001i001618106

01000

JL76

1

001

i0: 00010

0

以10rGF01CKnGi 0;0

teCi:

076

D 检 **0** **P**

**第10章** **数据可视化、可视分析与探索式数据分析**

**10.1** **什么是可视化**

数据可视化的研究由来已久，但是作为一个独立的领域，发轫于美国国家科学基金会 的“图形、图像处理和工作站”讨论组1987年发布的一篇里程碑式的报告——《科学计 算中的可视化》,这个报告明确提出了将可视化发展成为一个科学领域。自该报告发布以 来，可视化领域的研究蓬勃发展。

我们知道，计算的目的是从数据中获得洞察力。可视化技术是一种将数据转换成几何 图形表示的技术，它能够直观地展现数据，提供自然的人机交互的能力。可视化为各个领 域的业务专家提供了有力的工具，帮助他们组织、理解、探索数据，发现以前没有发现的 规律。在大数据时代，可视化在从数据中获取信息和洞察力的过程中扮演了重要的角色。

简而言之，可视化是一种数据的可视表现形式 (Visual Representation) 以及交互技 术的总称。它通过图形化的方式把数据表现出来，方便用户进行观察和理解，并且帮助用 户对数据进行探索 (Exploration), 发现 (Discover) 数据里隐藏的模式，获得对大量数据 的理解和洞察力 (Insight)。

**10.2** **可视化的强大威力**

人们常说，“一幅图胜过千言万语”,意思是说某些事物用文字来表达相当烦琐，很不 直观，但是用图形来表现，则非常容易把握和理解。下面举一个例子。

通过百度地图，查找从“中国人民大学”到“国贸三期”的公交路线，总共查出5条 线路，其中一条线路是.首先在苏州街乘坐地铁10号线，往西三环方向走，到达公主坟 站，换乘地铁1号线，到达国贸站，然后从E1 出口出来，步行1.1公里到达国贸三期写 字楼。路线的文字描述如图10-1 (a) 所示。展示这条线路的另外一种方式是在地图上对

这条路线进行可视化，如图10-1 (b) 所示。



禁 价¥5 地 铁 1 0 号 线 → 地 铁 1 号 线

1小时10分钟 24 .6公里 | 步行21公里

◎ 我的位置

六步970米

口 苏 州 街 站 上 车

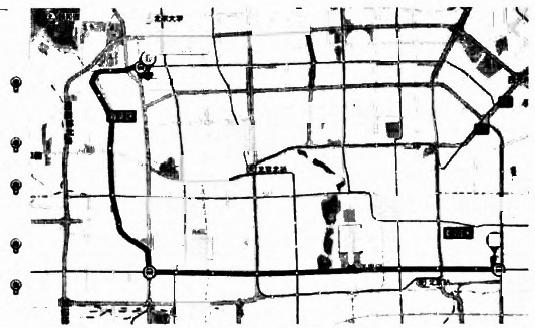
地铁10号线(外环) 日公主坟站下车

六 站 内 换 乘 步 行 3 0 米

日 公主坟站上车

地铁1号线(四虑东方向) 国贸站 (E1 口 出 ) 下 车

六 步行1 . 1公里



n

买

中 人民太季

⑧\*

d

7站

Sw

学 1 号

添

年720分忡

12站

Z

◆章率

激必均10号线中 车程约19分钟

工 人 o

(a) 文字描述 (b) 地理信息可视

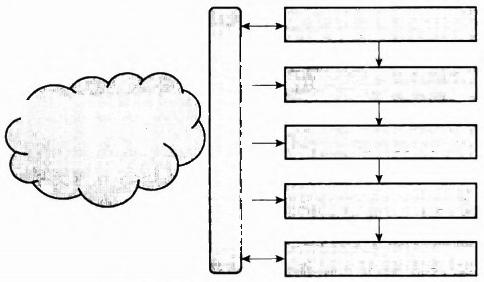
**图10** **-** **1** **从中国人民大学到国贸三期的公交路线之一**

在百度地图上，我们可以一目了然地看到线路的全貌。通过对关键的换乘点进行点 击，就可以获得进一步的换乘信息。这个实例生动地展示了可视化的直观性。

在大数据时代，数据来源多样，数据的规模巨大，可视化技术可以帮助我们对数据进 行观察、理解、探索和发现。

**10.3** **可视化的一般过程**

数据可视化的一般过程包含如下几个主要步骤：数据的过滤、映射、渲染和观察等 (见图10-2)。通过可视化软件的用户界面，用户可以参与到这些步骤中，影响和控制可 视化过程。(1)当数据集很大时，对数据进行过滤 (Filtering) 是必要的。过滤是选取原 始数据集 (Raw Dataset) 的一部分进行可视化，而不是对整个数据集进行可视化。在这 个过程中，有可能需要对数据进行必要的转换 (Transformation)。 原始数据经过过滤，获 得有待可视化的数据。(2)映射 (Mapping) 是指将抽象数据转换为可视化表示的过程。



**数据源(Data Source)**

过滤(Filtering)

用户认知过程(User

映射(Mapping)

渲染(Rendering)

观察(Viewing)

Cognitive

Processing)

用 户 界 面 ( c n f r e n m l s e u (

**图10** **-** **2** **可视化的一般过程**



比如，通过颜色映射表示数值之间的关系。我们通过巧妙设计的映射，可以将复杂抽象的 数据形象直观地展现到一张或者多张图形中，有利于用户快速准确地理解数据。(3)渲染 (Rendering) 是通过图形渲染库和显示卡的帮助，把经过映射的数据以二维或者三维图形 的形式绘制出来，获得可视化的展现效果。(4)最后用户通过图形用户界面，对可视化结 果进行观察和操控。

交互 (Interaction) 是指用户和计算机进行交流。用户做出某种动作，计算机做出反 应。通过设计便捷有效的交互方式，用户可以自由地对复杂数据进行探索。比如，通过捕 抓用户的手势，可视化软件适时翻转三维空间里的对象，方便用户对三维空间里的对象进 行不同角度的观察等。图10-3展示了微软公司的Hololens 系统，支持用户利用手势对可 视化对象进行操控。



**图10-3** **可视化与交互**

资 料 来 源 ：[http://thedaily.case.edu/interactive commons team-displayshololens exhibit-world-economic-forums-47th-an](http://thedaily.case.edu/interactivecommonsteam-displayshololensexhibit-world-economic-forums-47th-an)- nual-meeting/.

可视化是人们与海量数据之间的媒介，它要求我们对上述流程的每一个步骤进行优 化，以获得满意的可视化效果。对大规模数据进行可视化的关键在于概括性以及自然的交 互能力，其中，映射和交互两个环节是达成这个目标的关键。

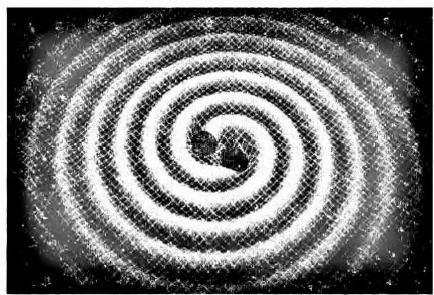
**10.4** **科学可视化与信息可视化**

目前，可视化领域包括三个主要分支，分别是科学可视化 (Scientific Visualization)、 信息可视化 (Information Visualization) 以及可视分析 (Visual Analytics)。可视化领域 的顶级学术会议是IEEE VIS, 它包含三个分会，正好分别对应数据可视化的三个分支， 分别是科学可视化会议 (IEEE Scientific Visualization) (起源于1990年)、信息可视化会 议 (IEEE Information Visualization) (起源于1995年),以及可视分析会议 (IEE Con- ference on Visual Analytics Science and Technology) (起源于2006年)。在这里，我们首 先简单介绍科学可视化和信息可视化，在本章的末尾对可视分析进行简单介绍。

在数据可视化领域，科学可视化是其中最成熟的一个研究分支，它主要面向自然科学 实验、探测活动(如天文观测)、计算机模拟所产生的数据进行建模、操作和处理。科学 可视化是针对特定领域的，比如天文观测、地震研究、医学研究、核物理研究、石油勘探 等，其数据类型较为单一.数据中一般带有物理和几何结构(比如风洞数据、磁共振成像

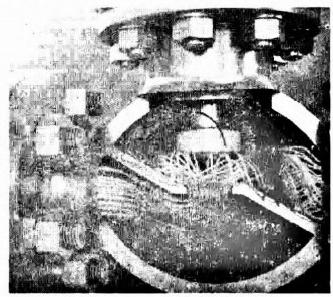
数 据 (Magnetic Resonance Image) 等),可视化的任务 一 般是固定的。

20世纪90年代以来，随着互联网的发展和信息的爆炸，数据可视化的另外一个分 支 — — 信息可视化逐渐兴起。信息可视化需要处理的数据类型丰富多样，可以是数值型数 据，也可以是类别数据，数据具有不同的结构，如层次结构、网状结构等。具体包括时间 序列数据、文本、地图、社交网络等，数据来自不同的行业，比如新闻、电商、股票市 场、社交网站等。图10-4和图10-5分别给出了科学可视化和信息可视化的两个有意思 的实例 。



(a) 黑洞碰撞与引力波

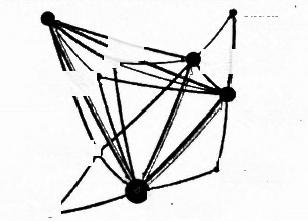
<https://www.sciencenews.org/article/spin>- may-reveal-black-hole-history/.



(b) 液体的流动

<https://www.autodesk.com/solutions/> simulation/cfd-fluid-flow.

**图10-4** **科学可视化实例**



**MAPPING THE STRUCTURE OF SCIENCE**

Analysis of citatons from 17 million computor-soence pubbicabons in Microsoft Academic Search data reveals the relabve importance of research fields,and the fiow of citations betwen them.

inteligence

mathematics

Domoutational

e

Oper

rese

Computer

scence,theory

Computer vsion

Computer

graphics

Intomation theory and ngtworks

Computer

science,software

Computer

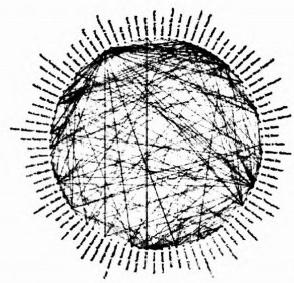
mathematcs

Discrete

Arificial

(a) 计算机科学的结构

<http://www.nature.com/news/2011/110802/> full/476018a/box/1.html



(b) 企业家在社交媒体上的六度距离 <http://www.geekpreneur.com/six-degrees>- of-social-media-for-entrepreneurs.

**图10-5** **信息可视化实例**

表10 - 1列出了科学可视化和信息可视化在目标任务、应用领域、数据类型、主要方

法、面向的用户等方面的主要区别。

**表10-1** **科学可视化与信息可视化的比较**

|  |  |  |
| --- | --- | --- |
| 对比项 | 科学可视化 | 信息可视化 |
| 目标任务 | 研究科学问题，深入理解自 然界中的现象 | 探索、发现信息之间的关系，发现隐藏的模式 |

续前表

|  |  |  |
| --- | --- | --- |
| 对比项 | 科学可视化 | 信息可视化 |
| 应用领域 | 气象、高能物理、天文学、生 物学、医学、地质学、流体力 学 …… | 传感器网络、电子商务、金融、社交网络、新闻、 博客、反恐 …… |
| 数据来源和类型 | ●来自科学实验、观测、仿真  ●结构化数据，具有物理、几 何属性 | ●来自各个领域  ●结构化数据和非结构化数据。一般不具有物理、 几何属性 |
| 主要方法与要求 | ●预处理、映射、渲染、交互  ●准确反映数据中的物理几何 关系 | ●数据挖掘与机器学习、映射、渲染、交互以及可 视分析  ●把抽象复杂的信息及其关系映射为有效的可视化 表示(Visual Representation),寻找合适的可视 化形式 |
| 面向的用户 | ●面向科学家 | ●面向非技术人员、普通用户、管理人员 |

可视化技术已经应用到各个行业和领域，包括软件可视化、数据挖掘和机器学习可视 化、地理信息可视化、生物数据可视化等。

对于可视化应用来讲，不存在科学可视化、信息可视化、可视分析的严格区分。针对 具体的应用，有可能需要结合不同来源的数据，需要科学可视化和信息可视化的相关技术 进行展现。此外，可视分析技术可以帮助用户对数据进行建模 (Modeling) 、 建立假设 (Hypothesis) 以及进行假设检验 (Hypothesis Testing) 等。

**10.5** **数据可视化的原则**

可视化的目的是把复杂数据有效地展示出来，首要的原则是准确 (Precision) 和清 晰 (Clarity) 。 准确是指可视化结果反映的是数据的本来面目或者本质 (Substance), 清 晰是指可视化结果所表达的含义要明确。(1)我们希望在更小的空间里 (Less Space), 用最少的图形(并非越烦琐越好而是越简洁越好， Less Ink),在最短的时间里 (Less Time), 传达给用户最多的信息 (More Ideas)。对可视化效果进行合理简化，突出重点。

(2)可视化的结果，需要阐明事物之间的相互关系，以及事物的变化趋势，对于类似的 事物要方便用户进行比较。需要结合时间、空间因素进行设计，包括使用箭头，创造流 动感等。比如图10-6通过箭头表示从世界各国流向美国的旅游者。(3)使用用户熟悉 的事物，对需要比较的数据进行比较。比如图10-7中，对不同的数据量进行形象化的 比较。(4)构建实物场景，生动展现数据。比如图10-8中，通过实物场景，生动形象 地展现了土地的使用情况。(5)在可视化设计过程中，要考虑把交互方式和动画效果加 进去。动画效果可以从时间和空间维度对事物的发展变化过程进行刻画，以便给用户创 造沉浸式的体验。





**图10-6** **来自世界各国的旅游者(美国)**

资料来源：<http://bloggerspath.com/20-well-designed-informative-infographics/.>



**Yottabyte?**

TERABYTE

ru

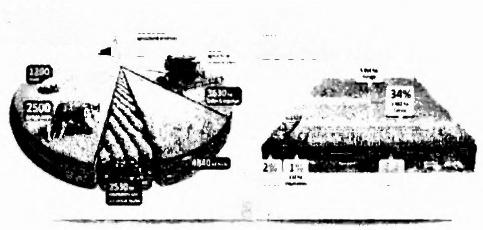
A 两地a

ZETTABYTE 10Au

YOTTABYTE

**图10** **-** **7** **Yottabyte到底有多大**

资料来源：<https://gizmodo.com/5557676/how-much-money-would-a-yottabyte-hard-drive-cost>

Lod bak

A

120000.

