大数据可视化技术及应用

沈恩亚

清华大学软件学院;大数据系统软件国家工程实验室,北京 100084

摘要 随着人类产生数据量的增加,数据可视化需要处理的数据规模、类型及需求都发生了显著变化。在大数据时代,数据可视化面临诸多新的挑战。从大数据本身的特点及其应用需求出发,结合数据可视化的研究现状,介绍了适用于大数据的数据可视化技术;分析在大数据条件下数据可视化所要解决的8个关键问题;讨论了针对大数据可视化应用需求自主研发的交互式可视化设计平台AutoVis及其应用。

关键词 数据可视化;大数据;时序数据可视化;可视化系统

20世纪90年代以来,随着计算机计算能力的提高,人类进入数据爆炸时代。以数据规模为例,20世纪末数值模拟结果以GB(10^3 MB)为单位,21世纪初逐渐开始以TB(10^3 GB)为单位,近10年来诸多数值模拟达到PB(10^3 TB)级以上。以数据类型为例,随着互联网和物联网的兴起以及各类互联网产品和电子设备的出现与发展,产生了大量的网页、图片、视频、音频、传感器时间序列等结构化与非结构数据。社会已经进入数字时代,从科学研究与技术实践的角度来说,进入了大数据时代。这一时代的技术性标志在于人类产生的数据超过了传统的数据处理工具的处理能力。

这一时代也给人们提供了新的机遇,如图灵奖获得者Jim Gray所说,数据密集型科学发现是继实验归纳、逻辑推演、仿真模拟之后的第4类科学方法[1],这一方法作为前3种科学范式的补充,进一步促进人类科技的进步。在基础科学研究领域如此,在科技应用领域亦是如此。以近10年蓬勃发展的深度学习为例,前所未有的数据规模是其核心推动力之一,结合算力的提高与算法的精进,一些与日常息息相关的技术逐渐达到实用级别,从原有的数据驱动的商品、好友、新闻辅助推荐,逐渐发展到日常的人机语音对话、人脸识别等各类应用,具有代表性的自动驾驶技术也开始进入道路实测阶段。

数据推动着诸多科学领域与各行各业发展的同时,也带来了前所未有的挑战。如何有效地让人去理解数据,避免“big data”成为“big rubbish”。“我们需要开发更好的工具以支持整个研究过程,包括数据捕捉、数据治理、数据分析以及数据可视化”[1]。数据可视化技术作为人理解数据的途径之一,已广泛应用于各个领域,例如医学(图1)、航空航天(图2)、基础物理(图3)等,也已渗透到人们日

收稿日期:2019-11-11;修回日期:2020-01-08

基金项目:北京市科技计划项目(Z111100067311053);国家重点研发计划项目(2016YFB0501504);国家自然科学基金项目(U1509213)

作者简介:沈恩亚,博士,研究方向为大数据、数据可视化、可视分析及人机交互,电子信箱:shenenya@mails.tsinghua.edu.cn

引用格式:沈恩亚. 大数据可视化技术及应用[J]. 科技导报, 2020, 38(3): 68-83; doi: 10.3981/j.issn.1000-7857.2020.03.004

常生活的各个角落,例如常用的地铁路线图。

从促进科学研究的科学计算可视化、到在信息领域应用广泛的信息可视化,再到逐渐深入的可视化分析领域,数据可视化技术伴随着数据规模、类型以及应用需求的不断发展,亦在不断演进。在大数据时代,数据可视化技术在广泛应用的同时,也面临诸多新的挑战。大数据可视化是一个面向应用的研究领域,本文重点从应用实践的角度,讨论在大数据背景下大数据可视化内涵、研究进展、相关技术与产品以及所面临的一系列挑战。

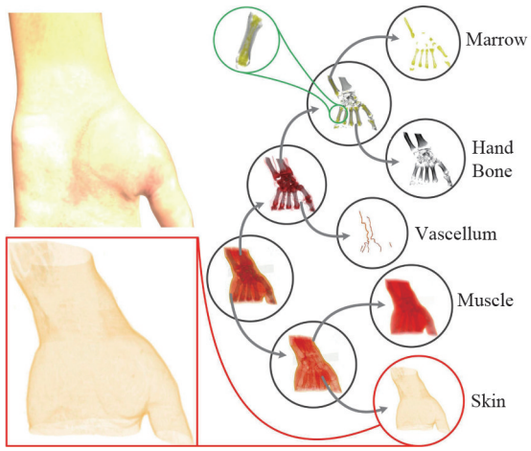


图1　手部MRI数据可视化[2]

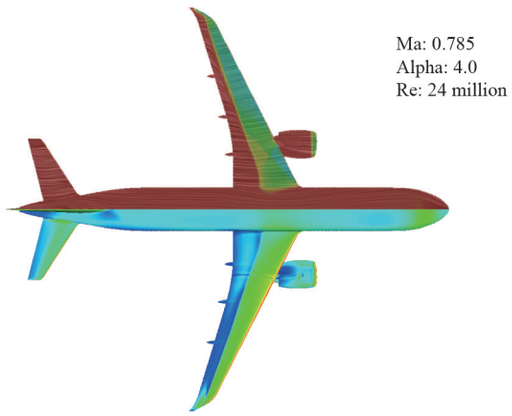


图2　C919气动外形流场与温度场数据可视化[3]

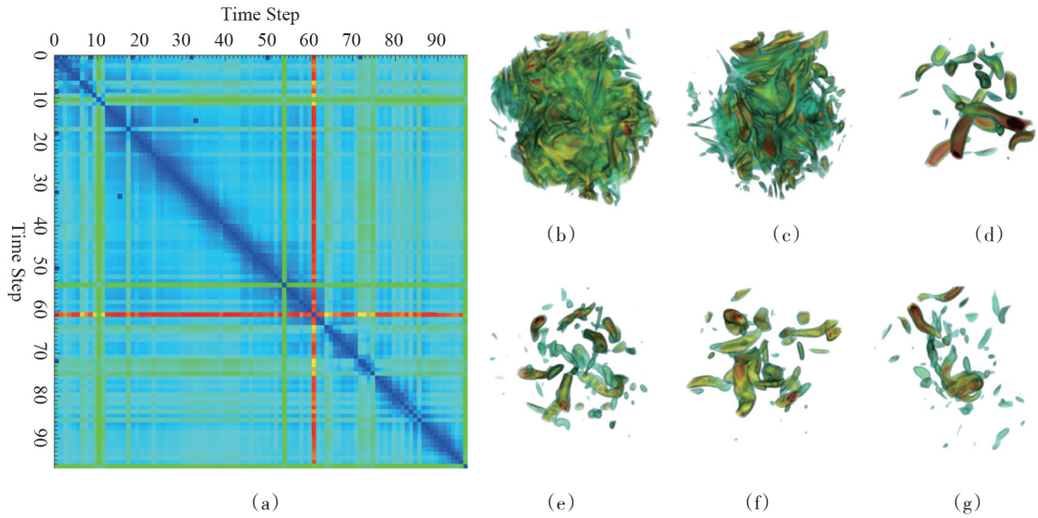


图3　时变湍流数据可视化[4]

1　大数据可视化内涵

大数据是大数据可视化所处理的对象,通常指数据规模、产生速度以及复杂度超过了传统方法处理能力的数据[5]。这一概念在20世纪90年代逐渐被使用,21世纪初开始流行,其中,《Nature》和《Science》分别于2008年和2011年推出特刊,讨论大数据带来的影响与挑战[6-7]。于此同时,大数据技术从网络搜索服务逐渐应用到更多领域。大数据的三个典型特征是数据规模大(volume)、类型多样(variety)和更新速度快(velocity)(3Vs)[5]。

1)数据规模大:大数据需要大规模存储空间,

一般为分布式存储架构。

2）类型多样：大数据通常包括多种类型的数据，包括图片、音频等非结构化数据。

3）更新速度快：大数据更新快速，例如传感器常常达到s级或ms级采样。

一般而言，可视化指将抽象之物形象化，所谓一图胜千言。研究表明，人每天所接受的信息中约83%通过视觉获得[8]，可视化将不可见的事物（如气流）通过可见的形式表达，从而让人可以去观察和理解相应事物，获得更多信息。可视化的思想在中国已有几千年的历史，例如为了“可视化”表达《易经》的主要思想所发明的太极图，形象直观地表达了中国古代哲学中的阴阳概念以及“你中有我，我中有你”的哲学思想。

数据可视化是将抽象的“数据”以可见的形式表现出来，帮助人理解数据。现代可视化利用计算机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。它涉及计算机图形学、图像处理、计算机视觉、计算机辅助设计等多个领域，成为研究数据表示、数据处理、决策分析等一系列问题的综合技术。这一概念自1987年正式提出，经过30余年的发展，逐渐形成3个分支：科学计算可视化（scientific visualization）、信息可视化（information visualization）和可视分析（visual analytics）。

