浙江农林大学

本 科 生 毕 业 设 计（论文）

外文翻译

（ 2019 届）

题　　目：产品流通的可视化表达与数据挖掘

学生姓名：倪畅

学　　号：201505010315

专业班级：计算机科学与技术 153

学　　院：信息工程学院

指导教师：陈文辉　职称：讲师

2018 年 12 月 23 日

数据可视化方法在大数据应用中的分析评述

Evgeniy Yur'evich Gorodov and Vasiliy Vasil'evich Gubarev

新西伯利亚国立技术大学，St.Karla Marksa，Novosibirsk，俄罗斯

摘要：本文从数据表示和可视化两个方面阐述了大数据这一术语。在大数据可视化中，存在一些特定的问题，因此有对这些问题的定义和一组避免这些问题的方法。同时，我们回顾了现有的数据可视化方法在大数据应用中，并考虑到所描述的问题。对结果进行了总结，提出了在大数据量应用中的可视化方法的分类。

关键词：大数据；数据可视化；ECharts

1 介绍

客户需要处理二级数据，而二级数据并不直接与客户业务相连，这导致了所谓的大数据（Big Data）现象。我们将提供大数据项的定义。正如 Gubarev Va.y Vasil'evich 所提到的，大数据是没有明确边界的现象，并且可以以无限甚至无限的数据累积形式呈现。而且，累积的数据可以以各种数据格式呈现，其中大多数不是结构化数据流。

通常，在“大数据”这个术语下，我们理解一个大数据集，其容量呈指数增长。对于关系数据库理论中使用的经典数据处理方法，这个数据集可能太大、太“原始”或太非结构化。该数据的应用领域[1]。

它用于在不同的分析文献源中提供以下大数据属性：大数据量(Volume)、多格式数据表示(Variety)和高数据处理速度(Velocity)。因此，现在有以下大数据类：“Volume-Velocity”类、“Volume-Variety”类、“Velocity-Variety”类和“Volume-Velocity-Variety”类。

大数据处理根本不是一件小事，它需要特殊的方法和途径。图形思维是人类一种非常简单而自然的数据处理方式，因此可以说，图像数据表示是一种有效的方法，它使数据易于理解，并为决策提供足够的支持。但是，对于大数据，大多数经典的数据表示方法变得不那么有效或者甚至不适用于具体的任务。对大数据的一个具体类的适用性分析是主题领域的一个热门问题，因为以前没有这样的案例研究。根据现有可视化方法适用于所描述的大数据类之一的标准。

为了对描述过的大数据类之一进行分类决策，需要从以下几个方面进行分析：对大容量数据的适用性、以不同数据格式表示的数据可视化的可能性，数据呈现的速度和性能。

2 大数据可视化问题

注意所描述的大数据属性，我们可以识别以下问题，让大数据可视化不是一项简单的任务。

2.1 视觉噪声

正在研究的整个数据数组的简单表示可以成为屏幕上的完全混乱，我们只看到一个大点，由点组成，表示每个数据行。这个问题来自于数据集中的大多数对象彼此相对而言，屏幕上的观察者不能将它们划分为单独的对象。所以，有时候，分析师不能连一点从整个数据可视化的有用信息，而无需任何预处理的任务。必须提到的是，在噪声在这个话题我们不了解任何数据损坏或变形，它就应该被认为是一个数据的可视性缺失现象。

2.2 大图像感知

作为上述问题的解决方案，出现了一种方法，其结论是在更大的屏幕之上的数据分布。但是，偶尔，它会导致另一个问题，即大图像感知。对于不同的数据可视化，人类有一定的感知水平。尽管图形数据可视化的这个级别要高得多，但是与表数据可视化相比，它有自己的局限性。在达到这种感知水平之后，人类就失去了从数据过载视图中获取任何有用信息的能力。所有可视化方法都受到负责可视化输出的设备分辨率的限制，因此每个可视化显示的

点数是有限的。当然，我们可以将可视化设备替换为用于局部数据可视化的现代设备或一组设备，允许我们用更多的数据点呈现更详细的图像，但是即使我们可以无限次地重复这个过程，我们也会遇到人类感知的限制。随着数据量的迅速增长，人类对数据的理解和分析将遇到困难。