**70000**

**10855**

1300

55%

81

**图10-8** **土地使用情况**

资料来源：<https://blog.kurtosys.com/storytelling-data-visualization/.>

总之，我们需要从空间布局、颜色、用户的交互方式、适应用户的感知和认知 (Per- ception &.Cognition) 习惯等方面，对可视化的效果进行优化设计，使得可视化的结果具 有视觉上的冲击力，对用户产生吸引力 (Attractive)。

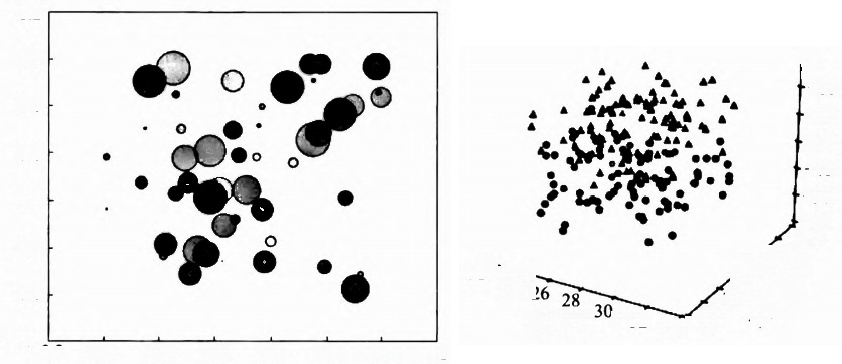


**10.6** **可视化实例**

**10.6.1** **散点图与直方图**

散点图 (Scatter Plot) 是对点数据 (Point Data, 即向量)的集中趋势、分布形 状、离散趋势进行把握的基本的可视化形式。集中趋势是指数据向中心点靠拢的趋势。 分布形态包括数据的分布是对称的还是非对称的，平缓的还是比较陡峭的等。离散趋势 指的是数据离开中心点的趋势。对于1维、2维和3维数据，我们可以使用1维、2维 和3维的散点图，对数据的分布特点进行观察。图10-9展示了2维和3维散点图的

ZLabel



实 例 。

1.2

1.0

0.8

-20

0.6

-40

0.4

-60

120

0.2

22

0.0

-0.2

-0.2

(a)2 维散点图 (b)3 维散点图

0.0 0.2 0.4 0.6 0.8 1.0 1.2

X*La****b***e*e*l 32 34

*0og*

20 *Label*

242

t⁰

-30

-50

-20

600

-10

0

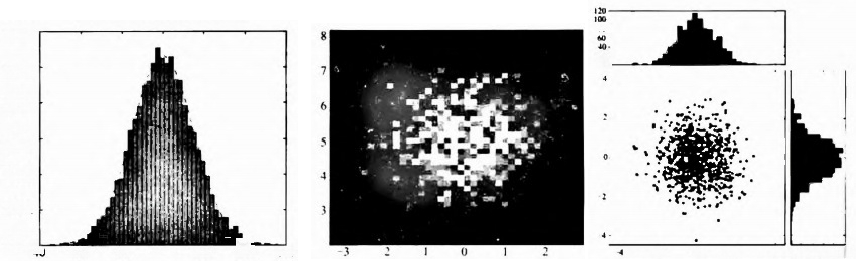
**图** **1** **0** **-** **9** **散点图**

直方图，也称为频率直方图 (Frequency Histogram), 它是统计学中用于表示频率分 布的图形。在直角坐标系中，横坐标表示随机变量的取值，横轴一般划分成一系列的小区 间，每个区间对应一个分组(每个分组也称为一个桶 (Bin)), 作为小矩形的底边。纵坐 标表示频率，每个分组的小矩形的高度表示随机变量取值落入该区间的频率。一系列的小 矩形构成频率直方图 。

我们平时使用的直方图有1维直方图、2维直方图等，频率直方图可以和散点图结合 使用。图10-10展示了三个直方图的实例，分别是1维直方图、2维直方图以及1维直方 图和散点图的组合，在这个组合图中，在平面上以散点图的形式展现了一组2维数据点， 并同时展示了x 方向和y 方向的直方图。



Probability



**Histogram of IQ:μ=100,o=15** 80

0.030

20-

0.025

0.020

0.015

0.010

0.005

0.006 60 80 100 120 140 160 2 0 1 4020406080100 Smarts

**(a)1 维直方图** (b)2 维直方图(不同灰度/ **(c)1 维直方图和散点图**

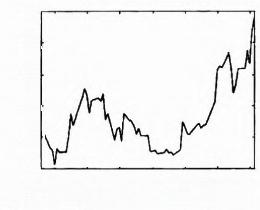
颜色表示不同频率) **的组合**

**图10-10** **直方图、直方图和散点图的组合**

**10.6.2** **线图**

线图通过画直线或者样条曲线，把若干数据点连接起来。线图分单线图 (Line Graph) 和多线图 (Multiple Line Graph)。多线图把多个数据点集合在一个坐标系里进 行绘制，为了区分不同的数据点集合，可以采用不同的线宽 (Thickness) 、 颜 色 (Col- or) 或者线型 (Line Pattern, 比如实线、虚线、点划线等)。图10- 11展示了单线图和

多线图的实例。



30-

29

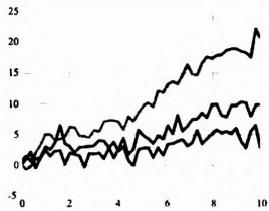
28

27

26

25L

*Jun 252003*9920032320*A*0*ug*3*062* *202*032*172003*



High

Close Open

Low

High

Open Close

Low

(a) 单线图 (b) 多线图 (c) 蜡烛图

**图10-11** **单线图和多线图**

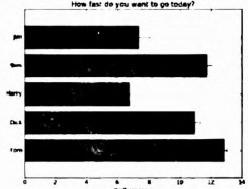
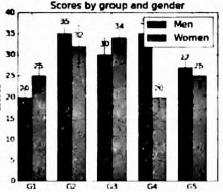
时间序列数据(比如一段时间内的温度变化),可以用单线图来表示。对于某些特 殊类型的时间序列数据，可以使用特定的可视化形式。比如一段时间内每天的股票价 格， 一般用蜡烛图 (Candlestick) 来表示。每天的股票价格被浓缩成了开盘价、收盘价、 最高价、最低价四个价格，用一个单纯的线图无法全面表示价格的变动情况。蜡烛图的 每根蜡烛则可以表示这些信息。当某日的收盘价高于开盘价，蜡烛渲染成红色，蜡烛的 底边表示开盘价，顶边表示收盘价。当某日的收盘价低于开盘价，蜡烛渲染成绿色，蜡 烛的底边表示收盘价，顶边表示开盘价。蜡烛两侧引出的线，分别表示当天的最高价和 最低价。

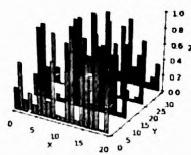


**10.6.3** **柱状图与饼图**

柱状图和饼图一般用来显示一个数据系列里各个数值之间的相对大小关系。柱状图的 各个柱子的高度的比例关系以及饼图的各个扇面的大小的比例关系，反映了数据系列中各 个数值之间的大小关系。

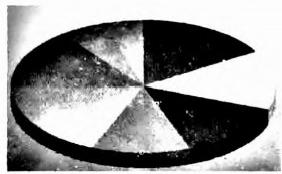
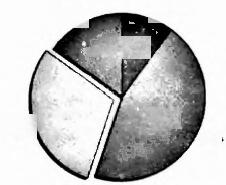
柱状图和饼图也可以显示多个数据系列，对应的柱状图为复式柱状图或者复式饼图。 此外，柱状图和饼图可以渲染为2维的效果，也可以渲染为3维的效果。图10-12显示 了柱状图和饼图的实例。





5

(a) 柱状图 **(b) 复式柱状图** **(c) 复式柱状图(3维渲染)**



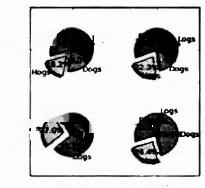
*Loy*

ro

马.6%

30.0%

Deg



rog

m

中

Srmy Cudy

ep

umy

10gt

(d) 饼图 (e) 复式饼图 (f) 饼图(3维渲染)

**图10-12** **柱状图和饼图**

**10.6.4** **解剖图、切片、等值面**

人体和动物的解剖图可以是手绘的，也可以是利用3维建模技术重建的。图10-13 展示了心脏及血液循环系统的解剖示意图。

CT(Computed Tomography) 设备和MRI(Magnetic Resonance Imaging) 设备可以 对人体进行断层扫描，获得高分辨率的人体切片图像(图10-14为脑部切片图像)。这些 图像连续播放，可以帮助医生观察和判断人体内部是否发生病变。

等值面是指空间中的一个曲面.在该曲面上函数f(x,y,z) 的值等于某一给定值 f,. 即等值面是由所有点S={(x,y,2):f(x,y,z)=f,} 组成的一个曲面。等值

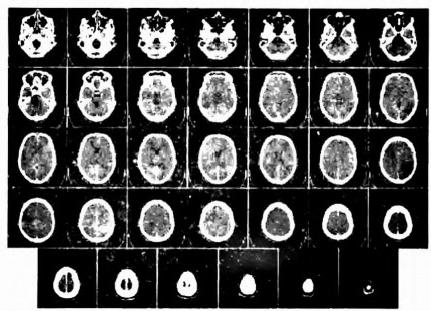
面技术在可视化中的应用非常广泛.许多标量场的可视化问题都可归纳为等值面的抽取和 绘制，如各种等势面、等压面、等温面等。





**图10-13** **人体血液循环系统解剖示意图**

资料来源：[http://www.freedigitalphotos.net/images/download-free.php?id=446952 &.key=0427e3d4.](http://www.freedigitalphotos.net/images/download-free.php?id=446952&.key=0427e3d4.)



**图10-14** **脑部的断层扫描图像**

资料来源：<https://en.wikipedia.org/wiki/CT_scan.>

**10.6.5 表现层次关系：树、圆锥树、** **Tree Map、信息立方体**

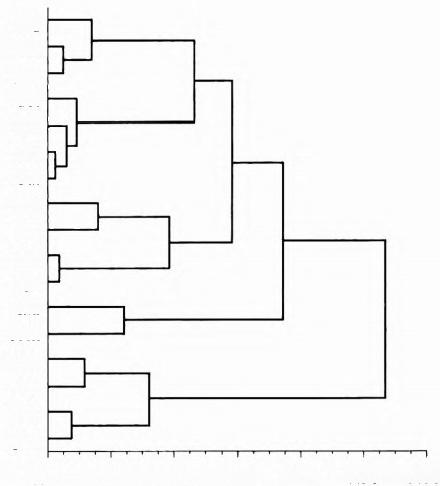
**1. 树** **(Tree)**

树状结构是可视化中应用最广泛的一种图形结构，它一般用于表现某种层级关系，比 如某个组织的各个部门、某个家族的族谱等。凝聚层次聚类是一种自底向上的聚类方法， 各个类簇逐步基于相似度，凝聚成更大的类簇。类别结构存在天然的一种层次关系，可以 用树状结构表示出来。

图10-15是凝聚层次聚类的可视化实例。这个凝聚层次聚类基于Los Angeles County 的17个学区(17 School Districts) 的若干指标，对学区进行聚类。这些指标包括5年级 学生的阅读课考试分数、5年级学生的数学课考试分数、5年级学生的语言课考试分数、 LEP 学生的比例①等。



①Limited English Proficiency(LEP)is refered to a person who is not fluent in the English language.often be- cause it is not their native language.



ing lan plm

**dor**

**abc**

**ccu**

bur

pas

lbu

**gln**

sgu

lau

com

*smm*

tor

bhu

pvu

0.00 1.50 3.00 4.50 6.00 7.50 9.00

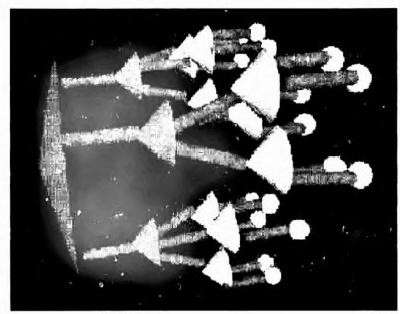
**图10-15** **凝聚层次聚类的可视化**

资料来源：<http://docs.roguewave.com/imsl/java/6.1/manual/default.htm?turl=worddocuments%2Fdendrogr>-

amchart.htm.

