大数据时代的可视化与协同创新

屈华民

大数据时代的来临

我们毫无疑问已经处在一个大数据的时代。各行各业都在快速产生和积累数据。广泛认为大数据具有四个特点，也就是大数据的四个V：数据量[Volume]惊人，前所未有；这些数据产生的速度[Velocity]极快；而数据的品种[Variety]非常多，包括文本、图像、视频、传感器等多种数据种类；并且，这些数据里面包含潜在的价值[Value]。关于数据对人类历史的影响可以从下面三段引言中看出。

“我怀疑二百年后的后人书写我们这段历史时，他们会发现我们所处时代人类思考方式发生了重大的变化，那就是我们比以往的任何时代在更多的事情上变得更加理性，我们更多以数据为依据分析思考问题。”（哈佛前校长 Lawrence Summers）

“得益于计算机技术和海量数据库的发展，个人在真实世界的活动得到了前所未有的记录……社会科学将脱下‘准科学’的外衣，在21世纪全面迈进科学的殿堂。”（雅虎首席科学家沃茨）

“大数据的影响，就像四个世纪前人类发明的显微镜一样……而大数据，将成为我们下一个观察人类自身社会行为的‘显微镜’。”（麻省理工[MIT]教授Erik Brynjolfsson）

从数据，到海量数据，再到大数据，对人类的做事和思维方式都有很大的影响。在《大数据时代：生活、工作与思维的大变革》[1]一书中，笔者将其归结为三个特点：（1）更多：不是随机样本，而是所有的数据；（2）更杂：不是精确性，而是混杂性；（3）更好：不是因果关系，而是相关关系。对大数据的研究涉及计算机、数学、生物学等多个领域。大数据尤其是对数据存储、数据挖掘等提出了重大挑战。而数据的可视化也将在大数据时代扮演一个重要的角色。设计正是数据可视化中非常重要的一环，需要计算机科学家和设计师紧密合作，协同攻关。本文讨论什么是数据可视化，数据可视化和平面设计的关系，并用多个案例来示例说明数据可视化系统是如何设计的。

可视化在大数据时代的作用

在美国奥巴马政府的大数据计划中，专门有一个项目：“向一个研究培训小组发放200万美元的奖金，用于支持一项大学生培训计划，教授他们如何利用图形和可视化工具解析复杂数据。”数据可视化的重要性由此可见一斑。 在《大数据》[2]一书中，在讨论商务智

1 [英]维克托·迈尔舍-恩伯格［Vikitor Mayer-Schönberger］、肯尼思·库克耶［Kenneth Cukier］著，盛扬燕、周涛译，《大数据时代：生活、工作与思维的大变革》，浙江人民出版社，2012年。

2 涂子沛著，《大数据:正在到来的数据革命，以及它如何改变政府、商业与我们的生活》，广西师范大学出版社，2013年。

能时，专门提到数据可视化的“化蝶”作用：

数据可视化把美学的元素带进了商务智能。一幅好的数据图像不仅能有效地传达数据背后的知识和思想，而且华美精致，如一只只振动翅膀的彩蝶，刺激视觉神经，调动美学意识，留下栩栩如生的印象。

数据可视化的这种“导航”作用也极大地推动了商务智能的大众化。通过把复杂的数据转化为直观的图形，并呈现给最普通的用户，商务智能已经不再是少部分高级分析人员的专利，而是贴近大众生活，浅显易懂，人皆可用的工具和手段。

对大数据的分析不外乎两种类型：使用机器（尤其是计算机），利用复杂精妙的算法进行自动分析，或者是让人利用他们的领域知识进行交互式的分析。如果人成为数据分析的重要一环，那么有必要为人提供直观易懂的界面，来帮助人了解数据里面隐藏的信息。这种界面往往就是数据可视化系统。

什么是可视化?

可视化，简要地讲，就是把数据转换为图形图像的方式，帮助人们理解大量的和复杂的数据。可视化有三个主要的分支：科学可视化、信息可视化、可视分析。科学可视化，主要研究如何可视化科学研究中产生的大量数据，比如流体动力学模拟产生的数据，医学图像如CT/MRI数据，向量场和张量场等。这些数据本身往往包含在真实世界中存在的几何结构。信息可视化主要研究的是抽象数据如文本、图像、网络、股票、社交媒体等。这些数据本身并没有看得见摸得着的几何结构。人们只是把它们转换为图形图像的方式便于理解。最近兴起的可视分析更多地集成了数据挖掘等自动算法，加重了系统中的分析含量。可视化的目的可以概括为记录信息，分析推理，证实假设，交流思想。[3]