因此，可以说，数据可视化方法不仅受到器件的纵横比和分辨率的限制，而且受到物理感知的限制。

2.3 信息丢失

另一方面，可以使用减少可见数据集的方法。但是，尽管解决了上述问题，这些方法导致了另一个问题，即信息丢失。这些方法以数据聚合和过滤为基础，基于一个或多个标准的具体数据集中对象的相关性。使用这些方法可以误导分析员，当他不能注意到一些有趣的隐藏对象时，有时，复杂的聚合过程会消耗大量的时间和性能资源，以获得准确和所需的信息。

2.4 高性能要求

图形分析不仅仅停留在静态图像可视化上，因此上述问题在动态可视化中变得更加重要。另外还有一个问题，在静态可视化中很难注意到，因为可视化速度要求较低——高性能要求。在行为分析任务中，分析人员通常希望访问整个数据阵列，即使不需要频繁刷新，这个过程也会消耗大量时间。它最终导致计算资源的不断增加，或者过滤越来越多的数据。通常，这两种方法由于组织、支持和经济成本高，或者在监控过程中有用的信息丢失等原因，在实际中都可以得到广泛的应用。第二种方法很难定制和适应，因为通常分析系统或分析人员不知道传入数据的性质。因此，此方法中的过滤任务只能由简单的步骤组成，例如排除每行数据或从数据中删除一些因素。

2.5 高速图像变化

最后一个问题是图像变化率高。当观察数据的人不能对数据的数量变化或显示的强度作出反应时，这个问题在监视任务中变得尤为重要。改变速率的简单降低不能提供期望的结果，因为人的反应速度直接取决于它。

通过文章的这一部分，可以说大数据可视化导致分析质量下降，这突出了本文的主题性。

3　大数据可视化途径

图形可视化方法有很多种，但多维数据可视化的研究还比较少，是一个热门的研究课题。

图形可视化已经用于人类活动的各个方面，但是随着数据量的增长和数据生产速度的提高，方法的有效性甚至适用性可能成为真正的问题。所描述的问题来自以下几点：

（1）需要人工制备数据切片，进行部分数据可视化；

（2）感知数据因素数量的视觉限制。

我们需要概述现有的数据可视化方法，并提供解决这些问题的方法。这些方法必须提供更加直观、信息丰富的数据表示，以帮助分析人员在大数据中发现隐藏的关系。大多数数据可视化方法通常不从无到有，但它们成为早期现有方法的发展。

分析工具最多必须满足以下要求：

（1）分析人员应该能够同时使用多个数据表示视图；

（2）用户与可分析视图之间的主动交互；

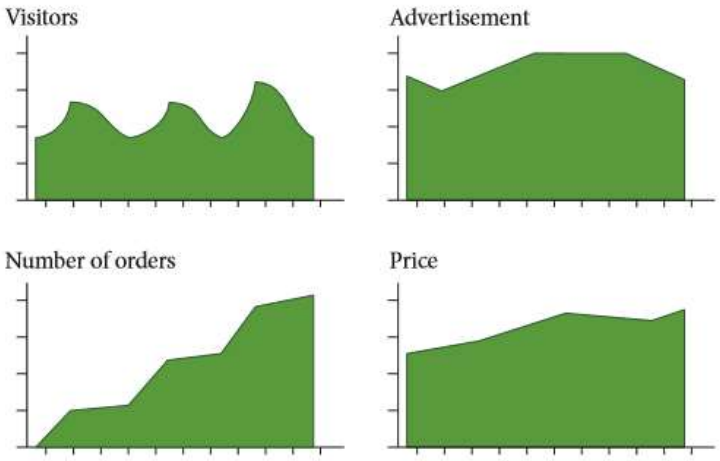

图 3-1　同一视图上的多个数据表示

（3）展望了工作过程中因素数量的动态变化。

下面我们将更清楚地描述这些需求。

3.1 每个表示显示多于一个视图

为了达到全面的数据理解，分析人员通常使用简单的方法，当他放置不同的经典数据视图时，这些视图仅包括有限的一组因素，以便他能够容易地在这些视图之间或者在一个具体视图中找到一些关系[4，5]。

