



“十三五”规划教材
13th Five-Year Plan Textbook



“十三五”规划全媒体人才培养丛书·数据科学系列

数据可视化原理与实例

DATA VISUALIZATION THEORY
AND EXAMPLES

李春芳 石民勇 著

中国传媒大学出版社

非
外
借



“十三五”规划教材

13th Five-Year Plan Textbook



“十三五”规划全媒体人才培养丛书·数据科学系列

数据可视化原理与实例

DATA VISUALIZATION THEORY
AND EXAMPLES

李春芳 石民勇 著

中国传媒大学出版社

·北京·

图书在版编目(CIP)数据

数据可视化原理与实例 / 李春芳, 石民勇著. —北京: 中国传媒大学出版社, 2018.7

(“十三五”规划全媒体人才培养丛书·数据科学系列)

ISBN 978-7-5657-2354-4

I. ①数… II. ①李… ②石… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2018) 第 147447 号

数据可视化原理与实例

SHUJU KESHIHUA YUANLI YU SHILI

著 者 李春芳 石民勇

策划编辑 阳金洲

责任编辑 黄松毅

特约编辑 李克俭

责任印制 日新

封面设计 风得信设计·阿东

出版发行 中国传媒大学出版社

社 址 北京市朝阳区定福庄东街1号 邮编: 100024

电 话 86-10-65450528 65450532 传真: 65779405

网 址 <http://www.cucp.com.cn>

经 销 全国新华书店

印 刷 北京中科印刷有限公司

开 本 787mm×1092mm 1/16

印 张 16

字 数 304千字

版 次 2018年9月第1版 2018年9月第1次印刷

书 号 ISBN 978-7-5657-2354-4/TP·2354 定 价 69.00元

“十三五”规划全媒体人才培养丛书编辑委员会 >>>

编委会主任 胡正荣

编委会成员 (按姓氏笔画)

丁俊杰 王永滨 王 宇 乔保平 刘大年 刘剑波
许一新 李佐文 李怀亮 张鸿声 张毓强 范 周
段 鹏 高晓虹 隋 岩 鲁景超

“十三五”规划全媒体人才培养丛书·数据科学系列 >>>

主 编 石民勇

前言

2015年9月5日国务院发布了《促进大数据发展行动纲要》(以下简称《纲要》),意味着中国大数据发展迎来顶层设计,正式上升为国家战略。《纲要》在政策机制中指出,鼓励高校设立数据科学和数据工程相关专业,重点培养专业化数据工程师等大数据专业人才。中国传媒大学计算机学院在石民勇院长筹划下,2014年开始招收“计算机科学与技术(大数据技术与应用)(080901)”专业本科生,调整原来的计算机科学与技术本科专业方向,着重培养大数据技术与应用方向人才,目前已经招生4年,在校本科生共计134人,2018年,招生方向与教育部专业目录一致,改为“数据科学与大数据技术(080910T)”本科专业。中传的大数据招生可以说走在了全国前列,体现了中传在新专业发展上的预见性,在国内算是较早开办大数据本科专业的高校。

大数据时代,海量数据的理解成为极富挑战性的难题。数据可视化是大数据技术体系的重要内容之一,可视化在压缩数据、解释问题、发现知识、归纳结论中发挥着重要作用。

本书是大数据专业的必修课——数据可视化及暑期实践课使用的教材,作者本着理论与实践结合,以提高软件开发工程能力为目标,立足可视化并兼顾外围数据采集处理问题。

本书适用于本科或研究生的数据可视化教学、软件开发中数据可视化应用以及启发开发新的数据可视化算法。第一、二章是Web可视化基础,第三至八章是D3的API各种布局与数据处理,包括饼图类、比例尺、动画与交互、力导向图、地图,第九章是音乐可视化,第十章是JavaScript图像处理,第十一章是可视化数据采集,第十二章是包括分词的词云图绘制,第十三章的主要内容——一种基于占用矩阵的词

云图可视化算法——是第一作者在博后期间和导师刘连忠教授合作的一个专利，非常感谢刘老师同意将算法收入书中。其中第十章对于理解流行的深度学习之卷积神经网络小有帮助，第十三章主要是启发开发新的数据可视化算法，内容稍有难度，可以选学。

