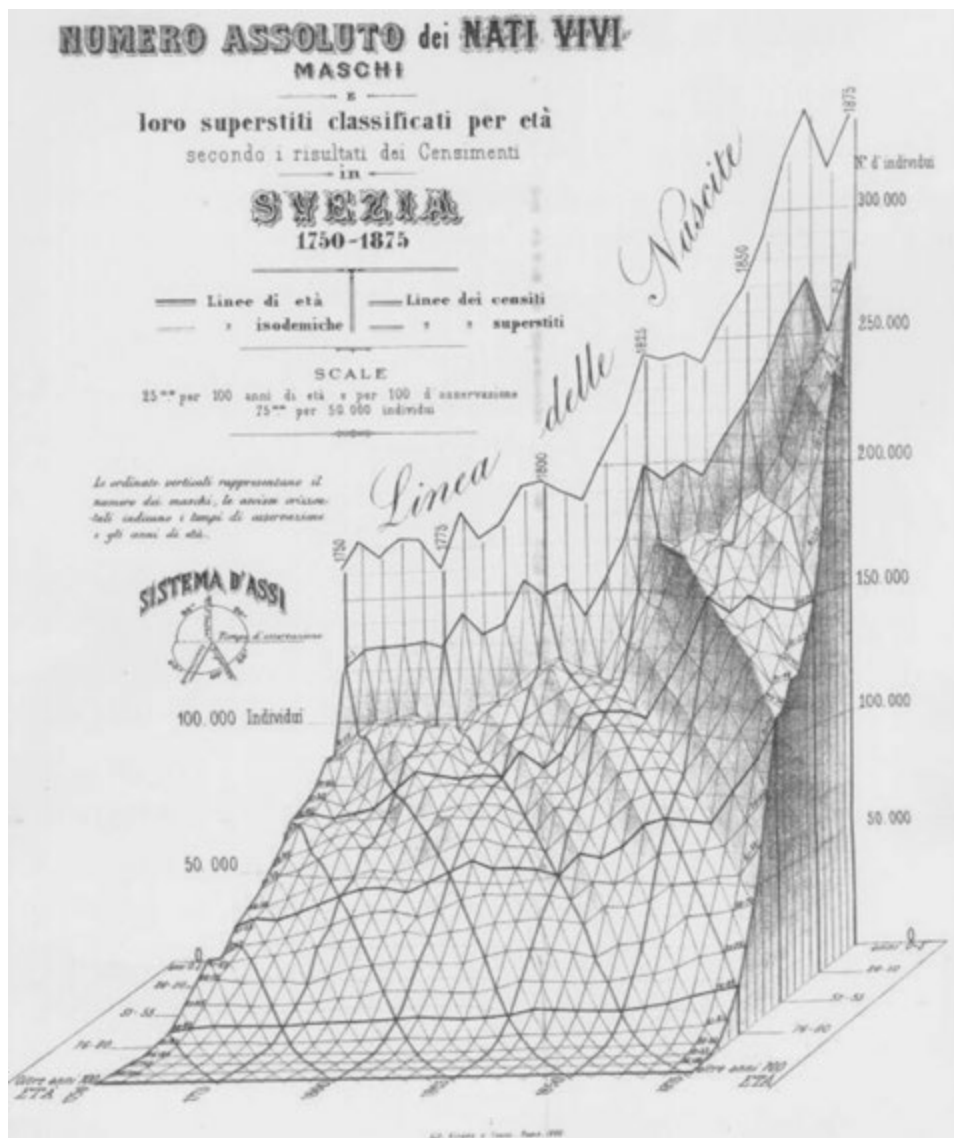


图 5-5 Luigi Perozzo 绘制的立体可视化图

19 世纪末期至 20 世纪初期,随着统计图形的普及,数据可视化开始在政府、商业和科学等领域广泛使用。1904 年,英国天文学家爱德华·蒙德(E. W. Maunder)针对太阳黑子随时间的变化绘制了蝴蝶图,如图 5-6 所示。

20 世纪 20 年代,教育学家纽拉特开发了一套国际印刷图形教育系统(或称为通用视觉语言符号系统)。1931 年,英国工程师 Harry Beck 关于伦敦地铁图的设计是一个比较经典的可视化图实例。

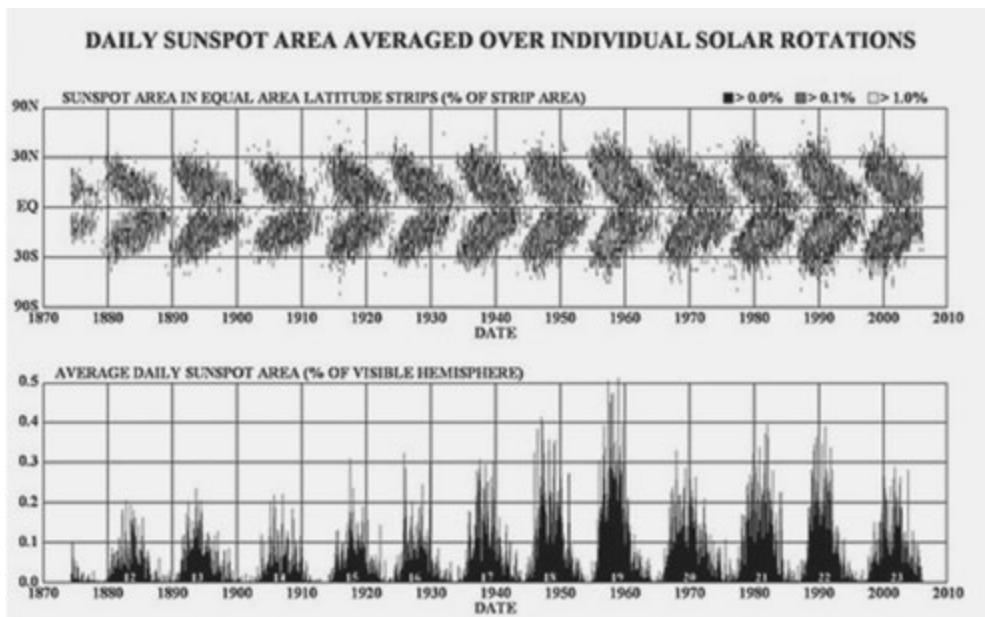


图 5-6 关于太阳黑子随时间扰动的蝴蝶图

1967年,法国设计师雅克·贝尔汀出版了《图形符号学》(*Semiology of Graphics*)一书,并确定了构成图形的基本要素。1972年,德国设计师奥特·艾希尔通过奥运标识系统的设计,奠定了现代图形设计教育的基础。1973年,计算机专家赫尔曼·诺夫发明了表达多维变量数据的脸谱编码。1983年,耶鲁大学统计学教授爱德华·塔夫特出版了专著《定量信息的视觉化》(*The Visual Display of Quantitative Information*),由此奠定了信息可视化的基础。1985年,Alfred Inselberg提出了高维数据的可视化平行坐标图,如图5-7所示。

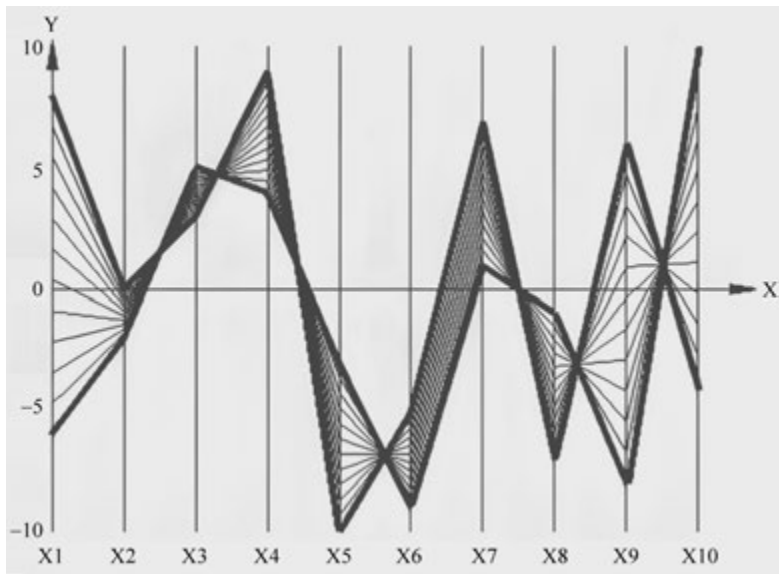


图 5-7 Inselberg 高维数据平行坐标图

1988年,著名的统计图形学者威廉·克莱弗兰在其著作《统计动态图形学》(*Dynamic Graphics for Statistics*)中,给出了多变量统计数据的动态可视化手段。进入21世纪后,现有的可视化技术已难以应对海量、高维、多源异构数据,需要研究人员结合可视化技术、计算机图

形学、数据挖掘理论与方法,提出新的理论模型、可视化方法和交互手段,辅助用户从多源异构数据中快速挖掘出有用的价值信息,为用户决策提供参考依据。

5.1.3 数据可视化的类型

数据可视化是指将大型数据集中的数据以图形图像的形式表示,并利用数据分析和开发工具发现其中未知信息的处理过程。数据可视化的类型主要有科学可视化、信息可视化和可视化分析等多种类型,如图 5-8 所示。



图 5-8 数据可视化类型

1. 科学可视化

科学可视化(Scientific Visualization)是数据可视化类型中的一个主要类型,是科学之中的一个跨学科研究与应用,其应用领域主要有物理学、生物学、化学、气象学、医学以及航空航天等。科学可视化主要关注的是三维现象的可视化,重点在于对体、面以及光源等的逼真渲染。美国计算机科学家布鲁斯·麦考梅克将科学可视化阐述为利用计算机图形学来创建视觉图像,帮助人们理解科学技术概念或结果的那些错综复杂而又往往规模庞大的数字表现形式。1982 年 2 月,美国国家科学基金会在华盛顿召开了科学可视化的首次会议,提出了科学家不仅需要分析由计算机得出的计算数据,而且需要了解在计算过程中的数据变换,而这些都需要借助于计算机图形学以及图像处理技术。科学可视化面向科学和工程领域数据,如三维空间数据、计算机模拟数据和医学影像数据等,重点探索如何以几何、拓扑和形状特征呈现数据中蕴含的规律。利用科学可视化相关软件工具 Origin 可以生成科学可视化效果图。科学可视化示例图如图 5-9 所示。

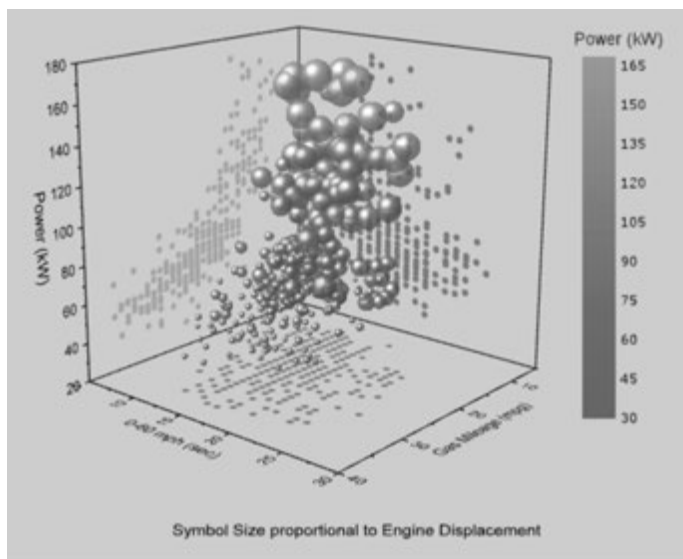


图 5-9 科学可视化示例

按照数据的类别,科学可视化可以大致分为标量场可视化、向量场可视化以及张量场可视化。其中,标量场可视化方面,标量是指单个数值,标量场每个数据点记录一个标量值。标量场可以看成显示数据分布的隐函数表示,即 $f(x, y, z)$ 代表了在点 (x, y, z) 处的标量值。向

量场可视化方面,向量场每个采样点记录一个向量,向量代表某个方向和趋势,如实际测得的风向和漩涡等,向量场可视化主要关注点是其中蕴含的流体模式和关键特征区域。张量场可视化方面,可以分为基于纹理、几何、拓扑三类,基于纹理的方法将张量场转换为一张或动态演化的图像,图示张量场的全局属性,其思路是将张量场简化为向量场,进而采用线积分法、噪声纹理法等方法显示;基于几何的方法显示生成刻画某类张量场属性的几何表达;基于拓扑的方法可有效生成多变量场的定性结构,快速构造全局流畅结构,比较适合数值模拟。

2. 信息可视化

信息可视化是指将数据信息和知识转换为一种视觉形式,在信息可视化中充分利用了人们对可视模式快速识别的自然能力。信息可视化是一个跨学科领域,旨在研究大规模非数值型信息资源的视觉呈现,通过利用图形图像方面的技术与方法帮助人们理解和分析数据。信息可视化的处理对象是抽象的和非结构化的数据集合(如文本和图表以及地图等)。传统的信息可视化起源于统计图形学,与信息图形及视觉设计等现代技术相关,其表现形式通常在二维空间,因此,其关键问题是在有限的展示空间以直观的方式传达抽象信息。与科学可视化相比,信息可视化更加关注于抽象和高维的数据。在工作中使用的如流程图和趋势图等都属于信息可视化,这些图形的设计都将抽象的概念转换为可视化信息。利用信息可视化相关软件工具 Tableau 可以生成信息可视化效果图。信息可视化示例图如图 5-10 所示。

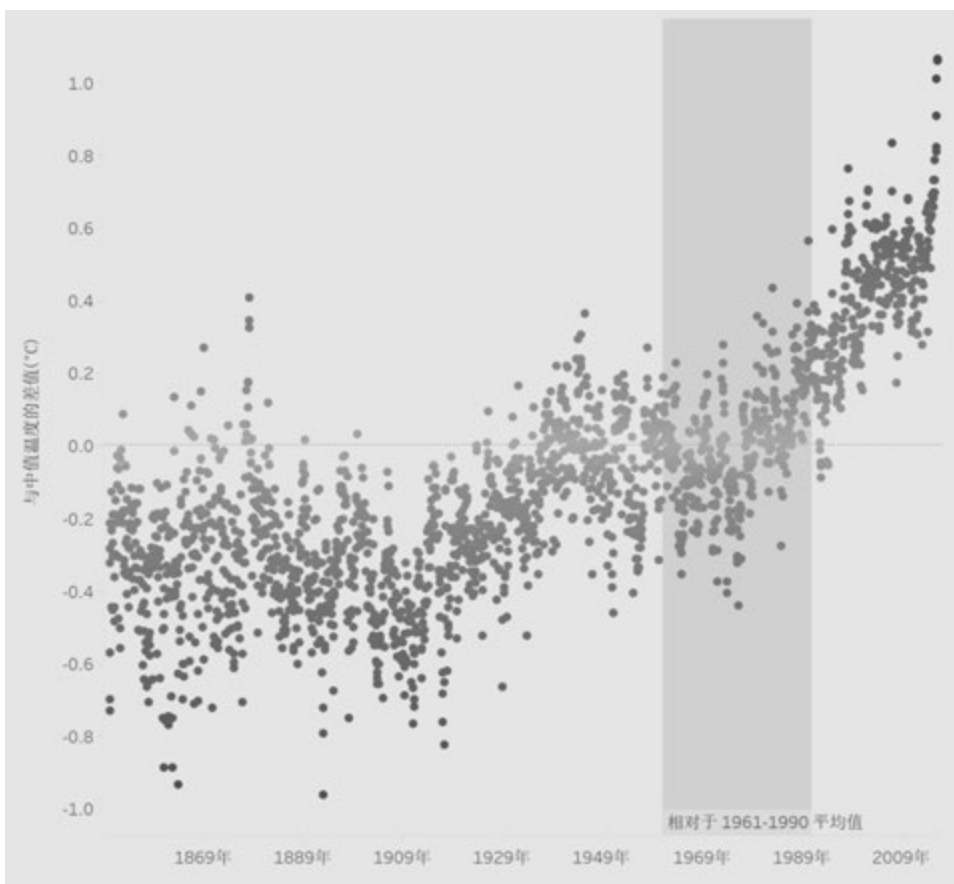


图 5-10 信息可视化示例

3. 可视化分析

可视化分析是科学可视化与信息可视化领域发展的产物,侧重于借助交互式的用户界面进行数据的分析与推理。可视化分析综合图形学、数据挖掘和人机交互等技术,是一个多学科领域。

