软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(5):1091−1102 [doi: 10.13328/j.cnki.jos.004956]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种层次结构中多维属性的可视化方法*

陈 谊[1,2], 甄远刚[1,2], 胡海云[1,2], 梁 婕[3,4], Kwan-Liu MA[5]

[1](食品安全大数据技术北京市重点实验室(北京工商大学),北京 100048)

[2](北京工商大学 计算机与信息工程学院,北京 100048)

[3](机器感知与智能教育部重点实验室(北京大学),北京 100871)

[4](北京大学 信息科学与技术学院,北京 100871)

[5](Department of Computer Science, University of California-Davis, Davis, USA)

通讯作者: 陈谊, E-mail: chenyi@th.btbu.edu.cn

摘 要: 在很多领域的统计分析中,通常需要分析既具有层次结构又具有多维属性的复杂数据,如食品安全数据、股票数据、网络安全数据等.针对现有多维数据和层次结构的可视化方法不能满足对同时具有层次和多维两种属性数据的可视分析要求,提出了一种树图中的多维坐标 MCT(multi-coordinate in treemap)技术.该技术采用基于 Squarified 和 Strip 布局算法的树图表示层次结构,用树图中节点矩形的边作为属性轴,通过属性映射、属性点连接、曲线拟合实现层次结构中多维属性的可视化.将该技术应用于全国农药残留侦测数据,实现了对全国各地区、各超市、各农产品中农药残留检出和超标情况的可视化,为领域人员提供了有效的分析工具.MCT 技术也可用于其他领域的层次多属性数据的可视化.

关键词: 信息可视化;层次结构;多维属性;树图;平行坐标

中图法分类号: TP391

中文引用格式: 陈谊,甄远刚,胡海云,梁婕,Ma KL.一种层次结构中多维属性的可视化方法.软件学报,2016,27(5):1091−1102. http://www.jos.org.cn/1000-9825/4956.htm

英文引用格式: Chen Y, Zhen YG, Hu HY, Liang J, Ma KL. Visualization technique for multi-attrbute in hierarchical structure. Ruan Jian Xue Bao/Journal of Software, 2016,27(5):1091−1102 (in Chinese). http://www.jos.org.cn/1000-9825/4956.htm

Visualization Technique for Multi-Attrbute in Hierarchical Structure

CHEN Yi[1,2], ZHEN Yuan-Gang[1,2], HU Hai-Yun[1,2], LIANG Jie[3,4], Kwan-Liu MA[5]

[1](Beijing Key Laboratory of Big Data Technology for Food Safety (Beijing Technology and Business University), Beijing 100048, China)

[2](School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

[3](Key Laboratory of Machine Perception of Ministry of Education (Peking University), Beijing 100871, China)

[4](School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

[5](Department of Computer Science, University of California-Davis, Davis, USA)

Abstract: Nowadays, there is increasing need to analyze the complex data with both hierarchical and multi-attributes in many fields such as food safety, stock market, and network security. The visual analytics appeared in recent years provides a good solution to analyze this kind of data. So far, many visualization methods for multi-dimensional data and hierarchical data, the typical data objects in the field

* 基金项目: “十二五”国家科技支撑计划(2012BAD29B01-2); 国家科技基础性工作专项(2015FY111200); 虚拟现实技术与系统国家重点实验室(北京航空航天大学)开放基金(BUAA-VR-14KF-04)

Foundation item: “Twelfth Five Year Plan” National Key Technology R&D Program of China (2012BAD29B01-2); Basic Research Project of the Ministry of Science and Technology of China (2015FY111200); Open Funding Project of the State Key Laboratory of Virtual Reality Technology and Systems (BeiHang University) of China (BUAA-VR-14KF-04)

收稿时间: 2015-07-31; 修改时间: 2015-09-19; 采用时间: 2015-11-10

of information visualization, have been presented to solve data analyzing problems effectively. However, the existing solutions can't meet requirements of visual analysis for the complex data with both multi-dimensional and hierarchical attributes. This paper presents a technology named Multi-Coordinate in Treemap (MCT), which combines rectangle treemap and multi-dimensional coordinates techniques. MCT uses treemap created with Squarified and Strip layout algorithm to represent hierarchical structure, uses four edges of treemap's rectangular node as the attribute axis, and through mapping property values to attribute axis, connecting attribute points and fitting curve, to achieve visualization of multi-attribute in hierarchical structure. This work applies MCT technology to visualize pesticide residue detection data and implements the visualization for detecting excessive pesticide residue in fruits and vegetables distributed in each provinces of China. This technology provides an efficient analysis tool for field experts. MCT can also be applied in other fields which require visual analysis of complex data with both hierarchical and multi-attribute.

Key words: information visualization; hierarchical structure; multi-attribute; treemap; parallel coordinate

随着数据采集技术的革新与进步,经济社会各领域获得的数据体量和复杂性不断激增,为数据分析提出了极大的挑战.信息可视化与可视分析作为一种大数据分析的有效方法,已逐渐显示出其强大的生命力.近年来,针对具有多维属性的多维数据、具有层次结构的层次数据、具有网络关系的网络数据[1]和具有时间空间特征的时空数据等各种类型的数据,人们已提出了相应的可视化方法[2].

层次数据的可视化算法主要有节点-链接法和空间填充法 2 种.节点-链接法采用不同形状的点表示对象(内容信息),点之间的连线表示对象间的关系(结构信息).其优点是层次结构表现清晰,其缺点是空间利用率低,展现层次和节点数目有限,通常不能表现属性值的大小等.空间填充法则利用空间嵌套形式表示层次结构,利用多边形面积表示对象属性值的大小[3].其优点是空间利用率高,展现的层次和节点数比节点-链接法多,且可展现节点权值的大小;其缺点是层次结构展现没有节点-链接法清晰.多维数据的可视化方法主要有散点图[4]、散点图矩阵[5]、平行坐标[6]、雷达图[7]、堆叠图[8]等,其目的是将多维数据从抽象的高维空间上映射到可视的二维或三维空间上,便于人们理解数据和发现其中的信息.

而在实际应用领域中,数据的组成不断趋于复杂,大部分数据不是仅具有单一的数据特征,而是同时具有多种数据特征,本文将其称为复杂数据,如农药残留侦测数据、股票数据和网络安全数据就都具有多种数据特征.对于这类复杂数据,现有针对单一数据特征的可视化和可视分析方法已不能满足对其分析的需求,一般采用由 2 种或以上的可视化方法结合而成,例如 SONHC[9],Elastic Hierarchies[10]和 TreemapBar[11]等.

