

6-5-2020

PCNL: A Visual Analysis Method for Comparing Homogeneous Hierarchical Data

Dong Yu

Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;

Chen Yi

Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;

Yuehong Sun

Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;

Follow this and additional works at: <https://dc-china-simulation.researchcommons.org/journal>



Part of the [Artificial Intelligence and Robotics Commons](#), [Computer Engineering Commons](#), [Numerical Analysis and Scientific Computing Commons](#), [Operations Research, Systems Engineering and Industrial Engineering Commons](#), and the [Systems Science Commons](#)

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

PCNL: A Visual Analysis Method for Comparing Homogeneous Hierarchical Data

Abstract

Abstract: According to the requirements of comparing homogeneous hierarchical data in different fields, *a visual analysis method called PCNL is presented in this paper*. Homogeneous information is extracted from the multiple levels of the heterogeneous hierarchical data set, *and isomorphic node-link trees are constructed*. Parallel coordinates are combined with node-link trees, *which displays attribute values of leaf nodes in multiple isomorphic trees*. *Interactive techniques are introduced to help users to compare and analyze multiple attributes of each leaf node in multiple homogeneous hierarchical structures*. Applying this method to compare multi-national MRL data in food safety field can help users to analyze the general trend and detail difference of the multinational MRL effectively, and find out the existing problems of Chinese MRL standards.

Keywords

information visualization, parallel coordinates, homogeneous hierarchical data, MRL standards, food safety

Recommended Citation

Dong Yu, Chen Yi, Sun Yuehong. PCNL: A Visual Analysis Method for Comparing Homogeneous Hierarchical Data[J]. Journal of System Simulation, 2017, 29(11): 2700-2708.

一种同构层次数据的对比可视分析方法 PCNL

董禹, 陈谊*, 孙悦红

(北京工商大学计算机与信息工程学院食品安全大数据技术北京市重点实验室, 北京 100048)

摘要: 针对不同领域中同构层次数据的对比分析需求, 提出了一种同构层次数据的对比可视分析方法 PCNL。对多个异构层次数据集进行对比统计, 抽取同构信息, 用节点-链接法构建同构树; 将平行坐标与同构树结合, 展示多个同构树中叶子节点的属性值, 通过对节点进行布局形成多个同构层次数据的平行坐标可视化表达; 通过交互手段可有有效的对比分析多个同构层次结构中每个叶子节点的多个属性。将该方法应用于食品安全领域的多国 MRL 标准数据的对比分析, 可帮助用户有效地对比分析多国 MRL 的总体趋势和细节差异, 找出中国 MRL 标准的问题。

关键词: 信息可视化; 平行坐标; 同构层次数据; MRL 标准数据; 食品安全

中图分类号: TP391

文献标识码: A

文章编号: 1004-731X (2017) 11-2700-09

DOI: 10.16182/j.issn1004731x.joss.201711015

PCNL: A Visual Analysis Method for Comparing Homogeneous Hierarchical Data

Dong Yu, Chen Yi, Sun Yuehong

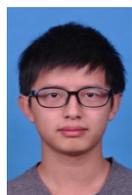
(Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, 100048, China)

Abstract: According to the requirements of comparing homogeneous hierarchical data in different fields, a visual analysis method called PCNL is presented in this paper. Homogeneous information is extracted from the multiple levels of the heterogeneous hierarchical data set, and isomorphic node-link trees are constructed. Parallel coordinates are combined with node-link trees, which displays attribute values of leaf nodes in multiple isomorphic trees. Interactive techniques are introduced to help users to compare and analyze multiple attributes of each leaf node in multiple homogeneous hierarchical structures. Applying this method to compare multi-national MRL data in food safety field can help users to analyze the general trend and detail difference of the multinational MRL effectively, and find out the existing problems of Chinese MRL standards.