5.1.4 数据可视化的目标与作用

数据可视化使得数据更加易懂,提高了数据资产的利用效率,更好地支持用户对数据的认知、数据表达、人机交互和决策支持等方面的应用,在教育和医学等领域发挥着重要作用。

1. 数据可视化的目标

数据可视化的目标可以从宏观角度和应用角度来理解。从宏观角度方面来理解,数据可视化的主要目标是信息分析、信息记录以及信息传播。从应用角度方面来理解,数据可视化的主要目标是通过数据可视化来揭示事物内部的客观规律以及数据之间的内在联系,辅助人们理解事物的概念和过程,通过数据可视化来有效呈现数据中的重要特征。

2. 数据可视化的作用

数据可视化是数据加工和处理的基本方法之一,可以通过计算机图形图像技术来更为直观地展示数据,展示数据的基本特征和隐藏于数据背后的有价值信息,辅助用户认识和理解数据。数据可视化的作用主要有数据分析、数据操作以及数据表达。

1) 数据分析

数据分析一般具有比较明确的目标,可以根据数据分析得出的结果做出适当的判断,用来为以后的决策提供依据。数据分析是通过数据计算获得多维、多源、异构和海量数据所隐含信息的核心手段,它是数据存储、数据转换、数据计算和数据可视化的综合应用。可视化作为数据分析的最终环节,直接影响着用户对数据的认知和应用,友好且易懂的可视化结果可以帮助用户进行信息推理和分析,方便用户对相关数据进行协同分析,也有助于信息和知识的传播。

2) 数据操作

数据操作是以计算机提供的界面、接口、协议等条件为基础完成人与数据的交互需求。数据操作需要友好且便捷的人机交互技术和标准化的接口以及通信协议来完成对多数据集的操作。以可视化为基础的人机交互技术快速发展,包括自然交互、可触摸、自适应界面和情景感知等在内的新技术极大地丰富了数据操作方式。

3) 数据表达

数据表达是通过计算机图形技术来更加友好地展示数据信息,以方便用户理解和分析数据。常见的形式主要有图像、文本、图表、网络图、树结构、符号和二维图形等。

5.1.5 数据可视化的理论基础

数据可视化的处理对象是数据,是信息记录与表达、信息推理与分析、信息传播与交互的重要手段。数据可视化是以统计学、计算机图形学、颜色理论、视觉感知、格式塔理论、视觉通道为基础,利用图形学和图像处理技术,将数据转换为图形或图像在屏幕上显示出来,并进行交互处理的理论、方法和技术。

1. 统计学

统计学是进行数据可视化的基础。统计学是通过搜索、整理、分析、描述数据等手段,以达到推断所测对象的本质,甚至预测对象未来的一门综合性科学。数据统计是把握数据全貌,了解数据分布状况的有力工具。数据经过收集和整理后,可以用直观的形式将其展示出来,常用的工具有统计表和统计图。利用统计软件可以生成可视化图形,通俗易懂,并且能够直观地显示数据的变化趋势和各种现象之间的关系。数据的基本统计特征可以分为集中趋势度量、离中趋势度量和数据分布形态。集中趋势度量表示数据的集中位置,寻找数据的中心值或代表值,主要有均值、中位数、众数等。离中趋势度量表示数据的分散程度,反映了数据远离中心值的程度,描述一组数据的波动性,主要有标准差、极差等。数据分布形态指图表化数据后呈现的形态,有助于用户更好地理解数据的特征。

统计学中的数据最直观的表达方式就是图表的呈现。统计图表中最常用的主要有直方图、长条图、圆饼图和折线图等。通过可视化效果图,复杂的多维数据得以更加直观的方式呈现给用户,使得人们能够在短时间内从大量信息中提取出关键模式和趋势。统计图是根据统计数据,利用计算机技术将数据转换为图形图像,展示数据的基本特征和隐含规律,发现数据背后有价值的信息,实现对稀疏、杂乱、复杂、多源异构数据的深入洞察。

2. 计算机图形学

数据可视化起源于计算机图形学。在计算机图形学的基础上,人们开始使用计算机创建图形图表,可视化提取出来的数据,将数据的各种属性和变量呈现出来。美国电气与电子工程师协会(Institute of Electrical and Electronic Engineers, IEEE)对计算机图形学的定义为“Computer graphics is the art or science of producing graphical images with the aid of computer”。

计算机图形学(Computer Graphics, CG)是研究如何利用计算机来处理图形的方法、原理、技术的一门学科。计算机图形学的主要研究内容就是研究如何在计算机中表示图形,以及利用计算机进行图形的计算、处理和显示的相关原理与算法。计算机图形学关注数据的空间建模、外观表达与动态呈现,为可视化提供数据的可视编码和图形呈现的基础理论与方法。计算机图形学在数据可视化中的应用主要体现在将复杂的数据转换为直观的图形表示,帮助用户更好地理解和分析数据。计算机图形学通过使用数学算法和计算机技术,将二维或三维图形转换为计算机显示器上的图像,从而实现对数据的可视化展示。

3. 颜色理论

颜色是由物体发射、反射或透过的光波通过视觉所产生的印象。颜色是数据可视化与视觉设计中的关键要素。颜色包含丰富的信息,比较适合信息编码——数据信息到颜色映射。数据可视化结果的表现力与视觉美感依赖于设计者对于颜色的准确使用。颜色理论在数据可视化中的应用方面,通过使用鲜明的色彩和对比度,可以引导注意力,使用户关注重要的数据或趋势。在数据可视化中,利用颜色理论使图表数据看起来美观,让用户更容易理解数据和记住数据,增强数据信息传递的效率。

4. 视觉感知

视觉感知在数据可视化中起着关键作用。视觉感知是指客观事物通过人的视觉器官在脑中形成的直接反映,是人类认知世界的基础。在数据可视化中,视觉感知帮助用户通过视觉

方式理解和分析数据,从而揭示数据的内在规律和趋势。人类感觉器官包括眼、鼻、耳以及遍布身体各处的神经末梢等,相应的感知能力分别称为视觉、嗅觉、听觉和触觉等,其中,与可视化密切相关的主要是视觉感知。人类感知系统的工作原理取决于对所观察事物的相对判断。在观察时通常会选取参照物,用参照物的外观来对比描述。

视觉感知包括对物体的深度、形状、边界和表面材质的感知。在数据可视化中,这些感知帮助用户理解数据的空间结构和属性。颜色和纹理是视觉感知的重要组成部分。在数据可视化中,颜色和纹理用于表示数据的不同特征,帮助用户区分和识别不同的数据点。格式塔原则是视觉感知的基本准则,主要包括对称性、连接性、闭合性等。这些原则在数据可视化设计中用于提高信息的可读性和理解性。在可视化设计中,设计者需要充分考虑人类感知系统的规律,避免设计的可视化视图存在误导用户的元素。

5. 格式塔理论

格式塔理论起源于格式塔心理学,强调经验和行为的整体性。格式塔理论最基本的法则是简单精炼法则,并认为人们在进行观察的时候,倾向于将视觉感知内容理解为常规的、简单的、相连的、对称的或有序的结构。格式塔法则又称为完图法则,主要包括贴近原则、相似原则、连续原则、闭合原则、共势原则、好图原则、对称性原则、经验原则等。

(1) 贴近原则方面。在数据可视化中,贴近原则强调将视觉元素在空间上相近时,人们通常会将其归为一组。这一原则在数据可视化设计中非常重要,因为它帮助观众理解数据的层次结构和相互关系。通过将相关的数据点、标签和注释靠近在一起,可以确保它们被视为相互关联的部分,从而提升数据的可理解性和视觉效果。

(2) 相似原则方面。人们在观察事物的时候,会自然地根据事物的相似性进行感知分组。在数据可视化中,相似原则的应用主要体现在通过视觉上的相似性来组织信息,帮助用户更好地理解和分析数据。这一原则强调将视觉上相似的元素放在一起,从而帮助用户快速识别模式和关系。

(3) 连续原则方面。在数据可视化中,连续原则是指将相关的数据元素在视觉表示上保持连续性,以使用户可以更容易地识别和理解数据之间的关系。这一原则强调的是将具有相似或相关属性的数据元素在视觉展示上保持一致性和连续性,从而帮助用户更好地理解和分析数据。具体应用场景包括时间序列数据、地理位置数据、流程数据等。例如,在股票价格走势图中,时间轴的连续性表示可以帮助投资者更好地理解股票价格的变动趋势,从而做出更明智的投资决策。

(4) 闭合原则方面。在某些视觉影像中,物体可能是不完整的或不闭合的,然而格式塔心理学认为,只要物体的形状足以表征物体本身,人们会很容易地感知整个物体而忽视未闭合的特征。

6. 视觉通道

可视化将数据属性采用不同的视觉通道进行编码。视觉通道在可视化中的应用是通过将数据信息映射到可视化元素上,以展示数据的各种属性和变量。将数据以可视化图呈现的核心内容是可视化编码,它是将数据映射成可视化元素的技术。可视化编码主要由几何标记(图形元素)和视觉通道两个部分组成。可视化中的几何标记主要是一些几何图形元素,如点、线、面、体等。视觉通道用于控制几何标记的展示特性,为标记提供视觉特征,包括标记的位置、大小、形状、方向、色调、饱和度和亮度等。

- (1) 位置：在可视化中,通过点的位置来展示数据的位置信息。
- (2) 大小：通过点的大小展示数据的数量或强度。
- (3) 形状：通过不同的形状展示数据的类别或属性。
- (4) 方向：通过箭头的方向展示数据的流向或趋势。
- (5) 色调、饱和度、亮度：通过颜色的变化展示数据的不同维度或类别。

视觉通道的类型决定了可视化不同的数据时可能采用的视觉通道,视觉通道的表现力和有效性则可以指导可视化设计者如何挑选合适的通道,实现对数据信息完整而具有目的性的呈现。视觉通道的表现力要求视觉通道准确编码数据包含的所有信息,即视觉通道在对数据进行编码时,要尽量忠于原始数据。使用颜色、形状、大小等多种视觉通道来展示不同的数据维度,例如,在地图上用不同颜色表示不同地区的温度,用大小表示人口数量,用形状表示不同类型的地理特征。以可视化图形式展出,不仅提高了数据的可读性,还增强了信息的传达效果。

5.1.6 数据可视化的主要技术

数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示,大量的数据集构成数据图像,同时将数据的各个属性值以多维数据的形式表示,可以从不同的维度观察数据,从而对数据进行更深入的观察和分析。

数据可视化技术主要有基于几何的技术、面向像素技术、基于图标的技术和基于层次的技术。基于几何的技术主要有 Scatter plots、Landscapes、Projection Pursuit 和 Parallel Coordinates,采用的方法为散点矩阵、平行坐标和多维切片;面向像素技术主要有查询独立和查询依赖,采用的方法为独立于查询和基于查询;基于图标的技术为 Shape Coding、Stick Figures 以及 Chernoff-face,采用的方法为彩色图标;基于层次的技术主要有 Dimensional Stacking、Treemap 以及 Cone Trees,采用的方法为多维堆积图和树图。数据可视化的主要技术比较如表 5-1 所示。

表 5-1 数据可视化的主要技术比较

数据可视化技术分类	关键技术	采用的方法
基于几何的技术	Scatter plots、Landscapes、Projection Pursuit、Parallel Coordinates	散点矩阵、平行坐标、多维切片
面向像素技术	查询独立、查询依赖	独立于查询、基于查询
基于图标的技术	Shape Coding、Stick Figures Chernoff-face	彩色图标
基于层次的技术	Dimensional Stacking、Treemap Cone Trees	多维堆积图、树图

通过比较多种可视化技术可得知,每一种可视化技术都有其不同的关键技术和采用的方法,可使用户从多角度、多层面将可视化技术应用于数据可视化。

5.1.7 数据可视化的流程

数据可视化的实施是一系列数据的转换过程,主要以数据为基础,其核心包括数据采集、数据处理和变换、可视化映射和用户感知等多个步骤,如图 5-11 所示。



图 5-11 数据可视化的流程

1. 数据采集

数据采集是数据可视化流程中的关键一步,是进行数据处理和数据变换的基础。数据可以通过仪器采样及调查记录等方式进行采集,采集的数据涉及数据格式、维度、分辨率和精确度等重要特性,这些对数据可视化的效果起着关键作用。

2. 数据处理和变换

数据处理的目的是提高数据质量。数据可视化之前需要将原始数据转换成用户可以理解的模式和特征并显示出来,因此,数据处理和变换是比较关键的,其流程包括去噪、数据清洗和提取特征等。通过数据处理和转换能够有效保证数据的完整性、有效性、准确性、一致性和可用性。

3. 可视化映射

可视化映射是数据可视化流程中的核心环节,其目的是让用户通过可视化结果去理解数据信息以及数据背后隐藏的数据价值。可视化映射主要用于将不同的数据之间的联系映射为可视化视觉通道中的不同元素,如标记的位置、大小、长度、形状、方向、亮度以及饱和度等。可视化映射是与数据、感知、人机交互等方面相互依托、共同实现的。

4. 用户感知

用户感知是指从数据的可视化结果中提取有用的信息、知识和灵感。可视化映射后的结果只有通过用户感知才能转换为知识和灵感。用户从数据的可视化结果中进行信息融合、提炼、总结知识和获得灵感,并从中发现数据背后隐藏的现象和规律。

5.1.8 数据可视化设计的基本原则

数据可视化设计的基本原则包括美学原则、准确性原则、可读性原则、对比原则、简洁性原则、交互性原则等,这些原则共同作用,确保数据可视化设计既美观又实用,能够有效地传达数据信息。

1. 美学原则

在可视化设计中,图形的构建要素包括坐标轴、网格、布局、形状、色彩、线条等,是实现可视化之美的关键,具体体现是构图美、色彩美和布局美,这些原则在数据可视化设计中起着至关重要的作用,能够提升图表的美观性和易读性,从而更好地传达数据信息。

2. 准确性原则

数据可视化在遵循美学原则的同时,还需要考虑到数据图表的准确性,要忠于原始数据,如实地反映数据信息的变化特征,避免数据被夸大处理,使用户能够准确地获取数据信息。例如,科学研究数据的可视化方面,准确的可视化数据能够帮助研究人员和其他同行验证实验结果的可靠性。

3. 可读性原则

在数据可视化设计中,要考虑文字的大小、颜色是否便于阅读。如果是在移动端展示可视化数据,需要注意适配小屏幕,保证数据和标签能够清晰地展示。

4. 对比原则

在可视化设计中,适当的对比可以使关键数据更加明显,从而引导用户关注特定的数据点或趋势。例如,在交通大数据可视化设计中,使用强烈的颜色对比来突出交通拥堵时的数据异常信息。

5. 简洁性原则

在可视化设计中,要坚持简洁性原则,避免过多的图形信息,选择合适的图形类型和颜色,避免过多的标签和注释。

6. 交互性原则

可视化设计应具有交互性,包括放大、缩小、筛选、排序、联动等功能,使用户能够更好地探索和分析数据。

5.1.9 数据可视化的意义

数据是用来描述科学现象和客观世界的符号记录,是构成信息和知识的基本单元。数据可视化是通过计算机图形图像等技术手段展现数据的基本特征和隐含规律,辅助用户更好地认识和理解数据,进而从多源异构的数据中获得需要的数据信息。数据可视化的意义主要体现在以下4个方面。

