信息见解引导的可视化推荐

陈滨, 张钰涵, 胡瑞珍*

(深圳大学可视计算研究中心 深圳 518061)
(ruizhen.hu@szu.edu.cn)

摘 要: 对高维数据的探索和分析需要在数据中提取有价值的信息见解并生成合适的可视化图表. 为降低生成有效图表的门槛, 以信息见解为中心自动生成可视化图表, 提出一个可视化推荐框架, 包含数据片段推荐和视觉编码推荐2部分. 在数据片段推荐中, 相比现有方法使用单一指标, 所提框架结合多种分析方法和机器学习算法进行信息见解的挖掘和数据片段的排序, 为用户推荐信息见解丰富的数据片段; 在视觉编码推荐中, 该框架结合规则和机器学习方法, 基于信息见解的特征建立视觉编码映射, 并对图表中的关键信息进行高亮以增强用户的理解. 在图表类型预测任务中, 使用用户绘制的真实可视化图表数据进行实验的结果表明, 所提框架比现有方法具有更高的预测准确率; 用户调查的结果则证明, 该框架能帮助用户更好、更快地了解一个陌生的数据表格中蕴含的关键信息.

关键词: 信息可视化; 可视化推荐; 信息见解; 可视分析
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023.19312

Insight-Guided Visualization Recommendation

Chen Bin, Zhang Yuhan, and Hu Ruizhen*

(*Visual Computing Research Center, Shenzhen University, Shenzhen* 518061)

Abstract: Exploration and analysis of high-dimensional datasets require extracting valuable insights and generating proper charts from data. To lower the barriers of creating effective visualizations, a new insight-centric visualization recommendation framework is presented that recommends both what to visualize (data query recommendations) and how to visualize it (visual encoding recommendations). For the former, instead of using single indicators, multiple analysis methods and learning-based algorithms are utilized to automatically rank data queries based on the insights and the most informative ones are provided to the users. For the later, rule-based and learning-based methods are combined to generate visual mappings from data queries to charts and further highlight the key information in the generated charts to enhance their readability. It is demonstrated that proposed method achieves higher accuracy than existing methods for the chart type prediction task with real-world charts dataset. A user study is also conducted to show that proposed method can help users quickly explore new data tables and find important insights.

Key words: information visualization; visualization recommendation; insights; visual analytics

随着现代社会中的团体或个人产生的数据越来越多, 对数据进行可视化分析以指导人们生产生活的需求变得越来越强. 传统的探索性数据分析工具需要数据分析师手动筛选数据、选择数据的

收稿日期: 2021-08-29; 修回日期: 2022-03-16. **基金项目**: 国家自然科学基金(61872250); 广东省自然科学基金(2021B1515020085).
陈滨(1996—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为信息可视化、计算机图形学; **张钰涵**(2000—), 男, 在校学生, 主要研究方向为信息可视化; **胡瑞珍**(1988—), 女, 博士, 副教授, 硕士生导师, CCF 高级会员, 论文通信作者, 主要研究方向为计算机图形学.

组织方式以及可视化方式, 这就要求数据分析师对可视化工具以及数据本身都有较深入的了解, 才能构建出有意义的可视化图表, 给数据分析带来了较高的门槛. 近年来, 很多可视化推荐系统尝试简化这一过程, 自动为用户推荐可视化图表. 可视化推荐可拆分为数据片段推荐和视觉编码推荐 2 个子任务, 前者决定要对什么数据进行可视化, 后者决定如何可视化[1]. 现有的可视化推荐系统大多仅集中于解决这 2 个任务之一. 对于前者, 现有方法[2-3]大多通过定义多个信息见解类型, 每类信息见解使用一个特定的衡量指标对提取的数据片段进行筛选和对比, 但单一的衡量指标难以适用于不同的数据. 而对于后者, 现有的方法或基于人工定义的规则[4-5]难以推广至更多不同情景, 或使用机器学习[6-7]直接将数据映射得到视觉编码, 无法解释为什么选择了特定的可视化方式, 也就无法向用户直接传递信息, 使用户信服.

本文提出一个可视化推荐框架, 以信息见解为中心, 同时进行数据片段推荐和视觉编码推荐. 相比现有方法使用单一的衡量指标, 本文对每类信息见解都使用多种基于统计分析或机器学习的方法进行挖掘并综合排序, 更全面地提取每个信息见解的特征; 基于不同类型信息见解的特征, 使用结合规则和机器学习的方法为推荐的数据片段建立视觉编码映射, 通过分析不同见解特征对模型的贡献程度使该映射过程具备可解释性; 同时, 为生成的图表高亮标注出其蕴含的信息见解, 使推荐的可视化结果能直接向用户传递有价值的信息; 为方便解读, 设计了一个可交互的界面, 使用户可以选择要查看的信息见解类型和排序偏好, 筛选感兴趣的数据维度, 协助其快速熟悉一个陌生数据表格并渐进式地挖掘其中有价值的信息.

1 相关工作

1.1 数据片段推荐

对一个输入的完整数据表格, 数据片段推荐系统需要从中提取出有分析或解读意义的数据片段. 通常需要先对表格的数据变换和维度组合进行枚举, 再对枚举得到的候选数据片段进行分析, 找出其中有价值的部分进行推荐或排序.

早期的方法通过为特定的分析任务设计指标判断数据片段是否有分析意义. 例如, Profiler[8]会在所有候选数据片段中找出相对异常的片段; $S_{EE}DB$[9]则会让用户指定一个参考数据片段, 将所有与之差异较大的数据片段进行推荐.

近几年, 更多工作使用信息见解作为推荐数据片段的依据. Foresight[2]中定义了离群值、线性相关等 10 种信息见解类型, 每个类型使用一个特定的衡量指标, 用于判断数据片段是否属于该类型, 并进行评分; 之后, QuickInsights[3]定义了标准化的信息见解模型, 以及 12 种信息见解类型; DataShot[10]和 Calliope[11]分别对 QuickInsights 提出的信息见解模型进行修改和完善, 并使用生成的图表合成数据海报或数据故事流; SpotLight[12]提出对每种信息见解类型使用多种算法进行更全面的分析, 为平衡不同算法, 其对相同算法的结果进行归一化作为得分, 从而对各信息见解类型及每个类型内的数据片段进行排序.

这些基于信息见解的方法将所有推荐的数据片段按见解类型分组, 然而有些数据片段可能同时蕴含不同类型的信息见解, 使它们会在不同组中重复出现. 相比之下, 本文将数据片段按其涉及的维度数量和类型进行分层, 更加符合认知习惯, 也更便于在不同层次内使用多种挖掘算法并进行对比分析.