针对具有层次、多维、时空特征的农药残留侦测复杂数据,本文提出了一种树图中的多维坐标布局算法 MCT.该算法将平行坐标的思想应用于树图布局之中,充分结合树图布局矩形填充的特点,利用有限的可视化空间,同时展示数据的层次结构和多维属性信息,帮助领域可视分析.将该算法应用于农药残留侦测数据,用树图来表示农产品分类、农药分类和地域的层次结构,用矩形节点中的多维坐标表示农产品、农药、残留量、限量标准值等多维属性,取得了较好的效果.

1 相关研究

1.1 基于树图的层次数据可视化方法

树图是一种典型的层次数据可视化方法,矩形树图通过矩形嵌套填充展示层次结构,通过矩形面积大小展示节点权值.近年来,人们对其布局算法进行了大量研究.Johnson 等人[12]于 1991 年提出了一种树图布局算法,命名为 Slice and Dice.该布局算法容易出现长宽比恶劣(长宽比与 1 相差很多)的矩形,导致矩形难以识别,不便于交互,影响算法可应用的范围.Bruls 等人[13]针对该缺点于 1999 年提出了 Squarified 布局算法,确保填充的矩形更接近正方形,易于人眼对矩形的识别.为了使无序的 Squarified 布局算法部分有序,陈谊等人[14]于 2013 年提出了一种分块排序的正方化树图布局算法 Squarified-SP.Shneiderman 等人[15]于 2001 年提出了 Pivot 布局算法.Pivot 布局算法采用分治的思想,运用 pivot 节点将数据集 T 的填充区域划分为 R_1,R_P,R_2 和 R_3 这 4 部分,在当前的封闭矩形内进行布局.Bederson 等人[16]于 2002 年提出了以带状形式进行矩形填充的 Strip 算法,Tu 等人[17]于 2007 年提出了以螺旋状进行矩形填充的 Spiral 算法.这 3 种算法在保证树图布局矩形平均长宽比的情况下,能

够达到更高的连续性和可读性[18].针对具有地理位置属性的特殊层次结构的数据,Wood 等人[19]于 2008 年提出了 Ordered-Squarified 和 Spatially Ordered 这两种算法.胡海云等人[20]于 2014 年针对节点权值差异大的层次数据,提出了一种有序的正方化树图布局算法 SOTLA,在保证平均长宽比的情况下兼顾了连续性.各种算法的优势与缺陷见表 1.当数据层次关系较为复杂时,传统树图算法无法有效布局,此时可以引入 Cushion[13]和 Framing[12]算法,将嵌套的矩形以层叠或缩进的形式进行优化,以突显数据的层次结构.虽然上述算法在树图布局的基础上进行了层次结构表现力的优化,但树图布局算法依然不适用于表现层次关系过于复杂的数据.树图布局算法的可视化效果随层次的增加而减弱.在使用树图布局算法时,需要控制数据层次显示的深度.

Table 1 Treemap visualization methods' comparison

表 1 树图可视化方法比较

发表年份	方法名称	优势	缺陷
1991	Slice and Dice	最早提出的布局算法,简单且易于实现	容易出现长宽比恶劣的狭长矩形
1999	Squarified	矩形更接近正方形,易于是别饿和交互	排列无序,用户查找节点困难
2001	Pivot	节点部分有序,便于对随时间变化而更新的数据进行追踪	算法复杂,不能保证所有节点有序
2002	Strip	以带状形式进行矩形填充,保证所有节点有序	距离相关性较差
2007	Spiral	以螺旋状进行矩形填充,能够保证所有节点有序	无明显缺陷,但每个指标均不优秀
2008	Ordered-Squarified, Spatially Ordered	解决地理位置等特殊层次数据布局顺序问题	只适用于地理位置等特殊数据,没有通用性
2013	Squarified-SP	算法简单,保证平均长宽比,且节点部分有序	不能保证所有节点有序
2014	SOTLA	兼顾节点有序和平均长宽比	在几种算法中,可读性和稳定性指标居中

1.2 基于平行坐标的多维数据可视化方法

平行坐标(parallel coordinate)[6]是最经典的多维数据可视化方法之一,它通过多个平行坐标轴表示数据的各个属性,从下到上数据值逐渐变大,通过各个坐标轴之间的连线,能够快速定位某条数据记录的各个属性值.经典平行坐标的缺点是坐标轴固定,当数据量过多时,边重叠交叉覆盖严重.

TimeWheel 由 Tominski 等人[21]提出,将 6 个属性轴以六边形的形式围绕时间轴,以不同的颜色连线代表不同的属性与时间轴的映射关系.Jarry 等人[22]提出一种灵活连接坐标,通过改造平行坐标的轴,可帮助用户自定义不同的形状的坐标轴表示不同属性.

平行坐标能够对多维数据进行表示,直观地展示多维属性之间的关系,帮助用户发现多维数据间隐藏的关联和多维大数据集的变化规律.但是单一的平行坐标不能展示数据集其他方面的特征,并且在数据较多的情况下,传统的平行坐标会存在大量交叉、重叠的线段.

因此,许多专家学者致力于对平行坐标的改进,较为常见的方式是将平行坐标与聚类、交互和其他可视化技术结合.Fua 等人[23]利用层次聚类算法构造分层聚簇树,提出采用分层显示的方法在平行坐标中对数据进行分层显示,从而能够从不同抽象层次上表示数据,有效地减轻了平行坐标中的视觉杂乱问题.Hong 等人[24]在前人基础上提出了新的平行坐标可视化聚类方法,利用 Catmull-Rom 样条曲线来替代原始直线平行坐标,以最小化曲率和最大化相邻边缘的平行技术来优化样条曲线,可视化效果得到进一步提升.此后,局部交互聚类方法也被 Guo 等人[25]提出并应用于平行坐标当中,这种方法可以让用户选中平行坐标中感兴趣的区域,通过交互的方式,直线向该区间聚拢.此外,Slingsb 等人[26]将平行坐标与密度图思想结合在一起,通过计算平行坐标中直线分布区域的密度,以颜色明亮度来表示线条密度的情况,从而减少了异常数据的干扰.Claessen 等人[22]提出了将坐标轴灵活移动的方案,用户可以通过自由地设定坐标轴的布局位置,从而使得不受显示区域的限制,人们可以通过不断改变平行坐标轴的位置来找寻数据隐藏的规律.

1.3 基于树图的混合布局算法

树图布局算法可以很好地应用于层次数据的可视分析.对于较为复杂的数据,单一的树图布局无法满足其所有属性可视化的需求.选用混合布局算法能够实现可视化算法之间的互补,更好地显示数据的复杂特征.根据树图布局的矩形布局特点,节点通过与节点链接、标签云和柱状图等可视化方法相结合,可以形成新的混合布局算法.

