2017 届研究生专业技术报告

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：51151500102

華東師範大學

East China Normal University

专业技术报告

报告题目：信息可视化与可视分析

院　　系：计算机科学与软件工程学院

专　　业：　软件工程

研究方向：　信息可视化

指导教师：　王长波　教授

学位申请人：　李柯林

2016 年 9 月 18

目 录

第1章 可视化与可视分析

1.1 信息可视化概述

大数据在我们的生活中无处不在，随着数据总量的不断增加，数据分析在大数据研究中地位越来越重要。数据可视化即通过借助图形图像学知识，结合人机交互、美学知识，将复杂的多维数据用图形化的形式表示。它可以将多维数据以不同的视觉角度来进行展示，通过人机交互的方法对数据进行多视角分析观察，将数据隐藏的信息有效地传达。对大数据进行可视分析，需要以下工作：

（1）数据的采集。对于已有的历史数据、网络数据或者传感器数据等，不同的数据需要通过不同的方式进行采集来获得，这些惊人的数据中有大量无用或者冗余的信息，使用正确过滤器和处理方法过滤无用字段信息，同时保留有效的数据字段。

（2）数据分析。对所得多维数据进行研究，通过对数据的分析发现数据中的新的特征或者属性。其中又包含了两类分析手段：数据的探索性分析和定性分析。探索性分析是基于统计学之上的一种对数据提出的假设验证分析，从而挖掘出新的数据特征；定性分析则是对特定的如词语、句子和照片等非数值型数据进行的分析。

（3）数据处理。对将要可视化的数据进行冗余信息的清理，同时在进行可视化过程中，要保持数据的安全性和信息的可靠性，将正确的数据通过合适的可视化方式进行展示。

（4）数据管理。对于大量的可视化数据，我们需要使用一定的数据库对其进行管理，可能会在数据库的基础上对其进行管理和研究。特别是对于需要及时读取的实时可视化系统，对于大量数据读取和管理，需要对数据库数据进行快速搜索引擎。

（5）数据的挖掘和可视分析。数据挖掘需要完整的、可信的并且有用的数据，通过挖掘和统计等方法，获取有用的信息。而数据可视化是一项结合图形学知识和美学知识的工程，通过将大量的数据设计成美观的图形图像，将数据的多

维信息展示出来，让使用者通过对数据多维度分析，从大量数据信息中挖掘出隐藏着的有用信息。数据可视化技术在商业、金融、医学、化工等不同行业的应用，促进了行业的发展，同时也推动了可视化技术的进一步提升。

大数据推动了可视化技术的发展与探索，随着微博、淘宝、商业、社交网络、地理信息数据等的增加，越来越多的人投身到数据挖掘和可视分析中。例如微博数据中就包含了各种热门话题和热门人物，其中 Twetter 和 Tweets 话题就通过意见流的形式进行展示，从意见流中很容易分析在某个时间最热门的话题或人物，并且可以很容易看出一些话题的开始和结束。现有的数据可视化技术主要有文本信息可视化、网络信息关系可视化、时空数据可视化、多维数据可视化。

（1）文本信息可视化。主要将文本信息进行分析处理，对文本网络语句进行分词、统计和信息挖掘，通过不同形式的标签云、事件河流图、层次结构树等以不同的形式对文本信息进行聚类分类，从而获取文本中的关键信息。

（2）网络信息关系可视化。各个互联网站信息，微博、社交网络、通信等关系的信息都是数据源。通过使用不同结构和形式的树、力导向布局图、空间树、和弦图等方式可以有效地对网络关系进行可视化。同时对于这种网络关系，我们可以使用适当的边绑定和聚类分析等技术，将复杂的网络关系聚集并且进行可视化，同时结合用户交互的方式。

（3）时空数据可视化。主要是针对地理位置数据进行的可视化，例如交通流通数据、天气指数数据、空气质量数据等等。这些数据具有较强的地理位置属性，它包含了大量的地点时间标签，其中可以通过使用 Flow map 将事件流和地图进行融合，其中可以使用二维地图、三维地图等，结合边绑定和统计分析等技术，可以实现地理数据的实时可视化。现有的热力图也是时空数据分析的一种非常可行的技术，通过热力图对整体的地理数据进行分析，更加能够获得整体信息。

（4）多维数据可视化。多维数据即涵盖了各个维度和信息的数据，这些数据具有多个属性，其数据来自于各种企业、金融、政府等，这些数据以惊人的速度与日俱增，通过分析多维数据可以挖掘商业价值，给企业提供有效的商业指导。其中最为常见的方法有平行坐标、散点图、投影等。这些方式通过将多维信息投

影到二维空间中，从不同的视角对数据进行观察分析，可以更加概括地获得多维数据隐含的信息。

以上的一些数据可视化的技术，都是平常可见且应用比较多的技术。随着可视化技术的越来越发展，越来越多的人投身的一些技术的创新和可视化系统的设计上来。基于对大数据有一个较好的分析和理解，就需要选择合适的可视化方法对数据进行可视分析和展示，结合人机交互等技术，让可视化技术越来越受到各个行业的关注，可视化以直观的展现形式深入我们的生活，客观形象地刻画数据信息。

1.2 数据收集、清洗与匹配

在科研或者商业领域，数据的收集是相当重要的，而在今天，硬件设备的发展使得存储海量数据成为现实，但是数据的收集仍然是一个难题。在这么大规模的数据集中，很难获得有用的信息，那么数据的清洗和匹配显得尤为重要。

数据来自什么地方？数据是够准确？数据是否全面？以及如何将数据转换成可以使用的图表和文件？都是当前高校和企业所关注的中点。

现在所谈论最多的是以需求为导向的大数据分析，这就要求我们有明确的目标，有了目标才能知道我们如何获取数据、使用数据以及从何种角度分析。假如已经有了清晰的目标，那么接下来的工作就是数据的收集和准备，这往往是很困难并且耗时很大的工作。在可视化领域，面相图表的数据几乎无法直接使用，其格式也无法被很好地解析。

可视分析中第一个难点是确定要收集什么数据。首先，识别有哪些与目标相关的数据可用，这些数据是否能够很好地解决任务目标。每一个可视化视图都是有若干个相关的数据所构成，而分析这些数据之间的联系是很有必要的。

在通过各种手段拿到数据时候，下一步要面临的问题就是数据的清洗。数据的格式很少能够直接使用，大部分时候，数据是杂乱的。很可惜的是，几乎所有的图表无法展示杂乱的数据，所以数据的清洗工作在整个数据分析和可视化的流程中是十分重要的，只有得到相对干净的数据，才能把其提供给绘图工具完成可视化视图的绘制。

原始数据集中导致出现数据质量问题的原因可能有以下几种：1）不一致的命名方式，由于不同数据集的来源可能不同，那么命名方式的差异几乎成为了一个必然，而这会导致数据清洗与分析过程中的困难呈倍数增长，这就需要我们规范不同数据集的命名方式，并且把这些数据合并到一个数据集中；2）同样多个数据来源会导致重复数据的出现，这就引申出了数据处理中的很重要的一个技术，那就是去重，选择合适的去重算法不经能够很好的处理数据，也能够节省很多时间；3）还很常见的数据问题还有空数据或者无效数据，这可能是由于硬件问题或者数据库管理人员的测试导致，这些数据可能是空的或者是伪造的，这对数据的处理与分析造成了很大的困难，我们在进行可视分析之前应该删除这些数据；4）除此之外还有一些例如单位不统一、数据缺失等问题，都需要根据实际情况去选择删除或者重新收集。

在经过数据收集和数据清洗之后，我们能够得到相对质量不错的数据集，在一些简单图表中能够展示一些数据之间的联系，但这是远远不够的，因为数据之间的内在联系很难直接获取，这就要求我们对数据与数据间的匹配，数据与图表间的匹配花费更多的精力。

现在的可视化绘图工具有很多，但是机会所有的这种类型的工具对数据输入格式都有着严格的要求，而能够很好地进行订制图表绘制的工具相对的学习成本有很高，这对可视图表的绘制造成了一定困扰。

在绘制相对复杂的可视化视图时，我们通常需要对数据与数据进行匹配，不仅仅是简单的连接，而是数据分析层面上的匹配，这就要求我们找寻数据的内在关联。在这之后需要进行数据与图表的匹配，对于同一个数据集，可能有多种图表可以对其进行展示，但是每种图表类型都具有各自的优势，如何选择合适的方式是需要长时间尝试与积累的。

1.3 统计分析

统计分析是指运用统计方法及与分析对象有关的知识，从定量与定性的结合上进行的研究活动。同时运用统计方法、定量与定性的结合是统计分析的重要特征。随着统计方法的普及，不仅统计工作者可以搞统计分析，各行各业的工作者

都可以运用统计方法进行统计分析。大数据的发展也为统计分析领域带来了新的机遇和挑战。

大数据分析的第一步就是统计分析，统计数据可以提供丰富的信息，一些高层面的统计数据将回答几乎全部数据浅层规律和部分深层规律。在实际工作当中，月度、季度、年度，都需要拿出这段时间的数据来做工作汇报，在数据收集、数据整理和数据分析的过程当中，难度很大。特别对于基层团队管理人员，数据收集统计还可能多次反复，感觉耗时耗力。

而在实践中，大数据统计分析有一些现实的困境：1）基础数据可比性差，按统计分析的三个步骤来看，第一个步骤(收集数据)就出现了问题，但是在大数据时代，这个问题能够得到不错的解决；2）流程工具偏多，统计学工具有很多，统计学方法也有很多，但是这些方法的选择和合理使用是很困难的；3）对数据的掌握程度低，在大数据时代，统计学的发展离不开数据的发展，但是传统的统计学方法只能得到部分数据之间的联系规律，这就导致了对数据的掌握程度低。

在统计分析的研究中得出的结论依旧具有足够的价值，但是也会出现很多问题，关键在于几个方面：1）基础数据的整合，现有的数据，因为历史的原因，可能比较粗糙。所以需要深入挖掘才能体现它的价值，互联网公司把数据作为一种资源，我们也可以向它们看齐，做数据资源的整合；2）数据的统一，同一规则下的数据能够很好地进行整合，从数据本身的角度看，统一数据规范显得更加重要；3）指标预测模型的建立，数据分析的几个阶段：常规报表、查询、多维分析、报警、统计分析、预报、预测模型、优化。随着高层质量意识的加强，大家都越来越关注到客户发生的异常，特别是重大异常。

1.4 图布局

图布局是可视化视图绘制中极其重要的环节，如何选择合适的布局方式对可视分析的成功与否有直接的联系，现有的布局方式大致有以下几种：地图、弦图、邻接矩阵、树图、平行坐标、力导向图、饼图、柱形图、线形图以及散点图等。

1.4.1 弦图

弦图是展示多个节点之间的联系的图表，而两点之间的连线则表示两者之间

具有联系，线的粗细则表示权重，如图 1.1 所示。

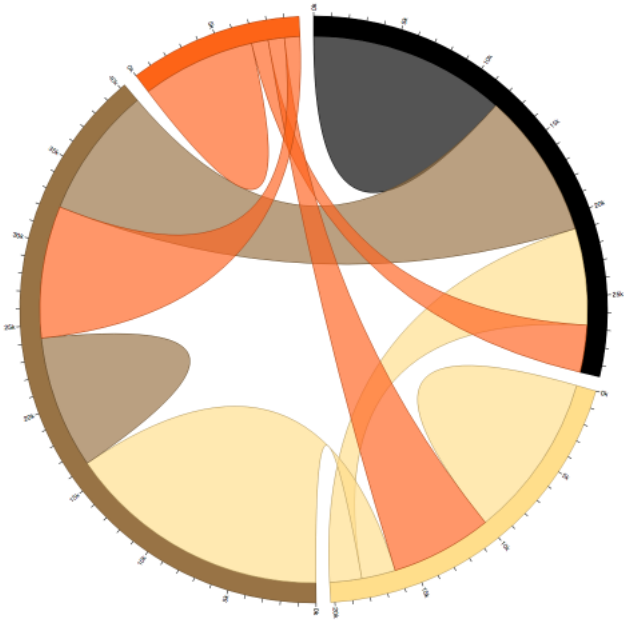

图 1.1 弦图

弦图（Chord Diagram）通过用一条弦连接对应的两个节点，来表示每个方向上的数据流动。其中弦的粗细指出了该节点流出的流量。通过在视觉上比较弦两端的大小，观察者能够感受到每个方向上的总流量。在视觉上比较两端能够看出两者的区别。

1.4.2 地图

基于地图（Map）绘制可视化视图是展示时序地理空间信息最好的方式，他能够很直观的表现数据在空间上的位置与变化，但是，同样由于地图的局限性，很难挖掘一些多维数据间隐藏的规律。一些地图可视化视图如图 1.2 和图 1.3 所示。

图 1.2 使用 D3js 绘制的世界地图轮廓

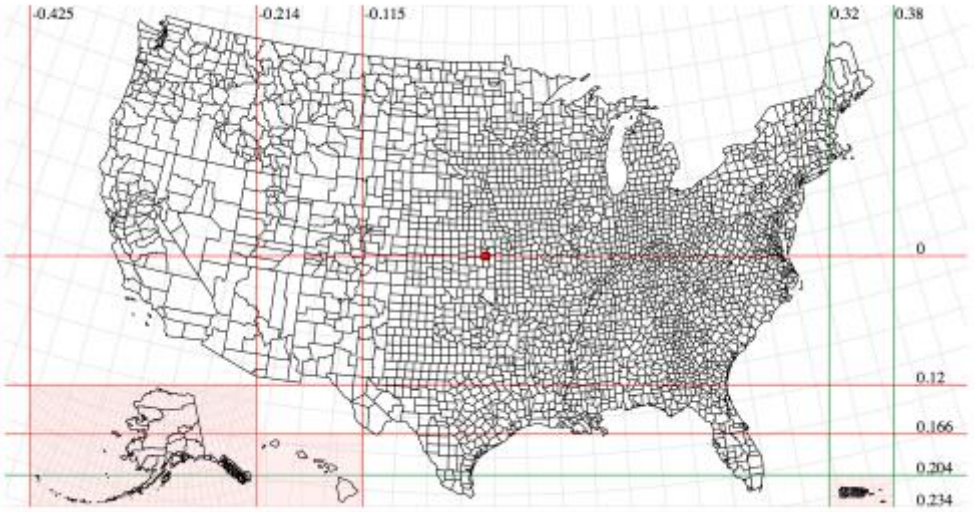

图 1.3 美国人口密度网格化视图

地图视图的展示形式与其余布局方式比较来说要相对固定，但是由于其能够直观的显示地理空间信息，仍旧成为最广泛使用的可视化绘图布局方式之一。

1.4.3 树图

树状图（Tree）用于表示层级、上下级、包含和被包含的关系，原始的树图布局属于基础统计图表之一，但是由于树图的特性，使其的布局方式发展迅速，原始的树状图布局如图 1.4 所示，而更多树形布局方式如图 1.5 和图 1.6 所示。