科学计算可视化是指将具有空间维度属性的数据（例如医学、计算流体力学和气象学）进行可视化的方法[9]，是可视化研究中传统的研究领域，在上述3个分支中得到的研究也最多，近年研究相对减少。信息可视化伴随互联网兴起于信息爆炸而诞生，主要用于相对抽象的非空间数据可视化[10]，这是大众接触相对较多的可视化形式。可视分析在科学计算可视化和信息可视化的基础上，更加注重分析推理与交互[11]，近年研究逐渐增加。值得注意的是，在诸多数据可视化应用中，三者的界限逐渐模糊，例如传统的数值模拟科学计算可视化常常结合风洞实验的传感器和拍照数据；可视分析所处理的对象也不限于抽象的信息数据。3个子领域出现了逐渐融合的趋势。本文统称为“数据可视化”。

大数据可视化源于传统的数据可视化，其核心要义依然是将数据映射为图表等可见的形式，不同在于大数据可视化相对传统的数据可视化，处理的数据对象有了本质不同，在已有的小规模或适度规模的结构化数据基础上，大数据可视化需要有效处理大规模、多类型、快速更新类型的数据。这给数据可视化研究与应用带来一系列新的挑战。因此，在传统数据可视化基础上，尝试给出大数据可视化的内涵如下：大数据可视化是指有效处理大规模、多类型和快速变化数据的图形化交互式探索与显示技术。其中，有效是指在合理时间和空间开销范围内；大规模、多类型和快速变化是所处理数据的主要特点；图形化交互式探索是指支持通过图形化的手段交互式分析数据；显示技术是指对数据的直观展示。

大数据可视化是大数据系统必要组成之一。以美国国家标准及技术研究所提出的大数据参考架构[12]（图4）为例，该架构将数据可视化列为大数据应用组件。其中根据目的不同，将可视化分为可视化探索（exploratory visualization）、可视化评估（evaluative visualization）和可视化解释（explanatory visualization）（3Es）。可视化探索即大数据场景下的“探索式数据分析”，通过可视化对原始数据进行交互式分析，例如数据分布；可视化评估即大数据场景下的“数据与模型调试”，评估数据分析和机器学习方法的有效性；可视化解释即大数据场景下的“信息可视化”，用于知识的交流与传播；可视化解释即大数据场景下的“信息可视化”，用于知识的交流与传播。

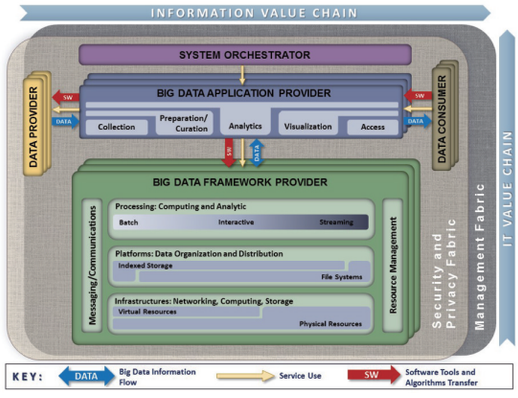


图4 NIST大数据参考架构[12]

2　大数据可视化技术

根据不完全统计,数据可视化方法超过了百种。如果将不同的方法进一步细分,则达到数百上千种,以折线图为例,即存在数十种变种,适用于不同场景。首先从方法层面介绍基本满足常用数据可视化需求的通用技术,根据可视化目标分类介绍,然后根据大数据的特点,重点介绍相关的大规模数据可视化、时序数据可视化、面向可视化的数据采样方法和数据可视化生成技术。

2.1　常用的数据可视化技术

数据可视化技术在应用过程中,多数非技术驱动,而是目标驱动。例如分析飞行器气动特性,关注其周围涡结构,其目的是通过涡结构可视化分析飞行器流场数值模拟准确性及其反应的动力学特征。再比如某公司需要查看公司近期业绩情况,其目标是对比不同时刻公司业绩数据。图5显示了目前业界广泛使用的根据目标分类的数据可视化方法[13],数据可视化目标抽象为对比、分布、组成以及关系。

1)对比。比较不同元素之间或不同时刻之间的值。对于不同元素,又可以根据元素包含的变量数目分为单元素多变量和单元素单变量。如果是单元素多变量,例如两个企业自身不同产品销量对比,可以采用多变量柱状图。如果是单元素单变量,例如多个企业产值比较,可以采用柱状图。比较不同时刻之间的值,可以根据时间长短细分,如果是长期时序数据,又可以根据是否有周期性分别采用周期面积图和折线图。如果是短期时序数据,根据类别多少可以分别采用折线图和柱状图。

2)分布。查看数据分布特征,是数据可视化最为常用的场景之一,常用于数据异常发现、数值过滤和数据基本统计性特征分析。单个变量的分布,根据数据点数量多少分别采用折线图和柱状图。两个变量的分布可以采用散点图。多个变量的分布可以采用平行坐标方法。

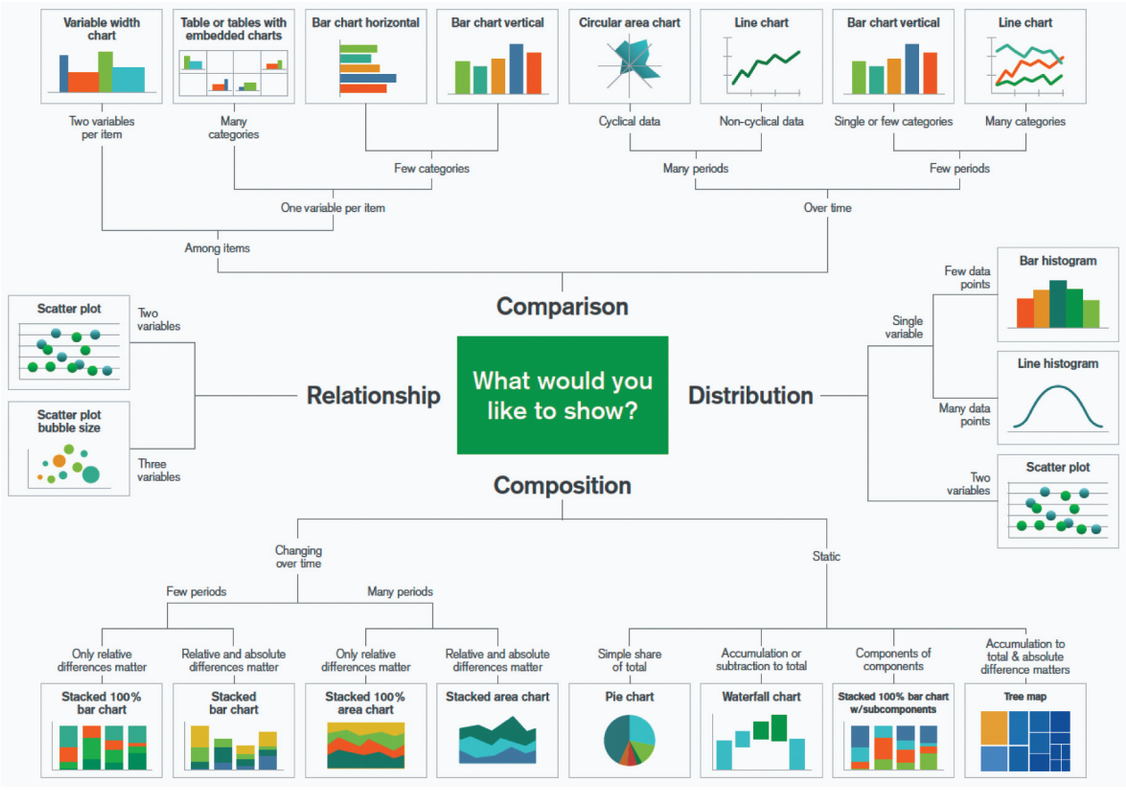


图5　按目标分类的常用数据可视化方法[13]

3）组成。查看数据静态或动态组成。动态组成可以根据数据特点分为短期和长期数据。对于短期数据,根据关注相对比例或绝对组成可以分别采用堆叠比例柱状图和堆叠柱状图;对于长期数据,同样根据关注相对比例或绝对组成可以分别采用堆叠比例面积图和堆叠面积图;对于静态组成,简单的总体组成,可以采用饼状图;若关注相对整体的增减可以采用瀑布图;若组成元素包含子元素,可以采用堆叠比例柱状图;若关注组成及其绝对差,可以采用树图。