Zhao 等人[10]于 2005 年结合节点-链接的形式和树图布局算法提出了 Elastic Hierarchies 算法,用“焦点+上下文”的交互技术作为算法之间的衔接方式,布局效果不仅突显层次结构,并且极大程度地提高了空间利用率;Linsen 等人[27]于 2011 年提出了 Linked treemap,将节点链接和树图算法的结合方式扩展到 3D 的维度;Vliegen 等人[28]于 2006 年提出了 Generalized Treemap,把树图布局算法嵌入柱状图、饼图和节点链接图中,更适用于商业数据的可视化分析.Huang 等人[11]于 2009 年提出了 TreemapBar,在柱状图中加入树图结构,充分利用柱状图中的矩形空间,并通过 tablelens 的缩放技术,使其能够查看数据中更细节的内容.相反地,为了充分利用树图填充矩形中的空白位置,Kobayashi 等人[29]于 2012 年提出了 Edge Equalized Treemaps,在树图的布局矩形中嵌套显示柱状图,在展示树图节点的层次结构的同时,实现了对叶子节点的属性对比.

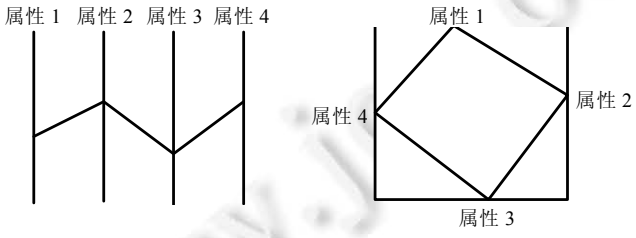
2 MCT 布局算法

2.1 设计目标

为了可视化具有层次结构和多维属性的数据集,充分利用树图中节点矩形的内部空间,本文提出了一种称为 MCT 的可视化技术.该技术将树图与平行坐标可视化思想相结合,把树图节点的 4 条边定义为 4 个属性轴,通过属性值映射和规格化,属性值首尾顺次相连,曲线拟合形成完整统一的视图.使用 MCT 方法可以实现对具有层次结构的多维属性数据进行可视化,从而实现如下的分析任务:(1) 比较层次结构中各节点权值的大小;(2) 分析各属性间的关联;(3) 根据多维属性比较分析各节点的综合指标,发现异常.

2.2 树图节点中的多维属性轴

首先使用基于 Strip 的树图布局算法将层次结构可视化为树图,则层次结构中每个节点被表示为一个矩形.参照平行坐标的思想,将矩形每一条边类比为平行坐标中的一个属性轴,如图 1(a)所示.在一个节点矩形中,可以用其 4 条边分别表示为除节点权值外的 4 个属性,如图 1(b)所示.规定从最上边开始以顺时针的方向,依次排列属性 1~属性 4.属性值从小到大的方向分别为从左至右和从上至下,其中,离散值以首字母顺序作为排序的基准.由于每个节点同时具有多个属性值,所以矩形边之间会存在多条线条连接的情况形成一个闭合图形,如图 1(b)所示.



(a) 具有 4 维属性的平行坐标 (b) 平行坐标到多维坐标的映射

Fig.1 Multi-Attribute axis of in treemap nodes

图 1 树图节点中的多维属性轴

2.3 连续属性值的映射

若该矩形边对应的属性值为连续数值,则可以根据属性值的最小值 M_n 和最大值 M_x 确定矩形中属性值在该

矩形边上映射的位置.设位置 P_n 和 P_x 分别对应属性权值的最小值 M_n 和最大值 M_x,则属性值为 V_a 的点在矩形边上对应的位置 P_a 记作:

$$P_a=\frac{(V_a-M_n)\times(P_x-P_n)}{M_x-M_n}+P_n \tag{1}$$

如果直接将矩形边的两端设为 P_n 和 P_x,当相邻属性处于最大点或最小点时,属性值的点将会处于矩形两边相交的顶点处,会影响该点的归属判定,无法得知该点是属于哪个属性.而当相邻属性分别取得最大值和最小值时,连线有可能与矩形边重叠,如图 2(a)的红线边所示.因此,需要对边的有效位置重新设定.采取百分比缩进的形式重新设置矩形边位置 P_n 和 P_x,为每一条的两端留出空白位置作为临界区域,不进行利用.

若该属性所映射的矩形边边长为 L,缩进百分比为 q,则 $P_n=\frac{L\times q}{2}$,$P_x=L\times(1-q/2)$.通过将点布局的位置进行前后两端的缩进,避免了属性值点归属不清和不同属性之间的连线与矩形边重合的问题.

(a) 属性连线与矩形边重叠情况 (b) MCT 算法中连续属性值刻度设定

Fig.2 Continuous attribute values scale set and connection diagram

图 2 连续属性值刻度设定与连线示意图

2.4 离散属性值的映射

若属性取值为离散值,将不能采用与连续值属性相同的映射方式,需要重新计算该矩形边下不重复的离散值的个数.本文采用平均划分的方式对离散属性值与矩形边位置进行映射.

若该 A 属性具有离散值 $X\{x_1,x_2,\ldots,x_s\}$,总计共有 s 个,计算其去重后的离散值为 $Y\{y_1,y_2,\ldots,y_u\}$,共计 u 个.该属性对应的矩形边边长为 L,则计算两个离散属性值之间的间隔距离 $d=L/u$.设每个离散属性值在矩形边的开始端位置为 $D\{d/2,d/2+d,\ldots,d/2+(u-1)d\}$,分别对应属性值 $\{y_1,y_2,\ldots,y_u\}$.为了防止某一离散刻度上的线过于密集,对不同属性值对应的相同的离散刻度需要进行位置偏移.首先,计算每个点位移的位置为 $c=d/s$.具有相同离散值的点,其在矩形边上的位置以 $D[i]$为中心,按照顺序依次从左向右或从上至下进行位置映射.以第 3 个刻度为例,若有 5 条线对应 $5d/2$ 这个刻度,则开始以第 $5d/2-2c$ 点的位置为基础,逐条直线的连接点向右偏移 c,如图 3 所示,最后一条直线的连接点为 $5d/2+2c$.当且仅当离散值 Y 的个数 u 为 1 时,才会出现属性值点的位置位于矩形顶点的情况.若出现上述极端情况,或由于屏幕像素限制引起的连线与矩形边重叠时,则可采用图 2 中所提到的缩进形式进行优化,避免节点与矩形顶点重合,影响数据分析.离散值刻度的设定如图 3 所示.

