参观可视化动物园

Jeffrey Heer, Michael Bostock and Vadim Ogievetsky, Stanford University

由于传感、网络和数据管理方面的进步, 我们的社会正在以惊人的速度生产数字信息。据一项估计, 仅在 2010年, 我们将产生 1200 艾字节——60万倍的国会图书馆的内容。在这海量的数据中蕴藏着大量宝贵的信息, 说明我们如何开展业务、政府和个人生活。为了使信息得到良好的利用, 我们必须找到方法来探讨、联系和交流数据。

可视化的目标是通过利用人类视觉系统的高度调谐能力来帮助我们理解数据, 从而看到模式、斑点趋势和识别异常值。设计良好的视觉表征可以用简单的知觉推论代替认知计算, 提高理解力、记忆力和决策能力。通过提高数据的可访问性和吸引力, 视觉表征也可以帮助更多不同的受众进行勘探和分析。挑战在于创造出适合于数据的有效和引人入胜的可视化效果。

创建可视化需要一些微妙的判断。必须确定要问的问题, 确定适当的数据, 并选择有效的可视化编码将数据值映射到图形功能, 如位置、大小、形状和颜色。面临的挑战是, 对于任何给定的数据集, 可视化编码的数量--因而可能的视觉设计的空间--非常大。为了指导这一进程, 计算机科学家、心理学家和统计人员研究了不同编码对数据类型 (如数字、类别和网络) 的理解程度。例如, 图形感知实验发现, 空间位置 (如散布图或条形图) 会导致对数值数据进行最精确的解码, 通常比视觉变量更可取, 如角度、一维的长度、二维的面积、三维体积和颜色饱和度。因此, 最常见的数据图形 (包括条形图、折线图和散点绘图) 应使用位置编码, 这一点不足为奇。然而, 我们对图形感知的理解仍然不完整, 必须适当地与交互设计和美学平衡。

这篇文章通过 "可视化动物园" 提供了一个简短的教程, 展示了各种数据集的视觉和交互技术。在许多情况下, 简单的数据图形不仅是足够的, 他们也可能是可取的。在这里, 我们专注于一些比较复杂的和不寻常的技术, 处理复杂性数据集。毕竟, 你不去动物园看娃娃和浣熊;你去欣赏雄伟的北极熊, 优雅的斑马, 和可怕的苏门答腊虎。类似, 我们包括一些更具异国情调的 (但实际上是有用的) 视觉数据表示形式, 从最常见的时间序列数据开始;继续统计数据和地图;然后使用层次结构和网络完成访问。一路上, 记住所有的可视化共享一个共同的 "DNA"--数据属性和视觉属性之间的一组映射, 例如位置、大小、形状和颜色--而且通过改变这些编码, 可以构建出自定义的形象化物种。此处显示的大多数可视化效果都伴随有交互式示例。live 示例是使用 Protovis (http://vis.stanford.edu/protovis/) 创建的, 这是一种用于基于 web 的数据可视化的开放源代码语言。要了解有关如何进行可视化 (或复制并粘贴到您自己的使用) 的详细信息, 请简单地在页面上 "查看源"。所有示例源代码都被释放到公共域中, 并且对重用或修改没有任何限制。但是, 请注意, 这些示例仅在支持 svg (可伸缩矢量图形) 的现代标准兼容浏览器上起作用。支持的浏览器包括火狐、safari、chrome 和歌剧的最新版本。不幸的是, ie 8 和早期版本不支持 svg, 因此不能用来查看交互式示例。

时间序列数据

时间序列数据集——变量集随时间的变化——是最常见的记录数据形式之一。时变现象是许多领域的核心, 如金融 (股票价格、汇率)、科学 (温度、污染水平、电力潜力) 和公共政策 (犯罪率)。人们通常需要同时比较大量的时间序列, 并且可以从多个可视化项中选择这样做。

索引图表

在某些形式的时间序列数据中, 原始值比相对变化更重要。考虑那些对股票的增长率比其具体价格更感兴趣的投资者。多个股票可能有显著不同的基准价格, 但可能有意义的比较, 当正常化。索引图表是一个交互式折线图, 它根据所选的索引点显示时间序列数据集合的百分比变化。例如, 图1A 中的图像显示了在2005年1月购买时所选股票价格的百分比变化: 人们可以看到那些投资亚马逊、苹果或谷歌的人在这段时间里所享受的坎坷的上涨。

堆积图

其他形式的时间序列数据可以更好地看到的总和。通过将区域图堆叠在一起, 我们到达了时间序列值的可视化求和-一个堆积图。这种类型的图 (有时称为流图) 描述了聚合模式, 通常支持深入到单个系列的子集中。图1B 中的图表显示了过去十年中美国失业工人的数量, 按行业细分。虽然这些图表在近几年已经被证明是受欢迎的, 但它们确实有一些明显的局限性。堆积图不支持负数, 对于不应求和的数据 (例如, 温度) 毫无意义。