4）关系。查看变量之间的相关性,这常常用于结合统计学相关性分析方法,通过视觉结合使用者专业知识与场景需求判断多个因素之间的影响关系。根据变量的多少进行划分,若是两个变量可以采用散点图;若是3个变量,可以采用气泡图,用散点半径表征第3个变量;超过3个变量可以采用平行坐标方法。

2.2 大规模数据可视化

大规模数据可视化[14]一般认为是处理数据规模达到TB或PB级别的数据,常用于科学计算数据,例如气象模拟、数值风洞、核模拟、洋流模拟、星系演化模拟等领域。以图6为例[3],该数据模拟了航空领域三段翼周围流场结构,单时间步数据规模达到30 GB,通过大规模数据可视化,可以有效显示机翼周围各尺度涡结构、分布和变化趋势。经过数十年的发展,大规模数据可视化经过了大量研究,重点介绍其中的并行可视化和原位(*in situ*)可视化。

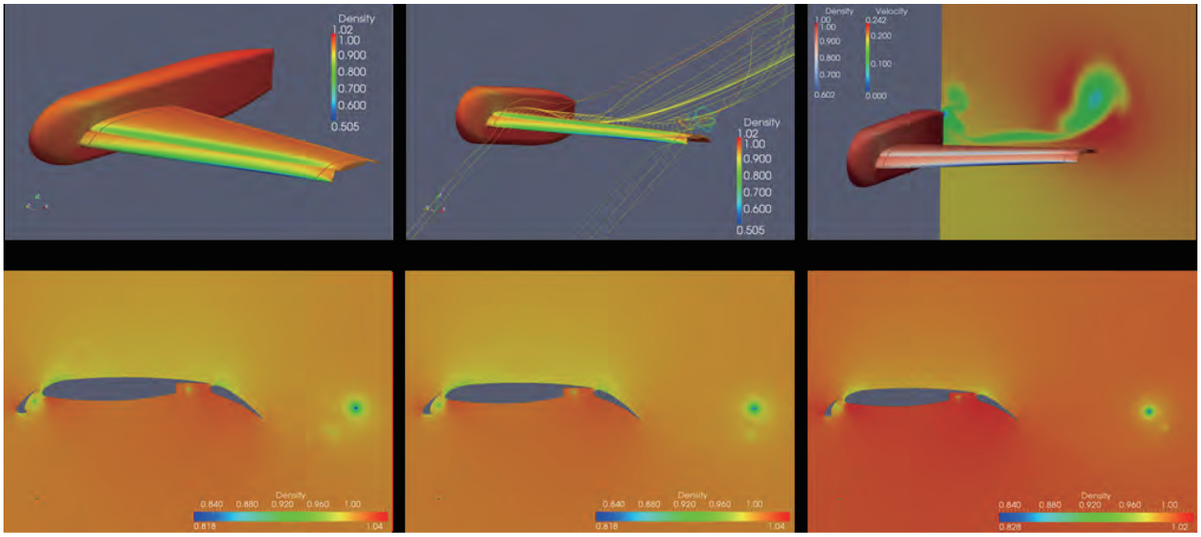


图6 大规模数据可视化实例[3]

1）并行可视化。并行可视化通常包括3种并行处理模式,分别是任务并行、流水线并行、数据并行[15-16]。任务并行将可视化过程分为独立的子任务,同时运行的子任务之间不存在数据依赖。各部分分别进行并行的可视化处理,最后进行合成。这一模式的优点是可以根据任务划分得到的子任务并行处理,缺点是当子任务不均匀时需要等待长耗时进程,存在资源浪费。如果子任务时间存在较多的数据依赖,也将显著影响并行性能。流水线并行采用流式读取数据片段,将可视化过程分为多个阶段,计算机并行执行各个阶段加速处理过程。其优点是可以充分利用计算机的硬件资源,缺点是处理过程受限于最慢阶段的耗时。数据并行是一种"单程序多数据"方式,将数据划分为多个子集,然后以子集为粒度并行执行程序处理不同的数据子集。其优点是可以实现高并行度,缺点是当数据之间的处理存在依赖时将导致等待耗时。

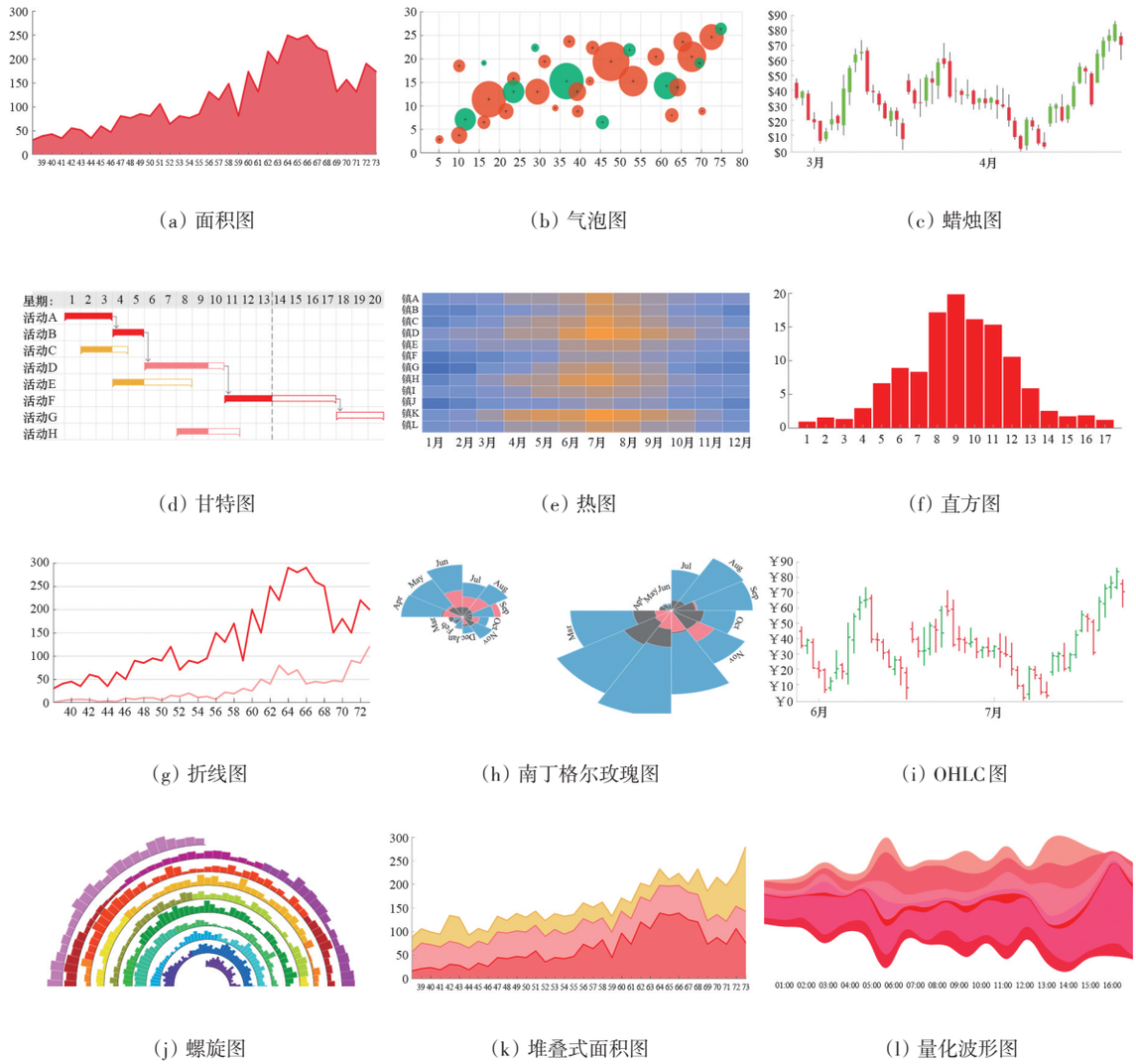
2）原位可视化。数值模拟过程中生成可视

化,用于缓解大规模数值模拟输出瓶颈[17-18]。根据输出不同,原位可视化分为图像、分布、压缩与特征。输出为图像的原位可视化,在数值模拟过程中,将数据映射为可视化,并保存为图像[19-20]。输出为分布数据的原位可视化,根据使用者定义的统计指标,在数值模拟过程中计算统计指标并保存,后续进行统计数据可视化[21];输出为压缩数据的原位可视化采用压缩算法降低数值模拟数据输出规模,将压缩数据作为后续可视化处理的输入[22-23];输出为特征的原位可视化采用特征提取方法,在数值模拟过程中提取特征并保存,将特征数据作为后续可视化处理的输入[24]。

