《数据可视化分析相关》

百度 EFE 宿爽

主持人：好的感谢大家踊跃提问。还有问题的小伙伴不要着急，在咱们下半场圆桌讨论环节，可以来发问，我看现场群已经满了还没有加群可以勾搭一下，中场有茶歇环节群里互动。接下来有请第二位嘉宾，百度 EFE 宿爽老师来来数据可视化相关分析分享。

宿爽：我先自我介绍一下，我是视觉团队，这个题目叫数据视觉编码交互，这三个名词实际上是可视化里头主要三个基本的概念。之所以举这个，可能是想谈一谈，在做这 ECHARTS 以及可视化的理解。因为我们现在数据非常庞大的时代，数据如果躺在磁盘硬盘中没有价值。真正有价值经过人脑或者是机器对数据进行一些转移的理解，从中提取信息或者是向人传达出一些信息，这才能有价值。

如何把这些很难理解的数据，转换成人脑理解的东西，这个是所研究的点。讲可视化从这堆字母中找出到底有几个 V 字符，大家能数得出来吗？就是说标出来很明显，四个，另外一个例子给了八个数字，哪一个数值是属性是年轻、年老的，哪个高是低，是不是辅助能看出来，这画出来图来，现在有三个下降的线，一个上升的线，上升的线黄线是女性的低收入人群，就这人群随着年龄增长，这是外国的统计不见得符合咱们能做到这样。都是通过视觉来接受，我们脑子首先是进行视觉的感知，这是第一。第二更进一步的是要进行信息的提取分析归纳解码，这些形成概念、形成知识，这就是视觉的认知。因为我们接受的信息大部分的视觉，所以我们进行认知、进行分析这样活动也大部分在处理视觉的信息。所以实际上我们更擅长于在视觉上面进行处理，而不是对数字进行一些逻辑上的思考。这就是为什么可视化在数字分析和信息传播中会有一个很重要德威治。再一个例子很常见，有几个数据，也就是说某种线性比例的关系，如果在统计学中进行分析的话，那可能常常用去计算一些相关系数，或者做一些直线。可以直接说答案这四组数据计算相关系数他们得到答案是一样，如果一个直线也是完全的直线。但是统计学得出来这个直线是不是能说明这些数据真的四组数据都是一样的，可视化的数据可以看到，这是有很大的差异，这每一组是一样的数据，但是这数本身很不一样，比如说统计学意义上来说是比较理想，和数据本身的特征是相对符合的，而第二组数据实际上类似一个抛物线，但是统计得出来的直线和它没有能够反应出来这方面的特征。

第三个也是一个直线，但是它的直线和你统计量得到的直线是有偏差原因就是有一个比

较大的离群点，第四个根本就是一个竖线，也是一个离群点，导致了偏差，所以这个说明，当时做这例子先进行一种可视的方式画一个图，有一个重要性。

先说一下数据，我们就是在做数据可视化的时间，可能会面临的数据各种各样的，不可能对于哪一种数据的形态做分别处理，可能对数据进行一些分类和抽象，根据我们常见的类型，做一种对应的处理，可能有哪些类别，一些类型的时间，是连续的，它就能够采用很多已经有了代数运算，比如说你缺失哪个值我就能很明白的知道，那个值在数轴上什么位置。

但是也有可能离散的数据，离散而有序这样采用能力的偏弱一些，但是能够有所计算。还有完全离散，有可能需要能够对他进行处理的话，也需要知道它的整个全集，否则的话对它进行形象的描述，并不能进行处理。现实中有能够数据有很多纬度，可能相应的纬度，看看他们之间有什么联系和规律提取一些结论。还有数据之间的关系常常会用数来反应。把这种数据影射到各种途径元素活动。最常用编码大概可以分为两件事情，或者是两类，一个是标记就是图片元素，另外一个视觉通道。首先标记是什么呢？很简单点、线、面。这个是非常容易理解。大家也见过很多，你像知道这些东西可能会代表，可能别人要表达什么意义。但是大家可能不会去有意思的视觉通道颜色、亮度、饱和尺寸形象伦理方向、动画等等。这些东西实际上是一个补充一个辅助，能够在可视化的东西上传达出来更多的信息。