很多时候，可视分析和数据挖掘的最终目标是一致的，即理解数据。但数据挖掘更偏重于研究各种自动算法来充分利用计算机的强大计算能力，而可视化则更偏重于设计交互的图形展示，以便利用人的强大的视觉处理能力和领域知识。我个人认为，可视化和数据挖掘之间的关系就像风景照片里面山与水的关系。就像一个好的风景往往同时包含山和水，一个好的大数据解决方案必然同时拥有强大的数据挖掘能力和充满灵气的可视化展示。

可视化中的美学元素

1.可视化的“信达雅”

可视化也可以看作是一种翻译，即将数据（语言）翻译成图形图像（语言）。大家知道，翻译的最高标准是“信达雅”。严复提

3 陈为、张嵩、鲁爱东著，《数据可视化的基本原理与方法》，科学出版社，2013年。

出：“翻译作品内容忠实于原文谓信，文辞畅达谓达，有文采谓雅。”同理， 可视化系统也要做到“信达雅”，力求忠实、有效和优美。

信：从数据转化到可视表示时不歪曲，不误导，不遗漏。也就是说，可视化系统要忠实地反映数据里面包含的信息。

达：可视化的表现方式自然有效，清楚易用，容易上手，帮助用户达成目标。也就是说，可视化系统要有效地帮助用户找到有用的信息。

雅：系统要充满美感，给用户优雅的体验。也就是说，系统一定要优美。

2.可视化中美的含义

可视化的主要目的是展示数据中隐藏的知识。另一方面，可视化呈现也需要美观。在《可视化之美》[4]这本书里，提到什么是漂亮的可视化。漂亮的可视化有下面这些标准：

（1）**美感**：美感很难形容，但你看到了，你就会知道。

（2）**新颖**：普通的图形表示很难让人兴奋，它们已经变成了陈词滥调。漂亮的可视化，往往有新奇的元素，能让人兴奋。

（3）**简单有效**：没有太多华而不实的元素。能有效地表达出数据里的故事。简单有效[Simple and effective]就是所谓的科技的优雅。

书中也提到：“可视化中的美学概念远远不止是漂亮的图片。当然，使用舒心是一项重要且一直被低估的因素……但是，正如史蒂夫·乔布斯的一句名言：‘设计不在于产品的外观和感觉，而是它如何工作。’一个真正的审美可视化，除了必须美丽外，而且必须能够表达现有的潜在隐含特征，并能够激励用户、读者去探索更丰富多彩的世界。”这些都是很有见地的看法。

可视化和设计的关系

可视化系统的设计经常需要遵循一些原则。这些原则有些是从别的领域（如人机交互）借鉴过来的，有些则是大量实践的过程中总结出来的。其中设计领域中的很多原则都可以在可视化系统的开发中得到应用。事实上，很多可视化系统本身就是直接受到平面设计作品，尤其是信息图[Infographics]的启发。在可视化领域，有一类研究就是如何自动生成信息图[Infographics]。下面试举一例。

图1显示的是设计师兰德尔·芒罗[Randall Munroe]手绘的《星球大战》电影的故事线[Storyline]。[5]里面显示了《星球大战》电影中的主要角色和一些角色共同出现的场景。水平轴显示的是时间。每一个线条代表一个演员。如果两个或多个演员共同出现在一个场景中，那么这些线条就会画得很靠近，并且一起穿过一个代表该场景的色块。这个信息图无疑非常直观而