**2. 圆** **锥** **树** **(Cone Tree)**

圆锥树用于对层次结构进行可视化展现。在圆锥树中，层次结构通过3维方式进行展 现，以利于最大化使用屏幕空间，以及展现整个层次结构。每个节点表示为圆锥的顶点， 它的子节点在该圆锥的底面周围绘制出来。用户可以通过交互式动画 (Interactive Anima- tion) 的方式，重点查看层次结构的不同部分。圆锥树把一部分用户认知的压力，转移到 人类的感知系统。图10- 16是圆锥树的一个实例。



**图10-16** **圆锥树实例**

资料来源：[http:i/www.infovis-wiki.net/index.php?title=Cone\_Trees.](http:i/www.infovis-wiki.net/index.php?title=Cone_Trees.3.TreeMap)

[**3.Tree Map**](http:i/www.infovis-wiki.net/index.php?title=Cone_Trees.3.TreeMap)

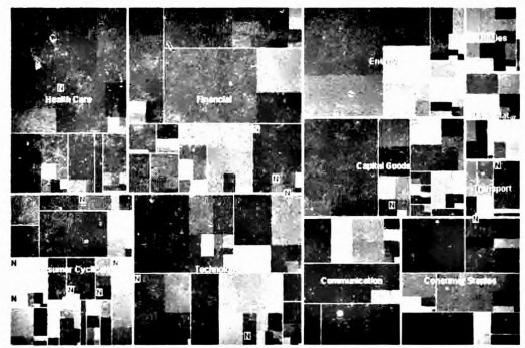
Tree Map是由马里兰大学的Ben Shneiderman教授于20世纪90年代提出的，其最初 的目的是找到一种有效了解磁盘空间使用情况的方法。 Ben 教授首先想到的是用树状结构 来表示，但是这样的图形占用太大的空间。接着他想利用面积来表示文件大小.但是用矩

形、三角形还是圆形都有一定的问题。最后，他想到了将屏幕交替切分为水平和垂直方向 的矩形，用递归来实现层级关系。目前该方法已经用到同时包含层级关系和相对大小关系 的事物的可视化应用中(见图10-17)。



**图10-17** **Shneiderman 教授和** **Tree Map作品**

Smart Money 杂志对市值排名靠前的股票进行了展示，他们使用了Tree Map (见图 10-18)。矩形的大小表示行业、公司的市值，矩形的颜色表示股票的涨跌程度，用从红 到绿的渐变的各种颜色表示。这是一个从设计上来讲非常优秀的可视化实例，可以帮助用 户迅速了解市场状况。



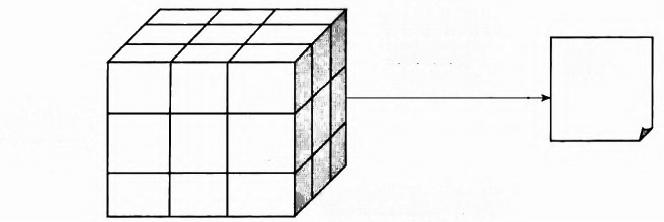
**图10-18** **Smart Money:Map of the Market**

资料来源：<htp://www.cs.umd.edu/class/spring2005/cmsc838s/viz4all/viz4all_a.html.>

**4. 信息立方体** **(Infor Cube)**

信息立方体是一种多维的数据结构，用于从不同维度对数据进行汇总和观察。图10-19 展示了一个关于销售额的信息立方体，它有三个维度，分别是客户维度、产品维度和地区 维度。一般来讲维度具有层次结构，比如产品维度具有产品大类→产品小类→产品这样的 层次结构。用户可以对各个维度基于某种层次对销售数据进行过滤，然后进行汇总和观 察。比如，我们可以查看某个产品大类各个小类在亚洲地区的销售情况。





在各个维度的

某个层次上

进行过滤

地区维度

客户维度(客户1,2,3)

产品维度(产品A,B,C)

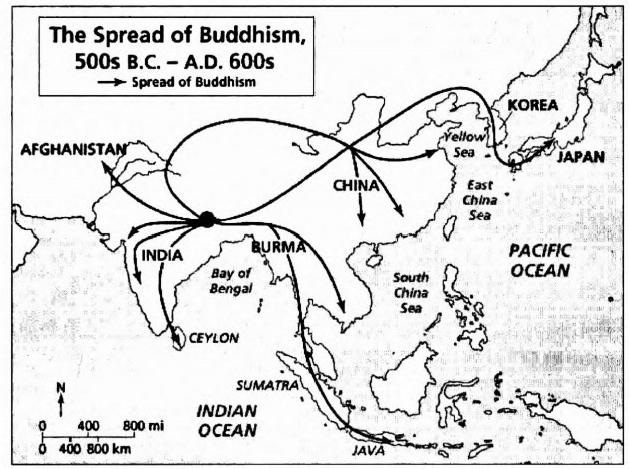
报告

**图10-19** **信息立方体实例(地区维度包括美洲、欧洲、亚洲三个地区)**

**10.6.6 地图** **(Map) 和地球** **(Earth)**

在地图上进行可视化，可以展示事物的发展过程所涉及的不同地理位置。

图10-20展示了佛教在亚洲的传播过程。从图中可以清晰地看到佛教从印度起源， 传播到阿富汗、中国，进而传播到朝鲜和日本。如果在这张图上在关键点上增加交互的能 力，当用户点击时，可以展现什么时间、什么人实现了这个传播过程，对我们了解佛教的 传播将大有益处。图10-21展示了英国约克郡可能发生洪水泛滥的危险地区。点击某个 检测点，可以查看该检测点最近7天检测到的水位变化情况，这些可视化效果和交互能 力，对于防洪救灾具有显著的辅助决策的实际应用价值。



**图10-20** **佛教的传播过程(公元前500年一公元600年)**

图10-22展示了在谷歌地球 (Google Earth) 上渲染飞机航线的可视化效果。我们可 以进行深度开发和集成，在飞行模拟器中集成航线数据和谷歌地球，那么整个飞行模拟过 程的用户体验将获得极大的提升，如图10-23所示。



Up d

24 ncd

240ec00 26Dke 0058 20 Du00 M Do.C000

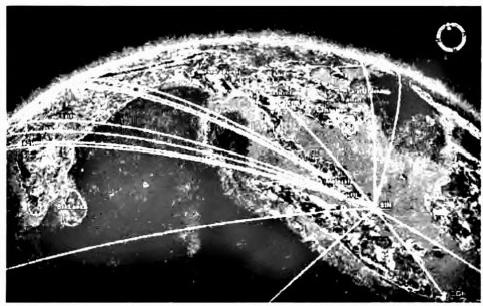
**UK Flood Map &River Levels** i p/y

l 7 ]

0warea rsy-Lu 1210

**图10** **-** **21** **洪水泛滥灾害风险区域与水位监测**

资料来源：<http://www.theinformationlab.co.uk/2015/12/31/rain-destruction-modern-mapping-stack/.>



**图10** **-** **22** **在谷歌地球上显示航线数据**

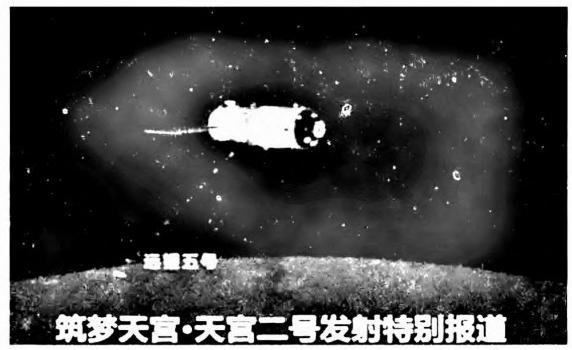
资料来源：<https://blog.openflights.org/2009/02/23/give-your-flights-a-spin-in-3d-on-google-earth/.>



**图10** **-** **23** **利用谷歌地球和航线数据进行飞行模拟**

资料来源：<https://originaldougal.com/up-and-running-with-google-earth-flight-simulator/.>

我们国家在空间站和载人飞船任务中，某相关单位研发了飞行任务轨道3维可视化 软件 · 为保障空间站和载人飞船任务的圆满完成发挥了重要作用。飞行任务轨道可视化 软件实现了对空间站、飞船在轨飞行任务的图形化展示。它直观展示空间站、飞船的运 行状态和运行环境，包括对飞船和空间站对接过程的展示和监控，对实施空天任务运行 过程的全程监控、实现有效载荷运行管理具有重要意义。图10-24为飞行任务轨道可 视化效果图。



**图10-24** **飞行任务轨道可视化(电视屏幕截屏)**

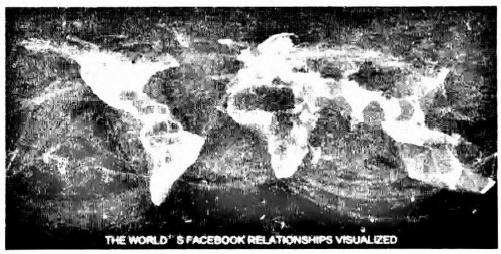
另外一个相关单位研发的航天测控3维可视化系统，集实时仿真、实时监控、遥感数 据处理和信息可视化功能于一身，为航天任务的实施保驾护航。地面测控人员通过该系 统，可以直观、快捷地了解飞船运行状况，并及时调整它的运行轨道，实时监控。该系统 的计算结果和绘制的画面和实际的地球、宇宙空间环境(根据星历精确显示太阳、月球， 乃至其他行星的精确位置)、时间、位置、形状等各种参量和数值基本实现了零误差，也 就是真实反映了航天器及其所处空间的当前状态。

**10.6.7 社交网络** **(Social Network)**

可视化提供了观察社交网络的有力工具，下面举两个例子。

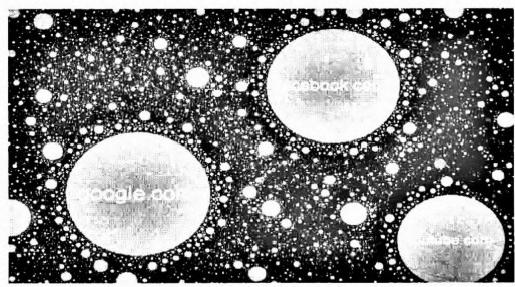
图10-25对世界范围内Facebook 的好友关系进行了可视化。Facebook 的用户规模超 过5亿，如何展现这些用户之间的关系是一个不小的挑战。这张图的作者对数据集进行了 采样，选取1000万对好友关系，并且引入了用户的位置信息 (Locality), 然后确定用户 所在的城市和地区，并且把每对城市之间的好友关系汇总起来，利用城市和地区的经度和 纬度信息绘制两者之间的连接线。这张图可以展现不同地理区域 (Geography) 之间、国 家 (Country) 之间、城市 (City) 之间好友关系的多寡 (Friendships between Them)。

图10 - 26是Internet Map项目(作者 Ruslan Enikeev) 的可视化结果。这张图展示 了各个主要网站的规模和网站之间的关系。每个网站在图中表示为一个圆形，其大小由用 户访问量 (Website Traffic) 决定，即访问量大的网站，圆形的半径越大。圆形的颜色则 代表网站所在的国家。用户通过链接在网站之间切换，这个链接越强.那么在图上两个网 站越互相靠近。



**图10-25** **世界范围内Facebook 好友关系的可视化**

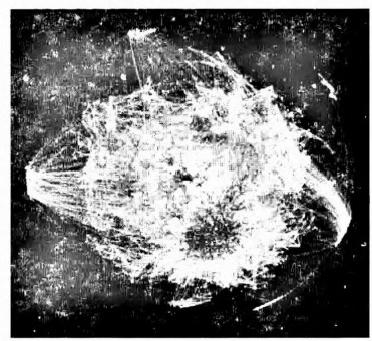
资料来源：<http://www.huffingtonpost.com/2010/12/14/facebook-friendship-map_n_796448.html.>



**图10-26** **Internet Websites Map**

资料来源：<http://internet-map.net/.>

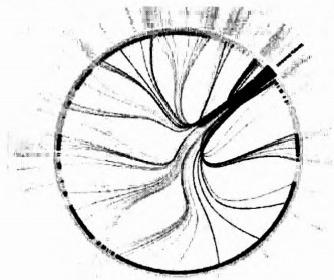
图10-27是对博客空间的可视化效果。节点的大小表示博客的入链 (In Link) 的 多 少，颜色则代表隶属同一个域名的博客。深色的边表示双向的互相引用关系 (Reciprocal Link,A Cited B and B Cited A), 浅色的边表示单边引用关系 (A-reciprocal Link)。图 中 部的密集区域表示博客空间里的社会/政治讨论区，上部的密集区域是技术讨论区，下部 的密集区域是关于家用小工具 (Gadgetry) 的讨论区。



**图10-27** **博客空间的可视化** **(Visualization of Blog Sphere)**

资料来源：<http://datamining.typepad.com/gallery/blog-map-gallery.html.>

图10-28是引文网络的可视化结果 (Radial Graph)。其选用的数据集是1997—2005 年 Thomson Reuters'Journal Citation Reports数据集的一个子集。该子集涵盖了最近10 年来7000个期刊之间的60000000个引用关系。这张图给出了整个引文网络的一个概览 (Overview)。颜色代表了杂志的4大分组 (Group of Journals),在杂志分组底下分成领域 (Field), 也就是图中的外环 (Outer Ring)。内 环 (Inner Ring) 的分段则表示不同的杂 志，内环的大小根据特征向量因子分析的得分 (Eigen FactorR Score)来确定。图中显示 了强度排名 Top 1000的引用连接，线宽和透明度表示连接的强度。



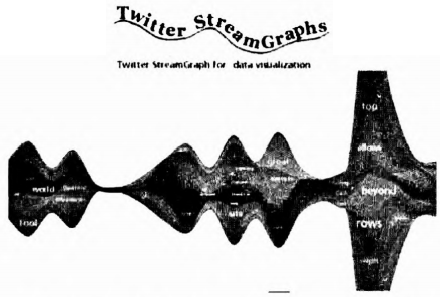
**图10-28** **引文网络** **(Citation** **Network)** **的可视化**

资料来源：<http://well-formed.eigenfactor.org/projects/well-formed/radial.html#/?id=.>

当我们选择某本杂志或者某个领域，针对该杂志或该领域的流入 (Coming In) 和流 出 (Coming Out) 的引用关系高亮起来，连线的颜色根据初始节点的颜色确定，比如计算 机科学领域的颜色为黑色，黑色的线表示从计算机学科领域流出的(指向其他领域，即计 算机科学领域的杂志对其他科学领域杂志的引用关系)引用。

