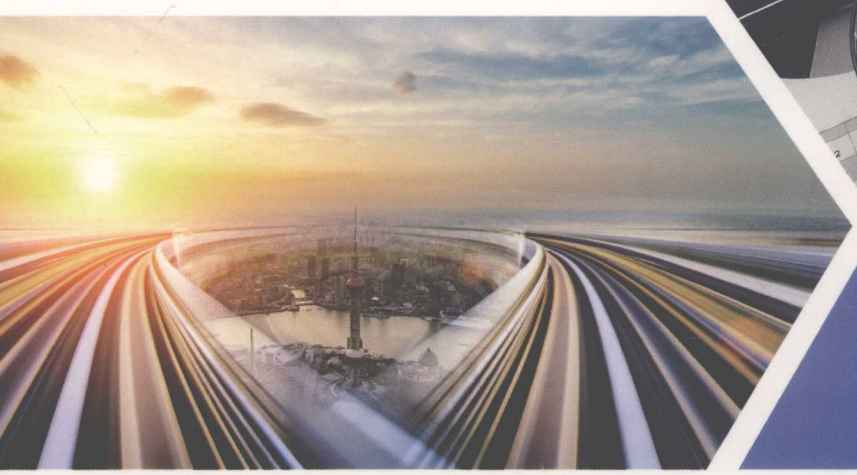
可视化的美之基于R语言的

大数据可视化分析与应用

KESHIHUA DE MEI ZHI JIYU R
YUYAN DE DASHUJU KESHIHUA
FENXI YU YINGYONG



陈凌云 著

电子科技大学出版社
University of Electronic Science and Technology of China Press

可视化的美之基于R语言的

大数据可视化分析与应用

陈凌云 著

电子科技大学出版社
University of Electronic Science and Technology of China Press

图书在版编目(CIP)数据

可视化的美之基于R语言的大数据可视化分析与应用 / 陈凌云著. -- 成都: 电子科技大学出版社, 2018.5
ISBN 978-7-5647-6097-7

Ⅰ. ①可… Ⅱ. ①陈… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第079325号

可视化的美之基于R语言的大数据可视化分析与应用

陈凌云 著

策划编辑 李述娜

责任编辑 熊晶晶

出版发行 电子科技大学出版社

成都市一环路东一段159号电子信息产业大厦九楼 邮编 610051

主 页 www.uestcp.com.cn

服务电话 028-83203399

邮购电话 028-83201495

印 刷 定州启航印刷有限公司

成品尺寸 185mm × 260mm

印 张 14.25

字 数 330千字

版 次 2019年1月第一版

印 次 2019年1月第一次印刷

书 号 ISBN 978-7-5647-6097-7

定 价 52.00元

版权所有，侵权必究

前　言

这是一个大数据爆发的时代。面对信息的激流，多元化数据的涌现，大数据已经为个人生活、企业经营，甚至国家与社会的发展都带来了机遇和挑战，成为 IT 信息产业中极具潜力的蓝海。大数据可视化这种新的视觉表达形式是应信息社会蓬勃发展而出现的——因为我们不仅要呈现世界，更重要的是要通过呈现来处理更庞大的数据，理解各种各样的数据集合，表现多维数据之间的关联。换句话说，就是归纳数据内在的模式、关联和结构。复杂数据可视化既涉及科学也涉及设计，它的艺术性实际上是使用独特手法展示万千世界的某个局部，从而提出问题。大数据可视化，位于科学、设计和艺术三学科的交叉领域（准确地说，应该是位于三个不同维度的人类活动的交叉领域），蕴藏着无限可能性。

本书针对计算机、信息管理和其他相关专业学生的发展需求，系统、全面地介绍了关于大数据技术及其可视化的基本知识和技能，详细介绍了数据可视化之美，数据可视化方法、应用、管理与计算、分享与发布等内容。希望本次研究能够对我国大数据行业的发展起到正面的影响，也希望能够帮助到对大数据技术感兴趣的读者。

