可视化的基础知识

可视化的定义

可视化是一种将**抽象的数据转化为可见的几何特征**，帮助使用者观察数据的计算结果和数据分布特征。可视化提供了一种**呈现数据中不可见特征的方法**。

可视化研究的是人或者机器如何感知、处理、交互可视化的信息。

可视化的三个基本原则：

1、基于真实数据进行可视化

2、可视化需要生成一张图像

3、生成的可视化图像是方便使用者观察的

可视化和数字图像处理、计算机视觉、计算机图形学的分辨

数字图像处理：图像----->图像

计算机视觉: 图像----->对应的3D模型，识别图像，转化成模型

计算机图形学：3D模型------>图像，将一个模型渲染到图像中。

可视化：数据----->图像

科学可视化

科学可视化是使用计算机图形学和可视化的相关知识，构建虚拟的图像，用于一些复杂且庞大的概念或者试验结果的呈现

可视化的重点

可视化分析是一个解析推理，可视交互的过程，重点是要将**数据信息呈现给人(使用者)，将人放在可视化分析的回路中**。

数据特征的分类

- 定量的数据：数据可以量化，且能够比较次序

  例子是：长度、宽度
- 定序的数据：数据不可以量化，但是能够比较次序

  例子是：等级、先后顺序
- 名词性数据：数据不可以量化也不能比较，只包含文字性意义

  例子是：姓名和性别

箱型图

辛普森悖论-----美国航空准点率问题

当来自多个组的数据合并成一个组时，且比较关系或者关联关系发生反转。

核心因素：每个组的数据样品数量等分配可能并不均衡，因此才产生了数据合并之后的比较关系与合并前分组比较关系相反的情况发生。

箱型图

箱型图是一种五点图：

- 最小值：围栏内的最小值
- Q1第一个四分位点：前1/4的数据值
- Median中位数：中位数
- Q3 第三个四分位点：前3/4的数据值
- 最大值：围栏内的最大值

要计算的关键数据：

IQR=Q3-Q1

围栏：围栏值得时数据的上限和下限，超过围栏的数据点属于离群点。

上围栏：max fence=Q3+1.5*IQR

下围栏：min fence=Q1-1.5*IQR

注意：围栏只是为了找到离群点，在最后图像上并不显示出来。

箱型图的生成过程：

1、先根据数据分别确定出第一个四分位点Q1，中位数、第三个四分位点Q3，并且根据这三个数据画出基本的箱子

2、计算出IQR=Q3-Q1。使用IQR计算出围栏的值：

max fence=Q3+1.5*IQR min fence=Q1-1.5*IQR

根据上下围栏，找出离群点。使用圆圈标出离群点

3、根据围栏，确定出最大值和最小值，画线，并且从箱子向两侧画线。

最后在图上显示出来的是围栏内的最大值和最小值。

平行坐标系

平行坐标系的定义

平行坐标系是一种数据的展现方式，横轴是数据的多个属性维度，纵轴向是数据在该属性维度上的值。在平行坐标系中，一条有多个属性维度的数据，表示成一条折线，从第一个属性坐标开始到最后一个属性坐标结束。

平行坐标系的应用场景

平行坐标系能够比较快捷的反映出一个数据多个属性之间的对应关系，在多维问题中有比较好的效果。

如果一个数据有多个维度不方便显示，可以使用平行坐标系来进行可视化处理

平行坐标系与基本分布分坐标系的对应关系

对于散点图中的一个点，可以映射到平行坐标系上的一条横跨两个维度的线。

对应规则

1、散点图中的反比关系-----x维度上越高，y维度上越低

2、散点图中的随机分布-----x维度和y维度的变化关系并不十分明显。

3、散点图中的正比关系------x维度和y维度的变化方向一致，x越高，y也随之变高。

平行坐标系存在的问题

当使用平行坐标系进行可视化时，如果数据的条数过多就会出现平行坐标系的折线过于密集导致难以分辨的问题。(数据条数太多，形成一片黑色，难以看出对应关系)