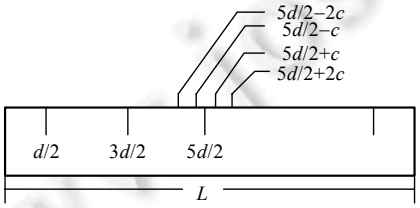


Fig.3 Discrete attribute value calibration set diagram

图 3 离散属性值刻度设定示意图

2.5 属性连接与曲线拟合

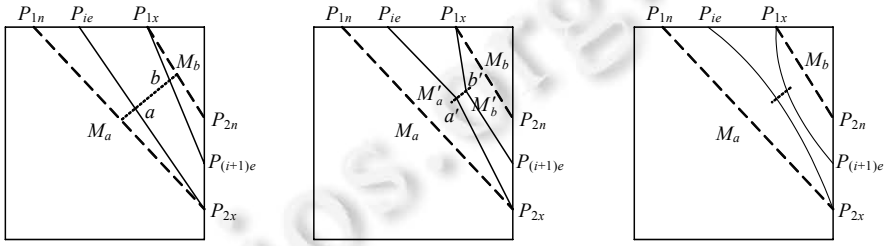
对于两个相邻的属性,可以使用直线对属性值的点进行点对点的连接,将 4 个维度的属性值连成一个四边

形,如图 1(b)所示.由于不同属性之间的连线有可能出现交叉的情况,布局的混乱程度将随连线数量的增加而不断加大,即使使用了颜色对线条进行区分,也会对数据的分析造成不同程度影响.针对该情况,本文采用基数样条曲线的算法,将直线优化为中间收缩的曲线.

以矩形的最上边和最右边的属性之间的连接线为例,若此相邻属性之间的最小值和最大值坐标分别对应 P_{1n},P_{1x},P_{2n} 和 P_{2x},则这 4 个点形成不规则四边形 $P_{1n}P_{1x}P_{2x}P_{2n}$.分别取 $P_{1n}P_{2x}$ 和 $P_{1x}P_{2n}$ 中点 M_a 和 M_b,如图 4(a)所示.根据用户自定义的收缩系数 r,以 M_a 和 M_b 的中点对线段 M_aM_b 进行长度的收缩,使得 $M'_b-M'_a=(M_b-M_a)\times r$,$M'_a$ 与 M'_b 为收缩后的对应 M_a 和 M_b 的点,如图 4(b)所示.若属性 i 和属性 i+1 的 2 点 P_{ie} 和 $P_{(i+1)e}$ 分别与 P_{2x} 与 P_{1x} 存在连线,该连线与线段 M_aM_b 的交点为 a,b,则经过缩放后系数 r 改变后的交点位置为(a',b'),如公式(2)所示:

$$a'=\frac{(a-M_a)\times(M'_b-M'_a)}{M_b-M_a}+M'_a,\ b'=M'_b-\frac{(M_b-b)\times(M'_b-M'_a)}{M_b-M_a} \tag{2}$$

a',b'点位置的公式可以通用.过点 P_{ie},$P_{(i+1)e}$,P_{1x},P_{2x} 和 a',b'可利用基数样条函数画曲线,如图 4(c)所示.



(a) 直线链接 (b) 收缩后直线链接 (c) 收缩后曲线链接

Fig.4 Curve connection process between the different attribute values

图 4 不同属性值之间曲线连接处理过程

曲线由两个顶点和基础函数组来确定.用于计算的点包括曲线顶点 P_a 和 P_b,2 个张力相关的变量 T_1 和 T_2,其中,$T_i=\frac{(P_{i+1}-P_{i-1})\times g}{2}$,而 $g\in(0,1)$为用户自定义的张力系数.而当 P_i 不存在时,P_i=0.T_1 和 T_2 分别作用于点 P_a 和 P_b 上.基础函数组为公式组(3),将点 P_a,P_b 和中间的任意一个点 P_{ab} 带入公式(4)中,计算每一个节点的位置.

$$\begin{cases}h_1(s)=2s^3-3s^2+1\\h_2(s)=-2s^3+3s^2\\h_3(s)=s^3-2^2+s\\h_4(s)=s^3-s^2\end{cases} \tag{3}$$

其中,$s\in[0,1]$.根据方程组(3),并保证曲线上的每一个点 p_s 都符合公式(4):

$$p_s=h_1(s)\times P_a+h_2(s)\times P_b+h_3(s)\times T_1+h_4(s)\times T_2 \tag{4}$$

此时,将直线分割为 n_p 段,当 n_p 趋近于无穷大时,点与点之间连线无线趋近于曲线,具体过程如图 5 所示.

```
for (int t=0; t < np; t++)
{
  float s=(float)t/(float)np;
  vector p=h1(s)*Pa+h2(s)*Pb+h3(s)*T1+h4(s)*T2;
  connectto (p);
}
```

Fig.5 Curve fitting process based on spline function pseudo code

图 5 基数样条函数曲线拟合过程伪代码

以图 5 中 $P_{ie}P_{2x}$ 连线为例,分别代表属性 i 取值的点 P_{ie} 和属性 i+1 取值的点 P_{2x},以及其收缩后中线的交点

a',组成节点向量 $S(P_{ie},a',P_{2x})$,对每一条曲线的节点向量依次绘制两点之间的曲线,绘制过程如图 6 所示.

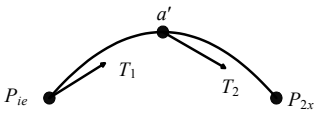


Fig.6 Curve fitting process

图 6 曲线拟合过程

根据上述曲线的绘制方式,在两个属性点位置确定的条件下,曲线的形状由辆个用户自定义的控制变量组成,分别为缩放系数 r 和曲线的张力系数 g.这两个系数的取值不同,曲线的连接形状将会有所区别,如图 7 所示.当收缩系数 q 为 1 时,3 点位于同一条直线上,连线为一条直线线段;当 $0<q<1$ 时,q 越小,收缩情况越明显曲,当收缩系数为 0 时,两个属性之间的连线将收缩于同一个中点;张力的大小将影响曲线的弯曲程度,g 越接近 1,曲线弯曲程度越大.

使用 MCT 算法对数据进行可视化布局后,用户可以根据曲线的密集程度进行收缩系数 r 和张力系数 g 的自定义设置,使曲线的弯曲程度更加适用于当前结果图的可视分析.针对不同 r 和 g 的系数值,利用 MCT 布局算法布局后的可视化效果如图 7 所示.