4 朱莉·斯蒂尔[Julie Steele]等编，祝洪凯、李妹芳译，《数据可视化之美》，机械工业出版社，2011年。

5 R. Munroe. Xkcd #657: “Movie narrative charts”, http://xkcd.com/657 , December 2009.

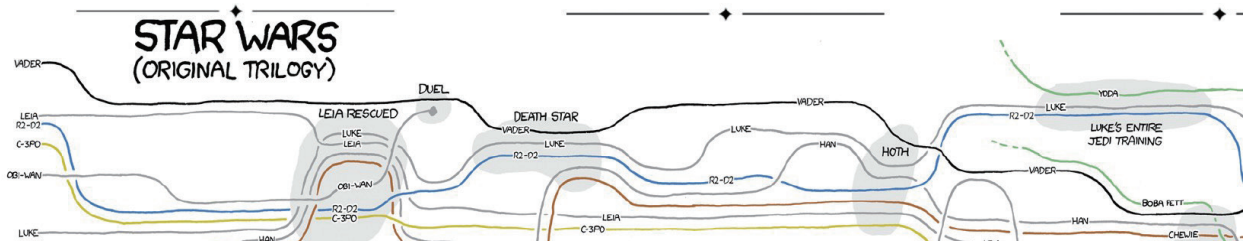

图1 设计师R. Munroe设计的星球大战电影的故事线

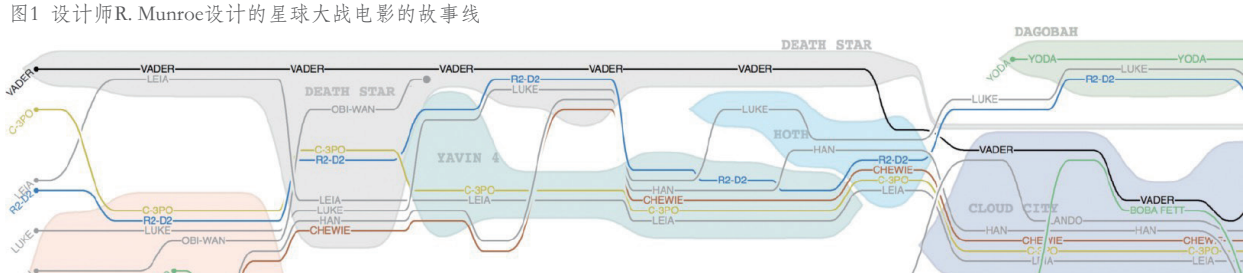

图2 可视化系统自动生成的《星球大战》电影的故事线

且包含了大量的信息。但因为原图是设计师手绘的，如果换了一个电影，又得重新绘制，比较费时。而且一般用户也不能很方便地用这样的可视化来探索他们自己喜欢的电影。所以可视化领域就有研究者开发了一个可视化系统，[6]可以自动生成这样的图。图2显示的是该可视化系统生成的《星球大战》的故事线。将两个图进行比对，可以看出，虽然自动生成的图艺术性可能稍差一些，但完全可以传达手绘图中同样的信息。这样的系统可以很方便地表现别的电影的故事线。相对于手工绘制的故事线，该可视化系统具有很好的扩展性，可以根据数据快速自动生成各种电影、电视或是小说的故事线，而且支持用户交互，并且可以很容易加载别的信息。

基于隐喻的可视化设计

可视化系统开发中最关键的一环是如何根据数据和应用来设计美观有效的视觉呈现。除了上文提到的信息图，隐喻也被广泛用来设计可视化呈现。下面介绍香港科技大学可视化小组的三个工作。这些工作广泛使用隐喻[Metaphor]来帮助人们理解可视化中图的含意。

案例一：基于钟表隐喻的交通轨迹数据可视化

图3显示的是一个出租车轨迹数据的可视化系统。[7]目前很多城市的出租车都安装了GPS系统。这些GPS系统可以提供车辆在不同时刻的位置。装有GPS系统的出租车就像移动的传感器一样，可以提供一个城市动态的交通状况和人群的移动特征。其中一个有意思的问题是如何做路径推荐。从一个地方到另一个地方经常有多个路径可选。如何根据交通状况给用户推荐一个省时不堵车的路径无疑具有很大的实用价值。一个解决方案是根据历史数据，从出租车司机在某个时间段从A地到B地所走的路径中选取一个用时最短的路径推荐给用户。因为出租车司机是最了解城市交通状况的群体，他们的

6 Yuzuru Tanahashi, Kwan-Liu Ma: "Design Considerations for Optimizing Storyline Visualizations". IEEE Trans. Vis. Comput. Graph. 18(12): 2012, pp. 2679-2688.

7 He Liu, Yuan Gao, Lu Lu, Siyuan Liu, Huamin Qu, Lionel M. Ni: "Visual analysis of route diversity". IEEE VAST 2011, pp. 171-180.