2.3　时序数据可视化

快速变化是大数据的典型特征,可视化数据的时间维度特征是一个有趣的问题,首先,由于人类存在直接感知空间的器官,例如视觉和触觉,但缺少直接感知时间的器官;其次,在空间维度,人可以前进和后退,但在时间维度,人暂时只能向前,却不能后退。这些构成了时序数据可视化对于人的互补性,帮助人类通过数据的视角观察过去,预测未来,例如建立预测模型,进行预测性分析和用户行为分析。

常用时序数据可视化方法如图7[25]所示。其中,面积图可显示某时间段内量化数值的变化和发展,最常用来显示趋势,而非表示数值。气泡图可



(a) 面积图　(b) 气泡图　(c) 蜡烛图

(d) 甘特图　(e) 热图　(f) 直方图

(g) 折线图　(h) 南丁格尔玫瑰图　(i) OHLC图

(j) 螺旋图　(k) 堆叠式面积图　(l) 量化波形图

图7　常用时序数据可视化方法[25]

以将其中一条轴的变量设置为时间,或者把数据变量随时间的变化制成动画来显示。蜡烛图通常用作交易工具,用来显示和分析证券、衍生工具、外汇货币、股票、债券等商品随着时间的价格变动。

甘特图通常用作项目管理的组织工具,显示活动(或任务)列表和持续时间,也显示每项活动何时开始和结束。热图通过色彩变化来显示数据。当应用在表格时,将其中一行/列设为时间间隔,热图也可用于显示数据随时间的变化。直方图适合用来显示在连续间隔或特定时间段内的数据分布,其中每个条形表示每个间隔/时间段中的频率。直方图的总面积也相等于数据总量。

折线图用于在连续间隔或时间跨度上显示定量数值,最常用来显示趋势和关系。南丁格尔玫瑰图绘制于极坐标系之上。每个数据类别或间隔在径向图上划分为相等分段,每个分段从中心延伸多远(与其所代表的数值成正比)取决于极座标轴线。适用于周期性时序数据。与蜡烛图类似,OHLC图通常用作交易工具,显示和分析证券、货币、股票、债券等商品随时间的价格变动。

螺旋图沿阿基米德螺旋线绘制基于时间的数据。图表从螺旋形的中心点开始往外发展。螺旋图十分多变,可使用条形、线条或数据点,沿着螺旋路径显示。堆叠式面积图的原理与简单面积图相同,但它能同时显示多个数据系列,每一个系列的开始点是先前数据系列的结束点。量化波形图可显示不同类别的数据随着时间的变化,其形状类似河流,因此量化波形图看起来相对美观。

另外,具有空间位置信息的时序数据,常常将上述可视化方法地图结合,例如轨迹图。以上从方法层介绍了常用的时序数据可视化方法,结合具体数据与应用需求常常需要个性化定制,这也是诸多研究工作的创新之处,例如,Morrow等[26]设计了一种内外复合图(图8),将折线图、条形图等与柱状图结合,通过中间区域图表显示原始时序数据随时间变化趋势,外围统计图查看数据统计特征,实现两者的同时显示,结合区域刷选功能,达到交互式时序周期分析功能。Tominski等[27]汇总了100余种时序数据可视化方法。

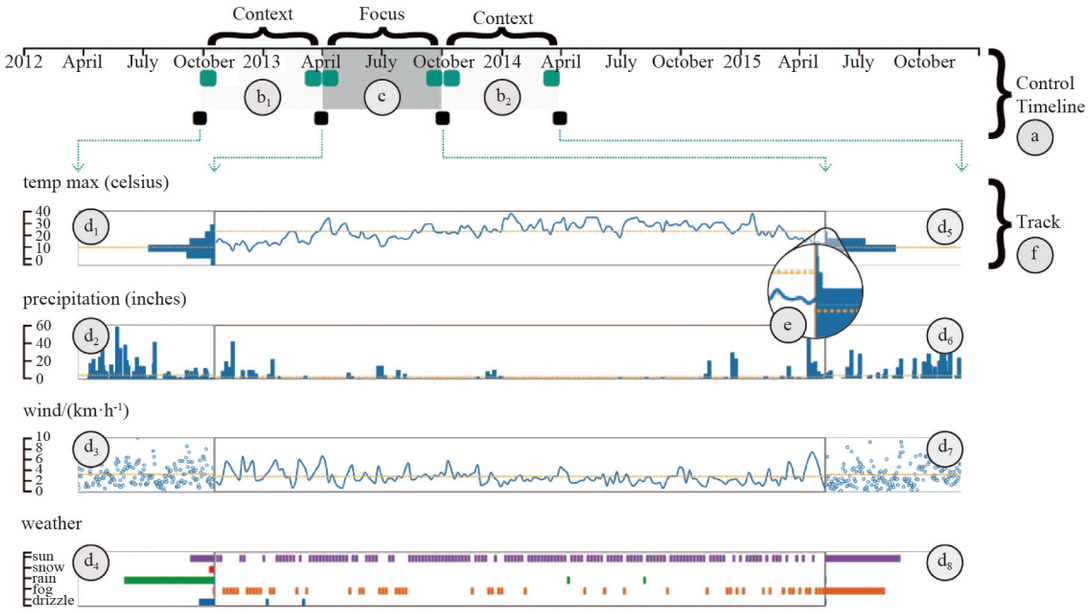


图8 内外复合图[26]

2.4 面向可视化的数据采样方法

大规模是大数据的显著特点,“在百万像素的屏幕空间显示数亿甚至更多的记录”[28],数据采样成为一种合理降低可视化数据规模的途径,与通用

的数据采样方法不同的是,这里需要考虑面向可视化的需求,例如如何在有限的屏幕空间显示数据的关键特征。面向可视化的时序数据采样,主要针对时序数据的折线图视觉效果进行优化。此类研究的主要目标为,从时序数据中选择小部分时序数据,利用折线图上的点与连线的视觉效果,使得选取数据的折线图视觉效果与原始数据的可视化结果尽可能接近。Steinarsson[29]总结了一些基于折线图的时序数据采样算法,认为折线图中的每个数据点都存在各自的视觉权重,例如峰值与谷值的数据点往往是人们最容易注意到的图形特征,具有最高的视觉权重。采样策略为先将时序数据按照时间区间划分为不同的桶,在每个桶中选择出视觉权重最高的数据点作为采样结果。Kehagias[30]提出了M4 aggregation时序数据采样算法。其首先考虑到可视化图表分辨率的限制,将可视化图表中位于同一个水平像素的所有数据点归为一个桶。进而在每个桶内,选择最大数值点、最小数值点、第一个数值、最后一个数值4个极值点作为采样结果,以保证样本数据覆盖了最大的水平范围与垂直范围——最大的视觉权重。

Guo等[31]总结了4种针对空间数据的可视化约束:代表性约束、可视性约束、平移一致性与缩放一致性,并基于可视化约束提出了贪心策略采样算法。可视性约束意味着限制采样结果中数据点彼此的最小间距以保证人可以在视觉上进行区分。平移一致性与缩放一致性则意味着在用户改变数据查询范围时不会产生某个采样数据点此时有彼时无的矛盾结果。Wu等[32]评估了近年来的图采样算法的可视化效果,认为针对可视化进行需要提出具有可视化性质的权重函数。Zhang等[33]提出了一种基于集合匹配的损失函数度量,并采用了类似于聚类算法的思路实现了对流式图数据的采样。