Keywords: information visualization; parallel coordinates; homogeneous hierarchical data; MRL standards; food safety

引言

在很多领域存在着对比分析多个层次数据或



收稿日期: 2016-5-31 修回日期: 2016-10-28;
基金项目: 国家科技支撑计划(2012BAD29B01-2),
国家科技基础性工作专项(2015FY111200), 北京市科
技计划(Z161100001616004);
作者简介: 董禹(1993-), 男, 北京, 硕士生, 研究方
向为信息可视化与可视分析。

多棵树的复杂分析需求, 如在经济领域中对比分析多个国家各领域的生产总值贡献, 在食品安全领域中对比分析多个国家各类农产品中农药和食品添加剂的限量值标准等。由于这类数据既具有层次结构又具有多维属性, 使用传统的手工对比或表格工具对比方法费时费力、效率低下、直观性差, 而且当数据维度和数据量增大时, 很难分析多维间的复杂关系。近年来出现的信息可视化与可视分析充分

<http://www.china-simulation.com>

• 2700 •

利用人的视觉高通道认知能力,以图的形式将复杂数据展现出来,辅以交互手段将人的智能与计算机的超强处理能力有机融合,为数据分析和科学决策提供了新的方法和有效手段。

本文以世界多国或地区组织(下简称“多地区”)食品中农药残留最大限量标准(MRL, Maximum Residue Limit)对比分析为需求,提出了一种基于平行坐标和节点-链接树的同构层次数据对比可视分析方法 PCNL(Parallel Coordinates and Node-Link),兼顾层次结构的展示和多维数据的对比分析,较好地解决了同构层次数据的对比分析问题。将 PCNL 方法应用于世界多地区农药残留 MRL 标准的对比分析,实现了对比单一或各类农产品对应的各地区 MRL 标准数值,挖掘同类农产品对应的各国 MRL 设定模式,有助于找出中国 MRL 标准的优势和存在的问题。

1 相关工作

1.1 平行坐标

平行坐标是一种典型高维多元数据的可视化方法^[1],它将多维数据在两维平面显示出来,通过交互人们可以更容易地发现数据背后隐含的规律^[2]。1985 年 Inselberg 提出了多维数据的平行坐标表示法,并将其应用于计算领域^[3]。1990 年美国统计学家 Wegman 将平行坐标应用于多维数据分析^[4],至此,平行坐标开始走进多维数据统计领域。如今,平行坐标可以和其它几种高维数据可视化方法融合,例如:直方图,散点图矩阵^[5]等。多视图联动分析高维数据已在信息可视化^[6]、可视分析^[7]、数据挖掘^[8]等领域得到广泛使用。

1.2 节点-链接树

节点-链接树是一种典型的用来表示层次数据的可视化方法,它用节点表示实体,用节点之间的连线表示个体之间的层次关系。代表技术有空间树^[9],圆锥树^[10]等。这种方法直观清晰,擅长表示节点间的层次关系。但当节点数目过多,特别是节

点-链接树形结构的广度和深度相差较大时,可读性较差^[11]。故选择合适的广度和深度有助于更好地发挥节点-链接法的优点。2015 年,Chen 等使用放射环结合节点-链接树的可视化方法,在显示农药、农产品层次结构的同时,探索两者之间的关系,取得了良好效果^[12]。这说明节点-链接树方法可以发挥自身优点融合其他可视化方法。

平行坐标可以良好地展示多维数据,但在层次数据展示方面却表现不佳^[5];节点-链接法可以清楚地显示层次数据,但在数据对比方面略有欠缺。因此,如何将节点-链接树和平行坐标相结合,在兼顾显示层次信息的同时进行数据对比值得深入研究。

1.3 MRL 标准数据

农药最大残留限量(MRL, Maximum Residue Limit)是指法定的某种农药在某种农产品中残留的最高允许浓度,以每千克采样农产品中农药残留的毫克数(mg/kg)计量^[13]。

MRL 标准是衡量食品安全水平的重要标准,其制定及修改是基于科学的风险评估数据,在国际食品法典委员会(Codex Alimentarius Commission,简称 CAC)、欧盟、美国和日本等国际组织及国家受到广泛关注。各国针对本国农产品、农药制定的 MRL 标准不仅从一定程度上反映了一个国家对食品中农药残留的监管水平,还制约着国家之间的贸易。中国 MRL 标准的制定起步晚,存在标准数量少,数据不严格等问题,所以,对比分析我国 MRL 标准与世界各国或地区组织的 MRL 标准的差异,完善我国 MRL 标准是一个亟待解决的问题。

2 PCNL 算法原理

由于在现实生活中层次数据大多是异构的,故 PCNL 算法首先需对要对比的多棵异构树进行统计并抽取同构信息,构造同构树;其次对同构树的节点进行布局;然后进行可视化;最后实现交互分析。PCNL 算法的流程如图 1 所示。