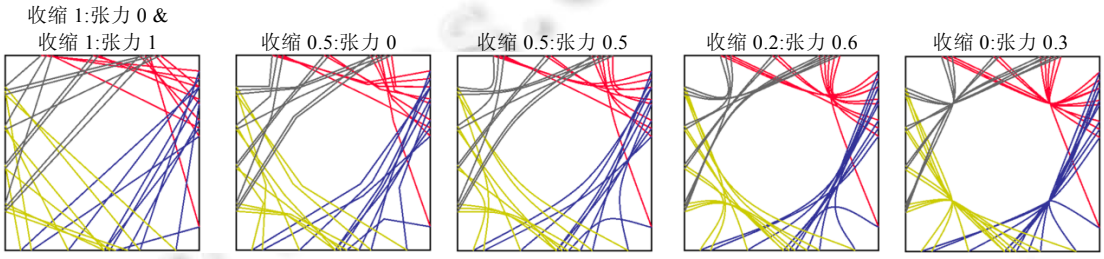


Fig.7 Under different scaling coefficient and tension coefficient curve fitting effect

图 7 不同缩放系数和张力系数下的曲线拟合效果

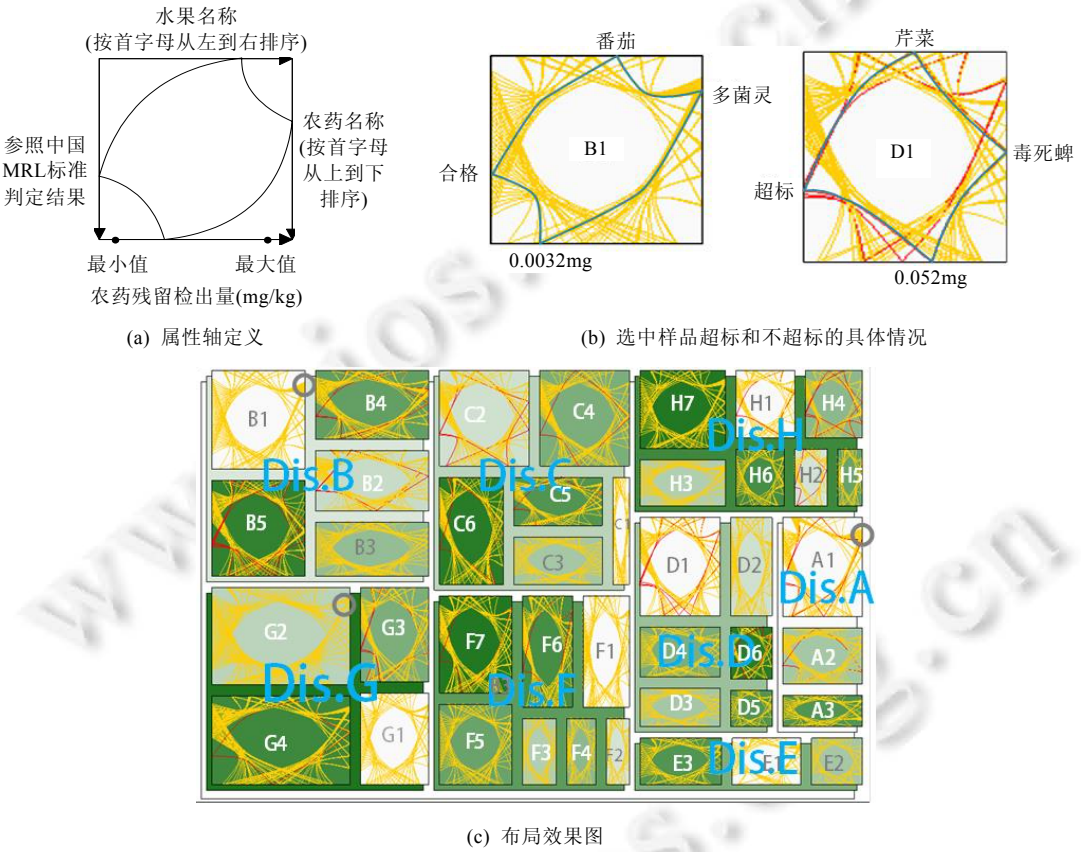
3 实例分析

利用 MCT 算法对农残数据进行可视化分析.农残数据具有多维特征,针对其层次结构,采用 Strip 算法进行树图布局,分别突出检出频次的多少和数据的顺序特性.顺序特性体现在树图节点的底色上,节点的底色越浅,表示节点的布局顺序越在前面.节点的面积表示为检出农药的占比大小(占比大小=检出农药样品数/样品总数).在节点内,采用颜色标注连线的方式找到数据异常值,即,一旦出现异常值,则与其相关的闭合曲线都呈警戒色(本文实例中超标数据的警戒色采用深色,而检测数据正常值用浅色表示).

图 8 为使用模拟的农残数据,使用 MCT 算法与 Squarified 树图算法结合布局的可视化效果图.数据集具有典型的地理区域层次特征.数据源 X 为连锁超市,在 Y 城市 8 个区所有分店的抽样检测结果,每个节点矩形的边依顺时针分别对应 20 种水果名称(离散值)、18 种农药名称(离散值)、农药残留检出量(连续值)、中国 MRL(最大残留限量 maximum residue limit)标准判定结果(离散值)这 4 项属性,如图 8(a)所示.其中,中国 MRL 标准判定结果有超标和未超标两种情况,对应矩形的右侧边点从上至下为未超标和超标.对于超标的数据,采用深色线条连接矩形 4 条表对应的属性值形成闭合曲线.按照经验,刻度缩进 s=0.8,收缩系数 r=0.2,张力系数 g=0.5.整体的布局效果图如图 8(b)、图 8(c)所示,分别展示了当选中不同超市的样品时,样品超标和不超标的具体情况.以超市 B1 和 D1 为例,选中曲线高亮显示.图 8 中的树图采用 Squarified 布局算法生成,每一个树图节点表示一个超市中各水果样品中农药残留的检出情况,包括水果名称、检出农药名称、检出量和参照中国 MRL 标准的判定

结果这 4 种属性.

由图 8(b)可以分析出:B 区的检出的农药超标频次占比最高,E 区的检出农药最小.由于每个超市检出检测的水果种类和农药种类都不一定都有检出,因此存在抽样不完全或者农药未检出的情况.从图中可以看出,每种水果对应的检出农药超标结果数据较为平衡,而农药存在相对于集中于某 1~2 个农药中.通过查看灰色圈中线条的原始数据可以发现,其对应的农药为多菌灵和嘧霉胺,均为杀菌剂.通过观察各矩形右侧边的交点,根据 MCT 算法可以推断出来线条密集区域为未超标的数据,则可以推断出大部分抽样结果属于未超标范围,仅有少量的农产品检测结果超标,部分超市的采样样品不存在农药超标的情况.同时也可以看出,所有区域都存在所有农产品均检测合格的超市.图 8 的区域 G 中,左上角第 1 家超市 G1 虽然检测的数据量较大,但是整体检测效果良好,不存在超标的检测数据.