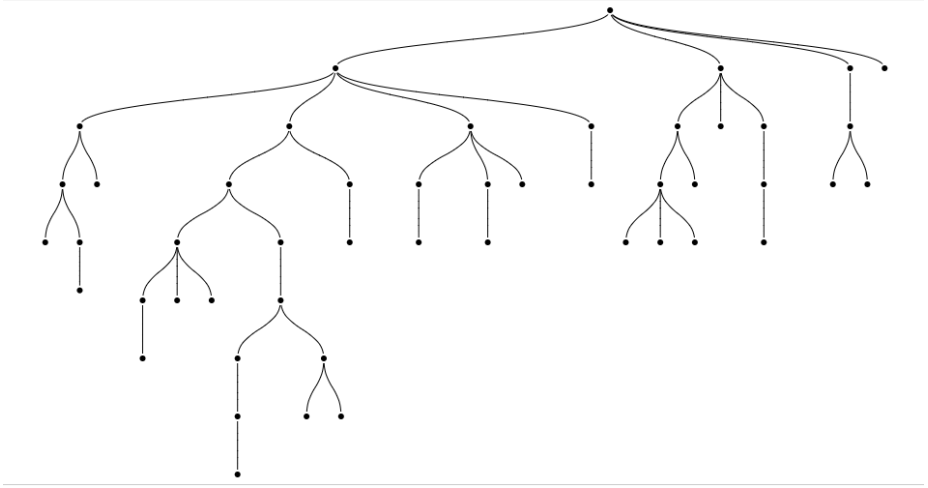

图 1.4 基础树状图布局

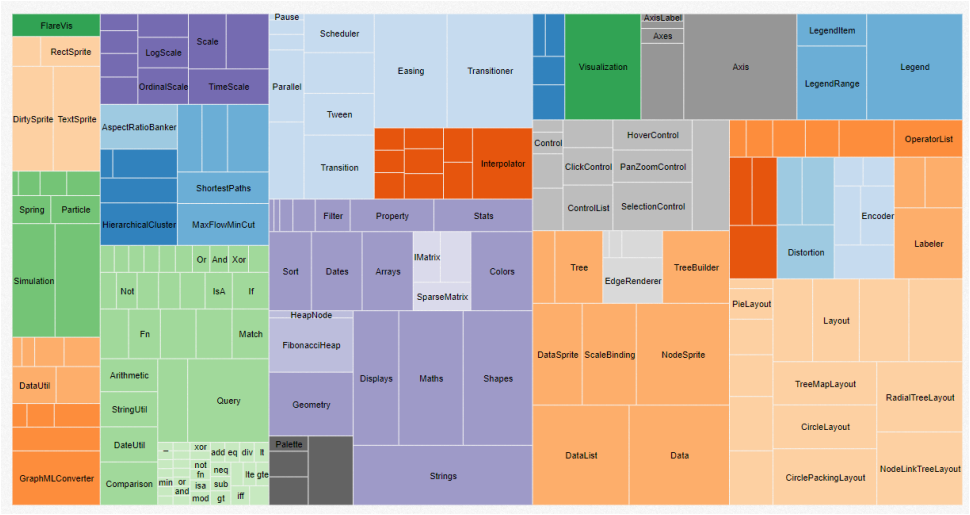

图 1.5 可视化绘图技术树形图

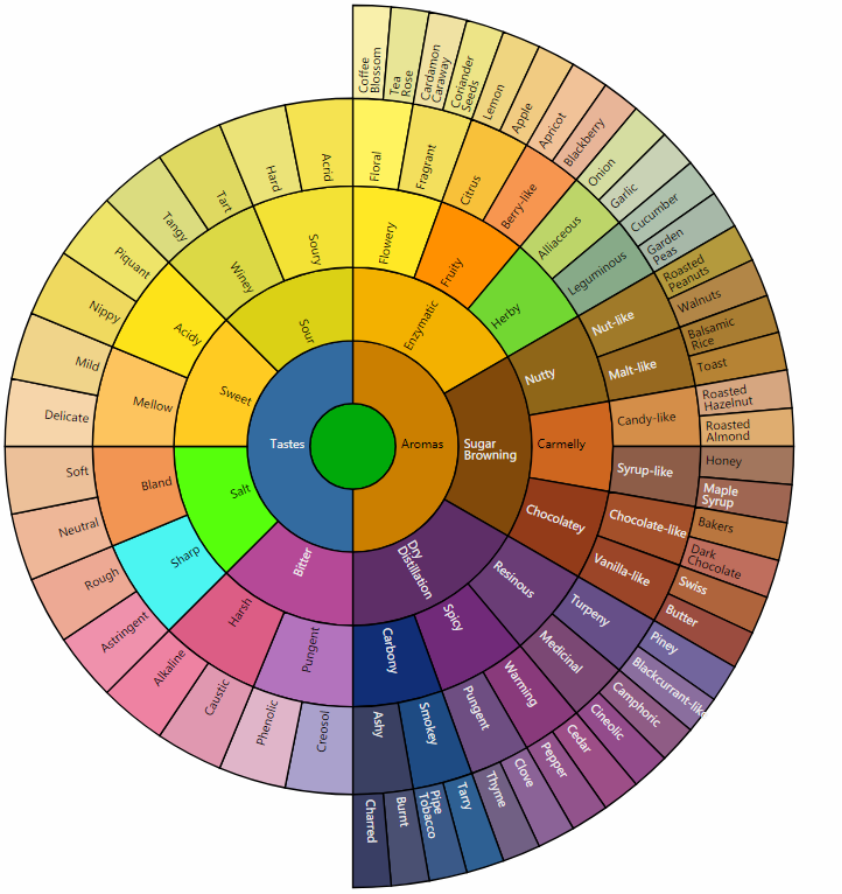

图 1.6 咖啡口味树形图 •

1.4.4 邻接矩阵

邻接矩阵（Adjacency Matrix）是表示顶点之间相邻关系的矩阵。它最初是图的一种数据存储方式。在可视化领域，邻接矩阵视图通常用来展示数据间的统计关系，譬如相关性或者相关因子，如图 1.7 所示。

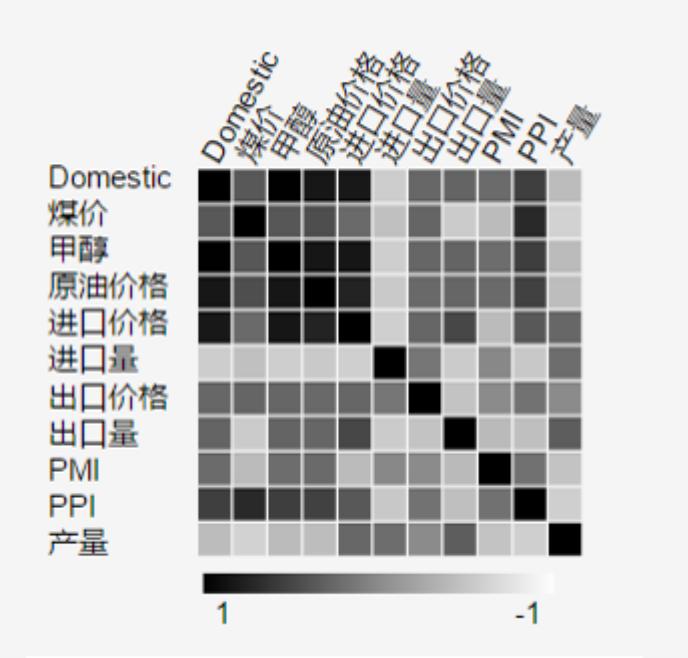

图 1.7 甲醇价格影响因子邻接矩阵视图

1.4.5 平行坐标

平行坐标（Parallel Coordinates）是一种通常的可视化方法，用于对高维几何和多元数据的可视化。为了表示在高维空间的一个点集，在 N 条平行的线的背景下，（一般这 N 条线都竖直且等距），一个在高维空间的点被表示为一条拐点在 N 条平行坐标轴的折线，在第 K 个坐标轴上的位置就表示这个点在第 K 个维的值。

平行坐标是信息可视化的一种重要技术。为了克服传统的笛卡尔直角坐标系容易耗尽空间、难以表达三维以上数据的问题，平行坐标将高维数据的各个变量用一系列相互平行的坐标轴表示，变量值对应轴上位置。为了反映变化趋势和各个变量间相互关系，往往将描述不同变量的各点连接成折线。所以平行坐标图的实质是将维欧式空间的一个点 $Xi(xi_1,xi_2,...,xi_m)$映射到维平面上的一条曲线。

平行坐标图可以表示超高维数据。平行坐标的一个显著优点是其具有良好的数学基础，其射影几何解释和对偶特性使它很适合用于可视化数据分析。图 1.8 展示了平行坐标布局的常见方式。

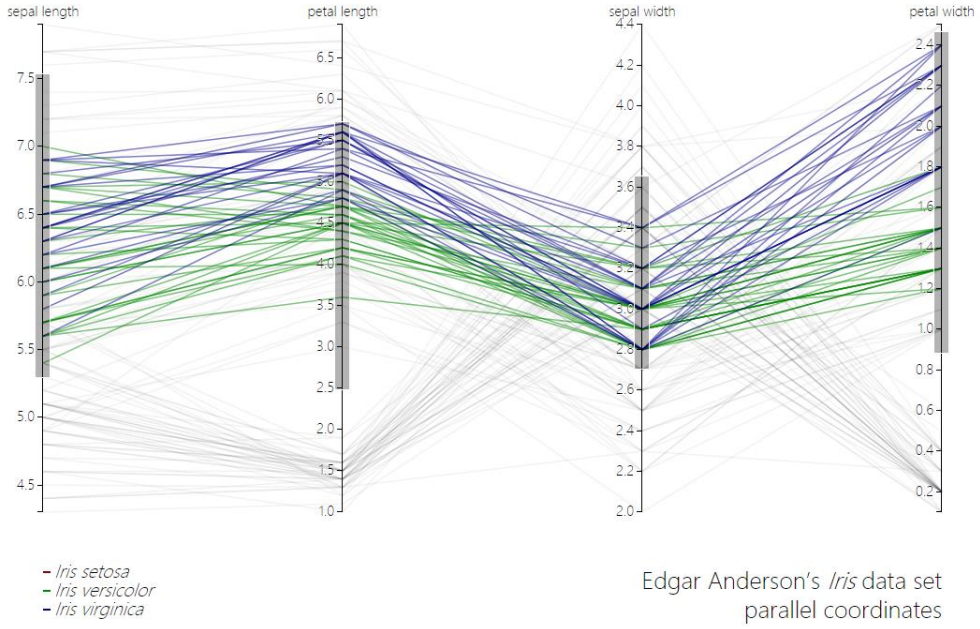

图 1.8 关于虹膜数据的平行坐标视图

1.4.6 力导向图

力导向图（Force-Directed Graph），是绘图的一种算法。在二维或三维空间里配置节点，节点之间用线连接，称为连线。各连线的长度几乎相等，且尽可能不相交。节点和连线都被施加了力的作用，力是根据节点和连线的相对位置计算的。根据力的作用，来计算节点和连线的运动轨迹，并不断降低它们的能量，最终达到一种能量很低的安定状态。

力引导算法主要应用与复杂网络可视化。力引导布局最早由 Peter Dades 在 1984 年的“启发式画图算法”文章中提出。目的是减少布局中边的交叉，尽量保持边长一致。主要引入库伦斥力和胡克弹力，考虑阻尼衰减（这就是为什么我们拉动力引导图，它能很快稳定回来的原因）。事先定义好图里的点，边的权重等信息，力引导图可以根据实时状态自动完成很好的聚类，方便地看出点之间的亲疏关系。力导向图布局如图 1.9 所示。

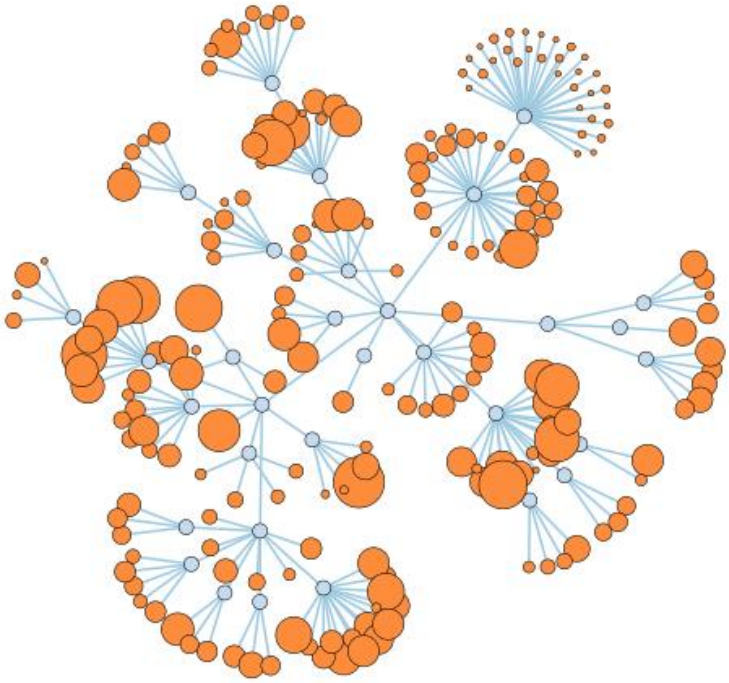

图 1.9 耀斑大小力导向图布局

1.4.7 基础统计图表

基本统计图表的应用有悠久的历史，其中使用最为广泛的有扇形图、饼状图、柱状图和散点图等。其中不同基础统计图表可以相互融合，用来同时展示多个数据集或者一个数据集中的不同属性。基础统计图表布局如下图所示。