此外, 堆叠可能使得很难准确地解释其他曲线上的趋势。交互式搜索和筛选通常用于补偿此问题。

小倍数图

代替堆叠, 可以在同一轴内绘制多个时间序列, 如在索引图表中。然而, 在同一空间中放置多个系列可能会产生重叠曲线, 从而降低可读性。另一种方法是使用小倍数: 在自己的图表中显示每个系列。在图1C 中, 我们再次看到失业工人的数量, 但在每个行业类别中被规范化。我们现在可以更准确地看到每个部门的总体趋势和季节性模式。当我们在考虑时间序列数据时, 请注意, 可以为任何类型的可视化构造小的倍数: 条形图、饼图、地图等。这通常会产生一个更有效的可视化, 而不是试图将所有数据强制转换为单个绘图。

地平线图

当你想一次比较更多的时间序列时会发生什么？地平线图是一种在保持分辨率的同时增加时序视图的数据密度的技术。考虑图1D 所示的四图。第一个是标准面积图, 正值值呈蓝色, 负值为红色。第二个图 "镜像" 负值与正值相同的区域, 使面积图的数据密度加倍。第三个图表--地平线图--通过将图形划分为带区并将其分层以创建嵌套形式, 再次使数据密度加倍。结果是保存数据解析的图表, 但只使用了四分之一的空间。虽然地平线图需要一些时间来学习, 但是当图表尺寸变得很小时, 它已经被发现比标准图形更有效。

统计分布图

其他的可视化设计是为了揭示一组数字是如何分布的, 从而帮助分析人员更好地理解数据的统计属性。分析师通常希望将其数据与统计模型相匹配, 以测试假说或预测未来的价值, 但错误的模型选择可能会导致错误的预测。因此, 可视化效果的一个重要用途是探索性数据分析: 了解数据的分布方式, 以通知数据转换和建模决策。常用的技术包括直方图, 它显示了分组到垃圾箱中的值的流行, 以及箱和晶须图, 它可以传递统计特征, 如平均值, 中值, 四分之一的边界, 或极端的异常点。此外, 还有一些其他技术用于评估分布和检查多个维度之间的交互。

茎叶图

为了评估一个数字的集合, 一个替代的直方图是茎叶的情节。它通常根据第一个有效数字将数字分类, 然后将每个 bin 中的值堆叠到第二个重要数字。这种简约表示使用数据本身来绘制频率分布, 替换传统直方图条形图中的 "信息空" 条, 并允许评估每个 bin 的总体分布和内容。在图2A 中, 茎叶图显示了在亚马逊机械特克完成人群任务的工人完成率的分布情况。注意多个集群: 一组集群在高完成级别 (99-100%);另一个极端是一群 Turkers 人, 他们只完成了少数几个任务 (约 10%)。

q-q 图

尽管直方图和词干图是评估频率分布的常用工具, 但 q q (位-位) 图是一种更强大的工具。q-q 图比较两个概率分布的图形他们的位 (http://en.wikipedia.org/wiki/Quantile) 互相反对。如果两者是相似的, 绘制的值将大致沿中央对角线。如果两个线性相关, 值将再次沿着一条线, 虽然有不同的斜坡和拦截。

图2B 显示了相同的机械特克参与数据与三统计分布。请注意, 数据在与统一和正常 (高斯) 分布相比, 如何形成三不同的组件: 这表明, 一个具有三个分量的统计模型可能更合适, 而且我们确实在最后的图中看到, 三正态分布的拟合混合物提供了更好的拟合。q q 情节虽然强大, 但其有效使用要求观众具备一定的统计知识, 这是一个明显的局限性。

SPLOM (散点图矩阵)

其他可视化技术试图表示多个变量之间的关系。多变量数据经常发生, 众所周知, 很难表示, 部分原因是在三维以上的精神想象数据的困难。克服这个问题的一种方法是使用散布图的小倍数, 显示变量间的一组两两关系, 从而创建 SPLOM (散点积矩阵)。SPLOM 允许对任意对变量之间的相关性进行目视检查。

在图2C 中, 散布图矩阵用于可视化汽车数据库的属性, 显示马力、重量、加速度和位移之间的关系。此外, 交互技术, 如刷牙和 linking-in, 其中一个图表上的点的选择突出了相同点的所有其他图形-可以用来探索模式内的数据。