(a) 属性轴定义 (b) 选中样品超标和不超标的具体情况

(c) 布局效果图

Fig.8 Every market in *Y* city eight districts (A,B,…,H) pesticide residue detection visual result with MCT technology

图 8 用 MCT 技术对 *Y* 市 8 个地区(A,B,…,H)中各市场检出农药残留情况的可视化结果

为了更加详细地分析数据的特征,本文将借助柱形图辅助 MCT 算法进行可视化分析.对比中国、欧盟和日本的 MRL 标准,图 9 为采用了模拟的农残数据,在 MCT 与 Squarified 树图算法结合布局的可视化系统的效果图,通过下拉列表可以选择不同的属性,若想要显示的属性少于 4 个,则树图中的对应边不利用.采用数据的水果种类作为层次关系,数据源为图 8 采用的 *Y* 市 8 个地区(A,B,…,H)中各市场检出的农药残留数据.4 个多维属性依顺时针对应水果名称、农药名称、农药检出量(mg/kg)和标准判定结果.选取刻度缩进 s=0.8,收缩系数 r=0.4,张力系数 g=0.6.4 种颜色分别代表不同属性之间的连线,矩形颜色的深浅代表节点的存储顺序.通过选择不同国家

的标准,会生成不同的视图表示依照该国家的超标情况.深色线条连接的四边形表示该条数据有超标情况.通过点击选择视图①上的节点(以 D1 为例),右侧会出现放大视图②,会同时展示出矩形各条边所对应的值,视图③表示的是各种检出农药 MRL 的合格占比情况,如图 9 所示.其中的树图采用 Squarified 布局算法生成,每一个树图节点表示一种被检水果样品中农药残留的超标情况,包括水果名称、农药名称、农药检出量(mg/kg)和标准判定结果这 4 种属性.

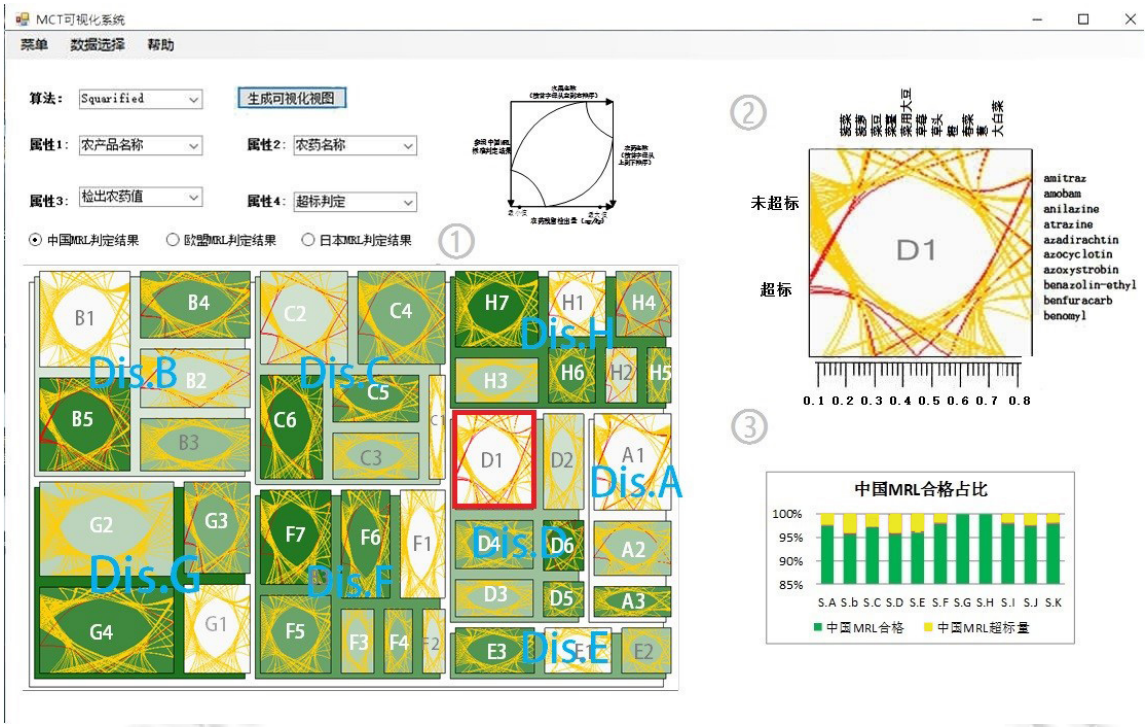


Fig.9 With MCT technology in seven areas of *Y* area were found in all fruit in the detection of pesticide residue levels of visual result

图 9 用 MCT 技术对 *Y* 地区 7 个区域中所有被检水果中检出农药残留超标情况的可视化结果

由图 9 可以分析出,不是每个超市都对 12 种水果进行采样,采样的水果品种数并不相等.面积大的矩形,即检出农药频次占比多的超市不一定存在超标样品数据,如超市 B1,B3;而面积小的矩形,即检出农药频次占比小的超市却可能存在超标样品数据,如超市 H2 和 H3.针对超标的数据进行分析,分别查看在中国、欧盟和日本 MRL 标准下的检测判定情况.通过分析柱状图可以发现:按照中国 MRL 标准,本批检测样品的合格率达 95%以上;日本 MRL 标准判定较为严格,但可以看出,检测区域的农产品也可以达到 86%以上的合格率.这证明我国水果中农药超标情况仍在控制范围.从树图中查看每一个超标数据在 4 个属性之间的连线,可以发现:在中国 MRL 标准下属于超标的农药残留量检测结果数据,在欧盟 MRL 和日本的 MRL 标准下也属于超标,反之则不然.同时,欧盟与日本的 MRL 标准也存在类似的情况.可以分析出:就 MRL 评价标准的严格性而言,再次证明了日本 MRL 标准最为严格,其次为欧盟 MRL 标准,中国 MRL 标准相对来说比较宽松.这说明我国的食品安全检测标准还需要加强.

4 评价与用户体验

为了对本文提到的 MCT 算法进行评估,本文拟采用 Y 城市 8 个区所有分店的抽样检测结果的可视化视图进行实验,并设计相关问卷.根据市场调查与检疫检验局相关人员沟通所知,人们较为关心的食品安全问题主要围绕在农药超标的农产品、超标农药以及超标量以及超标较为严重的区域.因此设计了问卷,见表 2.