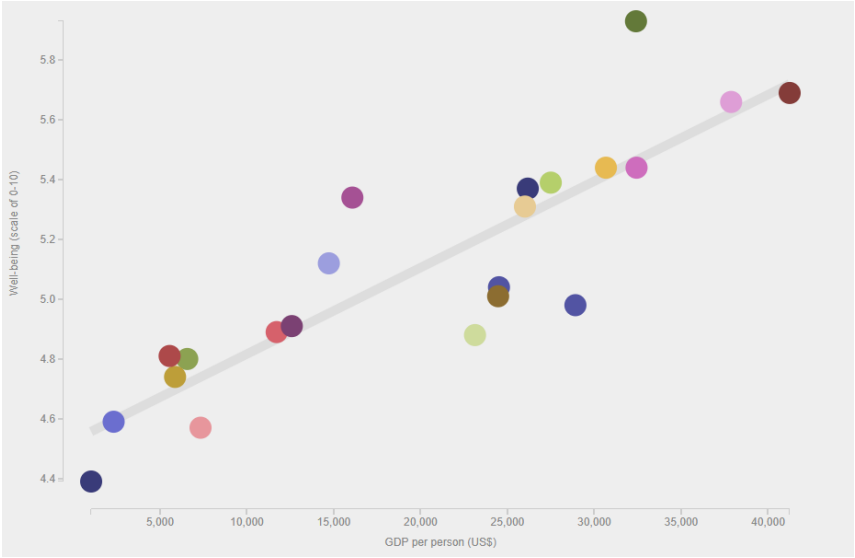

图 1.10 GDP 与幸福度散点图

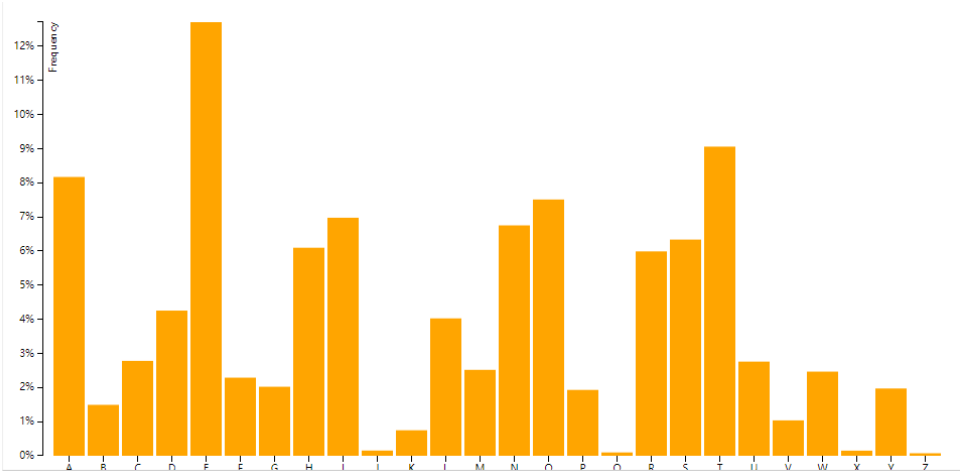

图 1.11 简单的柱状图布局

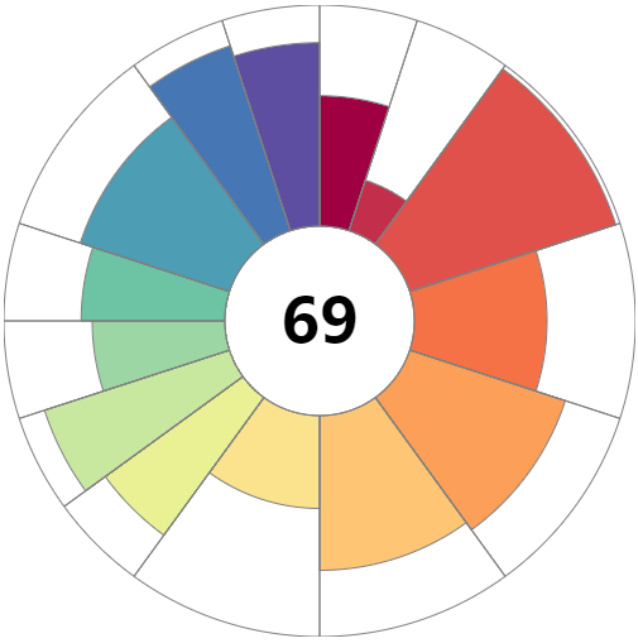

图 1.12 环境情况扇形图布局

1.5 视觉探索

当有一个基本的图和一个合理的布局之后，下一步就是应用视觉特性进行视觉探索。在商业环境中，通常存在大量统计性数据，例如年龄、收入、性别等。通过利用视觉手段，例如颜色、线条、标签等，可以展示这些信息。选择正确的视觉展示方式能够利用人们的感知能力。例如，亮色的视图在整体灰暗的背景下会更加突出，更大的图表比小的图表更能引起人们的注意等。

视觉探索中能够利用的视觉特性主要有：1）节点特性，不同的节点的属性

有差异，运用合理的颜色等手段可以解释节点的特征；2）连接特性，属性之间的强弱联系能够通过视觉手段更加突出的展示，最简单的方式就是用线条的粗细代表联系的紧密程度；3）在此之外，标签的应用能够使用户更加快速的理解图表内容，在视觉探索中也占据很重要的地位。

节点是对具有特定意义的名词的统称，例如人、计算机、国家或者某一个网站都可以用节点表示。这些名词中常常有额外的特性，但是在图的布局中是看不到的，但是还需要把这些特征展示出来。这就要求我们合理的使用视觉特征来完成这一工作。

节点的大小是用来表示量值的数据，大小是很直观的一个视觉特征。统计数据中的计数、和等数据非常适合用大小来表示。但是如果一个数据集有负值或者零值，就要考虑使用比例等方式来展示这些属性，或者使用绝对值来表示。设计者可能希望人们看到图后，能够在视距上估计节点的相对大小，这样就要使用一些手段，例如用更大的节点表示两倍的数据，这可能与实际值有差别，但是在视觉感受中能够更加直观。相对大小同样十分重要，尤其是当动态范围超过了两个量级的时候。特别大的范围和精确的大小存在的问题是，一些节点是极小的点，而其他的点非常大，以至于会覆盖掉其他节点，导致观察值理解产生偏差，这种情况下，使用相对大小会显得很有用。

除了节点的大小、节点的颜色同样十分重要。在表示数据时，颜色是一种强大的视觉指示器。颜色通常有不同的特性，例如色调、亮度和饱和度等，这些特性各不相同。例如，使用红色表示温度高、警告或者错误，使用蓝色表示温度低或者夜晚等。

标签是图中很重要的一部分。标签能够清晰的标示具体的内容。而标签的使用也会有一些技巧。例如在图表中标注姓名，把整个名字显示出来是不现实的，这就要求我们使用合适的短标签。如果标签数目过多也会长生互相覆盖的问题。

节点之间的联系，即边的特性也能够从不同的角度展示数据信息，其中最简单的是边的权重。在视觉上，边的权重就是两个节点间边的粗细程度。除此之外边的颜色也很重要，它去节点颜色的使用情况类似。和节点不同的是，边还具有

独特的类型和弧度，这都是数据展示过程中需要注意的地方。

在可视化领域有一项关于边的研究被称为边绑定，这种技术很好的解决了复杂网络图中边交织不清的现象，但是也在一定程度上模糊节点之间的联系，图 1.13 是边绑定的展示。

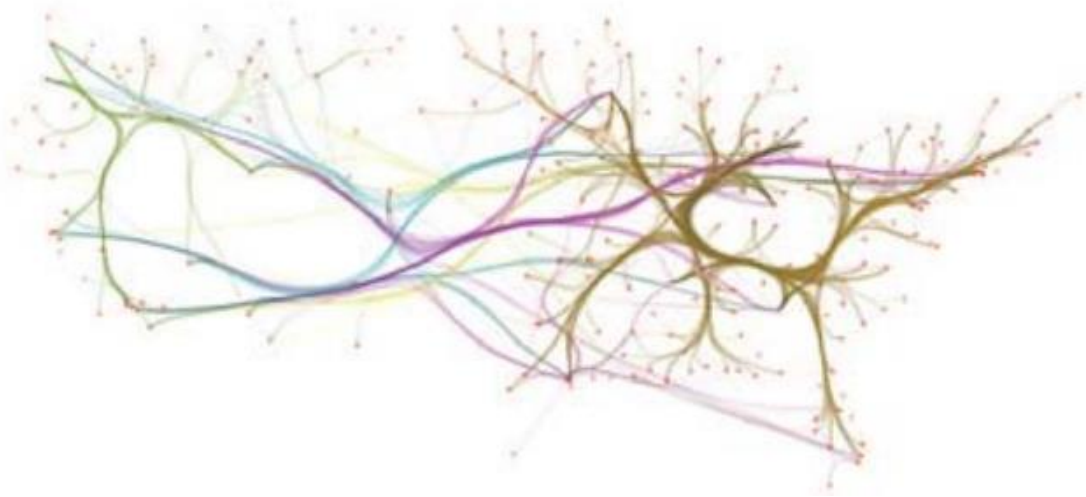

图 1.13 美国航空航线图

1.6 可视化分析

可视分析是大数据分析的重要方法。大数据可视分析旨在利用计算机自动化分析能力的同时，充分挖掘人对于可视化信息的认知能力优势，将人、机的各自强项进行有机融合，借助人机交互式分析方法和交互技术，辅助人们更为直观和高效地洞悉大数据背后的信息、知识与智慧。主要从可视分析领域所强调的认知、可视化、人机交互的综合视角出发，分析了支持大数据可视分析的基础理论，包括支持分析过程的认知理论、信息可视化理论、人机交互与用户界面理论。

当前,我们的世界已经迈入大数据时代。随着互联网、物联网、云计算等信息技术的迅猛发展，信息技术与人类世界政治、经济、军事、科研、生活等方方面面不断交叉融合，催生了超越以往任何年代的巨量数据。遍布世界各地的各种智能移动设备、传感器、电子商务网站、社交网络每时每刻都在生成类型各异的数据，他们体量巨大、类型繁多、时效性高以及价值高密度低,给人们带来了新的机遇与挑战。《Nature》于 2008 年出版了大数据专刊“big data”,专门讨论了巨量数据对于互联网、经济、环境以及生物等各方面的影响与挑战。《Science》

也于 2011 年出版了如何应对数据洪流的专刊“Dealing with Data”，指出如何利用宝贵的数据资产推动人类社会的发展。

如今，大数据已成为新兴的学术研究热点，并被认为是继云计算和物联网之后又一个具有革命性的信息技术。大数据分析是大数据研究领域的核心内容之一。Google 首席经济学家、UC Berkeley 大学 Hal Varian 教授指出：“数据正在变得无处不在、触手可及；而数据创造的真正价值，在于我们能否提供进一步的稀缺的附加服务。这种增值服务就是数据分析。”数据的背后隐藏着信息，而信息之中蕴含着知识和智慧。大数据作为具有潜在价值的原始数据资产，只有通过深入分析才能挖掘出所需的信息、知识以及智慧。未来人们的决策将日益依赖于大数据分析的结果，而非单纯的经验和直觉。美国《时代》杂志于 2012 年 11 月指出，奥巴马的成功连任背后所依托的关键即是两年来对大数据的分析与挖掘，例如，通过对海量选民微博的分析得出选民对总统候选人的喜好。

当前，大数据分析方法论以及支撑技术的研究成为大数据领域的核心焦点之一。通常，数据的分析过程往往离不开机器和人的相互协作与优势互补。从这一立足点出发，大数据分析的理论和方法研究可以从两个维度展开：一是从机器或计算机的角度出发，强调机器的计算能力和人工智能，以各种高性能处理算法、智能搜索与挖掘算法等为主要研究内容，例如基于 Hadoop 和 MapReduce 框架的大数据处理方法以及各类面向大数据的机器学习和数据挖掘方法等，这也是目前大数据分析领域的研究主流；另一个维度从人作为分析主体和需求主体的角度出发，强调基于人机交互的、符合人的认知规律的分析方法，意图将人所具备的、机器并不擅长的认知能力融入分析过程中，这一研究分支以大数据可视分析为主要代表。人类从外界获得的信息约有 80%以上来自于视觉系统，当大数据以直观的可视化的图形形式展示在分析者面前时，分析者往往能够一眼洞悉数据背后隐藏的信息并转化知识以及智慧。大数据可视分析是大数据分析不可或缺的重要手段和工具。事实上，在科学计算可视化领域以及传统的商业智能领域，可视化一直是重要的方法和手段。然而，这些研究领域并未深入地结合人机交互的理论和技术，因此难以全面地支持可视分析的人机交互过程。同时，大数据本身的新特

点也对可视分析提出了更为迫切的需求与更加严峻的挑战。总体而言，当前对于大数据可视分析的研究仍十分初步，对于这一研究领域的理论、方法和技术体系至今尚未形成。

1.7 本章小结

本章介绍了数据可视化项目的流程与视图类型，并分析了视觉探索的意义和方式，最后对可视化分析进行了介绍。这是对我几年可视化研究中所使用的方法和遇到的问题的一个总结。

第2章 交通数据分析

本文综合了数据分析和数据挖掘的相关知识以技术，并结合可视分析手段，进行了关于交通数据和商业数据的一系列分析工作，并设计了可视化系统来帮助用户理解数据内涵以及进行交互式的可视分析。