感谢“媒体大数据处理与应用关键技术研究”项目资助本书出版，并资助数据可视分析应用于中传如艺剧本系统，如果您对影视行业感兴趣，欢迎访问<http://www.ruyi.cool>。

鉴于作者水平所限，更兼时间和精力不足，书中错谬之处在所难免，若蒙读者老师不吝指正，不胜感激。

李春芳

石民勇

2018年1月16日

全书配套代码链接：<http://cuc.yingshinet.com>



目 录

第一章 基于Web的数据可视化基础

第一节	数据可视化概述	001
第二节	HTML文档	003
第三节	层叠样式表CSS	006
第四节	JavaScript脚本	012
第五节	SVG矢量图	022

第二章 基于JavaScript和SVG的绘图

第一节	直方图	026
第二节	递归二叉树	034
第三节	树状词云	040

第三章 D3 数据可视化基础

第一节	D3 可视化API 概述	049
第二节	D3 数据绑定	052
第三节	D3 直方图	057
第四节	JSON文件格式	060

第四章 D3 饼图、环图、玫瑰图和弦图

第一节	D3 绘制饼图	064
第二节	从原生数据到绘制D3 饼图	068
第三节	弦图	077

第五章 D3 比例尺

- 第一节 D3 比例尺 090
- 第二节 比例尺 API 093
- 第三节 序数比例尺与 D3 配色 096

第六章 D3 动画与交互

- 第一节 D3 动画 099
- 第二节 鼠标交互 102
- 第三节 键盘交互 109

第七章 D3 力导向图

- 第一节 D3 力导向算法 113
- 第二节 《人民的名义》人物关系图 118
- 第三节 基于路径绘制力导向图 127

第八章 地图可视化

- 第一节 D3 地图可视化 134
- 第二节 D3 地图可视化实例 138
- 第三节 雄安新区区位优势可视化分析案例 141

第九章 音乐可视化

- 第一节 音乐可视化 API 150
- 第二节 JavaScript 音乐可视化直方图 152
- 第三节 D3 音乐可视化南丁格尔图 154

第十章 Canvas 与简单图像处理

- 第一节 Canvas 功能概述 158
- 第二节 Canvas 基本图像处理 161
- 第三节 Canvas 复杂图像处理 166

第十一章 数据可视化之数据采集

第一节 Python爬虫基础	174
第二节 数据解析	180
第三节 数据存储与知识图谱可视化	186

第十二章 分词与词云图可视化

第一节 词云图	196
第二节 D3 词云图	199
第三节 基于Python的分词实例	205
第四节 影视剧小说词云图案例分析	211
第五节 三维动态词云	216

第十三章 基于占用矩阵的词云算法

第一节 算法设计概述	220
第二节 基于占用矩阵的词云图可视化系列算法	225
第三节 基于占用矩形的词云图可视化算法	240
第四节 算法性能分析	243

参考文献	246
------	-----

后 记	247
-----	-----

第一章

基于 Web 的数据可视化基础



基于 Web 的数据可视化开发，需要网页程序设计的基础知识，本章在概述 HTML、CSS、JavaScript 和 SVG 基础上，快速进入上述脚本的互操作。

第一节 数据可视化概述

一、数据可视化是大数据技术体系的重要组成

中国计算机学会 CCF 大数据专委会首次发布《中国大数据技术与产业发展报告（2013）》^[1]，提出大数据的技术体系通常分为大数据采集与预处理、大数据存储与管理、大数据计算模式与系统、大数据分析与挖掘、大数据可视化计算以及大数据隐私与安全等方面。2016 年 12 月，该专委会发布的《2017 年大数据发展趋势预测》中预测的趋势是：可视化技术和工具提升大数据分析工具的易用性，并指出可视化连续多年成为十大发展趋势预测的选项，2016 年曾占据榜首，2017 年投票关注度有所下降，但还是占据了十大趋势的最后一席^[2]。2017 年 12 月发布的《2018 年大数据发展趋势预测》^[3]中，第十项基于知识图谱的大数据应用成为热门应用场景，此项与数据可视化中力导向图布局密切相关。数据可视化是大数据技术体系六大模块之一，在大数据分析挖掘中占有举足轻重的地位，可视化使晦涩的数据成为一种人人易懂的显学。