1.2 视觉编码推荐

要将一个数据片段映射为最终的可视化图表, 视觉编码推荐中最重要的是确定图表类型, 其次还可以在视觉通道等方面进行调整. 现有方法可以分为 2 类: 基于规则的方法和基于机器学习的方法.

基于规则的方法通常通过一系列条件判断语句来对要使用的图表类型进行推荐. 文献[4]通过衡量表现力和视觉感知效果从枚举的图表中筛选结果. 文献[13-14]等工作实现了更丰富的推荐规则, 支持对更多的数据属性、视觉编码、统计信息等进行分析推荐. 但人工设计的规则难以涵盖多种不同的需求, 往往需要一定的用户输入以决定最终的输出结果.

基于机器学习的方法则有很多不同的尝试. 例如, VizML[6]将视觉编码推荐问题转换为多个可视化设计选择的分类问题, 通过训练神经网络 (neural network, NN), 为输入的数据片段预测图表类型、视觉通道等设计选择, 得到可视化图表; 但其定义的 5 个设计选择往往不足以决定数据片段的具体绘制方式, 有时甚至会相互冲突. Data2Vis[7]则是将问题转换为一个自然语言翻译问题, 训练端到端的模型直接输出 Vega-lite[15]可视化语句, 也因此更多关注于输出结果的语法正确

性，而非信息的表达效果．不同于上述方法，本文使用规则和机器学习相结合的方式，基于数据片段所表达的信息见解为其推荐可视化映射方式．

1.3　综合可视化推荐

近年来，有部分工作尝试同时对数据片段和视觉编码进行推荐，其中，有的需要用户的输入作为指导．VizDeck[16]通过几项统计指标对枚举的可视化图表进行初步排序，再让用户进行搜索和挑选；Voyager[17-18]提出一个基于多项统计指标和约束条件对图表进行分析的框架，在为用户选择的数据片段推荐视觉编码的同时，额外推荐部分包含没被选择的数据维度的图表；Draco[19]则在 Voyager 的基础上使用机器学习方法为推荐的约束条件学习权重，从而对所推荐的图表进行排序．还有一些工作不需要用户输入．DeepEye [20]使用基于维度类型、数据长度等指标的简单规则决定视觉编码，并收集用户的选择偏好训练模型对所有图表进行排序；Table2Charts[21]则使用强化学习方法为每种特定的图表类型从数据表格中挑选维度组合和变换方式．这些方法并没有对推荐的数据片段进行信息的挖掘和分析．

2　本文框架概述

信息可视化的最重要目的是为用户提供信息见解[22]．信息见解是用户在对数据进行分析任务的过程中得到的新知识，或者说，见解是在分析任务中展现出或意外、或复杂、或深度、或与任务息息相关的特征的数据片段[23]．本文基于挖掘算法定义信息见解：信息见解是数据片段在挖掘算法下体现出的值得关注的特征，如有离群值、服从正态分布等．

本文提出一个基于信息见解的综合可视化推荐框架，在不需要用户输入的情况下同时进行数据片段推荐和视觉编码推荐．对于一个给定的数据表格，自动寻找其中的信息见解，并为每个信息见解生成合适的图表推荐给用户，帮助用户快速、准确、全面地了解一个陌生数据表格中蕴含的重要信息．图 1 所示为本文设计的可交互界面．图 2 所示为本文进行可视化推荐的基本流程，分为数据片段枚举、信息见解挖掘及视觉编码映射 3 个阶段．

对于一个输入的数据表格，第 1 步是对其进行划分，得到多个数据片段并分为不同层次，如图 2b 所示．首先判断数据表格各维度的数据类型，并据此选择数据变换方式；然后通过组合不同的维度和变换方式枚举数据片段．与现有方法根据见解类型进行分组不同，本文根据涉及维度的数据类型将所有数据片段划分为多个层次，使相同的数据片段不会在不同组别中重复出现，方便为不同层次的数据片段选择信息见解挖掘算法．