尽管事实上可以完全使用数据可视化的每种方法，但是，当分析人员仅使用一些相似或接近于相似图形对象时，我们常常可以看到一种方法。例如，线性或点图（图 1）。当然，分析员可能对比较相同数据的完全不同的可视化感兴趣，但是在这种情况下，可视化分析的整个过程变得更加困难。现在，研究者不仅要比较相似的图形对象，还要根据不同的因素清楚地区分不同的数据，做出决策[6]。

因此，可以说，这种方法可以引导分析人员进入期望的位置，并提供足够的支持以在研究的第一阶段作出决定。因此，有可能出现这种情况，当这个阶段成为当前研究的最后阶段，驱使分析师远离完全错误的决定。

此外，这种方法的另一个关键点是能够在所有相关表示中选择所需的数据区域，如图 2 所示。

分析人员可能希望以各种方式协调视图：在一个视图中选择项目可以突出显示其他视图中的匹配记录，或者提供过滤标准以从其他显示中删除信息。链接导航提供了一种附加的协调形式：滚动或缩放一个视图可以同时操作其他视图[5,7]。

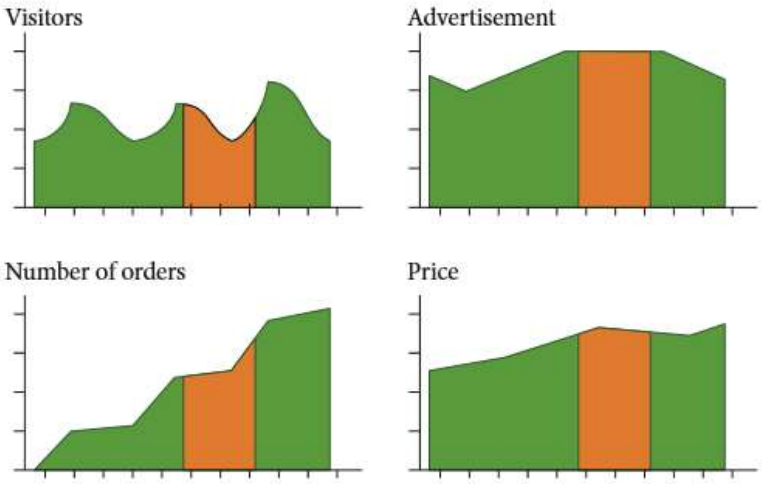

图 3-2　在相关表示上的数据区域选择。

3.2 因素数量的动态变化

也许可视化分析中最基本的操作包含在数据可视化规范中。分析人员必须指出应该显示哪些数据以及应该如何显示这些数据以减轻信息感知。

任何图形可视化都可以绝对地应用于任何数据，但是为了获得任何有用的信息，所选方法是否正确地应用于数据集始终是一个热门问题。通常，对于大数据，分析员不能观察整个数据集，不能发现其中的异常，也不能从第一眼就发现任何关系[6]。因此，另一个主题的方法是因素数量的动态变化。在分析师选择了一个因素之后，他愿意看到一个经典的直方图，该直方图显示了记录号码根据记录类型的分布。

在分析师选择了另一个因素之后，例如支持费用，图表类型也变成了点图。图 3 的底部显示了每个现金收集或单元的支持费用的分布。

继续下去，我们可以相应地改变因素的数量，降低或增加可见因素的数量，我们将看到图表中的变化。这个过程是迭代的，并且可以重复，直到尚未找到所需的模式。

3.3 过滤

价值识别问题一直是视觉分析的热门话题，在大数据的情况下它变得更加重要[6]。即使我们只显示 60 个惟一值，更不用说数百万个惟一值，在一个图表上，也很难为每个值放

置标签。

而且，在一个数据集中可能存在完全不同的值范围。因此，一些值只是由具有较高振幅水平的其他值所支配。因此，整个图表的感知会很复杂。例如，一些每天 24 小时工作的组织可以具有不同的客户流，并且每小时显示出客户的依赖性，当值接近相等并且具有比白天小时低得多的振幅时，我们将失去对一组夜间小时的感知能力。