(1) 数据可视化有助于提高人们理解与处理数据的效率。

可视化图形相对于传统的数字表格或文本更能够吸引人们的注意力,并让人们能够更快速的理解数据所表达的信息。从人的认知规律来看,人的大脑处理图形、图像等视觉符号的速度远远高于对文字处理的速度。因此,通过将复杂的数据转换为图形图像等直观形式,可以有效提高数据处理的效率。

(2) 辅助数据分析与决策。

在可视化的图形图像中,数据的规律和趋势能够更加直观地呈现出来,通过观察可视化图表,可以更容易地发现数据中的异常值、周期性变化、趋势走向等,从而为深入的数据分析提供方向和线索,有助于决策者及时把握发展态势,做出更加科学和准确的决策。

(3) 实时监控与预警。

通过大数据可视化在线平台,可以对业务数据进行实时监控,及时发现数据的异常值和潜在风险。例如,在工业生产中,可以对生产设备的运行数据进行实时监控,及时发现并预警设备故障问题。

(4) 数据探索与创新。

在数据探索过程中,可视化图形可以让人们更加直观地感受数据的变化和特征,从而产生新的创意和想法,为业务创新提供支持。在科学研究领域,大数据可视化可以帮助科研人员更好地理解和分析实验数据、观测数据等,发现新的科学现象和规律,推动科学研究的进展。

5.2 大数据可视化的关键技术

大数据可视化的关键技术主要包括海量数据采集技术、数据信息的符号表达技术、数据预处理技术、可视化渲染技术、可视化交互技术、数据表达模型技术、智能分析技术等。

5.2.1 海量数据采集技术

数据采集是进行数据可视化的基础。数据采集涉及从多种数据源获取数据,涵盖结构化数据、半结构化数据和非结构化数据。数据采集的准确性和精准性会直接影响到后续可视化的效果。随着大数据技术和人工智能的快速发展,数据采集技术也在不断进步,数据采集逐步向着自动化、智能化、实时化和交互性等方面发展。海量数据采集技术是在数智时代针对海量数据的采集技术,主要包括利用网络爬虫技术实现的数据采集、利用大数据软件工具或大数据平台实现的数据采集等。数智时代,海量数据采集的自动化就是要在数据采集和处理过程中的自动化,减少人工干预,提高数据处理效率。海量数据采集的智能化就是要通过人工智能技术,提高数据处理和可视化的智能化水平,使数据分析更加智能和精准。海量数据采集的实时化就是要实现数据采集和处理的实时性,能够实时获取和处理数据,提高数据分析的时效性。海量数据采集的交互性就是要实现数据可视化的交互性,用户可以通过交互操作,动态调整和分析数据,提高数据分析的灵活性和效果。

5.2.2 数据信息的符号表达技术

数据信息的符号表达技术是一种将数据信息转换为符号信息的技术,以便于人们更好地理解 and 处理。数据信息的符号表达就是要将原始的数据信息转换为更易于理解、处理和传递的符号形式。数据信息的符号表达技术主要包括图形符号技术和色彩符号技术等。图形符号技术主要是指以直观的图形符号来代表信息的技术,涵盖柱状图、条形图、饼图、折线图、散点图、地图符号、箱型图、关系网络图、预测图等。色彩符号技术主要是利用不同颜色来区分不同类别的数据或通过色彩的深浅来表示数据的数量变化的技术。色彩符号技术在大数据可视化中的应用主要有数据分类与分组、数据数值的表达、突出重点数据与异常数据、数据趋势的呈现等。

5.2.3 数据预处理技术

数据预处理技术与数据可视化密切相关,良好的数据预处理是有效可视化的前提。通过数据清洗和标准化处理,可以减少噪声和异常值对可视化的干扰,使得可视化结果更加准确和可靠。在可视化数据集中,如果存在缺失值可能会导致可视化结果出现偏差或者不完整。例如,在分析交通数据时,如果部分日期的交通数据缺失就会在绘制交通流量趋势折线图时出现断层。通常可以采用删除包含缺失值的记录,用均值或中位数等统计量填充缺失值等方法来处理,确保数据的完整性,使可视化能够准确反映数据的真实情况。

通过降维和特征选择可以减少数据的维度和复杂度,提高可视化的效率和性能。通过标准化和缩放处理,可以使不同量级的特征在同一量级上进行比较和分析,增强可视化的解释性和可理解性。数据预处理技术在可视化中的应用不仅可以提高数据的质量和可用性,还可以优化可视化的效果和性能,增强可视化结果的解释性和可靠性。

5.2.4 可视化渲染技术

大数据可视化渲染研究的是如何将可视化后的图形图像更快、更准确地显示出来。渲染是指在计算机图形学中,将三维场景中的模型按照设定好的环境、灯光、材质及渲染参数生成二维图像或视频的过程,其本质是将虚拟的三维世界以逼真的方式呈现在二维屏幕上。渲染过程中需要考虑光照模型、材质属性、相机角度等因素,以确保生成的图像能够准确地呈现出

设计师的意图。随着技术的发展,新的算法和工具不断涌现,使得渲染过程更加高效和逼真。可视化渲染技术包括实时渲染技术、云渲染技术等。实时渲染技术可以将复杂的数据集转换为直观的图形和图表,使用户能够更容易地理解和分析数据。云渲染技术方面,例如,渲云通过强大的并行计算能力,能够处理大规模的数据并生成精细的可视化图表,帮助用户快速生成动态效果图,提升数据分析的效率和准确性。

利用可视化渲染技术可以生成基础可视渲染图、地理空间可视渲染图等。基础可视渲染图主要包括单态图、统计图、分布图、关系图等。单态图用于对数据简单状态的变化进行呈现;统计图常见的有柱形图、折线图等,便于直观地观察数量的多少,比较数据间的差别;分布图可以表现数据的分布态势,如散点图、雷达图等;关系图用于呈现海量的数据间的关联关系。地理空间可视渲染图是在二维、三维态势图基础上叠加可视化分析图表,以呈现数据变化及规律。

在金融领域,可视化渲染技术通过实时渲染生成动态效果图,帮助金融分析师快速生成涵盖股票、债券、外汇等多维度市场波动的图表,这些图表不仅提升了数据分析的效率,还使复杂的数据变得更加直观易懂,帮助金融机构和新闻媒体准确把握市场走势。可视化渲染技术在多个领域中发挥着重要作用,不仅提升了工作效率,还为用户提供了更加直观和逼真的视觉体验。

5.2.5 可视化交互技术

可视化交互技术是一种通过图形化表示和交互式操作,将复杂的数据转换为易于理解的形式,帮助用户探索、分析和理解数据的技术。可视化利用计算机将数据转换成可交互的图形化表示。可视化交互主要是对交互式图表进行缩放、平移、旋转、过滤、选择数据点等操作。通过缩放(放大或缩小)图表,可以查看数据的细节或整体情况。通过平移可以对图表中不同部分的数据进行查看。缩放与平移操作是由鼠标与键盘作为主要交互设备进行控制,平移用于改变信息空间的位置,缩放用于改变信息空间的比例。用户可以根据任务目标对不同尺度层次对象信息密度先进行预定义,当浏览过程中选择该对象时就可以实现自动缩放至目标尺寸,并平移至显示区的中央,以提高用户任务完成时间效率。对于大型信息空间的探索还可以采用平移的同时实现自动缩放技术,一种方法是平移的同时根据平移的速度自动缩放,平移的速度越快则对象缩放得越小;另一种方法是使用滚动条平移的同时不断连续放大缩放的比例,当鼠标离开滚动条时尺度随之减少。

旋转方面,在三维可视化场景中,旋转功能可以让用户从不同角度观察数据对象,更好地理解其空间结构和特征。过滤方面,是指用户可以根据特定的条件对数据进行筛选,只显示符合条件的数据。选择数据点方面,用户可以单击图表中的数据点,获取该数据点的详细信息或与其他数据点进行关联分析。此外,用户还可以对可视化的数据进行标记、编辑和添加注释,以便更好地记录和传达用户对数据的分析结果和想法。

5.2.6 数据表达模型技术

数据表达模型技术在可视化中的应用主要体现在数据建模及可视化系统中,通过多源异构的数据采集、数据治理、模型构建和可视化展示等功能,帮助用户高效管理和分析数据。数据表达模型技术在可视化中的应用不仅提高了数据处理和分析的效率,还通过直观的图形化展示,帮助用户更好地理解 and 传达信息,推动数据科学的发展和应用。

5.2.7 智能分析技术

智能分析技术可以用于预测未来的趋势和结果,帮助用户做出更明智的决策。智能分析技术与交互式可视化工具相结合,可以提供更加动态和交互式的用户体验,用户可以通过拖曳、缩放、过滤等操作,实时查看和分析数据,进一步探索数据中的模式和趋势,这种交互式可视化不仅提高了用户的参与度,还使得数据分析更加灵活和高效。

智能分析技术可以对多维度数据进行综合分析,帮助用户从不同角度理解数据。通过可视化工具,用户可以同时查看多个维度的数据,如时间、地点、类别等,从而获得更全面的数据洞察。在制造业中,智能分析技术可以帮助企业优化生产流程、预测设备故障和维护需求。通过可视化工具,企业可以直观地监控生产线的运行状态,及时调整生产计划。智能分析技术可以协助医生分析病例、预测疾病发展,并通过可视化展示辅助医生和患者更好地理解病情。医疗领域的智能分析系统能够提高诊断的准确性和效率。智能分析技术在可视化中的应用不仅提高了数据分析的效率和准确性,还为用户提供了更加直观和互动的数据洞察工具,推动了各行业的智能化发展。

5.3 大数据可视化方法

大数据可视化通过丰富的视觉效果,将数据以直观、生动和易理解的方式呈现给用户,可以有效提升数据分析的效率。大数据可视化技术涵盖了传统的科学可视化和信息可视化两个方面,它以从海量数据分析和信息挖掘为出发点,信息可视化技术将在大数据可视化中扮演着重要角色。根据信息的特征可以把信息可视化技术分为一维、二维、三维、多维信息可视化,以及层次信息可视化、网络信息可视化和时序信息可视化。大数据可视化方法主要包括文本数据可视化方法、关系数据可视化方法、网络数据可视化方法、空间信息可视化方法、时间数据可视化方法和复杂数据可视化方法等,如图 5-12 所示。



图 5-12 大数据可视化

5.3.1 文本数据可视化

文本信息是大数据时代非结构化数据类型的典型代表,是互联网中最主要的信息类型。与图形、语音和视频信息相比,文本信息体积更小、传输更快,并且更容易生成。将互联网中广泛存在的文本信息用可视化的方式表示,能够更加生动地表达蕴含在文本中的语义特征,如词

频、动态演化规律等。因此,针对一篇文章,文本可视化能够更快速地告诉读者文章在讲什么;针对社交网络上的发言,文本可视化可以帮助读者将所有信息归类;针对一系列的文档,读者可以通过文本可视化找到它们之间的联系;针对一个新闻,文本可视化可以帮助读者捋顺事情发展的脉络等。图 5-13 展示了文本可视化的示例。



图 5-13 文本可视化示例

1. 文本可视化的类型

文本可视化的类型,除了包含常规的图表类,如饼图、柱状图以及折线图等形式,在文本领域使用较多的还有基于文本内容的可视化、基于文本关系的可视化、基于多层面信息的可视化等。其中,基于文本内容的可视化研究包括基于词频的可视化和基于词汇分布的可视化,常用的可视化形式有词云、分布图和 Document Cards 等;基于文本关系的可视化主要研究文本内外的关系,帮助人们理解文本内容和发现规律,常用的可视化形式有树状图和节点连接的网络图以及叠式图等;基于多层面信息的可视化主要研究如何结合信息的多个方面帮助用户从更深层次理解文本数据,发现其内在规律,如地理热力图以及基于矩阵视图的情感分析可视化等。

2. 文本可视化的流程

文本可视化流程主要包括信息收集、文本信息挖掘、视图绘制和交互设计等多个部分。文本可视化流程如图 5-14 所示。



图 5-14 文本可视化流程

文本信息挖掘等技术充分发挥了计算机的自动处理能力,将无结构的文本信息自动转换为有结构信息,而视图绘制使人类视觉认知、关联和推理的能力得到充分的发挥。文本信息挖

掘依赖于自然语言处理,主要通过分词、抽取和归一化等操作提取出文本词汇及相关内容。视图绘制方面,可视化呈现是将文本分析后的数据用视觉编码的形式来处理,其中涉及的内容有尺寸、颜色、形状、方位和纹理等,并使用各种图表来描述。交互设计方面,为了使用户能够通过可视化有效地发挥文本信息的特征和规律,通常在可视化设计中根据使用的场景为系统设置一定程度的交互功能。

3. 文本可视化的实现

文本可视化的实现需要经过多个步骤才能完成,其主要步骤为:首先要在文本中进行分词计算,提取关键词,并去掉冗余的文字;其次是要为提取出来的关键词计算权重,即决定哪些词着重显示,一般来说,权重较高的词会显示在较引人注目的地方;最后是为可视化显示布局,在布局中要计算出每个词或英文单词的摆放位置,并最终呈现在用户面前。

5.3.2 关系数据可视化

关系数据可视化是指将数据之间的关系通过图形等直观方式展现。关系数据可视化能够更直观地展示复杂的数据关系,帮助用户发现数据中的模式、趋势和关联,从而做出更明智的决策。

在社交网络分析中,关系数据可视化可以帮助研究人员理解社交网络的结构和动态。在生物信息学中,关系数据可视化用于展示基因、蛋白质之间的相互作用网络。通过绘制节点和边,可以展示基因表达、蛋白质相互作用等复杂关系,帮助研究人员理解生物系统的复杂性和动态变化。在交通流量管理中,关系数据可视化可以展示交通流量的动态变化和拥堵情况。通过绘制节点和边,可以展示不同路段的交通流量、拥堵情况,帮助交通管理部门进行流量调控和拥堵缓解。在市场营销中,关系数据可视化可以展示客户之间的关系、购买行为等。通过绘制节点和边,可以展示客户群体的细分、购买行为模式等,帮助企业进行精准营销和客户管理。此外,还有货物贸易关系数据可视化等,如利用 Origin 软件绘制的全球木材贸易关系可视化数据,如图 5-15 所示。

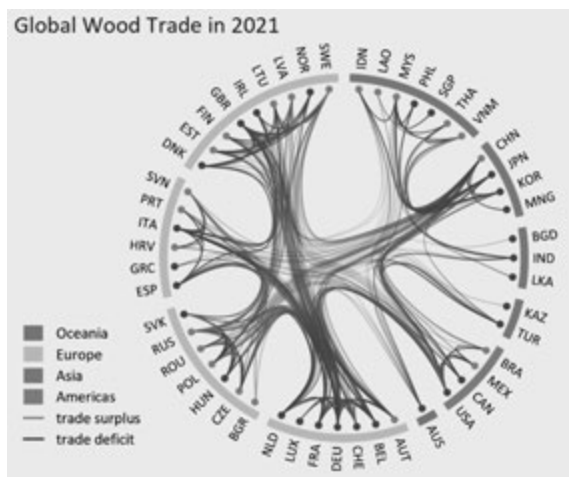


图 5-15 关系数据可视化示例

5.3.3 网络数据可视化

网络可视化通常展示数据在网络中的关联关系,一般主要用于描绘互相连接的实体,如社交网络等。社交网络是一个网络型结构,其典型特征是由节点与节点之间的连接构成,这些节

点通常代表个人或组织,节点之间的连接关系有朋友或亲属关系、关注或转发关系、支持或反对关系,或者拥有共同的兴趣爱好等。层次结构数据也属于网络信息的一种特殊情况。图 5-16 展示的是网络可视化示例。