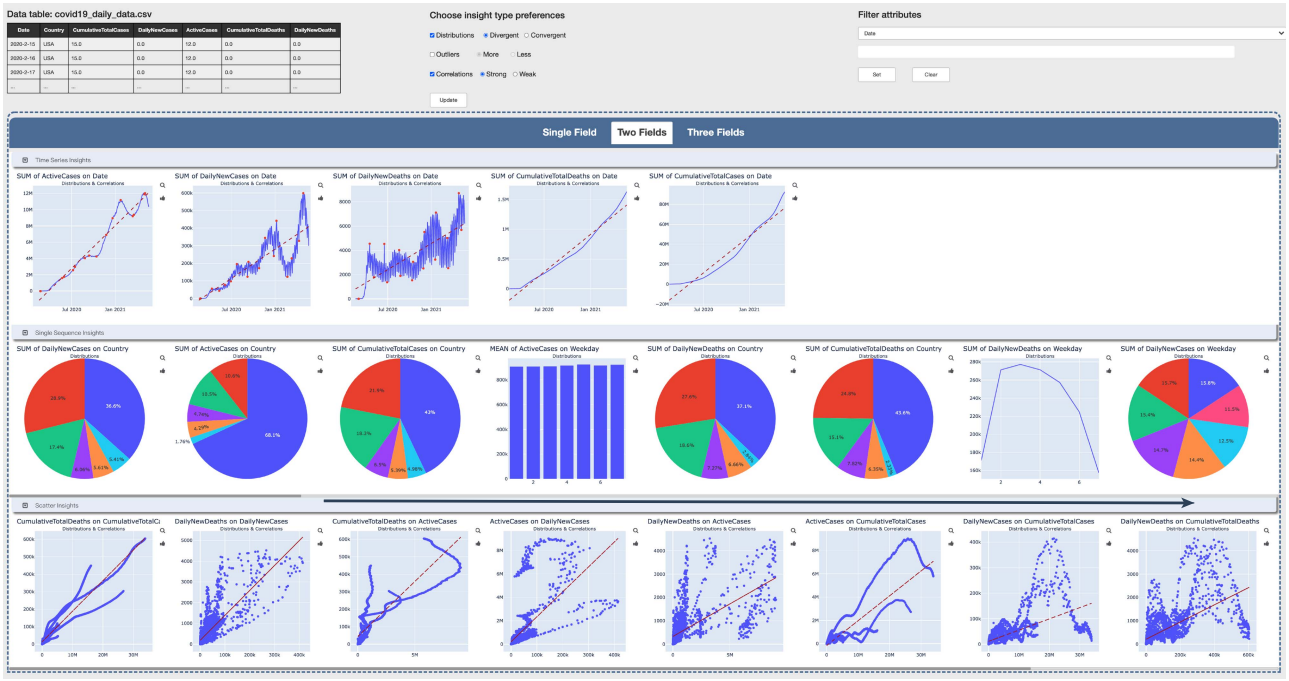


图 1　本文可视化推荐交互界面

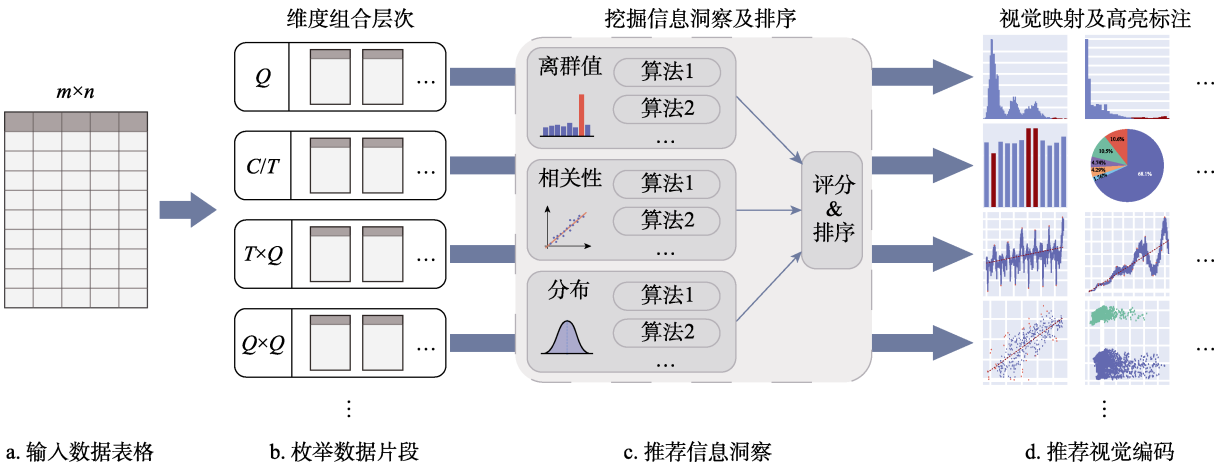


图 2 本文框架流程

得到枚举的数据片段. 本文使用多种算法挖掘信息见解, 并进行排序. 如图 2c 所示, 本文定义了离群值、相关性和数据分布 3 大类信息见解类型. 由于数据类型不同, 不同层次的数据片段需要使用不同的挖掘算法. 对于同层次同类型的信息见解, 本文也使用多种算法进行挖掘分析. 用户可以通过交互选择同时查看一种或多种类型的信息见解. 经过多层次、多算法的信息见解挖掘过程, 本文综合多种类型和多个方法的评价对每个层次内的信息见解进行排序, 用户可以通过界面交互决定排序的偏好.

基于信息见解的层次、类型和评价结果, 可以为其生成可视化图表, 如图 2d 所示. 不同于现有方法, 本文使用规则和机器学习相结合的方式决定每个信息见解的视觉编码, 使用的数据来自 VizML[6]提供的可视化图表数据集, 其覆盖的图表类型包含柱状图、饼图、条形图、折线图、堆叠条形图、复式折线图及散点图; 在此基础上, 在图表中为每个信息见解蕴含的关键信息进行突出标记, 增强用户的理解.

3 数据片段枚举

3.1 表格数据变换

一个 m 行 n 列的数据表格可以视为一个有 m 个记录点、n 个维度的高维数据. 其中, 每个维度根据其值的数据类型可以划分为 3 种维度类型: 分类维度 C 、时间维度 T 和数值维度 Q . 将这些维度的值按照一定的规则进行变换, 可以得到新的维度或生成新的表格. 本文支持分组变换、聚合变换以及排序变换 3 种维度变换方式.

3.1.1 分组变换

分组变换根据某个维度的值将原表格的 m 个记录点分为若干个组别的变换. 对于分类维度, 将在该维度取值相同的记录点分为同一组; 对于时间维度, 除同样可以将取值相同的点分为一组外, 还可以按时间单元分组, 如按年、月或星期分组; 数值维度则可以通过分段进行分组, 成为新的分类维度.

3.1.2 聚合变换

将原始数据进行分组变换后, 通常需要对每个分组内的记录点进行聚合操作, 为每个分组得到一个最终数值, 如此便能得到一个数据片段. 聚合变换包括计数以及求和、求均值等操作. 计数即以每个分组包含数据点的数量作为其度量值, 而除计数外的聚合变换都需要应用在一个数值维度上进行计算.

3.1.3 排序变换

排序变换根据一定的顺序对数据片段中的数据进行排序, 以更好地展示信息, 如按照时间顺序或按数值大小等排列数据.

3.2 枚举及分层

通过对表格维度及变换方式的遍历和组合, 可以枚举得到大量的数据片段, 作为该数据表格信息见解的候选.

现有的基于信息见解的可视化推荐系统[2-3,12]会在判断见解类型之后, 将数据片段根据见解类型进行分组. 然而很多情况下, 同一个数据片段有可能同时拥有不同的见解类别的信息, 根据见解类型分组将导致这些数据在不同组中重复出现, 也使用户无法高效地获取来自同一个数据片段的多类信息. 另外, 不同维度类型组合的数据片段需要使用不同的见解检测方法进行分析, 支持使用

的图表类型也不同, 相互之间无法直接进行对比.

基于以上考虑, 本文根据数据片段的维度类型组合将所有数据片段分为不同层次. 首先只使用单个维度的数据片段, 单个 Q 维度的数据片段将使用柱状图进行可视化, 单个 C 维度或 T 维度的数据片段会经过分组变换及计数聚合变换, 通过饼图、条形图或折线图展示该维度的分布情况. 随后是使用 2 个或 3 个维度的数据片段, 如表 1 所示. 由于本文使用 VizML[6]提供的数据学习图表映射, 各层次的数据片段所支持的图表类型由数据集的图表类型决定.