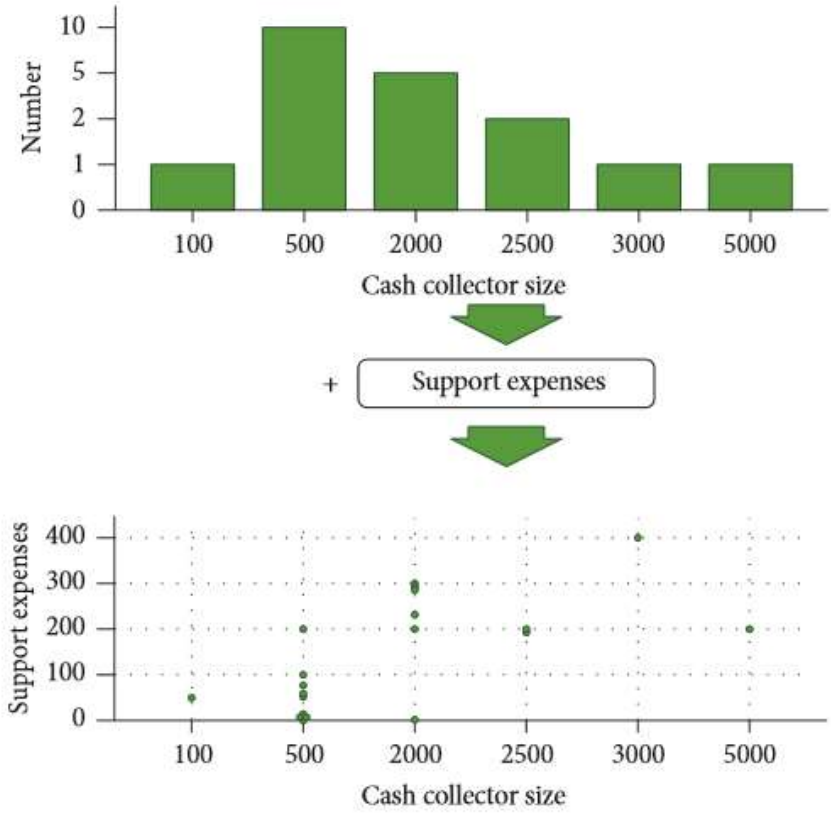

图 3-3　因素数量的动态变化

分析人员通常希望看到整个数据表示和部分更详细的数据表示都位于他感兴趣的区域，而且，他感兴趣的区域不是静态的，可以在研究过程中动态变化。

4　大数据可视化方法

本段包含大数据可视化方法描述。每个描述都包含用于方法分类到大数据类之一的参数。我们假设以下数据标准：

（i）数据量大；

（ii）数据多样化；

（iii）数据动态。

4.1 树图

该方法基于层次数据的空间填充可视化。根据定义，对于必须分层链接的数据数据对象有严格的要求。树图由根矩形表示，根矩形被分成若干组，也由较小的矩形表示，这些矩形对应于来自集合[10]的数据对象。

这种方法的例子是硬盘可视化上的自由空间、来自不同组织及其附属机构的盈利能力。

方法可以应用于大数据量，迭代地表示每个层次结构的数据层。在设备分辨率出现异常的情况下，分析人员总是可以前进到下一个块，继续对较低层级的更详细数据进行研究。因此，满足大数据量的准则。

因为该方法基于从一个或多个数据因子计算的形状体积估计，所以每次数据变化之后都会针对当前可见的层次结构重新绘制整个图像。较高级别的更改不需要重新绘制图像，因为其中包含的数据对分析员是不可见的。

该方法获得的可视化结果只能显示两个数据因子。第一个是用于形状体积计算的因素。第二种是颜色，用于对形状进行分组。此外，用于体积估计的因素必须由可计算数据类型表示，因此不能满足标准的数据变化。

最后一个标准也不能满足，因为树图只在一个时刻显示数据表示。

方法优点：

（i）分层分组清楚地显示数据关系;

（ii）使用特殊颜色可以立即看到极端异常值。

方法缺点：

（i）数据必须是层次化的，而且树图更适合于分析数据集，数据集至少有一个重要的数量维度，并且变化很大;

（ii）不适合考察历史趋势和时代模式;