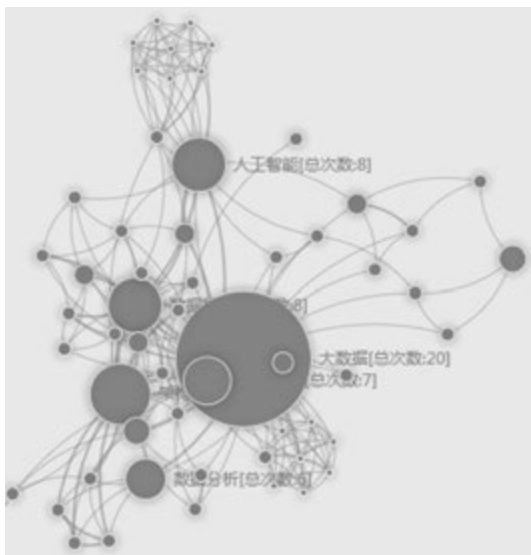


图 5-16 网络可视化示例

5.3.4 空间信息可视化

空间信息可视化源于可视化研究的基本原理。空间信息可视化是指运用计算机图形图像处理技术将复杂的科学现象和自然景观及一些抽象的概念图形化的过程。具体地说,是利用地图学、计算机图形图像技术,将地学信息输入、查询、分析、处理,采用图形、图像,结合图表、文字、报表,以可视化形式,实现交互处理和显示的理论、技术和方法。

空间信息可视化的目的是以可视化方式显示输出空间信息,通过视觉传输和空间认知活动,去探索空间事物的分布及其相互关系,以获取有用的知识,并进而发现规律。空间信息可视化的主要表现形式有动态地图、多媒体信息、三维仿真地图和虚拟现实等。其中,动态地图是一种处于运动状态的数字地图,借助于计算机综合处理多种媒体信息的功能,将图形、图像、动画、声音和视频技术相结合,使多种信息逻辑地连接并集成为一个有机的具有人性化操作界面的空间信息传输系统。多媒体信息主要是使用图形、图像、视频、动画等各种形式综合形象地表现空间信息,是空间信息可视化的重要形式。三维仿真地图主要是利用地图动画技术直观而又逼真地显示地理实体运动变化的规律和特点。虚拟现实主要是以视觉为主,并结合触觉、听觉等来感知环境,使人们犹如进入真实的地理空间环境之中并与之发生交互作用。

5.3.5 时间数据可视化

时间是一个非常重要的维度和属性。时间数据也称为时间序列数据,即同一现象或数据在不同时间点或时间段的数据序列。著名的语言学家拉科夫和约翰逊在经典著作《我们赖以生存的隐喻》(Metaphors We Live)一书中解释说,我们往往是用运动物体或容器的概念来描述时空体验的。历史时间通常是以时间线的图形形式表示的,时间线是相关历史事件发生的时间顺序的标尺。时间数据可视化的方法主要有时间序列图、柱状图、散点图、折线图、面积图等。

时间序列图是将多个时间点的数据以线段连接,从而形成一系列连续的线段图形。通过时间序列图,用户可以快速发现数据中存在的一些基本模式,如趋势、周期性等。时间序列,通俗的字面含义为一系列历史时间的序列集合。时间序列通常由趋势、季节性、残差等部分组成。趋势是指时间序列在较长一段时间内呈现出来的持续向上或者持续向下的变动;季节性是指时间序列在一年内重复出现的周期性波动,如气候条件、生产条件、节假日等;残差也称为不规则波动,指除去趋势、季节性、周期性外的随机波动。

柱状图可以用来比较两个及两个以上可比较对象的数值大小。柱状图能够直观地看出各个数据的大小,易于比较数据之间的差别,能够清楚地展示出数量的多少。例如,中国互联网络信息中心(CNNIC)发布的第43次《中国互联网络发展状况统计报告》中关于2011—2018年网民平均每周上网时长的统计图中就以柱状的可视化图形展示,如图5-17所示。



图 5-17 柱状可视化图示例

散点图是数据点在直角坐标系平面上的分布图,可以用来粗略观察数据点的分布情况,能够大致推断出变量间的关系,包括正相关、负相关、不相关;是线性还是指数或U形关系;相关性是强还是中或者弱。在散点图的基础之上可以进一步做量化分析,得出比较精准的结果。时间散点图按连续时间间隔展示数据点。时间散点图提供了针对基本统计信息和检验的时间序列数据的特征的初步分析,因此在建模之前生成了有关数据的有用洞察。时间散点图包含分解、扩展 Dickey-Fuller 检验(ADF)、相关性(ACF/PACF)和谱分析之类的分析方法。时间散点图通过在二维坐标系中绘制数据点,直观展示数据随时间的变化情况,便于观察和分析。通过时间散点图,可以识别数据的总体趋势、季节性变化和异常值,帮助分析人员发现潜在的规律和问题。时间散点图能够提供丰富的信息,帮助决策者制定科学的策略和方案,提高决策的准确性和有效性。时间散点图主要应用于金融市场分析、市场营销以及气象研究等。

折线图也称为趋势图。根据数量的多少描出各点,然后将各点用线段顺次连接起来形成折线图。折线图的特点是易于显示数据变化趋势及变化幅度,可以直观地反映这种变化以及各组之间的差别。折线图的绘制与散点图相似,只是需要用线条将圆点连接起来,以便观察数据的变化趋势。例如,中国互联网络信息中心(CNNIC)发布第45次《中国互联网络发展状况统计报告》中关于城乡地区互联网普及率就可以用折线图的形式进行可视化效果展示,如图5-18所示。

折线图利用线段可以描述数据在时间上的变化趋势,如图5-19所示。

利用 Origin 软件可以绘制出时间数据变化的可视化棒图,如图5-20所示。

面积图又称为区域图,是在折线图基础之上形成的。将折线图中折线与自变量坐标轴之



图 5-18 城乡地区互联网普及率可视化效果图

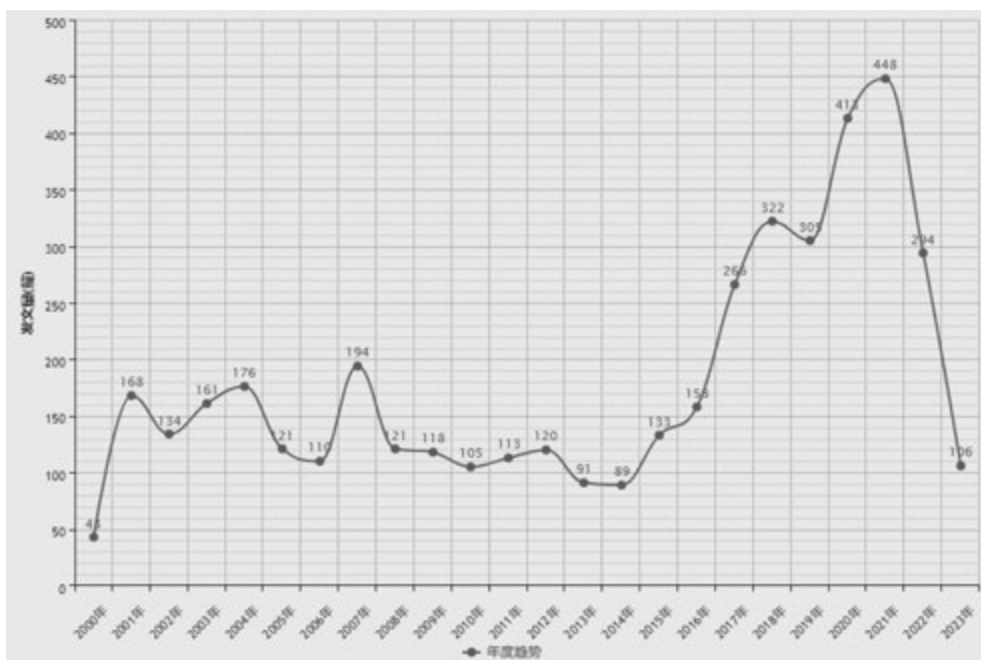


图 5-19 时间数据折线图示例

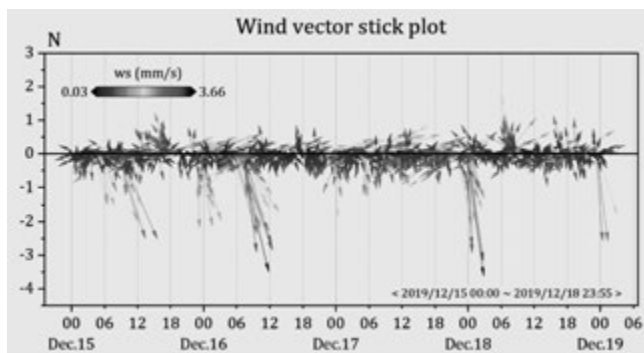


图 5-20 风矢量可视化棒图

间的区域使用颜色进行填充,就形成了面积图。面积图可以更好地突出趋势信息。利用 Origin 软件绘制的面积图可视化示例如图 5-21 所示。

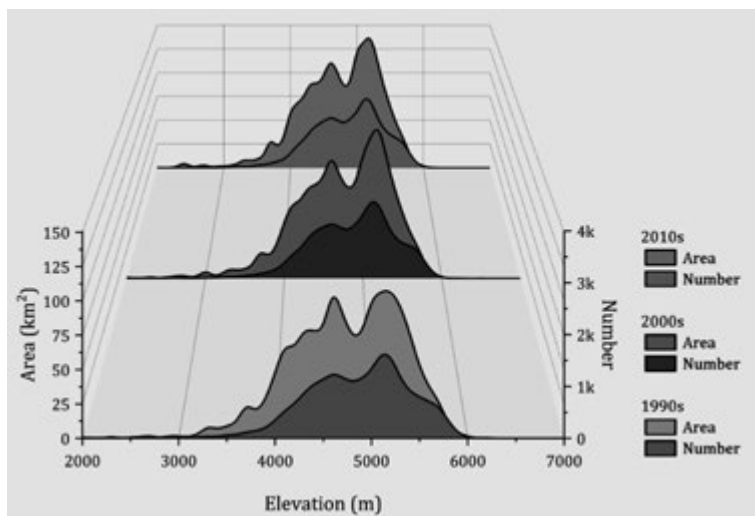


图 5-21 面积图可视化示例

5.3.6 复杂数据可视化

复杂数据可视化是指将海量、多维度、异构、复杂的数据以直观易懂的图形方式展示出来的方法。复杂数据包括海量的视频影像数据、网络数据、空间信息数据等,对该类复杂数据进行分析处理、呈现和应用是当前大数据技术领域研究的重要课题。复杂数据往往是抽象的,如金融交易数据、气象变化数据、空间信息数据等。将复杂数据以图表、图形、信息图等方式呈现,用户可以快速直观地了解数据的模式、规律、趋势、关系等关键信息。例如,在航空航天领域,由于航空航天需要对广阔的空间进行探索,数据量大且要求高,大数据可视化可以帮助进行测发、测控设备控制、战斗管理体系仿真模拟演练、战斗评定等。通过可视化手段,可以实现对航天飞机数据的统计分析、情况监控等。多维数据可视化示例如图 5-22 所示。

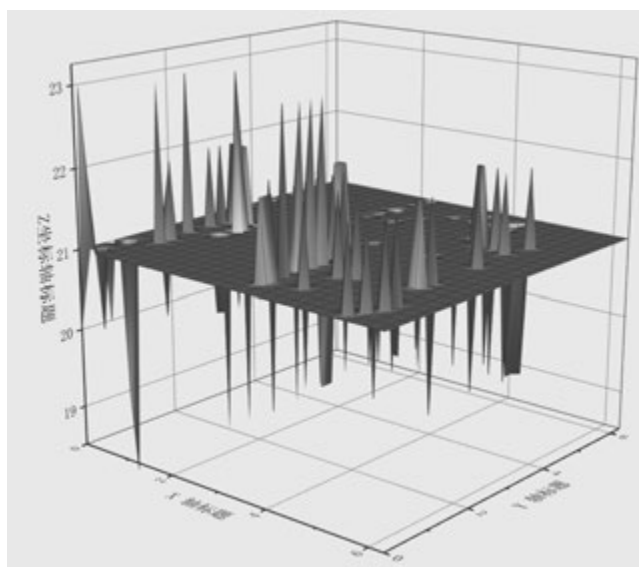


图 5-22 多维数据可视化

5.4 数据可视化平台

数据可视化是对数据的一种形象直观的解释,从不同的维度来观察数据,可以获得用户需要的价值信息。数据可视化将抽象的、复杂的、不易理解的数据转换为人眼可以识别的图形、图像、符号、颜色、纹理等,这些转换后的数据通常具备较高的识别效率,能够有效地传达出数据本身所包含的有用信息。比起枯燥乏味的数值,用户能够更好、更快地认识大小、位置、形状、颜色等物体的外在直观表现。数据可视化平台通过将复杂的数据以图形或图像的形式呈现,使得数据更加易于理解和分析。

5.4.1 ImageGP 平台

ImageGP 是一个专为生物信息学设计的在线作图可视化平台,提供了多种图形绘制选项和高级分析功能,能够帮助相关研究人员简化数据处理过程,提高研究效率。ImageGP 以其全面的功能和用户友好的界面,成为生物信息学领域的可视化平台工具。ImageGP 可以在线生成常见的柱状图、堆积柱状图、线图、散点图、箱线图、韦恩图、集合图、直方图等多种可视化图形。ImageGP 的操作步骤如下。

(1) 访问 ImageGP 平台。用户首先需要访问 ImageGP 的在线平台,打开后的页面如图 5-23 所示。

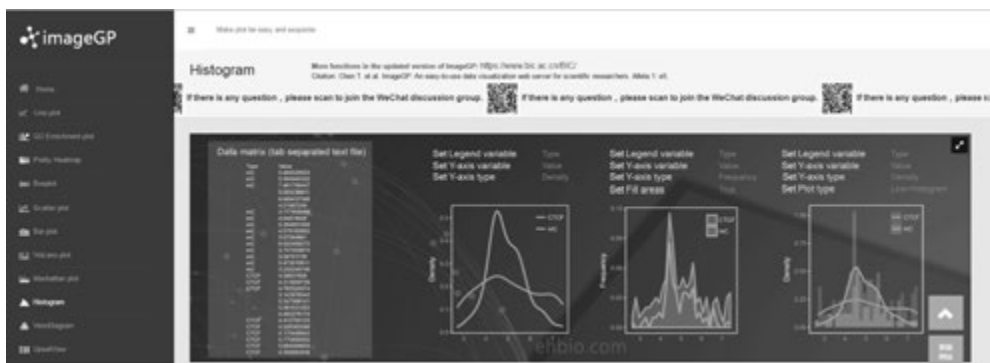


图 5-23 ImageGP 平台首页

(2) 选择需要生成的图形类型。在线平台提供了丰富的图形选项,可以根据数据类型和需求选择合适的图形。

(3) 在线上传数据和调整参数。可以根据所选图形的需要上传相应的数据文件,并根据需求调整图形的参数设置。

(4) 生成和下载图形。完成设置后,可以预览图形,并进行进一步的调整,直至满意为止,最后可直接在平台上生成可视化图形效果。

5.4.2 EVenn 平台

EVenn 是一个在线可视化平台,能够支持多种图表类型。EVenn 在线可视化平台首页如图 5-24 所示。

EVenn 图是可以用于分析和展示多组数据集的交集、并集和差集关系的可视化图。在 EVenn 平台首页,选择 Input numbers for each intersection 选项,打开该页面,选择数据集数量,在数据集数量下方的文本框中输入相应数据,即可完成多组数据集关系的可视化效果图。