数据样本数量过多的改进方法

使用聚类算法和增强视觉效果的可视化技术，例如对数据进行聚类再分别使用不同的颜色进行可视化。

可视化设计

可视化的设计需要遵从爱德华的设计原则：

图表的完整性：

1、如果是条形图，需要注意零刻线的使用，如果不带有零刻线地可视化一个条形图可能会产生认知上的错误。

2、如果是线形图则不需要注意零刻线，因为线形图比较的是**线的角度**。线形图需要遵从x坐标轴和y坐标轴的45度法则。使x和y轴的数据变化呈现45度的变化规律能够减少可视化的视觉错误。

同时在一个线形图上，多段数据的变化率应该是相同的，一条直线，相同的x轴变化量，应该对应相同的y轴变化量。

图表的真实性

可视化图表中显示出来的特征应该和原本数据的数值特征成比例。

重点：**Lie Factor说谎因素的计算**

Lie Factor=图中的尺寸变化百分比/实际数据变化百分比

变化百分比的计算方法：从V1变化到V2

变化百分比=｜V1-V2｜/ V1

注意：**说谎因素的计算时两个百分比相除**

说谎因素的值越大，说明图上对数据的体现相比于数据本身的特征越明显，图表的设计者可能夸张了图表。

最大化使用图像

为了使图像最大化表现出数据的特征，需要最大化使用图像

使用Data-Ink-Ratio数据的油墨占比

数据的油墨占比=反应数据的油墨/图像的总油墨

一般来说，二维图像的数据油墨占比要高于三维图像的数据油墨占比。也就是版面上用于反应数据信息的部分更大。

避免图表中出现垃圾Chart JUNK

Chart Junk的定义：哪些会使观察者从数据信息、数据特征中分散注意的无意义的可视化元素称为图表垃圾。

需要正确的判断Chart junk：

并不是图表中所有的附加元素都是Chart Junk，部分附加元素在对理解没有损害的基础上，能够加深使用者对图表的印象，这种附加元素是有意义的。

我们认为要避免的是有害的图表垃圾。

关于附加信息的理论

可视化图表中的每一部分不应该只有数据、图表本身的意义。可视化图表中携带的元素应该包括携带数据信息的部分和展现可视化设计效果的部分。

总结：爱德华的图表设计原则

1、保证图表的完整性：保证例如零刻线、45度原则等使用，使图表能够反映出数据的完整情况。

2、保证图表的真实性：使用说谎因素来衡量可视化图表的真实性，要保证图表对数据的反映是真实的

3、最大化利用图表：使用数据的油墨占比来衡量一个可视化图表中数据的显示占比，尽可能地反映出更多使用者可以获取的信息。

4、避免有害的图表垃圾：引入图表垃圾这个概念，要注意的是需要分辨有害的图表垃圾和可行的附加元素，图表需要同时展现数据特征和可视化设计的效果。

配色问题

图表中颜色起到的作用

- 标注数据的类型
- 度量数据的大小
- 利用颜色展示某种数据模型或者真实情况(用于科学可视化)
- 装饰图表，体现图表的设计特征

颜色的感知过程

1、光

2、眼睛的视锥细胞感光

3、形成对应的颜色信号

4、颜色的感知---大脑知道有了颜色

5、颜色的呈现---颜色呈现在大脑中

6、颜色的认知---大脑对颜色进行认知，说出颜色种类

Rainbow Color Map

彩虹颜色图表是基于可见光的颜色光谱顺序形成的。

存在的问题

1、人为的将颜色分类，色调不是自然排列的

2、不同的亮度值表示的数据是相同的

3、低频数据低亮度的颜色(蓝色)会遮盖掉高频数据高亮度的颜色(红色)，使高频数据不明显。

颜色使用的法则

分类数据-----可以使用定性的颜色来表示(例如：rainbow color map表示)

有序、定量的数据------使用顺序方案来表示(渐进的颜色)