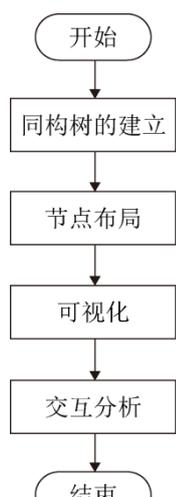


图 1 PCNL 算法流程图
Fig.1 Flowchart of PCNL algorithm

2.1 同构树的建立

首先将要对比的多个异构层次数据集进行对比统计, 构建多个空间树形结构, 然后通过整合和减枝, 抽取其中的同构信息, 形成同构树, 其过程如图 2 所示。

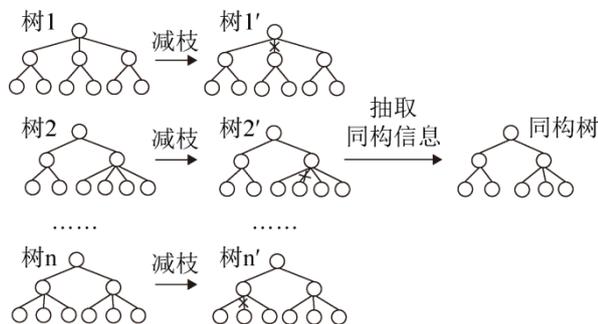


图 2 同构树的建立过程
Fig.2 Establishment process of the homogeneous trees

2.2 PCNL 布局算法

可视化布局效果对用户从中获得有价值的信息起着十分重要的作用。在层次数据可视化中, 用户的注意力往往集中在图中的中心区域^[3]。为减少数据交叠, 使布局更加美观, 需对布局进行重新设计。

本文将与根节点相连的一族子树称为一个子树簇。统计同构树中每个子树簇的叶子节点个数, 从大到小排序。将叶子节点个数最多的子树簇(称

为第一个子树簇)将其布局在图形中央, 剩余的子树簇按照叶子节点个数从大到小依次排列在第一个子树簇的上下两端, 具体布局如下图 3 所示。

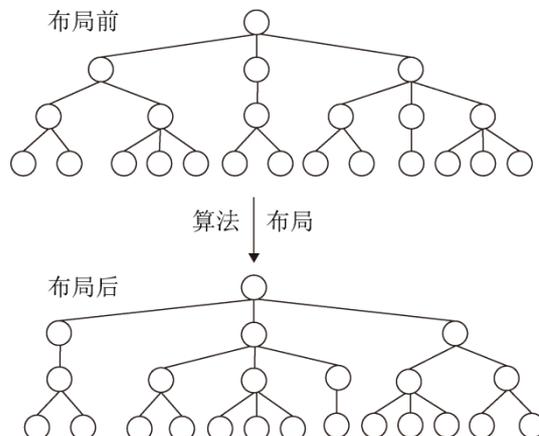


图 3 PCNL 算法布局结果示意图
Fig.3 PCNL algorithm layout result figure

在描述本算法前, 先定义如下参数:

假设: n 为同构树中叶子节点的总数; m 为同构树中子树簇的总数; k 为同构树的深度。

定义 S_{ij} 为第 i 颗子树中第 j 个叶子节点的位置(在平行坐标系中的纵坐标), 其中 i 表示子树的编号, j 表示叶子节点的编号, 如 S_{1a} 表示第 1 颗子树中第 a 个叶子节点的位置; 设 S_h 表示当前已布局叶子节点的上边界位置; S_l 表示当前已布局叶子节点下边界的位置。

定义 $H_{i'j'}$ 为第 i' 颗子树中第 j' 个子节点位置, 其中 i' 表示子树的编号, j' 表示子节点的编号。

定义 $F(H_{i'j'}, k)$ 为位于 k 层的第 i' 颗子树中第 j' 个子节点的定位函数。

具体布局算法描述如下:

Step1: 将坐标轴分为 n 等份, 方便度量。将第 1 个子树簇中的所有叶子节点按照其子树编号。初始化 S_{11} 的位置为:

$$S_{11} = \begin{cases} \frac{n+1}{2}, n \text{ 为奇数} \\ \frac{n}{2}, n \text{ 为偶数} \end{cases}$$

Step2: 布局其余的叶子节点, 直至遍历该子树的所有叶子节点:

$$S_{l(a+1)} = \begin{cases} S_{la} + 1, n \text{ 为奇数} \\ S_{la} - 1, n \text{ 为偶数} \end{cases}$$

