# 大数据可视化技术探讨

武丽芬[1] 严学勇[2]
（1. 晋中学院，山西 晋中 030619；
2. 中国联通晋中分公司，山西 晋中 030600）

摘要：数据可视化作为大数据研究领域一个异常活跃的分支越来越受到学者们的重视。本文从大数据可视化的发展历程、可视化常用技术和前沿技术、可视化流程、可视化工具及可视化技术面临的挑战五个方面做了详细阐释。

关键词： 大数据；可视化技术；可视化流程；可视化工具

当前，在大数据研究领域，数据可视化是一个异常活跃的分支。一方面，数据可视化以数据挖掘、数据采集、数据分析为基础；另一方面，它还是一种新的表达数据的方式，是对现实世界的抽象表达。数字永远是枯燥的、抽象的，而图形、图像却富有生动性和表现力。数据可视化是关于数据视觉表现形式的科学技术研究，它为大数据分析提供了一种更加直观的挖掘、分析与展示手段，从而让大数据更有意义，更贴近大多数人，因此大数据可视化是艺术与技术的结合。数据可视化将各种数据用图形化的方式展示给人们，是人们理解数据、诠释数据的重要手段和途径，已被广泛应用于诸多领域。

数据可视化所涵盖的技术方法广泛，它是以计算机图形学及图像处理技术为基础，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。它涉及计算机视觉、图像处理、计算机辅助设计、计算机图形学等多个领域，并逐渐成为一种研究数据表示、数据综合处理、决策分析等问题的综合技术。

## 一、数据可视化的发展历程

“可视化”的概念最初由相关学者在 1986 年的国家科学基金会举办的图形和图像处理研讨会上提出，其范围随着信息技术的发展而不断拓展，并根据数据处理对象和处理目的的不同而与数据挖掘、决策理论、认知科学和信息论等领域相融合，目前主流观点认为可视化技术主要包含四大类，分别是科学计算可视化、数据可视化、信息可视化和知识可视化。

进入 19 世纪以后，随着科技迅速发展，工业革命从英国扩散到欧洲大陆和北美。社会对数据的积累和应用的需求日益剧增，现代的数据可视化慢慢开始成熟，统计图形和主题图的主要表达方式。

进入 21 世纪以来，计算机技术获得了长足地进展，随着数据规模不断成指数量级的增长，数据的内容和类型也比以前要丰富得多，这些都极大地改变了人们分析和研究世界的方式，也给人们提供了新的可视化素材，推动了数据可视化领域的发展。数据可视化依附计算机科学与技术拥有了新的生命力，并进入了一个新的黄金时代。

现在，大数据可视化已经注定成为可视化历史中新的里程碑，VR、AR、MR、全息投影等这些当下最热门的数据可视化技术已经被应用到游戏、房地产、教育等各行各业。因此，人们应该深刻地认识到数据可视化的重要性，更加注重交叉学科的发展，并利用商业、科学等领域的需求来进一步推动大数据可视化的健康发展。

## 二、大数据可视化技术

大数据可视化技术是指将大规模数据集合以可视化形式呈现。通过图表、图像、地图等视觉元素来直观地理解和分析数据的工具和手段。通过数据可视化，可以更加深入地了解数据背后的模式、规律和趋势，进而作出更有价值的决策和应用。

大数据可视化的本质特征是数据规模大、数据类型多、数据更新快，在这个基础上，大数据可视化技术可以被概括为在合理的时间和空间范围内，对大规模、多类型和快速更新的数据进行交互式的处理、分析和展示的技术。

### （一）常用的数据可视化技术

数据可视化技术在应用过程中，常以目标驱动为主，在目标驱动下，数据可视化可抽象为对比、分布、组成、关系。

对比。比较不同元素之间或不同时刻之间的值。可以采用柱状图、多变量柱状图。根据不同元素包含的变量，分为单元素多变量和单元素单变量。如果是单元素多变量则用多变量柱状图；如果是单元素单变量，则采用柱状图。如果比较的是不同时刻之间的值，如长期时序数据，且有周期性，则采用周期面积图；长期时序数据如无周期性则采用折线图。如果是短期时序数据，则根据类别多少确定使用折线图还是柱状图。

分布。分布是数据可视化 最为常用的场景之一，用来查看数据分布特征，常用于数据异常发现、数值 过滤和数据基本统计性特征分析。根据不同变量的分布情况采用不同的方法。如果是单个变量分布，根据数据点多少分别采用折线图和柱状图；如果是两个变量的分布则采用散点图；如果是多个变量的分布则采用平行坐标方法。