本章主要介绍了关于交通大数据的分析与可视化，主要分为两个部分，分别为地铁站站点评级分析和交通人流数据分析。

2.1 轨道交通站点评级

国内地铁站分级方法的研究已有很久的历史，但是都有各自的局限性。同时由于轨道交通管理的复杂性，对政府制定相应策略造成困扰。为了能够提出适合当前时代的分级方式，提高政府策略制定效率，我们研究了上海市轨道交通刷卡数据进行人流量分析，并通过区分人群类型进行进一步研究。我们结合人流量、人群类别、站点特性并结合轨道交通和商场一体化现象提出了一种站点的分级方法。最后我们进行了分级方式对比。

上海市是国际型大都市，平均客流量在 2000W 以上，其中占据轨道交通占据了 50%，约有 950W 人次/天。同时商业的发展使上海市轨道交通站点和商业场所一体化程度越来越高，很多大型的枢纽站与大型商场合为一体，这对轨道交通站点评级造成了很大的困扰。同时分级方式的缺陷导致了很多资源的浪费，以及部分重要站点重视程度不足的现象。

我们基于轨道交通刷卡数据进行了地铁站繁荣程度的研究，并结合轨道交通和商场一体化的现象以及轨道交通站点固有属性对其进行了修正，最后提出了一种新的分级方式，能够有效地划分国际型大都市的轨道交通站点级别。

最后我们结合前人的研究成果，使用多种方法进行站点等级划分，并与政府公布站点等级进行对比。

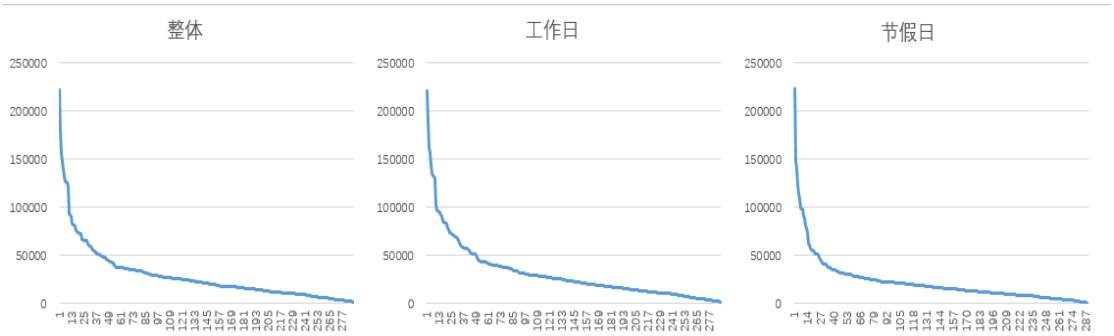

图 2.1 上海地铁站人流累计图

图 2.1 上海市 288 个轨道交通站点的人流累计图，其中不区分工作日和节假日的话，大概在 40000 人/天的节点处会形成差别很大的图形断点，而工作日在 45000 人/天，节假日在 30000 人/天。同时经过统计分析，我们发现断点位置大概处于所有站点的前 20%出，而占据了上海市每天人流总量的 50%。

我们使用 Yun Wei 等人提出的方法进行了站点评分计算。设立分界点 F_x=40000，评分中心点S_x=50，以及F_{min}=687（最小人流量—华夏中路站），F_{max}=222405（最大人流量--人民广场站），S_{min}=0，S_{max}=100 进行计算，计算方法如下：

如果当前站点人流量<F_x，那么当前站点评分为：

$$S_i = \frac{F_i - F_{min}}{F_x - F_{min}} S_x \qquad (1.1)$$

如果当前站点人流量>=F_x，那么当前站点评分为：

$$S_i = \frac{F_i - F_x}{F_{max} - F_x} (S_{max} - S_x) + S_x \qquad (1.2)$$

通过计算我们得到了站点人流量因素评分表，如表 2.1 所示。

同时我们还分别计算了区分工作日和节假日的评分结果，对实验结果（图 1）对比发现，在单独计算工作日与整体计算结果对比时，评分误差都在 5 之内，而计算节假日时误差则扩大到 [-22,16] 之间，因此我们认为根据节假日人流差异进行研究是很有必要的。

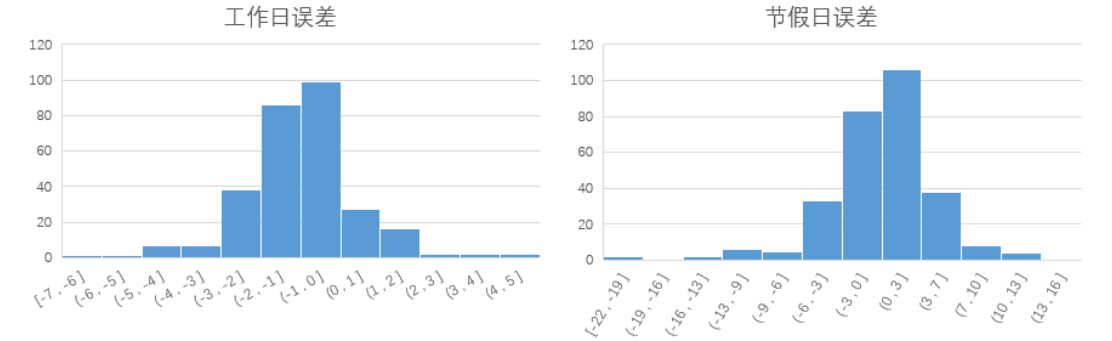

图 2.2 计算误差

表 2.1 站点等级表

上海地铁 2 号线	站点因素			人流量因素		站点评分	归一化	站点等级
	类别	等级	修正评分	人流量因素	流量评分			
人民广场	B	b	1.1	222405	100	110	100	A
北新泾	C	d	0.9	35952	45	40.35709	36.68826	C
川沙	C	d	0.9	25062	31	27.86615	25.33287	D
创新中路	C	d	0.9	2614	2	2.119017	1.926379	D
东昌路	C	d	0.9	62492	56	50.54895	45.95359	C
广兰路	C	c	1	60571	56	55.63883	50.58075	C
海天三路	C	d	0.9	2380	2	1.851446	1.683133	D
虹桥 2 号航站楼	A	a	1.2	30065	37	44.80676	40.73342	C
虹桥火车站	A	b	1.1	81770	61	67.5949	61.44991	B
华夏东路	C	d	0.9	16850	20	18.44737	16.77033	D
江苏路	B	c	1	72158	59	58.81498	53.46816	C
金科路	B	c	1	59197	55	55.26225	50.23841	C
静安寺	B	b	1.1	140566	78	85.3233	77.56664	B
凌空路	C	d	0.9	4738	5	4.555906	4.141733	D
龙阳路	C	b	1.1	67056	57	63.15824	57.41658	C
娄山关路	B	c	1	73817	59	59.26985	53.88169	C
陆家嘴	B	b	1.1	126356	74	81.03863	73.67148	B
南京东路	B	a	1.2	126734	74	88.53007	80.48188	A
南京西路	B	b	1.1	91430	64	70.50737	64.09761	B
浦东国际机场	A	a	1.2	16528	20	24.10438	21.91308	D
上海科技馆	B	c	1	29024	36	36.01168	32.73789	C
世纪公园	B	c	1	29577	37	36.71715	33.37923	C
淞虹路	C	d	0.9	75332	60	53.71659	48.83326	C
唐镇	C	d	0.9	26846	33	29.91332	27.19393	D
威宁路	C	d	0.9	33973	42	38.087	34.62455	C
徐泾东	C	b	1.1	83272	62	68.04769	61.86154	B
远东大道	C	d	0.9	3091	3	2.666446	2.424042	D
张江高科	B	c	1	59996	55	55.48133	50.43757	C
中山公园	B	b	1.1	132038	75	82.75205	75.22914	A

我们发现由于上海市属于国际经济中心，因此简单实用人流量评分并通过静态因素（站点位置、规模等）进行修正的方法不是很有效的。因此我们引进了人群类型区分的方法以及轨道交通与商业一体化因素对这个问题进行深入研究。

我们认为一些轨道交通站点具有很明显的特征，总结之后我们把站点累心分为 3 类：商业区域，工作区域和普通区域。同时一个轨道交通站点可能同时属于

多个类别。我们没有单独划分旅游景点的原因在于，上海市几乎所有的景点都实现了与大型商场的一体化。同时上海市轨道交通站承担了不同程度的枢纽任务，因此我们加入了枢纽等级进行修正。等级与修正指数如表 2.2。

表 2.2 地铁站评级修正值

	商业区域		工作区域		居住区域		枢纽等级	
级别	描述	修正值	描述	修正值	描述	修正值	描述	修正值
A	国际商业中心	1.5	金融中心	1.5	无	无	国际客运中心	1.5
B	中央商业区	1.4	国家商务区	1.4	特大型居住区	1.4	国家客运中心	1.4
C	市级商业中心	1.3	市级商务区	1.3	大型居住区	1.3	城市交通枢纽	1.3
D	区级商业中心	1.2	区级商务区	1.2	中型居住区	1.2	区域交通枢纽	1.2
E	社区级商业区	1.1	无	无	小型居住区	1.1	小型交通枢纽	1.1

上海市作为国际经济中心，其发展现状与其他类型城市有很大不同，其中很明显的一点就是，轨道交通枢纽站与大型商场的一体化。XR Lin [2] 等人对上海市五角场地区进行了研究，五角场作为上海市杨浦区最大的商业中心，同时具有交通枢纽、商业商务等多种功能。因此在对这类站点进行研究时，单一分类进行评分修正是不准确的，因此我们认为进行累计修正能够更好地为轨道交通站点评级。我们认为评级公式应如下：

$$\mathrm{R}_i = \mathrm{S}_i \times \mathrm{Business}_i \times \mathrm{Work}_i \times \mathrm{Resident}_i \times \mathrm{Hub}_i$$

其中如果当前站点不具备某一项功能，那么此项修正值为 1。

我们通过计算得到了一组评分值，结果如表 2.3。

通过对比可以看出，我们的方法能够更加符合大众共识，但是评分标准更加具体，能够很好地为其分级，同时我们提供了三级标准和四级标准。通过对上海市交通状况进行分析，我们发现由于人流集中现象相对明显，因此我们认为当$\mathrm{R}_i\epsilon(50,100]$时为 A 级，同时三级情况下 B 级$\mathrm{R}_i\epsilon(25,50]$，C 级$\mathrm{R}_i\epsilon(0,25]$，四级情况下 B 级$\mathrm{R}_i\epsilon(30,50]$，C 级$\mathrm{R}_i\epsilon(20,30]$，D 级$\mathrm{R}_i\epsilon(0,20]$。

通过对以往的研究进行分析，同时基于交通卡刷卡数据进行居住于工作区域划分，轨道交通与商业一体化现象进行商业区域与枢纽等级划分，提出了一种适用于上海市的轨道交通站点分级方法。我们使用了对人流数据、商业等级、枢纽等级等因素进行评分，最后提出了三级和四级两种等级评定标准。最后通过实验结果对比，我们认为我们的方法能够有效地评定轨道交通站点等级，为行政管理与相关政策制定起到很好的辅助作用。

表 2.3 上海轨道交通 2 号线站点评级

名称	商业		工作		居住区域		交通	
	等级	评分	等级	评分	等级	评分	等级	评分
人民广场	B	1.4	B	1.4	/	1	C	1.3
北新泾	/	1	/	1	D	1.2	E	1.1
川沙	E	1.1	/	1	/	1	E	1.1
创新中路	/	1	/	1	/	1	/	1
东昌路	/	1	D	1.2	E	1.1	/	1
广兰路	/	1	/	1	/	1	D	1.2
海天三路	/	1	/	1	/	1	/	1
虹桥 2 号航站楼	/	1	/	1	/	1	A	1.5
虹桥火车站	/	1	/	1	/	1	B	1.2
华夏东路	E	1.1	/	1	/	1	/	1
江苏路	E	1.1	/	1	/	1	D	1.2
金科路	/	1	/	1	/	1	/	1
静安寺	C	1.3	C	1.3	/	1	/	1
凌空路	/	1	/	1	/	1	/	1
龙阳路	C	1.3	D	1.2	D	1.2	D	1.2
娄山关路	C	1.3	/	1	E	1.1	/	1
陆家嘴	C	1.3	A	1.5	/	1	/	1
南京东路	A	1.5	/	1	/	1	C	1.3
南京西路	B	1.4	C	1.3	/	1	C	1.3
浦东国际机场	/	1	/	1	/	1	A	1.5
上海科技馆	D	1.2	/	1	/	1	/	1
世纪公园	/	1	D	1.2	/	1	/	1
淞虹路	/	1	/	1	D	1.2	/	1
唐镇	/	1	D	1.2	C	1.3	E	1.1
威宁路	/	1	/	1	E	1.1	/	1
徐泾东	/	1	/	1	B	1.4	D	1.2
远东大道	/	1	/	1	/	1	/	1
张江高科	D	1.2	C	1.3	/	1	/	1
中山公园	B	1.4	/	1	/	1	C	1.3

通过计算我们得到轨道交通站点划分表，并与其余方法进行对比，如表 2.4 所示。

表 2.4 评级结果对比

站名	媒体-共识(3)	OUR(3)	OUR(4)	Yun Wei(4)
人民广场	A	A	A	A
北新泾	C	B	C	C
川沙	C	C	D	D
创新中路	C	C	D	D
东昌路	B	B	C	C
广兰路	C	B	B	C
海天三路	C	C	D	D
虹桥 2 号航站楼	B	B	C	C
虹桥火车站	B	B	B	B
华夏东路	C	C	D	D
江苏路	B	B	B	C
金科路	B	B	C	C
静安寺	A	A	A	B
凌空路	C	C	D	D
龙阳路	B	B	B	C
娄山关路	B	B	B	C
陆家嘴	A	A	A	B
南京东路	A	A	A	A
南京西路	A	A	A	B
浦东国际机场	B	C	D	D
上海科技馆	B	C	D	C
世纪公园	C	C	D	C
淞虹路	B	B	B	C
唐镇	C	B	C	D
威宁路	C	C	D	C
徐泾东	B	B	B	B
远东大道	B	C	D	D
张江高科	B	B	B	C
中山公园	A	A	A	A