EVenn 多组数据集关系的可视化图如图 5-25 所示。

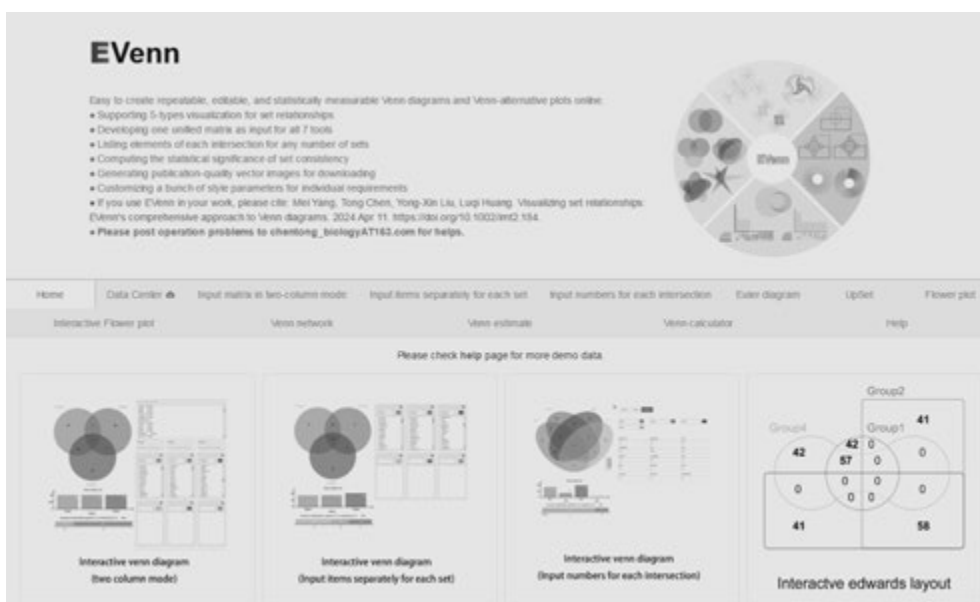


图 5-24 EVenn 平台首页



图 5-25 EVenn 多组数据集关系的可视化图

5.4.3 华为云可视化平台

华为云数据可视化平台是一个在线可视化平台,适配云上云下多种数据源,提供丰富多样的 2D、3D 可视化组件,采用拖曳式自由布局,旨在帮助用户快速定制和应用属于自己的数据大屏。华为云可视化平台首页如图 5-26 所示。

华为云数据可视化平台提供了丰富的可视化组件,包括常用的数据图表、图形、控件等,支持绘制热力分布、地域区块、3D 地球等效果,支持地理数据多层叠加,拖曳即可完成组件自由配置与布局、所见即所得,无须编程就能轻松搭建可视化大屏,并且依据投放设备分辨率,自由



图 5-26 华为云可视化平台首页

定制大屏尺寸,无缝集成华为云数据仓库服务、数据湖探索、关系数据库、对象存储服务等,支持本地 CSV、在线 API 及企业内部私有云数据。

单击华为云可视化平台首页右上角的“注册”按钮,进入注册页面,填写手机号进行注册账户,注册完成后就可登录进入华为云数据可视化平台,进入后可以打开相关示例,如图 5-27 所示。



图 5-27 华为云平台在线可视化效果示例

5.4.4 腾讯云图数据可视化平台

腾讯云图数据可视化(Tencent Cloud Visualization)平台是在线数据可视化展示平台,旨在帮助用户快速通过可视化图表展示大量数据,低门槛快速打造出专业大屏数据展示。精心预设多种行业模板,极致展示数据魅力。采用拖曳式自由布局,全图形化编辑,快速可视化制作。腾讯云图数据可视化支持多种数据来源配置,支持数据实时同步更新,同时基于 Web 页面渲染,可灵活投屏多种屏幕终端,不必预先导入数据,用户可以直接将所要呈现的组件拖曳到画布上进行自由配置和布局,通过单击拖曳即可调整图层顺序,全程图形化编辑操作,所见即所得。腾讯云图数据可视化支持从静态数据、CSV 文件、公网数据库、腾讯云数据库和 API 等多种数据源接入数据,通过配置可以动态实时更新。统一的资源管理使得数据和图片可以

方便快捷地重复使用。腾讯云图数据可视化首页如图 5-28 所示。



图 5-28 腾讯云图数据可视化首页

5.4.5 百度 Sugar BI 数据可视化平台

Sugar BI 是百度智能云推出的敏捷 BI 和数据可视化平台,支持直连多种数据源(Excel/CSV、MySQL、SQL Server、PostgreSQL、Oracle、GreenPlum、Kylin、Hive、Spark SQL、Impala、Presto、Vertica 等),还可以通过 API、静态 JSON 方式绑定可视化图表的数据,简单灵活。大屏与报表的图表数据源可以复用,用户可以方便地为同一套数据搭建不同的展示形式。支持丰富的图表组件(折线、柱图、饼图、拓扑图、地图、3D 散点图等)和过滤条件(单选、多选、日期、输入框、复杂逻辑等),酷炫的 3D 地图效果,充分满足复杂的可视化需求。百度 Sugar BI 数据可视化平台网站首页如图 5-29 所示。



图 5-29 百度 Sugar BI 数据可视化平台网站首页

Sugar BI 基于百度 Echarts 提供丰富的图表组件,不需要 SQL,全流程智能化操作,用户可不写一行代码,分钟级即可完成自助 BI 报表分析和可视化大屏,提升数据分析效率,帮助企业精准快速决策。百度 Sugar BI 数据可视化平台在线可视化示例如图 5-30 所示。



图 5-30 百度 Sugar BI 数据可视化平台在线可视化示例

5.5 大数据可视化工具

大数据时代,利用传统的可视化技术已很难满足对海量、高维、多源和动态数据的分析,需要利用大数据可视化工具来实现对海量数据的可视化。大数据可视化工具主要有 Tableau、Infogram、ChartBlocks、Datawrapper、Plotly、RAW、Visual.ly、Ember、Charts、NVD3、Google Charts、FusionCharts、Highcharts、Chart.js、Leaflet、Chartist.js、n3-charts、SigmaJS、Polymaps、Processing.js 以及大数据魔镜等。利用 Tableau 可对海量数据信息进行快速分析及可视化呈现,其中,中国东方航空公司利用 Tableau 分析市场研究、优化“出发地到目的地”航线并实现了增收。使用 Tableau 一年之后,该航空公司的营业收入就增加了两亿美元,增幅达到了百分之二。

大数据可视化常用的工具主要有 Tableau、DataWatch 和大数据魔镜。Tableau 的数据源为 Oracle、MySQL、SAP、SAS、QlikTech、Text、Files、Microsoft Access、Netezza、ParAccel 和 IBM DB2。数据处理过程为数据采集—处理—计算—挖掘—可视化呈现,作用为实现交互的、可视化的分析和仪表盘分析,应用领域为商务服务、能源、电信、金融服务、医疗保健、制造业、媒体娱乐和教育。开发机构为 Tableau Corporation。DataWatch 的数据源主要有 Active MQ、One tick CEP、Oracle CEP、SAP、Sybase、StreamBase、KXkdb+、SAP Sybase、IQ、Apache Qpid、SAP、Netweaver Gateway。数据处理过程为数据采集—处理—计算—挖掘—可视化呈现,作用为快速分析多维度数据,将多种数据源(包括顶级的 Database、kdb+、OData)可视化,更立体地分组显示数据,更流畅地显示快速变化的数据,快速地挖掘潜在的实时性的数据信息,应用领域主要在金融、零售业、医疗。开发机构为 DataWatch Corporation。大数据魔镜的数据源有 SQL Server、Oracle、Access、NoSQL、MongoDB、DB2、Spark、Google Analytics,以及微信、微博、淘宝、京东等第三方社会化数据源。数据处理过程为数据采集—

处理—计算—挖掘—可视化呈现,作用为全景分析、跨表分析和多源图表、内存分析。魔镜通过数据获取、数据清洗、数据整合的技术,针对企业不同需求,为企业建立数据仓库,包括传统数据仓库、Hadoop 数据仓库、新一代动态数据仓库,应用领域包括电商、制造业、政府、金融、医疗、银行、保险、电信、高校、大中型企业,如中石油、一号店、中国移动等大型知名企业,开发机构为苏州国云数据科技有限公司。

5.5.1 Tableau

Tableau 是一款数据可视化的商业智能工具,可以用仪表盘、工作表的形式实现交互式 and 可视化,允许非技术用户轻松创建自定义的仪表盘,为用户提供更好的数据观察角度。Tableau 的操作十分简单,使用者不需要精通复杂的编程和统计原理,只需要把数据直接拖放到工作簿中,通过一些简单的设置就可以得到自己想要的数据可视化图形。Tableau 侧重于数据分析可视化层级,可以实时动态、人机交互地分析数据,探索规律和查找问题。Tableau 官方网站首页如图 5-31 所示。

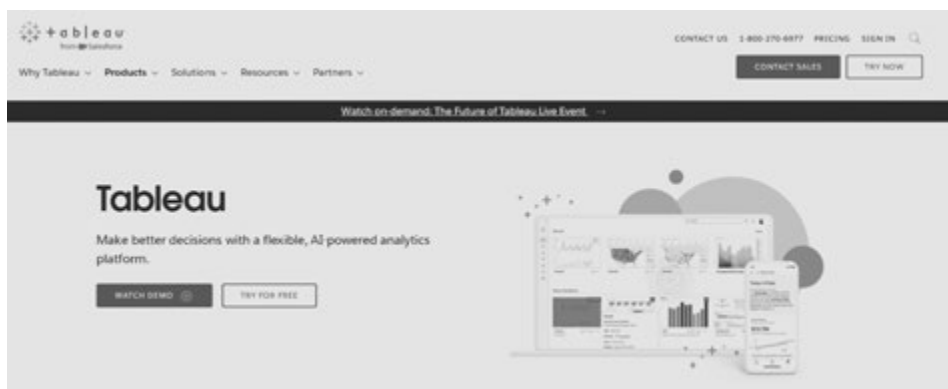


图 5-31 Tableau 官方网站首页

Tableau 可以将各种图表整合成仪表盘在线发布,但为此必须公开自己的数据,把数据上传到 Tableau 服务器。用户可以基于软件 UI 交互创建和发布交互式及可共享的仪表板,以图形和图表的形式描绘数据的趋势、变化和密度。它支持连接本地或云端数据、关系数据源和大数据源来获取及处理数据。利用 Tableau 可以绘制各种精美的图表,图 5-32 显示了利用 Tableau 绘制的“细分市场每月销售额”图。图 5-33 显示了利用 Tableau 绘制的“每小时天气趋势”图,图 5-34 显示了利用 Tableau 绘制的条形图,图 5-35 显示了利用 Tableau 绘制的折线图,图 5-36 显示了利用 Tableau 绘制的分布类图。

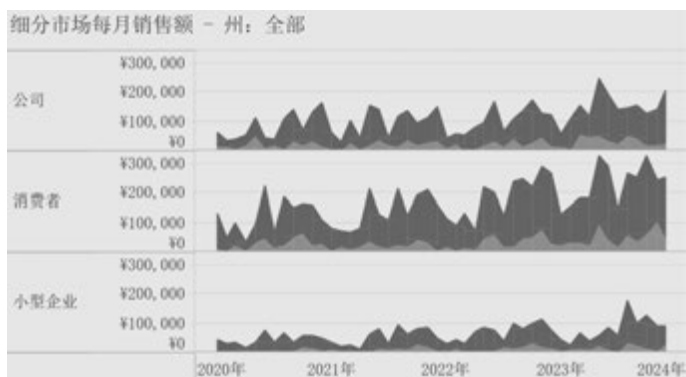


图 5-32 利用 Tableau 绘制的“细分市场每月销售额”图

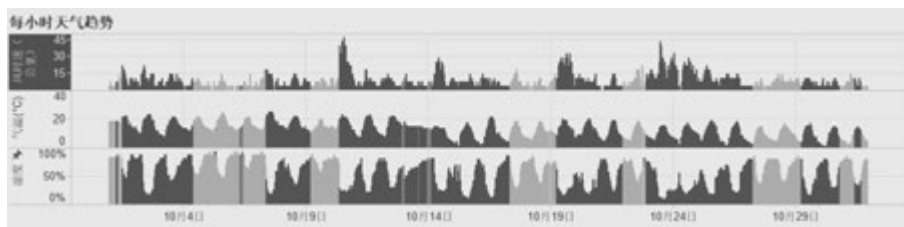


图 5-33 利用 Tableau 绘制的“每小时天气趋势”图

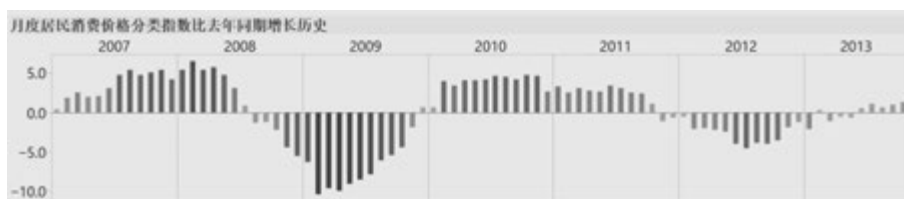


图 5-34 利用 Tableau 绘制的条形图



图 5-35 利用 Tableau 绘制的折线图

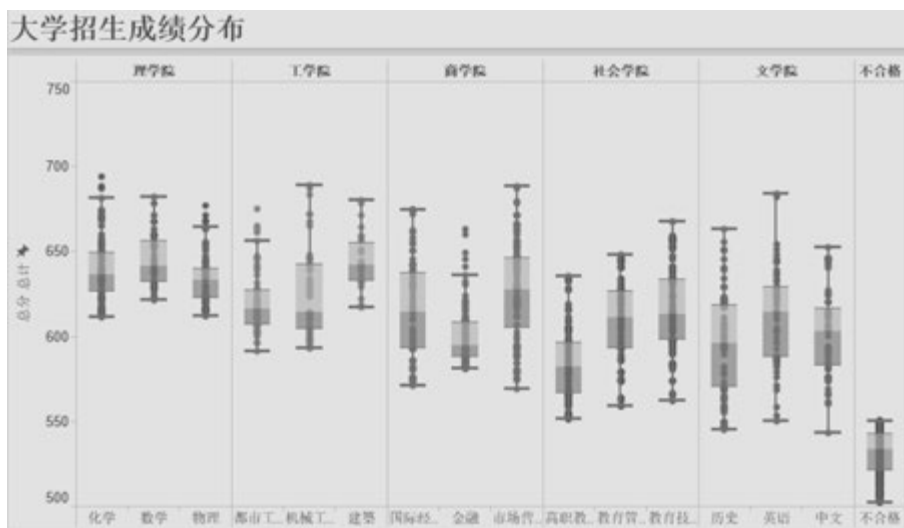


图 5-36 利用 Tableau 绘制的分布类图

Tableau 可以为多个领域或相关部门提供解决方案,其功能具有多样性。Tableau 的产品主要有 Tableau Desktop、Tableau Server、Tableau Online 和 Tableau Reader 等。

5.5.2 大数据魔镜

大数据魔镜是中国的一个大数据可视化分析平台,该平台积累了大量来自内部和外部的数据,用户可以自由地对这些数据进行整合、分析、预测和可视化。大数据魔镜首页如图 5-37 所示。



图 5-37 大数据魔镜首页

进入大数据魔镜官网即可浏览相关信息,也可进入官网注册试用该平台,对海量数据进行可视化分析。大数据魔镜的优点如下。

(1) 该平台拥有丰富的数据公式和算法,让用户可以真正探索和分析数据,如通过一个直观的拖曳界面就可创造交互式的图表和数据挖掘模型。

(2) 用户通过简单的拖曳操作就能一步生成分析模型,如精准营销、客户分析、用户画像等,有力支持管理者进行商业决策,提高核心竞争力。

(3) 用户可以邀请团队成员到自己的项目,进行合作探索和分析,通过分享让整个公司变得智能,并进一步发挥数据的威力。

5.5.3 ECharts

ECharts 是商业级数据图表(Enterprise Charts)的缩写,是百度公司开发的一个开源的数据可视化工具,是一个使用 JavaScript 实现的开源可视化库,可以流畅地运行在计算机和移动设备上,并能兼容当前绝大部分浏览器。进入 ECharts 的官方网站,单击“下载”按钮即可下载最新版本,如图 5-38 所示。