比如说颜色，颜色可以去编码类别数据，最常见的就是用图，不同的颜色一目了然就能知道不同的类别。当然实际上它也是一个面子也表达了一些信息。比如说 GDP 的值这个是 R 值上这些一目了哪里值高、哪里 值低。这样筛选的话也可以沿海高很多，内陆的地方非常低。我们处理的数据刚才说过有很多的纬度，这只是一个比较少纬度的例子。这个数值是一个天气情况的数据。第一列是日期，就取了一个月的日子 1 号到 31 号，第二列就是空气质量数字，第二列数值越高代表空气质量越差。空气是一些细节二氧化碳、PH 值、二氧化硫这些。然后最后是它的一个官方共用的评价，与轻度污染严重这些。我们常常会把第一列定到 X 轴，第二运到 Y 轴这样就能清晰的表示出来污染的程度。

但是如果需要把后面的这些信息也反应进去，对需要关注的人让他能够查大这些信息，那可能就需要借助颜色通道，多种视觉通道的辅助，比如说 Y 是 AQ 指数，S 轴是日期。颜色表示城市的类别。可能 PM2.5 最大的日期是这些日期，也可能二氧化硫最大的这些值，它这明暗表示二氧化硫，二氧化硫最大是这些日期。如果再考虑一些三维可能用高度表示一些额外的信息，高度成为一个视觉通过道，这个例子是地球的例子，表示的是人口数量，各个城市的人口数量，这个柱子越高那就是人口数量越大。那看出来东京人口这是十分高的，也可以做一些筛选。

但是很多视觉通道我们在选举它的时候，我们把这优放成了一个菱形，但是实际上能看出来并不明显，我们最开始注意颜色的不同，以及大小的不同，所以在可视编码的时候，可能需要把最重要的一些数据，就是最重要用户需要关心的属性，然后影射最有区分的通道中，而这种形状辅助和明暗度一样，辅助的一些筛选。可视编码这一块就说完了。

下面说一下交互，为什么单独提这个事情呢？就是之前我们能看到可视化的，比如说互联网上往往是一个图片比较缺乏的交互，但是事实上进行可视化的过程，如果只是一个静态，没有交互的话，就屏蔽了人和数据进行互动的过程。实际上很难发现一些数据或者是规律，或者更深层的得到数据的一些细节信息。所以在可视化的研究中，这是可视分析的模型图，这是比较偏学术，这模型图反应出一些问题，这可视分析的过程先有数据，把数据定到视觉，或者数据得到一些模型或者模式，最后在这互动循环中得到了一些知识，反馈修改，在这过程中进行了一些分析。实际上整个的过程都是人是重度参与，这参与的途径就是交互。所以说之前正在做分享的时候，题目是深度交互的数据可视化，就是说提出交互的重要性。在可视化的设计原则中，对交互有很多种可以归类为一些任务，比如说总览、缩放、过滤、按需细节，还有相关的历史操作。在五可视化中总览韦先还有缩放过滤还有按需察看细节，考虑一些已有的模式，是不是能够满足这需求。

比如说这是降水量的图。横轴是时间、纵轴是流量，如果交互的话，看整体的趋势，再看里面每一个细节需要关注的话，比如说去掉一些比较高的值，看一些基础的值变化趋势等等。

最后说一下动画，动画拿出来说的意思是，因为大家很多关注前端的，实际上在浏览器前端里面，相对来说动画不是一个很重要的过程，大部分的一些能看到的网站或者管理系统，就是很简单的动画。因为动画可以给人带来什么呢？这系统里带来视觉抓眼求的效果。但是它会带来不同的复杂性，可能是软件大得多，也可能代码比较复杂，代码进行一些优化的话，可能直接破坏代码的结构。所以说如果要采用动画，像 CS300，分为 CSS 里也要解决这些问题，使这动画更容易，但是你要采用更复杂的动画，可能收益就不高。但是对于可视化来说动画这一件事情，实际上我们认为还是非常重要，它能够表达在交互中，各种数据图形元素，在电话中的一些联系，它能帮助理解。如果你是一个突变的，你可能很容易使人对于这个视觉的迷失。所以我们也会花了很多的精力去在动画上面做一些优化和效果的实现。

然后有了前面三个数据，视觉编码还有交互，那我们可以就在这个基础上可以实现很多的效果。在这里面按照各种的数据类别，来区分一下，回到数据上面看看能够做一些什么。比如说点数据的可视化，就是可以标记在地理上面、位置上面，没有大小尺寸，但是可以反