2.5　数据可视化生成方式

经过数十年的发展,数据可视化形成了从底层编程到上层交互式定制的多层次生成方式,如图9

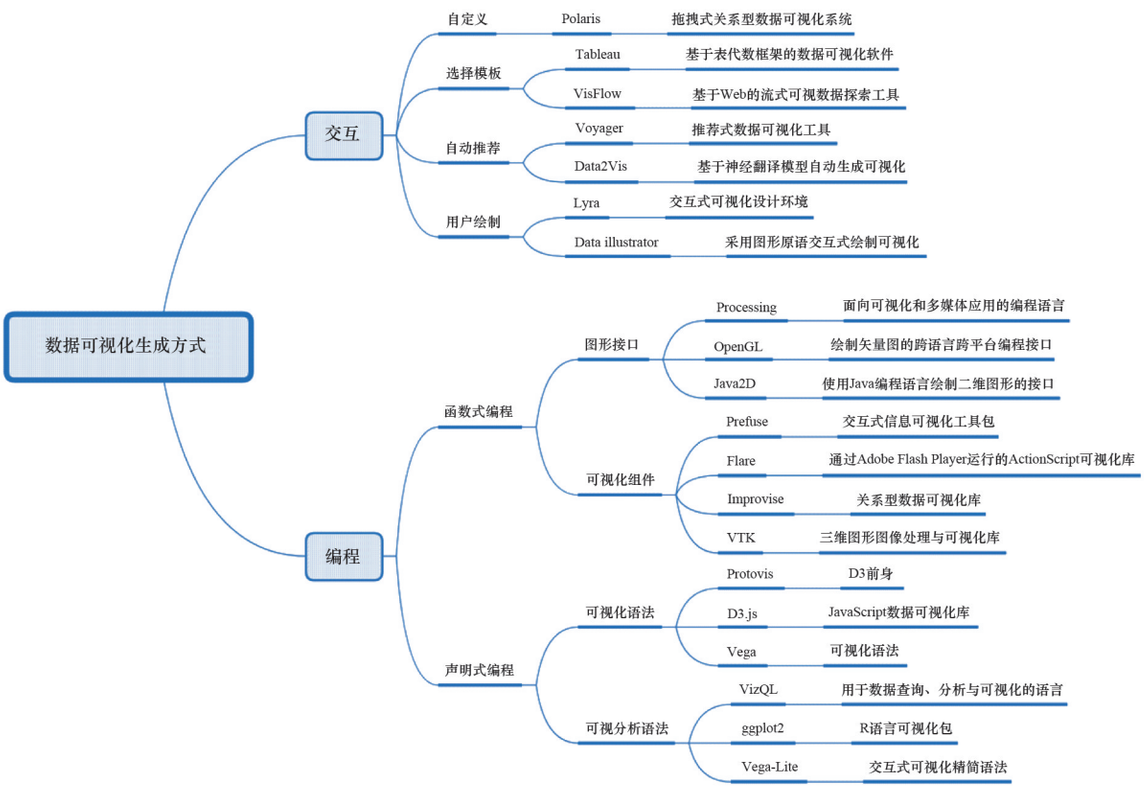


图9　数据可视化生产方式

所示。编程方式根据语言类型可以分为函数式编程与声明式编程。函数式编程可以根据图表元素封装层级分为更基础的图形编程接口,其中包括历史悠久且使用广泛的OpenGL[34],及封装为可视化方法组件的开发库,例如在科学计算可视化领域应用广泛的VTK[35]。声明式编程出现时间相对较晚,其中采用图形语法思想的可视化语法,包括代表性的D3[36]及Vega[37]。可视分析语法库在可视化语法基础上增加了更多的数据统计分析功能,例如Vega-Lite[38]在Vega基础上,增加了数据聚合能力,其语法更加简练。

交互式数据可视化生成方式通过交互接口,使得用户不用编程即可定制可视化图表。其中,Polaris[39]与Tableau[40]采用数据列拖选方式交互生成可视化图表;Voyager[41]和Data2Vis[42]则提供了根据数据自动生成可视化图表的能力,如图10所示,引入目前流行的LSTM深度学习方法,训练数据到图表JSON脚本的"翻译"模型,实现了基本的数据至图表的自动化映射。Lyra[43]和Data Illustrator[44]则提供了一种类似于Visio的交互方式,从图表元素拼装图表的能力。VisFlow[45]在提供多个可视化模板的基础上采用数据流的思想,将可视化图表作为数据处理流的中间步骤,通过可视化的方式进行数据处理。

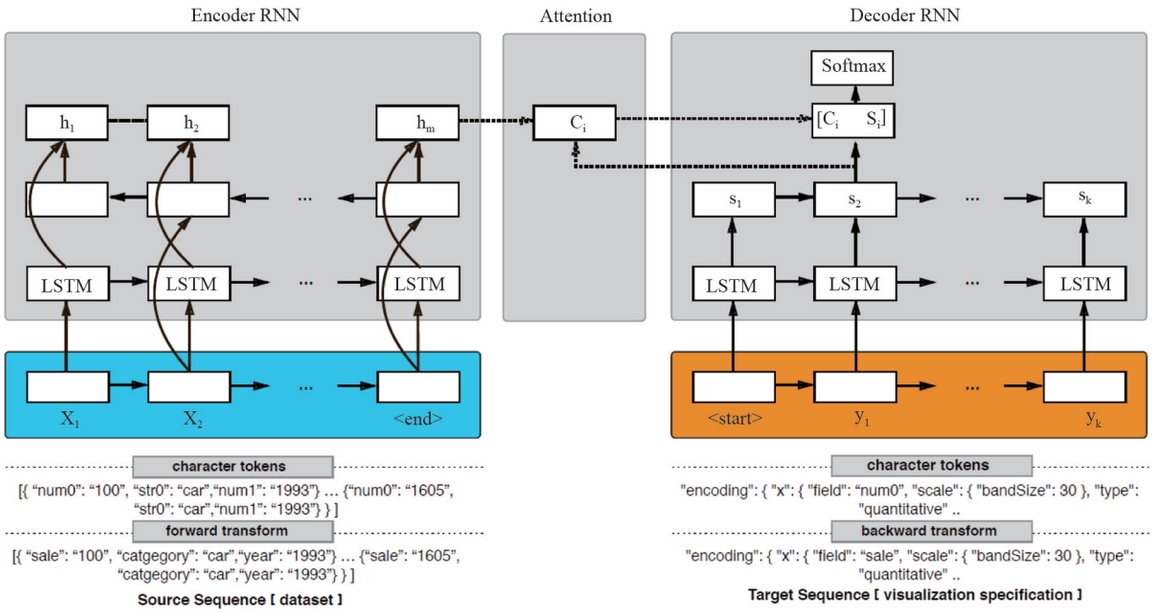


图10 Data2Vis数据可视化自动生成模型[41]

以上从方法学的角度介绍了常用的可视化生成方式,包括近期具有代表性的研究工作。整体而言,编程方式的优点在于丰富的表现能力与个性化定制能力,缺点是缺乏直观性,要求使用者具有编程能力,且相对需要更多的人力与时间成本;相对而言,交互方式的优点在于直观,用户无需编程即可定制图表,使用更为广泛,缺点是表达能力有限,系统功能和性能常常无法满足使用者个性化需求。

3 大数据可视化产品

第2节介绍了已有的数据可视化方法与工具,在此基础上重点面向大数据,介绍相关的大数据可视化产品,包括适用于一定大数据场景的传统数据可视化产品及面向大数据的数据可视化产品。

3.1 传统数据可视化产品

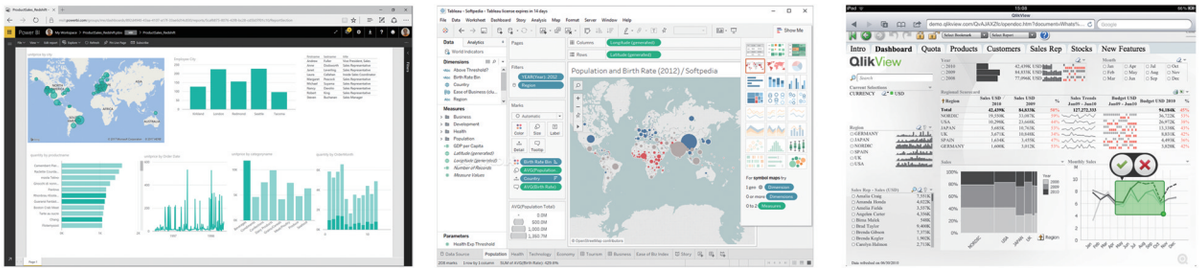
目前,存在诸多数据可视化产品,特别是结合

具体应用领域产生了多个使用广泛的数据可视化软件(图11)。例如计算流体力学领域的Tecplot、Ensight和ParaView。限于篇幅,本节重点介绍与大数据相关的通用数据可视化产品,不再讨论面向领域的相关产品。

PowerBI[46]作为微软推出的数据可视化产品,在2019年的Gartner BI象限中排名首位。优点在于易用性,交互方式类似于Excel。可以与微软Azure、Excel、SQL Server等产品快速集成。通用性突出,支持常用的云端和本地数据源;缺点在于性能相对较弱,缺少数据准备于清洗工具。

Tableau[40]基于关系型代数理论研发,随着关系型数据的普及应用而不断发展,是目前使用最为广泛的数据可视化产品之一。优点在于基于拖放的交互方式,丰富的功能以及支持Hadoop和Google BigQuery等大数据平台;缺点是仅支持结构化数据,大数据实时响应较慢,权限约束有限。

QlikView[47]为新兴的数据可视化产品,使用越来越广泛。优点在于数据关联查询与钻取能力,图表绘制快速;缺点在于易用性不足,作为内存型的数据可视化产品,数据处理速度依赖于内存大小,对硬件要求较高。



(a) PowerBI　(b) Tableau　(c) QlikView

图11　已有的数据可视化产品[40,46-47]