另因撰文历时漫长，书中或有些许不足，望读者朋友们不吝赐教，大家共同进步。

陈凌云

2018 年 3 月

目 录

试读结束：需要全本请在线购买： www.ertongbook.com

第一章 绪 论

第一节 大数据可视化研究背景综述

在过去的20年中，各个领域都出现了大规模的数据增长，包括医疗保健和科学传感器、用户生成数据、互联网和金融公司、供应链系统等。数据的历史悠久，它以一种记录符号为人们所熟知。如今，它已经不再被局限于统计图里的数字、符号或表格，也仅仅是储存在电脑中等待人选取分析的资源，随着近年来发展迅猛的智能手机、可穿戴智能设备的发展和兴起，人们日常生活中的一个表情、一句问候，举止行为、所在方位甚至生理变化等的每一个细微的改变都能被转换为数据进行记录和分析。在过去的十年中，由于在全球范围内使用互联网的快速发展和所连接设备的数量飞速增长的刺激，整个世界一直处于超级信息爆炸之中。我们正在经历的“增长的数据”增长的速度比人类整个历史上任何时候都快。企业应用程序数据以及机器生成的数据继续成倍增长，行业专家和研究人员正面临巨大的挑战——开发新的创新技术来对硬件和软件技术和产品做出评估和基准测试。有研究估计，在2008—2020年，企业数据总量将从约0.5字节（Zettabyte，简称ZB，$1ZB=10^9TB=10^{12}GB$）增长到35字节。据官方相关资料显示，互联网络上的数据每年将上升50%，每两年则翻一倍；据悉，如今全球高于90%的数据都来自十年内人们的各种生活、工作等活动。互联网发展到今天，发微博、写博客、网上购物、浏览网页这些都是日常生活常做的事。在与互联网、信息系统等的交互中，我们创造了海量数据，加上无处不在的传感器和微处理器收集、处理的数据，互联网上的数据量越来越庞大，传统的数据库和数据架构无法及时对这些数据集进行处理、管理和分析。与大数据相关的科学、技术和应用迅速发展成为信息科学领域的热点，引起相关部门、学术领域专家的极大兴趣。也正因如此，“大数据时代”已然到来，“大数据”也成了最热门的词汇之一。

数据时代正处于热火朝天的形势，而人类面对的不仅仅是“一切皆有可能”，更是前所未有的大量挑战。除了政府机构、媒体（传统媒体）、企业等提供的越来越多的数据以外，新兴的社会化媒体（如SNS社区）、物联网技术（如智能手机）等的应用普遍化、介入人类生活的深入化也是数据量急剧增长的主要原因。通过处理这

些海量数据资源，研究人员能更好地、有所依据地对相关领域的发展、趋势做出总结和预测；人们也能对自身情况有更好的了解与评估，以此制定下一阶段的改进计划，适时调整自己的目标。在这些美好蓝图的背后，数据分析、数据挖掘技术得到了充分的关注和研究，计算机技术的飞快发展使数据的收集和存储等处理的过程变得更高效，而互联网犹如开辟了一个新天地——时间和空间都不再限制人类的思维和行为。从前的数据统计和数据分析等工作由统计学方面的专家、数据分析师和科学研究者们全权负责，但在如今的大数据背景下，海量数据只有在被合理采集、解读与表达之后才能完美展现它们的瑰丽与深奥，而可视化则无疑是让数据变得亲切和便于理解的最有效的途径。所以，各种跨专业、跨学科的知识都被迫切地需要，大量的非计算机领域的人才都逐渐加入了数据可视化的行列，他们发挥各自领域的专业优势，通力合作，将数据可视化推向了当代的热潮。大数据主要来自互联网渗透人们日常生产、生活等方面使用网络留下的印迹，如浏览网页、网上购物等。大数据技术旨在从庞大的数据中提取出有价值的数据信息。随着发展，大数据可能会在未来成为最大的商品，数据的大量使用将会使大数据变成一个大产业。大数据产业实现盈利的关键，在于提高大数据的信息含量和价值。

