数据可视化与领域应用案例

Data Visualization Design and Application

陈红倩 著

机械工业出版社
CHINA MACHINE PRESS

数据可视化与领域应用案例

陈红倩　著

机 械 工 业 出 版 社

本书针对不同领域的多种数据集，包括仿真领域数据集、影视领域数据集、服务器管理数据集、空气质量数据集、农药残留检测数据集等，通过分析数据中的分析需求，设计适用有效的可视化方法，并通过可视化方法和技术手段对数据进行分析，得出了相应的分析结论，所提出的可视化方法包括过程可视化、时序可视化、空间可视化、实时可视化、对比可视化、倾向性可视化和多关系可视化等。

本书可作为计算机、信息等相关专业的教师、研究生和大学高年级学生的参考书，也适合于从事数据可视化、大数据分析、数据挖掘、知识发现等方面的研究人员和技术人员阅读使用。

图书在版编目（CIP）数据

数据可视化与领域应用案例/陈红倩著. —北京：机械工业出版社，2019.5

ISBN 978-7-111-62537-7

Ⅰ.①数… Ⅱ.①陈… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2019）第 072557 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

策划编辑：吕 潇 责任编辑：吕 潇

责任校对：陈 越 封面设计：马精明

责任印制：张 博

三河市宏达印刷有限公司印刷

2019 年 6 月第 1 版第 1 次印刷

184mm×240mm · 11 印张 · 4 插页 · 195 千字

0001—2500 册

标准书号：ISBN 978-7-111-62537-7

定价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务

服务咨询热线：010-88361066

读者购书热线：010-68326294

网络服务

机 工 官 网：www.cmpbook.com

机 工 官 博：weibo.com/cmp1952

金 书 网：www.golden-book.com

教育服务网：www.cmpedu.com

封面无防伪标均为盗版

前　言

数据可视化（Data Visualization）是利用计算机图形学和图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。2013 年孟小峰教授[1]明确指出："可视化技术是数据分析与信息获取的重要手段。" 2013 年麦肯锡咨询报告[2]指出："可视化技术已经成为处理数据的关键技术。"

数据可视化以直观方式表达抽象信息，使得用户能够目睹、探索以至快速理解大量的信息，能有效吸引人们的注意力，其已经被证明为一种提高信息获取能力的有效方法，并在实践中得到了广泛的应用。数据可视化的发展，让数据的呈现更及时、更直观、更简单。

随着数据容量和复杂性的与日俱增，大数据可视化的需求越来越大，成为人类对信息的一种新的阅读和理解方式。通过大数据可视化手段进行数据分析，可以实现从错综复杂的数据中挖掘信息，再通过可视化的方式展示出来，使读者对数据的空间分布模式、趋势、相关性和统计信息一目了然。

本书以作者近几年来的研究工作为基础，阐述了作者本人及研究团队在可视化方面的技术研究工作，并结合这些可视化技术，针对不同领域的数据特征和分析目的，提出了多种可视化方法，数据可视化方法涉及层次数据可视化、空间数据可视化、时序数据可视化、多维数据可视化，以及多关系数据可视化方法。

本书的内容共包含 9 章，各章内容描述如下：

第 1 章：总结可视化相关技术基础，包括根据数据类型不同而分类的可视化技术，如层次数据可视化、多维数据可视化、时序数据可视化、地理数据可视化；然后针对基于可视化的数据分析——可视分析技术进行总结；最后对基于可视化技术的用户交互技术和领域知识结合的可视分析技术进行了简要介绍。

第2章：针对控制过程数据的复现与数据分析问题，提出了一种针对过程数据分析的可视化方法，从而能够对动态状态数据进行实时展示，有效提高过程数据分析的直观性和效率，同时提出的数据处理方法和处理效率能满足实时交互需求。

第3章：针对电视剧收视率在播放过程中的影响因素分析需求，提出了一种时空特征可视化方法，从而能够快速获取不同电视台和电视剧类别在收视率和观众两个方面的对比可视化分析，总结出各目标电视台的差异性特征，从而帮助电视台在制作、购买和编排电视剧等方面做出决策。

第4章：针对网络考场中的常规监考手段无法及时发现网络作弊的问题，提出了一种针对网络考场监控日志数据流的可视化方法。将整个考场中所有学生的考试状态呈现在同一可视化结果中，一旦有考生存在作弊等异常行为，可视化结果中能够实现提醒，并能对考生进行行为分析。

