

JavaScript 多模态信息探索器

Fenja Schweder,^{1,2}, Sebastian Trujillo-Gomez², and Kai Polsterer,²

¹*University of Bremen, Bremen, Germany; fen_sch@uni-bremen.de*

²*HITS gmbH, Heidelberg, Germany*

Abstract. 天文数据在数量、信息和方面上都非常丰富。尽管这提供了多样的研究视角，处理这些数据仍然是一个挑战。用于分析、检查、探索和数据通信的基础设施是必需的。为了解决这个问题，我们引入了 Jasmine，即 JavaScript Multimodal INformation Explorer (JavaScript 多模态信息探索器)。Jasmine 允许用户打开不同的数据显示模式，展示一组数据中的特定数据点。当前查看器支持图像数据以及点云对象。用户可以决定他们希望显示的数据点的相关信息。点云是交互式的，并允许放大、翻转和旋转。通过提供一组按照关键属性排列的结构化视图来启用选择数据点的功能。这种排列是通过自动编码实现的。

1. 介绍

大规模的调查和广泛的模拟为天文学家提供了不仅数量庞大而且涵盖各个方面、领域和维度的信息。在分析和处理数据方面的努力是由 Zhang & Zhao (2015) 广泛描述的一个遗留问题，目前仍在审查中 (Faaique (2023))。这不仅提高了对用于处理和分析数据的算法的期望，还要求基础设施来应用这些算法并使用结果。

可视化和交互性在提到的基础设施中起着关键作用。它们已在早期关于互动探索框架的提案中得到解决。Sciacca et al. (2014) 引入了一个平台，以促进对大型天文数据的访问、可视化和探索。另一个例子是 Lochner & Bassett (2021) 的天文学异常，这是一个用于在大数据集中检测异常值的网络应用。超粉红 (Kollasch & Polsterer 2024) 将数据集中的常见结构可视化自组织 Kohonen 图。

所述的解决方案解决了在大量数据集中提供结构化概览的难题。然而，额外的挑战是显示具有多个字段的单个数据点。单一的多维数据点可视化无法在不超载的情况下展示所有包含的信息。此外，由于某些方面的维度或行为不同，需要不同的可视化技术。理想的是一个数据检查器，能够切换感兴趣的维度。

2. 多模态数据探索的概念

在这项工作中，我们引入了 JavaScript 多模态信息探索器。JASMINE 是一个原型网络应用，能够对大数据集中的多维数据点进行检查和交互。JASMINE 的关键特性是收集并比较天文对象在多种模式下的不同视图。以下我们将解释底层概念以及如何实现该概念以满足天文数据的要求。

2.1. 设计一个大数据浏览器

由 Bederson & Shneiderman (2003) 编写的书籍信息可视化的技艺描述了如何创建用户友好且易于理解的科学可视化的有用指南。在这项工作中, Plaisant et al. (2003) 提供了设计信息浏览器的全面策略。我们多维数据探索者的方法遵循他们对多视图浏览器的分类。

基于 Shneiderman 的座右铭概览第一, 然后放大和过滤, 最后按需查看详细信息 (Shneiderman 2003), 启动窗口显示整个数据集的概览。选择一个数据点将打开该点的单独模态, 在其中可以进行进一步和更详细的检查。模态之间的通信是从概览窗口到详细窗口的单向通信。通过这种方式, 与概览窗口的交互可能会改变详细窗口的视图, 但反之则不行。

Plaisant et al. (2003) 重新推荐多窗口浏览器, 并提出窗口布局和对齐策略。将窗口管理留给用户也会给他们带来额外的工作负担。另一方面, 很难找到一种适当的战略来满足用户在多种情况下的需求。这尤其对于探索性研究的信息浏览器来说是一个问题, 在这种情况下, 任务重点最初可能不明确。

2.2. 大规模和多变量数据

由于我们预计数据量将会非常大, 因此可以选择适当的表示方法, 必要时可以借助机器学习。Kollasch & Polsterer (2022) 简要说明如何将聚类、回归和降维等技术应用于创建一个大数据集的交互式可视化。这些技术能够生成数据中的概览, 突出显示强特征。表示与实际数据之间的关系得以保留。在一个可视化的单一视图中展示多个字段的天文对象是无法实现的, 除非不使该可视化过载。针对这种多变量数据, 我们设计了一种数据类型, 允许将每个字段作为独立可访问的部分存储。原始集中的每个数据点都以数据立方体格式保存。立方体的每一面包含属于底层天文现象的不同字段。数据立方体可以嵌套, 以确保属性之间的层级关系, 这意味着每一面都可以有任意大量的子面。用户可以选择在详细窗口中显示某一面。所选的一面必须是最低层级的。由于采用多模态方法, 因此可以直接比较各个立方体侧面以及相应的数据字段。

3. 原型

我们概述了在 JavaScript 中实现引入概念的一个典型实例, 以及一个大规模多变量数据集的用例示例。结果是一个由浏览器驱动的客户运行端运行的网络应用程序。尽管 Plaisant et al. (2003) 不建议这样做, 但我们并未实施窗口管理策略, 而是将网页浏览器窗口用作模态对话框。由于种类繁多, 找到一种适当的方法来自动组织、对齐和调整窗口大小需要特别关注, 并将在未来得到解决。相反, 用户现在可以根据自己的需求自由设计浏览窗口。

3.1. 示例数据和准备

作为浏览极大数据集的一个案例, 我们以伊利斯图斯项目 (Nelson et al. 2018) 的快照为例。Illustris TNG 是一组大型宇宙学模拟, 预测包括许多复杂物理过程的星系形成和演化。这些模拟跨越了整个宇宙历史中的代表性宇宙体积, 追踪数千个星系的演化。每个快照包含不同类型的粒子。每个粒子场又具有多种物理属性。