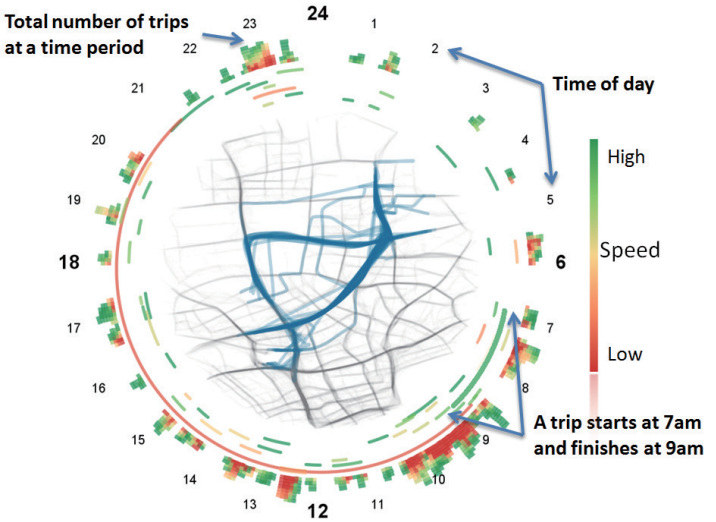

图3 用来呈现不同行车路线上不同时间段的车辆数目和堵车情况的可视化

最佳选择往往好过目前机器自动计算选择出来的路径。在我们的可视化系统中，我们希望能表现在每个时间段，选择某个路径的车辆的数目，以及这些车辆的平均速度等。这样用户就能直观地判断哪个路径在什么时间段比较好。为了表现时间，我们采用了一种基于钟表的隐喻。一天的24小时就像在钟表上一样分布在一个圆环上。每一个时间点上车辆的多少用柱状图的高低来表示。而每个柱状图中不同颜色则代表车辆行驶速度的不同。而颜色的设计也采用了红绿灯的隐喻。因为红色让大家想到红灯，所以对应于慢速行驶的车辆；而绿色则让大家想到绿灯，对应于高速行驶的车辆。从图中非常明显地能看出来早上9点到10点，车辆非常多，而且很多车的行驶速度很慢，而下午4点则车辆稀少而且速度比较快。

案例二：基于向日葵隐喻的信息传播的可视化

图4显示的是如何呈现信息在社交媒体，尤其是Twitter[推特]（或微博）上的传播。[8]信息在社交网络上的传播有三个关键因素：被传播的信息本身，传播信息的人，以及信息传播的过程及影响。为了表现这三个因素，我们基于向日葵的隐喻设计了一个可视化系统。向日葵的花盘边缘是舌状花，而花盘内侧则是管状花。管状花成熟后变成种子。这些种子可以被风、动物、或是人带到别的地方，成长为新的向日葵。信息的传播和向日葵种子的扩散有类似之处。当一个Tweet刚出来的时候，就像没有成熟的管状花，如果有人转推了这个Tweet之后，它就成熟了，代表着它有了新的生命，会离开这个向日葵的花盘，到转发它的用户那里。整个设计分成三部分：主题盘、用户组和扩散路径。主题盘类似于向日葵的圆形花盘。未成熟的管状花代表未被转发的Tweet，处在非常中心的位置。而周围成熟的管状花则代表活跃的Tweet，它们都在某个时间段被转推过。一个Tweet如果第一次被转推，它就会从中心移动到外围，并最终移动到

8 Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, Huamin Qu: "Whisper: Tracing the Spatio-temporal Process of Information Diffusion in Real Time", IEEE Trans. Vis. Comput. Graph. 18(12): 2012, pp. 2649-2658.

转推它的用户组。一段时间内没有被转推的Tweet就会慢慢消失，这样主题盘内剩下的就是和该主题有关的比较活跃的Tweet了。用户组是那些环绕在主题盘旁边的圆形图标。可以根据用户的关系或是地理位置等将类似的用户组合到一起，变成一个用户组。扩散路径则表现为那些连接了代表Tweet的管状花和代表转推了这些Tweet的用户组的路径。在主题盘和用户组之间，画有一圈圈的等值线，代表时间。每一个小的三角线段代表在某个时间被某个用户组转推的来自该主题盘的Tweet。颜色代表情感。三个预先选定的颜色（红色、橙色和绿色），分别用来表示负面、中立和正面的意见。从图中可以很直观地看出哪些Tweet比较热门，哪些Tweet在哪个时间段被哪个用户组转推过，以及这些用户组对该Tweet的情感。

案例三：基于河流隐喻的文本数据中主题的合并与分离的可视化

了解隐藏在大量文本数据中的主题的演变非常重要。它可以帮助政治家、商务人士和社会学家及时了解他们领域中的新的和热门的话题，以及这些话题是如何随着时间变化的。话题的演变包括新话题的产生，话题的分裂和合并，以及话题的消失。用户经常需要了解新话题是怎样产生的，什么触发了新话题的发展，以及它们怎么逐步瓦解或融入其他的话题。我们和微软亚洲研究院合作开放了TextFlow系统。[9]该系统用文本挖掘的方法从大量文本数据中抽出主要的话题，并且建立起它们之间可能的分裂与合并的关系。在用可视化来表现话