组成。查看数据静态或动态组成。动态组 成可以根据数据特点分为短期和长期数据。对于 短期数据，根据关注相对比例或绝对组成可以分别 采用堆叠比例柱状图和堆叠柱状图；对于长期数据，同样根据关注相对比例或绝对组成可以分别采 用堆叠比例面积图和堆叠面积图；对于静态组成， 简单的总体组成，可以采用饼状图；若关注相对整 体的增减可以采用瀑布图；若组成元素包含子元 素，可以采用堆叠比例柱状图；若关注组成及其绝 对差，可以采用树图。

关系。查看变量之间的相关性，这常常用 于结合统计学相关性分析方法，通过视觉结合使用 者专业知识与场景需求判断多个因素之间的影响 关系。根据变量的多少进行划分，若是两个变量可 以采用散点图；若是 3 个变量，可以采用气泡图，用 散点半径表征第 3 个变量；超过 3 个变量可以采用 平行坐标方法。

### （二）可视化前沿技术

可视化前沿技术是指在可视化领域中最新、最先进的技术和方法。这些技术通常采用了最新的计算机图形学、数据处理和交互技术，旨在提供更高质量、更真实、更交互性强的可视化体验。可视化前沿技术的发展使得可视化应用能够更好地满足用户的需求，提供更丰富、更直观的视觉体验，拓展了可视化在各个领域的应用范围。如分子可视化，能够为新型蛋白质设计以及药物研

制等多方面的研究工作提供重要线索，复杂分子及其特性的可视化一直是分子生物学和相关领域研究的基础之一。如分子可视化，分子可视化中高质量的三维渲染效果能有效地帮助科研工作者更好地观察蛋白质分子的结构以及理解蛋白质分子结构与功能之间的联系。如光线追踪，光线追踪是一种计算机图形学的渲染技术，通过模拟光线在场景中的传播和交互，生成逼真的图像。它可以产生真实的光照效果、阴影和反射，使得可视化结果更加真实和生动。如体积渲染，体积渲染是一种用于可视化三维立体数据（如医学图像、地质数据等）的技术。它通过对具体数据进行采样和渲染，生成具有透明度和颜色的体绘制效果，使得用户可以直观地观察和分析体数据。

**三、大数据可视化流程**

可视化流程分为数据空间和可视化空间，如图 1 所示。数据空间对原始数据进行一系列的数据转换操作，如清洗、选择、聚集等，提炼原始数据的主要特性并将原始数据转换成易于可视化的形式。可视化空间将数据映射为合适的可视化形式，以帮助用户直观理解数据蕴含的知识和规律。

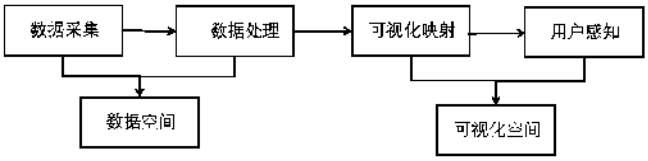


**图 1 可视化流程图**

数据采集是数据可视化的基础，数据可以通过仪器采样、调查记录等方式进行采集。数据采集又称为“数据获取”或“数据收集”。采集得来的原始数据不可避免含有噪声和误差，同时数据的模式和特征往往被隐藏。因此为保证数据的完整性、有效性、准确性、一致性和可用性，需要进行数据处理。 数据处理是可视化前期工作，其目的是提高数据质量。可视化映射是可视化流程的核心环节，它用于把不同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、色调、饱和度、亮度等。用户感知是从数据可视化结果中提取有用的信息、知识和灵感。用户可以借助数据可视化结果感受数据的不同，从中提取信息、知识和灵感，并从中发现数据背后隐藏的现象和规律。

**四、大数据可视化工具**

（一）初级工具 Excel。

Excel 是 Microsoft 为使用 Windows 和 Apple Macintosh 操作系统的计算机用户编写的一款电子表格软件。直观的界面、出色的计算功能和图表工具，再加上成功的市场营销，使 Excel 成为最流行的个人计算机数据处理软件。初学者可以使用 Excel 制作各种精美的图表，包括了条形图、饼图、气泡图、折线图、仪表图以及面积图等。