毫无疑问，数据可视化顺应大数据时代的到来而兴起，而只有在理解了数据可视化概念的本质之后，才能通过对其原理和方法进行研究和合理运用，获取数据背后隐含的价值。

第二节　大数据可视化发展沿革

一、数据与大数据

数据，英文名是 Data，是用来描述科学现象和客观世界的符号记录，是构成信息和知识的基本单元。数据是没有进行加工处理的事实，也就是说单个数据之间互不相干，并无瓜葛，独立存在，人们用约定俗成的方式将其排序或表达就使之有了意义，以此来供特定领域内的专业人员进行交流沟通——包括描述、解读和保存。一般来说，常见的是从数据表现的角度来进行数据分类——分成数字数据和模拟数据。数字数据是指各种统计或量测数据等是不连续的、离散的；模拟数据则是由连续函数组成的、在一定范围内连续变化的物理量，一般分类包含几何图形或空间图形数据（如点、线、面），以及符号、文本和图像数据等，如声音的分贝数和热度的高低等。

大数据比较公认的概念是含有 4 个 V 的特点的数据——数据量大（Volume）、变化速度快（Velocity）、数据类型多样化（Variety）与价值密度低（Value）。从计

算机技术的专业角度来说，大数据是结构复杂、数量庞大、类型众多的数据的集合，包括非结构化数据、半结构化数据和结构化数据；从更容易被人理解的角度上来说，大数据就是海量资料，在效率至上的时代，其规模巨大，所以无法由人工在较短的时间内采集、管理、处理、分析并整理成为普通人所理解的内容，因而必须借助计算机技术进行一系列处理，精准结果的同时节省时间，最终获得数据背后的信息与价值。此外，一些学者认为大数据该被定义为大数据技术而非单纯的多类型的数据。但本文认为，大数据就是指信息时代的数据，若非如此，大数据的类型和数量等属性同理都应该归于大数据的概念，成为一个复杂的混杂体，那就更混乱了。所以，大数据技术（包括大数据处理技术、大数据管理技术等）应该与大数据这个名词分开来看，它们都是大数据时代的产物。

了解数据的定义有助于我们挖掘数据与大数据的相关性，从而更好地认识大数据的概念。如果单从数据的字面意思上延伸，我们可以将大数据理解为数以千计的数据的集合，人们多用它来描述信息爆炸时代产生的海量数据和与之相关的技术创新。当然大数据的概念并非如此简单，让我们先来看看目前学术界出现的有关大数据的定义："大数据（Big Data），又称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯。"一般意义上，大数据是指无法在可容忍的时间内用传统 IT 技术和软硬件工具对其进行感知、获取、管理处理和服务的数据集合。大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

二、大数据时代的新特点

（一）数据化

舍恩伯格在《大数据时代》一书中说：明天，我们的下一代，一群被"大数据观念"陶冶长大的家伙，会发自肺腑地认为"量化一切"并从中学习社会是至关重要的。把各种各样的现实转化为数据，对今天的我们而言也许是新奇而有趣的，但在不久的将来，这将变成如同吃饭、睡觉一样与生俱来的能力——这又让我想起了"数据"这个词语的拉丁语原意。我们可以想象未来将会是一个量化一切的世界，也就意味大数据时代我们的生活将会趋于智能化、数据化，用数据分析并对事物进行量化处理将会给人类未来的社会生活带来极大的便利。也许还会出现一种自我量化，即对自己身体的每一个部位，每一次神经反应都能够进行精准的测量，也就是说大数据让量化一切事物成为一种可能。

（二）数据更加个人化

大数据时代的数据被赋予了强烈的个人色彩，笔者之所以这么说是因为大数据

的数据有很大一部分来源于网民每日上网浏览所留下的信息，例如日志、照片、地理位置、交易记录等。如今的信息追踪和数据处理技术可以轻松地记录保留这些数据，通过计算处理便可以获得用户的基本数据资料，在无数个数据信息的相互关联下就可以绘制出针对每个个体的数据资料库。所以说这些看似没有生命的数据实际上拥有着鲜明的个人主义色彩。而个人的数据信息是具有极高的利用价值的，它可以应用到医疗、交通、传媒、银行等各个社会服务行业中，为人们的日常生活带来便利，也能更好地帮助企业进行自动化运营管理。