第5章：针对空气质量数据的分析需求，提出了时空数据可视化方法，帮助用户全方面、多角度发现雾霾污染源头及与时间、城市等之间的关系。对时空数据的可视化方法研究还是对空气质量的规律发现、污染源发现都十分具有现实意义。

第6章：针对食品安全领域农药残留检测数据的可视分析需求，提出了一种针对多判定标准的对比可视化方法，可有效实现农药残留检测数据的可视化，并可实现地理位置、农药、农产品维度上的多尺度、多标准的对比分析。

第7章：针对农药残留检测数据多统计量的对比展示及安全风险评估需求，提出了一种针对农药残留数据的时序分组可视化方法，从而能够在可视化结果中一次性表现多种统计量数据，并能实现时间维度上的数据对比。

第8章：针对数据集中两类互相关联的研究对象，通过可视化布局方法的设计突出一类研究对象之于另一类研究对象的倾向性，展现了用户重点关注属性的倾向性分布模式，提出了一种基于极坐标的旋转布局可视化方法，在突出展现数据倾向性关联分布特点的同时，展现数据的多统计量。

第9章：针对多关系数据的表达中超边的可视化效果不直观、描述不准确的问题，提出了两种基于 Catmull－Rom 插值算法的超图可视化方法，从而能够直观、有效地表达超图中的超边。超图中的各条超边可以保持很高的区分度，绘制效率能满足实时交互的要求。

本书中内容以作者近几年来的研究工作为基础，属于可视化及可视分析领域最新

的研究成果，对于系统地了解、学习和研究数据可视化、信息可视化方面的前沿知识，具有较好的帮助作用。

本书可作为计算机、信息等相关专业的教师、研究生和大学高年级学生的参考书，也适合于从事数据可视化、大数据分析、数据挖掘、知识发现等方面的研究人员和技术人员阅读使用。

本书所涉及的所有研究工作由作者本人及所属研究团队、多名研究生协力完成，本书所涉及的研究工作得到了北京工商大学陈谊教授、孙悦红副教授、刘瑞军副教授的大力帮助，在此致以深深的感谢！本书中所述研究成果相关的研究生包括：北京工商大学2014级研究生方艺、2015级研究生杨倩玉和樊亚慧、2016级研究生程中娟和温玉琳，在此对全体研究生同学的辛勤工作表示感谢。

本书的研究工作得到了国家自然科学基金（31701517）、北京市自然科学基金（4154066、9164028）、北京市社会科学基金（17GLC060）、北京市属高校青年拔尖人才培育计划项目（CIT&TCD201704039）、北京市教委科技计划面上项目（KM201410011004）、北京市优秀人才培养资助青年骨干个人项目（20140000 20124G029）的资助。本书的出版得到了“北京工商大学科研创新服务能力建设项目（食品类专项）（No. PXM2018_014213_000033）”的资助，在此表示深深的感谢！并对所有关心与支持本研究工作的领导、专家和各位老师表示感谢！

由于作者水平有限，书中错误和不足之处在所难免，恳请读者予以指正。

作者

2018年12月

目　录

第 1 章
可视化相关技术基础

1.1 数据可视化的技术分类

数据可视化以直观的方式表达数据信息，已被可视化领域众多专家证明为一种高效获取信息的方法。对于可视化技术的研究，Neumann[3]给出了一个可视化信息分析技术的框架。可视化的表现形式主要可分为基于几何的可视化技术、基于像素的可视化技术和基于图标的可视化技术。在可视化技术方面，根据数据特征可分为层次数据可视化技术、多维数据可视化技术、时序数据可视化技术等。

1.1.1 层次数据可视化

层次数据是一种常见的数据类型，具有的层次结构是一种抽象的树形结构，注重表达数据间的层次关系。层次关系主要分为包含和从属两类，也可以表示逻辑上的承接关系。在一个树形结构中，只有根节点没有父节点，其余节点有且仅有一个与之相连的父节点。每个节点对应一条数据，节点值代表数据属性值，父节点下的分支代表数据集的逐级向下的分类。

对于树形层次数据的可视化方法一般包含两类[4]：一类是节点 - 链接法（Node - Link，又称点线法），如双曲树（Hyperbolic Tree）；另一类是空间填充法（Space - Filling），如树图（TreeMap）。在点线法中，随着数据规模的增大，一般需要通过边绑定的方式[5]降低连线交叉问题。Ghani[6]通过点线式图像扩大数据点携带的信息量，提高了数据表达能力。在空间填充法中，也出现了很多的改进和变种，袁晓如等人[7]从树图的布局算法、交互方法、改进和变种、应用领域和用户评价研究等角度，对树图