数据的采集、提取和理解是人类感知和认识世界的基本途径之一，数据可视化作为人类洞察数据的内涵、理解数据蕴含的规律提供了重要手段。从宏观角度看，可视化包括三个功能：信息记录、支持对信息的推理和分析、信息传播和协同^[4]。从奇数和的可视化、勾股定理的图形化证明到著名的“鬼图”曾帮助发现霍乱流行原因、蛋白质折叠游戏、大国高层政治人物社交网络和揭示批量文本主题的词云图，大量实例表明可视化在知识发现、分析、理解和传播中扮演着重要作用。

二、数据可视化简史

可视化的全称“科学计算可视化”(Visualization in Scientific Computing, ViSC)是1987年美国国家科学基金会的“科学计算可视化研讨会”报告中正式提出的概念^[4]。大数据浪潮下,海量数据的分析、挖掘、传播和理解成为信息技术发展的巨大挑战。可视化在大数据挖掘中的作用体现在多个方面,如揭示想法和关系、形成论点或意见、观察事物演化的趋势、总结或积聚数据、存档和汇整、寻求真相和真理、传播知识和探索性数据分析等。

可视化是认知的过程,即形成某个物体的感知图像,强化认知理解,其终极目的是对事物规律的洞悉。人脑50%的信息通过视觉感知。人眼是一个高带宽的巨量视觉信号输入并行处理器,最高带宽为每秒100MB,具有很强的模式识别能力。因此,领域学者将可视化简明地定义为“通过可视表达增强人们完成某种任务的效率”,这种任务包括:发现、决策、解释、分析、探索和学习^[4]。

信息科学领域面临的一个巨大挑战是数据爆炸。在信息管理、信息系统和知识管理学科中,最基本的模型是“数据、信息、知识、智慧”(Data, Information, Knowledge, Wisdom, DIKW)的诺兰模型。从数据到智慧,可视化在每一层递进转化中都起着重要的作用。

三、本书重点

可视化从传统的直方图、饼图、雷达图、箱图,发展到玫瑰图、关系图、词云图、力导向和弦图等多种方式,新的可视化算法、形式和工具仍不断出现。现实世界复杂系统的复杂性根源是多维度和多粒度的关联性。另一方面,概率、密度和频率构成纵深复杂性,这些概念也是统计学认识世界的重要方法,基于词频的词云图开创了文本数据可视化的新局面,提供了海量文本的探索性数据分析方法。然而以表达关系的力导向图和表达频次特征的词云图为代表的较复杂的可视化算法在实际的Web软件开发中有广泛需求,这些算法都有可用的API,并且基于最简单的HTML、CSS、JavaScript、SVG。

数据新闻专家沈浩认为数据可视化是一种数据分析、叙事手段和批判思维^[5-6]。在有内容表达的前提下,数据可视化的形式大于内容。本书着重围绕D3.JS讨论数据可视化的形式如何实现。从学术角度看,D3.JS中比较有算法特点的可视化图是力导向和词云图,用途也非常广泛。力导向有效表达了关系和联系,在社交网络、知识图谱中广泛应用。词云图成为大数据的一个可视化代名词,本书第十三章是作者的一个词云图专利,着重论述了一种基于占用矩阵的词云图布局算法的原理。本书在论述基于D3的可视化原理上,着重对力导向和词云图做了比较深入的介绍,配书代码见cuc.yingshinet.com。

第二节 HTML 文档

HTML (Hyper Text Markup Language) 是用来描述网页的一种语言, 即超文本标记语言, 它不是一种编程语言, 而是一种标记语言, 使用标签来描述网页^[7]。

HTML 标签由尖括号包围的关键词构成, 比如<html>, 标签通常是成对出现的, 比如表示一个段落的<p>和</p>, 标签对中的第一个是开始标签, 第二个是结束标签。