应直观的信息。这个例子是一个微博签到图，每一个点相当于一个数值，点越多越密越亮，亮就表明这程度会高。这里是 11 万多个点。线数的可视化，一个是表明两个实体之间的关系，另外它可能表明距离、长度。线上面还可以增加一些属性。这里做的一个非线航线图的例子。现在展示的是美国航空公司的航线。也可以看看中国航空公司，国航主要是在中国。

那看一下场和向量，场在更深的地方用的比较多。比如说这个例子是说在洋流，洋流大家知道实际上看不见，你要表达这个洋流视觉你就得做一些视觉的处理。美国拉萨（音）在 2005 年做了一些视觉上的效果，取梵高的星空图。也做了这效果的尝试，实际上是从向量上表达出这洋流的走势，这是南极。

时间数据的可视化这个非常常见。就像刚才那个例子横轴就是时间，直接把时间投影到最基础的轴上这是非常常见。有时候也需要时间并不占用这么基础的通道。而是说放在更高的纬度。比如说之前做的这个例子，本身它是作为地图的数值，数值是经济指标可以看到视觉哪个省份高、哪个省份低，如果在高纬度增加时间，就可以在时间变化中看到颜色的加深，就可以看到 GDP 变化的趋势。

最后说一下关系的数据，关系数据里，比较常见的是树和图。首先是数是层级数据没有回落，最简单的直接把树画出来，树很大，这种方式很直观。但是如果有时候在树的某个结点上有更多的信息想表达，想要更清楚的表达，你可能需要找一些别的方式，所以国外比较流行的。

这也是一个树，每一块都是树的结点，每一块还有子结点，实际上从树的根结点往下看看这个树，这里最重要的通道是面积，把这树里每一个结点里的数值，影射到面积就能很明白的树值的。这个预算是奥巴马预算的例子，这是医疗，如果要进行一下 2011 年的比较，但是看不出来区别是什么？我们可以把这 2012 年比 2011 年预算的增长率也影射到视觉通道上，比如说影射明暗上面，有一些比较亮就表示增长率很高，其他地方只增长了百分之三点多。这一块是个人安全，反正相当于这个。另外还可以进行下算，比如说这个是一个硬盘里面，空间占用的例子，这一块一个目录占用了大概 1G 左右，其他的几百 K，如果想看一下详情里面具体什么，就可以进行下算，看到里面具体的一些内容。

下面再说一下网络数据，常常就是用图来表示，网络数据并不是互联网，而是说点和信息，还有点之间的关系信息，然后怎么样可视化来表示，然后点的大小表达一定的含义，就是表达更多的视觉。比如说数据新闻里的话会用很多，因为实际上既有美观的效果，而且也很清楚的反应出来一些社会的数据情况，但是用图的时候，经常会提到不足这概念。我们可以采用更多的布局方式。

最后说一下力导向图，这个图是 Webkit 内核里面的依赖，就是性能优化或者代码处理的时候，我们无从下手，力导向进行了一些聚类，形成这些类的相关性比较大。我们可能能够发现一些问题，从这里这一片就能尝试，力导向图本身它是一个弹簧的模型。它把没两个点中引出一个实例，它的引力和事例的大小和距离平方成正比和反比，两个点离的近相吸、离的远相斥。它是直接把这种算法利用到这图象上面来达到分析的效果。

今天我讲的就这么多，然后谢谢大家。

主持人：谢谢宿总精彩的分享。哪位小伙伴有问题，我们先看看举手的人。

提问：你好我想问一下，咱们在浏览器端展示？我有一个场景，比如说在一定时间内经过一些轨迹这数据量非常大，我就想量级的情况下要达到一个很好的展示效果？

宿爽：看这量级有多大，如果超出了这范围也不得不进行一些处理。

提问：数据本身的处理？它原始的数据可能就会比较数据量展示的情景？

宿爽：原始的数据如果说需要展示的话，经常就是不能够直接拿原始数据直接放在最终的展示终端，还是进行一些预先的处理。比如说原始的话，可能会不光是数据量级的问题，而且不不见得能够很好的表达出来展示的效果。

主持人：OK 还有哪位？

提问：你好，刚才听到很多数据可视里面的展现。但是在我们场景当中，也有很多是在展现完以后，就是用户想在展现的过程中操作，就是一些入口和接口？

宿爽：现在的版本比较偏重于暴露在外面操作，现在这算是非常非常后发展比较重要的点。

主持人：好，我们再次感谢宿爽精彩的分享。接下来我们是茶歇时间。我们稍后休息一下，10 分钟之后要在微信群里面发红包抽奖了。