有序、定量的数据且中间有一个具有意义的分界点----使用离散的对比方案(对比反差的两种颜色，中间颜色是白色代表的分界点)

定量数据的分类方法

定量数据进行分类时，其核心目的就是最小化组内差异，最大化组间差异。

- 等间隔分类(算术方法的分类)
- 分位数分类----比较好的一种分类方法
- 标准差分类
- 聚类算法分类

可视化中的设计原则与感知

可视化的设计原则(两个方面)

表达方面：在一个可视化方案中，如果可视化表达了数据集中的全部事实，并且只说了事实，这组数据是被良好表达的。

简化后的表达：如果可视化说的全是事实，而且只说事实，那么这个可视化方案是表达良好的。

效率方面：当一个可视化方案传达的信息比其他可视化方案更容易被感知，这个可视化方案是更有效率的。

简化后的表达：如果一个可视化方案使用人们更加容易接受的方式来传达信息，那么这个可视化方案效率高。

常见的可视化的维度

位置，长度，区域面积体积，亮度，色调，角度，形状

不同数据种类下，比较好的数据维度

定量的数据：位置、长度、角度

定序的数据：位置、密度(浓度)、色彩的饱和度

名词性数据：位置、色调、质地、纹理、连接关系

图像感知的概念

图像感知是图像的观察者理解并且能够解释图像，从而获取相应的信息的能力

JND最小可视差异

JND最小可视差异指的是人类或者动物，对于某一个特定的感官刺激，所感感受到的最小改变

Gestalt分组原则

Gestalt分组原则描述的是大脑在哪些情况下可能会将物体认为是一个整体，即大脑是怎样对物体进行分组的。

主体、背景原则、接近性原则、相似性原则、对称性原则、连通性原则、封闭性原则、共同命运原则、透明度原则

降维方法

降维的概念

为了方便于对数据进行处理和可视化，将数据从高维数据映射到低维的数据分布中，并且尽可能多的保存数据的特征。

使用降维方法的意义：最小化测量的高维数据和映射到的低维数据分布之间的差异性。

降低数据维度之后的优点

1、存储数据的花费更小

2、数据维度比较低，更加容易计算

3、能够去除一部分数据的噪声，提高数据的质量

4、二维、三维空间的可视化更加容易实现

降维方法

传统方法：

- 主成分分析法PCA：

  基本思想；通过正交变化将一组可能存在相关性的变量转变成一组线性不想关的变量。找到一个单位向量u，将数据整正交投影到与u对应的方向上，并且保证投影数据$u^t x$方差最大

- 线性判别分析LDA

  基本思想：将多维数据分成多个类别，在进行降维时，使同类的样本尽可能相近，不同类的样本尽可能远离。假设有D维样本，其中N1属于w1，N2属于w2.将样本投影到一条线上且保证同类的样本降维后仍然接近。

  特点：能够保存尽可能多的类别信息

- 多维缩放MDS

  要求原始空间中样本间的距离在低维空间得以保持

传统方法的问题：需要监督才能产生集群，而且PCA和LDA对可靠性数据没有作用。

改进方法(SNE和t-SNE技术)

