基于网络大数据分析的可视化技术

北京工商大学　翁彬月，黄今慧

（北京工商大学计算机与信息工程学院，北京 100048）

本课题得到国家级大学生科学研究与创业行动计划项目资助。

论文发表于2017年第23期《信息与电脑》，刊号：ISSN 1003-9767

指导教师：黄今慧副教授

摘要：随着大数据产业的快速发展，每天全球产生EB级的数据流量。面对这些数据量大，产生速度快，维度高、多来源、多结构的大数据信息，研究通过借助人机交互式分析方法，运用可视化分析技术充分挖掘大数据背后的关键信息，为人们提供更为直观高效的数据认知。本文首先探讨大数据分析目前最为关注的多维数据可视化、层次数据可视化和时序数据可视化的技术发展。其次，指出在大数据可视化分析领域面临的数据规模大、维度高、多源异构等的瓶颈问题，以及对动态数据进行实时分析可视化和数据预测分析的技术挑战。

关键词：大数据分析；可视化；多维数据；层次数据；时序数据

Abstract : With the rapid development of big data industry, EB level data traffic is generated worldwide. Faced with the large amount of data, high-speed, high-dimensional, multi-source, multi-structure of large data information, research through the use of human-computer interactive analysis method, the use of visual analysis technology to fully excavate the key information behind the large data, to provide more intuitive and efficient data cognition. Firstly, this paper discusses the development of multi-dimensional data visualization, hierarchical data visualization and sequential data visualization, which are the most concerned technologies in large data analysis. Secondly, it points out the bottlenecks of large-scale, high-dimensional, multi-source and heterogeneous data in the field of large-scale data visualization analysis, and the technical challenges of real-time analysis visualization and data prediction analysis of dynamic data.

Key words: big data analysis; visualization; multidimensional data; hierarchical data; time series data

一 引言

随着大数据云时代的来临，各种智能移动设备、传感器、穿戴设备等的兴起，大众产业向着数字化、信息化快速发展，在世界互联网上每时每刻都会产生结构类型各异且数量庞大的数据。这样产生的海量的数据流，IBM公司提出大数据的5V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）。

然而，大数据能力在企业应用时，需要以非常简单易用的方式来呈现，才能让更多的数据用户使用。对于现在大部分的企业数据用户（往往是业务、产品、营销负责人等非大数据专业人士）而言，网络数据具有结构复杂，数据量庞大，产生速度快，关键信息分散等难点，使信息使用者理解起来非常困难，不利于对数据信息的使用。同时，海量的网络数据无法直接分析，降低数据的使用效率。但是，通过大数据分析可视化技术的呈现，凸显数据之间的关键联系或差异，就可以使这些企业数据用户能更容易、更快速地从中获得想要的信息。

可视化技术起源于20世纪80年代出现的科学计算可视化（Visualization in Scientific Computing），它是指利用计算机图形学、计算机图像处理、计算机信号处理等方法对数据、信息、知识的内在结构进行表达[1]。同样的，在大数据背景下的信息可视化技术则是利用人眼的感知能力和人脑的智能，对数据进行交互的可视表达，从而达到增强认知的一门学科[2]。Shneiderman 根据信息的特征把信息可视化技术分为[1]一维信息(1-dimensional)、二维信息(2-dimensional)、三维信息(3-dimensional)、多维信息(multidimensional)、

作者简介：翁彬月（1996-），女，贵州人，本科生，主要研究领域为大数据可视化分析。

黄今慧（1966-），女，副教授，长沙人，主要研究领域为数据仓库与数据挖掘、物联网、网络信息安全。

层次信息(tree)、网络信息(network)、时序信息(temporal)可视化[3]。本文就目前大数据可视化技术研究的主要方向：多维信息、层次信息、时序信息这三种信息可视化进行分析比较。

二 大数据可视化技术分析

大数据可视分析是指在大数据的背景下进行数据挖掘、建立模型，同时，利用支持信息可视化的用户界面运用人机交互方式与技术对数据进行可视化映射和模型可视化，有效融合计算机的计算能力和图形图像的表达能力，提高数据使用者对于大规模复杂数据集的认知力和洞察力。