Step3: 定义该子树簇中其余子树的叶子节点位置:

$$S_{ij} = \begin{cases} S_h + 1, i \text{ 为奇数} \\ S_l - 1, i \text{ 为偶数} \end{cases}$$

Step4: 计算上一层父节点位置的定位函数 $F(Hi'j', k)$:

$$H_{i'j'} = \frac{\sum_{j=1}^p S_{ij}}{p}$$

Step5: 令 $Hi'j'$ 为新的子节点, 重复 step4, 逐层向上递归计算, 直至遍历整个子树簇。

Step6: 重复 step3 计算新的叶子节点位置, 遍历一个完整的子树簇后, 重复 step4, step5, 直至形成新的子树簇。

Step7: 重复 step6, 直至 m 个子树簇被遍历完, 输出结果。

该布局算法有三个优点:

(1) 将叶子节点较多的子树布局在图形中央, 符合美学标准, 使用户的注意力集中在可视化结果的主要区域;

(2) 避免了层次结构中线条的交叉, 使分类的层次更加清晰;

(3) 新加入的子树布局在图形两侧, 不会引起整体结构的改变。

2.3 同构层次结构的平行坐标可视化表达

节点-链接树可以清晰的展现数据的层次结构, 但很难对比分析各节点的属性值。为此, 引入平行坐标, 将平行坐标的列抽象显示为不同层次结构叶子节点的值, 如图 4 所示。平行坐标中第一列的值代表第一棵树中各叶子节点的属性值, 以此类推, 第 n 列的值代表第 n 棵树中各叶子节点的属性值, 这 n 棵树为同构的。最靠近右端的一列用来显示不同数据的属性值。这样布局兼顾了展现树的层次结构, 又考虑了对多棵同构树叶子节点属性值的对比分析。

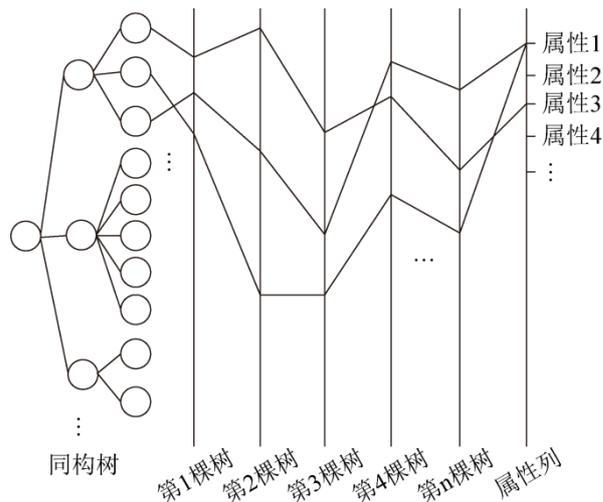


图 4 同构层次结构的平行坐标可视化表达
Fig.4 Visual expression of homogeneous hierarchy combined with parallel coordinates

3 PCNL 算法在多地区 MRL 标准对比分析中的应用

使用 PCNL 方法对各国 MRL 标准数据进行可视化。通过简单的对比和交互手段可以较为容易地对 MRL 标准数据进行评估, 得出相应结论。

3.1 MRL 标准数据的预处理

为对比各国 MRL 标准中对同一农产品中各种农药的限量值, 需先筛选出在各国 MRL 标准中都存在 MRL 值的农产品, 即抽取同构信息。本文通过对不同国家 MRL 标准数据进行数据整合, 筛选出符合要求的农产品。如表 1 中的葡萄就是一种, 过滤出葡萄在各个农药中的 MRL 限量值。除日本外, 未规定的值用 0 标识, 日本用 0.01 标识(日本的“一律标准”规定, 在标准中未规定的的数据均认为 MRL 检出量为 0.01mg/kg)。表 1 为葡萄中 7 项农药残留最大限量表。

3.2 农产品同构树的构建

本文只关注中国与各国农产品同构树的 MRL 标准值对比, 也尝试找出不同分类农产品的 MRL 值是否有关联性, 因此需要将农产品按分类信息显示。2015 年李等提出了农产品分类信息的目录层次结构^[14], 提供了一种四层的层次结构来对所有