HTML 文档实际上等同于网页, 它描述网页, 包含 HTML 标签和纯文本, Web 浏览器的作用是读取 HTML 文档, 并以网页的形式显示。浏览器不显示 HTML 标签, 而是使用标签来解释页面的内容。

网页文件可以在任何文本编辑器中编写, 如记事本、写字板, 然而作为一个程序员, 推荐使用稍微高级的文本编辑器, 如 NodePad++、UltraEdit 等。如本书使用 NodePad++, 一个类似于各种语言的 HelloWorld 程序的 HTML 网页见代码 CH1/Hello.htm。

程序编号: CH1/Hello.htm

```
<html>
  <head>
    <title>
      苏轼诗词精选
    </title>
  </head>
  <body>
    <h1>定风波·三月七日</h1>
    <h3>【宋】苏轼</h3>
    <p>
      <h5>三月七日, 沙湖道中遇雨。雨具先去, 同行皆狼狈, 余独不觉。
        已而遂晴, 故作此。</h5><br>
        莫听穿林打叶声, 何妨吟啸且徐行。<br>
        竹杖芒鞋轻胜马, 谁怕? 一蓑烟雨任平生。<br>
        料峭春风吹酒醒, 微冷, 山头斜照却相迎。<br>
        回首向来萧瑟处, 归去, 也无风雨也无晴。<br>
      </p>
    </body>
</html>
```



代码中<html>与</html>之间的文本描述这个网页, <body>与</body>之间的文本是可见的页面内容, <h1>与</h1>之间的文本被显示为标题, <p>与</p>之间的文本被显示为段落。
表示换行。该页面的运行结果见图 1-1。

127.0.0.1:8080/DVIZ2018/CH1/Hello.htm

定风波·三月七日

【宋】苏轼

三月七日，沙湖道中遇雨。雨具先去，同行皆狼狈，余独不觉。已而遂晴，故作此。

莫听穿林打叶声，何妨吟啸且徐行。
竹杖芒鞋轻胜马，谁怕？一蓑烟雨任平生。
料峭春风吹酒醒，微冷，山头斜照却相迎。
回首向来萧瑟处，归去，也无风雨也无晴。

图 1-1 纯 HTML 的示例网页

更多的 HTML 标记可参见 <http://www.w3school.com.cn/html/index.asp>^[7] 上的文档，该网站涵盖的文档内容如图 1-2 所示，包括 HTML、CSS、JavaScript、XML、SVG、JSON、JQuery 以及服务器端的 SQL、ASP、PHP 等文档，几乎涵盖了 Web 数据可视化前端的基础文档，作为一个程序员要记得经常去看官方文档。

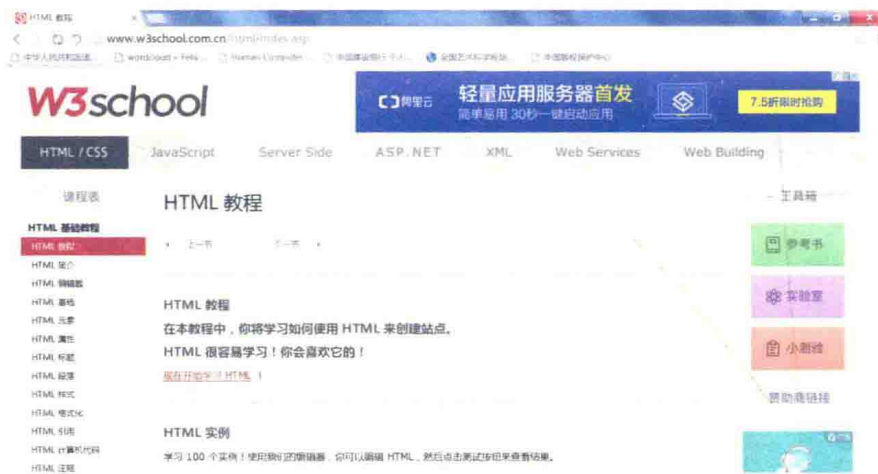


图 1-2 W3School.com.cn 上的 HTML 文档

从图 1-1 看，实际上这个网页文件是放在了 Tomcat 服务器下，本书的数据可视化大都放在 Tomcat 下，关于 Tomcat 的下载安装，请参考它的官网 <http://tomcat.apache.org/>，当然安装 Tomcat 之前还要安装 Java JDK，本书使用 JDK7+Tomcat9。