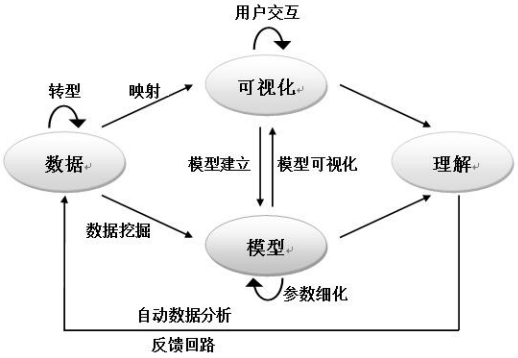


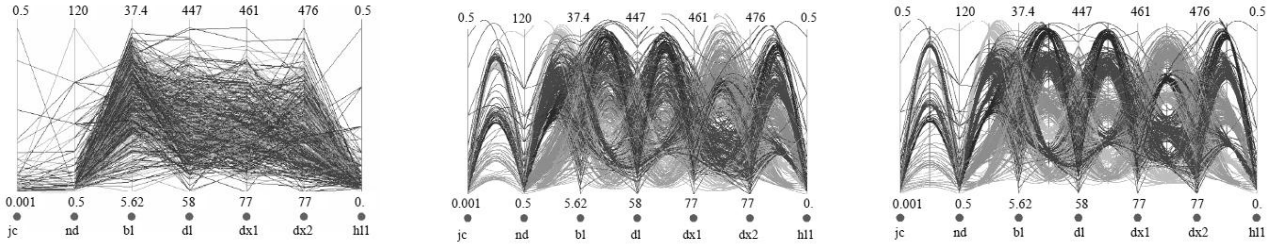
图1 可视分析的运行机制

因此，从图形图像的表达角度和对数据理解力的角度，将多维信息、层次信息和时序信息可视化的普遍使用方案技术与近5年内优化的可视化技术进行比较分析。

（一）多维信息可视化

传统的高维数据可视化算法，以Inselberg提出的平行坐标技术[4]最为代表。将高维数据转换二维或三维的空间中进行标示，算法复杂、耗时长适用性较差。

在近几年的研究中，陈谊等提出将数据聚类后的平行坐标可视化[5]，其主要思想是用K-means算法对原始大数据进行聚类处理，重新分配类区间的宽度，对不同的类区间按其权重大小来排序作图。



（a）原始数据的平行坐标图　　(b) 原始类区间的平行坐标图　　(c) 经过类区间划分后的平行坐标图

图2 基于类区间的多维数据可视化示意图[5]

这样相比于传统的平行坐标可视化技术，基于类区间的平行坐标可视化就能有效的降低数据线段的杂乱程度，使得数据线段的分布清晰明了，便于用户查看数据趋势。

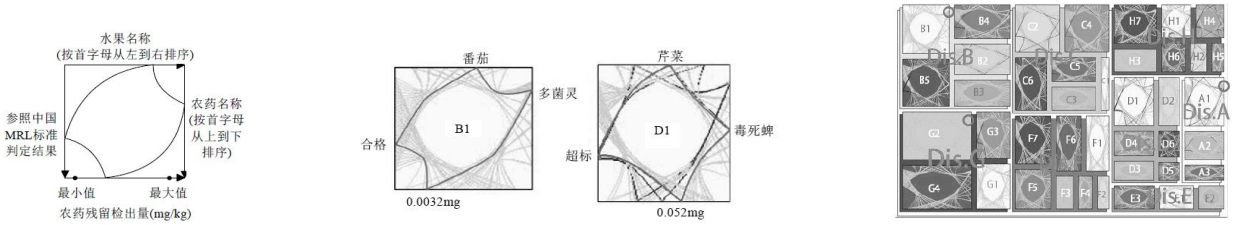
（二）层次信息可视化

层次信息是很常见的结构信息之一，例如：系谱图、计算机文件系统、组织结构图等。层次信息可视化最为常见的形式就是基于Johnson提出的树图布局算法[6]，但是，在大数据层次信息中，这种表示方式会因为横向的每层节点数和纵向的树深层数扩展比例失调、分支拥挤等问题导致层次结构的表示模糊。

杨如意等人于2014年提出大众标注层次可视化算法[7]，依赖全局上下文定位信息位置，快速直观的显示当前位置的层次结构，提高空间利用率。但这种算法有明显的缺点，即不能满足跨层级、多继承的数据要求。

陈谊等人于2016年提出一种树图多维坐标MCT(Multi-coordinate in Treemap)技术[8]。采用Squarified和Strip布局算法的树图表示层次结构，将树图节点的4条边定义成4条属性轴，计算属性值将其映射与属性轴

上，首尾顺次相连，进行曲线拟合。这样层次信息就能分区域进行可视化表达，避免了树形结构的比例失调问题，将层次信息清晰的呈现给用户。



（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

图3 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[8]

（三）时序信息可视化

在信息可视化领域中，时序信息可视化始终是一个难点。因为时间信息基于数据的其他属性信息而言，具有不可改变的顺序性。2000年Susan Havre 等人提出ThemeRiver模型[9]，以河流为模型表达随者时间改变数据发生变化的可视化算法。这种算法的优点是能够显示不同的时序信息的演化过程，缺点则为不能在图中直观的表现时序信息的具体数值和相关属性。

2007年Johansson 等人提出Temporal Density Parallel Coordinates(TDPC) 算法和Depth Cue Parallel Coordinates (DCPC) 算法[10]，通过转换函数得到绘图颜色，在平行坐标系上绘制多边形来代替折线表示数据根据时间的变化，可以观察到数据发生变化时的时刻位置，但这种算法存在着数据遮挡问题。

2015年董重和魏迎梅提出多变元时序数据可视化方法[11]，通过时间维分段、视觉聚类和颜色绘制三个步骤，对时序数据进行可视化处理，可以观察到数据信息的趋势变化和突变时刻，数据遮挡问题得到改善。