可视化及其扩展方法的基础和研究前沿进行了综述。

1. 节点－链接法（Node－Link）

节点－链接法将单个个体绘制成一个节点，节点之间的连线表示个体之间的层次关系，该方法清晰直观，外形接近于树的结构，对于表达层次关系有显著的优势，但是当数据的广度和深度较大时，空间利用率较低并且可读性差，因而不利于广度和深度相差较大时的布局。

节点－链接法的核心问题是如何在屏幕上放置节点，以及如何绘制节点间的链接关系。设计一个清晰有效的节点－链接图需要考虑以下四点：

1）节点位置的空间顺序和层次关系应一致。

2）减少连线之间的交叉。

3）减少连线的总长度。

4）可视化应该有一个合适的长宽比，以便优化空间的利用。

节点－链接法常用的布局方法包含正交布局（Axis－parallel）和径向布局（Radial），代表技术有双曲树[8]、径向树[9]。

（1）正交布局

正交布局节点放置都按照水平或垂直对齐，如 Kerr·B[10] 在 2003 年提出的 Thread Arcs 可视化方法，用来可视化邮件的答复时间顺序以及邮件之间的答复关系。如图 1-1 所示，每个节点表示一个消息，节点根据消息到达的先后顺序均匀分布于水平线上。发送消息的节点称为父节点，接收消息的节点称为叶节点，用螺旋弧表示消息之间的答复关系。Thread Arcs 可视化方法强调了消息的时间顺序，并且更加稳定和紧凑。

为了克服树状图层次较大时不便于浏览与理解的情况，SONG H 等人[11] 在 2010 年提出了一种带滚动条的一级列表和多级列表可视化方法。当层次结构中某一级子节点较少时，可以采用传统的方式（图 1-2a），用焦点＋上下文方式，通过放大关注的节点以及收缩非焦点节点来显示内容。当子节点较多时，则可以用带滚动条的一级列表，通过鼠标滚动来交互式的显示更多节点信息。然而，也可以选择第三种方式，如图 1-2c所示，即当节点较多时通过多级列表来显示更多内容，而不是通过滚动条查看。

相比较而言，第三种方式更易于用户理解层次结构，并且能在提升空间利用率的同时显示更多的节点信息。Ploeg A[12] 在 2014 年设计了一种 non－layered 树，针对层次较大时水平排列的节点占据较大空间问题，提出改变节点位置，有效利用空余空间展