下载到本地的 ECharts 文件是个名为“echarts.min”的 Script 文件,在编写网页文档时将该文件放入 HTML 页面中即可制作各种 ECharts 开源图表。

ECharts 底层依赖矢量图形库 ZRender。在功能上,ECharts 可以提供直观、交互丰富且可高度个性化定制的数据可视化图表;在使用上,ECharts 为开发者提供了非常炫酷的图形界面,提供了包含柱状图、折线图、饼图、散点图、雷达图、树图、漏斗图、气泡图以及象形柱图等在内的一系列可视化图表。ECharts 示例图如图 5-39 所示。

ECharts 侧重于统计数据图表化层面,即使用传统的统计性图表来表示数据,用户可以通



图 5-38 ECharts 下载页面



图 5-39 ECharts 示例图

过其看到历史数据的统计和解读。ECharts 提供了常规的统计图表,可以用于地理数据可视化的地图、热力图、线图,关系数据可视化的关系图、旭日图,多维数据可视化的平行坐标,还可以用于 BI 的漏斗图、仪表盘,且能够支持图与图之间的混搭。除了已经内置的包含丰富功能的图表外,ECharts 还可以自定义图形,用户只需要传入一个 `renderItem` 函数,就可以直接从数据映射到任何想要的图形。ECharts 提供了图例、视觉映射、数据区域缩放、数据刷等交互组件,可以进行多维度数据筛选、视图缩放、细节展示等交互操作。ECharts 由数据驱动,数据的改变驱动图表展现改变,因此,动态数据的实现变得比较容易,只需要获取并填入数据,ECharts 就会找到两组数据之间的差异,然后通过合适的动画去表现数据的变化。

利用 Echarts 制作图表的步骤如下。

- (1) 新建 HTML 页面。
- (2) 在 HTML 页面头部中导入 JavaScript 文件。

(3) 在 HTML 页面正文中用 JavaScript 代码实现图表显示。

【例 5-1】 制作 ECharts 图表。

通过标签方式直接引入构建好的 ECharts 文件。

```
<!DOCTYPE html>
<html>
<head>
  <meta charset="utf-8">
  <!-- 引入 ECharts 文件 -->
  <script src="echarts.min.js"></script>
</head>
</html>
```

在绘图前需要为 ECharts 准备一个具备高宽的 DOM 容器。

```
<body>
  <!-- 为 ECharts 准备一个具备大小(宽高)的 DOM -->
  <div id="main" style="width: 600px;height:400px;"></div>
</body>
```

然后就可以通过 echarts.init()方法初始化一个 echarts 实例并通过 setOption()方法生成一个简单的柱状图,下面是完整代码。

```
<!DOCTYPE html>
<html>
<head>
  <meta charset="utf-8">
  <title>ECharts</title>
  <!-- 引入 echarts.js -->
  <script src="echarts.min.js"></script>
</head>
<body>
  <div id="main" style="width: 600px;height:400px;"></div>
  <script type="text/javascript">
    //基于准备好的 DOM,初始化 echarts 实例
    var myChart = echarts.init(document.getElementById('main'));
    //指定图表的配置项和数据
    var option = {
      title: {
        text: 'ECharts 入门示例'
      },
      tooltip: {},
      legend: {
        data:['销量']
      },
      xAxis: {
        data: ["笔记本","手绘板","打印机","网卡","键盘","鼠标"]
      },
      yAxis: {},
      series: [{
        name: '销量',
        type: 'bar',
        data: [5, 20, 36, 10, 10, 20]
      }]
    };
    //使用刚指定的配置项和数据显示图表
    myChart.setOption(option);
  </script>
```

```
</body>  
</html>
```

测试后,显示 ECharts 柱状图效果,如图 5-40 所示。

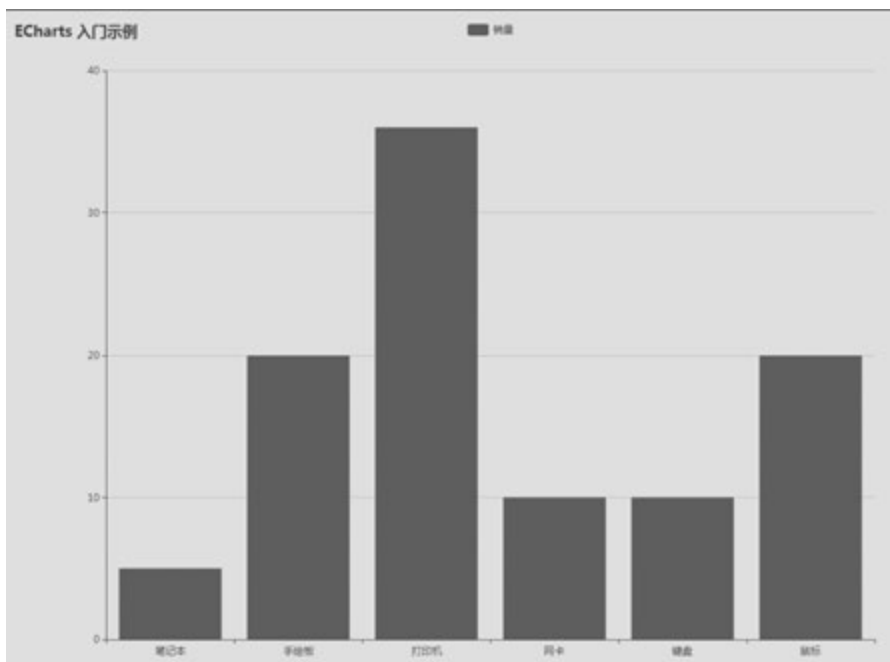


图 5-40 ECharts 柱状图

5.5.4 Pyecharts

Pyecharts 是一个用于生成 ECharts 图表的 Python 库,是一款将 Python 与 ECharts 相结合的强大的数据可视化工具。Pyecharts 作图灵活美观,相比其他可视化软件工具更容易上手,可以绘制柱状图、折线图、饼状图、散点图、热力图、漏斗图和极坐标图等多种图形。使用 Pyecharts 可以在 Python 中安装相应库,安装命令为 `pip install pyecharts`。

Pyecharts 可以在官方网站下载,下载完成后即可进行安装,安装完成后的界面如图 5-41 所示。

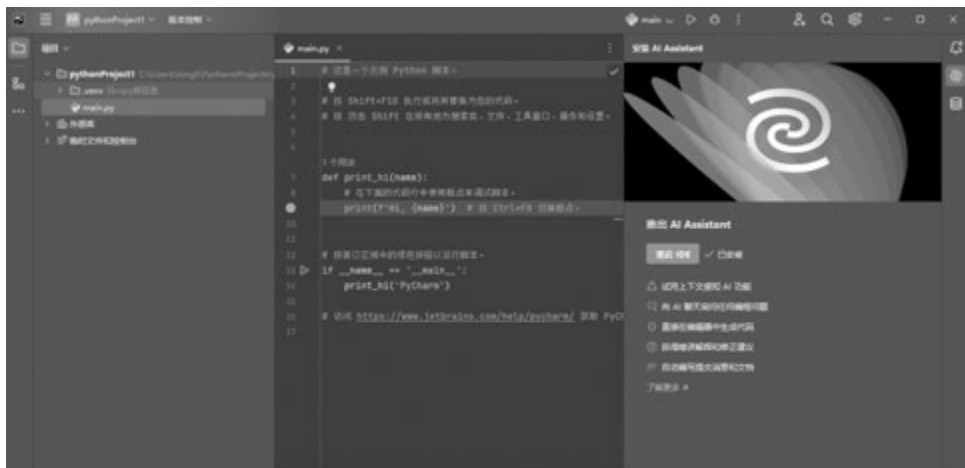


图 5-41 Pyecharts 界面

Pyecharts 包含绘制基本图表、地图、三维图、多子图等功能,具有简洁的 API 设计,包含 30 多种常见的图表函数。Pyecharts 可以为用户提供直观生动、可交互、可高度个性化定制的数据可视化图表。

5.5.5 D3.js

D3 是一种数据可视化工具,可以帮助用户将数据转换成视觉元素,以便更容易地理解和处理。D3 的全称是 Data-Driven Documents,顾名思义,它是一个被数据驱动的文档。D3 是一个基于 Web 标准(HTML、SVG 和 CSS)的 JavaScript 库。由于 JavaScript 文件的扩展名通常为 .js,所以 D3 也常叫作 D3.js,主要用于将数据绑定到文档的元素上,并通过这些元素进行数据可视化。D3.js 的核心是 SVG,允许开发者创建高质量的矢量图形。D3.js 支持多种交互功能,如缩放、拖动、单击事件等,可以增强用户体验和数据展示效果。D3 可以绘制的图形种类比较丰富,主要包括树状图、柱状图、可视化网络图、散点图、饼图、折线图、条形图等。

D3.js 可以用于创建动态更新的图表,如实时股票价格图表,可以根据最新的数据动态更新图表内容。利用 D3.js 的地理数据处理能力,可以创建交互式的地图,用户可以通过缩放、拖动等方式探索数据。D3.js 支持复杂的可视化效果,如力导向图、热力图等,适用于需要高度定制化可视化的场景。

5.5.6 Python

Python 是一门跨平台、开源、免费通用型面向对象的解释型高级动态编程语言。利用 Python 可以制作医疗数据、金融数据、教育数据等多种数据的可视化图表。Python 数据可视化的基本工具和方法主要有 Matplotlib、Seaborn 和 Pandas。

Matplotlib 是一个基于 Python 的绘图库,可以设计和输出二维以及三维数据,提供常规的笛卡儿坐标、极坐标、球坐标等。利用 Matplotlib 可以绘制出二维和三维图表,其绘制的图片质量清晰度高,能够高效地展示数据结果及相关数据间的关系。Pyplot 是 Matplotlib 的子包,提供了一个类似 MATLAB 的绘图框架,可以快速绘图,并设置图表中的各个细节。Seaborn 是基于 Matplotlib 的统计图形库,具有多种特性,如内置主题、调色板,可以可视化单变量数据、双变量数据、线性回归数据和数据矩阵等,能够创建富含信息量和美观的统计图形。Pandas 是基于 NumPy 的数据分析模块,提供大量标准数据模型和高效操作大型数据集所需要的工具,可以结合 Matplotlib 展现其绘图能力,实现数据可视化。在 Matplotlib 中,调用 figure() 函数能够创建画布并可以在画布上面绘图。创建多个 figure() 对象可以同时展示多个图。子图可以通过 add_subplot() 或 subplot() 方法来创建,add_subplot() 先创建一个画布,然后在该画布基础之上建立子图填充图表,subplot() 可以直接利用 Matplotlib 下的 Pyplot 创建子图。

【例 5-2】 使用 Matplotlib 中的 figure() 函数创建画布。

函数名称: figure()

函数功能: 创建画布

调用签名: figure(num, figsize, dpi, facecolor, edgecolor, frameon)

参数说明如下。

num: 主要是规定当前图形的编号或名称,数据为整数或字符串,默认为 None。

figsize: 主要是规定宽度和高度,数据为浮点数二元组,默认为 [6.4, 4.8]。

dpi: 主要是规定图形的分辨率,即每英寸的像素数,数据为浮点数,默认为 72。



视频讲解

facecolor: 主要是规定图片的背景颜色,数据为颜色值,默认为 white。

edgecolor: 主要是规定图片的边界颜色,数据为颜色值,默认为 white。

frameon: 主要是规定是否显示边框,数据为布尔值,默认为 True。

示例代码如下

```
import matplotlib.pyplot as plt
import numpy as np
# 创建一个新的绘图画布,并设置画布的大小和分辨率
fig = plt.figure(figsize=(9, 7), dpi=100)
# 在画布上绘图
plt.plot([4, 5, 6], [7, 8, 9])
# 显示画布
plt.show()
```

运行后,显示可视化图效果,如图 5-42 所示。

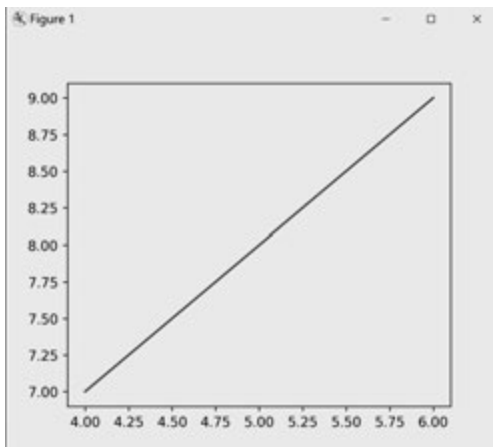


图 5-42 创建画布示例

【例 5-3】 使用 Matplotlib 中的 add_subplot() 函数创建子图。

函数名称: add_subplot()

函数功能: 创建子图

调用签名: fig.add_subplot(nrows,ncols,index)

参数说明如下。

nrows: 主要是规定当前画布划分几行,数据为整数。

ncols: 主要是规定当前画布划分几列,数据为整数。

index: 主要是规定选定具体的某个子区域,数据为整数,默认为 1。

示例代码如下。

```
import matplotlib.pyplot as plt
# 创建一个新的 figure
fig = plt.figure()
# 添加第一个子图
ax1 = fig.add_subplot(2, 2, 1)
ax1.plot([3, 4, 5], [6, 7, 8])
# 添加第二个子图
ax2 = fig.add_subplot(2, 2, 2)
ax2.bar([2, 3, 4], [5, 6, 7])
# 添加第三个子图
ax3 = fig.add_subplot(2, 2, 3)
```

表示 2 行 2 列的子图布局,当前的子图是第 1 个
在子图 1 上绘制一个简单的折线图

在子图 2 上绘制一个简单的条形图



视频讲解

```
ax3.scatter([2, 3, 4], [7, 8, 9])           # 在子图 3 上绘制一个简单的散点图
# 添加第四个子图
ax4 = fig.add_subplot(2, 2, 4)
ax4.pie([3, 4, 5], labels=['A', 'B', 'C']) # 在子图 4 上绘制一个简单的饼图
# 显示图形
plt.show()
```

运行后,显示可视化图效果,如图 5-43 所示。

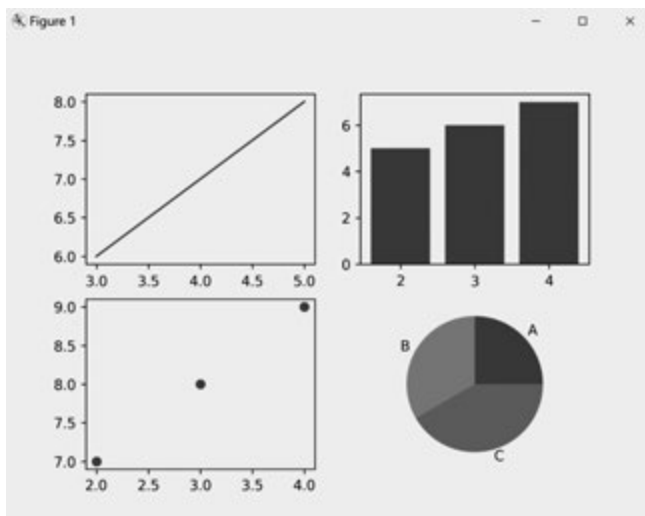


图 5-43 使用 add_subplot() 函数创建子图

【例 5-4】 使用 Matplotlib 中的 subplot() 函数创建子图。

函数名称: subplot()

函数功能: 创建子图

调用签名: plt.subplot(nrows,ncols,index)

参数说明如下。

nrows: 主要是规定当前画布划分为几行,数据为整数。

ncols: 主要是规定将当前的画布划分为几列,数据为整数。

index: 主要是规定选定的某个子区域,数据为整数,默认为 1。

示例代码如下。

```
import matplotlib.pyplot as plt
# 创建一个新的图像,大小为 2×2
fig, axes = plt.subplots(2, 1)
# 在上面的子图中绘制一条线
axes[0].plot([2,12, 29, 37], [13, 26, 20, 31])
# 在下面的子图中绘制一条线
axes[1].plot([3, 5, 17, 31], [6,13, 11, 28])
# 显示图像
plt.show()
```