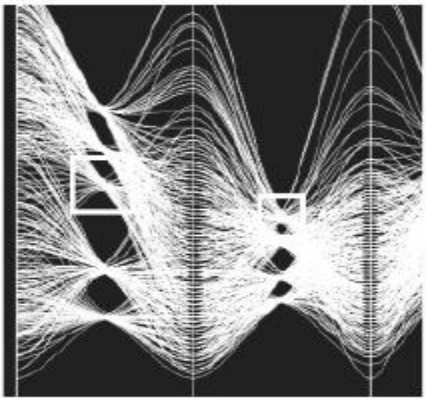


图4 中国气象科学数据的多变元时序数据可视化[11]

三 大数据可视化面临的问题和挑战

（一）海量高维数据集可视化的遮挡

从本文第2节的描述中可知，在大数据背景下，大量复杂的高维数据无论是在数据分析还是数据可视化都存在巨大的挑战，并且无可避免的会出现界面显示混乱不清、数据曲线严重重叠的现象。因此，对现有的高维数据可视化方法清晰化，或者是对高维数据可视化的布局优化都是此类问题研究的方向。

（二）多源异构数据的可视化

大数据时代数据来源广，种类多，且多来自于异构环境。所以，在面对多源异构的数据进行数据分析时，即使获得了数据源，也无法保证数据的完整性和准确性。导致对这类数据进行分析是可视化技术的一大难点，可以通过研究不同类型的数据接口进行可视化分析来解决。

（三）实时数据分析可视化

越来越多的实时数据在互动论坛、企业系统、穿戴设备等地方产生，且产生速度快、密度高。将这些数据经过合理的快速分析后，进行实时数据可视化成为时下的技术难点，也是时序数据可视化与实际运用相结合的严峻挑战。

（四）预测分析

目前，随着行为分析相关产业的应用发展兴起，大数据预测模型的需求在与日俱增，但支持预测分析的系统却寥寥无几。由此可见，先进的、准确的、可定制的可视化预测分析是未来的发展趋势。

（五）可扩展性

大数据的数据规模当下已经呈现爆炸式的增长，持续积累的数据量无限扩大，以至于普通计算机的处理能力难以跟上数据的新增速度。同时，设备显示屏幕范围有限，而需要表达的数据量无限，在这样极端的数据规模条件下，可视化技术的优点在不断减弱。提高可视化技术的扩展性成为解决问题的关键。如何降低数据规模，如何提高人机交互技术，如何结合大规模并行处理方法和超级计算机，这些问题都将是未来大数据发展中最为核心的挑战。

四 结束语

在可视化的过程中，数据可以变得更具可塑性、可行性，最终更加人性化。在可视化的呈现中，使海量数据集中或者汇总展示，让信息使用者可以快速聚焦在数据的关键点。本文针对数据信息的特征，将多维信息可视化、层次信息可视化和时序信息可视化，这三类可视化技术的常用算法和近5年的新提出的研究算法进行了详细的比较。同时，指出面临的困难与挑战，做出5点归纳，强调数据预测和可视化技术的扩展性将是未来大数据可视化发展的重大挑战。

在本文接下来的工作中，应在云平台上结合相关产业数据，围绕可视化分析的实践运用，进行深入的大数据可视化研究与探索。

参考文献

[1] 曾悠. 大数据时代背景下的数据可视化概念研究[D]. 浙江大学,杭州, 2014

[2] MUNZNE R T. Visualization analysis and design [J]. Wiley Interdisciplinary Reviews Computational Statistics, 2015, 2 (4): 387 —403.

[3] Card SK, Mackinlay JD, Shneiderman B. Readings in Information Visualization: Using Vision To Think. San Francisco: Morgan-Kaufmann Publishers, 1999. 1 712.

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

[5] 陈谊, 李潇潇, 蔡进峰, 陈红倩, 蔡强. 基于类区间的多维数据可视化方法[J]. 系统仿真学报, 2013, 25(10):2418-2423.

[6] 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]

[7] 杨如意, 刘东苏. 基于大众标注的层次信息可视化算法研究[J]. 现代图书情报技术, 2014(7):71-76.

[8] 陈谊, 甄远刚, 胡海云, 梁婕, Kwan-Liu MA. 一种层次结构中多维属性的可视化方法[J]. 软件学报, 2016(5):1091-1102.

[9] Havre S, Hetzler B, Nowell L. ThemeRiver: Visualizing theme changes over time [C]// Information Visualization, 2000, InfoVis 2000, IEEE Symposium on. USA: IEEE, 2000: 115-123.

[10] Johansson J, Ljung P, Cooper M D. Depth cues and density in temporal parallel coordinates[C]. Euro Vis, 2007, 7: 35-42.

[11] 董重, 魏迎梅. 运用平行坐标系的多变元时序数据可视化方法[J]. 小型微型计算机系统, 2015, 36(10):2408-2411.