提示

在 NodePad++ 中编写的 HTML 文件可能会出现乱码，如果遇到了，请在 NodePad++ 中，点击“编码”菜单，将代码格式修改为“转为 UTF-8 编码格式”，保存文件，并重新刷新页面即可。

瞭望

>>

提起 w3school.com.cn，不得不说 W3C (World Wide Web Consortium)，即万维网联盟，该组织作为欧洲核子研究机构的一个项目由 Tim Berners-Lee (蒂姆·博纳斯·李，见图 1-3) 倡导发展起来。1994 年 10 月，博纳斯·李在麻省理工学院创建了万维网联盟，他邀请微软、网景、苹果、IBM 等 155 家公司以及一些研究机构，协同研究 WWW 的技术标准化，为推动万维网走向世界做出了巨大贡献。博纳斯·李由此被称为万维网的发明人，并担任 W3C 的主任。2012 年伦敦奥运会开幕式邀请博纳斯·李出场，他现场在键盘上敲出“This is for everyone”，意喻万维网是送给世界上每一个人的礼物。2017 年，他因“发明万维网、第一个浏览器和使万维网得以扩展的基本协议和算法”而获得 2016 年度图灵奖。

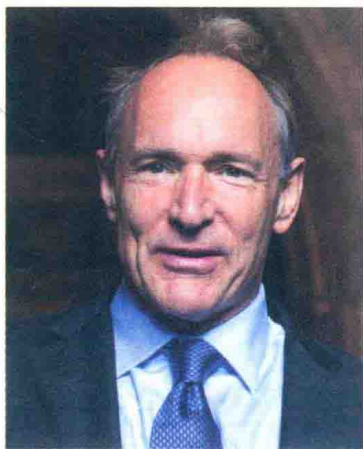


图 1-3 蒂姆·伯纳斯·李

W3C 在 1994 年被创建的目的是，为了完成麻省理工学院 (MIT) 与欧洲粒子物理研究所 (CERN) 之间的协同工作，并得到了美国国防部高级研究计划局 (DARPA) 和欧洲委员会 (European Commission) 的支持。W3C 由美国麻省理工学院计算机科学和人工智能实验室 (MIT CSAIL) 以及总部位于法国的欧洲信息数学研究联盟 (ERCIM) 和日本的庆应大学 (Keio University) 联合运作，并且在世界范围内拥有分支办事处。

万维网联盟中国办事处成立于 2006 年 4 月 1 日。现设址于北京航空航天大学，由北航计算机学院计算机新技术研究所负责日常的运营工作。W3C 中国办事处致力于促进国内外万维网标准领域信息沟通互动，为国内企业、高校及科研机构参与国际信息技术标准化的研究、集成和推广提供服务。

第三节 层叠样式表 CSS

为了解决内容与表现分离的问题，万维网联盟（W3C），在HTML 4.0之后定义了样式（Style），即CSS（Cascading Style Sheets）层叠样式表，以单独定义如何显示HTML元素，样式通常存储在样式表中，存储在CSS文件中外部样式表提高了网页开发的效率。

当同一个HTML元素被不止一个样式定义时，使用最近优先的样式。一般而言，样式由外到内的定义级别包括：

- a) 浏览器缺省设置；
- b) 外部样式表；
- c) 内部样式表（位于<head>标签内部）；
- d) 内联样式（在HTML元素内部）。

内联样式（在HTML元素内部）拥有最高的优先权，即它将优先于<head>标签中的样式声明、外部样式表中的样式声明以及浏览器中的样式声明（缺省值）。

一、内联样式

在Chrome浏览器中，默认的<h1>的字体是“黑体”，<p>默认的字体是“微软雅黑”，因此代码CH1/Hello.htm的运行结果是图1-1。如果使用了样式，见代码CH1/CSS.htm，根据最近样式优先原则，运行结果见图1-4，内联样式起作用，在<head>中的样式不起作用。