表 1　数据片段分层

维度数量	维度组合	支持的可视化图表
1	Q	柱状图
	C / T	饼图/条形图/折线图
2	$C\times Q$	饼图/条形图/折线图
	$T\times Q$	折线图
	$Q\times Q$	散点图
3	$C\times C\times Q$/$C\times T\times Q$	堆叠条形图/复式折线图
	$C\times Q\times Q$	分类散点图

随着维度数量增加, 数据片段的组合空间快速增长, 由 k 个维度所组成的数据片段数量能达到 $O(n^k)$ 复杂度, 具体受维度类型及变换的结果影响, 维度数量由少到多的分层方式能让系统优先处理简单的见解并呈现给用户, 不会出现需要长时间等待处理结果的问题; 每个片段所蕴含信息的复杂度也随着维度数量的增加而增加, 当面对一个陌生的数据集时, 维度数量由少到多的呈现方式能更好地帮助用户理解数据表格的信息, 并通过主动的筛选交互渐进式地进行探索; 如此分层后, 同一层次内的数据片段将使用相同的算法进行分析, 方便进行对比和排序.

4　信息见解挖掘及排序

通过枚举得到分层次的多个数据片段, 可以对所有数据片段进行分析, 判断其是否属于某一类信息见解, 根据信息的重要程度为每个层次的见解进行排序. 表 2 所示为本文使用的挖掘信息见解的算法及其所属的信息见解类型和适用的维度组合.

4.1　多层次的信息见解类型

已有方法大都为每种见解类型指定 1 种或 2 种维度类型组. 实际上, 各种不同的维度类型组合都可能存在这些类型的见解, 只是可能有不同的内涵, 或需要使用不同的算法进行挖掘, 因而不能直接进行比较. 只有为这些不同层次的数据片段都定义相应的见解类型并指定相应算法, 才能挖掘到这些见解, 更全面地找到数据中重要的信息. 如 Foresight[2], QuickInsights[3]和 Calliope[11]都定义了离群值见解类型, Foresight 的离群值类型仅支持 Q 维度的数据片段, QuickInsights 和 Calliope 则是只支持 $C\times Q$ 或 $T\times Q$ 维度. 本文则为所有单维度或双维度层次的数据片段都定义了离群值类型, 可以视作单维度离群值、时间序列离群值、散点图离群值等.

表 2　本文自动挖掘信息见解算法

见解类型	算法	适用维度组合
离群值	拉依达准则 IQR z-score 格拉布斯检验	Q / C / T / $C\times Q$ / $T\times Q$
	LOF[24] iForest[25]	$Q\times Q$
相关性	线性回归 自相关检测	$T\times Q$
	皮尔逊相关系数 斯皮尔曼相关系数 肯德尔相关系数 PCA	$Q\times Q$
	卡方检验 互信息 克莱姆关联系数	$C\times C\times Q$ / $C\times T\times Q$
数据分布	偏度 峰度	Q
	变异系数 基尼系数 信息熵 卡方独立性检验 夏皮罗-威尔克检验 KS 检验	Q / C / T / $C\times Q$
	ADF 检验 CH 检验 峰值检验	$T\times Q$
	k-means DBSCAN[26] 夏皮罗-威尔克检验	$Q\times Q$
	KS 检验 单因素方差分析	$C\times C\times Q$ / $C\times T\times Q$
	轮廓系数	$C\times Q\times Q$

4.2　多算法信息见解挖掘

已有的可视化推荐系统大都为每种见解类型设定单一的算法和衡量指标, 但对同样的见解类型往往有多种不同算法可以进行评估, 而根据“没有免费午餐定理”[27], 不同算法适用于不同的数据. 因此, 本文对每个维度层次和每种见解类型使用多种挖掘算法, 保证其能适用于用户指定的不

同数据集; 同时, 避免了使用固化的模板寻找信息见解, 只要数据片段在某种算法下被挖掘出值得关注的特征, 就认为该片段蕴含信息见解.

本文为每种见解类型使用大量不同的挖掘算法, 如表 2 所示, 这些算法有的是基于统计学的(如皮尔逊相关系数和夏皮罗-威尔克检验等), 有的是基于信息论的(如信息熵和互信息等), 还有的是基于有监督或无监督学习的(如 LOF, iForest, *k*-means 等). 其中, 有的算法需要对数据片段进行额外的变换操作, 如克莱姆关联系数需要将三维数据转换为矩阵; 有的算法需要尝试多组参数设置以找到最佳方式, 如 *k*-means 需要设置多个 k 值进行尝试(本文根据数据集取 $k\in[2,3,4,5]$), CH 检验需设置周期数参数 $m\in[4,12,30,52]$, DBSCAN 需设置参数组 $\varepsilon\in[0.1,0.3,0.5,0.7]$.

所有算法都有 2 个任务: 一是判断输入的数据片段是否属于该见解类型(任务 1); 二是输出一个得分以评价该见解, 用于对见解进行排序(任务 2).

对于任务 1, 统计检验的算法可以直接使用其 p 值作为判断依据, 如使用夏皮罗-威尔克检验判断数据片段是否服从高斯分布时, 若 p 值小于 0.05, 则拒绝原假设, 认为数据片段并不服从高斯分布. 其他算法则需要通过设置阈值作为判断依据, 如使用皮尔逊相关系数判断 2 个数值维度的相关性时, 使用 0.6 和−0.6 作为阈值[28].

对于任务 2, 大多数算法都能得到一个数值作为评分依据, 如皮尔逊相关系数等. 还有部分统计检验算法的输出统计值没有恒定的取值范围, 本文参考 QuickInsights[3]和 Calliope[11]的算法, 取其 p 值设计评分函数. 如线性回归算法, 本文参考 Calliope[11]的 Trend 见解类型设计了时间序列的线性相关性得分函数 $s=r^2\times(1-p)$. 其中, r^2 为相关系数, 衡量拟合的好坏程度.

值得注意的是, 同一个数据片段有可能属于多个见解类型. 如果根据见解类型对图表进行分组呈现, 就会使同样的数据重复出现. 本文则是在同一图表中将数据片段所蕴含的多类见解信息同时呈现, 避免了冗余, 也更符合人们分析数据获取信息的习惯.

本文设计的信息见解检验框架是非常灵活的, 每种算法都是互相独立且可替换的, 可以根据实际需求调整各见解类型的适用范围, 修改各算法的设置, 或者减少或新增更多的挖掘算法.

4.3 信息见解排序

得到信息见解后, 下一步便是在各数据片段层次内对这些见解进行排序, 以展示给用户. 不同于现有的大多数方法对每个见解类型使用单一的算法, 本文使用多种算法对数据进行分析和挖掘, 因此无法直接使用单个算法的结果对同类型的信息见解进行评分和排序; 此外, 由于本文支持用户同时查看同一个数据片段的不同见解类型, 还需要考虑结合多种见解类型的结果进行排序, 使得对信息见解的排序变得更加复杂.