测试过程如下:邀请本专业与其他专业学生各 15 名,首先给出 15 分钟对本可视化方法进行培训;然后对问卷的 10 个问题进行一一解答,记录回答者的答案与回答问题所用的时间;然后与正确答案进行比对,计算出平均准确率与平均所用时间.

Table 2 Questionnaire survey question table

表 2 问卷调查问题表

序号	测试问题
Q1	*Y* 城共有几个区,每个区有多少个超市?
Q2	*Y* 城哪个区检出农药超标率最严重?
Q3	*Y* 城哪个超市检出农药最多?
Q4	*Y* 城哪种农药被检出最多?
Q5	*Y* 城哪种农产品超标现象最严重?
Q6	超标最严重的农药是哪个?
Q7	哪些农产品检出农药居多,但是超标现象不明显?
Q8	检出农药最多的农产品是哪一个?
Q9	检出超标农药哪些具有共性(总是在同一种农产品检出)?
Q10	超标农药占比是否超过检出农药的一半(50%)?

图 10、图 11 为 15 个被测试者回答问题平均正确率与平均时间的统计结果,可以清楚地看出,每道问题正确率都在 80%以上,而回答每道题的平均时间都控制在 1 分钟.这说明在本实验中,MCT 算法能够有效地帮助使用者正确、快速地找到想要的答案,证明了本方案的有效性和实用性.

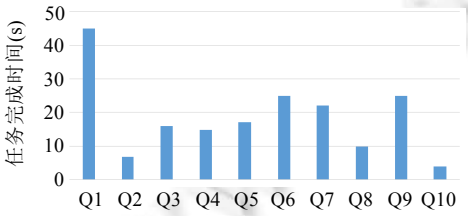


Fig.10 Task completion time (s) and the comparison result

图 10 任务完成时间(s)对比结果

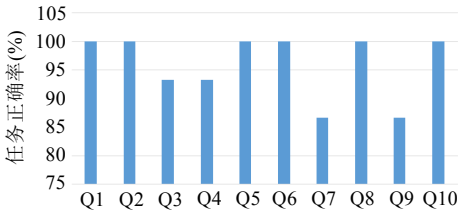


Fig.11 Task accuracy (%) and the comparison result

图 11 任务正确率/(%)对比结果

5 结论与未来的工作

本文针对具有层次结构和多维属性的复杂数据,提出了一种基于树图与平行坐标结合的 MCT 可视化技术,通过对树图中的矩形 4 条边的利用,提高了整体的空间利用率,使树图布局在显示数据的层次结构和节点权值大小的同时,可以展示更多维度的数据信息,从而可以实现如下的分析任务:(1) 比较层次结构中各节点权值的大小;(2) 分析各属性间的关联;(3) 根据多维属性比较分析各节点的综合指标,发现异常.通过使用由收缩系数和张力系数调节的基数样条插值曲线进行不同属性点之间的连接,用户可以通过对系数的调节,改变曲线的收缩和弯曲程度,有助于改善边杂乱的情况,使布局效果更有利于可视分析.

将 MCT 技术应用于水果蔬菜中农药残留检出情况和超标情况的可视分析,根据分析需要,此应用选择 Squarified 和 Strip 树图布局算法来生成树图,用树图中节点矩形表示农药残留检出量和超标量以及相关属性,在展示水果蔬菜中农药残留检出和超标情况按地域层次结构分布的同时,还展示了各地区检出农药残留的农产品名称、残留量、超标情况等属性信息,实现了帮助用户:(1) 比较各地区农药残留检出情况;(2) 分析检出农药残留较多的异常农产品和异常地区等功能.用户体验结果表明,MCT 技术在可视分析具有层次结构的多维属性数据上是有效的.

MCT 技术虽然解决了具有层次结构的多维属性数据的可视分析问题,但它仍有如下局限性:(1) 当节点中多维属性的数据量很大时,会出现线条杂乱问题;(2) 当节点中的属性超过四维时,目前没有给出解决方案.针对

这些局限,未来我们将采用边绑定的方法[30]根据聚类将同类属性联系捆绑解决边杂乱的问题;通过属性过滤降维、多边形树图[31,32]等方式解决属性维多于四维的情况.

References:

[1] Chung H, Cho YJ, Self J, North C. Pixel-Oriented treemap for multiple displays. In: Proc. of the IEEE Conf. on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2012. 289−290. [doi: 10.1109/vast.2012.6400512]

[2] Fenstermaker WH, Day AC. Method for visual presentation of key performance indicators of a business utilizing a squarified tree map including selectively displaying additional node data: U.S. Patent 8,660,887. 2014.

[3] Zhang X, Yuan XR. Treemap visualization. Journal of Computer-Aided Design & Computer Graphics, 2012,24(9):1113−1124 (in Chinese with English abstract).

[4] Zhao Y, Liang X, Fan XP, Wang YW, Yang MJ, Zhou FF. MVSec: Multi-Perspective and deductive visual analytics on heterogeneous network security data. Journal of Visualization, 2014,17(3):181−196. [doi: 10.1007/s12650-014-0213-6]

[5] Tatu A, Albuquerque G, Eisemann M, Bak P. Automated analytical methods to support visual exploration of high-dimensional data. IEEE Trans. on Visualization and Computer Graphics, 2011,17(5):584−597. [doi: 10.1109/tvcg.2010.242]

[6] Inselberg A. The plane with parallel coordinates. The Visual Computer, 1985,1(2):69−91. [doi: 10.1007/BF01898350]

[7] Liu W, Wang B, Yu JX, LI F, Wang SX, Hong WX. Visualization classification method of multi-dimensional data based on radar chart mapping. In: Proc. of the 2008 Int'l Conf. on Machine Learning and Cybernetics. IEEE, 2008. 857−862. [doi: 10.1109/icmlc.2008.4620524]

[8] Byron L, Wattenberg M. Stacked graphs-geometry & aesthetics. IEEE Trans. on Visualization and Computer Graphics, 2008,14(6): 1245−1252. [doi: 10.1109/tvcg.2008.166]

[9] Chen Y, Zhang XY, Feng YC, Liang J, Chen HQ. Sunburst with ordered nodes based on hierarchical clustering: a visual analyzing method for associated hierarchical pesticide residue data. Journal of Visualization, 2015,18(2):237-254 [doi: 10.1007/s12650-014-0269-3]

[10] Zhao S, McGuffin MJ, Chignell MH. Elastic hierarchies: Combining treemaps and node-link diagrams. In: Proc. of the IEEE Symp. on Information Visualization (INFOVIS 2005). IEEE, 2005. 57−64. [doi: 10.1109/infvis.2005.1532129]

[11] Huang ML, Huang TH, Zhang J. Treemapbar: Visualizing additional dimensions of data in bar chart. In: Proc. of 2009 the 13th Int'l Conf. on Information Visualisation. IEEE, 2009. 98−103. [doi: 10.1109/iv.2009.22]

[12] Johnson B, Shneiderman B. Tree-Maps: A space-filling approach to the visualization of hierarchical information structures. In: Proc. of the IEEE Visualization. Los Alamitos: IEEE Computer Society Press, 1991. 284−291. [doi: 10.1109/visual.1991.175815]

[13] Bruls M, Huizing K, Van Wijk JJ. Squarified Treemaps. Heidelberg: Springer-Verlag, 2000. 33−42. [doi: 10.1007/978-3-7091-6783-0_4]

[14] Chen Y, Jia YJ, Sun YH. A squarified treemap layout algorithm based on sorting by parts. Journal of Computer-Aided Design & Computer Graphics, 2013,25(5):731−737 (in Chinese with English abstract).