程序编号：CH1/CSS.htm

```
<html>
  <head>
    <title>
      苏轼诗词精选
    </title>
    <style type="text/css">
      body{background-color: #495A80}
      h1{color:#37C6C0}
      h3{color:#1DB0B8}
      h5{color:#00343F}
      p{font-family: 华文新魏;color:#011935;font-size:24}
    </style>
  </head>
  <body style="background-color:#C3BEDC">
```

```

<h1 style="color:#FFFEA0">定风波·三月七日</h1>
<h3 style="color:#FEE388">【宋】苏轼</h3>
<h5 style="color:#BA874C">三月七日，沙湖道中遇雨。雨具先去，同行皆狼狈，余独不觉。
    已而遂晴，故作此。</h5>
<p style="font-family: 华文仿宋;color:8F1D78;font-size:20px;">
莫听穿林打叶声，何妨吟啸且徐行。<br>
竹杖芒鞋轻胜马，谁怕？一蓑烟雨任平生。<br>
料峭春风吹酒醒，微冷，山头斜照却相迎。<br>
回首向来萧瑟处，归去，也无风雨也无晴。<br>
</p>
</body>
</html>

```



图 1-4 使用 CSS 的内联样式（在 HTML 元素内部）

图 1-4 内联样式使用的配色、内部样式配色和外部样式配色对比见图 1-5。



图 1-5 本例用于对比配色（内联样式、内部样式和外部样式）

二、内部样式

如果把内联的样式去掉，则 <head> 中的样式起到修饰作用，代码见 CH1/CSS-2.htm，运行结果见图 1-6，字体为“华文新魏”。



程序编号: CH1/CSS-2.htm

```
<html>
  <head>
    <title>
      苏轼诗词精选
    </title>
    <style type="text/css">
      body{background-color: #495A80}
      h1{color:#37C6C0}
      h3{color:#1DB0B8}
      h5{color:#00343F}
      p{font-family: 华文新魏;color:#011935;font-size:24}
    </style>
  </head>
  <body>
    <h1>定风波·三月七日</h1>
    <h3>【宋】苏轼</h3>
    <h5>三月七日，沙湖道中遇雨。雨具先去，同行皆狼狈，余独不觉。已而遂晴，故作此。</h5>
    <p>
      莫听穿林打叶声，何妨吟啸且徐行。<br>
      竹杖芒鞋轻胜马，谁怕？一蓑烟雨任平生。<br>
      料峭春风吹酒醒，微冷，山头斜照却相迎。<br>
      回首向来萧瑟处，归去，也无风雨也无晴。<br>
    </p>
  </body>
</html>
```

① 127.0.0.1:8080/DVIZ2018/CH1/CSS-2.htm

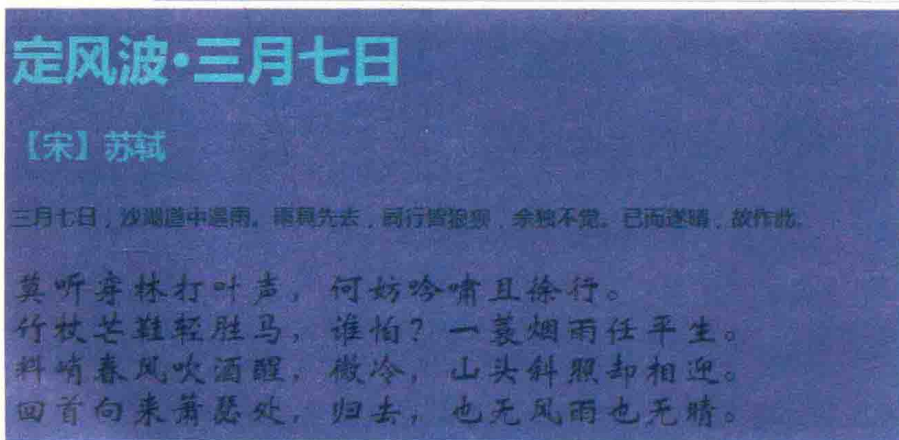


图 1-6 使用 CSS 的内部样式表（位于 <head> 标签内部）