（三）决策理性精准化

数据的个性化为社会各行各业提供了精准的数据信息，从而促进行业决策的精准性。首先，在大数据时代我们不需要依赖随机采样，便可以分析相关现象的所有数据。其次，一切事物的数据化与量化让我们可以直接在宏观层面把握事物的精确度，数据分析处理技术会替我们自动追寻因果关系。这样既减小了数据统计分析的烦琐程度，也为企业运营提供了决策制定的理性与便捷。大数据作为一种堪比自然资源、社会资源的新生资源，其带给人类的无限的价值正在被人们慢慢正视，技术的不断创新与发展能够对数据进行全面感知、收集、分析、共享，而这些活动将为人们提供一种看待世界的全新方法。在不久的将来，人们会基于数据和现实分析来做决策，所有的事物都可以通过量化达到精准的状态，而这样的未来都是大数据带给人类社会的颠覆性变革。互联网广告是数字化时代的重要信息产业，它的变革也必将发生，也更能体现大数据时代变革的先进性。

“信息时代，万物数化”。人们对数据概念的传统印象似乎已经脱离了当代信息发展的大环境而不再适用，但大数据的本质依旧不变，是数据，是构成信息和知识的基本单元，只是它的获取方式、涉及范围和展现形式被扩展、放大和多样化了。

三、大数据关键技术

随着互联网、云计算和物联网的迅猛发展，无所不在的移动设备、RFID（射频识别）、无线传感器每分每秒都在产生数据，数以亿计用户的互联网服务时时刻刻在产生巨量的交互。要处理的数据量越来越大，而且还将更加快速地增长，同时业务需求和竞争压力对数据处理的实时性、有效性也提出了更高的要求，传统的常规数据处理技术已无法应付，大数据带来了很多现实的难题。为了解决这些难题需要突破传统技术，根据大数据的特点进行新的技术变革。大数据技术是一系列收集、存储、管理、处理、分析、共享和可视化技术的集合。适用于大数据的关键技术包括以下几方面。

（一）遗传算法

借鉴生物界的进化规律（适者生存，优胜劣汰遗传机制）演化而来的随机化搜索方法。采用概率化的寻优方法，自动获取和指导优化的搜索空间，不需要确定的

规则，自适应地调整搜索方向，已被人们广泛应用于组合优化、机器学习、信号处理、自适应控制和人工生命等领域，是现代有关智能计算中的关键技术。应用实例包括制造业改善作业调度，以及优化投资回报率，等等。

（二）神经网络

受生物神经网络结构和运作的启发，模拟动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。应用实例包括识别高价值客户离开特定公司的风险，以及识别欺诈性的保险理赔行为，等等。

（三）数据挖掘

结合统计数据和机器学习、使用数据库管理技术从大型数据集中提取有用信息和知识的技术。根据其他属性预测特定（目标）属性的值，如回归、分类、异常检测等，或寻找概括数据中潜在的联系模式，如关联分析、演化分析、聚类分析、序列模式挖掘等。

（四）回归分析

确定当一个或多个独立变量值被修改时相关变量如何变化的统计方法，通常用于预测或预报。应用实例，如基于不同的市场和经济变量，或通过确定何种制造业参数对客户满意度影响最大来预测销售量等用于数据挖掘。

（五）分类分析

在训练集包含的数据点已经被归类的基础上，确定新的数据点所属类别的方法。典型应用是在明确假设或客观结果前提下，预测部分特定客户行为（例如购买决策、流失率、消费率等）。因为使用训练集，属于监督学习，是无监督学习类型聚类分析的反面。用于数据挖掘。

（六）聚类分析

一种多元化群体的分类统计方法。在事先不知道的前提下，将一个集合分成较小的对象组，组内对象具有相似特点。聚类分析的典型例子是将消费者分割成具有相似性的群体做针对性营销。因为不使用训练数据，属于无监督学习类型，是监督学习类型分类分析的反面。用于数据挖掘。