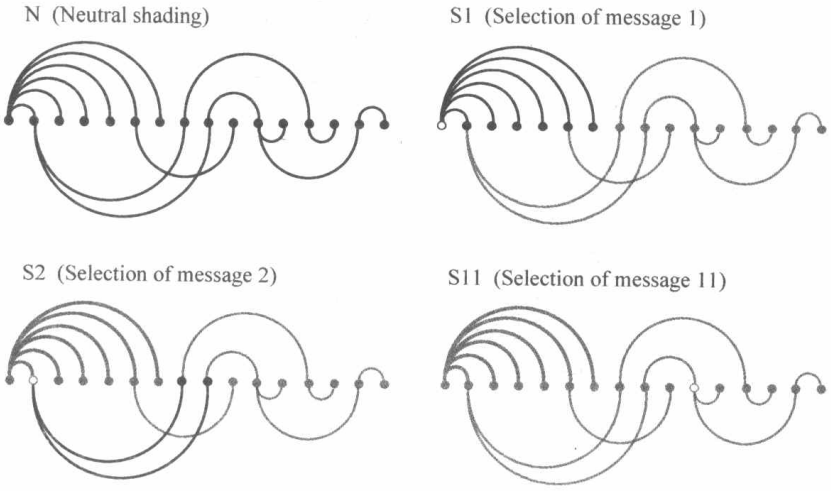
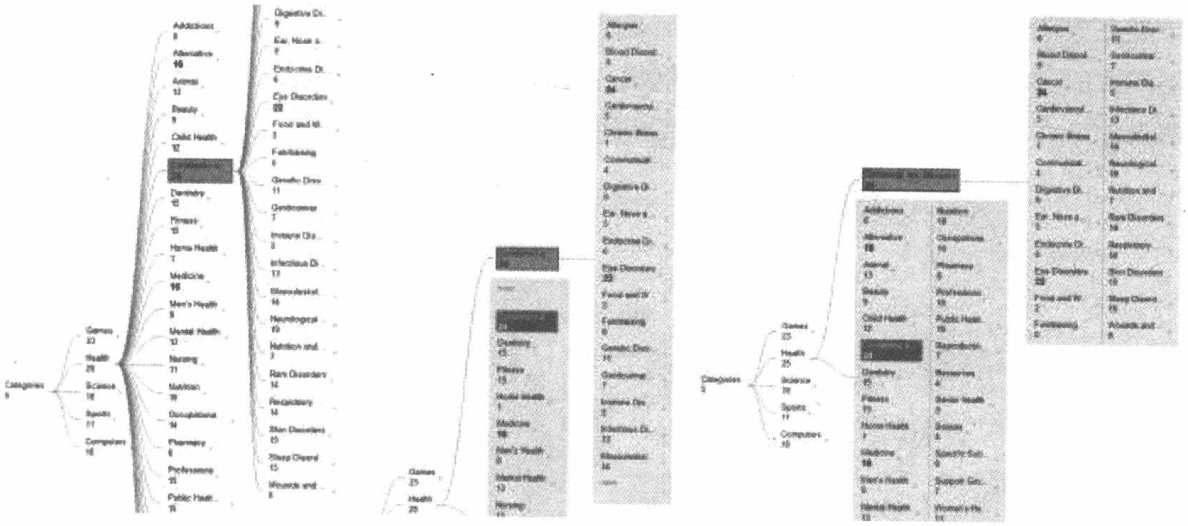


图 1-1　Thread Arcs 布局



a) 传统　　b) 列表　　c) 多级列表

图 1-2　多级列表布局

示兄弟节点，其效果图如图 1-3 所示。节点 - 链接不仅可以表达层次关系，还能表示网络图中边的链接关系，如 Sallaberry A[9] 在 2016 年设计了一种 Contact 树，用来可视化网络中的节点和边。

（2）径向布局

正交布局方法虽然直观，但是对于大型层次结构会造成数据显示空间不足和屏幕

空间浪费，所以径向布局被学者更为广泛地研究、使用。径向布局将根节点置于圆心，不同层次的节点被放置在半径不同的同心圆上，节点到圆心的距离对应于它的深度。这样的布局方式可以容纳更多的节点，并且克服了空间浪费的问题。

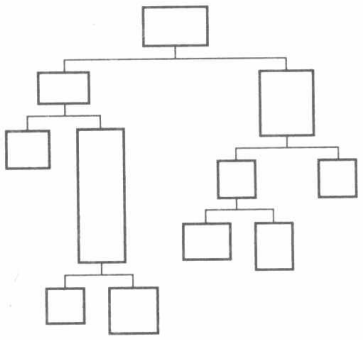


图 1-3　Non－layered 树

径向布局对于大的层次结构，在树的底层空间显示不足，会造成重叠现象，而双曲树则采用双曲空间作为信息的显示空间，使有限空间上可容纳信息节点更多了，例如 1996 年 John 提出的 Hyperbolic Browser，如图 1-4 所示[13]。

SCHULZ HJ[14] 于 2011 年提出一种环状径向树方法，与以往的方式不同，它在侧重于布置较大节点的同时还能保证较好的空间利用率。该方法的布局如图 1-5 所示，将节点布置于网格中，首先将根节点位于屏幕中心，然后再将根节点的前 4 个子节点放置在根节点的周围，其后按每四个节点为一组按一定比例旋转放置，其他的子树按照同样的方法递归布局。对于子树的规模大小采用颜色编码来显示。该方法通过将子节点显示在不同层空间来有效改善空间利用率，但不利于观察节点的父子关系。