的农产品进行分类。由于农产品分类比较复杂，构建树形结构时子树数量较多，故本文在此基础上用多叉赫夫曼树来划分各国的所有农产品，如图 5 所示，并对所有节点用赫夫曼编码准确映射，获得农产品编码，如表 2。考虑到同一层中有个数大于 9 的情况出现，所以使用两位数字来对一个节点编码。

表 1 葡萄中 7 种农药残留最大限量

Tab. 1 Maximum limit of 7 pesticide residues in grape						
农药名	中国	香港	美国	日本	欧盟	CAC
腈菌唑	1	1	1	1	1	1
多菌灵	3	3	0.01	3	0.3	3
甲霜灵	1	1	2	1	2	1
腈苯唑	1.5	1.5	1.5	1.5	1.5	1.5
烯酰吗啉	5	2	3	0.01	3	2
虫酰肼	2	2	3	2	3	2
戊唑醇	2	2	3	0.5	2	3

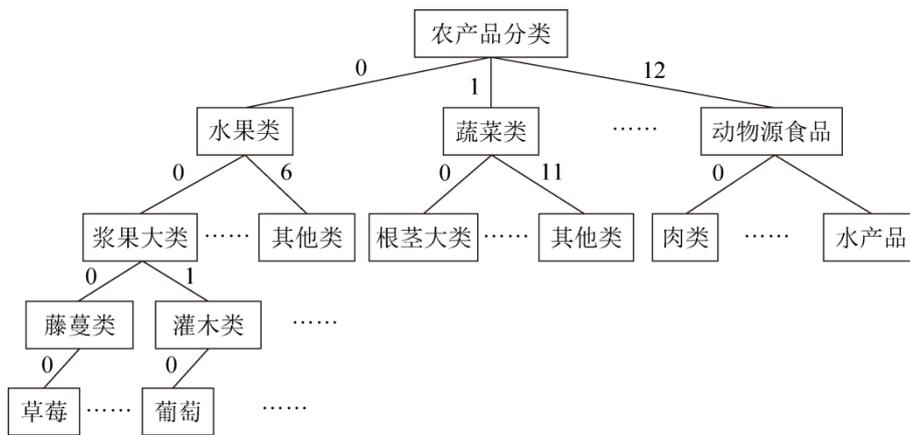


图 5 农产品分类信息的多叉赫夫曼树

Fig.5 Multiway Huffman tree of agriculture product's classification information

表 2 农产品分类编码举例

Tab. 2 Examples of agricultural products' classified coding		
农产品名	农产品分类	农产品编码
草莓	水果; 浆果大类; 藤蔓类;	00000000
葡萄	水果; 浆果大类; 灌木类;	00000100
梨	水果; 仁果大类; 仁果类	00020000
苹果	水果; 仁果大类; 仁果类	00020001

本文设计的农产品分类编码，有助于统计叶子节点个数和算法布局。使各国农产品同构树的构建变得简单易行。

3.3 颜色设计

颜色选择是可视化重要部分，好的颜色选择可以让用户更直观地了解可视化内容。本文对比研究 MRL 标准，只关注中国与其它国家的差距。因此，我们使用从红到蓝的颜色代表中国 MRL 限量值的大小。如图 6、图 7、图 8 所示的图例中，色带颜色越红，代表中国 MRL 限量值越高；反之代

表 MRL 值越低，取值范围为 0.01~20。

3.4 可视化及交互

依照本文 2.2 节中所述算法，以各国 MRL 标准数据中常见的水果为例，对以水果为根节点的农产品同构树进行布局，加入平行坐标显示农药信息和各国 MRL 标准值，如图 6 所示。可以看出各国 MRL 标准值的总体趋势，在日本和欧盟处有明显降低。

以各国 MRL 标准数据中常见的水果为例，左端的水果分类呈现类树形结构，布局算法保证了各个子树的叶子不交叉，较好地表现了水果分类的层次结构，右端显示了某种水果在各国 MRL 标准中的检出值。如图 6 所示，能够清晰地看出各国 MRL 标准的变化趋势。中国 MRL 标准限量值普遍高于日本和欧盟，与香港、美国、CAC 基本持平。