3.2　面向大数据的可视化产品

大数据背景下产生的数据可视化产品(图12)如下。

Apache Superset[48]是基于Flask-Appbuilder构建的开源数据可视化系统,B/S架构,集成了地图、折线图、饼图等可视化方法,提供了一种方便的看板定制方法。优点是系统可扩展性与权限控制机制;缺点是系统稳定性和大数据处理能力不足。

Apache Zeppelin[49]是面向大数据的交互式数据分析与协作记事本工具,开源项目,B/S架构。优点是与不同大数据框架的集成能力与系统可扩展性;缺点是需要编程,不支持异步,对于大规模数据,客户端可能需要等待较长时间。

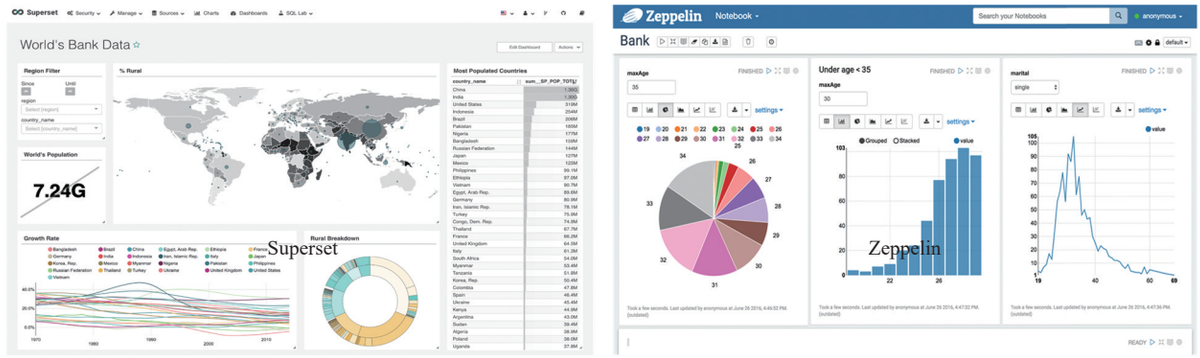


图12　面向大数据的可视化产品[48-49]

4 大数据可视化挑战

数据可视化在大数据场景下面临诸多新的挑战[50-53],包括数据规模、数据融合、图表绘制效率、图表表达能力、系统可扩展性、快速构建能力、数据分析与数据交互等。以下分别进行讨论。

1)数据规模。大数据一方面规模大,另一方面,价值密度降低。于此同时,受限于屏幕空间,所能显示的数据量有限。因此,为了有效显示使用者所关注的数据和特征,需要采用有效的数据压缩方法。目前已有的方法针对数据本身进行采样或聚合,未考虑数据可视化的显示特性。近期一些学者提出了针对特定可视化场景的数据压缩方法,如2.4节所述。但是目前依然缺少通用的面向可视化的数据压缩方法,也缺少实际应用的产品。

2)数据融合。大数据的另一个表现是数据类型多样,常常分布于不同的数据库。如何融合不同来源、不同类型的数据,为使用者提供统一的可视化视角,支持可视化的关联探索与关系挖掘,是一个重要的问题。其中涉及数据关联的自动发现、多类型数据可视化、知识图谱构建等多个技术问题。

3)图表绘制效率。随着数据规模的增加,图表可视化的效率问题越来越凸显。目前,有些可视化产品开始采用WebGL借助GPU实现平行绘制,例如Deck.gl[54]实现了地理空间数据的高效绘制。目前越来越多的数据可视化产品采用B/S架构,其性能一定程度上优先于浏览器;另外,由于跨终端需求越来越普遍,也对图表绘制提出了更多挑战。

4)图表表达能力。随着产生数据的来源增加,数据类型不断增加,与此同时,数据使用者对于数据的交互需求越来越多,已有的数据可视化产品完全无法满足使用者的可视化需求,时常出现需要的可视化形式产品不支持或支持不够等问题。这就对于系统的图表表达能力提出了更高的要求,同时对于系统支持使用者的个性化定制提出了新的要求。

5)系统可扩展性。大数据对于数据可视化系统的扩展能力提出了新的挑战,例如不同数据源,数据处理算法、可视化图表、交互方式及系统布局等多个维度的可扩展性提出了挑战。系统的可扩展性将成为衡量一个大数据可视化系统的重要指标。如3.2节所述,从近期研发的大数据可视化产品也可以发现,越来越多的产品开始将可扩展性作为系统架构设计与实现的核心要求。

6)快速构建能力。大数据伴随着快速变化与增加的数据,如何帮助用户及时理解数据,发现问题,离不开数据可视化的快速构建能力,即根据使用者数据驱动的图表快速定制能力。数据在s级甚至ms级更新的情况下,有没有可能实现图表的秒级更新与快速定制。另外,图表定制后的快速共享与响应功能也将成为必要的系统功能。

7)数据分析。传统的BI工具主要集中在数据筛选、聚合及可视化功能,已经不能满足大数据分析的需求,Gartner提出了“增强分析”[55],数据可视化只有结合丰富的大数据分析方法,例如,采用LDA[56]进行流场分类,区分不同的流场空间,如图13[57]所示。将数据的探索式分析形成一个闭环,才能实现完整的大数据可视化产品,有效帮助使用者理解数据。预测性分析是大数据的趋势,数据可视化有效结合预测方法,将有助于使用者的决策。

8)数据交互。大数据可视化与传统数据可视化的不同将是,使用者将不再仅仅是图表的受众,满足于看,而是要通过可视化与图表背后的数据和处理逻辑进行交互,由此反应使用者的个性化需求,帮助用户用一种交互迭代的方式理解数据。另外,在传统的交互手段基础上,更加自然的交互方式,例如立体显示与高自由度交互(图14)[3,58-60]、语音交互,将有助于使用者与数据更好的交互,也有助于拓展大数据可视化产品的使用范围与应用场景。

大数据可视化技术与产品所面临主要挑战的同时也对其发展带来了新机遇,例如Yu等[61]提出的面向数据流式可视化的自然语言交互接口,如图15所示,通过自然语言与可视化常见操作的映射实现。微软Excel软件集成自然语言交互,如图16所示,其中的AnnaParser算法[62]将数据表进行抽象并结合表格知识理解实现语义理解。

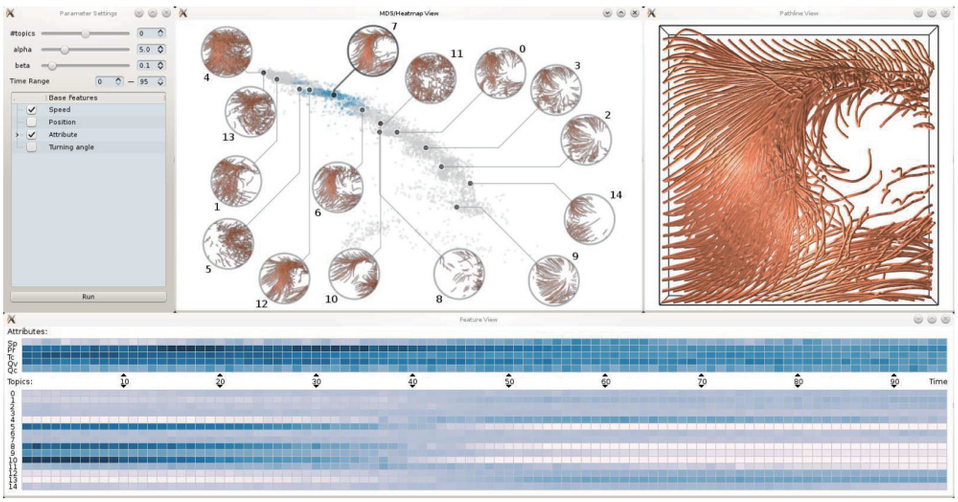


图13　流体数据分析与可视化[57]

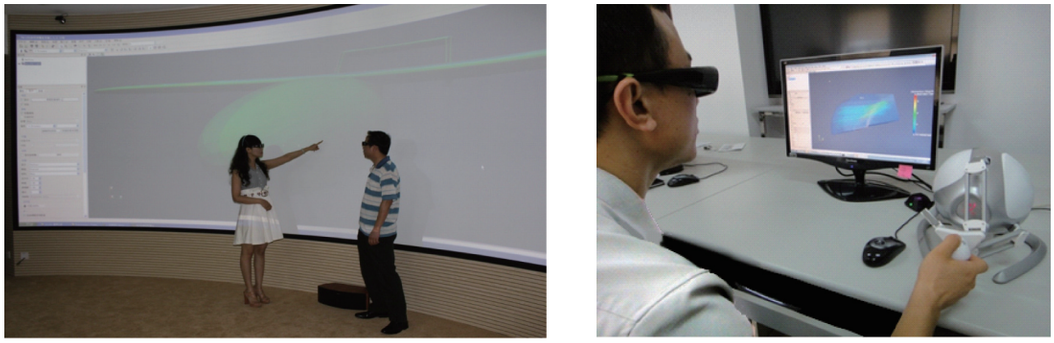


图14　立体显示与六自由度交互[3]