为了解决这些问题，发明了SNE和t-SNE技术

优点是：可以无监督的产生集群，可以处理可靠性数据

SNE使用条件概率计算两个点之间的相似度，使用高斯分布将距离转化为概率分布。

SNE的问题：1、难以优化。2、存在拥挤问题

t-SNE

对比于高斯分布，t分布受异常值影响更小，因此t-SNE使用t分布将距离转化为概率分布。

高维空间下，t-SNE使用高斯分布进行距离到概率分布的转换

低维空间下，t-SNE使用长尾t分布进行距离到概率分布的转换。

困惑值：一个点周围有效邻接点的个数。

力导向图

力导向图的定义

力导向图时一种描述关系图节点之间的关系的方法，在绘制时，只需要知道两个点之间是否有关系。

力导向图的基本算法

注意力导向图之的计算只针对于存在关系的两个节点

- 初始化：

  每个节点作为系统中的一个粒子，被初始化到一些随机的位置

- 力的作用过程：

  节点收到某些力的作用，被逐渐作用到某些位置上，通过这种方式形成的系统就是力的导向图。

  两种力：前提两点间的距离为d

  斥力是仿照排斥力的计算公式$F = \frac{Kr}{d^2}$

  引力是仿照弹簧拉力的计算公式,L是弹簧的原长$F = Ks(d - L)$

- 迭代过程：在一个循环中不断迭代，每次计算两种力作用导致的位移并且更新节点的位置，直到收敛到一些比较好的位置

- 迭代结果：两种力不断的作用在节点上，节点在不断位移之后逐渐趋于平衡，达到一个稳定的状态，这种稳定状态后形成的就是力导向图。

力导向图中需要调节的参数

计算力的参数Kr、Ks、弹簧原长d，迭代过程的步长Δt

基本力导向图的问题

- 迭代的步长不易确定：在力导向图的生成过程中，迭代步长是一个十分重要的参数：如果步长过小会导致需要迭代步需要合并；如果步长太大会导致形成的合力太大，可能会造成系统的震荡，不易达到平衡稳定的体系。
- 原始算法的时间负责度过高：原始算法的时间复杂度是$O(n^3)$
- 如果图中的节点和连接边的数量过多会导致边的交叉问题(**较大网络图中的问题所在**)

  如果网络中的节点数量太多，且连接的边较多的话，在力导向图中边的多次交叉会使得使用者很难判断与节点连接的是哪一条边。因此过于密集的网络图会导致很多个节点的连接信息无法全部显示。

较大力导向图需要解决的问题

从上面的部分可以了解，对于较大的网络图，由于系统中边的多次交叉，会导致使用者很难看到一些边的连接情况。因此，在较大的网络图中使用者很难获取到每个节点的全部信息。

在较大的力导向图中还需要解决三个问题：

1、可读性上的问题：美学上进行优化

2、计算性能的问题：在算法复杂度上进行优化

3、视觉复杂程序上的优化：需要交互工具允许对显示在屏幕上的网络信息进行限制。

力导向图的原始算法存在性能问题，对原始算法的优化：

1、优化距离的计算(减少平方根的计算)：在进行节点位移的比较时，将原先的两个距离的比较，改为两个距离的平方的比较，减少了计算平方根的时间

2、加入温度概念：加入一个温度概念来表示图像的绘制进展，为了使得算法更好的趋于稳定，允许节点在迭代初期移动比较远的距离，之后逐渐限制节点的移动距离

3、解决两个点邻居节点完全相同，收到的力也因此完全相同，会卡在一起的情况：如果两个点具有相同的邻居节点，则他们可能会卡在一起。这个时候系统生成一个随机的排斥力来推开他们。

4、允许用户通过改变参数Kr、Ks、d，控制力导向图的最后形状。

5、最快的优化方法----加入弹簧等效替代排斥力的计算，有效提高算法性能。

在力导向图的计算中，排斥力的计算时算法的瓶颈，其计算的时间复杂度时O(n2)。使用弹簧替代排斥力的方法优化计算。

使用弹簧替代排斥力的算法：去掉排斥力，而是使用弹簧进行等效的替代，如果相邻的节点认为存在L长度的弹簧，中间存在两条表的节点则认为存在2L长度的弹簧，以此类推。使用这些多余的弹簧能够等效替代掉斥力的作用效果，将节点分散开。系统中弹簧如果较少，则时间复杂度可以得到很大的优化。

数据透视图(pivot graph)

特点：每一个节点比如m1m2，在图上可能出现很多次，因此两个节点类型之间的边(比如m1到m2点边)也可能出现很多次

绘制的步骤：

1、先对图上的属性进行聚合：

记录每个节点的出现频率和两个节点之间边的出现频率

节点的出现频率：这个点在图中有多少个

边的出现频率：这两个节点之间的边出现多少次

利用这些信息形成一个表

2、利用这个表绘制pivot graph

树状图的绘制

树状图的种类

常见的树状图种类：

逐层缩进图、点线连接图、封闭包围图、层次结构图

本章的树状图的种类是：封闭包围图

基本的RT“点线连接树图”