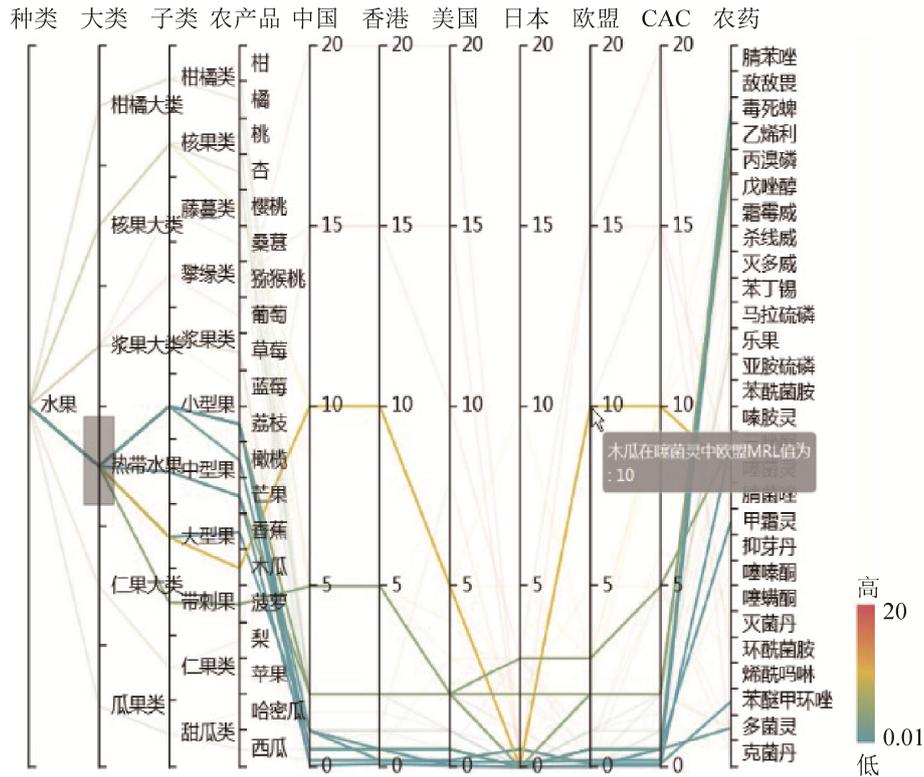


图 8 筛选出热带水果大类中各种农药的 MRL 值
Fig.8 MRL values distribution of fruit by filtering tropical fruit in 6 regions

好的交互手段能够帮助用户容易地对农产品及其分类进行 MRL 值的对比分析。如图 7 所示，用户可以通过颜色编码直接看出不同种类的水果在中国 MRL 标准中规定的大小，有利于对比分析中国 MRL 标准和各国 MRL 标准之间的差异。还可以通过鼠标拖拽的交互手段筛选各类水果信息。如图 8 所示，该图为筛选出的热带水果大类下所有的农产品信息。可以看出，除木瓜之外，大中小型果整体 MRL 值在各国的 MRL 标准中均偏低，只有木瓜达到了 10。这种对同构树的对比分析有助于用户找到同类农产品的不同 MRL 标准分布情况与趋势，与不同农产品在相同 MRL 标准下不同限量值的直观表达，有助于发掘农产品分类之间 MRL 限量值的关系，帮助领域人员进行相关决策。

4 用户评价

可视化图形的评价是用户体验的重要环节，主要有两个方面。其一，功能性^[15]，其二，美学标准。功能性指的是用户是否能够迅速了解可视化内

容并快速获取结果。本文采用平行坐标结合节点-链接的方法表现数据层次结构，同时通过交互手段使用户可以快速地了解 MRL 标准限量值的大小，充分满足了可视化的功能性。

为了对本文的方法进行评估，对水果中 MRL 标准数据用改进的平行坐标进行可视化，邀请了 20 名用户测试。实验过程设计如下：首先给出 30 分钟供测试者对该方法进行熟悉；然后，要求每位测试者通过查看平行坐标可视化结果，回答问题列表(表 3)所提的问题。记录问题完成的效率和正确率，实验结果如图 8 和图 9 所示。

图 9 为回答表 3 中 8 个问题这 20 个测试者所用的平均时间，所有任务均在 30 s 内完成。其中，Q2 到 Q6 均小于 10 s。这体现了这种基于平行坐标的同构层次数据对比可视分析方法不仅能够帮助用户高效地理解数据，在短时间内完成对简单信息的查找和筛选，而且可以快速地对分析同构层次数据下各个叶子节点的值。图 10 为回答表 3 中 8

