✦前瞻技术

专栏作家简介：陈明，男，中国石油大学计算机科学与技术系教授，博士生导师，特聘教授，研究方向为分布式并行计算、计算智能、软件工程、大数据计算等，chenming@cup.edu.cn。

文章编号：1672-5913(2015)05-0094-04 **中图分类号**：G642

大数据可视化分析

陈 明

（中国石油大学 计算机科学与技术系，北京 102249）

摘 要：人类利用形象思维获取视觉符号中所蕴含的信息并发现规律，进而获得科学发现。文章介绍科学可视化、信息可视化和数据可视化的内涵，阐述大数据可视化分析方法。

关键词：大数据；可视化分析；并行计算

0 引 言

人类的创造性不仅取决于逻辑思维，还与形象思维密切相关。人类利用形象思维将数据映射为形象视觉符号，从中发现规律，进而获得科学发现。期间，可视化关键技术对重大科学发现起到重要作用。在大数据时代，大数据可视化分析的研究与发展将为科学新发现创造新的手段和条件[1]。

数据可视化于20世纪50年代出现，典型例子是利用计算机创造出了图形图表。1987年，布鲁斯·麦考梅克等撰写的《Visualization in Scientific Computing》促进了可视化技术的发展，将科学计算中的可视化称之为科学可视化[2]。20世纪90年代初期，出现了信息可视化。目前将科学可视化与信息可视化都归为数据可视化。

2 科学可视化

2.1 问题的提出

传统的科学可视化技术已成功应用于各学科领域，但如果将其直接应用于大数据，将面临实用性和有效性问题，这说明需要对科学可视化技术重新审视与深入研究。

2.2 分布式并行可视化算法

可扩展性是构造分布式并行算法的一项重要指标。传统的科学可视化算法应用在小规模的计算机集群中，最多可以包括几百个计算节点，而实际应用是要在数千甚至上万个计算节点上运行。随着数据规模的逐渐增大，算法的效率逐渐成为数据分析流程的瓶颈，设计新的分布并行可视化算法已经成为一个研究热点。

2.2.1 并行图像合成算法

传统的并行图像合成算法主要包括前分割算法、中间分割算法和后分割算法3种类型，前分割算法主要分为如下3步骤：

（1）将数据分割并分配到每个计算节点上；

（2）每个计算节点独立绘制分配到的数据，在这一步，节点之间不需要数据交换；

（3）将计算节点各自绘制的图形汇总，合成

最终的完整图形。

从上述步骤中可以看出，由于节点之间可能需要大量的数据交换，尤其是步骤（3）可能成为算法的瓶颈。解决这个问题的关键是减少计算节点之间的通信开销，可以通过对数据进行划分并在各计算节点间进行分配来实现。划分和分配方案需要与数据的访问一致，原则是计算节点只使用驻留本计算节点的数据进行跟踪，从而减少数据交换。

2.2.2 并行颗粒跟踪算法的研究

传统的科学可视化研究对象主要集中在三维标量场数据。在科学大数据中，经常使用三维流场数据，其原因如下所述。

将二维的流场可视化方法直接应用在三维流的结构不可能都成功，每个颗粒虽然可以单独跟踪，但是可能出现在空间中的任何一个位置，这就需要计算节点之间通过通信交换颗粒。同时，当大量的颗粒在空间移动时，每个计算节点可能处理不同数量的颗粒，从而造成计算量严重失衡。解决这些问题的关键是减少计算节点之间的通信开销，其基本思路同并行图像合成算法。

2.2.3 重要信息的提取与显示技术

科学大数据可视化的另一个重要研究方向是如何从数据中快速有效地提取重要信息，并且用这些重要信息来指导可视化的生成。从可视化的角度来看，一方面需要可视化设计表达数据中特定信息的定义，通过人机交互工具，由用户来调整参数，观察和挖掘数据中的重要信息；另一方面需要根据用户的反馈信息调整可视化，以更好地突显重要信息，淡化非重要信息，方便用户对重要信息及其背景的观测。整个信息的提取过程是个典型的交互式可视分析过程。基于这一思想的两个技术是流场可视化的层次流线束技术和用于标量数据的基于距离场的可视化技术。

2.2.4 原位可视化

传统的科学可视化采用科学计算后进行处理的模式。随着计算机系统计算速度的提高，I/O速度与计算速度之间的差距增大。随着计算规模越来越大，而相应生成的数据规模也越来越大，现有的存储系统无法把所有的计算数据都保存下来。解决上述问题的常用方法是采用空间或者时间上的采样方法，最后只保存部分数据，造成结果数据的丢失，不能保证高精度数值模拟。