我们抽取了 1000 个带有恒星模拟图像的星系样本。为了生成可探索表面, 我们使用了 *SphErinator & HiPSter* 框架 (Polsterer et al. 2024)。球化器从模拟图像中学习星系的基本形态特征。它使用交互式分层球形表示来投影任意大的数据集。所学的星系形态空间的潜在表示是可以解释的。潜在空间中的 3D 位置反映

了每个星系的形态属性。为了探索它们, *HiPSter* 创建了一个可以使用 *Aladin Lite* (Baumann et al. 2022) 查看的分层视图作为 *HiPS* 瓦片。

最后一个准备步骤是将样本中的每个星系组织成一个堆叠数据立方体, 每种粒子类型对应一侧, 并且每个数据字段对应一个子侧。为了在 *HiPS* 瓦片生成过程中实时完成这一操作, 我们对 *HiPSter* 源代码进行了轻微修改。

3.2. 数据驱动的方法

全局数据集层面的交互方法直接继承自 *Aladin Lite*。放大会增加 *HealPix* 单元的数量, 因此投射到球面的星系数量也会增多。一个下拉菜单允许用户在潜在空间重建和原始图像之间切换视图层。此外, 还可以叠加所有对象的目录坐标, 并通过 *TNG API* 直接访问该对象的所有模拟数据, 以交互方式检查星系元数据。

右键点击一个单元格会打开相应星系的详细窗口。该原型可以配置为一次显示一个详细窗口, 或同时显示相同或不同星系的多个窗口。用户可以根据需要排列、调整大小和移除模态框。

3.3. 数据点驱动的方法

详细模型显示了数据立方体中选定的一个子集。默认情况下, 用于训练形态表示的模拟图像是可见的。由于 *Illustris TNG* 提供了多个粒子云来代表气体、恒星或暗物质分布, 用户可以选择查看银河系的 3D 结构。*JASMINE* 允许通过旋转和缩放与粒子云进行交互。颜色编码显示了活动粒子云的场值。一组单选按钮使用户能够确定应显示哪个字段。为了使差异清晰, 每个字段都有不同的颜色比例。

4. 结论与展望

JASMINE 提供了从全局到详细视角探索非常大的数据集的基本功能。该原型展示了多模态数据浏览器所带来的潜力, 并提供了未来工作的多个想法。一个适合可变数据的窗口管理策略将大大提升用户体验。细节窗口与全局视图之间的双向通信为数据探索提供了新的可能性。最后, 我们计划开发一个通用界面, 使我们可以将 *JASMINE* 作为抽象数据浏览器附加到任何类型的数据处理管道上。

References

- Baumann, M., Boch, T., Pineau, F.-X., Fernique, P., Bot, C., & Allen, M. 2022, in *Astronomical Data Analysis Software and Systems XXX*, edited by J. E. Ruiz, F. Pierfederici, & P. Teuben, vol. 532 of *Astronomical Society of the Pacific Conference Series*, 7
- Bederson, B. B., & Shneiderman, B. 2003, *The Craft of Information Visualization* (San Francisco: Morgan Kaufmann). URL <https://www.sciencedirect.com/science/article/pii/B9781558609150500573>
- Faaigue, M. 2023, *International Journal of Mathematics, Statistics, and Computer Science*, 2, 96 – 113. URL <https://ijmscs.org/index.php/ijmscs/article/view/8561>
- Kollasch, F., & Polsterer, K. 2022, in *Astronomical Data Analysis Software and Systems XXX*, edited by J. E. Ruiz, F. Pierfederici, & P. Teuben, vol. 532 of *Astronomical Society of the Pacific Conference Series*, 11
- Kollasch, F., & Polsterer, K. 2024, *Ultralink – new possibilities to explore self-organizing kohonen maps*. 2406.03832, URL <https://arxiv.org/abs/2406.03832>
- Lochner, M., & Bassett, B. 2021, *Astronomy and Computing*, 36, 100481. URL <http://dx.doi.org/10.1016/j.ascom.2021.100481>

- Nelson, D., Pillepich, A., Springel, V., Weinberger, R., Hernquist, L., Pakmor, R., Genel, S., Torrey, P., Vogelsberger, M., Kauffmann, G., Marinacci, F., & Naiman, J. 2018, *MNRAS*, 475, 624. 1707.03395
- Plaisant, C., Carr, D., & Shneiderman, B. 2003, in *The Craft of Information Visualization*, edited by B. B. Bederson, & B. Shneiderman (San Francisco: Morgan Kaufmann), Interactive Technologies, 353–363. URL <https://www.sciencedirect.com/science/article/pii/B9781558609150500457>
- Polsterer, K. L., Doser, B., Fehlner, A., & Trujillo-Gomez, S. 2024. URL <https://arxiv.org/abs/2406.03810>
- Sciacca, E., Pistagna, C., Becciani, U., Costa, A., Massimino, P., Riggi, S., Vitello, F., Bandieramonte, M., & Krokos, M. 2014, in *2014 International Conference on High Performance Computing & Simulation (HPCS)*, 351
- Shneiderman, B. 2003, in *The Craft of Information Visualization*, edited by B. B. Bederson, & B. Shneiderman (San Francisco: Morgan Kaufmann), Interactive Technologies, 364–371. URL <https://www.sciencedirect.com/science/article/pii/B9781558609150500469>
- Zhang, Y., & Zhao, Y. 2015, *Data Science Journal*, 14