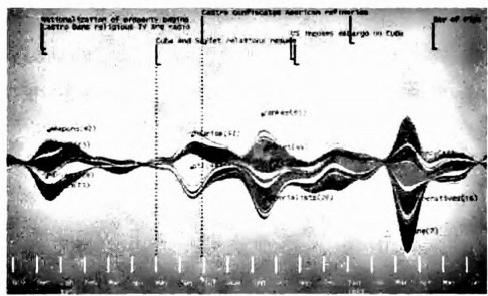
**10.6.8 堆叠的河流** **(Stacked River)**

Jeff Clark 创建了Twitter Stream Graph 可视化效果。他通过堆叠的河流(见图10- 29),显示Twitter 数据流里流行的关键字 (Top Trending Keywords) 随时间变化的情 况。通过这张图，我们可以了解到不同的时间段里，Twitter 数据流里最流行的一些关键 字及其频率对比。



**图10-29** **堆叠的河流** **(Twitter Stream Graph)**

在堆叠的河流上，不仅可以简单地显示关键字，而且可以在文本分析的基础上抽取文 本的话题信息，然后用堆叠的河流显示出来。早在2000年. Havre,Hetzler,Nowell 等 就创建了Theme River技术，表示社交媒体上新的话题的变化情况。每个话题是一个沿着 时间流动的河流，宽度随时间变化。不同话题用不同颜色表示。整个图像(见图10-30) 提供了一个话题演变的概览，显示了我们关心的所有话题，使我们可以在任意时刻找出最 重要的一些话题。

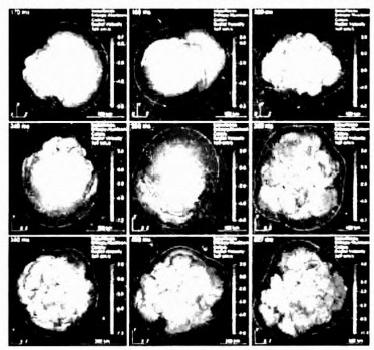


**图10-30** **话题河流** **(Theme River)**

**10.6.9** **多维数据的展示**

如果数据本身的维度为2维或者3维，我们可以方便地设计可视化效果。当数据为4 维以上而且维度不高时，我们可以固定某些维度为常数值 ·进行数据的可视化，然后改变 这些维度的值，再进行数据的可视化，最后获得一系列可视化结果。我们通过观察一系列 的可视化结果，可以了解事物的发展过程。

图10-31展示了关于恒星坍塌模拟的一系列体绘制 (Volume-Rendering) 结果。该 可视化效果展示了20个太阳质量的恒星走向坍塌，然后引起中微子驱动的爆炸 (Neutri- no-Powered Explosion) 的过程中，剧烈的非球面的质量运动情况 (Non-spherical Mass



**图10-3120个太阳质量的恒星的坍塌过程的可视化**

资料来源： <http://wwwmpa.mpargarching.mpg.de/mpa/research/current_research/hl2015-8/hl2015-8-en.html.>

Motion) 。 白色的核心球体表示新生成的中子星 (Neutron Star), 笼罩着中子星的蓝色的 表面表示超新星的冲击波 (Supernova Shock)。这些可视化的画面可以以动画 (Anima- tion) 的形式展示出来。

数据的可视化是为数据集创建一个2维或者3维的图像。如果数据的维度很高，那么 我们需要对数据进行降维 (Dimension Reduction), 然后进行可视化。下面的实例对高维 的用户行为数据集进行可视化。①

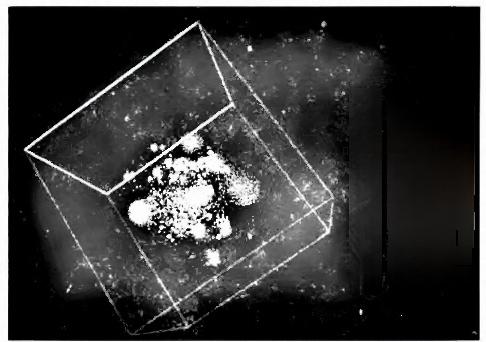
该实例使用的用户行为 (User Behavioral Dataset) 数据集包含大量的用户，每个用 户从上万个维度进行描述。我们可以把该数据集看作一张2维表，行方向表示不同的用 户，用户数量达到上百万个，列方向代表用户的行为属性，属性数量达到上万个。如果某 个用户具有某个属性，相应的单元格 (Cell) 为1,否则为0。这张2维表是一张大型的稀 疏的矩阵。每个用户可以看作高维空间的一个点。

这样的高维空间的一系列的点，很难进行描述 (Describe)、想 象 (Think About)和 可视化 (Picture), 直接在上面运行机器学习算法，效率也不高。机器学习流水线上的一 个重要步骤是降维。可以使用奇异值分解 (Singular Value Decomposition,SVD) 算法进 行数据降维。经过SVD 算法处理以后，数据从上万维降低为上百维(比如200维)。在这 个200维的空间里，每个用户对应一个点，行为相似的用户 (Behave Similarly) 在这个空 间里互相靠近 (Close to Each Other),但问题是200维仍然不容易进行可视化。

为了进一步降低维度，以便进行可视化，可以继续运行另一个算法t-SNE(t-Sto-

chastic Neighbor Embedding)。该算法接受高维数据，然后创建低维的可视化图像，并且 保证高维空间中互相接近的数据点，在低维空间(可视化结果)里也是互相接近的。注 意，空间里数据点互相接近，表示用户的行为相似。

最后的可视化效果如图10-32所示。三维空间中的每个点代表一个用户，点的颜色 表示用户隶属的类簇 (Cluster) 。 互相靠近的点，表示对应的用户的行为是相似的。用户 所属的类簇 (Cluster) 是通过在高维空间里运行K-Means 算法 (K=10), 赋予每个用户 一个类簇标签来确定的。在高维空间运行聚类算法，避免降维过程引起的信息丢失对聚类 结果造成的影响。降维过程则帮助我们对数据进行可视化。

[+] [+] +】 [+1 【+]

,

4 [+] [+I

AL1

cluster

Cluster

cluster

cluster

clustor

Cluster

Clustor

Cluster

Clustor

cluster

o 1 2 3 4 5 6 7 8

9

**图10-32** **高维数据的可视化实例**

资料来源：<https://github.com/datacratic/data-projector.>

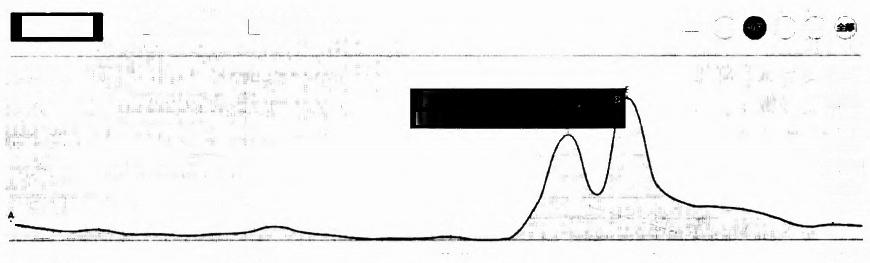
①<http://datacratic.com/site/blog/visualizing-high-dimensional-data-browser-svd-t-sne-and-threejs.>



最后的可视化结果显示，不同类簇的用户(数据点)通常是紧密靠近在一起的。也就 是，在高维空间中相互靠近的(相似的)用户，经过降维以后，在低维空间仍然互相靠近 (相似)。

**10.6.10** **特色可视化应用**

百度指数 (<http://index.baidu.com>) 和360指数 (<http://index.so.com/>) 可以 根据用户输入的关键字，显示该关键字的搜索趋势(搜索热度)。图10-33显示了通过百 度指数查询到的“天宫二号”最近30天(2016.9.28—2016.10.27)的搜索趋势，图中的 E点表示2016年10月19日神舟十一号与天宫二号成功实现自动交会对接当天，搜索热 度达到最高值。



最近 7天

口平脑

:5

·神舟十一与天度二号成功实预当动交台诗接

33,000

27,000

21,000

15.000

C

10月13日

**PC 趋** **势** **□** **移** **动** **趋** **势**

9ha as

2016年9月28日

9,00o

10月27日

45,000

39,000

按射

90天(半年

■天言二号

10月20日

10月6日

B

**图10-33** **百度指数——** **“天宫二号”**

图10-34显示了通过360搜索指数查询到的“Hadoop” 从2013年1月到2016年10 月的搜索趋势，从图中可以看出这段时间内，Hadoop 的搜索热度总体上处于上升趋势， 显 示 了Hadoop 作为大数据处理平台，其强大的处理能力逐渐为人们所了解和认可。

**搜索指数趋势②2013-1-11至2016-10-27** **国** **便** **改**



|  |  |
| --- | --- |
| ●Hadoop |  |
| 2013年1月 2013年6月 2013年11月 2014年4月 2014年9月 2015年2月 2015年7月 | 770  660  00  110  2015年12月 **2016年5月** 2016年10月 |

**图10-34360指数** **—** **—Hadoop**

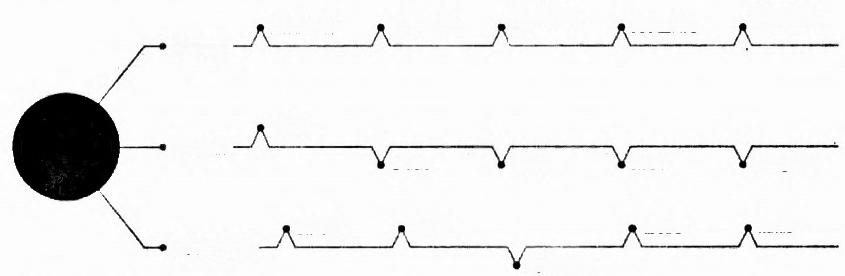
360指数还能显示某个关键字的相关搜索关键字及其近期搜索热度变化趋势，图10- 35展示的是 “Hadoop” 的相关搜索趋势。

百度公司基于其掌握的大数据以及百度地图，打造了若干具有酷炫可视化效果的应 用，包括百度迁徙、百度景区热力图、百度通勤图、百度慧眼等(见图10- 36)。百度迁 徙可以让用户观察到全国范围内的迁徙最热路线、迁入和迁出的最热城市和地区。百度景 区热力图则让用户实时查看各个热点景区的拥挤程度，提前为出行做出安排。百度通勤图 (北京版)展现了上下班时间白领的通勤路线及其热度，对于公交线路规划、地铁规划都



**相关搜索趋势?2016-9-28至2016-10-27** **国** **Hadoop**

驰 ~~1 诞 哪~~



大数据培训 猜悠关注

大数据 相关书籍

Spark

云计算 中国通信学会

OpenStack

数据分析工 具

阿里云服务 器

尚学堂官网

MongoDB

html5

Hadoop

云计算培训

**物联网**

数据库

Oracle

Redis

人\_趋势上升 Y 趋劈下降

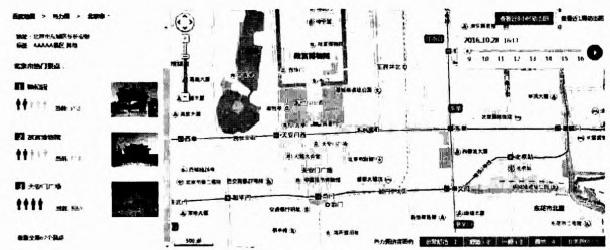
**图10-35360指数——Hadoop的相关搜索趋势**

具有参考意义。百度慧眼则是一项商业服务，它是一款位置大数据获取、管理、分析和可 视化平台，集成了商业信息、地理信息、人口信息。它的一个应用是帮助商业企业以热力 图的方式，对商场的各个楼层的客流进行监控和引导。

通过手机搜索记录，百度能够知道人们在不同时间的位置信息。把大量用户的信息收 集起来，进行汇总和分析，数据的价值就体现出来了。

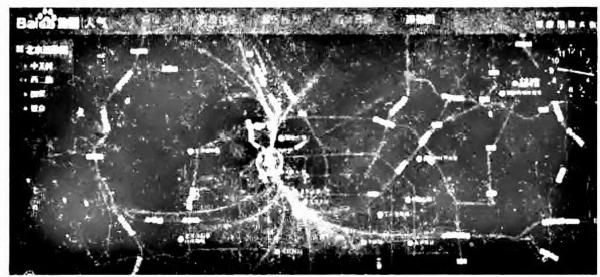


(a) 百度迁徙

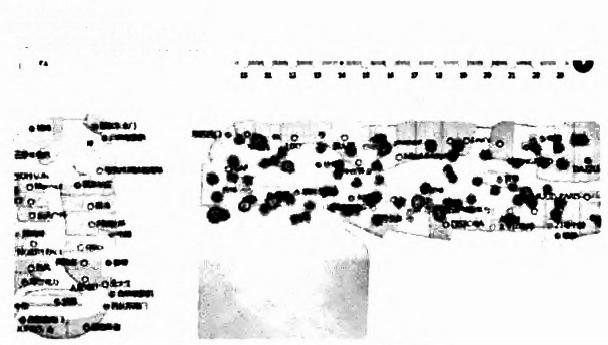


(b) 百度景区热力图





(c) 百度通勤图



I实射说热力●

9 场 **2016年10月28日**

(d) 实时客流热力图

**图10-36** **百度地图应用**

**10.7** **可视化的挑战和趋势**

在大数据时代，可视化面临新的挑战和机遇。在传统的数据可视化应用中，由于数据 规模不大，我们可以设计良好的映射和交互技术，达到很好的可视化效果。大数据时代， 海量异构数据的可视化对算法设计和硬件基础设施都提出了更高的要求，它要求我们必须 对可视化过程的各个环节进 一 步优化。此外，高维的 (High Dimensional)、多 元 的 (Multivariate) 、 多模态的 (Multimodality) 、 时变的 (Time Varying) 数据，以及数据不 完 整 (Incomplete Data)、有噪声 (Noise) 等特点，都对数据的可视化提出了严峻的挑 战，需要数据可视化领域的研究人员提出应对的办法。

在大数据时代，有几个可视化技术的发展趋势值得注意：(1)各种新硬件 (New

Hardware) 应用到可视化领域，可视化系统将支持更高的显示分辨率 (Higher Resolu- tion);(2) 可视化技术应用到更多的业务领域 (Application Domain);(3) 可视化技术支 持更多样的数据类型 (Data Type);(4) 新的研究热点是基于可视化以及可视化分析结果