原位可视化的基本思想是：

（1）将可视化与科学模拟集成在一起。在科学模拟的过程中，每个时间片的结果生成之后，可以立刻调用可视化模块，直接与科学模拟程序集成。为了减少数据的冗余，可视化程序与科学模拟程序共享数据结构。

（2）由于数据的分割和分配优先满足科学模拟的需求，可视化程序的工作分配有可能是不均衡的，需要重现可视化的工作量在各个计算节点上分配算法，减少数据传输。

（3）可视化程序的开销不能太高，要保持集成系统的高效能，必须提高可视化程序的效率，其可扩展性必须与科学模拟一致，可以应用上万个、上10万个或更多的计算节点。

3 信息可视化

自18世纪后期数据图形学诞生以来，抽象信息的视觉表达手段一直被用来揭示数据及其他隐匿模式的奥秘。20世纪90年代期间出现的图形化界面则使得人们能够直接与可视化信息进行交互，从而推动了信息可视化研究。信息可视化通过人类的视觉能力，来理解抽象信息的意思，从而加强人类的认知活动，达到能够驾驭日益增多的数据的能力。

信息可视化是跨学科领域的大规模非数值型信息资源的视觉展现，能够帮助人们理解和分析数据。信息可视化中的交互方法能够实现用户与数据的快速交互，更好地验证假设和发现内在联系。信息可视化技术提供了理解高维度、多层次、时空、动态、关系等复杂数据的手段，与科学可视化相比，信息可视化更侧重于抽象数据集，如对非结构化文本或者高维空间中不具有固有的二维或三维几何结构的点的视觉展现。信息可视化适用于大规模非数字型信息资源的可视化表达。

信息可视化与科学可视化的不同之处是，信息可视化所要可视化的数据并不是某些数学模型的结果或者是大型数据集，而是具有自身固有结构的抽象数据。

科学可视化主要处理具有地理结构的数据，信息可视化主要处理像树、图形等抽象式的数据结构，可视化分析则主要挖掘数据背景的问题与原因。更进一步说，科学可视化技术是指空间数据的可视化技术，而信息可视化技术则是指非空间数据的可视化技术。

4 数据可视化

4.1 概 念

数据可视化技术是指运用计算机图形学和图像处理技术，将数据转换为图形或图像在屏幕上显示出来，并利用数据分析和开发工具发现其中未知信息的交互处理的理论、方法和技术[3]。

数据可视化不仅包括科学计算数据的可视化，而且包括工程数据和测量数据的可视化。数据可视化是对大型数据库或数据仓库中的数据的可视化，它是可视化技术在非空间数据领域的应用，不再局限于通过关系数据表来观察和分析数据信息，还能以更直观的方式看到数据及其结构关系。

4.2 数据可视化技术的特点

数据可视化技术能够分析大量复杂和多维的数据，提供像人眼一样的直觉的、交互的和反应灵敏的可视化环境。数据可视化技术的特点如下所述。

（1）交互性。用户可以方便地以交互的方式管理和开发数据。

（2）多维性。对象或事件的数据具有多维变量或属性，而数据可以按其每一维的值分类、排序、组合和显示。

（3）可视性。数据可以用图像、曲线、二维图形、三维体和动画来显示，用户可对其模式和相互关系进行可视化分析。

数据可视化已经出现了许多方法，主要有基于几何技术、面向像素技术、图标技术、层次技术、图像技术和分布式技术等。

4.3 数据可视化技术的相关概念

（1）数据空间：是由 n 维属性和 m 个元素组成的数据集所构成的多维信息空间。

（2）数据开发：指利用一定的算法和工具对数据进行定量的推演和计算。

（3）数据分析：指对多维数据进行切片、分块、旋转等动作剖析数据，从而能多角度多侧面观察数据。

5 大数据可视化分析

5.1 概 念

大数据可视化分析需要应用有效的数据管理方法[4]。这也是创建混合环境的需要。在大数据环境下，人们利用各种技术分析数据，用形象直观的方式展示结果，这样能够快速发现数据中蕴含的规律特征。

可视化分析关注人类感知与用户交互的问题。大数据来自不同领域的模拟与观察实测。大数据可视分析通常应用高性能计算机群、处理数据存储与管理的高性能数据库组件及云端服务器和提供人机交互界面的桌面计算机。

5.2 大数据可视化分析方法

5.2.1 原位交互分析技术

在进行可视化分析时，将在内存中的数据尽可能多地进行分析称之为原位交互分析。对于超过 PB 量级以上的数据，将数据存储于磁盘进行分析的后处理方式已不适合。与此相反，可视分析则在数据仍在内存中时就会做尽可能多的分析。这种方式能极大地减少 I/O 的开销，并且可实现数据使用与磁盘读取比例的最大化。然而应用原位交互分析也会出现下述问题：①由于人机交互减少，进而容易造成整体工作流中断；②硬件执行单元不能高效地共享处理器，导致整体工作流中断。