RT算法的目的：使用点线的结构描述出一棵树状图的结构，且更加合理的利用空间，保证节点分布的密度和图形的对称性。

RT算法设计的核心点在于：

- 清楚的标名每个节点的深度
- 保证边不出现交叉
- 子树的绘制保证有序且对称
- 使用尽可能小的空间。

RT算法的基本思路：

每个节点的坐标是(x，y)，y由这个节点的深度决定；x通过在该深度值上的位移决定。

步骤一：自下而上，从子节点开始生成每个节点的位移

从最下层的子节点开始，每次生成父亲节点时，合并左右两棵子树，保证两棵子树在不发生交叉的基础上尽可能的近，保证子树的形态不改变，记录下每个节点的位移值shift

步骤二：自上而下，确定每个点的坐标

从父节点向子节点，对每个节点位置上的位移值shift求和，计算出x坐标。

根据每个节点的(x,y)数值即可画出整棵点线树形图。

基本的Tree Map(封闭包围图)

基本的Tree Map是一个封闭包围树形图。

Tree Map的特点：每一棵树都以矩阵的形式呈现。矩阵被划分成多个更小的矩形对应其子节点。每次对举行切片获得子矩形。**为了保证切片的均匀性，需要不断的改变切片的方向。**

重点：基本的Tree Map绘制算法

Draw()

{

从父级更改切片方向(水平or垂直)

读取该目录下的所有节点信息和子目录

为每一个节点分配矩形，并且按照数据的比例进行缩放

选择合适的颜色对举行进行上色

对于每个子目录

递归调用Draw()函数

}

改进的Tree Map----正方形树图

改进之后的Tree Map就是正方形树图。正方形树图所维护的最重要的概念是：**宽高比。**

思路：维持切片出来的矩形宽高比始终是接近1:1的正方形矩形，当框高比变差时就会改变切片方向。

重点：正方形树图的优点

1、使用正方形表示树图棵得到更小的周长，减少边界的墨水。

2、易于用户对树状图上的矩形进行选择和查看

3、宽高比接近1:1的矩形更易于比较大小

文本的可视化

文本的可视化管线(流程)