（二）开源可视化工具 ECharts。

ECharts 是百度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 PC 和移动设备上，并能够兼容当前绝大部分浏览器。在功能上，ECharts 可以提供直观、交互丰富，可高度个性化定制的数据可视化图表。ECharts 官网上提供了大量的可视化图表，如折线图、柱状图、饼图、散点图、雷达图、关系图、热力图、树图等。

（三）商业智能工具 Tableau。

Tableau 是用于可视分析数据的商业智能工具，也是目前全球最易于上手的报表分析工具。商业智能的概念最早在 1996 年提出，当时将商业智能定义为一类由数据仓库（或数据集市）、查询报表、数据分析、数据挖掘、数据备份和恢复等部分组成的、以帮助企业决策为目的技术及其应用。

（四）可视化编程语言 Python。

使用 Python 中的扩展库，可以较为轻松地实现数据可视化。一般来讲，Python 可视化的实现以 numpy 库和 matplotlib 库为基础，除此以外，还有 pandas 库、seaborn 库、Bokeh 及 pyqtgraph 库等。

matplotlib 可视化库是 Python 下著名的绘图库，是 Python 可视化库的基础库。matplotlib 库的功能十分强大，matplotlib 通过 pyplot 模块提供了一套和 Matlab 类似的绘图 API，将众多绘图对象所构成的复杂结构隐藏在这套 API 内部，只需要调用 pyplot 模块所提供的函数就可以实现快速绘图以及设置图表的各种细节。

numpy 库是 Python 做数据处理的底层库，是高性能科学计算和数据分析的基础，在数据可视化中需要用到 numpy 中的数组存储以及矩阵运算等功能，掌握 numpy 库对数据可视化十分重要。

pandas 库是 Python 下的数据分析库，主要功能是进行大量的数据处理，同时可高效的完成绘图工作。与 matplotlib 库相比，pandas 库绘图方式更加简洁。

seaborn 库是基于 matplotlib 的 Python 可视化库。它提供了一个高级界面来绘制有吸引力的统计图形，使得数据可视化既方便又美观。

Bokeh 库是一款针对现代 Web 浏览器呈现功能的交互式可视化库，其通过 Python 以快速简单的方式为超大型数据集提供高性能交互的多功能图形。

pyqtgraph 库是一种建立在 PyQt4/PySide 和 numpy 库基础之上的纯 Python 图形 GUI 库，在数学、科学和工程领域都有着广泛应用。尽管该库完全用 python 编写，但内部由于使用了高速计算的 numpy 信号处理库以及 Qt 的 GraphicsView 框架，在大数据量的数字处理和快速显示方面有着巨大的优势。

**五、可视化技术面临的挑战**

数据可视化技术在广泛应用的同时，也面临诸多新的挑战。1. 大数据规模大、价值密度低，受限于屏幕空间，所能显示的数据量有限。为有效显示使用者所关注的数据和特征，需有效压缩数据。近期一些学者提出了针对特定可视化场景的数据压缩方法，但是依然缺少通用的面向可视化的数据压缩方法，也缺少实际应用的产品。2. 数据融合难。大数据数据类型多样，常分布于不同的数据库。如何融合不同来源、不同类型的数据，为使用者提供统一的可视化视角，支持可视化的关联探索与关系挖掘是一个重要问题。3. 图表可视化的效率问题凸显，表达能力有限。随着数据源、数据类型的不断增加，数据使用者对于数据的交互需求越来越多，已有的数据可视化产品无法满足使用者的可视化需求，对于系统的图表表达能力和表达效率提出了更高的要求。

**参考文献：**

[1] 黄源，蒋文豪，徐受蓉 . 大数据可视化技术与应用 [M]. 清华大学出版社，2020.6

[2] 吴琼 . 基于大数据可视化技术的金融精准扶贫审计研究——以 X 银行精准扶贫项目为例 [D]. 南京审计大学，2021.

[3] 沈恩亚 . 大数据可视化技术及应用 [J]. 科技导报，2020，38（3）：16.

**项目编号：晋中学院大数据技术应用创新团队，编号 jzxyjscxtd202105，2021 年 9 月。**

1. 一作者：武丽芬（1978.11-），女，汉族，山西省太原市，研究生学历，教授，研究方向：数据挖掘、大数据可视化。

2. 二作者：严学勇（1977.01-），男，汉族，山西省晋中市，研究方向：大数据技术、云计算。