图15　FlowSense自然语言交互[54]

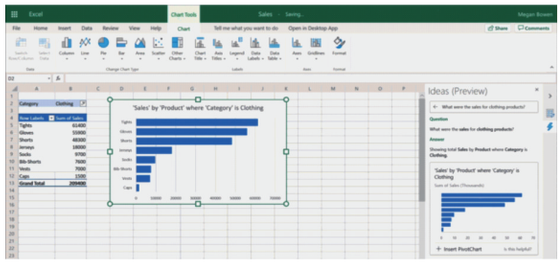


图16　Excel自然语言交互[57]

5　AutoVis

如前所述,大数据可视化面临一系列挑战。为此,课题组自主研发了数据感知的交互式可视化设计平台AutoVis,目标是让大数据的可视化过程更加简单,如图17所示。AutoVis的核心是辅助使用者快速完成从数据到图表的设计过程,包括数据定义、图表设计、映射过程、图表交互与看板服务。

1）数据定义。AutoVis支持IoTDB、PostgreSQL、MySQL、SQL Server、SQL Lite等常用数据库类型,以及提供RESTful API接口的数据服务。设计实现了抽象数据集构建与计算技术,支持不同

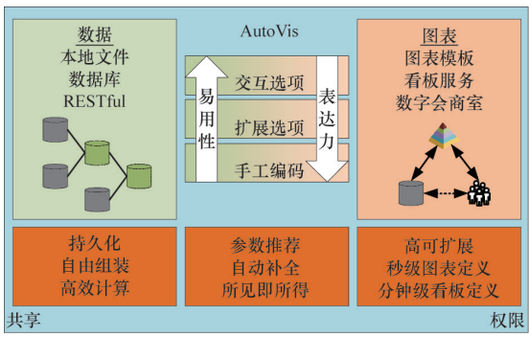


图17 数据感知的交互式可视化设计平台AutoVis

数据的自由组合,通过抽象数据集归一化,实现数据集的快速生成。

2)图表设计。AutoVis采用模板化思想,提供了百余个覆盖常用可视化技术的图表模板,支持即时模板扩展及拖拽即用,达到秒级图表定义。另外,AutoVis提供了所见即所得的图表组合定制看板能力,实现了分钟级看板定义。

3)映射过程。为了达到图表定制易用性的同时实现实时可扩展性,即融合编程方式的表达能力和交互方式的易用性,AutoVis设计实现了3种互补的数据至图表的映射方式:交互选项、扩展选项、手工编码。交互选项提供了图表常用选项交互定义窗口;扩展选项在交互选项的基础上支持用户扩展交互内容,实现图表的个性化定制;手工编码方式提供了一种在线脚本扩展能力,实现用户针对特定数据快速实现新型图表模板。

4)图表交互。图表交互能力在大数据场景下愈发重要。AutoVis的图表模板提供了常用的交互功能,包括点选、悬浮、刷选等。另外,AutoVis还实现了看板图表的自动关联,支持跨图表跨数据的钻取能力。

5)看板服务。AutoVis在支持常用的看板链接共享基础上,提供了看板服务能力,即使用者不仅可以将看板共享,或集成到其他系统,还可以动态向看板传递参数,动态调整看板可视化内容。另外,AutoVis围绕看板提供了“数字会商室”功能,使用者可以围绕数字看板进行数据驱动的讨论与决策。

AutoVis平台目前已应用于多个应用场景。例如,图18显示了AutoVis知识图谱可视化模板的一个应用,该图实现了某企业内部信息系统的交互式可视化。AutoVis内置了新型图布局算法,通过该算法解决了知识图谱多子图布局问题。结合新颖的EMOJI图形化节点设计,实现了知识图谱的合理布局与显示。此外,提供了节点悬浮查看详细信息、节点单击点选、平移调整画布、滚轮缩放图表以及图例交互筛选等交互功能。

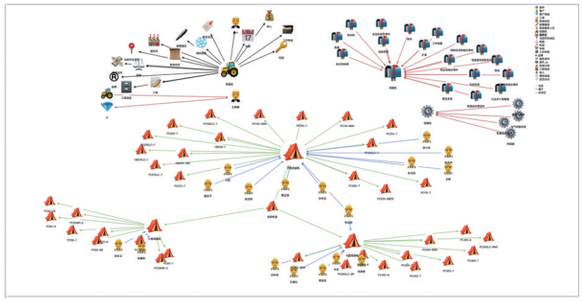


图18 某企业内部知识图谱交互式可视化

图19显示了AutoVis在地铁场景的应用,基于内置的折线图、柱状图及热力图模板,快速构建了该看板,折线图显示了一天不同时刻地铁人流情况;柱状图显示了每天该地铁线路人流总数;热力图显示了每天不同车次的人流情况,颜色从蓝色至黄色代表人数增加。另外,由于地铁运维人员重点关注车辆是否存在超载情况,通过AutoVis的热扩展能力,在热力图基础上,增加了超载标记,如图中红色圆圈所示。当某天某车次出现超载时,所对应的表格将显示红色圆,圆的半径代表超载次数。

图19 地铁断面客流统计监控

6　结论

数据可视化已成为人理解数据的重要途径,在大数据时代,数据量的增加和快速变化,人们更加需要有效的数据可视化工具直观分析大规模数据,快速捕捉数据变化。因此,其应用越来越广泛。但是相对传统的数据可视化,大数据也带来了数据规模、数据融合、图表绘制效率、图表表达能力、系统可扩展性、快速构建能力、数据分析与数据交互等多个方面的挑战。有效应对这些挑战将有助于大数据可视化随着大数据和数据科学的普及,推动其应用到更多领域。

参考文献(References)

[1] Hey T, Tansley S, Tolle K. The fourth paradigm: Data-intensive scientific discovery[J]. Proceedings of the IEEE, 2009, 99(8): 1334-1337.

[2] Shen E Y, Xia J Z, Cheng Z Q, et al. Model-driven multicomponent volume exploration[J]. Visual Computer, 2015, 31(4): 441-454.

[3] 沈恩亚, 王攀, 李思昆, 等. 大规模数据并行可视化与交互环境[C]//2012全国高性能计算学术年会论文集. 北京: 中国计算机学会, 2012: 1-7.

[4] Shen E, Wang Y, Li S. Spatiotemporal volume saliency[J]. Journal of Visualization, 2016, 19(1): 157-168.

[5] McAfee A, Brynjolfsson E, Thomas H, et al. Big data: The management revolution[J]. Harvard Business Review, 2012, 90(10): 60-68.

[6] Doctorow C. Big data: Welcome to the Petacentre[J]. Nature, 2008, 455(7209): 16-21.

[7] Reichman O J, Jones M B, Schildhauer M P . Challenges and opportunities of open data in ecology[J]. Science, 2011, 331(6018): 703-705.

[8] Rosenblum L D. See what I'm saying: The extraordinary powers of our five senses[M]. London: W.W. Norton & Company Ltd., 2011.

[9] Foley T A, Lane D A, Nielson G M, et al. Scientific Visualization[J]. IEEE Computer Graphics and Applications, 1990, 10(1): 32-40.

[10] Ware C. Information visualization: Perception for design [M]. San Francisco: Morgan Kaufmann Publishers Inc., 2012.

[11] Keim D, Andrienko G, Fekete J D , et al. Visual analytics: Definition, process, and challenges[M]//Information Visualization. Berlin: Springer, 2008.

[12] Chang W L, Grady N. NIST big data interoperability framework: Volume 6, big data taxonomies[R]. Gaithersburg: NIST, 2019.

[13] Abela A. Advanced presentations by design: Creating communication that drives action[M]. New York: John Wiley & Sons, 2008.

[14] Ahrens J, Brislawn K, Martin K, et al Large-scale data visualization using parallel data streaming[J]. IEEE Computer Graphics and Applications, 2001, 21(4): 34-41.

[15] Singh J P, Gupta A, Levoy M. Parallel visualization algorithms: Performance and architectural implications[J]. Computer, 1994, 27(7): 45-55.

[16] Moreland K. A survey of visualization pipelines[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(3): 367-378.

[17] Ma K L. In situ visualization at extreme scale: Challenges and opportunities[J]. IEEE Computer Graphics and Applications,2009, 6: 14-19.