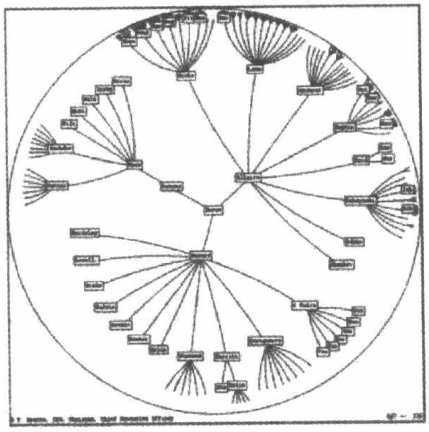


图 1-4　双曲线布局

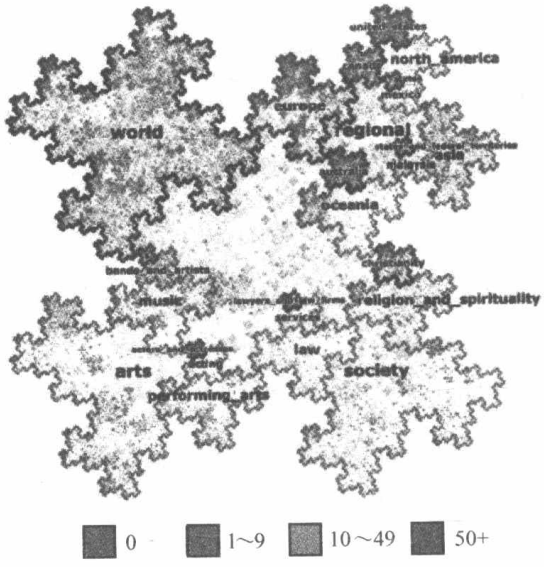


图 1-5　径向树布局图

Urribarri D K[15] 于 2013 年提出将双曲树布局引入到三维空间，设计了一种 Gyrolayout 布局方法，可以支持不同细节等级（Level - of - Detail）技术，以帮助用户交互式地探索分析大型数据集，其效果图如图 1-6 所示。

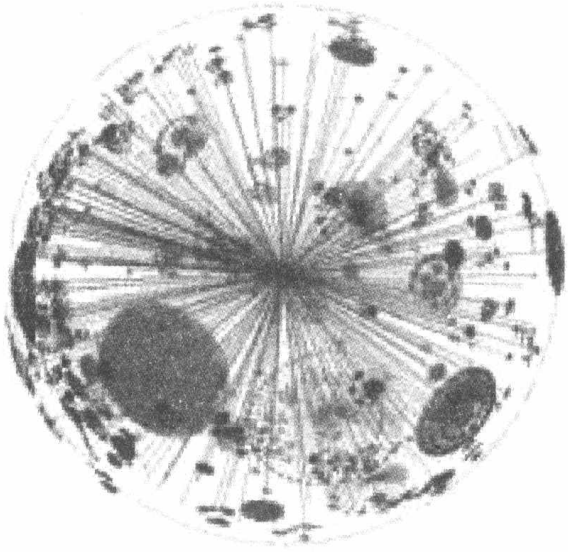


图 1-6　Gyrolayout 布局

Lott S C[16] 于 2015 年提出 CoVenn 树（加权维恩树）方法，它可利用维恩图的三色圆表示三类数据集，并将其聚集采用节点 - 链接法表示节点间的关联关系，且将圆的大小映射属性值。该方法可以展示大量数据集，并可同时展示比较多个数据集，且利用径向布局可有效利用空间，其效果图如图 1-7 所示。