2.2 交通流量分析

由于公共地铁系统的便捷性，使它成为大多数上班族的首选出行方式。然而，随着交通流量数据在数量和种类上的急剧增加，使得设计有效的可视分析方法成为挑战。本章着重分析基于地铁刷卡数据的人群移动行为，并提出交互式的可视分析视图，旨在分析不同群体的移动行为和展示时序的交通流量信息。在流量快照可视化模块中，允许用户选择不同时段的地铁流量信息和分析上班族的居住地点和工作地点。地铁站之间流量变化展示在流量关系视图中，并可以分析不同人

群的出行特征。流量时序视图展示了整体的地铁流量数据。最后和交通研究者做了两个案例分析验证该系统。

本章的目标是结合可视分析技术从大量的地铁刷卡数据中分析不同上班族群体的移动行为特征以及展示城市地铁系统不同时段的交通流量变化。首先，从大量的地铁刷卡记录中发现上班族群体，按照出行持续时间将他们划分为常规上班族和非常规上班族。其次，根据上班族的刷卡数据推测近似的居住地点和工作地点。最后，本章设计了三个可视化分析模块来解决本章提出的三个分析任务。流量快照可视化模块展示地铁系统中不同时段的流量变化以及进站和出战流量展示，方便用户根据流量分布来推断上班族的居住地点和工作地点，用户可以选择感兴趣的地铁站进一步分析。站点流量关系可视化模块展示了选定站点之间的流量变化，该模块可以呈现上班族群体不同时段的出行特征，其中弦图表示选定站点与其他地铁线路的整体流量关系。地铁流量时序可视化模块展示了每一个地铁站为期一个月的交通流量情况，能够让用户针对不同的地铁站比较工作日和非工作日的地铁交通流量变化。

本章中使用的是上海市公共交通系统中的地铁刷卡数据。这些刷卡数据记录了每一个乘客的行程，为期一个月，其中包括工作日和非工作日。在公共交通系统中，乘客通过刷卡来进站或者出站，读卡机记录乘客的每一次刷卡行为。其中，每一条刷卡记录包括一个匿名的卡号、进站、出站、进站时间、出站时间和价格。每条记录仅有出站和进站，为了尽可能准确地获取每个乘客的完整行程，即从始发站到目的站途经的所有站点，本章利用上海市地铁交通网络数据来推断乘客的完整行程。一部分乘客的形成会涉及到换乘，换乘是指乘客在车站内进行跨线乘坐列车的行为。

本章把上海市地铁交通网络数据看作一个有向图。一个交通网络包含一些站点和路线信息，交通网络中的站点可以看作有向图中的节点，路线可以看作连接节点的边。利用构造的有向图能够推测出乘客的完整行程，从始发站到目的站可能存在多条线路，我们选用时间成本最少路线作为乘客的选择，上海市地铁系统按照乘坐里程计费，时间成本较少等同于价格较低，这比较符合人们的实际乘车经验。

地铁调度数据包括上海地铁系统中各个线路的进站和出站信息。

由于地铁的便捷性，乘坐地铁成为了大多数上班族的首选出行方式。从地铁刷卡记录数据中能够探索不同人群的移动行为特征。此外，不同时段的地铁交通流量需要以合适的方式展示出来，以便用户能够了解流量变化。本章的目标是结合可视分析技术来发现和分析不同人群的移动行为特征以及地铁交通流量变化。本章的分析任务总结如下:

1)如何分析和展示地铁系统中不同时段的流量变化。这有助于用户了解不同地铁线路和站点的流量变化。

2)如何从刷卡记录数据中发现上班族的居住地和工作地。确定的位置信息能够为分析不同人群的移动行为带来便利。

3)如何从刷卡记录数据中区分出不同的人群并能够多维度地展示人群的行为特征。

2.2.1 上班族行为分析

每天生成的上千万条刷卡记录中包含着不同的群体，如上班族，老人和游客等。本文的目标是分析上班族群体的移动行为特征和对应的交通流量变化。与其他群体相比，上班族群体的出行往往更有规律。他们的行程在工作日有着连续性，比如从周一到周五都有刷卡记录,并且在工作日有着相同的始发站和目的站。如果一个人的刷卡记录满足

$$W = \{W_i \mid |S_{in}| \geq 4, |S_{out}| \geq 4\} \quad (2.1)$$

其中，W表示所有满足条件的上班族集合；W_i表示对第i个乘客若满足至少在工作日连续四天存在刷卡记录并出现在相同的始发站和目的站，则认为该乘客属于上班族群体。

此外，为了进一步分析上班族的地铁移动行为特征，本文定义了常规的上班族和非常规的上班族。常规的上班族是指那些一周最多连续工作五天的人；非常规的上班族是指连续工作大于五天的人，通常会在周六或者周日加班。

2.2.2 居民聚集行为分析

确定的位置信息包括上班族的居住地点和工作地点，有助于进一步分析该群体的出行行为特征。对于上班族居住地点和工作地点的推断基于这样的假

设：上班族群体总会选择距离家较近的地铁站作为起始站；并会选择距离公司较近的地铁站作为终点站。该假设与我们的实际经验吻合。关于发现上班族的居住地与工作地的公式如下：

$$L_r = \{S_{in} |\ 5{:}00 < T_{in} < 10{:}00, w_i \in \mathrm{W}\} \quad (2.2)$$

$$L_w = \{S_{out} |\ 17{:}00 < T_{out} < 20{:}00, w_i \in \mathrm{W}\} \quad (2.3)$$

L_r表示所有上班族的居住地集合，对每一个上班族 w_i，如果其进站时间在上午五点到十点之间，那么进站 S_{in}就近似为w_i的居住地。L_w表示所有上班族的工作地集合，对每一个上班族 w_i，如果时间在下午五点到八点之间，那么出站S_{out}就近似为w_i的工作地。

基于人流数据的研究已经有很多年的历史，但是由于关注重点不同，分析相同数据所得出的结论也有所差别。这项工作是根据人流数据进行人群划分以及聚集地划分的研究。

2.3 本章小结

本章通过对交通数据的分析与研究，分别对地铁站等级划分以及人群类型划分进行了研究。主要数据集使用了轨道交通刷卡数据，但是由于这项数据无法完全等同于出行数据，因此这两项工作还有很大改进空间。

第3章 商业数据分析

3.1 商圈引力研究

我们通过分析不同模型的优劣以及相关因素的影响程度进行商圈引力模型的构建。具体做法是通过实际人流数据计算出商圈对所有地点的吸引力程度，用概率表示。之后使用经典模型进行计算，分析实验结果，找出实验结果产生误差的原因。最后分析相关影响因素的影响因子，构建引力模型。

商圈是零售业聚集的区域，通常是一个地理位置范畴。广义上来说就是城市中的各类零售商店的聚集而成的商业街区，包含餐饮，服饰，金融等各式各样的店铺；而狭义上来说是一家或者多家店铺的覆盖范围。本文基于城市的轨道交通数据和商业数据，对大型城市的核心商圈（广义）来进行的研究，探索了商圈的吸引力与辐射范围。

传统商圈研究中很大程度上是基于抽样与专家经验来完成，但是在如今复杂的商业环境下，已经无法满足需求。传统商圈分析主要考虑人口特征，经济基础特点，竞争状况和市场饱和度等因素，但是在大型城市，商圈遍布整个城区，经济基础特点、市场竞争等因素已经没有很大的区分度，这就要求我们根据实际情况来进行研究。

零售商圈是零售交易区域的辐射范围，但是商圈的概念并没有很明确的定义，本文中，我们认为零售商圈的商业企业聚集所形成的空间范围。同样，在商圈级别划分标准下，本文中所研究的商圈是指由核心商业圈和次级商业圈组成的空间范围。商圈理论中应用最广的是雷利法则和哈夫模型以及其的演化模型。

由于大多数情况下，企业很难获取详细的商业信息，那么如何选择投资地成为了一个难题，而雷利法则最早为企业提供了容易实现的决策指导。雷利法则认为商业也具有相互吸引的特性，它以万有引力定律为核心，来确定商圈吸引力临

界范围。但是雷利法则是以商圈为中心的研究，并且需要有较严格的前提才能使结果有效。在我们对上海十九个大型商圈进行研究之后，我们发现使用雷利法则是很难确定商圈范围的，由于商圈差异和人群行为等因素。

我们计算了哈夫中两种主要变量对概率值的显著性与相关性，如表3.1，我们可以看到商业面积与时间成本对商圈吸引力来说都具有显著地相关性，同时时间成本的相关系数为-0.489，商业面积的相关系数为0.149。在与零售企业经理讨论后，我们认为所得系数是相对可靠的，因为在上海，交通相对发达，时间成本的影响程度已经远没有十几年前那么大，而由于研究样本都为核心商圈，商业面积的区别程度不大，因此得出的影响因子过小。在对大型商圈的研究中这也是可信的。

表 3.1 时间成本与商业面积对概率的相关性和显著性

变量	**概率**	**时间成本**	**商业面积**
相关性	1.000	-0.489	0.149
显著性	.	.000	.000
样本	5111	5111	5111

$$P_{ij} = \frac{S_j^{\mu} / T_{ij}^{\lambda}}{\sum_{j=1}^{n} S_j^{\mu} / T_{ij}^{\lambda}} \qquad (3.1)$$

由公式3.1可以看出，在哈夫模型中，μ和λ是模型调节指数，由于在商业方法中，这两个指数是由相关领域专家通过经验得到，为了对商圈吸引力的研究更加深入，我们邀请了相关领域专家，帮助我们给出两个调节指数值，作为主观指数值，同样，我们通过大样本相关分析，得到了相关系数，把相关系数作为一组调节指数，作为客观指数值。我们加入了一个约束条件，μ+λ=2，经过归一化处理和放大处理之后，我们得到了两组调节指数值。我们使用具有两种指数值得模型进行了商圈吸引力概率的计算，得到了商圈的辐射范围。两种调节指数值如下表：

表 3.2 哈夫模型调节指数

	主观调节指数		**客观调节指数**	
调节因子	λ(t)	μ(s)	λ(t)	μ(s)
原始数据	1.5	1.2	0.454	0.120
归一化	0.556	0.444	0.791	0.209

通过模型计算结果的可视化对比之后，我们可以清晰地看出，经过指数调节后的模型精度有了明显的提高，但是两种指数调节方法并没有很明显的优劣性，经过讨论后，我们认为这是由于哈夫模型仅仅使用商业面积和距离来进行计算的原因，而实际中，魅力和阻力的确定更加复杂。为了能够得到更准确地吸引力值，我们使用机器学习的方式对数据进行了训练，得到了一组影响因子的值，可能影响因子与训练结果如表3.3所示：

表 3.3 机器学习所得到的影响因子（12 个因素中最高的五个）

	时间成本	商业面积	商品档次	商场数目	商圈知名度
影响因子（%）	57.64570	36.85034	33.19354	28.09620	26.56424

使用机器学习方法得到的影响因子都没有很高，我们讨论后认为导致这个原因的可能因素是因为商圈的吸引力影响很复杂，同时主观情绪占有一定比重。

商圈是具有吸引力的，我们模型设计的基础同样是万有引力定律。面积越大，商品种类越多的商圈自然而然的吸引更多的消费者，但是在本文的研究中，商圈选取的都是大城市的核心商圈，商圈属性的差异比较小，在我们的相关性分析中，也能得到同样的结论：商业面积，商圈等级等因素对吸引力结果的影响因子都没有很高。但是一些因素，例如时间成本，换线次数等，对顾客选择商圈产生的影响要高得多。

在经过多次验证和分析之后，我们提出了一个适用于大城市大型商圈的商圈吸引力模型：

$$Attraction = Grade^{\alpha} * Mall^{\beta} * Area^{\gamma} * Reputation^{\delta} \qquad (3.2)$$

$$Attractive = \frac{Grade^{\alpha} * Mall^{\beta} * Area^{\gamma} * Reputation^{\delta}}{Time^{\mu}} \qquad (3.3)$$

其中，α，β，γ和δ是基于数据分析和挖掘所得出的调节指数。我们可以得到一个能够一定程度上表示商圈魅力的$Attraction$和一个能够表示商圈对某地吸引程度的$Attractive$。

我们认为影响顾客选择商圈的阻力因素主要为时间成本，尽管其受到一定的主观情绪影响（地铁线路图中的距离等因素），但是这是影响顾客选择商圈

的一个很重要因素。其次，商圈商品档次，商圈内商场数，商圈内商业面积，商圈知名度为主要魅力因素。

而某地顾客到各大商圈的概率值则为：

$$Probability = \frac{Attractive}{\sum_{i=1}^{n} Attractive_i} \quad (3.4)$$

因为在我们现阶段的工作中，无法验证调节因子的值是否适用于所有相似的大城市商圈。因此，我们通过多种分析方式所做的调节因子研究暂时只适用于上海市。

本文中，我们使用了两种方式对调节因子进行研究与确立，统计学相关性分析的方法与机器学习训练因子的方法。我们分别求出了不同因子的对应值，并对其进行对比研究。结果如图所示：

但是我们发现在某些情况下，计算出的值会有较大的误差（图3.1，图3.2），我们经过讨论认为，可能产生这些误差的因素主要有两个，一个是公交车对数据统计的影响，另外一个是换线次数对模型精度的购物阻力的影响。为了能够更加准确地预测出商圈的吸引力与辐射范围，我们对实验样本进行了分割，并对阻力值进行了优化。得到公式：

$$Attractive = \frac{Grade^{\alpha} \times Mall^{\beta} \times Area^{\gamma} \times Reputation^{\delta}}{Time^{\mu} \times \sqrt{Turn}} \quad (3.5)$$