设对于层次 L, $F_t=\{f_1,f_2,\cdots,f_n\}$ 是见解类型 $t\in T$ 适用于 L 的 n 种算法的集合, 那么每个信息见解 $i\in L$ 将对应使用 F_t 计算的 n 个分数.

为平衡同类型多种算法得分的方法, 本文对同层次同类型各算法的得分进行归一化处理. 对每个信息见解 $i\in L$, 有

$$S_{it}=\frac{1}{n}\sum_{j=1}^{n}g\left(f_j(i)\right) \tag{1}$$

其中, i 表示信息见解, 即一个从原始数据表格中提取或变换得到的数据片段; $g\left(f_j(i)\right)$ 表示该信息见解在该层次该类型的第 j 个算法中的得分, $g:R\to[0,1]$ 为归一化函数, 本文使用最小最大归一化方法.

式(1)为层次 L 中的信息见解赋予了见解类型 t 的信息评分. 对于不同的见解类型, 评分有不同的含义: 对于离群值类型, 更高的分数意味着更多、偏离程度更大的离群值; 对于相关性类型, 高分意味着更强的相关性; 而对于分布类型, 数值分布更不均匀、数值间有更大差异的数据将得到更高的分数. 但是这不一定就是用户感兴趣的, 如用户可能希望找到数据中离群值较少的信息, 或者呈正态分布的数据. 因此, 本文支持用户对排序的偏好进行交互, 当用户选择不同的偏好时, 将使用 $1-S_{it}$ 作为信息见解在该类型的信息评分.

当用户需要同时查看数据片段的多种信息见解类型时, 本文使用不同见解类型的信息评分均值作为其评分, 得到

$$S_i=\frac{1}{|T|}\sum S_{it} \tag{2}$$

其中, $|T|$ 表示所有见解类型的数量. 式(2)为层次 L 中的每个信息见解 i 赋予一个由多种信息见解类型的多种挖掘算法共同决定的分数. 层次 L 的信息见解根据该分数进行排序.

5　视觉编码映射

5.1　结合规则和机器学习方法的图表推荐

图表的选择取决于其是否能有效地传递信息片段所蕴含的见解信息，在很大程度上直接受到数据片段的维度类型约束. 例如，若要展示 2 个数值维度之间的分布关系，散点图是常见图表中唯一的选择，若要展示一个数值随时间变化的趋势，则通常会选择折线图. 因此，现有的基于信息见解的可视化推荐系统[2-3,12]都直接针对每种见解类型和维度组合指定使用的图表. DeepEye[20]等数据序列推荐系统也选择通过判断数据长度等简单的规则选择图表.

VizML[6]提供了一个可视化数据集，由大量网络用户绘制的可视化图表及对应数据片段组成. 本文通过对这些图表进行分析发现，即使属于相同见解类型，不同的数据也可能需要使用不同的图表以更好地展现其信息. 因此，本文使用基于机器学习的方法，从用户的选择中学习图表映射. 与现有的基于机器学习的视觉编码推荐方法(如 VizML[6]或 Data2Vis[7])不同，本文先对数据片段根据其维度组合进行分层，如表 1 所示，再在不同层次内根据信息见解的计算结果对视觉编码进行推荐，使用规则和机器学习相结合的方式提高推荐的有效性.

本文使用 VizML[6]提供的数据集，以第 4.2 节中的信息见解得分作为每个数据片段的特征，训练决策树模型使其为数据片段预测图表类型. 本文尝试了决策树模型和全连接 NN 模型，发现决策树模型能在数据集中取得更高的准确率，因此采用决策树模型作为决定图表类型的算法模型，详见第 6.1 节.

5.2　信息高亮标注

为更直观地将找出的见解信息传递给用户，本文在生成的图表上进行高亮标注，并生成文本描述该数据片段的所有见解信息. 图 3 所示为一个时间序列可视化示例的界面，其局部峰值被高亮标注出来，由于它具有随时间上升的趋势，同时为其拟合了趋势曲线；可以看出，其在 ADF 检验中表现为平稳序列，且具有自相关性.

应用本文框架还可以对已有的可视化图表进行优化. 对于一个现有的图表，可以提取其数据片段，进而应用相应的信息见解挖掘算法对其进行分析，判断其是否为用户传递了信息见解；获取相应特征后，可以为其高亮标注重要的信息，或者为其推荐更恰当的图表类型.

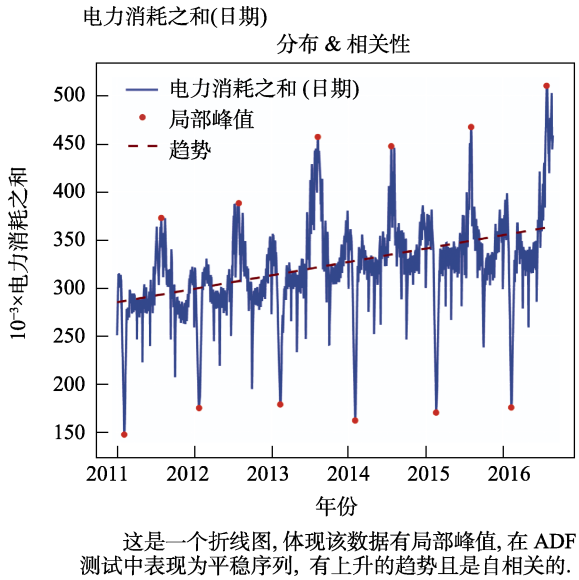


图 3　见解信息标注示例

6　实验与评估

为评估本文框架是否有效、准确，开展了一系列实验和用户调查.

6.1　图表类型预测

本文在 VizML[6]提供的数据集上验证视觉编码映射模块预测的图表类型是否准确，并与 VizML 方法进行对比.

6.1.1　数据处理

VizML 提供的完整数据共有 205 GB，收集自 Plotly 可视化社区，每条数据包括一个数据片段及其对应的图表. 由于直接收集自网络社区，数据集中有很大一部分重复、无效甚至错误的数据，需要进行清洗；数据集使用的表格数据格式与本文定义的数据片段格式不相同，需要进行转化；由于 Plotly 语言的图表类型不区分折线图和散点图，本文综合图表额外的属性描述及数据片段的特征对二者进行区分. 经过除重、清洗并转换为标准化的数据片段格式，共得到 120 330 组数据片段及对应图表，各类图表类型的数量和量级范围如表 3 所示. 另外，对数据进行预处理的代码及最终得到的训练数据，即本文为每个数据图表计算的信息见解特征已开源至 https://github.com/Bachery/VisGenData.