进行叙事，讲一个故事 (Telling a Story), 并且把故事讲完整、讲精彩；(5)可视化软件 提供更强大的可视分析能力 (Advanced Visual Analytics)。

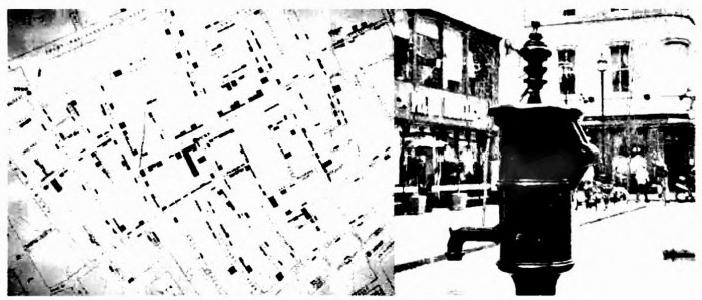
**10.8** **可视分析技术**

虽然可视分析技术(利用计算机)在21世纪初才开始兴起，但是可视分析的思想有 久远的历史。下面举个例子。

19世纪中叶，英国正处于工业革命的高潮。大量人口涌入城市，城市的基础设施不 堪重负，卫生状况堪忧，霍乱疫情频繁爆发，造成大量的人死亡。

当时医学界的主流理论是“瘴气论”,该理论认为霍乱是通过某种有毒气体传播的， 但是有一位科学家不这么认为，他就是斯诺。他认为这个理论不能解释诸多现象。斯诺猜 测，霍乱可能是通过水传播的，他需要证据来证实自己的假设。

1854年，霍乱又一次在英国伦敦爆发。斯诺仔细分析疫情的发展，为自己的假设寻 找证据。数据可视化以及可视分析发挥了重大作用，当然这个可视化是手工完成的。在伦 敦发生的这场霍乱疫情中，布罗德街和剑桥街的交界处是一个重灾区。斯诺创造性地使用 统计学方法，并且对结果进行了可视化，进而对可视化结果进行分析。他根据死亡病例数 据，绘制了著名的“死亡地图”。在这张图中，死亡病例的数量用柱状的长短来表示，并 且在地图上一一标注其位置。斯诺发现，大部分霍乱死者居住在布罗德街水泵附近(见图 10-37中间的小圆点),并且以这个水泵作为饮用水源，其他水泵附近的死亡病例则非 常少。



**图10-37** **斯诺的死亡地图和布罗德街的水泵(已经封存)**

斯诺还发现，布罗德街水泵附近不远的波兰街济贫院共有500多名贫民，但是只有少 数霍乱死亡病例。通过调查，他发现这是因为济贫院有自己的水井，饮水不需要从布罗德 街水泵中取水。

如果“瘴气论”是正确的，那么大家住在同一条街道上，吸着同样的毒气，感染霍乱 的概率应该差不多。因此，受污染的饮用水成为关注的对象。当局把水泵的手柄拆除，居 民不能从该水泵取水而是使用其他水源，布罗德街的霍乱疫情很快消退了。斯诺是历史上 第一个采用统计学的方法对流行病的传播途径进行调查的人，成为当之无愧的“流行病学 之父”。在这个过程中，可视分析可以说立下了汗马功劳。

2005年以来，可视化技术和数据分析技术结合，发展出数据可视化的一个新的分 支——可视分析学，它是借助计算机，以可视化交互界面为基础的分析科学。这个方向的 研究不断受到重视，经过十几年的发展，已经成为可视化领域的重要研究方向。可视分析 技术已经大量应用在科学、商业等领域。

可视分析综合大量的多模态信息(文本、语音、视频、图像、社交网络等),利用可 视化技术以及数据分析技术，帮助人们理解数据。简而言之，可视分析是可视化技术、交 互技术、分析技术的结合。可视化技术和交互技术是为分析和推理服务的。可视分析把若 干计算机科学分支的知识整合到一起，包括图形设计、交互方式设计、信息可视化(包括 数据的表示 (Representation) 和转换 (Transformation)) 、 感知和认知科学、社会学等， 可以认为可视分析是一门多学科交叉 (Multidisciplinary) 的学科分支。

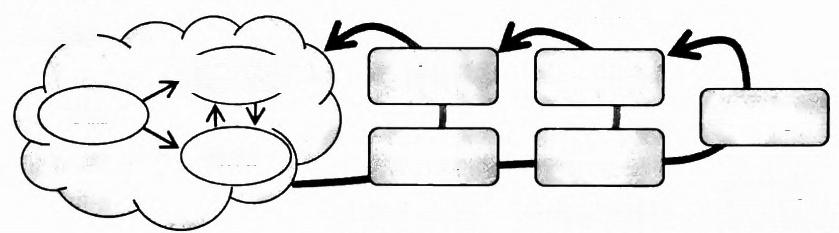
可视分析把交互式的可视化与具体分析技术(包括统计分析方法、机器学习/数据挖 掘方法)结合起来，帮助用户实现高层次的复杂的一系列活动，包括意义构建 (Sense Making) 、 推 理 (Reasoning) 、 决 策 (Decision Making) 等。通过推理分析，用户利用其 判断能力，基于掌握的证据和假设，获得对数据的理解，达成关于数据(数据反映了实际 业务)的结论，从而帮助用户完成评估 (Assessment) 和决策。

在可视分析过程中，用户通过如下方式增强对于数据的认知 (Cognitive):(1) 把大 量的数据在有限的空间里进行整体展示，使得用户对数据有一个总体的把握和初步的理 解；(2)在时间和空间维度上展示数据的变化，帮助用户对数据的模式 (Pattern) 进行感 知和认知；(3)把复杂的网络关系以可视化的方式展示出来，帮助用户基于感知进行事物 关系的推理 (Inference of Relationship);(4) 通过交互式分析的方式，用户可以调整参数 值 (Parameter Value), 并且及时改变可视化结果，对数据进行探索。

一般来讲，数据集的变量可以分为两类，分别是独立变量 (Independent Variable, 即 自变量)和非独立变量 (Dependent Variable, 即因变量)。自变量表示观察值的变化情 况，因变量依赖于自变量的变化而变化。为了对自变量和因变量的关系进行观察，我们一 般需要两个视图， 一个视图显示自变量，另一个视图显示因变量。这两个视图是连接在一 起的，也就是当用户选择了自变量视图的若干数据点以后，另一个视图的数据点将发生变 化，由此可以帮助我们直观地探索高维空间数据的性质。

可视分析是一个迭代的过程，如图10-38所示。可视分析引导用户进入分析流程， 让用户可以通过交互式的界面，将其经验和智慧输入到系统中，不断发现规律，建立假 设，进而肯定或者否定假设。可视分析的目的是引导用户完成给定的分析任务。我们将围 绕图10 - 38(该模型由Dominik Sacha等提出)介绍可视分析的关键组件 (Component)

和关键流程。



Human

Computer

Hyp othesis

Knowledge

**Insight**

Verification Loop

Finding

Exploration Loop

Knowledge Generation Loop

Visalization)

Action

Model

Data

**图10-38** **可视分析的过程(从数据到模型、到知识)**

**1. 数据** **(Data)**

数据是所有可视分析的基础。我们收集的结构化、半结构化、非结构化数据，描述了 自然和社会的事实 (Fact) 。 除了数据之外，我们还需要元数据来描述数据的模式 (Sche- ma) 或者结构.以便可视分析系统能够正确地存取数据，对数据进行可视化。

**2. 模型** **(Model)**

这里讲的模型，可以是描述数据的某个方面属性的一些统计值 (Descriptive Statis- tics), 也可以是复杂的数据挖掘和机器学习算法。比如，我们基于数据进行统计检验 (Statistical Test),达成是否接受一个假设的结论 (Trust a Hypothesis(r Not)。数据挖 掘算法，比如分类或者聚类的结果，可以通过可视化模块显示出来，能够展现数据内部的 一些模式和规律。此外，在数据上创建的模型，可以帮助我们获得一些对数据的理解。比 如，在数据上执行逻辑回归模型 (Logistic Regression Model), 使得我们可以确定最重要 的特征 (Feature) 或者特征组合 (Feature Combination)。

**3. 可视化** **(Visualization)**

在可视分析中，可视化结果一般依据模型的分析结果进行绘制，比如聚类分析模型可 以把数据点划分成不同类簇，通过可视化模块，就可以把不同数据点的类簇归属显示出来 (比如使用不同的颜色表示不同的类簇)。

可视化结果作为分析人员和可视分析系统 (Visual Analytics System) 的界面，用户 基于可视化结果，探测 (Detect) 数据中变量之间的关系。

**4. 探索回路** **(Exploration Loop)**

探索回路描述分析者如何和可视分析系统进行交互，目的是生成新的可视化结果或者 调整模型，并且据此对数据进行分析。它涉及两个方面，即动作 (Action) 和发现 (Find- ing)。

(1)动作 (Action)。这里所说的动作.包括创建模型 (Model Building)、应用模型 (Model Usage, 比如计算统计值或者对数据进行聚类)、把模型映射到可视化效果 (Map Model to Visualization),以及操控可视化效果 (Manipulate Visualization, 比如改变观察 的视角或者高亮部分数据)等。

(2)发现 (Finding) 。 发现指的是分析者使用可视分析系统，发现有趣的观察结果 (Interesting(bservation) 。 发现可以是数据展现出的模式，或者数据表现出的异常情况 等。模式指的是重复出现的事件。有时候，我们没有观察到某种模式，也可以认为是对数 据有所发现。有了初步的观察结果以后，可能需要进一步和系统进行交互，对数据进一步 分析(通过上述若干动作，操控可视分析系统),也可能直接导致对数据的新的理解。探 索过程构成了一个回路 (Exploration Loop), 分析者不断寻找有用的发现 (Finding), 以 解决预定的分析任务。

**5. 验证回路** **(Verification Loop)**

(1)假设 (Hypothesis) 。 在可视分析过程中，假设扮演了核心角色。假设是针对问 题领域构造了一个假说.然后进行验证性的分析。分析者试图找出足够的证据，支持或者 反对该假设，目的是看能否从数据中获得知识 (Gain Knowledge from Data)。从这个角度 来讲，可视分析过程是由假设引导的。

(2)理解和洞察力 (Insight) 。 洞察力定义为分析者对于上述的发现如何进行解释。 在解释过程中，需要用到领域的先验知识 (Previous Domain Knowledge), 产生新的信 息。分析者对数据的理解和洞察力用语言来表达，可能无须太多的篇幅。比如，通过分析



人们认识到数据中的某些属性具有某种关系，这种关系具有一定的显著意义 (Signifi- cant)。

上文中讲的发现 (Finding), 有可能支持某个假设，使得分析者确信假设是可靠的， 于是形成对数据的深入理解。

**6. 产生知识回路** **(Knowledge Generation Loop)**

在可视分析过程中，分析者为某个假设寻找证据，或者从数据中学习到新的知识。从 证据到知识，需要一个推理 (Reasoning) 的过程。在这个过程中，分析者根据收集的证 据，判别这些证据是否足够证明我们对数据的理解是可信的，可以提升为知识.或者需要 进一步检查。比如，用其他数据集进行验证，或者和领域专家进行讨论，然后确定等。

**10.9** **探索式数据分析**

当我们拿到一个数据集，但是对如何对其进行分析完全没有头绪，这个时候，探索式 数据分析可以大显身手。

探索式数据分析 (Exploratory Data Analysis,EDA) 是指对已有的数据在尽量少的 先验假定情况下进行探索，逐步了解数据的特点。当我们对数据的内在特点、所包含的信 息，没有足够经验，不知道应该用什么统计分析、数据挖掘、机器学习方法进行分析时， 探索式数据分析是一种有效的分析方式。

对数据进行探索式分析的主要理由是从数据得到灵感。用户可以在微观层面、宏观层 面进行自由切换，用交互的方式探索数据中潜藏的结构、模式、关系。

约翰 ·怀尔德 ·杜克 (John Wilder Tukey) 是美国贝尔实验室的一名统计学家，他开 拓了探索式数据分析 (Exploratory Data Analysis) 领域。杜克于1977年在《探索式数据 分析》一书中第一次系统地论述了探索式数据分析。他相对于验证性数据分析 (Confirm- atory Data Analysis) 来论述探索式数据分析 (EDA) 的不同之处。前者注重对数据模型 和研究假设的验证，后者则注重对数据进行概括性的描述，不受数据模型和研究假设的限 制。换句话说，探索式数据分析没有假设，也没有模型。杜克认为，统计分析不应该只重 视对模型和假设的验证，还应该充分发挥探索式数据分析的长处，在描述中发现新的理论 假设和数据模型。

探索式数据分析在对数据进行概括性描述，发现变量之间的相关性，以及引导数据科 学家提出新的假设方面，大显身手。从逻辑推理的角度讲，探索式数据分析属于归纳法 (Induction), 有别于从理论出发的演绎法 (Deduction)。

随着大数据研究的兴起，通过探索式数据分析，理解数据，进而挖掘数据的价值，逐 渐引起人们的重视。

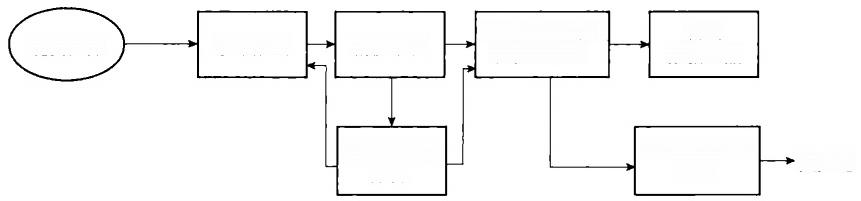
**10.10** **探索式数据分析的作用**

Rachel Schutt 和Cathy O'Neil在其著作《数据科学实战》(冯凌秉，王群锋译)中，把探



索式数据分析列为数据科学工作流的一个关键步骤，它能够影响多个环节，如图10-39所 示。在这里，探索式数据分析是为建立初步的模型服务的。

可视分析



构建

数据产品

探索式数据 分析

现实世界 收集数据 清洗数据

报告 做决定

机器学习算法 和统计模型

可视化/沟通/

探索式数据分析 数据可视化

**图10-39** **数据科学的工作流程和探索式数据分析的作用**

刚开始，我们对数据一无所知或者所知不多，经过探索式数据分析以后，了解了数据 的特点，于是开始建立模型，利用数据进行模型训练，然后评价该模型。由此可见，探索 式数据分析是建立算法和遴选模型的第一步。