（七）关联规则学习

在大数据集变量中发现感兴趣关系（即“关联规则”）的方法，包括多种生成和测试可能规则的算法。典型应用是市场购物篮分析，其中零售商可以决定哪些产品经常一起购买和如何使用这种营销信息。用于数据挖掘。

（八）数据融合与集成

集成和分析来自多个源的数据的方法。典型应用，如使用来自互联网的传感器数据综合分析炼油厂这样的复杂分布式系统的性能。使用社会媒体数据，经过自然语言处理分析，并结合实时销售数据，确定营销活动如何影响顾客的情绪和购买行为等。

试读结束：需要全本请在线购买： www.ertongbook.com

（九）机器学习

研究计算机怎样模拟或实现人类的学习行为，获取新的知识或技能，重新组织已有的知识结构并不断改善自身的性能，是人工智能的核心，是使计算机具有智能的根本途径。自然语言处理是机器学习的一个例子。

（十）自然语言处理

研究实现人与计算机之间用自然语言进行有效通信的理论和方法。典型应用是使用社交媒体的情感分析来确定潜在客户对品牌活动的反应等。

（十一）情感分析

从源文字材料中确定和提取主观信息的自然语言处理和分析方法的应用。分析的主要内容包括识别表达情感的特征、态势或作品。应用实例是分析社会化媒体（如博客、微博客或社交网络）确定不同客户群和利益相关者对其产品和行为的反应。

（十二）网络分析

在图或网络中描述离散节点之间特征关系的方法。在社会网络分析中，分析个人在社会或组织之间的联系，应用实例包括确定营销目标的关键意见负责人、确定企业信息流的瓶颈等。

（十三）空间分析

分析数据集拓扑、几何或地理编码性能技术的统计方法。数据通常来源于采集地址或纬度 / 经度坐标等位置的地理信息系统。应用实例包括空间数据的空间回归（例如，消费者是否愿意购买与位置相关的产品）或模拟（例如，如何将制造业的供应链网络分布到不同的地点）。

（十四）时间序列分析

分析数据点序列表示连续时间值，从数据中提取有意义特征的统计学和信号处理方法。一般通过曲线拟合和参数估计来建立数学模型。应用实例包括销售数字预测、气象预报、水文预报以及将诊断为传染性疾病人数的预测等。

（十五）可视化技术

可视化是支持大数据蓬勃发展的重要领域：可视化技术通过创建图片、图表或动画等，方便对大数据分析结果的沟通与理解。标签云即加权视觉列表，将其中出现频繁的词以更大的文本呈现，不经常出现的词用较小的文本呈现，帮助读者迅速感知大文本中最突出的概念；Clustergram 是一种聚类分析可视化技术，用于显示随着集群数量的增加，数据集的个别成员如何被分配到集群。使分析师能够更好地了解为何不同的集群数量产生不同的聚类结果；历史流用图形化的方法表示多个作者编辑文件的历史，在图中很容易发现不同的见解。空间信息流在视图中通过不同亮度、颜色等显示统计分析参数。如利用视图显示纽约和世界各地城市之间 IP 数据流的大小，在图中特定城市所在位置以不同亮度反映该城市和纽约之间的不同 IP 流量，

可以快速确定哪些城市与纽约的通信量大。

四、数据可视化的历史发展

可视化技术把数据变为图形展示给大众，注重技术的实现及其算法的优化，通过开发可视化工具变抽象为具象，便于理解的同时加深印象。它涉及计算机图形学、计算机仿真领域等，广为人知的实例包括虚拟现实技术、可视化仿真系统等。可视化技术是可视化表现的基础。可视化表现是指将晦涩难懂的数据进行“更友好的”图形、图像的表现，严格来说，并非局限于视觉，不仅仅是结合图表、文字、表格、录像等形式，亦可结合听觉、视觉、触觉等感觉，并加入交互处理的理论、方法和技术，让用户在互动中与数据交流，达到“易于理解”的目的。此处的可视化表现注重视觉表达、交互方式和人们的心理感知，通过对心理学、图形设计等知识的合理运用来展现数据并有效传达其隐含意义。普遍意义上的数据可视化被认为伴随统计学的诞生而出现。其实，用图形、图像描绘、记录量化信息的思想，从人们开始观察这个世界进而产生测量、管理的需要的时候就已经出现了。