6.1.2　图表预测

根据第 2.2 节的定义，将得到的数据片段分为多个层次，并根据第 3.2 节的信息见解算法计算数据见解特征. 由于部分层次所支持的图表类型唯

表 3 各类型图表数量及量级范围

图表类型	数量	量级范围
柱状图	11471	[2,7187912]
饼图	10086	[2,25]
条形图	23728	[1,39]
折线图	24465	[2,2475]
散点图	15555	[20,4020383]
堆叠条形图	13230	[3,75]
复式折线图	10769	[4,7537]
分类散点图	11026	[20,1622168]

一(如表 1 所示), 如维度类型为 $Q\times Q$ 的数据片段仅支持散点图, 可以直接为这些层次的数据片段确定图表类型. 而对于支持多种图表类型的层次, 由于同层次的数据片段使用的见解特征维度相同, 可以根据每个数据片段的特征为其选择图表类型.

本文使用决策树(decision tree, DT)作为机器学习模型, 采用 CART(classification and regression tree)决策树算法[29], 在构建每层节点时使用基尼系数作为特征的选择标准. 在为维度类型为 C, Q 或 $C\times Q$ 的数据片段从饼图、条形图或折线图中选择图表的任务中, 构建的决策树模型有 41 层, 预测准确率达到 97.53%; 在为维度类型为 $C\times C\times Q$ 或 $C\times T\times Q$ 的数据片段选择堆叠条形图或复式折线图的任务中, 构建的决策树模型有 11 层, 预测准确率达到 99.98%.

作为对比, 本文还使用 NN 作为分类模型进行训练, 网络包含 4 层全连接层, 使用 ReLU 作为激活函数, 隐藏层长度 1000, 结果如图 4 所示. 可以看出, 效果相比决策树较差, 主要原因是基于信息见解的特征层次较高且长度较短.

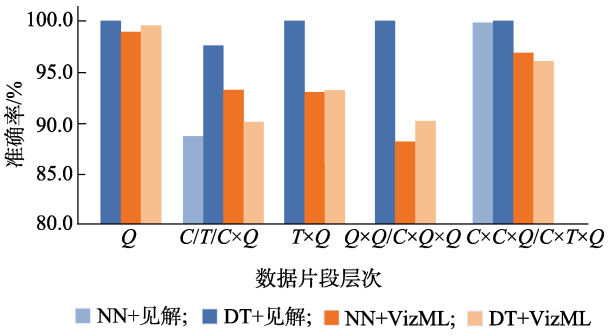


图 4 各信息见解层次图表预测准确率对比

将本文方法与 VizML 的方法进行对比. VizML 为所有类型的数据片段计算一样的特征, 得到长度为 841 的特征向量, 以此训练全连接 NN 预测图表类型, 网络结构同上. 本文还使用 VizML 提出的这些特征以同样的方式训练决策树, 结果如图 4 所示. 可以看出, 使用 VizML 的长特征能使 NN 取得更好的表现, 但本文方法在所有维度层次上都表现最好.

6.1.3 特征分析

通过分析训练得到的决策树模型各节点的权重, 可以比较各特征对决定图表类型的贡献程度, 既能帮助理解模型的决策, 也能指导今后对框架的改进或设计其他可视化系统. 表 4 所示为 2 个预测任务中最重要的 5 个特征.

表 4 训练的决策树模型中最重要的 5 个特征

排序	C / T / $C\times Q$	$C\times C\times Q$ / $C\times T\times Q$
1	信息熵	KS 检验
2	卡方检验	分组数
3	分组数	单因素方差分析
4	格拉布斯检验	克莱姆关联系数
5	夏皮罗-威尔克检验	互信息

对于维度类型为 C, Q 或 $C\times Q$ 的数据片段, 决定其图表类型最重要的是数值分布的均匀程度, 信息熵和卡方检验都可衡量一组数值的分布是否均匀, 仅有分布不均匀的数据片段会考虑使用饼图. 绝大多数饼图的分组数都小于 10, 而折线图的分组数则通常较多. 对于维度类型为 $C\times C\times Q$ 或 $C\times T\times Q$ 的数据片段, 同样是数值分布的均匀程度和分组数对决定图表类型的贡献最大, 分布较为均匀、分组数较少的片段会倾向于选择堆叠条形图, 反之则倾向于选择复式折线图.

6.2 用户调查

本文通过用户调查验证所提出的框架能帮助用户更好、更快、更全面地理解一个陌生的数据集, 邀请 15 名参与者用本文框架完成一些可视化分析任务.

6.2.1 实验设置

使用从 Kaggle 收集的 3 个数据集进行实验: (1) 有 8 个维度的关于好莱坞盈利最多电影的数据集(Hollywood); (2) 一个 8 维的奥运会历史奖牌数量数据集(Olympics); (3) 一个 7 维的各国新冠疫情数据集(COVID-19). 此外, 还有一个只有 3 个维度的关于中国华东几个省市电力消耗的数据集作为样例, 用于让参与者熟悉界面的使用.

实验中, 每个数据集设计 15 道单选题作为数据分析任务, 题目的设计参考 Amar 等[30]对信息可视化基本分析任务的讨论, 包括确定范围、寻找极值、计算派生值、发现相关性、描述分布等. 每道单选题都有 3~5 个选项, 包括一项“无法确定”. 每

个数据集的题目涵盖了表格中维度的各种组合, 如“该数据集统计的时间范围是多少?” 这样只涉及一个维度的问题, 或“美国的每日新增病例在哪一天达到峰值?” 这样涉及 3 个维度的问题等.

每次实验中, 参与者首先通过维度较少的电力消耗数据集熟悉界面及使用方式, 包括查看不同的见解类型、更改排序偏好等, 这里同样有几道单选题以便参与者学习, 其结果并不会纳入考量; 接下来正式的测试开始, 参与者逐个浏览 3 个数据集生成的图表, 并回答每个数据集的 15 道单选题, 每道题目的选项及花费的时间都被记录下来; 最后, 参与者还要完成一张李克特量表问卷, 分别评价系统对信息见解的分类和标注(Q1), 图表类型的选择(Q2), 层次的划分(Q3), 选择排序偏好(Q4)是否能帮助其更快、更好地理解数据并完成分析任务, 以及对系统的整体评价(Q5).

本文中, 所有参与者都是本科生或研究生, 在信息可视化方面没有或只有少量的经验.

6.2.2　实验结果

对信息见解的层次划分为用户提供了一个直观的工作流: 在进入一个新的数据集时, 几乎所有参与者都会先对数据表格进行了解, 对照浏览表格的各维度及第 1 个层次的图表, 再逐行确认各层次图表的内容后开始答题. 在答题过程中, 每位参与者在每个数据集上会进行 1~3 次图表更新操作, 更改见解类型或排序偏好.