[18] He W, Wang J, Guo H, et al. Insitunet: Deep image synthesis for parameter space exploration of ensemble simulations[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 26(1): 23-33.

[19] Ahrens J, Jourdain S, O'Leary P, et al. In situ MPAS-ocean image-based visualization[J/OL]. [2019-10-31]. http://sc14.supercomputing.org/sites/all/themes/sc14/files/archive/sci_vis/sci_vis_files/svs105s3-file4.pdf.

[20] Ahrens J, Jourdain S, O'Leary P , Patchett J , et al. An image-based approach to extreme scale in situ visualization and analysis[C]//SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway N J: IEEE, 2015: 10.1109/SC.2014.40.

[21] Dutta S, Chen C M, Heinlein G, et al. In situ distribution guided analysis and visualization of transonic jet engine simulations[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 23(1): 811-820.

[22] Di S, Cappello F. Fast error-bounded lossy HPC data compression with SZ[C]//2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Piscataway N J: IEEE, 2016.

[23] Lakshminarasimhan S, Shah N, Ethier S, et al. Isabela for effective in situ compression of scientific data[J].

Concurrency and Computation: Practice and Experience, 2013, 25(4): 524-540.

[24] Bremer P T, Weber G, Tierny J, et al. Interactive exploration and analysis of large scale simulations using topology-based data segmentation[J]. IEEE: Transaction on Visualization and Computer Graphics, 2011, 17(9): 1307-1324..

[25] The data visualisation catalogue[EB/OL]. [2019-11-08]. https://datavizcatalogue.com/search/time.html.

[26] Morrow B, Manz T, Chung A E, et al. Periphery plots for contextualizing heterogeneous time-based charts[J]. arXiv, 2019: 1906.07637.

[27] Tominski C, Aigner W. The timeviz browser[M/OL]. [2019-09-10]. https://vcg.informatik.uni-rostock.de/~ct/timeviz/timeviz.html, 2017.

[28] Shneiderman B. Extreme visualization: Squeezing a billion records into a million pixels[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008. New York: ACM, 2008, doi: 10.1145/1376616.1376618.

[29] Steinarsson S. Down sampling time series for visual representation[R/OL]. [2019-10-31]. https://skemman.is/bitstream/1946/15343/3/SS_MSthesis.pdf.

[30] Kehagias A. A hidden markov model segmentation procedure for hydrological and environmental time series[J]. Stochastic Environmental Research and Risk Assessment, 2004, 18(2): 117-130.

[31] Guo T, Feng K, CongG, et al. Efficient selection of geospatial data on maps for interactive and visualized exploration[C]//Proceedings of the 2018 International Conference on Management of Data. New York: ACM, 2018, doi: 10.1145/3183713.3183738.

[32] Wu Y, Cao N, Archambault D, et al. Evaluation of graph sampling: A visualization perspective[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 23(1): 401-410.

[33] Zhang J, Zhu K, Pei Y, et al. Clustering-structure representative sampling from graph streams[C]//International Conference on Complex Networks and their Applications. Berlin: Springer, 2017, doi: 10.1007/978-3-319-72150-7_22.

[34] Woo M, Neider J, Davis T, et al. OpenGL programming guide: The official guide to learning OpenGL, version 1.2 [M]. Boston: Addison-Wesley Longman Publishing Co. Inc., 1999.

[35] Schroeder W, Martin K, Lorensen B. The visualization toolkit: An object-oriented approach to 3D graphics[J]. Upper Saddle River: Prentice Hall Inc., 1998.

[36] Bostock M, Ogievetsky V, Heer J. D^3 data-driven documents[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2301-2309.

[37] Satyanarayan A, Russell R, Hoffswell J, et al. Reactive vega: A streaming dataflow architecture for declarative interactive visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 659-668.

[38] Satyanarayan A, Moritz D, Wongsuphasawat K, et al. Vega-Lite: A grammar of interactive graphics[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 341-350.

[39] Stolte C, Tang D, Hanrahan P. Polaris: A system for query, analysis, and visualization of multidimensional relational databases[J]. IEEE Transactions on Visualization and Computer Graphics, 2002, 8(1): 52-65.

[40] Tableau Inc[EB/OL]. [2019-11-08]. https://www.tableau.com/.

[41] Wongsuphasawat K, Moritz D, Anand A, et al. Voyager: Exploratory analysis faceted browsing of visualization recommendations[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 649-658.

[42] Dibia V, Demiralp Ç. Data2vis: Automatic generation of data visualizations using sequence to sequence recurrent neural networks[J]. arXiv, 2018: 1804.03126.

[43] Satyanarayan A, Heer J. Lyra: An interactive visualization design environment[J]. Computer Graphics Forum, 2014, 33(3): 351-360.

[44] Liu Z, Thompson J , Wilson A, et al. Data illustrator: Augmenting vector design tools with lazy data binding for expressive visualization authoring[C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2018, doi: 10.1145/3173574.3173697.

[45] Yu B W, Silva C T. Visflow-web-based visualization framework for tabular data with a subset flow model[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 251-260.

[46] Microsoft Inc[EB/OL]. [2019-11-08]. https://powerbi.microsoft.com/.

[47] Qlik Inc[EB/OL]. [2019-11-08]. https://www.qlik.com/us/products/qlikview.

[48] Apache software foundation[EB/OL]. [2019-11-08]. https:

//superset.incubator.apache.org/.

[49] MadhaviLatha A, Vijaya K A. Streaming data analysis using apache cassandra and zeppelin[J]. IJISET-International Journal of Innovative Science, Engineering & Technology, 2016, 3(10), http://ijiset.com/vol3/v3s10/IJISET_V3_I10_02.pdf.

[50] Wang L, Wang G, Alexander C A. Big data and visualization: Methods, challenges and technology progress[J]. Digital Technologies, 2015, 1(1): 33-38.

[51] Agrawal R, Kadadi A, Dai X, et al. Challenges and opportunities with big data visualization[C]//International Conference on Management of Computational & Collective Intelligence in Digital Ecosystems. New York: ACM, 2015.

[52] Ali S M, Gupta N, Nayak G K , et al. Big data visualization: Tools and challenges[C]//2nd International Conference on Contemporary Computing and Informatics (IC3I). Piscataway NJ: IEEE, 2016, doi: 10.1109/IC3I.2016.7918044.

[53] Bikakis N. Big data visualization tools[J]. arXiv, 2018: 1801.08336.

[54] Wang Y. Deck.Gl: Large-scale web-based visual analytics made easy[J]. arXiv, 2019: 1910.08865.

[55] Gartner Inc[EB/OL]. [2019-11-08]. https://www.gartner.com/en/information- technology/glossary/augmented- analytics.

[56] Balakrishnama S, Ganapathiraju A. Linear discriminant analysis—A brief tutorial[J]. Institute for Signal and Information Processing, 1998, 18: 1-8.

[57] Hong F, Lai C, Guo H, et al. Flda: Latent dirichlet allocation based unsteady flow analysis[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20 (12): 2545-2554.

[58] Shen E, Z Cheng, J Xia, and S Li. "Intuitive volume eraser[C]//1st International Conference on Computational Visual Media. Berlin: Springer, 2012: 10.1007/978-3-642-34263-9_32.

[59] Shen E , Li S , Cai X , et al. SAVE: saliency-assisted volume exploration[J]. Journal of Visualization, 2015, 18 (2): 369-379.

[60] Shen E, Li S, Cai X, Zeng L, et al. Sketch-based interactive visualization: a survey[J]. Journal of Visualization, 2014, 14(4): 275-294.

[61] Yu B, Silva C T. Flowsense: A natural language interface for visual data exploration within a dataflow system [J]. IEEE Transactions on Visualization and Computer Graphics, 2019, doi: 10.1109/TVCG.2019.2934668.

[62] Gao Y, Lou J, Zhang D. Annaparser: Semantic parsing for tabular data analysis[J]. arXiv, 2019: 1910.10363.

Big data visualization technology and applications

SHEN Enya

School of Software, Tsinghua University; National Engineering Laboratory of Big Data System Software, Beijing 100084, China

Abstract With the growth of data generated by human activities, the scale, the type and the demands for the data visualization have expanded greatly. In the big data era, the data visualization faces many challenges. In this paper, based on the characteristics and the requirements of the big data, and the current research states of the data visualization, the common data visualization techniques are reviewed. Eight important challenges that the data visualization has to deal with in the big data applications are highlighted. The AutoVis, a data-aware interactive visualization design platform, is specially discussed, as well as its applications.

Keywords data visualization; big data; time-varying data visualization; visualization system

（责任编辑　刘志远）