（一）可视化思想的起源（15—17 世纪）

15—17 世纪是欧洲中世纪的晚期，这段时间可以被看作是可视化的起始阶段。经济、技术的发展，文艺复兴的到来使人们开始了解人文和科学知识，对地球的新认识则使许多著名的航海家浮出水面，新的国家与地区开始被载入人类史册。天文学、测量学、绘图学等都快速起步以跟上对未知新世界的探索。三角测量技术、数学函数表相继出现了，人类也开始了对概率论和人口统计学的研究。这个时期是数据可视化的早期探索阶段。

（二）数据可视化的孕育时期（18 世纪）

在此期间，在数学和物理知识成了科学研究的基础，技术已经成为主力，社会管理的精确定量逐渐形成。伴随着早期统计学的萌芽，社会和科技的进步体现在数据表现的多样化，已经出现了很多现在被广泛使用的图形形式，直方图、柱状图、饼图、圆环图等也已经出现。

（三）数据图形的出现（19 世纪前半叶）

在 18 世纪至 19 世纪前半叶这几十年间，统计图、地图和主题图等这些如今依旧火热的数据可视化表达手法很多都开始被使用了，其中一个重要原因是很多公共领域的数据开始被政府部门重视，因而数据在这一时期极大地丰富起来，例如关于人口、教育、犯罪、疾病等数据都被系统性地收集和发布，已经从科学技术和经济领域扩展到社会管理领域。另外，正在萌芽的计算机、通信等提供了技术实现的可能性；书籍、报纸等媒体的出现和大量应用使印刷形式替代了手绘。重要的是，数据图形在这一时期在视觉表现上有了极大的进展，表达方式多样化了。现如今所常

见的统计图形的样式及其他表现图等都出现了。柱状图、饼图、地图等集成为这个年代展示数据信息的一种常用方式。

（四）第一个黄金时期（19 世纪中、末期）

前面的所有发展似乎都是为这个黄金时期所做的铺垫，数据可视化迎来了它历史上的第一段辉煌。欧洲逐渐意识到信息数据的作用，官方的统计机构也普遍建立起来了，数理统计成了一门新的学科，统计学的国际会议对可视化图形制定了分类和标准，各种图形、统计图表等都被广泛地应用和熟知起来了。

（五）低潮期（20 世纪前期）

20 世纪前期，数理统计成了数学的一个支派，统计学家们这个时期关注的主要是在准确的数学基础上扩展统计的疆域。数据的量和种类并没有太大的变化，于是黄金时期所出现的数据表示方式就已经够用，所以具有美观性和启发性的图形表达的研究就受到了冷落。

（六）新的黄金时期（20 世纪中末期至今）

现代电子计算机的诞生带来了强大的冲击，对数据可视化研究的再次兴起有了推波助澜的作用。计算机对数据分析的影响来自两方面——高分辨率的图形展现和交互式的图形分析都是手绘图形无法带来的革命性改变。同时，随着统计应用的发展，数据分析的应用扩展到了各行各业。当二者互相结合之后，就催生了统计计算工具、图形软件工具以及输入输出、显示技术等。

通过表 1-1，可以对数据可视化的发展状况有一个清晰的了解。

如今，海量数据使人类进入了大数据的背景，第二个黄金时期是数据的可视化，与第一次不同的是除了数据量的大幅度增加，另一个值得一提的方面是它不再只是专业学者们的工具，而是每一个普通人都能读懂它、感受它、应用它，每个人都能参与到数据可视化推动社会发展的进程中来。

表1-1 数据可视化发展历程

时期	数据可视化发展历程
15—17 世纪	可视化思想诞生初始，数据可视化的早期探索正式拉开序幕
18 世纪	数据可视化初步发展，直方图、柱状图、饼图、圆环图等开始出现
19 世纪前半叶	数据开始得到重视，数据图形出现
19 世纪后半叶	数据可视化第一个黄金时期，图形、图表等被广泛应用
20 世纪前期	前期的可视化表达方式已经够用，图形表达的研究并无新的进展
20 世纪中后期至今	数据可视化依附计算机科学与技术拥有了新的生命力，并将在不久的将来大放异彩