本文以完成数据分析任务的准确率衡量参与者在实验中的表现, 结果如图 5 所示, 其中, 误差棒表示标准差. 可以看出, 整体的准确率为 79.4%, Hollywood 的平均准确率最低, 为 70.6%, 另外 2 个数据集的平均准确率都超过 80%; 从任务的难度来看, 准确率随着每道题目所需信息涉及的维度数量增加而下降.

在涉及维度较少的任务中, 错误率较高的是计算派生值的任务, 如 Hollywood 中的题目“有多少电影的全球总营收高于 450?”, 需要参与者在提供的柱状图中寻找分界点并进行计算, 出错的原因往往是计算错误; 在涉及维度较多的任务中, 最常见的错误选项是“无法确定”, 这一方面是因为图表的数量大幅增多, 参与者可能找不到所需的图表, 另一方面则是因为随着视觉通道的增加, 图表的解读难度增加; 即使是错误率最高的题目, 也有至少 33%的参与者选择了正确的答案, 说明在这些基本的可视化分析任务类型[30]中, 本文提出的枚举方法能够满足绝大部分需求.

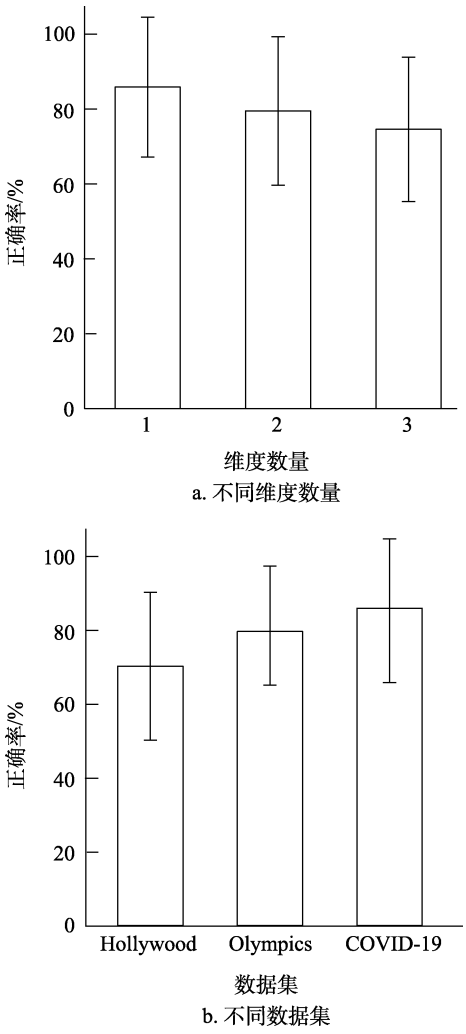


图 5　维度数量和数据集不同的题目平均准确率对比

李克特量表问卷评价结果如图 6 所示, 其中, 1 表示非常不同意, 5 表示非常同意, 误差棒表示标准差. 可以看出, 用户对系统各方面功能都给出了高评价, 在 5 级的量表中均取得了 4 分以上的评价. 其中, 对于系统主要功能模块, Q1 和 Q2 的评价最高且标准差较小, 说明本文框架有助于用户对陌生数据集的探索.

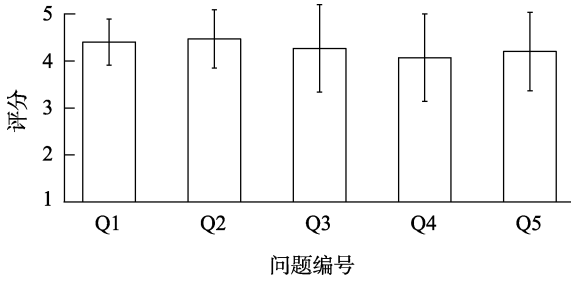


图 6　问卷结果

6.3　复杂度分析

本文使用 10 组从 Kaggle 收集的真实数据集测

试本文框架在不同情形下的表现并对算法的复杂度进行分析, 使用 M1 MacBook Pro, 数据集同样已开源至 https://github.com/Bachery/VisGenData. 实验结果如表 5 所示, 其中误差棒表示标准差. 由于不同数据集维度类型不同, 分组变换的结果不同, 所枚举的数据片段数量也有所不同, 一个维度数量为 m 的数据集将产生$[m^3,m^3]$个信息见解, 涵盖几乎所有维度组合和变换方式. 另外, 算法的耗时主要受到信息见解的数据类型而非见解的数量或数据集的量级影响, 如有上千信息见解的 Heart 数据集的总体耗时与仅有不到 300 个信息见解的 Covid19Daily 数据集相当. 由于本文对信息见解进行了分层, 交互界面无需等待所有计算完成就能先向用户展示使用单个维度的信息见解, 而所有数据集单个维度的信息见解计算耗时都不超过 1 s.

表 5 真实数据集实验结果

数据集	维度数量	数据量级	见解数量	总耗时/s
ForeignVistor	4	170	27	0.81
HollywoodsProfits	8	74	209	6.24
electricityConsump	3	10 320	36	6.31
SummerOlympic	8	2 238	236	12.36
Covid19Daily	7	2 784	292	24.23
Heart	14	303	1 271	25.11
WorldHappiness	20	149	433	32.01
HealthcareStroke	12	5 110	1 226	48.20
Covid19India	9	15 114	941	224.98
EcommerceShipping	12	10 999	1 103	324.77

为进一步分析算法的复杂度, 本文对不同算法的耗时占比进行了单独统计, 结果如表 6 所示. 可以看出, 除 CH 检验用于时序数据外, 耗时占比高的算法都用于对使用散点图的数据片段进行分析, 即算法的总体耗时取决于所枚举数据片段中维度类型为 $Q\times Q$ 的片段的数量和量级. 在对此类数据有较多需求的场景下, 或许可以先通过下采样快速获取初步的信息, 再根据需求选择进一步的计算.

表 6 不同算法耗时占比

算法	耗时占比/%
k-means	38.5
CH 检验	22.5
DBSCAN[26]	16.9
轮廓系数	8.6
皮尔逊相关系数	5.7
iForest	3.1
其他	4.8

7 结 语

本文提出一个基于信息见解和机器学习方法的可视化推荐框架. 首先对输入的数据表格进行遍历, 枚举数据片段并分组, 使用统计分析、机器学习等多种方法在数据片段中发掘不同类型的信息见解并进行排序; 然后使用结合规则和机器学习的方法推荐可视化图表, 并在图表上高亮标注出关键信息, 用户可根据分析需求选择查看不同的见解类型及排序方式. 该框架还可应用于对现有图表进行优化, 根据其展示的信息和使用的视觉编码提出优化建议. 实验及用户调查结果表明, 本文框架能有效、准确地提供可视化图表, 帮助用户快速、全面地了解一个陌生的数据表格.