如果模型的预测能力不理想，需要重新对数据进行探索式分析，以获得对数据更深入 的理解，帮助建立更加合理的模型。如果发现前期收集的数据不完整，还需要收集更多的 相关数据。

简而言之，探索式数据分析利用人机交互技术和数据可视化技术，通过不断揭示数据 的规律和数据间的关联，引导分析人员发现并认识以前不知道的数据模式或规律。对未知 的数据模式和规律的探索是其价值所在。

在探索式数据分析的过程中，目标是可变的，需要从多个角度观察数据，分析过程是 有选择、反复进行的。

**10.11** **探索式数据分析的基本方法**

探索式数据分析的基本方法分成两类，即计算一些汇总统计量(基于定量方法的)以 及制图和制表等。

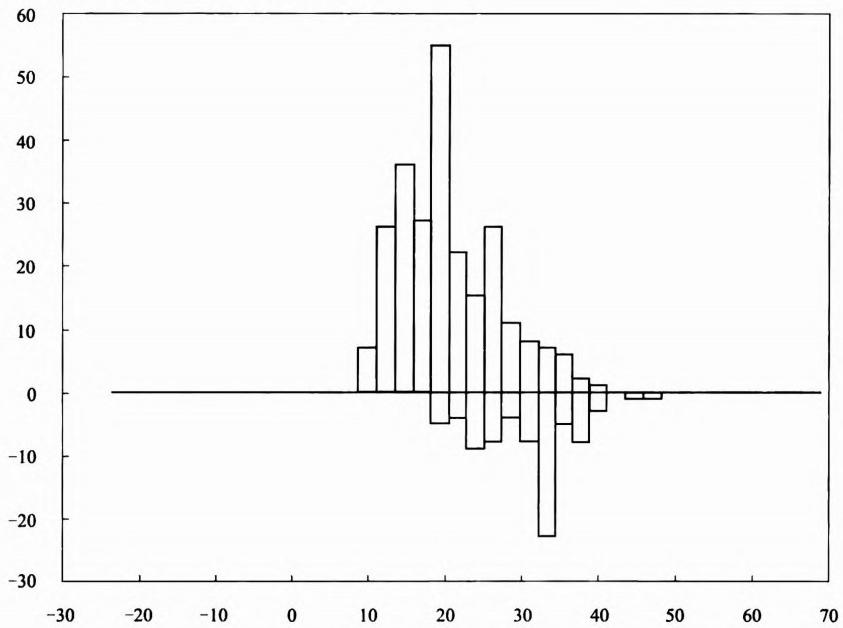
汇总统计量的计算包括计算均值、中位数、众数、切尾均值等，以便让人们了解数据 的典型值、最大值、最小值、上下四分位数、数据的范围等，还可以通过确定异常值，让 人们了解数据的异常情况。

制图和制表能够展示变量的分布情况、时间序列数据的变化趋势以及变量之间的关 系。有时候需要对变量进行变换，再进行制图。

由此看来，探索式数据分析涉及很多的数据可视化需求，但我们还是需要区分一下 数据可视化和探索式数据分析中的数据可视化。 一般来讲，数据可视化一般用于展示分 析结果，发生在数据分析之后。探索式数据分析中的制图，目的是对数据进行观察和把 握，发生在数据分析之前。数据可视化工具的日益成熟推动了探索式数据分析的快速 普及。

**10.11.1** **了解变量的分布情况，计算统计值**

为了解变量的分布情况，可以计算关于变量的一些简单统计值，也可以绘制直方图。 直方图可以直观显示变量的均值、范围、偏度、峰度，是否有异常值等。有一种特殊的直 方图，称为双直方图，如图10-40所示，它可以用于比较两个变量的分布，还可以绘制 某个变量的两个不同水平下因变量的分布情况，从而判断某个变量的不同水平是否对因变 量产生重要影响。比如，我们画出不同性别(具有“男/女”两个水平)的身高的分布， 从而比较性别是否对身高有显著影响。



**图** **1** **0** **-** **4** **0** **Bi-Histogram**

此外，可以使用箱线图，刻画一个变量的1/4分位数、3/4分位数、中位数、最大和 最小值以及异常值等。

**10.11.2** **了解变量之间的关系**

通过图形(比如散点图)可以了解变量之间的关系，使我们最大限度地获得对数据的 直觉。发现变量的各种关系，可以引导我们建立新的假设或者检验潜在的假设，比如将响 应变量与一组因子变量关联起来的最佳函数是什么,是不是线性关系等。

比如图10-41、图10-42、图10-43三个散点图，分别展示了变量X 和变量Y 的 线 性相关关系、二次曲线关系、指数关系等。

**Scatter Plot(Xvs.Y)**

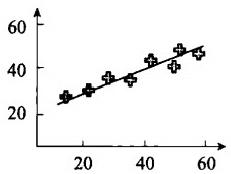
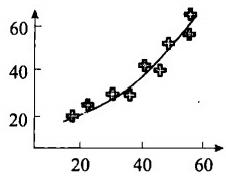


图10-41 线性相关的数据

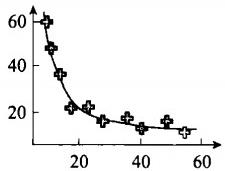
**Scatter Plot(Xvs.Y)**



10-42 二次曲线关系的数据



Scatter Plot(Xvs.Y)

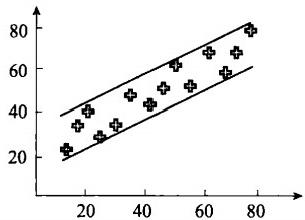


**图10-43** **指数关系的数据**

图 1 0 - 4 4 显 示 ，Y 和 X 大体呈现线性相关关系， Y 有一个变化范围，但是不同的X 值似乎对方差没有影响，这种情况称为与X 独立的同方差扰动。

图10 - 45则展示了X 和 Y 的不同关系，虽然大体上Y 和X 呈现线性相关关系，但是 Y 的变化范围随着X 的增大而扩大，这种情况称为与X 不独立的异方差扰动(随着X 增 大，噪声的方差也增大)。对于异方差，我们可以通过加权最小二乘法、 Box-Cox 变换、 对数变换等方法来进行数据预处理，基于处理后的数据建立模型。

Scatter Plot(Xvs.Y)



**图10-44** **与X 独立的同方差扰动**

Scatter Plot(Xvs.Y)

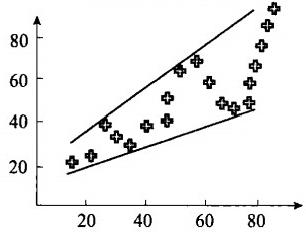


图10-45 与X 不独立的异方差扰动

**10.11.3** **了解因子变量的相对重要性**

在统计分析中，在自变量X,X₂, … ,X„ 中，对于因变量Y 来讲，哪个自变量是最 重要的，即最重要的因子是什么,我们需要判断因子的重要程度，然后给出因子重要性程 度的一个排名。常用的判断因子重要性的方法是DOE 均值图 (DOE Mean Plot)。

DOE 均值图一般用于分析试验数据 (Designed Experiment), 它用于对因子的重要性 给出一个判断，这些因子一般都有若干变量水平(也就是变量的可能取值)。针对每个因

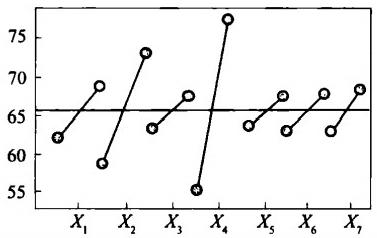


子 ，DOE 均值图显示该因子的不同变量水平对应的响应变量(即因变量)的均值，并且 用一条线把这些均值连接起来。DOE 均值图是传统的方差分析 (Analysis of Variance) 的 补充。当然，除了显示均值之外，也可以显示中位数 (Median)。

在 DOE 均值图上，横坐标表示各个因子，纵坐标表示针对每个因子的不同变量水平 的响应变量的均值(的连线)。 DOE 均值图可以用来回答如下几个问题：(1)哪个因子是 重要的?虽然DOE 均值图不能对该问题给出确定性的回答，但是它有助于把因子划分为 三类，分别是“明显重要”“明显不重要”“处于重要和不重要的临界线上”。(2)各个因 子重要性的排序是怎样的?

图10-46显示了7个因子和响应变量的DOE 均值图。从该图我们解读到各个因子的 重要性的排序情况，从最重要到不那么重要依次为， X₄,X₂,X₁,X₇,X₆,X₃,X;。

X₄不同的变量水平对应的因变量的均值变化范围大，表示 X₄ 具有对因变量的重要解释 作用。



Sensitivity

**图10-46** **DOE 均值图**

其中，X₄,X₂,X 似乎是“明显重要”(Clearly Important) 的因子，X₃ 和 X₅ 是“明显 不重要” (Clearly Unimportant) 的 因 子 ，X₆ 和 X₇ 是“处于重要和不重要的临界线上” (Borderline) 的因子，是否把它们纳入分析模型.有待于进一步的分析。

**10.11.4** **在探索式数据分析中对高维数据进行降维**

假设有一个数据集，包含数百个特征(变量),我们对数据所属的领域几乎没有什么 了解，这时候需要通过探索式数据分析识别数据中的隐藏模式，怎么办?降维是一种好办 法，降维是对数据进行约减的办法。降维以后的数据在2维和3维空间中进行可视化，可 以帮助我们掌握数据中体现的空间布局和时间变化趋势。

需要注意的是，不仅可以在数据分析结果的可视化中使用降维技术，也可以在探索式 数据分析过程中使用降维技术。

降维是把数据或特征的维数降低，一般分为线性降维和非线性降维。典型的线性降维 方法有：主成分分析 (Principal Components Analysis,PCA)、线性判别分析 (Linear Discriminant Analysis,LDA)、多维尺度分析 (Multidimensional Scaling,MDS) 等。典 型的非线性降维方法有：等距映射 (Isometric Mapping)、局部线性嵌入 (Locally Linear Embedding,LLE) 、 拉普拉斯特征映射 (Laplacian Eigenmap,LE) 、Sammon 映 射 (Sammon Mapping)、典型相关分析 (Canonical Correlation Analysis,CCA)、最大方差 展开 (Maximum Variance Unfolding,MVU)、随机邻域嵌入 (Stochastic Neighbor Em- bedding,SNE) 、t 分布随机邻域嵌入 (t-distributed Stochastic Neighbor Embedding,t-



SNE) 等。

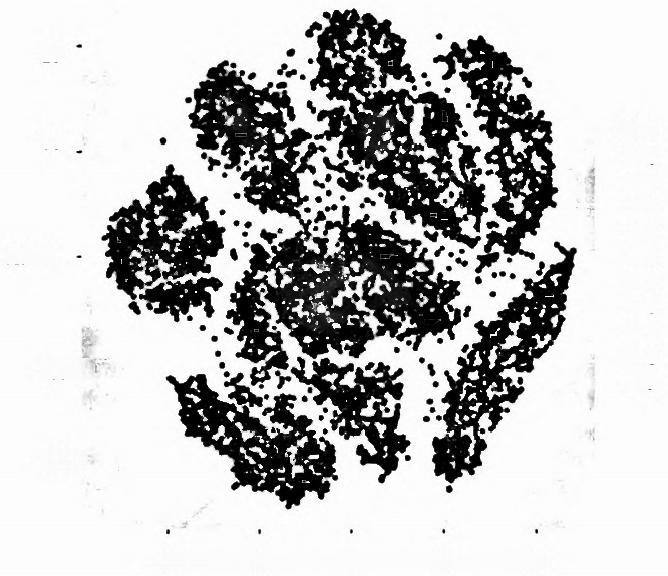
**1.t-SNE** **降维方法**

t-SNE 是一种非线性降维算法，它是对SNE 算法的改进。 SNE 算法的核心思想是高 维空间里数据点间的距离与映射到低维后数据点间的距离保持相似，靠近的数据点仍然靠 近，远离的数据点仍然远离。通过将数据降到2维或3维，我们就可以用一个散点图来表 示这些高维的数据，进而了解数据的结构。 t-SNE 与 SNE 的不同之处是它将低维空间中 数据点间的距离用t 分布来表示，这就使得高维空间中距离较近的点在低维空间中更近， 高维空间中距离较远的点在低维空间中更远。

MNIST 是一个手写体图片数据集，每个样本是一个图片，数据集里一共有10000个 样本，数据的第一列是分类标签(即到底是0～9中的哪个数字)。每个图片的大小是28× 28,展开后就一共有784个取值为0～254的像素特征。对该数据集运行t-SNE 算法后的 可视化结果如图10-47所示，可以看到，相同数字对应的样本都聚集在一起了，区分效 果非常好。

由此可见，t-SNE 可以用于寻找局部的近邻 (Local Neighborhood) 以及类簇 (Clus-

ter) 。 对数据的降维和可视化，帮助我们发掘数据里潜在的结构模式。



20

10

X20

-10

-20.