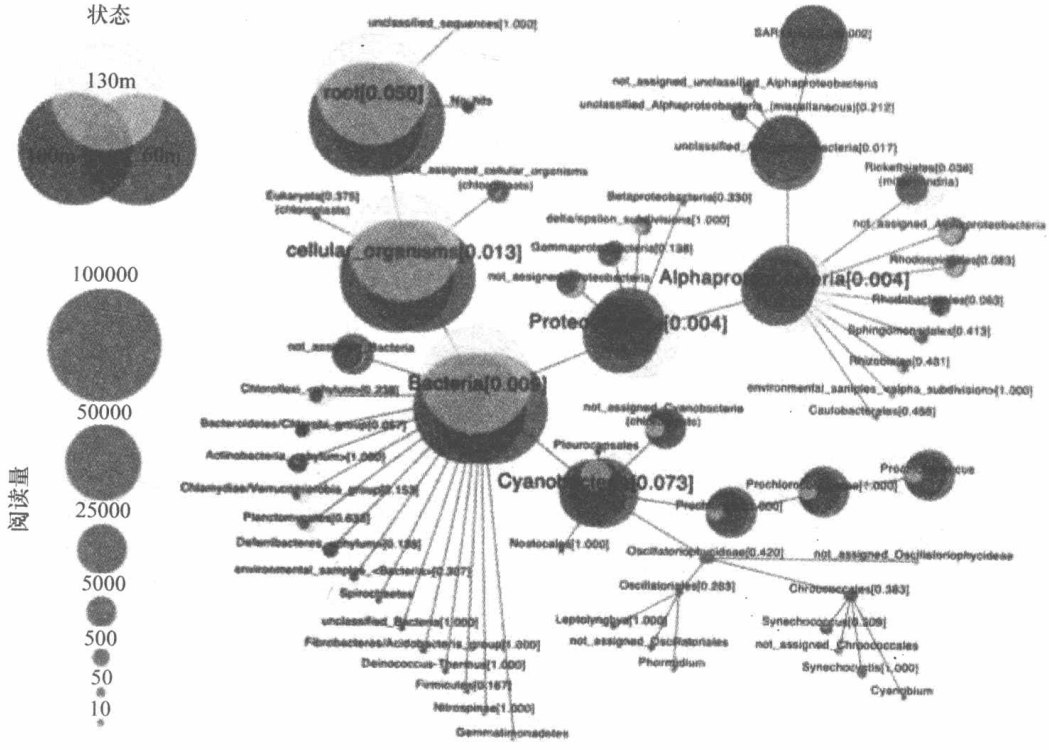


图 1-7　CoVenn 树布局

2. 空间填充法（Space - Filling）

空间填充法用空间中的分块区域表示数据中的个体，并用外层区域对内层区域的包围表示彼此之间的层次关系。相对节点 - 链接法弱于层次结构的表达，空间填充法提高了空间利用率。与节点 - 链接法相比，空间填充法更适合表示包含与从属关系，并且空间利用率较高，但不利于层次信息的表达。

空间填充法主要有树图（TreeMap）[17]和放射环（SunBurst，又称玫瑰图）[18]两种方法。树图采用矩形表示节点，通过矩形的嵌套表达父子节点关系。玫瑰图的径向布局类似于节点 - 链接法里面的径向树，但其采用放射环填充的形式改善了空间利用率，并且比树图更注重层次关系。

（1）树图（TreeMap）

Johnson 和 Schneiderman 等人[19]在 1991 年提出了树图，采用嵌套的矩形表示节点以及层次关系，在此基础上衍生出了多种树图布局改进算法，例如交替纵横切分法（Slice And Dice）、正等分法（Squarified）、有序布局（Pivot）、条形布局（Strip）等[7]。

Tak Susanne[17]等人在 2012 年提出了用希尔伯特（Hilbert）和摩尔（Moore）曲线来构建树图，如图 1-8 所示。布局方法分为两步：

1）根据节点权值将节点划分为权值总和大体相近的四个象限。

2）将每个象限内节点根据 Hilbert 和 Moore 曲线排列。这种布局方法保证了空间稳定性，当数据发生变化的时候节点位置不会发生太大的变化，并且在其他度量属性上也维持较好的效果，包括长宽比、最大连续性。

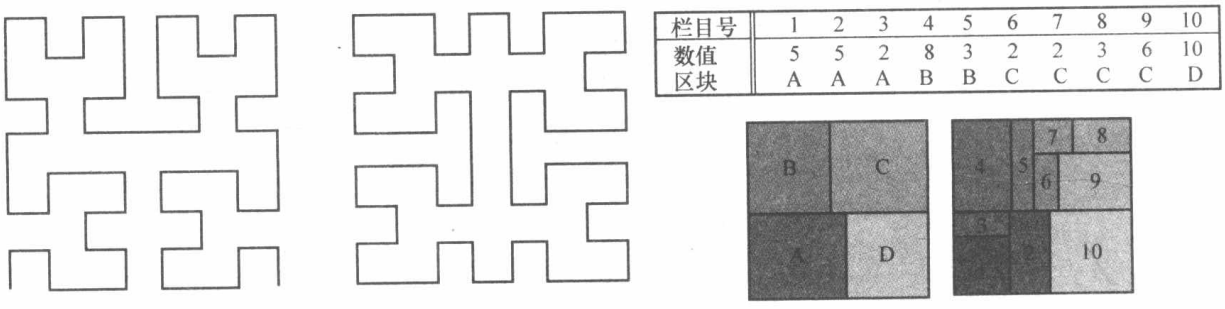


图 1-8　希尔伯特和摩尔布局

空间长宽比是树图布局好坏的一个重要评判标准，为了克服矩形空间长宽比，