- 源数据
- 对源数据进行NLP自然语言处理

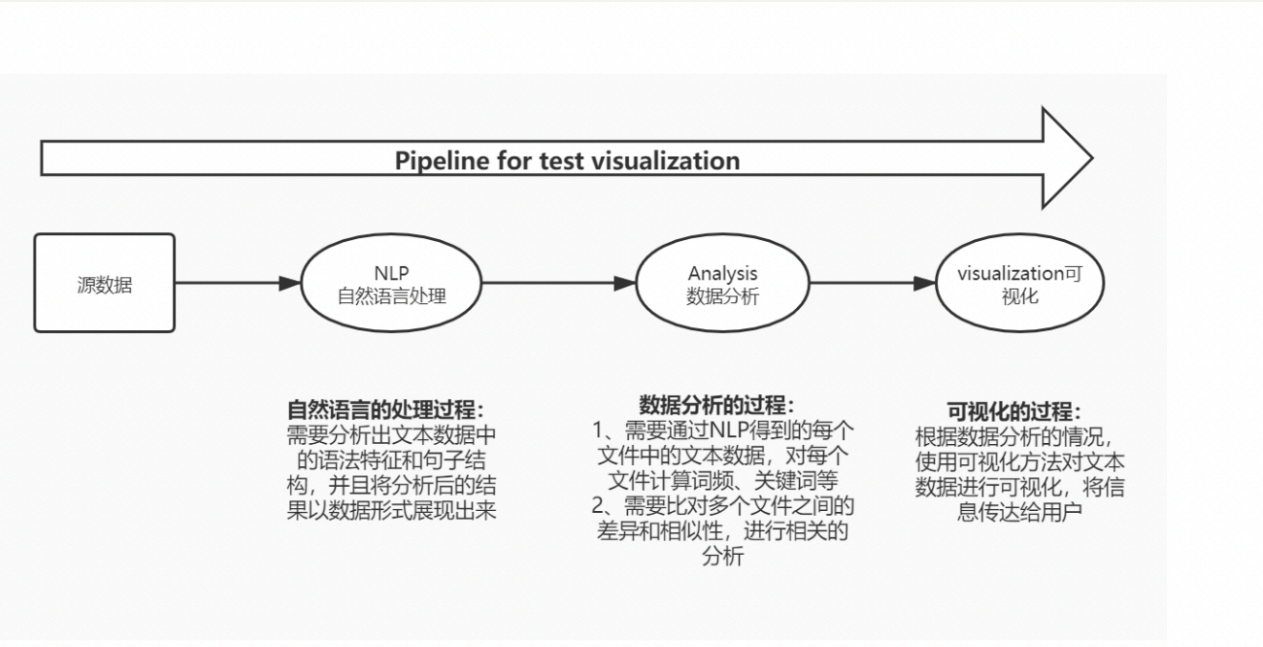
  需要对每个文本数据中的语法特征和句子结构进行分析，分解出每个文件中的词句等信息，并且以数据的形式展示出来
- 对自然语言处理获取的数据进行数据分析

  两个方面：

  1、宏观方面：对多个文件之间的联系、相似度等特征进行分析

  2、微观方面：对每个文件内部的词频、关键词等特征进行分析
- 可视化

  将数据分析得到的情况，使用可视化方法进行展示，将信息传达给使用者



文本可视化的两个层次：

- 宏观层次：对多个文档之间的联系和相似性等特征进行可视化。
- 微观层次：对一个文档内部的词频、关键词等特征信息进行可视化。

重点概念：元数据

元数据适用于描述数据的数据，其存储的是一些特征和结构信息，利于更加便利的管理资源。

常见的文本可视化方法

- 标签云图和词云图
- 气泡图
- 单词树图
- 结果图
- 瓷砖图
- 方面图
- paperlens

设计与交互

缩略-细节图(Overview-detail)模式

Overview-detail是一种将概览图和细节图相连接的交互方式，设计特点：两张图相互连接

1、概览图和细节图相互链接的链接导航之间需要空间分隔，特点是缩略图和细节图一定是两种分开的图进行链接形成的缩略-细节交互模式

2、具有快捷导航的功能，保证能够从概览图向细节图进行浏览，且不会改变图上的细节

3、细节图中做出的更改可能无法立刻在概览图上体现出来

焦点-上下文(Focus-Context)模式

Focus-Context焦点-上下文交互方式是将图中的基本信息和关注的信息(Focus)相结合的图一张复合图

基本思想是：Focus-Context模式能够使得使用者能够看到呈现的主要关注对象，同时获得周围的可用信息。

Focus-Context焦点-上下文图使用的场景，三个前提：

1、用户需要同时浏览上下文信息和详细信息

2、上下文信息和详细信息中显示的重点不相同

3、上下文信息和焦点信息在一个图中进行的组合，符合人类的视觉习惯。

Overview-Detail和Focus- context图之间的区别性

1、Overview-Detail概览-细节图中包括概览图和细节图两个图，并且在两个图之间进行链接。

2、Focus-Context焦点-上下文图中只包括一个图，是将焦点信息和上下文信息在一个图中进行组合的。

Brushing/Linking技术(画笔/链接)

brushing画笔/linking链接，都指代将同一数据的多个视图链接起来。基本功能：

1、在某一个视图中选择突出显示的案例，在其他视图中页突出显示

2、移动鼠标到案例上，可以显示同一个数据在多个视图之间的对饮关系

3、对一个视图中做的更改，在其他的视图中也会被修改。