-20 -10 0 10 20

label ·0 ● 1

**·** **2 ●** **3 ●** **4** **·5** **·6** **·7** **·** **8**

·9

X1

**图10-47** **t-SNE降维效果**

资料来源： <http://blog.csdn.net/a358463121/article/details/55003356>. 读者还可以通过 <http://projector.tensorlow.> org/ 和 <http://colah.github.io/posts/2014-10-Visualizing-MNIST/> 两个网址查看和操控MNIST 数据集的降维效果。

**2.** **主成分分析** **(Primary** **Component** **Analysis,PCA)** **降维方法**

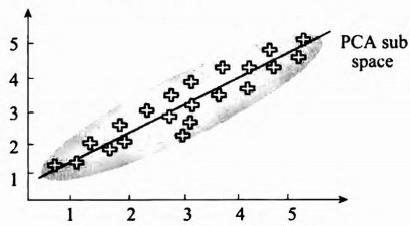
在一组样本 (Sample) 里，每个样本是一个多维向量，即每个样本包含多个特征 (Feature) 。 多维向量里的某些分量(特征)本身没有多大的区分度。比如，某个特征在

所有的样本中都取相同或者相近的值，那么这个特征就没有区分度。用这个特征来做分类 和回归分析，它的贡献就会很小。

我们希望找到变化很大的特征，也就是方差大的维度(特征),去掉那些变化不大的 特征，从而使得剩下的特征都是“精品”,对目标变量具有较强的解释作用。在分类和回 归中，需要的计算量也会相应变小。

比如，在图10-48中，样本分布在一个45度角倾斜的椭圆区域中，如果按照x,y

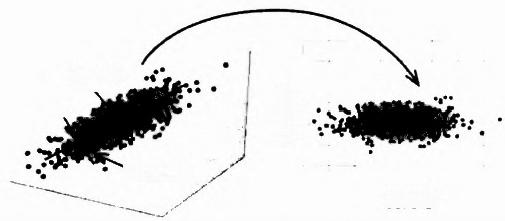
坐标来投影，单独使用这些点的x 或y 属性很难区分它们，因为它们在x,y 轴上坐标变 化的方差都比较大。我们可以根据这些点的x 属性和y 属性区分这些点，但是无法仅仅根 据这些点的x 属性或者y 属性来区分这些点。



**图10-48** **PCA 降维示例**

现在，将坐标轴旋转，以椭圆长轴为x 轴，椭圆短轴为y 轴。这些样本在长轴上的分 布比较宽，方差大，在短轴上的分布比较窄，方差小。我们可以只保留这些点的长轴属性 来区分椭圆区域上的点，在新的x 轴上的样本区分性比使用旧的x,y 轴好，实现了降维， 同时保留了特征的区分度。

我们再给出一个从3维到2维的PCA 降维的直观的实例。假设有一个3维数据集 (见图10-49的左侧子图),现在想要把它降维为一个2维的数据集(见图10-49的右侧 子图)。PCA 方法在原有的3维空间中找出主要的坐标轴 (Principal Axes), 在这些坐标 轴上，点和点之间的差别(方差， Variance) 最大。当找到能够解释绝大部分数据差异 (Most Variance) 的两个坐标轴(左侧子图中的黑线),就可以基于这两个新的坐标轴对 数据重新作图，如右侧子图所示。现在3维数据集变成了2维数据集，达到了降维的效 果。上述两个实例，给我们以直观的认识。需要注意的是， PCA 技术更多的是用于高维 空间数据集的降维。

1st PC

2nd PC

2nd PC

1st PC

**图** **1** **0** **-** **4** **9** **3D-2D PCA 降维的一个直观的实例**

资料来源：[http://blog.kaggle com/2017/04/10/exploring-the structure-of-high-dimensional-data-with-hypertools-in](http://blog.kagglecom/2017/04/10/exploring-thestructure-of-high-dimensional-data-with-hypertools-in)- kaggle-kernels/.



为了实现降维，需要求得一个n 维特征的投影矩阵，这个投影矩阵可以将特征向量 从高维降到低维，投影矩阵也称为变换矩阵。新的低维特征必须每个维都正交，特征向量 都是正交的。通过变化这些维的坐标系，从而使这个特征向量在某些维上方差大，在某些 维上方差很小。

本质上，主成分分析通过正交变换将原始的n 维数据集变换到一个新的被称做主成分 的数据集中。在该变换过程的结果中，第一个主成分具有最大的方差值.每个后续的成分 在与前述主成分正交条件限制下具有最大方差。降维时仅保存前n'(n'<n) 个主成分， 尽量保持最大的信息量。需要注意的是，主成分变换对正交向量是尺度敏感的，数据在变 换之前需要进行归一化处理。此外，新的主成分并不是由实际系统产生的.因此在进行 PCA 变换后，各个主成分一般就丧失了可解释性。

PCA 的过程包括如下几个主要步骤：(1)假设有m 个样本，每个样本都是一个 n 维 向量.也就是特征数量为n。(2)计算样本的协方差n×n 矩阵。(3)求出这个协方差矩阵 的特征值和特征向量。(4)根据特征值的大小，取前n 个特征值所对应的特征向量，构 成n×n 矩阵，这个矩阵就是我们要求的特征矩阵(投影矩阵)。(5)原有的m×n 样本矩 阵，乘以这个n×n′特征矩阵，得到新的m×n '样本矩阵，现在的维度为n', 达到降维的 效果。PCA 实际上是求得这个投影矩阵用高维的特征乘以这个投影矩阵就可以将高维特 征的维数下降到指定的维数。

协方差的计算公式为Cou(X.X₂)=E(X₁-E(X₁))(X₂—E(X)), 协方差表示特征X₁ 和特征X₂ 之间的相关性。比如，假设有3个样本.每个样本是4维的，如表10-2所示。

**表10-2** **包含3个样本的样本集(每个样本是一个4维向量)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 样本 | 属 性 X₁ | 属性X | 属 性 X₃ | 属性X |
| Samplel |  | 2 | 3 | 4 |
| Sample2 | 3 | 4 | [ | 2 |
| Sample3 | 2 | 3 | 1 | 4 |

那么,Cov(X₂.X₃)= (第2列一第2列均值)T× (第3列一第3列均值)/(样本数一 1)=(-1.1,0)×(1.33,一0.67,一0.67)/(3—1)=-1。

矩阵(n×n) 的 n 个特征向量就是n 个标准正交基，特征值的模(即绝对值大小)则代 表矩阵在每个基上的投影长度。特征值越大，说明矩阵在对应的特征向量上的方差越大， 信息量越多。

比如，一个数据集包含200个样本 (Sample), 特征是10维，构成一个200×10的样 本矩阵，求这个样本的协方差矩阵，得到一个10×10的协方差矩阵.然后求出这个协方 差矩阵的特征值和特征向量，总共有10个特征值和特征向量。根据特征值的大小，取前4 个特征值所对应的特征向量，构成一个10×4的矩阵，这个矩阵就是我们要求的特征矩 阵。200×10的样本矩阵乘以这个10×4的特征矩阵，就得到了一个200×4的新的降维之 后的样本矩阵，每个样本的维数降低了，也就是从10维降低为4维。当给定一个测试的 样本.是1×10的向量，乘以上面得到的10×4的特征矩阵，得到一个1×4的降维以后 的特征，可以用这个新特征去分类。

下面举一个PCA 实例。李毛侠对安徽省的消费水平进行了研究①,他把人均消费支出水



①<http:'/>[www.cnki.con.cn/Article/CJFDTotal-XDWX201002017.htm.](https://www.cnki.con.cn/Article/CJFDTotal-XDWX201002017.htm.)

平作为因变量Y, 自变量分别是人均生产总值 (X₁)、 城乡收入差距 (X₂)、 居民消费价格 指数 (X₃)、 就业率 (X₄)、 少年儿童抚养系数 (X₅)、 老年抚养系数 (X₆)、 住房制度改革 (X₇) 、 一年期存款利率 (X₈) 、 人均医疗支出 (X₉) 等。对影响消费需求的各因素X₁,X₂,

X₃,X₄,X₅,X₆,X₇,X₈,X9 进行主成分分析，分析结果如表10-3所示。

**表10-3** **主成分分析结果**

|  |  |  |  |
| --- | --- | --- | --- |
| 主成分 | 特征根 | 贡献率(%) | 累计贡献率(%) |
| 1 | 6.388 | 70.981 | 70.981 |
| 2 | 1.583 | 17.584 | 88.564 |
| 3 | 0.594 | 6.602 | 95.166 |
| 4 | 0.228 | 2.530 | 97.696 |
| 5 | 0.098 | 1.085 | 98.780 |
| 6 | 0.062 | 0.690 | 99.470 |
| 7 | 0.039 | 0.436 | 99.906 |
| 8 | 0.008 | 0.085 | 99.991 |
| 9 | 0.001 | 0.009 | 100.000 |

第一个主成分的特征根为6.388,解释了总变异的70.98%,第二主成分的特征根为 1.583,解释了总变异的17.58%。前两个特征根的累计贡献率达到88.56%,说明前两个 主成分已经反映原来九个指标88.56%的信息。前两个主成分的表达式分别如下：

F₁=0.37ZX₁+0.12ZX₂—0.19ZX₃+0.36ZX₄—0.38ZX5

+0.38ZX₆+0.35ZX₇—0.34ZXg+0.38ZX₉

F₂=0.16ZX₁+0.62ZX₂+0.62ZX₃+0.05ZX₄—0.10ZX₅+0.15ZX₆

—0.25ZX,+0.28ZX₈+0.13ZX₉

式 中 ，ZX,(i=1,2,3,4,5,6,7,8,9) 为 X 相应的标准化变量。

第一个主成分主要包含除变量X₂,X₃ 以外所有其他变量的信息，第二个主成分则主 要包含变量X₂,X₃ 的信息。

我们可以选择这两个主成分建立回归模型。用ZY 表示标准化后的因变量，以ZY 为 因 变 量 ，F₁,F₂ 为自变量进行多元线性回归分析，回归模型为ZY=0.382F₁+0.102F₂。

通过对新模型的判定系数计算.判断模型的拟合优度，以及通过 F 检验和t 检验，判断模 型设定是否有意义.每个自变量(每个主成分)是否需要纳入模型等。

PCA 算法的主要优点有：(1) PCA 以方差衡量信息量，不受数据集意外因素的影响；

(2)各个主成分之间是正交的，可以消除原始数据的各个成分间的相互影响；(3) PCA 的计算方法简单，主要运算是特征值求解，容易实现。

PCA 算法也有其固有的局限性：(1)各个主成分的含义解释性不强，不像原始样本 特征的解释性那么强；(2)方差小的非主成分.也可能包含对样本差异具有决定作用的重 要信息，降维有可能对后续数据处理产生影响。

正如上文所述，标准的PCA 算法，其各个成分的可解释性较差。人们研究了PCA 的 一些变种，以提高各个成分的可解释性，比如方差最大化旋转 PCA(PCA with Varimax Rotation) 等。方差最大化旋转是在主成分分析或因子分析中使用的一种方法，通过坐标 变换使各个因子载荷的方差之和最大。通俗地讲，就是：(1)要么任何一个变量只在一个 因子上有高贡献率，在其他因子上的载荷几乎为0;(2)要么任何一个因子只在少数变量 上有较高载荷，在其他变量上的载荷几乎为0。满足这个条件的因子载荷矩阵，具有简单

的结构。方差最大化旋转对载荷矩阵进行旋转，使之尽量接近简单结构。对于一组样本来 说，方差最大化旋转找到了一种表示主成分的最简单的方法，即每个主成分可以用少数变 量的函数的线性组合表示。方差最大化旋转的具体细节，请参考相关资料。

下面我们展示一个通过方差最大化旋转 PCA, 进行数据降维以后保持了因子的可解 释性的例子。①对关于学习韩语的内在驱动力 (Motivation Item) 的描述数据，进行方差 最大化旋转 PCA 处理以后.获得表10-4。表格的各行表示不同的人学习韩语的内在驱动 力的描述，表格的各列则表示进行 PCA 分析之后得到的各个主要成分。通过对各列较大 的系数对应的行进行分析，可以给各主要成分一个标签，分别是 “School” “Career” “Personal Fullfillment”“Ethnic Heritage” 等(每一行的一个语句可以看作一个文档，可 以用向量空间模型，表示成一个向量)。

**表10-4** **学习韩语的内在驱动力的描述数据的方差最大化旋转PCA结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Factor 1  School  Related | Factor 2  Career  Related | Factor 3  Personal  Fulfillment | Factor 4 Ethnic  Heritage |
| I learn Korean to transfer credits to college | 0.83 | 0.01 | 0.17 | 一0.11 |
| I learn Korean because my friend recommended it | 0.82 | 0.14 | 0.06 | 0.22 |
| I learn Korean because my advisor recommended it | 0.80 | 0.10 | 0.36 | 0.04 |
| I learn Korean because of the reputation of the pro- gram and instructor | 0.77 | 0.12 | 0.12 | 0.22 |
| I learn Korean for an easy A | 0.69 | 0.18 | 一0.29 | 0.16 |
| I learn Korean to fulill a graduation requirement | 0.63 | 0.11 | 0.19 | 一0.18 |
| I learn Korean to get a better job | 0.04 | 0.80 | 0.20 | 0.11 |
| I learn Korean because I plan to work overseas | 0.26 | 0.80 | 0.11 | 0.10 |
| I learn Korean because of the status of Korean in the world | 0.10 | 0.73 | 0.20 | 0.22 |
| I learn Korean to use it for my research | 0.38 | 0.48 | 0.44 | 0.10 |
| I learn Korean to further my global understanding | 0.16 | 0.33 | 0.71 | 0.08 |
| I learn Korean because I have an interest in Korean literature | 0.18 | 0.04 | 0.64 | 0.13 |
| I learn Korean because it is fun and challenging | 0.04 | 0.04 | 0.63 | 0.46 |
| I learn Korean because I have a general interest in languages | 0.11 | 0.03 | 0.57 | 0.51 |
| I learn Korean because it is the language of my family heritage | 一0.01 | 0.26 | 0.05 | 0.80 |
| I learn Korean because of my acquaintances with Ko- rean speakers | 0.10 | 0.21 | 0.20 | 0.70 |
| %of variance explained by each factor | 0.23 | 0.15 | 0.14 | 0.12 |