平行坐标图

平行坐标 (| |-coord), 如图2D 所示, 采用不同的方法来可视化多变量数据。我们不是在两个维度中绘制每对变量, 而是反复地在平行轴上绘出数据, 然后用直线连接相应的点。每一个折线表示数据库中的一行, 而在维度之间的行交点通常表示反向关联。重新排序维度可以帮助模式查找, 也可以通过交互式查询沿一个或多个维度进行筛选。并行坐标的另一个优点是它们相对紧凑, 因此可以同时显示许多变量。

地图

虽然地图似乎是一种自然的方式来可视化地理数据, 它有着悠久而丰富的设计历史。许多地图是基于制图投影: 一个数学函数, 把地球的三维几何映射到一个二维的图像。其他地图故意歪曲或抽象地理特征, 以讲述一个更丰富的故事或突出显示特定的数据。

流量图

通过在地图上放置描边线, 流程图可以描述空间数量的移动和 (含蓄地) 时间。流线通常对大量的多变量信息进行编码: 路径点、方向、线条粗细和颜色都可以用于向查看者显示信息的维度。图3A 是对查尔斯. Minard 对拿破仑在莫斯科不幸行进的描述的现代诠释。许多最伟大的流动图也涉及微妙的扭曲使用, 因为地理学被修改以容纳或突出流动。

等值线地图

数据通常由国家等地理区域收集和汇总。通信此数据的标准方法是使用地理区域的颜色编码, 从而产生等值线的映射。图3B 使用一种颜色编码来传达美国各州肥胖的流行情况。虽然这是一个广泛使用的可视化技术, 它需要一些护理。一个常见的错误是对原始数据值 (如填充) 进行编码, 而不是使用规范化的值来生成密度映射。另一个问题是, 一个人对阴影值的感知也可能受到地理区域的底层区域的影响。

渐变符号图

等值线地图的另一种选择是毕业的符号图, 它将符号放置在基础地图上。这种方法可以避免混淆地理区域和数据值, 并允许可视化更多的维度 (例如, 符号大小、形状和颜色)。除了简单的形状 (如圆圈) 外, 渐变符号图还可以使用更复杂的字形 (如饼图)。在图3C 中, 总圆大小代表一个州的人口, 每个切片表示具有特定 bmi 等级的人的比例。

变形地图

变形地图扭曲了地理区域的形状, 以便该区域直接对数据变量进行编码。一个常见的例子是重新绘制世界上每个国家按比例将其调整为人口或国内生产总值。许多类型的变形地图已被创建;在图3D 中, 我们使用林变形地图, 它代表每个地理区域的大小圆, 放置, 以便像真正的地理配置。在本例中, 循环区域对每个状态的肥胖者总数进行编码, 并对肥胖的总人口百分比进行颜色编码。

层次

虽然有些数据只是简单的数字集合, 但大部分都可以组织成自然的层次结构。考虑: 空间实体, 如县、州和国家;企业和政府的指挥结构;软件包和系统树。即使对于没有明显层次结构的数据, 统计方法 (例如, k-均值聚类) 也可用于按经验组织数据。特殊的可视化技术存在于利用层次结构, 允许快速多尺度推断: 微观观察的单个元素和宏观观察大群体。

节点链接图

词树与层次结构交替使用, 因为橡木的分形分支可能会反映数据的嵌套。如果我们采取一个 二维蓝图的树, 我们有一个普遍的选择, 可视化的层次结构: 节点链接图。设计了许多不同的树形布局算法;Reingold-蒂尔福德算法在软件类的包层次结构中使用, 在图4A 中用最少的浪费空间产生一个整洁的结果。

另一种可视化方案是图 (或簇) 算法, 它将树的叶节点置于同一级别。因此, 在图4B 中的图中, 类 (橙色叶节点) 在圆的直径上, 包 (蓝色的内部节点) 在里面。使用极性而非笛卡尔坐标具有令人愉悦的美感, 同时更有效地使用空间。

我们将忽略缩进树, 使用无所不在的操作系统来表示文件目录, 以及其他应用程序 (参见图 4C)。虽然缩进树需要过多的垂直空间, 不利于多尺度推断, 但它确实允许对树进行高效的交互式探索, 以找到特定的节点。此外, 它允许快速扫描节点标签, 而多变量数据 (如文件大小) 可以显示在层次结构的旁边。

邻接图

邻接图是节点链接图的空间填充变体;与其在层次结构中绘制父项和子级之间的链接, 不如将节点绘制为实心区域 (弧形或条形), 而相对于相邻节点的位置则显示它们在层次结构中的地位。图4D 中的冰柱布局与第一个节点链接图类似, 因为根节点出现在顶部, 下面是子节点。但是, 由于节点现在是空间填充的, 因此我们可以对软件类和包的大小使用长度编码。这将显示一个在节点链接图中很难显示的附加维度。