（iii）用于尺寸计算的因子不能具有负值[11]。

4.2 圆形包装

这种方法是树图的直接替代方法，除了它使用圆作为原始形状之外，还可以从更高层次上将其包括在圆中。这种方法的主要优点是，通过使用经典树图[12]，我们可以放置和感知更多的对象。

由于圆填充法是基于树图方法的，所以具有相同的性质。因此，我们可以假设该方法仅满足大数据量准则。

尽管如此，在方法的优点和缺点方面仍然存在如下差异：

方法优势：与树图相比，空间高效的可视化方法。

方法缺点：与树图方法相同的缺点。

4.3. Sunburst

这种方法也是树图的替代方法，但它使用树图可视化，转换为极坐标系。这些方法的主要区别在于可变参数不是宽度和高度，而是半径和弧长。这种差异使得我们不能在数据改变时重新绘制整个图表，而是通过改变其半径，只允许一个包含新数据的扇区。而且由于这个特性，这个方法可以用动画显示数据动态。

动画可以增加数据的动态性，只用太阳暴光半径进行操作，因此可以说，满足了数据动

态性准则。

与前两种方法相比，Sunburst 具有相同的优点和缺点：

方法优势：大多数人容易理解[13]。

方法缺点：与树图方法相同的缺点。

4.4 环形网络图

数据对象被放置在一个圆的周围，并根据其相关性的速率通过曲线连接。不同的线宽或颜色饱和度通常用作物体相关性的测量。

此外，方法通常提供交互，使得不必要的链接不可见，并突出显示所选的链接。因此，这个方法强调了多个对象之间的直接关系，并显示了它的相对性[14]。

对于该方法的典型用例，有以下示例：城市之间的产品转移图、不同商店中购买的产品之间的关系等。

该方法允许我们将聚集的数据表示为分析数据对象之间的一组弧，以便分析人员能够获得关于对象之间关系的数量信息。该方法适用于大数据量、按圆半径放置数据对象、改变对象方格面积等情况。还有，在圆弧附近可以显示附加信息，这些信息可以从数据对象的其他因素中提供，并且有必要补充的是，对于每个图只使用一个因素没有限制，我们总是可以放置对象的不同因素并在它们之间建立关系。虽然很难理解和理解，但是在某些情况下，这种方法会产生足够的信息来改变分析者的研究方向，或者做出最终的决定。

圆形形状鼓励眼睛沿着曲线运动，而不是以方形或矩形的曲折方式运动[15]。

而且，作为整个数据表示的结果，数据的每次更改都必须跟随图的重新绘制。

5 结果

因此，我们提供了表 1，显示了哪种方法可以处理各种数据、大容量数据以及处理时间数据的变化（表 5-1）。

TABLE 1: Properties of visualization methods.

	Large data volume	Data variety	Data dynamics
Treemap	+	−	−
Circle packing	+	−	−
Sunburst	+	−	+
Circular network diagram	+	+	−
Parallel coordinates	+	+	+
Streamgraph	+	−	+

通过分析表 1，可以说，基于树图方法的方法不能应用于 Big Data 类之一，因为它仅满足一个准则，而需要满足至少两个准则。

TABLE 2: Visualization methods classification.

Method name	Big data class
Treemap	Can be applied only to hierarchical data
Circle packing	Can be applied only to hierarchical data
Sunburst	Volume + Velocity
Circular network diagram	Volume + Variety
Parallel coordinates	Volume + Velocity + Variety
Streamgraph	Volume + Velocity

根据表 5-2，现在我们可以按照大数据类（表 5-2）清楚地对可视化方法进行分类。

6 结论

本文描述了大数据可视化的主要问题及如何避免这些问题的方法，并且基于对三个大数据类之一的适用性给出了大数据可视化方法的分类。

该领域的未来工作可以集中在以下几个方面：研究不同尺度的可视化方法的适用性，为具体大数据类的可视化方法选择做出决策和建议，以及形式化应用于一个或多个大数据类的可视化方法的要求和限制。

原文出处：Gorodov, Evgeniy Yur'evich, Gubarev, Vasiliy Vasil'evich. Analytical Review of Data Visualization Methods in Application to Big Data[J]. Journal of Electrical and Computer Engineering, 2013, 2013:1-7.