我们使用原始模型进行计算，得出了吸引力值，然后计算其与实际值得平均误差，如图4.4所示，由于计算结果为0~1之间的概率值，我们使用绝对误差展示：

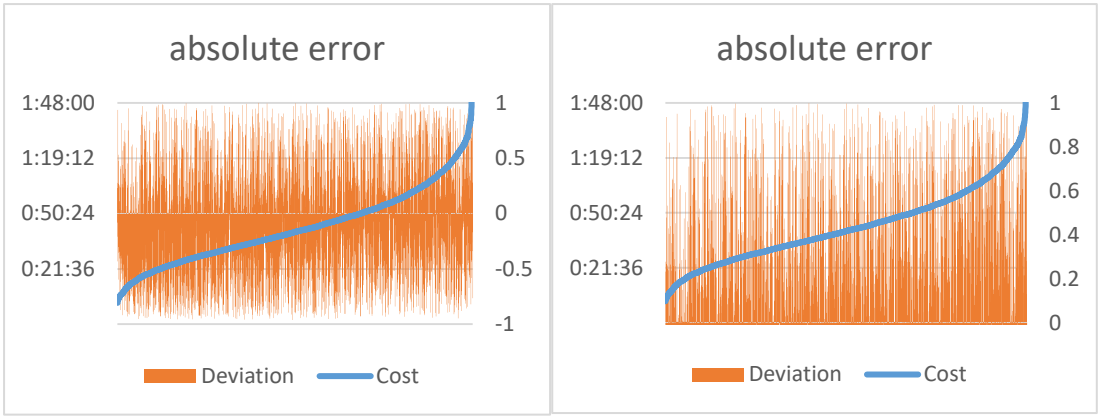

图3.1 原始模型计算结果绝对误差。

我们可以发现，在整体情况下，时间成本和误差正负与大小没有必然的联系。我们认为可能是由于原始模型的不适用以及误差过大导致：

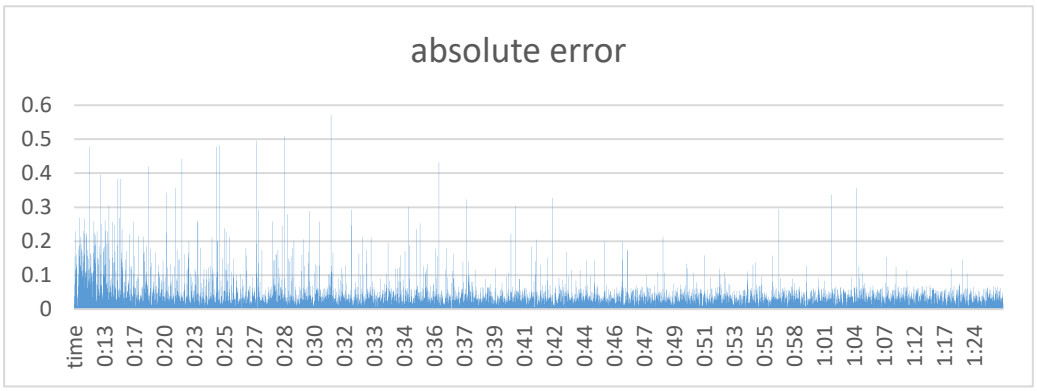

图3.2模型绝对误差。

在图中可以清晰的看出时间成本越低的位置产生的误差越大。同时，我们对具有不同调节指数的哈夫模型计算值进行比较，如图所示：

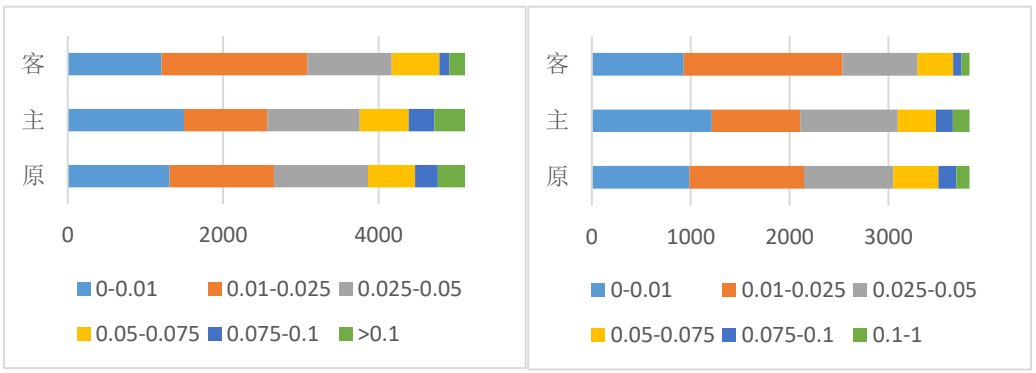

图3.3调节模型误差。

其中客观指数调节，很明显具有更好的准确度，因为其计算结果中误差值小于0.025的较其他两种要高很多，而主观指数调节得到的值能够得到更多的误差小于0.01的数据。

3.2 甲醇价格预测

价格预测是商业竞争的核心之一，特别是大宗化学产品（BCP）。影响BCP价格的国家政策或经济危机有很多因素，高分辨率因素很难分析和提取相关因素。与此同时，大数据时代的到来使得数据更加冗余，尽管它提供了更丰富的支持。

传统的预测算法通常集中在统计数据上，但BCP的价格是由许多主观因素决定的，结果不能准确地描绘未来。为了提高准确性，我们整合了传统的统计模型和机器学习方法，结合主观情绪分析。最后，我们改进预测模型，使其更适合于预测BCP的价格。

在与化学工业专家讨论之后，我们将重点放在分析和预测BCP的价格，并通过适当的视觉界面为决策者和企业经理显示结果。

为了解决这个问题，我们提出了以下分析任务：1）分析具有高影响因子的代表因素；2）分析网络信息，获取情感数据并量化专家的经验；3）选择BCP的预测方法并改进；以及4）结合情感因素和专家经验，进一步提高预测精度。

众所周知，嘈杂和不可靠的数据将影响预测的准确性。同时，冗余和不相关的数据可能导致预测结果不准确。因此，数据的预处理是很重要的。我们使用指数平滑（ES）方法用于数据校正，一些异常数据可能对最终结果产生巨大影响。我们使用Z-score方法尽量避免由数据偏差引起的预测误差。

$$\mathrm{Z(x)} = \frac{x - mean(x)}{StaDiv(x)} \qquad (3.6)$$

其中Z(x)是原始数据，mean(x)是所有数据的平均值，StaDiv(x)是标准差，然后我们得到平滑值。使用指数平滑方法优化数据的原因是数据中存在一些人为错误。

然后，我们检查不同模型的性能，并应用各种评估标准来评估预测结果。处理甲醇波动，复杂和不规则价格的GARCH模型比其他（如表3.4所示）更好，具有最佳拟合度和最小平均绝对百分比误差（MAPE），显然偏差率和均方误差（MSE）优于线性模型。

同时，通过跟踪2014年4月至2015年10月的甲醇价格，我们发现GARCH模型在短期预测中的误差最小，为4.7%，优于其他预测方法。组合GARCH与ARMA的长期预测具有最小预测误差，其第二预测误差达到7.80%，第三误差为10.1%，与其他预测方法相比，该方法具有最佳性能。

表3.4 不同预测方法的误差值

	方法	拟合度	MSE	MAPE	误差
单变量预测	ES	82%	222.0	4.62%	0.02%
	ARMA	87%	189.7	3.84%	0.01%
	SES	84%	213.5	5.19%	0.05%
多变量预测	BP	89%	146.3	3.08%	0.00%
	WNN	88%	163.3	3.49%	0.01%
	GARCH	93%	122.5	2.97%	0.00%

情绪数据是网络用户对化学产品价格趋势的判断，由于人员的专业程度，差异很大，难以分析情绪数据。传统预测模型中忽视的情绪因素有巨大的影响，我们发现情绪是通过分析研究预测最终价格的关键因素之一。在我们的分析和预测中，将网络情感因素整合到模型中。基于Web文本信息，数据清理，过滤，分类和文本分割等处理，文档分为短语列表。我们需要计算和分析具有价值感的文本是一个很大的挑战。由于情绪波动，情绪在不同时期不同，情感信息得到成功的爬取和分析，我们与传统的预测模型相结合，使预测精度大大提高。

3.3 本章小结

本章极少了关于商业数据的分析与研究，在大数据分析过程中，使用了数据清洗、相关性分析、多元线性方程以及数据挖掘的知识，这些不同的数据分析手段相结合能够很好的解决数据分析目标。

第4章 可视化系统设计

本章介绍了三个完善多的可视化系统，这三个系统都是基于 Web 端实现的，都具有很好的交互性与实用性。

4.1 甲醇价格预测可视化系统

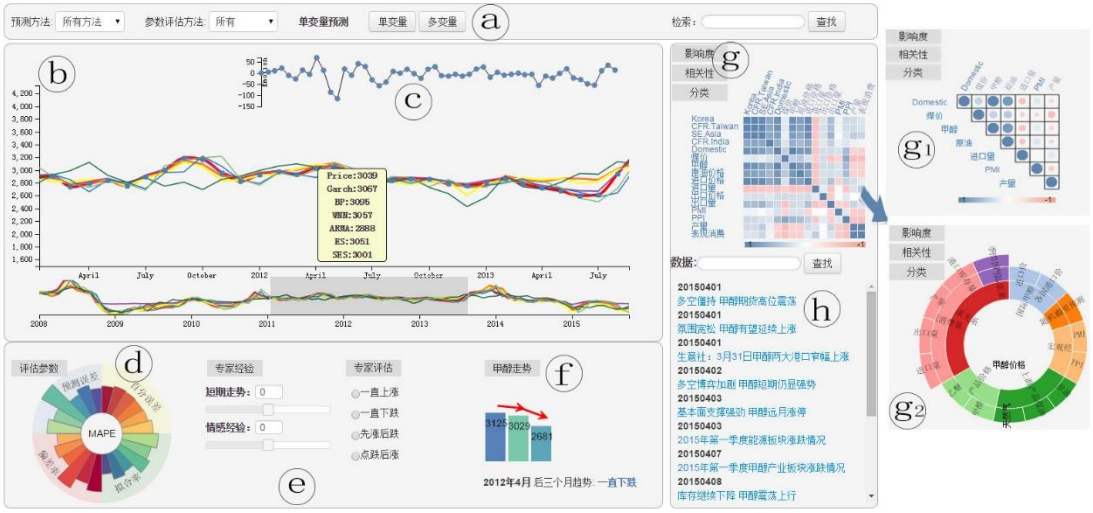

图4.1 甲醇价格预测可视化系统

4.1.1 相关性分析视图

图4.2是一个相关性分析图，展示了我们在数据分析方面的一些工作。相关性的分析并不是很困难，例如在数学中比较简单的正相关和负相关。但是随着数据维度的增加，相关性分析工作的难度也越来越大，同时数据间也不再具有简单的相关性。在多维化工产品的相关性分析中，产品的价格收到多种因素的影响，例如PMI、国家政策和港口环境等，每种因素又互相影响，最终形成了极其复杂的相关性。

在我们设计的可视系统中，用了一些交互完成了对相关性的展示，使用了一个相关性分析图展示了每两个元素间的想关心，同时用更加精准的相关性图展示了核心元素间的相关性。我们所有的工作都是在数据分析的基础上完成的，在数据分析中，相关性的分析研究占据了很重要的地位。通过我们的研究分析，对甲醇价格的影响因素以重要性区分，并通过对每个元素的分析找出了元素间的正负相关以及相关性的强弱，例如煤价与产量有着较大的负相关性，

这就表示了当煤价有所增加时，甲醇的产量有着明显的减少，同时由于产量与进口量呈正相关，因此煤价的增加还会导致进口量的减少。

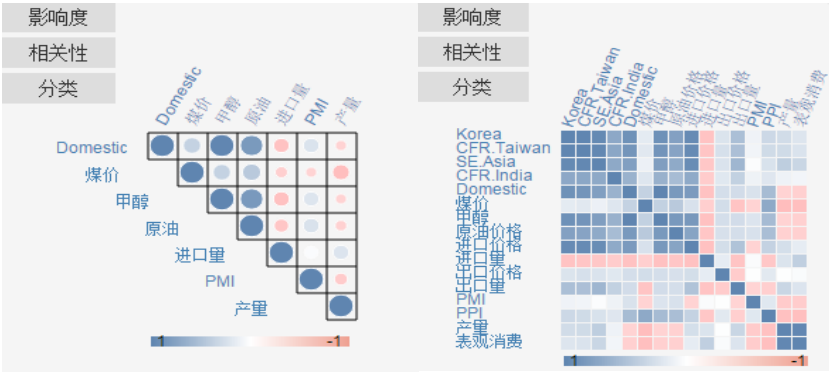

图4.2甲醇价格因素相关性分析视图

图 4.3 甲醇价格影响因素分类

我们通过d3.js设计相关性图如4.2所示，左图选择了代表性因素来进行与甲醇价格的相关性分析，从图中可以看出蓝色呈正相关、红色表示负相关，圆圈大小表示相关性强弱，正负相关性越强的因素圆圈大，相关性较弱的因素圆圈比较小。右图将化工产品价格因素的相互关系通过方块的方式来表示，其中蓝色到红色，表示相关性强到弱，通过右图我们可以明显发现，进口量与所有的因素都呈现较强的负相关，而进口量左边的所有因素相互之间都存在较强的正相关，从分析结果可知甲醇价格与各个港口的价格呈现出较强的相关性，专家可以在每个月月前通过分析每个港口的价格来对本月的甲醇价格进行一定程度的估计。同时当每个月的甲醇进口量增多时，对其他化工产品价格会造成一定

的影响，遇到这样的情况，应当适当减少进口、降低甲醇价格，加快甲醇的销量进而减少库存量。

4.1.2 情感分析图

情感分析图使我们的可视系统中比较独特的地方，情感数据以其主观性和不确定性一直被信息预测者所诟病，在这个时代，社会舆论所产生的情感对多维化工产品的价格影响很大。之前很多人做的预测由于没有正确认识到情感因素对预测结果的影响，因此产生了很大的误差，我们吸取了前人的经验，在使用多种模型对数据进行分析预测之后，加入了情感因素和专家经验的优化，使得预测结果更加准确。在这里，我们使用了标签云技术对情感数据进行了展示，通过对收集到的网络情感信息的提炼，把情感信息生动的展示给用户，把主观情感加入到客观预测中，使得预测结果更加精准。

情感数据的分析有着极大的困难，情感信息属于主观信息，在不同的角度，每个人对于同一个事情的看法也是不同的，这就造成了我们的处理困难。为了完成情感信息的融入，我们搜集了大量信息，然后逐步分析，在我们的可视系统中，在网评信息部分可以搜索到部分情感信息，其余大量的情感信息被我们整理之后在情感分析图部分通过标签云来显示，当用户对某一时间的社会情感信息感性兴趣的时候，可以很容易的得到在这个时间的社会情感信息。