运行后,显示可视化图效果,如图 5-44 所示。

Matplotlib 库中,使用 savefig() 函数保存图像,利用 plt.show() 或 fig.show() 函数显示图像。使用 Matplotlib 库中的 text() 函数可以添加文本注释。

在数据可视化中,雷达图也称为蜘蛛图、网络图或极坐标图,是一种将多个变量的数据点连接起来构成多边形的可视化图表。雷达图可以直观地展现多维数据集,查看哪些变量具有



视频讲解

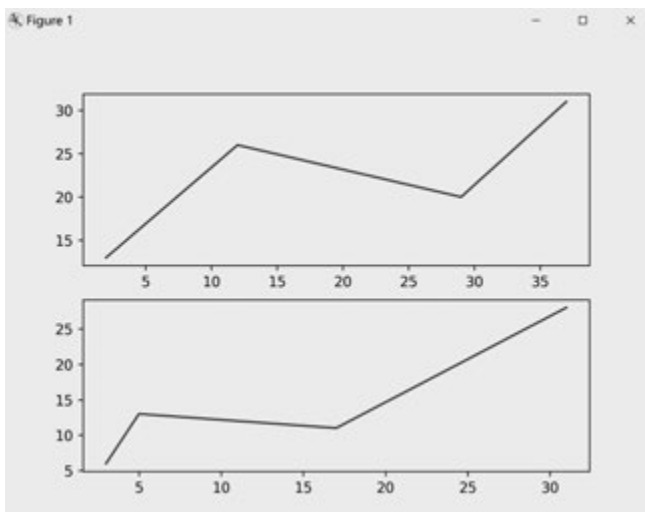


图 5-44 使用 subplot() 函数创建子图

相似的值,变量之间是否有异常值,也可以用于查看哪些变量在数据集内得分较高或较低。

【例 5-5】 根据学生的课程成绩清单绘制雷达图,代码如下。

```
import numpy as np
import matplotlib.pyplot as plt
# 某学生的课程与成绩
courses = ['高等数学', '大学英语', '计算机组成原理', '计算机网络技术', '操作系统',
            '机器学习', '人工智能', '大数据技术']
scores = [90, 89, 93, 85, 71, 77, 83, 98]
dataLength = len(scores)                                # 数据长度
# angles 数组把圆周等分为 dataLength 份
angles = np.linspace(0,                                # 数组第一个数据
                    2 * np.pi,                          # 数组最后一个数据
                    dataLength,                          # 数组中数据数量
                    endpoint = False)                   # 不包含终点
scores.append(scores[0])
angles = np.append(angles, angles[0])                   # 闭合
# 绘制雷达图
plt.polar(angles,                                     # 设置角度
         scores,                                       # 设置各角度上的数据
         'rv—',                                       # 设置颜色、线型和端点符号
         linewidth = 2)                                # 设置线宽
# 设置角度网格标签
plt.thetagrids(angles[:8] * 180/np.pi, courses, fontproperties = 'simhei')
# 填充雷达图内部
plt.fill(angles, scores, facecolor = 'b', alpha = 0.6)
plt.show()
```

运行后,显示可视化图效果,如图 5-45 所示。

【例 5-6】 利用 NumPy 生成数组以及对应的余弦数据,然后使用 Scatter 绘制散点图,代码如下。

```
import matplotlib.pyplot as plt
x = pl.arange(0, 2.0 * pl.pi, 0.1)                    # x 轴数据
y = pl.cos(x)                                           # y 轴数据
pl.scatter(x,                                           # x 轴坐标
```



视频讲解



视频讲解

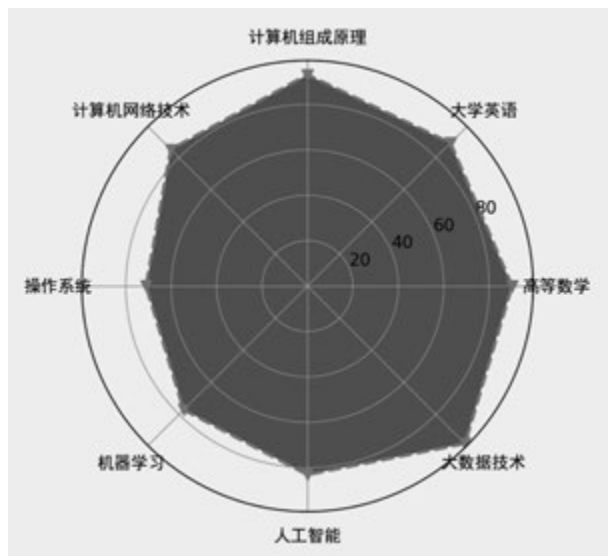


图 5-45 课程成绩分布雷达图

```

Y,                                     # y 轴坐标
s = 10,                               # 散点大小
linewidths = 5,                       # 线宽
marker = '+' )                       # 散点符号
pl.show()

```

运行后,显示可视化图效果,如图 5-46 所示。

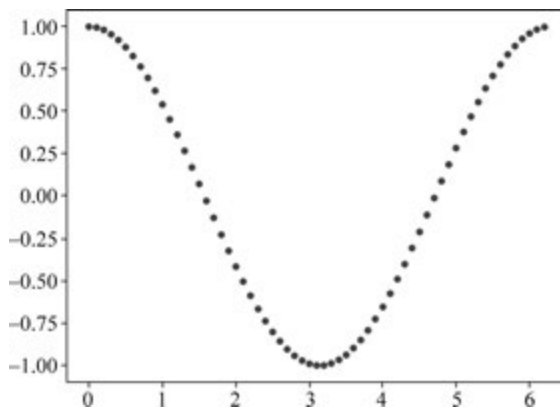


图 5-46 余弦散点图

【例 5-7】 利用 Matplotlib 绘制某公司销售额的柱状图,代码如下。

```

import pandas as pd
import matplotlib.pyplot as plt
data = pd.DataFrame({'月份': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
                    '手绘板': [12, 22, 21, 35, 41, 16, 46, 33, 41, 37, 52, 60],
                    '鼠标': [47, 30, 48, 73, 82, 80, 43, 25, 30, 49, 79, 60],
                    '键盘': [25, 40, 46, 50, 57, 76, 70, 33, 70, 61, 49, 45],
                    '扫描仪': [36, 45, 48, 56, 89, 92, 81, 84, 91, 85, 89, 86],
                    '打印机': [101, 95, 84, 99, 103, 89, 124, 125, 85, 80, 81, 77]})

# 绘制柱状图,指定月份数据作为 x 轴
data.plot(x='月份', kind='bar')
# 设置 x、y 轴标签和字体

```



视频讲解

```
plt.xlabel('月份', fontproperties = 'simhei')
plt.ylabel('销售额(万元)', fontproperties = 'simhei')
# 设置图例字体
plt.legend(prop = 'stkaiti')
plt.show()
```

运行后,显示可视化图效果,如图 5-47 所示。

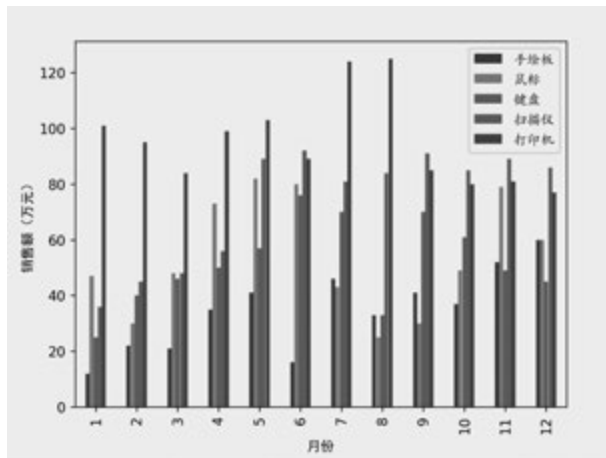


图 5-47 年度销售额柱状图

【例 5-8】 利用 Matplotlib 绘制折线图,展现某个城市一个月的天气情况,代码如下。

```
import matplotlib.pyplot as plt
plt.figure(figsize = (10, 10), dpi = 100)
# 绘制折线图
plt.plot([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28,
29, 30],
[8, 7, 8, 9, 11, 7, 5, 8, 9, 3, 5, 8, 12, 16, 8, 9, 5, 8, 12, 15, 14, 16, 18, 13, 17, 15, 12, 16, 9, 8])
# 添加标签
plt.xlabel("Date")
plt.ylabel("Temperature(℃)")
plt.show()
```

运行后,显示可视化图效果,如图 5-48 所示。

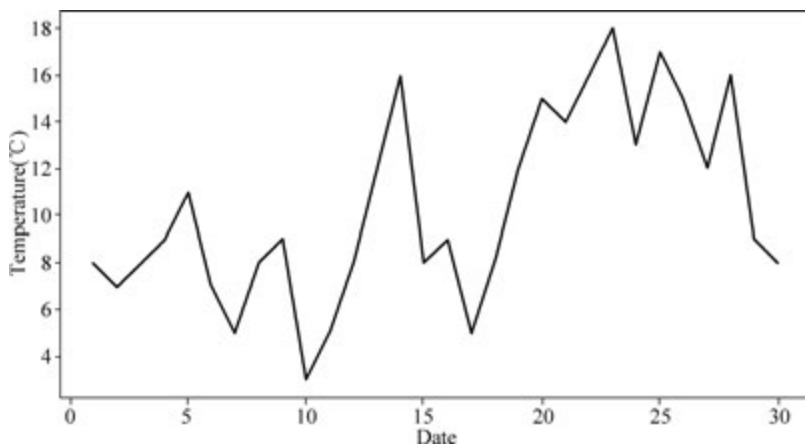


图 5-48 天气情况折线图



视频讲解

5.5.7 RAWGraphs

RAWGraphs 是一个在线的数据可视化工具,能够让用户轻松地将数据转换为各种可视化图表,可在线预览和修改,同时支持在线导入多种不同形式的数据格式,包括 TSV、CSV、DSV、JSON 等。RAWGraphs 支持用户自定义颜色、标签和图表风格,让用户能够创作出美观的可视化图表。数据图表的设计结果可以输出为多种图片格式,便于随时随地导入进行数据图表更新与维护。进入 RAWGraphs 官网后即可打开示例图效果,其可视化示例图如图 5-49 所示。

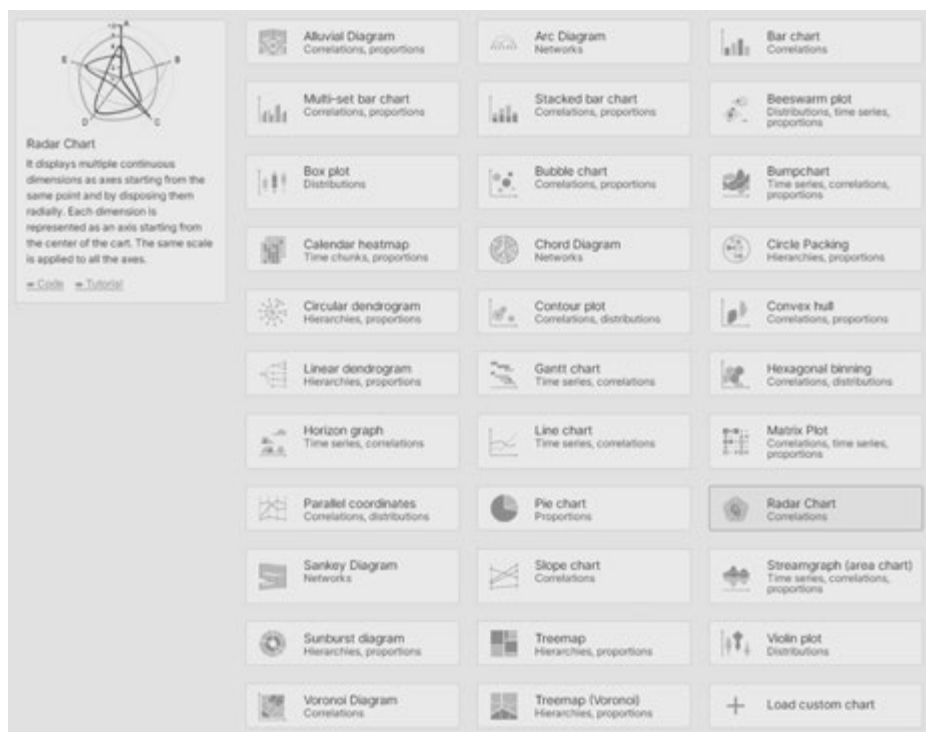


图 5-49 RAWGraphs 可视化示例图

打开官方网站后单击右上角的 Learning,即可进入数据可视化软件操作说明页面,如图 5-50 所示。

RAWGraphs 数据可视化的步骤:加载数据、选择图表、关联数据项与图表中可视化变量、定制图表配置项、输出图表文件。

(1) 选中 Use it now!选项,单击后可以进入第一步加载数据。RAWGraphs 提供了多种加载方式:复制粘贴数据、上传数据、数据样本、利用 URL 地址、上传保存的 RAWGraphs 工程文件。RAWGraphs 加载数据页面如图 5-51 所示。

(2) 选择样本数据中的 EU Index of consumer prices 数据并进行复制,将其样本数据粘贴至空白处,选择示例图中的 parallel coordinates。

(3) 在 DIMENSIONS 显示数据列中,CHART VARIABLES 是图表变量框,# 表示接收数字类型数据、Aa 表示接收字符类型数据、钟表图标表示接收日期类型数据、* 表示必与数据项关联的图表变量。用拖动方式(将 DIMENSIONS 下方的 Category、Date、HICP(Harmonised Index of Consumer P...)拖曳至右侧的对应位置)建立数据项与图表变量的对应关系,如图 5-52 所示。

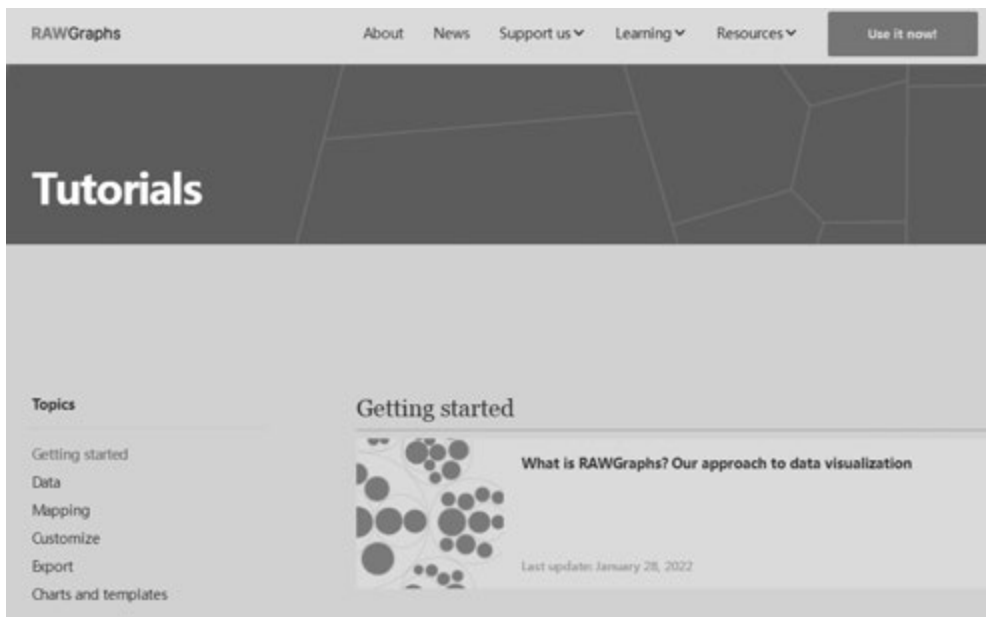


图 5-50 RAWGraphs 软件操作说明详细页面



图 5-51 RAWGraphs 加载数据页面

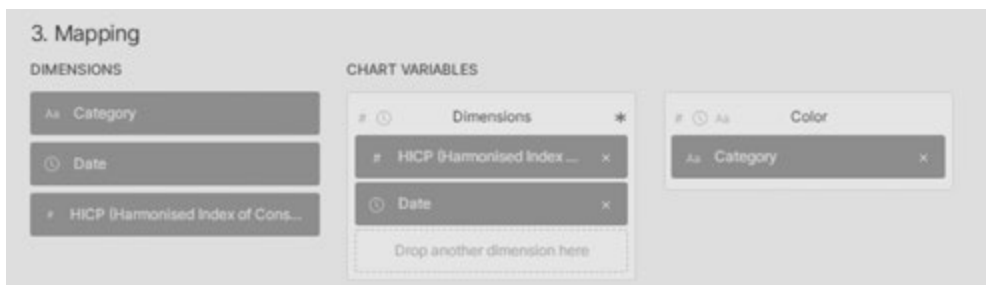


图 5-52 数据项与图表变量

(4) 对可视化图表大小、空白间隙宽度、图例、数据项系列、数据轴等配置项进行设置,完成后便可得到可视化图效果,如图 5-53 所示。

(5) 数据可视化效果图生成后,可将图表输出为 .svg、.png、.jpg 等图片格式,也可输出为 .rawgraphs 文件格式,便于数据图表的更新与修订。RAWGraphs 可视化图输出如图 5-54 所示。