个问题这 20 名测试者的正确率。可以看到, 8 个问题的正确率均在 90% 以上, 说明使用这种改进的平行坐标时, 用户使用不易出错, 功能性良好。

表 3 用户评价问题列表
Tab. 3 Question list of user evaluation

序号	问题
Q1	水果分类有几级? 各级又含多少子类或多少农产品?
Q2	哪个国家的 MRL 限量值普遍偏低?
Q3	热带水果中, 哪种农药的限量值偏高?
Q4	图中农药苯丁锡的 MRL 标准数据共有几条?
Q5	对比木瓜在噻菌灵中中国与日本 MRL 的大小。
Q6	中国 MRL 值在 10 至 20 之间的数据有几条?
Q7	农药乙烯利在农产品樱桃中, 各国 MRL 标准的限量值是多少? 最高的和最低的分别是哪个国家?
Q8	对比分析中国浆果大类下的所有农产品与其他国家的差距。

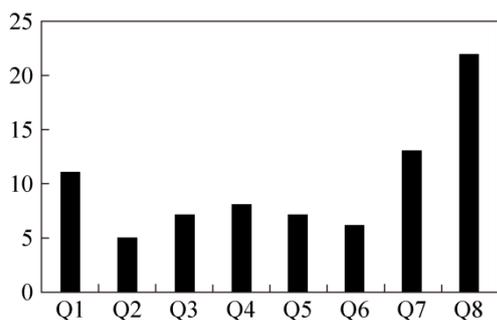


图 9 任务完成时间(单位: s)
Fig.9 Task completion time (unit: s)

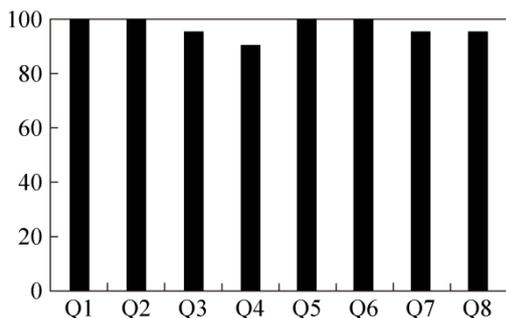


图 10 任务正确率(单位: %)
Fig.10 Task accuracy (unit: %)

美学标准是指可视化图形布局是否合理, 视觉设计是否符合用户的视觉体验。本文的可视化通过颜色分明的布局设计, 将主要的可视化结果集中显示在图形中央, 较好地利用了空间, 符合美学标准。

5 结论

本文将平行坐标和节点-链接树相结合, 辅以相应的交互手段, 提出了一种同构层次数据的对比可视分析方法 PCNL。该方法具有如下特点:

- (1)使用节点-链接法清晰的展现了同构树的层次结构;
- (2)使用平行坐标清晰的展现了多棵同构树中各节点的属性值;
- (3)支持对各节点属性值的对比分析, 包括最大和最小值、属性值高低的分布模式和变化趋势、属性之间的关联等等;
- (4)通过交互和颜色区别支持快速筛选感兴趣的属性, 高亮显示。

将该方法应用于世界各国(地区)的农药残留最大限量 MRL 标准的对比分析, 大大提高了分析的效率和精准度, 应用结果表明, 这种对比可视分析方法能够对世界多国(地区)的 MRL 标准进行深度对比分析, 效果良好。

在未来的研究工作中, 需要研究数据规模增大时带来的若干问题。如当子节点数量继续增大时, 可加入树的折叠和展开交互方法, 按子树逐个分析; 也可加入其它交互技术, 例如 Link Sliding^[16](链路滑动)和鱼眼视图^[17]等, 解决节点规模问题。

参考文献:

- [1] 翟旭君, 李春平. 平行坐标及其在聚类分析中的应用[J]. 计算机应用研究, 2005, 22(8):124-126.
- [2] Yuan X, Guo P, Xiao H, et al. Scattering points in parallel coordinates[J]. Visualization & Computer Graphics IEEE Transactions on(S1077-2626), 2014, 15(6):1001-1008.
- [3] Inselberg A. The plane with parallel coordinates[J]. The Visual Computer(S0178-2789),1985,1(4):69-91.
- [4] Wegman E. Hyper-dimensional data analysis using parallel coordinates[J]. Journal of the American Statistical Association(S0162-1459), 1990,85(411):664-675.
- [5] 陈谊, 蔡进峰, 石耀斌, 等. 基于平行坐标的多视图协同可视分析方法[J]. 系统仿真学报, 2013, 25(1): 81-86.