第三节 大数据可视化目标与作用

根据信息传递方式，传统的可视化方法可以大致分为两大类，即探索性可视化和解释性可视化。前者指在数据分析阶段，不清楚数据中包含的信息，希望通过可视化快速发现特征、趋势与异常，这是一个将数据中的信息传递给可视化设计与分析人员的过程。后者指在视觉呈现阶段，依据已知的信息或知识，以可视的方式将他们传递给公众。从应用的角度来看，可视化有多个目标：有效呈现重要特征、揭示客观规律、辅助理解事物概念和过程、对模拟和测量进行质量监控、提高科研开发效率、促进沟通交流和合作等。从宏观角度看，大数据可视化可分为以下四类。

一、文本可视化

作为大数据时期文本可视化数据的一个典型文本信息，实际上也是最主要的互联网数据信息，与此同时，也是物联网通过一定的传感器收集到的信息类型。在正常的工作和学习以及日常生活中，人们使用最多的就是文本形式的电子文档。文本可视化可以在一定程度上直观的体现文本主要优势和特点，例如，逻辑结构、动态演化规律以及主体聚类等。最基本和典型的文本可视化就是标签云，依据词频来合理地把关键词进行排序和归类，然后利用一定的颜色、大小等属性来进行文本可视化。现阶段，最主要的就是利用字体大小展现的关键词使用在互联网中主题热度的识别。随着关键词数量的不断增加，如果不能合理地进行设计阀值，就会出现重复覆盖以及局部密集的问题，这样就需要提供一定的交换窗口来操作。

二、网络可视化

在大数据分析中最常见的关系就是网络关联，例如，社交网络和互联网。实际上层次结构在一定程度上属于一种比较特殊的网络信息。依据连接拓扑和网络节点之间的关系，可以非常直观地体现出网络中隐藏的关系。例如节点，实际上是进行网络可视化的重要内容之一。怎样在大规模边和节点的网络中利用有限空间进行一定的可视化，是现阶段大数据研究的重点和难点。除了能够可视化静态拓扑关系，还具有相应的动态流动演化性，所以对动态网络进行一定的可视化也是不容忽视的内容。随着网络中边和节点数目的增多，很容易出现覆盖、重叠以及聚集等问题，不能很好地进行可视化，影响效果。因此，处理大规模可视化的主要方式就是图简化。可以分成两类：一类是利用多尺度和层次聚类进行交互，把大规模数据变化为具有一定层次的树结构，然后利用多尺度进行不同的可视化；另一类是对边进行适

当的聚集，保证具有清晰的可视化效果。这些都是简化的主要方式，也可以看出引入交互技术，是可视化技术未来发展过程中必不可少的方式。

（三）时空数据可视化

时空数据主要是指具有一定时间标签和地理位置的数据。移动终端与传感器发展非常迅速，因此，使得时空数据逐渐成为大数据发展过程中典型的数据类型。充分结合地理制图学以及数据可视化技术，分析和研究空间和时间对于可视化表征之间的关系，能够很好地展示空间和时间以及规律模式。大数据时代发展模式下，时空数据具有实时性和高维性，同时这也是数据可视化的重点。为了能够更好地体现信息随着空间和时间位置发生一定的变化，一般可以利用信息对象来逐渐实现数据可视化。流式地图是最典型的可视化方式，充分融合地图和时间事件流。为了打破二维数据的局限性，出现了时空立体方，是利用三维模式来展现空间、时间、事件。