①<http://jalt.org/test/PDF/Brown31.pdf.>



**10.11.5** **探索式数据分析案例**

下面通过建立垃圾邮件分类器的过程以及雾霾成因分析，考察探索式数据分析的 作 用 。

**1.** **通过探索式数据分析建立垃圾邮件分类器**

大量的垃圾邮件充斥着商业广告和色情信息，干扰了用户的正常工作和生活，给用户 带来很多不便。我们通过查看邮件的发送者邮箱、邮件标题等，根据直觉和经验，可以判 断哪些是垃圾邮件，有时候还需要仔细查看邮件内容才能判断垃圾邮件。人工判断和清理 是非常耗时的。建立垃圾邮件分类器，自动对邮件进行分类势在必行。

建立垃圾邮件分类器的第一步是从大量邮件中随机抽样出少量的邮件，比如100份， 然后人工标注哪些是正常邮件，哪些是垃圾邮件。

第二步，可以使用探索式数据分析，对筛选出的垃圾邮件进行分析，统计出哪类词汇 出现的频率较高，包括各类促销、诱惑的词汇等。根据这些词汇出现的频率，可以选出经 常出现的5～10个词汇。

第三步，利用选出的词汇，开始建立第一版邮件分类器，并且开发完整的邮件过滤软 件。然后使用这个软件，对一个较大的样本(包含1000份邮件)进行垃圾邮件的过滤。

第四步，对过滤器筛选出的垃圾邮件进行人工验证。然后用探索式数据分析计算过滤 的成功率以及正常邮件Top 10词汇及其频率，垃圾邮件的Top 10 词汇及其频率等。

第五步，根据过滤的成功率以及正常邮件Top 10 词汇及其频率，垃圾邮件的Top 10 词汇及其频率，改进过滤模型。比如，调整事先设定好的阈值 (Threshold), 增加或者减 少作为机器学习模型输入数据的词汇数量等。通过优化进一步提高垃圾邮件过滤器的成 功 率 。

在探索式数据分析过程的每次迭代中，都可以利用数据可视化技术，展示探索式数据 分析的结果，为邮件过滤器的开发随时提供参考，帮助对模型进行修正。

**2.** **雾霾成因的探索式分析**

某研究机构收集了大量关于雾霾的数据(包括企业排放数据、城市机动车辆数据、天 气数据、地理信息数据等),分析人员想利用这些数据分析雾霾的成因。看起来好像是诊 断性分析，但是由于分析人员对雾霾的成因还缺少深刻的认识，只能通过很多尝试性的分 析来了解雾霾和各种因素的关系。

在对比了雾霾天气和气温的关系后，发现冬天雾霾更多。进一步分析发现.气温和雾 霾存在相关性，但并不是决定性的，因为可以从不同季节、不同年份的分析结果否定气温 是主因的结论。分析人员还可以对不同地域、不同时段、不同类型数据进行进一步细致的 分析。这种探索过程能够不断帮助分析人员发现一些重要的规律，形成一些假设。分析人 员可以从地域、时间等维度分析燃煤情况和雾霾的关系，通过不断的探索，验证燃煤是不 是雾霾的一个可能的主因。当然，还可以对汽车保有量、排放量和雾霾天气的变化进行 探 索 。

在这个例子中，分析模型在数据分析开始阶段不是很明确。分析人员在探索过程中， 了解不同因素和雾霾的关系后，才能逐步建立起对各个因子和雾霾关系的理解，从而建立 适当的模型。在探索式分析过程中，分析人员根据对数据的已有理解调整探索的方向。

\

**10.12** **可视化工具介绍**

**10.12.1 D3.js**

D3 是一个开源项目，其作者是《纽约时报》的工程师。目前， D3 项目的代码托管于 GitHub(<https://github.com/d3/d3/wiki>)。GitHub 是一个全世界最流行的代码托管平 台，云集了来自世界各地的优秀的工程师和开源项目。

D3 的全称是Data Driven Documents, 顾名思义，它是数据驱动的文档。 D3 是 一 个 JavaScript 函数库，结合HTML,CSS 和 SVG 等技术，在网页上实现数据的可视化，并 且提供交互的能力。 D3 提供各种简单易用的函数，简化了JavaScript操作数据的难度。

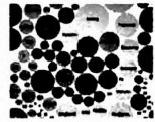
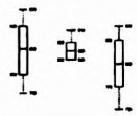
D3 将可视化的复杂步骤，包括数据的准备、坐标 (Axis) 、 形 状 (Shape) 、 尺 度 (Scale) 的控制等，精简到了若干简单的函数。用户只需输入必要的数据，就能够渲染出 各种炫丽的图形，大大减少用户的工作量。目前，D3.js 已经成为基于Web 的可视化的最 流行的开源软件库之一 。图10-50展示了部分D3.js 的可视化实例。①我们可以根据实际 应用的需要，参考这个实例网页上的代码，实现数据的可视化。

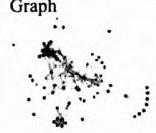
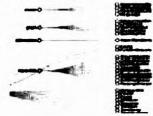
Calendar View

Bullle Charts

Bubble Chart

Box Plots



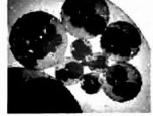
Non-contiguous Cartogram

Dendrogram

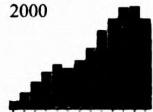
Force-Directed

Chord Diagram

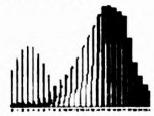


Circle Packing

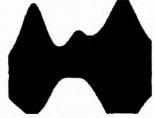
Population Pyramid



Stacked Bars



Streamgraph





**图10-50** **D3.js的可视化实例**



①<https://github.com/d3/d3/wiki/Gallery.>

**10.12.2 Processing.js**

Processing.js 是可视化编程语言项目Processing① 的姐妹项目。

Processing 项目由Ben Fry 和Casey Reas 两位作者发起，他们基于Java 研发了开源的 编程语言 Processing, 帮助人们进行数据的可视化 (Data Visualization)、创建电子艺术作 品 (Electronic Art)、可交互的动画 (Interactive Animation), 甚至视频游戏 (Video Game)。

Processing.js 把 Processing 的能力扩展到了网页上，它为实现基于Web 的可视化而 设计。用户使用Processing 语言编写可视化程序，然后包含在网页中，Processing.js 则负 责完成渲染和交互。也就是通过 Processing.js, 任何 HTML5 兼容的浏览器，包括Fire- fox.Safari,Chrome,Opera 以及 Internet Explorer等，都可以运行Processing 语言编写

的代码。

Processing.js 为 Processing 代码的开发者以及网页开发者提供了在网页上方便地进行 可视化编程 (Visual Programming) 的手段。读者可以访问 Processing.js 项目主页②查看 Processing.js 丰富的实例。

**10.12.3 Protovis**

Protovis③ 是一款基于BSD License的免费开源软件。由斯坦福大学可视化研究组

(Stanford Visualization Group) 的 Mike Bostock和Jeff Heer共同开发。它使用JavaScript 和 SVG 技术实现基于Web 的可视化。基于Protovis 函数库编写的可视化代码，可以直接 在浏览器里运行，渲染可视化效果，无须任何插件。 <http://mbostock.github.io/proto>-

vis/ex/上展示了60多个精彩的实例，用户可以参考这些实例，学习如何使用Protovis实 现数据的可视化。

Protovis 的最终版本是V3.3.1, 开发团队此后不再对这个项目进行主动的更新了。 Protovis 开发团队的精力转移到借助 Protovis 的经验，为 D3.js 提供更加强大的能力上， 包括动画和交互 (Animation and Interaction) 。Mike Bostock 也是D3.js 的作者之一。

**10.12.4 Prefuse**

Prefuse(<http://prefuse.org/>) 是一款使用Java 语言开发的开源 (BSD license) 软 件，用于开发交互式的数据可视化程序。它为 Java 语言提供了可视化的开发框架 (Framework) 。Prefuse 本身是用Java 语言编写的.使用了Java 2D图形库，开发者可以 基于Prefuse, 编写独立的应用程序，或者在大型应用程序中实现可视化组件，也可以开 发Java Applet, 用于在Web 网页上进行数据可视化。

Prefuse 支持各种结构化与非结构化数据的可视化，包括具有层次性结构的数据、具



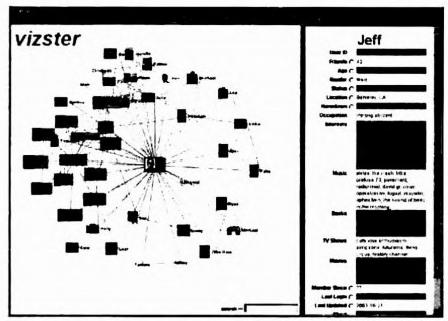
①<https://processing.org.>

②<http://processingjs.org/exhibition/.>

③<http://mbostock.github.io/protovis/.>

有网络结构的数据以及时间序列数据等。它提供一系列数据可视化所需的功能，包括数据 建模 (Data Modeling)、可视化 (Visualization) 以及用户交互 (Interaction) 等。在数据 建模方面，Prefuse 支持表格 (Table) 、 图 (Graph) 、 树 (Tree) 等数据结构。在用户交 互方面，支持动画效果 (Animation) 、 查询和搜索 (Query and Search) 等功能。此外， Prefuse 还可以从数据库里提取数据，进行实时的可视化。

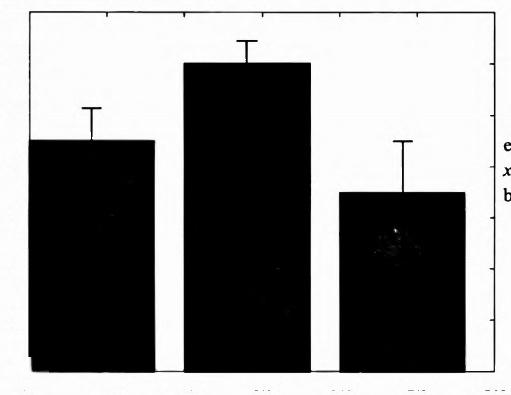
Prefuse 开发组对Prefuse 的功能进行了移植，提供了Prefuse Flare 工具包，它是一 个 ActionScript 库，用于创建可以通过Adobe Flash Player渲染的可视化效果。由于几乎 所有浏览器都支持Adobe Flash Player, 所以使用Prefuse Flare开发的可视化组件可以很 方便地集成到网页中。图10-51展示了基于Prefuse 实现的社交网络可视化效果。



**图10-51** **基于Prefuse的社交网络可视化**

**10.12.5 Matplotlib**

Matplotlib 是基于Python 语言的图形绘制库 (plotting library)。使用Matplotlib, 用 户编写少量代码就可以创建各种常用的图形，比如直方图、散点图、饼图、柱状图。这些 图形达到印刷的品质，用户可以导出成PDF 文档或者PNG 图像文件。图10-52展示了



14-

12

10

x=arange(len(y))

bar(x,y,yerr=errors,ecolor="red",capsize=10)

6

4

2

0.5 1.0 1.5 2.0 2.5 3.0

y=[9,12,7]

errors=[1.3,0.9,2]

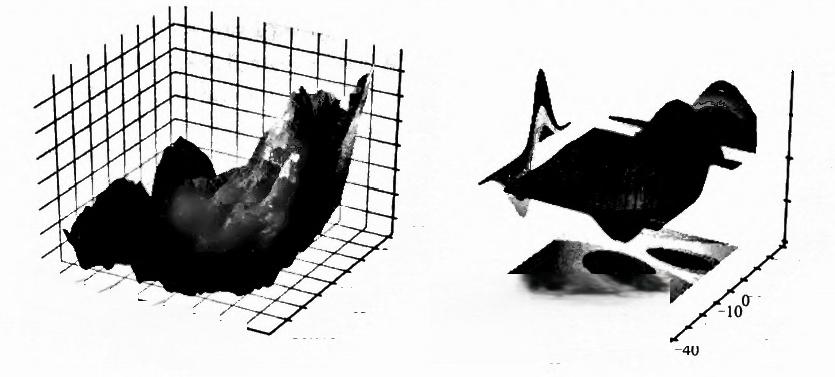
0.0

8

**图10** **-** **52** **Matplotlib实例(柱状图及其脚本)**

简单的柱状图及其Python 脚本，可以看到通过调用Matplotlib, 少量 Python 代码就可以 生成复杂的可视化效果。开发者可以通过面向对象的编程接口，对线型、字体、坐标、颜 色等图形属性进行精细控制，绘制出符合特定需要的图形。

Matplotlib网站上①提供了上百个利用Matplotlib实现的数据可视化实例。Matplotlib支 持2维和3维图形的绘制。图10-53展示了使用Matplotlib 绘制的3维图形的两个实例。



-84.390

700

650

600

550

0

500

450

400

36.730

36.725

363.170

36.710

36.705

36.700

36.695 -84,375

100

50

Z

-50

-100

3040

20

10

-40-30-20-10010203040 X

-20

Y

-84.385

30

(a) (b)

**图10-53** **Matplotlib的3维可视化实例**

资料来源：(a)<http://matplotlib.org/examples/mplot3d/custom_shaded_3d_surface.html.>

(b)<http://matplotlib.org/examples/mplot3d/contourf3d_demo2.html.>

用 于 Python 编程语言的可视化软件包，除了 Matplotlib, 还 有 geoplotlib② 和 ggplot③ 等。

除此之外，商用的可视化软件包，比如 IBM Many Eyes,Tableau等，提供更强大的 可视化和交互功能，厂家可以提供周到细致的培训和服务。

**10.13** **思考题**

(1)可视化的定义。

(2)可视化的一般过程。

(3)科学可视化与信息可视化。

(4)可视化的若干原则。

(5)可视化实例——层次关系、地图、社交网络、堆叠的河流。



①<http://matplotlib.org/gallery.html.>

②<https://github.com/andrea-cuttone/geoplotlib.>

③[http://ggplot.yhathq com/.](http://ggplot.yhathqcom/.)



**(6)高维数据可视化。**

**(7)可视分析。**

**(8)可视化的挑战和趋势。**

**(9** **)探索式数据分析。**

**(10)可视化工具。**