[15] Shneiderman B, Wattenberg M. Ordered treemap layouts. In: Proc. of the IEEE Symp. on Information Visualization. Los Alamitos: IEEE Computer Society Press, 2001. 73−78. [doi: 10.1109/infvis.2001.963283]

[16] Bederson BB, Shneiderman B, Wattenberg M. Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. ACM Trans. on Graphics, 2002,21(4):833−854. [doi: 10.1145/571647.571649]

[17] Tu Y, Shen HW.Visualizing changes of hierarchical data using treemaps. IEEE Trans. on Visualization and Computer Graphics, 2007,13(6):1286−1293. [doi: 10.1109/tvcg.2007.70529]

[18] Chen Y, Hu HY, Li ZL. Performance compare and optimization of rectangular treemap layout algorithms. Journal of Computer-Aided Design & Computer Graphics, 2013,25(11):1623−1634 (in Chinese with English abstract).

[19] Wood J, Dykes J. Spatially ordered treemaps. IEEE Trans. on Visualization and Computer Graphics, 2008,14(6):1348−1355. [doi: 10.1109/tvcg.2008.165]

[20] Hu HY, Chen Y, Zhen YG, Liu RJ. A squarified and ordered treemap layout algorithm. Journal of Computer-Aided Design & Computer Graphics, 2014,26(10):1703−1710 (in Chinese with English abstract).

[21] Tominski C, Abello J, Schumann H. Axes-Bbased visualizations with radial layouts. In: Proc. of the 2004 ACM Symp. on Applied Computing. ACM Press, 2004. 1242−1247. [doi: 10.1145/967900.968153]

[22] Claessen JHT, Van Wijk JJ. Flexible linked axes for multivariate data visualization. IEEE Trans. on Visualization and Computer Graphics, 2011,17(12):2310−2316. [doi: 10.1109/tvcg.2011.201]

[23] Fua YH, Ward MO, Rundensteiner EA. Hierarchical parallel coordinates for exploration of large datasets. In: Proc. of the Conf. on Visualization'99: Celebrating Ten Years. IEEE Computer Society Press, 1999. 43−50. [doi: 10.1109/VISUAL.1999.809866]

[24] Zhou H, Yuan X, Qu H, Cui WW, Chen BQ. Visual clustering in parallel coordinates. Computer Graphics Forum, 2008,27(3): 1047−1054. [doi: 10.1111/j.1467-8659.2008.01241.x]

[25] Guo P, Xiao H, Wang Z, Yuan XR. Interactive local clustering operations for high dimensional data in parallel coordinates. In: Proc. of the 2010 IEEE Pacific Visualization Symp. (PacificVis). IEEE, 2010. 97−104. [doi: 10.1109/PACIFICVIS.2010.5429608]

[26] Slingsby A, Dykes J, Wood J. Exploring uncertainty in geodemographics with interactive graphics. IEEE Trans. on Visualization and Computer Graphics, 2011,17(12):2545−2554. [doi: 10.1109/TVCG.2011.197]

[27] Linsen L, Behrendt S. Linked treemap: A 3D treemap-nodelink layout for visualizing hierarchical structures. Computational Statistics, 2011,26(4):679−697. [doi: 10.1007/s00180-011-0272-2]

[28] Vliegen R, van Wijk JJ, Van der Linden EJ. Visualizing business data with generalized treemaps. IEEE Trans. on Visualization and Computer Graphics, 2006,12(5):789−796. [doi: 10.1109/tvcg.2006.200]

[29] Kobayashi A, Misue K, Tanaka J. Edge equalized treemaps. In: Proc. of 2012 the 16th Int'l Conf. on Information Visualization (IV). IEEE, 2012. 7−12. [doi: 10.1109/iv.2012.12]

[30] Holten D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Trans. on Visualization and Computer Graphics, 2006,12(5):741−748. [doi: 10.1109/tvcg.2006.147]

[31] Balzer M, Deussen O, Lewerentz C. Voronoi treemaps for the visualization of software metrics. In: Proc. of the 2005 ACM Symp. on Software visualization. ACM Press, 2005. 165−172. [doi: 10.1145/1056018.1056041]

[32] Sud A, Fisher D, Lee HP. Fast dynamic voronoi treemaps. In: Proc. of the 2010 Int'l Symp. on Voronoi Diagrams in Science and Engineering (ISVD). IEEE, 2010. 85−94. [doi: 10.1109/isvd.2010.16]

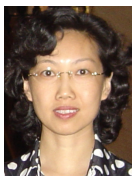
附中文参考文献:

[3] 张昕,袁晓如.树图可视化.计算机辅助设计与图形学学报,2012,24(9):1113−1124.

[14] 陈谊,贾艳杰,孙悦红.分块排序的正方化树图布局算法.计算机辅助设计与图形学学报,2013,25(5):731−737.

[18] 陈谊,胡海云,李志龙.树图布局算法的比较与优化研究.计算机辅助设计与图形学学报,2013,25(11):1623−1634.

[20] 胡海云,陈谊,甄远刚,刘瑞军.一种正方化有序树图布局算法.计算机辅助设计与图形学学报,2014,26(10):1703−1710.



陈谊(1963－),女,北京人,博士,教授,CCF高级会员,主要研究领域为信息可视化,可视分析,食品安全数据分析.

梁婕(1981－),女,博士,助理研究员,CCF专业会员,主要研究领域为数据可视化,可视分析.

甄远刚(1989－),男,硕士生,CCF 学生会员,主要研究领域为信息可视化,可视分析,食品安全数据分析.

Kwan-Liu MA(1959－),男,博士,教授,博士生导师,主要研究领域为可视化,计算机图形学,高性能计算,用户界面设计.

胡海云(1988－),男,硕士,主要研究领域为信息可视化,可视分析,食品安全数据分析.