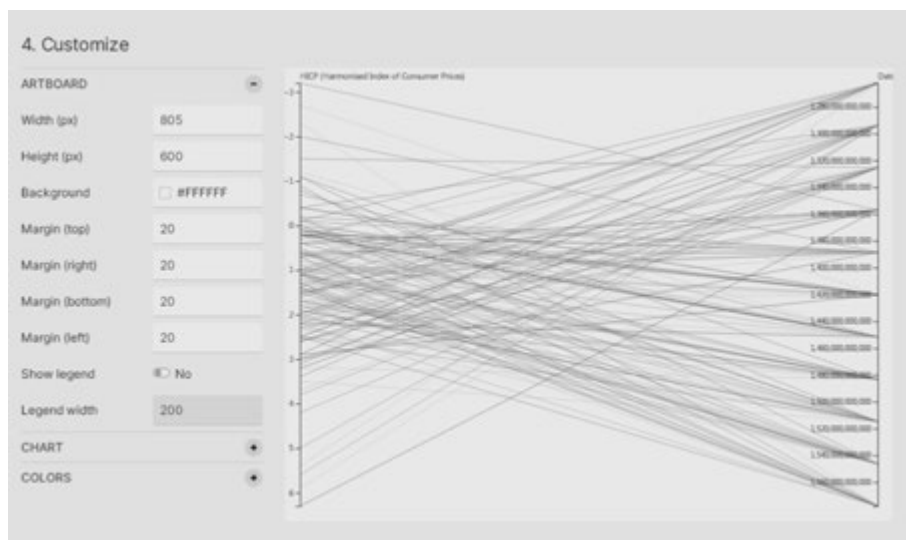


图 5-53 RAWGraphs 可视化效果图

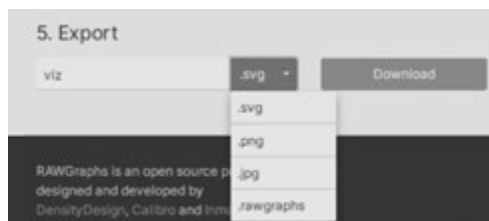


图 5-54 RAWGraphs 可视化图输出

5.5.8 Veusz

Veusz 是一款功能强大的科学可视化软件工具。Veusz 为科学研究提供一种简单而有效的方式来展示复杂的数据集。Veusz 提供了图形用户界面 (GUI)、Python 模块、命令行、脚本、DBUS 和 SAMP 接口,以访问其绘图功能。Veusz 可以运行在 Windows、macOS 和 Linux 系统上,通过组合各种绘图组件来构建图形,用户界面简洁且功能强大。它的主要特点是交互式界面,使得用户可以在实时预览中调整参数,直观地看到结果变化。Veusz 软件界面如图 5-55 所示。

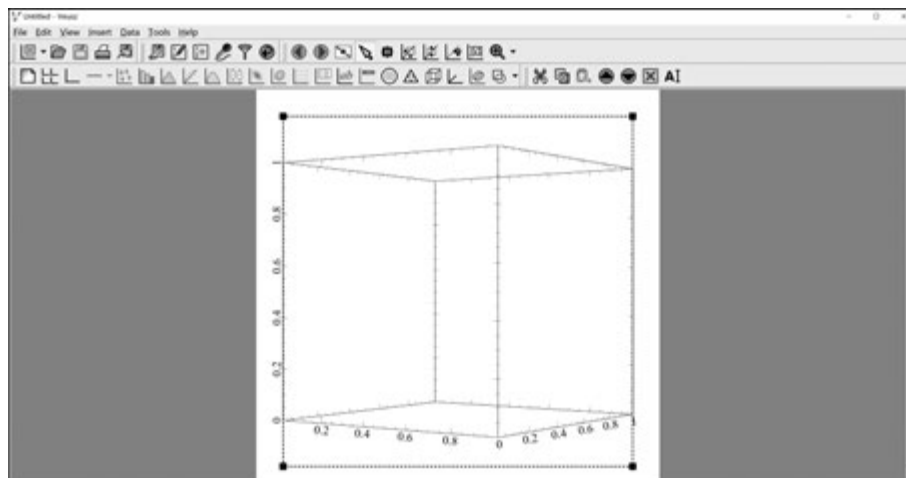


图 5-55 Veusz 软件界面

Veusz 支持直接导入多种数据格式,并可生成高质量的 PDF、SVG 和 PNG 图片。通过 Veusz 呈现的直观图表,可以快速理解大数据的分布和趋势。

5.6 大数据可视化应用

大数据可视化是正确理解数据信息的最好方法,大数据可视化让数据变得更加可信。当前,大数据可视化应用主要体现在教育、医学、电商、交通、气象等领域。

5.6.1 可视化在教育领域的应用

教育信息化的不断发展,产生了海量的教育数据,如何对教育大数据进行分析与可视化呈现,是教育领域中的重大问题。教育大数据对于如何更好地促进学生的个人发展,如何更好地发掘教育教学规律,如何更好地提升教育管理服务和教育治理水平,如何更好地制定教育决策起到了关键作用。

可视化在教育领域的应用主要体现在学生用户行为分析可视化、课程建设与学生学习分析可视化等多个方面。其中,学生用户行为分析可视化主要是指校园网学生用户行为分析可视化,通过对校园网络进行测量和分析,挖掘和发现网络中呈现出来的各种行为规律,同时识别一些网络异常行为,最后将学生用户行为分析进行可视化呈现。课程建设与学生学习分析可视化主要利用一些学习平台来实现,如利用学校的在线教育综合平台、爱课程网、MOOC(大规模开放在线课程平台)、SPOC(小规模定制在线课程平台)和智慧树、超星尔雅、腾讯课堂、网易云课堂等,利用在线教育平台中的登录学习资源数据和学习轨迹数据以及其他评价数据可以分析学习者学习的状况。对学习者的学习结果进行评价和反馈,更好地改进学习者的学习效率。

教育大数据可视化的数据类型主要包括时序数据、分类数据、文本数据、空间数据、多维数据、网络数据等。时序数据也被称为时间序列数据,是指某个指标按照时间顺序记录的数据列,例如,学习者每天在学习平台上的访问次数以及在线时长等。如图 5-56 所示是以散点图显示学生对在线课程的访问次数。

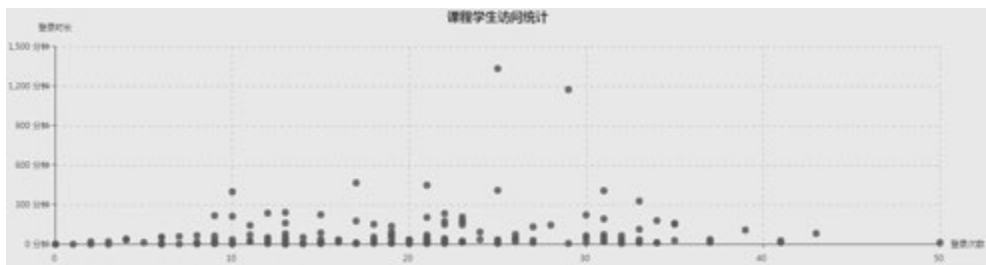


图 5-56 以散点图显示学生的访问次数

分类数据是指反映事物类别的数据,如学习者的考勤数据、互动文本数据、考试题目数据、在线测试数据、作业成绩分布数据等。如图 5-57 所示是以可视化形式显示学生作业成绩分布。

文本数据是教育大数据可视化数据类型的一种。文本数据可视化是指将文本数据以图形图像的形式呈现出来,增强数据的可读性和理解性。文本数据可视化具有直观性、形象性和易于理解等诸多优点。文本数据可视化能够将复杂的数据关系以简单的图形图像表现出来,便

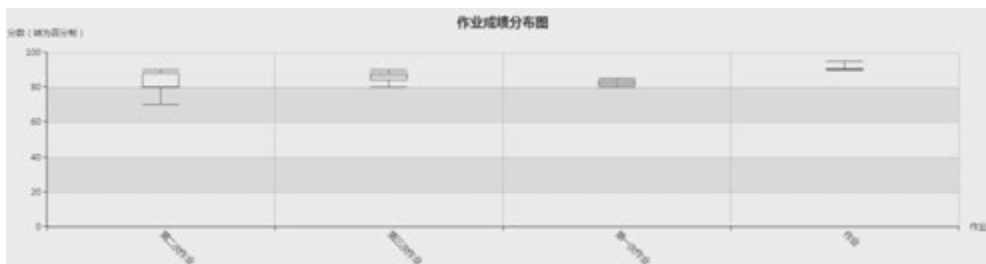


图 5-57 学生作业成绩分布

于分析和发现数据中的规律和趋势。教育领域文本数据的可视化可以分为基于知识图谱的可视化技术、基于词云的可视化技术、基于主题的可视化技术、基于网络的可视化技术等。知识图谱是一种基于语义网络的可视化知识表示方法,通过图形化的方式将复杂的知识结构进行清晰、直观的表达。基于词云的可视化技术是指将文本数据以词云的形式进行呈现;基于主题的可视化技术是指使用主题模型对文本数据进行聚类分析,识别其中的主题和子主题,然后将主题和子主题以可视化的形式呈现给用户;基于网络的可视化技术是指将文本数据中的单词、短语或句子映射为网络节点,将它们之间的关联或语义关系映射为网络的边,然后使用网络图可视化技术来展示文本数据的结构和关系。例如,智慧树在线教育中教育学课程的知识图谱环图可视化效果,如图 5-58 所示。



图 5-58 课程知识图谱环图可视化

智慧树在线教育中音乐与健康课程的知识图谱网图可视化效果,如图 5-59 所示。

空间数据是指用来表示空间实体的位置、形状、大小及其分布特征等方面信息的数据,可以用来描述来自现实世界的目标,具有定位、定性、时间和空间关系等特性,例如,学习者在不同地点学习的数据。多维数据主要是指具有多个维度的数据。多维数据可视化可以采用雷达图、堆叠面积图、箱线图、平行坐标系等多种方法来呈现。网络数据主要指利用网络处理和产生的数据。在教育领域,文本数据可视化可以更好地帮助学生理解复杂的教学内容。

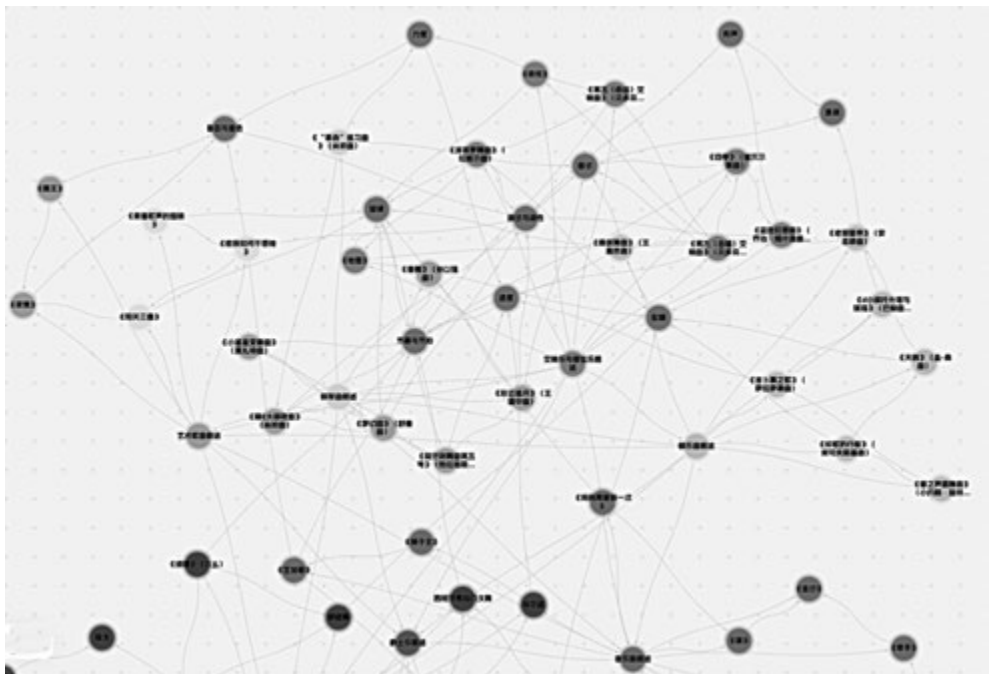


图 5-59 课程知识图谱网图可视化

5.6.2 可视化在医学领域的应用

可视化在医学领域的应用主要表现为医学图像数据可视化。医学图像数据可视化主要是利用现代计算机技术,将收集到的二维医学图像数据重构成物体的三维图像的技术。这种技术通常用于构造人体病变组织或器官的三维图像,通过这种手段获得的三维医学图像对临床应用具有较大的价值。

5.6.3 可视化在电商领域的应用

可视化在电商领域的应用主要体现在利用数据可视化挖掘出数据背后隐藏的价值信息,为业务的发展方向提供决策依据。在“互联网+电子商务”时代,通过可视化技术对海量的用户数据进行分析,可以快速确定用户的消费需求。

5.6.4 可视化在交通领域的应用

交通领域涉及的数据量庞大且复杂,包括车辆信息、道路状况、天气条件、交通事件等多个方面。借助数据可视化技术,可以将这些复杂数据以直观、易懂的形式展现出来,帮助决策者和公众更好地理解 and 掌握交通状况。通过将运输数据可视化,可以快速识别出运输路线、时间和成本等方面的瓶颈,进而针对性地优化运输策略,提高运输效率。

5.6.5 可视化在气象领域的应用

气象大数据可视化是将复杂的气象数据如温度、湿度、风速、气压等,以图形化的方式呈现,使得气象数据更为直观和便于理解。例如,通过温度曲线、降水图、风向图等形式,人们可以更好地做出生活和工作安排。此外,数据可视化在应对极端天气事件时也非常重要,能够帮助人们及时采取预防和救援措施,减少灾害损失。气象大数据可视化平台的交互功能方面,用

户可以通过触摸屏、鼠标等方式与画面进行互动,实现数据的交互式体验。未来,随着技术的不断进步和创新应用,气象大数据可视化将会发挥更加重要的作用,为人们的生活和经济发展带来更多的便利和价值。

习题

本书提供在线测试习题,读者扫描下方二维码可以获取本章习题。