本文框架还存在很多方面可以改进: (1) 在数据片段枚举过程中, 若某一维度跨度较大、可能存在多个局部显著模式时, 应考虑划分出其不同的子空间进行挖掘和分析; (2) 考虑不同挖掘算法之间的联系. 现实的数据分析任务往往要求算法串联, 如先检测离群值并剔除, 再进行相关性等分析; (3) 图表的丰富程度. 受到数据集的限制, 目前支持的图表类型较少.

今后还有一些方向可以考虑. 例如, 可以利用本文提取的见解特征对图表进行优化, 让图表的设计更好地服务于传递有效的信息, 如优化图表的配色[31]等; 在对数据片段进行排序的过程中, 本文对多种度量算法进行了等权重的线性组合, 在实际使用场景中, 可能还有更好的组合方式, 或许也可以根据用户反馈进行调整, 在未来的探索中进行尝试.

致谢 感谢腾讯医疗健康事业部医疗数据智能中心对本文工作的支持, 以及罗源、彭思翔和姜天意在项目进行过程中提出的宝贵建议!

参考文献(References):

[1] Wongsuphasawat K, Moritz D, Anand A, *et al*. Towards a general-purpose query language for visualization recommendation[C] //Proceedings of the Workshop on Human-in-the-Loop Data Analytics. New York: ACM Press, 2016: Article No.4

[2] Demiralp Ç, Haas P J, Parthasarathy S, *et al*. Foresight: recommending visual insights[J]. Proceedings of the VLDB Endowment, 2017, 10(12): 1937-1940

[3] Ding R, Han S, Xu Y, *et al*. QuickInsights: quick and automatic discovery of insights from multi-dimensional data[C] //Proceedings of the International Conference on Management of Data. New York: ACM Press, 2019: 317-332

[4] Mackinlay J. Automating the design of graphical presentations of relational information[J]. ACM Transactions on Graphics, 1986, 5(2): 110-141

[5] Mackinlay J, Hanrahan P, Stolte C. Show me: automatic presentation for visual analysis[J]. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(6): 1137-1144

[6] Hu K, Bakker M A, Li S, *et al*. VizML: a machine learning approach to visualization recommendation[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2019: 1-12

[7] Dibia V, Demiralp Ç. Data2Vis: automatic generation of data visualizations using sequence-to-sequence recurrent neural networks[J]. IEEE Computer Graphics and Applications, 2019, 39(5): 33-46

[8] Kandel S, Parikh R, Paepcke A, *et al*. Profiler: integrated statistical analysis and visualization for data quality assessment[C] //Proceedings of the International Working Conference on Advanced Visual Interfaces. New York: ACM Press, 2012: 547-554

[9] Vartak M, Rahman S, Madden S, *et al*. S_{EE}DB: efficient data-driven visualization recommendations to support visual analytics[J]. Proceedings of the VLDB Endowment, 2015, 8(13): 2182-2193

[10] Wang Y, Sun Z D, Zhang H D, *et al*. DataShot: automatic generation of fact sheets from tabular data[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 895-905

[11] Shi D Q, Xu X Y, Sun F L, *et al*. Calliope: automatic visual data story generation from a spreadsheet[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 453-463

[12] Harris C, Rossi R A, Malik S, *et al*. Insight-centric visualization recommendation[OL]. [2021-08-29]. https://arxiv.org/abs/2103.11297

[13] Roth S F, Kolojejchick J, Mattis J, *et al*. Interactive graphic design using automatic presentation knowledge[C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 1994: 112-117

[14] Casner S M. Task-analytic approach to the automated design of graphic presentations[J]. ACM Transactions on Graphics, 1991, 10(2): 111-151

[15] 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

[16] Perry D, Howe B, Key A M F, *et al*. VizDeck: streamlining exploratory visual analytics of scientific data[C] //Proceedings of iConference 2013. 2013: 338-350

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

[18] Wongsuphasawat K, Qu Z N, Moritz D, *et al*. Voyager 2: augmenting visual analysis with partial view specifications[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2017: 2648-2659

[19] Moritz D, Wang C L, Nelson G L, *et al*. Formalizing visualization design knowledge as constraints: actionable and extensible models in Draco[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 438-448

[20] Luo Y Y, Qin X D, Tang N, *et al*. DeepEye: towards automatic data visualization[C] //Proceedings of 2018 IEEE 34th International Conference on Data Engineering. Los Alamitos: IEEE Computer Society Press, 2018: 101-112

[21] Zhou M Y, Li Q T, He X Y, *et al*. Table2Charts: recommending charts by learning shared table representations[C] //Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2021: 2389-2399

[22] Chang R, Ziemkiewicz C, Green T M, *et al*. Defining insight for visual analytics[J]. IEEE Computer Graphics and Applications, 2009, 29(2): 14-17

[23] North C. Toward measuring visualization insight[J]. IEEE Computer Graphics and Applications, 2006, 26(3): 6-9

[24] Breunig M M, Kriegel H P, Ng R T, *et al*. LOF: identifying density-based local outliers[C] //Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2000: 93-104

[25] Liu F T, Ting K M, Zhou Z H. Isolation forest[C] //Proceedings of the 8th IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2008: 413-422

[26] Ester M, Kriegel H P, Sander J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise[C] //Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Palo Alto: AAAI Press, 1996: 226-231

[27] Wolpert D H, Macready W G. No free lunch theorems for search[R]. Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1996

[28] Muller K. Statistical power analysis for the behavioral sciences[J]. Technometrics, 1989, 31(4): 499-500

[29] Loh W Y. Classification and regression trees[J]. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2011, 1(1), 14-23.

[30] Amar R, Eagan J, Stasko J. Low-level components of analytic activity in information visualization[C] //Proceedings of the IEEE Symposium on Information Visualization. Los Alamitos: IEEE Computer Society Press, 2005: 111-117

[31] Wang Hui, Hu Ruizhen. Perception and harmony guided color assignment optimization for multi-charts[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(6): 815-825(in Chinese)
(王晖，胡瑞珍. 感知性与和谐度引导的多图表协同配色优化[J]. 计算机辅助设计与图形学学报, 2021, 33(6): 815-825)