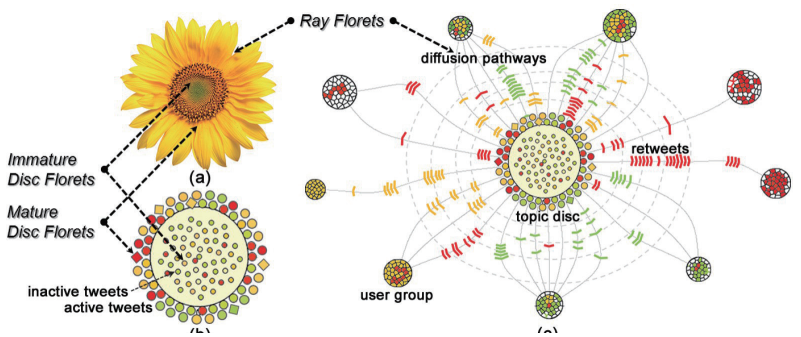

图4 用来呈现在推特[Twitter]上推文[Tweet]被不同群体转发的信息传播的可视化

题的演变时，我们使用了河流的隐喻。图5显示的我们系统中使用的可视化方法。图中的水平轴表现的是时间。图中每个不同颜色的色带代表一个话题。每一个话题就好比是一条河流，河流的宽度表示的是该话题的热度（有多少个文本在讨论这个话题），河流的分叉和合并代表话题的分裂和合并。这样，文本中包含的主要话题以及这些话题之间的关系如何随时间演变就变得一目了然了。更进一步，在关键的地方，引发主题变化的主要因素也可以被可视化出来并叠加到河流图上，便于大家更深入地了解变化的原因。

9 Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, Xin Tong: TextFlow: "Towards Better Understanding of Evolving Topics in Text", IEEE Trans. Vis. Comput. Graph. 17(12): 2011, pp. 2412-2421.

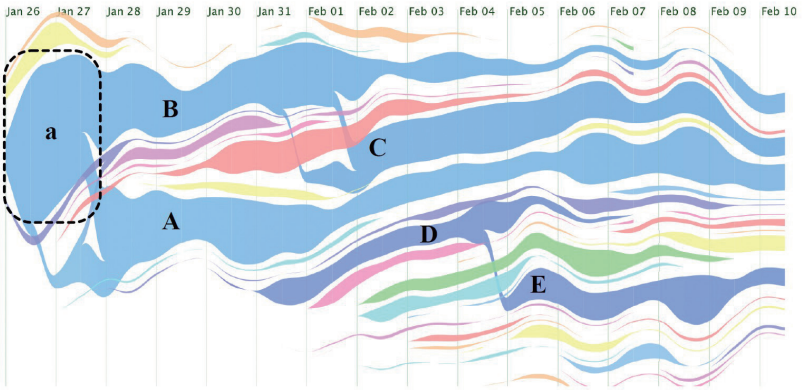

图5 用来呈现文本数据中不同主题的热度随时间变化以及主题的分裂和合并情况的可视化

大数据可视化中的一些研究课题

大数据的可视化系统开发面临一些前所未有的挑战。下面试举几例。一是大数据本身包含很多的噪音，如数据的不完整和不精确。如何将这种不完整、不精确性用可视化的方式传达给用户是一个难题。二是因为大数据本身的规模和复杂度，完全依靠专业的分析师进行分析不是完全行得通。如何利用众包的方式，以可视化为工具，让大家都成为数据分析中的一员，是一个值得研究的课题。三是大数据的“在位分析”[In-situ Visualization]。对于大量的动态数据，如何直接进行分析，而不是先把数据放到数据库里，然后再从数据库倒入内存中进行分析，是另一个非常值得探索的方向。四是异构数据（不同类型的数据如文本、视频、传感器数据）的可视化问题。

工程和艺术的协同创新

综上所述，在大数据时代，可视化变得异常重要。而可视化系统的设计与开发既需要工程方面的背景（如数据处理、挖掘，以及图形学方面的知识）又需要艺术方面的背景（如平面设计）。我国目前的教育体制，还比较难培养出兼具这两种背景的学生。目前拥有相关工程背景的研究人员和学生基本上集中在各大综合性大学或工科院校的计算机学院，而拥有相关艺术背景的人员基本集中在像中国美术学院这样的美术类院校。这就需要这两类人员密切合作，设计出既能高效直观地解决实践中的大数据问题，又能带给用户极大美感的可视化系统。从协同创新的角度讲，中国美术学院和香港科技大学的联合课程“设计思维”是一个良好的开端。