我们通过网络文本的爬取，同时对文本进行清理、分词、统计等工作，通过使用构造的化工产品情感词典，获得每个月中出现频率最高的单词，在系统中我们通过echarts将标签云以动态的方式呈现出来。用户可以通过鼠标交互去获得在每一个用户对甲醇走势的情感倾向，从图4.4中可以看出，在2012年6月，甲醇价格呈现下跌的状态，同时网评信息中，出现最多的词组有：下跌、跌、下滑、低迷、弱势等一些词语，而在2013年9月时当甲醇价格出现上涨时，网评中多出现：稳定、旺季、增加、恢复等词偏多，并且热度较大。由此可见，网络评论与甲醇价格的走势存在非常大的相关性。用户通过可视化的方

法，可以迅速获取在近几年中，在每个月网评信息对甲醇价格走势的判断。通过可视化方式来展示不同月份中呈现的规律信息。

图 4.4 网络情感数据搜索热度

4.1.3 预测误差分析视图

预测必然会产生误差，不同的预测方式在不同的环境下产生的误差也是天差地别。在我们的系统中，由于使用了多种模型来进行预测，所以误差的可视分析显得尤为重要，因为对多维化工产品价格的预测最重要的一点就是要精准。

在我们的可视系统中，使用了一个板块进行了对拟合率、偏差率等误差评估参数的展示，在这个部分，六种模型的误差分析详细具体的展现给用户，使用户可以很清晰的了解到各种预测模型对当前所预测的多维化工产品的优劣度。同时我们还给出了预测精度评价指数（MAPE），使得我们的预测更加有说服力。

本文中使用改进的饼状图，通过StartAngle和EndAngle来固定内圆和外圆的终始位置来显示预测误差。如图4.5所示，通过饼状图的形式，通过设置内半径和外半径的大小，将四种不同的预测误差显示出来，通过鼠标交互可以获取预测误差的大小，同时四种预测误差的鲜明对比，让使用者可以迅速了解哪种预测方法具有最好的拟合率和最小的偏差率。

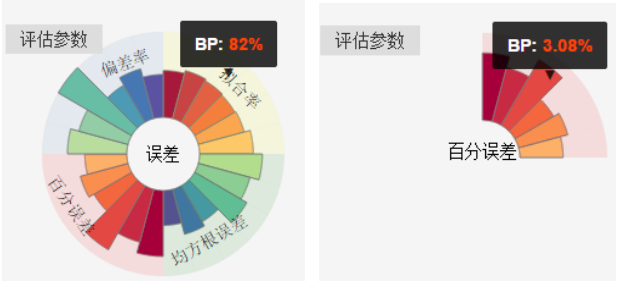

图 4.5 预测误差

4.1.4 统计分析视图

图4.6详细显示甲醇价格的十八个月的历史信息，以及六个传统算法的预测结果。用户可以通过浏览该图来识别传统预测算法的优缺点，提高对整个系统的认知度。

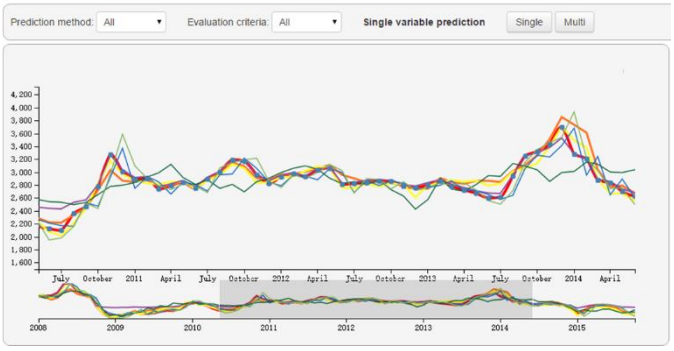

图4.6 历史信息的时序图

4.2 交通流量可视化系统

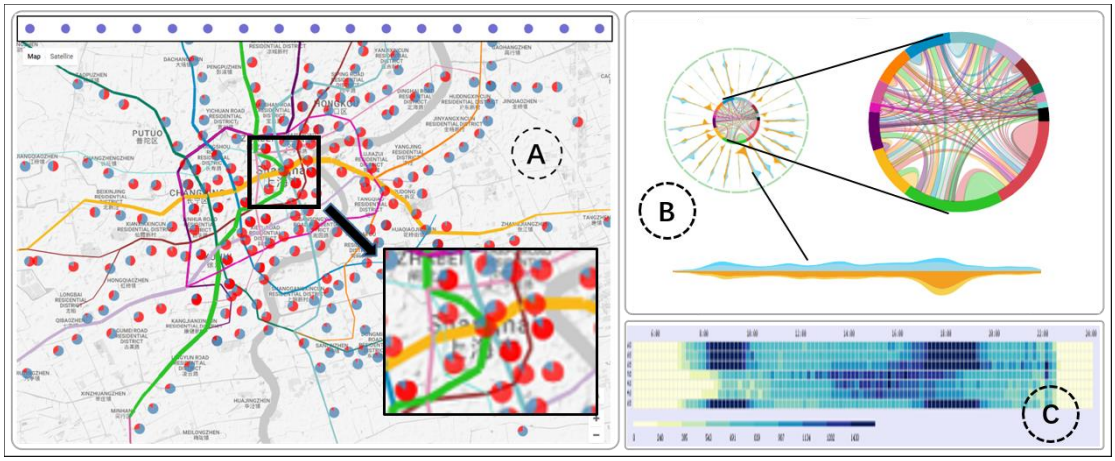

图4.6 交通流量可视化系统

4.2.1 流量快照可视化模块

本文设计地铁流量快照可视化模块。如图4.7所示，地铁流量快照可视化模块展示不同时间段的地铁交通流量变化，用户通过选择位于图4.7上部的蓝色圆圈展示不同时间段的流量信息。每个圆代表一个长度为半小时的时间片。具体的流量信息展示在地图中，同一条地铁线路的两个相邻站点被带颜色的线连接，不同的颜色表示不同的地铁线路，线的粗细表途经两站的流量大小。如图3所示，与其它线路相比，绿色的上海地铁2号线负载了较多流量。此外，散布在地图上的饼图展示了上班族群体在某一时刻的进站和出站数量，饼图中红色部分表示出站的数量，蓝色部分表示进站的数量。通过图中饼图分布情况，可以判断出大规模的居住地点和工作地点。如果一个区域进站的数量较大，即红色部分较大的饼图较多，表明该区域有较多的居民区。如果一个区域出站的数量较大，即蓝色部分较大的饼图较多，表明该区域可能是分布着大量公司的科技园区。

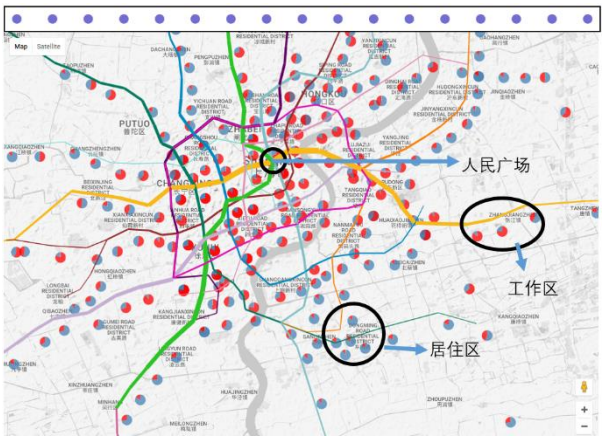

图4.7 流量快照可视化视图

4.2.2 地铁站人流流动可视化模块

我们设计了站点流量关系可视化模块，如图4.8所示，该模块展示了不同站点之间的流量。当用户在流量快照视图中选择一个站点，该站点到其他各个连接的站点的流量细节信息会展示在该模块中。图4.8a展示了与选定地铁站连接的流量较大的所有站点。其中的每一个弧段代表一个站点。图4.8b展示了站点之间的交通流量，并能够区别出不同上班族群体的地铁交通流量。如图4.9所示，本文运用流图展示两个站点之间不同时刻的流量，其中上半部分表示进站

流量，下半部分表示出站流量，外层的浅色部分代表非常规的上班族，内层的深色部分代表常规的上班族。该视图能够不同站点之间的交通流量信息，并直观地展示不同上班族群体之间的出行特征。图4.8c中展示的弦图表示选定的站点到其他地铁线路的流量关系，其中颜色与上海地铁线路的官方颜色一致，该图能够呈现出整体的连接趋势，方便用户分析不同群体的移动行为。

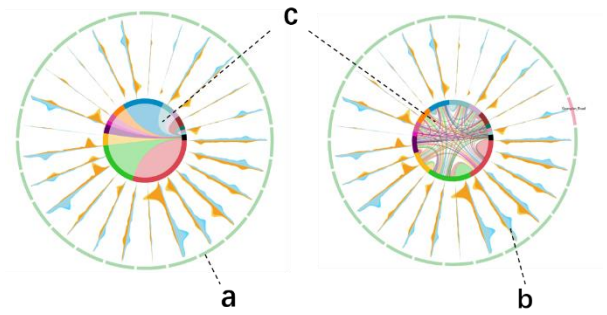

图4.8 地铁站流量关系可视化模块

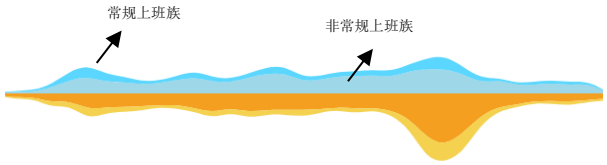

图4.9 显示地铁站流量的流图

4.2.3 流量时序可视化模块

图4.10展示了为期一周的流量变化，在该图中，我们使用较大的正方形映射流量信息，时间间隔为15分钟，横轴为时间轴，从早晨六点到晚上十一点。从该图中可以看出前三天和最后一天早晚高峰流量聚集明显，中间三天流量分布与其它四天不同，是因为这三天是法定假日，使得整天的交通流量比较分散。 为了探究流量变化的细节，我们采用了更细的力度来展示流量信息，每一条线表示每一分钟，用颜色来映射流量大小，颜色越深表示流量越大。如图4.11所示，流量时序可视化模块展示了每个地铁站为期一个月的进站和出站流量信息。这里用颜色的深浅映射一个地铁站在某个时刻经过的流量，颜色较深表示此时的流量较大，颜色较浅表示此时的流量较小。每一个长条代表每天从早上六点到晚上十一点的流量分布。通过该视图，用户能够直观地了解该地铁

站一个月的流量分布情况。在图4.11中，从上午八点到九点和下午五点到七点有明显的地铁交通流量聚集现象。这种现象只发生在工作日。在周末或者假日，整体的出行比较分散，并且大部分地铁交通流量分布在中午十二点以后。在该视图中，我们还可以发现一些有趣的现象，比如从下午五点到七点，在流量聚集区中会有一些间隔。这可能是因为不同的下班时间引起的。

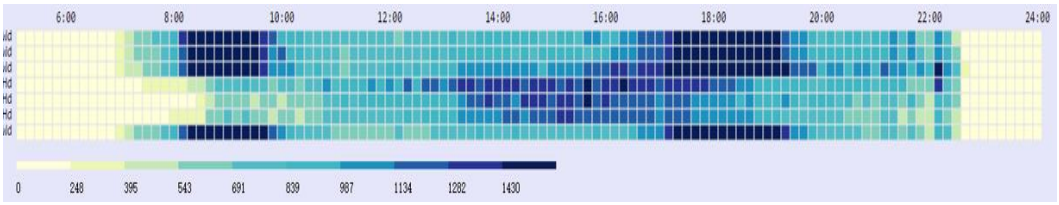

图4.10 地铁流量时序(一周)

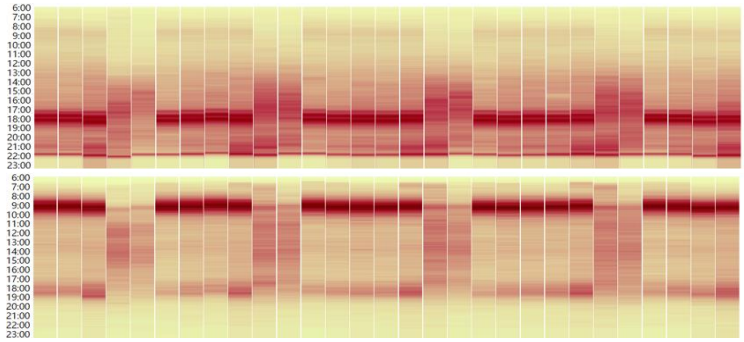

图4.11 地铁流量时序(一个月)

4.3 零售商店选址推荐可视化系统

4.3.1 商业影响力视图

地图视图在多维地理信息可视化研究中是常见的，在我们的研究中，我们使用地铁图来显示商业区和行政区划的影响范围。

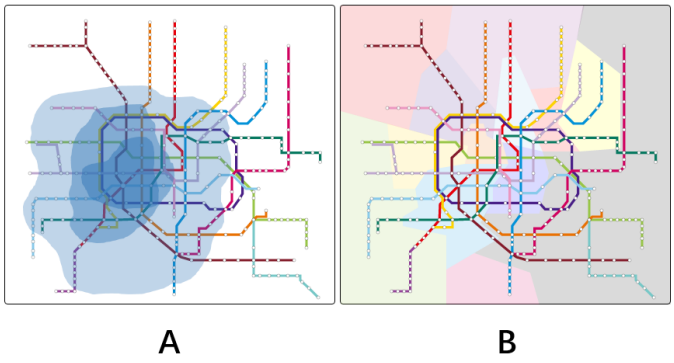

图 4.12 商业影响力视图

在这种观点下，我们使用不同程度的颜色代表影响范围（图4.12A），其中最深的颜色表示这里的居民有超过30％去商业区购物，较浅的是20％最浅的是15％。我们通过聚类所有站（集群中心是中心城市的商业区）得到这些数据。我们发现，居民的消费偏好有明确的规律：居民宁愿去商场购物时间较少，而商场的面积和知名度也有很大的影响。例如，徐家汇是上海最大的商业区之一，经过分析比较，发现居民还喜欢来这里，甚至有更高的时间成本。