5.2.2 数据存储技术

大数据是云计算的延伸，云服务及其应用的出现影响了大数据存储。流行的 Apache Hadoop 架构已经支持在公有云端存储 EB 量级数据的应用。许多互联网公司都已经开发出了基于 Hadoop 的 EB 量级的超大规模数据应用。一个基于云端的解决方案可能满足不了 EB 量级数处理。

一个主要的疑虑是每千兆字节的云存储成本仍然显著高于私有集群中的硬盘存储成本。另一个问题是基于云的数据库的访问延时和输出始终受限于云端通信网络的带宽。不是所有的云系统都支持分布式数据库的ACID标准。对于Hadoop软件的应用，这些需求必须在应用软件层实现。

5.2.3 可视化分析算法

大数据的可视化算法不仅要考虑数据规模，而且要考虑视觉感知的高效算法。需要引入创新的视觉表现方法和用户交互手段。更重要的是用户的偏好必须与自动学习算法有机结合起来，这样可视化的输出具有高度适应性。可视化算法应拥有巨大的控制参数搜索空间，减少数据分析与探索的成本及降低难度，可以组织数据并且减少搜索空间。

5.2.4 不确定性的量化

许多数据分析任务中引入数据亚采样来应对实时性的要求，由此也带来了更大的不确定性。数据中不确定性的来源对于决策和风险分析十分重要。随着数据规模不断增大，直接处理整个数据集的能力也受到了极大的限制。不确定性量化已经成为科学与工程领域的重要问题之一。不确定性的量化对未来的可视分析工具极端重要，新的可视化技术将提供一个不确定性的直观视图来帮助用户了解风险，从而帮助用户选择正确的参数，减少产生误导性结果。不确定性的量化将成为可视化分析任务的核心部分。

5.2.5 并行计算

并行处理可以有效地减少可视计算所占用的时间，从而实现数据分析的实时交互。多核的计算体系结构的每个核所占有的内存也将减少，在系统内移动数据的代价也将提高。为了发掘并行计算的潜力，许多可视化分析算法需要完全地重新设计。在单个核心内存容量的限制之下，不仅需要有更大规模的并行，也需要设计新的数据模型，需要设计出既考虑数据大小又考虑视觉感知的高效算法，需要引入创新的视觉表现方法和用户交互手段。

5.2.7 领域资源库、框架以及工具

由于缺少低廉的领域资源库、框架和工具，基于高性能计算的可视化分析应用的快速研发受到了严重阻碍。如用户界面、数据库等领域对于可视分析系统的开发至关重要。在绝大部分的高性能计算平台上，即使是最基本的软件开发工具也很少见。目前为高性能计算平台开发定制这样的软件，还是个耗时耗力的做法。

5.2.8 用户界面与交互设计

由于传统的可视化分析算法的设计通常没有考虑可扩展性，所以许多算法的计算过于复杂或者不能输出易理解的简明结果；加之数据规模不断地增长，以人为中心的用户界面与交互设计面临多层次性和高复杂性的困难；同时计算机自动处理系统对于需要人参与判断的分析过程的性能不高，现有的技术不能更充分发挥人的认知能力。利用人机交互可以化解上述问题。为此，在大数据的可视化分析中，用户界面与交互设计成为研究的热点，主要应考虑下述问题：用户驱动的数据简化、可扩展性与多级层次、异构数据融合、交互查询中的数据概要与分流、表示证据和不确定性、时变特征分析、设计与工程开发等一系列问题。

6 结 语

原位交互分析技术、数据存储技术、可视分析算法和用户界面与交互设计等多种技术的运用，使得人们可以通过交互可视界面来对大数据进行分析、推理和决策，这种将数据通过可视化变成图形的方法能更好地激发人的形象思维与想象力。

参考文献：

[1] 李国杰. 大数据研究的科学价值[J]. 中国计算机学会通讯, 2012, 8(9): 8-15.

[2] 俞宏峰. 大规模科学可视化[J]. 中国计算机学会通讯, 2012, 8(9): 29-36.

[3] 陈明. 大数据概论[M]. 北京: 科学出版社, 2014: 182-198.

[4] 黄伯仲, 沈汉威, 克里斯托弗·约翰逊, 等. 超大规模数据可视化分析十大挑战[J]. 中国计算机学会通讯, 2012,8(9): 38-43.

（编辑：彭远红）