（四）多维数据可视化

多维数据可视化实际上就是说拥有很多个维度的数据变量，在数据仓库以及数据库中具有广泛的应用，如商业智能系统、企业信息系统。进行多维数据的主要目的就是不断发现多维数据的模式和规律，合理展示不同纬度之间存在的关系。多维数据可视化具有多种方式，主要包括基于图标、基于图结构、几何图形、基于层次结构、基于像素、混合方式。近年来，随着大数据的不断发展，几何图形是研究多维数据可视化的重点。最常用的多维数据可视化的方式就是散点图，二维散点图可以适当利用多维度中的两个维度综合体现映射到两条轴上，利用不同的图形在二维平面内合理反映维度信息。例如，可以利用不同颜色、形状等来表示一定的离线或者连续性。投影是从多维度方面来体现可视化的一种方式，能够很好地体现出维度属性值的分布情况，还可以体现多维度之间的关系。

作为大数据分析的重要方式，可视化分析可以有效地弥补计算机自动化分析过程中出现的不足和缺陷。大数据可视化分析可以很好地融合计算机的分析能力和人们对信息的感知能力，在依据数据挖掘前提下进行的数据分析。

第四节　大数据可视化中R语言数据挖掘的应用

数据挖掘的应用极为广泛。易观智库以应用成熟度和市场吸引力作为两个维度，给出了当前数据挖掘的十大典型应用及其分布。

数据挖掘在电子商务领域的应用是最成熟和最具吸引力的，金融和电信行业紧随其后。政府公共服务领域的数据挖掘将有较大的发展潜力，其未来的应用成熟度将会有巨大的提升空间。

进一步，数据挖掘在电子商务中的应用价值主要体现在市场营销和个性化导购等方面。有效实现用户消费行为规律的分析，制订有针对性的商品推荐方案，根据用户特征研究广告投放策略并进行广告效果的跟踪和优化；金融行业中，数据挖掘主要应用于客户金融行为分析以及金融信用风险评估等方面；数据挖掘在电信企业的应用主要集中在客户消费感受等分析方面。目的是通过洞察客户需求，有针对性地提升网络服务的质量和安全；在政府公共服务中，数据挖掘的作用主要体现在智慧交通和智慧安防等方面，旨在实现以数据为驱动的政府公共服务；医疗行业的数据挖掘应用价值集中在药品研发、公共卫生管理、居民健康管理以及健康危险因素分析等方面。

尽管上述典型数据挖掘应用所解决的问题不同，但研究思路类似，且问题的切入也有很多共同点。若对上述各个应用问题分别展开论述，内容难免冗余、雷同。因此，这里仅对金融、电子商务、电信中的典型商业数据挖掘共性问题进行梳理并做详尽讨论，主要包括客户细分研究、客户流失预测、交叉销售、营销响应、欺诈甄别等方面。

一、数据挖掘在客户细分中的应用

客户细分的概念是美国著名营销学家温德尔·史密斯于20世纪50年代中期提出的。客户细分是经营者在明确其发展战略、业务模式和市场条件下，依据客户价值、需求和偏好等诸多因素，将现有客户划分为不同的客户群，属于同一客户群的消费者具有较强的相似性，不同细分客户群间存在明显的差异性。

在经营者缺乏足够资源应对客户整体时，由于客户间价值和需求存在异质性，有效的客户细分能够辅助经营者准确认识不同客户群体的价值及需求，从而制定针对不同客户群的差异化的经营策略，以资源效益最大化、客户收益最大化为目标，合理分配资源，实现持续发展新客户、保持老客户、不断提升客户忠诚的总体目标。

客户细分的核心是选择恰当的细分变量、细分方法以及细分结果的评价和应用等方面。

（一）客户细分变量

客户细分的核心是选择恰当的细分变量。不同的细分变量可能得到完全不同的客户细分结果。传统的客户细分是基于诸如年龄、性别、婚姻状况、收入、职业、地理位置等的客户基本属性。此外，还有基于各种主题的，如基于客户价值贡献度、需求偏好、消费行为的客户细分等。

不同行业因其业务内容不同，客户价值、需求偏好以及消费行为的具体定义也不同。需选择迎合其分析目标的细分变量。例如，电信行业4G客户细分，主要细分变量可以包括使用的手机机龄、自动漫游业务、月平均使用天数、月平均消费额、

试读结束：需要全本请在线购买： www.ertongbook.com