4.3.2 统计分析视图

这部分是一个多层圆形图，用于通过详细的统计信息显示固有的模式。

在本节中，我们将提供统计数据的主要介绍。图4.13A是我们系统中显示的图

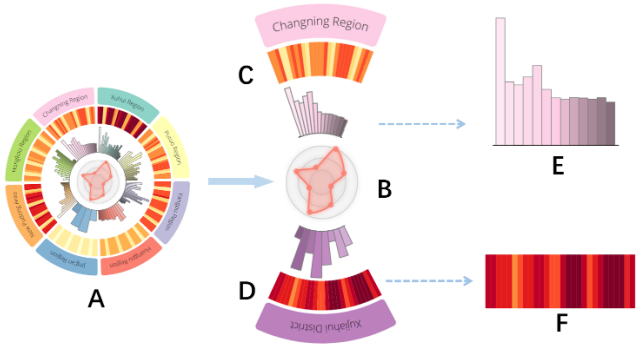

图 4.13 统计分析视图

在这种观点下，比萨饼图的每个部分代表商业区或行政区域。一个派别提供各种信息，例如一个月内的客户流量（图4.13F）和该商业区每个商场在区域视图中的销售量，以及一个月内的旅行者人数和每个车站的购物者。使用过境

通行证数据，我们假设过境通行证持有人经常在每个工作日的6：00-9：00和5:00-7:00之间通过，是在业务领域工作的员工，我们通过排除雇员。我们通过公司的销售数据和哈夫的模型获得销售量。此外，旅客在这方面表示了大概居民人数。

此外，我们还通过专家经验预测当前地区的业务前景，其范围在1到10（图4.13B）。

4.3.3 选址推荐视图

我们的系统可以在显示市场繁荣的热图上显示多达10个推荐位置（图4.14）。

我们还提供预期成本和利润的输入框，通过用户交互，系统将显示十个最符合预期的位置。此外，我们还提供了一个缩放功能，以更好地展示我们的研究。

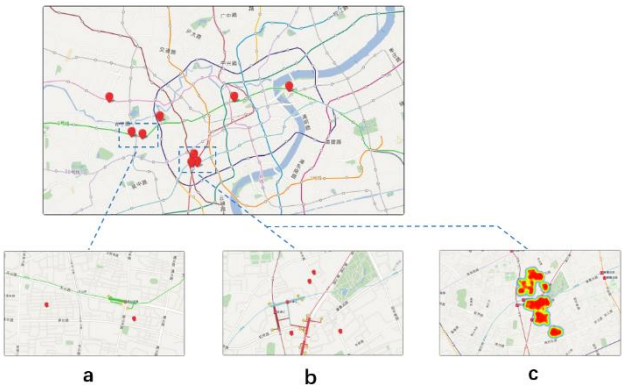

图 4.14 选址推荐视图

4.3.4 视觉比较视图

我们还设计了视觉比较视图（图4.15），它比较了不同位置的优点。在图4.15A中，我们为用户提供了每个解决方案的详细比较，并在图4.15B中进行了总体比较，最后，用户可以对这些解决方案进行排序（图4.15C）。

我们从二十多个因素中选出八个细节。首先，我们得到二十个因素，分为四个类别：消费因素（商业区偏好，过境时间成本），商场因素（劳动力成

本，租金，广告费用或无障碍），市场因素（行业竞争或市场饱和度）和社会因素（经济基础，经济政策或时尚潮流）。

在这些中，市场饱和因素，行业政策价值相同，主要商品，销售风格对我们的研究影响不大。因此，经过多次研究和专家的讨论，我们选择了十个因素进一步研究。

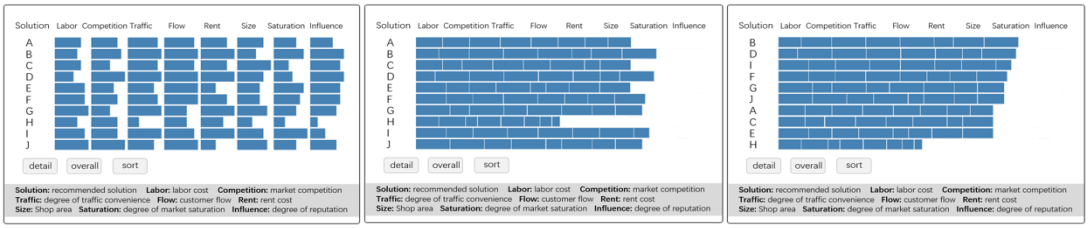

图 4.15 视觉比较视图

4.4 本章小结

本章介绍了三个可视化系统，每个可视化系统都是以科研或者实际应用为导向，具有很好的现实意义。同时可视化系统为数据分析提供了很好的展示手段，并能通过可视分析的方式发现一些隐藏在数据深处的规律。

第5章 总结

本文介绍了数据可视化与可视分析相关技术，并应用这些技术进行了数据分析、数据挖掘、可视分析和可视化系统设计。信息可视化是一个跨学科领域，旨在研究大规模非数值型信息资源的视觉呈现。通过利用图形图像方面的技术与方法，帮助人们理解和分析数据。信息可视化致力于创建那些以直观方式传达抽象信息的手段和方法。可视化的表达形式与交互技术则是利用人类眼睛通往心灵深处的广阔带宽优势，使得用户能够目睹、探索以至立即理解大量的信息。

本文所介绍的相关工作都是以科研或是实际需求而展开的，这也很符合当今社会对数据可视化的要求，就是要有明确的应用目的与需求。只有以实际出发，以需求为导向，才能更好地推动数据分析与可视化领域的发展。

参考文献

[1] Zou H, Zhao G, Liu H, et al. Bulk Chemical Production: Chemo- and Bio-integrated Strategies[M]// Sustainable Production of Bulk Chemicals. 2016.

[2] Trevor Boyns, Mark Matthews, John Richard Edwards. The development of costing in the British chemical industry, c.1870-c.1940[J]. Accounting & Business Research, 2004, 34(1):3-24.

[3] Goldstein E, Cobey S, Takahashi S, et al. Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method. [J]. Plos Medicine, 2011, 8(7):483-484.

[4] Yin X, Chen W, To A, et al. Statistical volume element method for predicting microstructure–constitutive property relations[J]. Computer Methods in Applied Mechanics & Engineering, 2008, 197(43–44):3516-3529.

[5] Anil Kumar KM, Anil B, Anand CU, Aniruddha S, Rajath Kumar U.Machine Learning Approach to Predict Real Estate Prices. Discovery, 2015, 44(205), 173-178

[6] Chen L, Ma J, Liu Y. Predicting the target coordinate(X) of the PVP using a mathematical model[J]. Chinese Journal of Stereotactic & Functional Neurosurgery, 2005.

[7] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140.

[8] Savojardo C, Fariselli P, Casadio R. BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes[J]. Bioinformatics, 2013, 29(4):504-5.

[9] Perlesribes J F, Ramónrodríguez A B, Morenoizquierdo L, et al. Research note: Economic crises and market performance – a machine learning approach[J]. Tourism Economics, 2016.

[10] Pompe P P M, Feelders A J. Using Machine Learning, Neural Networks and Statistics to Predict Corporate Bankruptcy: A Comparative Study[M]. Springer US, 1996.

[11] Dennis C, Marsland D, Cockett T. Central place practice: shopping center attractiveness measures, hinterland boundaries and the UK retail hierarchy ☆ [J]. Journal of Retailing & Consumer Services, 2002, 9(4):185-199.

[12] Wahlberg O. Small town center attractiveness: evidence from Sweden [J]. International Journal of Retail & Distribution Management, 2016, 44(4):465-488.

[13] Yao Lizhen, Yue Yang. Factor analysis of shopping center attractiveness based on principal component logistic model [J]. Journal of Geo-Information Science, 2016.

[14] Singla V, Rai H. Investigating the effects of retail agglomeration choice behavior on store attractiveness [J]. Journal of Marketing Analytics, 2016:1-17.

[15] Reilly T. Reilly's rules for value-added selling [J]. Official Board Markets, 2006.

[16] Honda M, Ushizawa K. Quantitative geographic study of buying behavior: Huff's model and its parameter estimation [J]. Sanno College Bulletin Department of Management & Informatics, 1982, 2:81-104.

[17] Anne Marie Doherty. Factors influencing international retailers' market entry mode strategy: qualitative evidence from the UK fashion sector [J]. Journal of Marketing Management, 2000, 16(1):223-245.

[18] Doherty A M. The internationalization of retailing: factors influencing the choice of franchising as a market entry strategy [J]. International Journal of Service Industry Management, 2007, 18(2):184-205.

[19] Lee H S, Griffith D A. Transferring corporate brand image to local markets: governance decisions for market entry and global branding strategy [J]. Advances in International Marketing, 2012, 23:39-65.

[20] Khakzar K, Blum R, Kohlhammer J, et al. Interactive product visualization for an In-Store sales support system for the clothing retail [C]. Human Interface and the Management of Information. Methods, Techniques and TOOLS in Information

Design, Symposium on Human Interface 2007, Held As. 2007:307-316.

[21] Yaeli A, Bak P, Feigenblat G. Understanding customer behavior using indoor location analysis and visualization[J]. Ibm Journal of Research & Development, 2014, 58(5/6):3:1-3:12.

[22] Nesbitt K V, Barrass S, Stephen Barrass@csiro. Evaluation of a multimodal sinification and visualization of depth of market stock Data [C]// R. Nakatsu and H. Kawahara. 2002:2--5.

[23] Macer T. Data visualization in market research [J]. Research World, 2014, 2014(47):10 - 19.

[24] Pryke A, Mostaghim S, Nazemi A. Heatmap visualization of population based multi objective algorithms[C]// International Conference on Evolutionary Multi-Criterion Optimization. Springer-Verlag, 2007:361-375.

[25] Hashimoto Y, Matsushita R. Heat map scope technique for stacked Time-series data visualization[C]// International Conference on Information Visualization. IEEE, 2012:270-273.

[26] Shibahara N. Realtime Panoramic Mobile Streaming Application for Assisting Visualization to Remote Person [J]. 2015.

[27] C. Shi, Y. Wu, S. Liu, H. Zhou, and H. Qu. LoyalTracker: visualizing loyalty dynamics in search engines. IEEE TVCG, 20(12):1733 - 1742, 2014.

[28] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: interactive visualization of hotel customer feedback. IEEE TVCG, 16(6):1109 - 1118, 2010.

[29] Gleicher M, Albers D, Walker R, et al. Visual comparison for information visualization [J]. Information Visualization, 2011, 10(4):289-309.

[30] Borowski K, Soh J, Sensen C W. Visual comparison of multiple gene expression datasets in a genomic context [J]. Journal of Integrative Bioinformatics, 2016, 5(2):94-103.

[31] Kim M C, Zhu Y, Chen C. How are they different? A quantitative domain comparison of information visualization and data visualization (2000–2014) [J]. Scientometrics, 2016, 107(1):123-165.

[32] Liu D, Di W, Li Y, et al. SmartAdP: Visual analytics of large-scale taxi trajectories for selecting billboard locations [J]. IEEE Transactions on Visualization & Computer Graphics, 2016:1-1.

[33] Collins C, Carpendale S. VisLink: revealing relationships amongst Visualizations [J]. IEEE Transactions on Visualization & Computer Graphics, 2007, 13(6):1192.

[34] Koleva, Vyara. Interactive visualization of the building of university of economics – varna via 3D modeling [J]. Digital Presentation & Preservation of Cultural & Scientific Heritage, 2013, 263-266(III):8.

[35] Kuang Y, Liang Z, Lei Y. Scientometrics research on spatial economics based on information visualization [J]. Science & Technology Management Research, 2012.

[36] Dennis C, Marsland D, Cockett T. Central place practice: shopping center attractiveness measures, hinterland boundaries and the UK retail hierarchy ☆ [J]. Journal of Retailing & Consumer Services, 2002, 9(4):185-199.

[37] Zheng D, Zhang J. A Comparative Research on High Speed Rail Business District Planning:Example of Shanghai Hongqiao and Jiaxing South HSR Station Area[J]. Planners, 2011.

[38] Kotthaus S, Grimmond S. Energy exchanges in a Central Business District - Interpretation of Eddy Covariance and radiation flux measurements (London UK)[C]// AGU Fall Meeting. AGU Fall Meeting Abstracts, 2013.

[39] Lin X R, Pan H X. The effects of the integration of metro station and mega-multi-mall on consumers' activities: a case study of Shanghai[J]. Transportation Research Procedia, 2017, 25:2578-2586.

[40] Weimer J A. The Role of Marketing in Business Attraction for Neighborhood

Business Districts: Case Study Research and Applied Findings[J]. 2011.

[41] Goel R, Tiwari G. Access-egress and other travel characteristics of metro users in Delhi and its satellite cities[J]. Iatss Research, 2016, 39(2):164-172.

[42] Reilly W J. Methods for the Study of Retail Relationships[M]// Methods for the study of retail relationships. University of Texas, 1929.

[43] Reilly W J. The Law of Retail Gravitation[J]. American Journal of Sociology, 1931.

[44] Huff D L. A Probabilistic Analysis of Shopping Center Trade Areas[J]. Land Economics, 1963, 39(1):81-90.

[45] Huff D L. Defining and Estimating a Trade Area[J]. Journal of Marketing, 1964, 28(3):34.

[46] Andrienko G, Andrienko N, Chen W, et al. Visual Analytics of Mobility and Transportation: State of the Art and Further Research Directions[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, PP(99):1-18.

[47] Xiong W, Zhang M. Distribution Research of Survival Road for Local Retail Businesses in Relatively Backward Areas - A Case Study on Luoyang Dazhang[J]. Advanced Materials Research, 2014, 989-994:5090-5093.

[48] Rüscher M. Business Improvement Districts - An Approach for Retail-Area Revitalization in American Downtowns[J]. Neurus Papers, 2005.

本篇技术报告由学生 李柯林 独立完成

导师评价意见：

导师签字：_____