森伯斯特布局, 如图4E 所示, 相当于冰柱布局, 但在极坐标。这两种方法都是使用分区布局实现的, 也可以生成节点链接图。同样, 以前的簇布局可用于在笛卡尔或极坐标中生成空间填充邻接图。

存储模块关系图

存储模块图也是空间填充, 使用包容而不是邻接来表示层次结构。由 ben Shneiderman 在1991年介绍, 树图递归地细分区域成长方形。与邻接图一样, 树中任何节点的大小很快就会被发现。图4F 所示的示例使用填充 (蓝色) 来强调机箱;有时使用替代的饱和编码。正方化树图使用近似正方形长方形, 比一个天真 "片和-骰子" 细分提供更好的可读性和大小估计。泰森多边形和竖锯树图等更高档的算法也存在, 但不太常见. 通过填充圆圈而不是细分矩形, 我们可以生成不同类型的存储模块

具有几乎有机外观的图。虽然它不像树图那样有效地使用空间, 但图4G 所示的圆形包装布局的 "浪费空间" 有效地揭示了层次结构。同时, 使用区域判断可以快速比较节点大小。

网络

除了组织, 我们可能希望通过可视化来探索的数据的一个方面是关系。例如, 给定一个社交网络, 谁是谁的朋友？谁是主角？什么派系存在？谁, 如果有人, 是在不同的群体之间的桥梁？抽象地说, 层次结构是一种专门化的网络形式: 每个节点都有一个指向其父级的链路, 而根节点没有链接。因此, 节点链接图也用于可视化网络, 但层次结构的丢失意味着需要不同的算法来定位节点。

数学家使用正式的术语图来描述一个网络。图形可视化的一个中心挑战是计算一个有效的布局。布局技术通常寻求在绘图中关闭紧密相关的节点 (在图形距离方面, 如节点之间的链接数或其他度量);重要的是, 不相关的节点也必须放在足够远的位置来区分关系。有些技术可能会寻求优化其他视觉特性, 例如, 通过最小化边缘交点的数量。

力导布局

对网络布局的一种常见而直观的方法是将图形建模为物理系统:

节点是相互排斥的带电粒子, 链接是将相关节点连接在一起的被挫伤的弹簧。然后对这些力进行物理模拟, 确定节点位置;避免计算所有成对力的逼近技术使大量节点的布局得以实现。另外, 交互性允许用户引导布局和微动节点消除链接的歧义。这种力定向布局是理解一般无向图结构的好起点。在图5A 中, 我们使用一个强制的布局来查看在维克多·雨果的经典小说《悲惨》的章节中的字符网络。节点颜色描述由社区检测算法计算的群集成员身份。

弧形图

图5B 所示的弧线图使用了节点的一维布局, 用圆弧来表示链接。虽然弧线图可能无法像二维布局那样有效地传达图的整体结构, 但通过对节点的良好排序, 可以很容易地识别出派系和桥梁。此外, 与缩进树布局一样, 多变量数据可以很容易地与节点一起显示。以一种显示底层集群结构的方式对节点进行排序的问题被正式称为系列化, 在可视化、统计、甚至考古学上有着不同的应用。

矩阵视图

数学家和计算机科学家经常用它的邻接矩阵来考虑一个图: 矩阵中第一行和列 j 中的每个值对应于从节点 i 到节点 j 的链接。给定这个表示, 一个明显的可视化是: 只显示矩阵!使用颜色或饱和度而不是文本, 可以更快地感知与链接相关的值。

序列化问题与图5C 所示的矩阵视图同样适用于弧形图, 因此行和列的顺序非常重要: 在这里, 我们使用由社区检测算法生成的分组来命令显示。虽然矩阵视图中的路径跟踪比节点链接图更难, 但矩阵有许多补偿优势。随着网络的大规模和高度的连接, 节点链接图经常被下放到巨大的毛的线交叉路口。然而, 在矩阵视图中, 线交叉是不可能的, 并且通过有效的排序可以快速地发现簇和网桥。允许矩阵的交互式分组和重新排序有助于更深入地探索网络结构。

结论

我们已经到达了我们的旅程结束, 希望读者已经找到的例子既有趣又实用。虽然我们参观了许多视觉编码和交互技术, 更多的可视化物种存在于野外, 而另一些则等待发现。生物信息学和文本可视化等新兴领域正在驱使研究人员和设计者不断地制定新的和创造性的表述, 或者找到更有力的方法来应用经典。在这两种情况下, 所有可视化的DNA都保持不变: 数据的原则性映射，变量的视觉特征, 如位置, 大小, 形状和颜色。当你离开动物园, 回到野外的时候, 试着解构穿越你的道路的各种可视化。也许你可以设计一个更有效的显示器？