面向精度可控的大规模结构化数据集约减方法

艾志玮^{1,2)}, 冷珏琳^{1,2)*}, 夏芳^{1,2)}, 王华维^{1,2)}, 曹轶^{1,2)}

¹⁾ (北京应用物理与计算数学研究所 北京 100094)

²⁾ (中物院高性能数值模拟软件中心 北京 100088)

(leng_juelin@iapcm.ac.cn)

摘要: 科学与工程模拟产生的数据规模可达 TB 甚至 PB 量级, 数据约减已成为降低 I/O 时间开销和存储成本的重要手段。为实现高精度的科学可视化和数据分析, 面向大规模结构化数据集, 提出一个精度可控的数据约减方法。该方法首先以可视分析数据的插值误差为约束条件, 根据物理场量的空间分布特征构造多层嵌套的自适应背景网格; 然后, 将原始数据插值映射到轻量化的背景网格, 减少冗余数据的存储; 最后, 将约减后的数据集并行输出至高效存储的可视化文件。数据约减算法基于并行编程框架 JASMIN 实现, 能够无缝对接基于 JASMIN 框架研发的数值模拟程序。经测试, 并行算法可扩展至上万 CPU 核。数据约减方法已成功应用于无人机辐照电磁模拟, 在不超过 10% 的相对误差范围内, 将千亿结构网格数据集的规模降低了 99.8%。基于约减数据绘制的图像与原始数据图像之间的峰值信噪比为 47.08 dB, 具有较高的相似度, 满足可视分析的分辨率要求。

关键词: 数据约减; 结构化数据集; 精度可控; 自适应细化; 多分辨率技术

中图法分类号: TP391.41

DOI: 10.3724/SP.J.1089.2021.19263

Error-Controlled Data Reduction Approach for Large-Scale Structured Datasets

Ai Zhiwei^{1,2)}, Leng Juelin^{1,2)*}, Xia Fang^{1,2)}, Wang Huawei^{1,2)}, and Cao Yi^{1,2)}

¹⁾ (Institute of Applied Physics and Computational Mathematics, Beijing 100094)

²⁾ (Software Center for High Performance Numerical Simulation, China Academy of Engineering Physics, Beijing 100088)

Abstract: The massive datasets generated by scientific or engineering simulations have reached terabytes (TB) or even petabytes (PB). Data reduction has thus become one of the most important tools for saving I/O and storage costs. In order to achieve high-precision visualization and analysis, an error-controlled data reduction approach is proposed for reducing structured large-scale datasets. Firstly, taken the difference between the resulting data and the original one as a constraint, a multi-level structured adaptively-refined background grid is constructed, according to the spatial distribution characteristics of the underlying physical fields. Secondly, the original data is interpolated and mapped to the background grid, and as a result, the data with much less cells is obtained and the storage cost is reduced. Finally, the reduced data is exported to the parallel file system in real time. The proposed data reduction algorithm is implemented based on the parallel programming framework named JASMIN. In this way, the algorithm can be directly coupled with the numerical simulation programs developed with JASMIN. Test results demonstrate that the parallel algorithm can be extended to tens of thousands of CPU cores in parallel. The proposed algorithm has been successfully applied to the electromagnetic simulation of unmanned aerial vehicle irradiation. The cell number of a structured dataset with one hundred billions cells is

收稿日期: 2021-07-07; 修回日期: 2021-08-23. 艾志玮(1973—), 男, 硕士, 副研究员, CCF 会员, 主要研究方向为大规模科学与工程可视化; 冷珏琳(1988—), 女, 博士, 副研究员, 论文通讯作者, 主要研究方向为计算几何、网格生成; 夏芳(1970—), 女, 硕士, 高级工程师, CCF 会员, 主要研究方向为大规模数据管理; 王华维(1976—), 男, 博士, 副研究员, 硕士生导师, CCF 会员, 主要研究方向为科学计算可视化; 曹轶(1975—), 男, 博士, 研究员, 硕士生导师, CCF 会员, 主要研究方向为科学计算可视化。

reduced by 99.8%, with the relative error less than 10%. The peak signal-to-noise ratio between the two images, rendered using the reduced data and the original one respectively, is equal to 47.08 dB, which means a high similarity and thus satisfies the precision requirement of visualization.

Key words: data reduction; structured datasets; error-controlled; adaptive refinement; multi-resolution techniques

随着计算能力的快速增长,科学与工程模拟产生的数据规模可达到 TB(10^{12})甚至 PB(10^{15})量级,超大规模数据集的高效存储和后处理可视分析面临着巨大挑战.纵观高性能超算环境的建设历程^①,高性能计算机存储和 I/O 带宽的增长率远低于计算能力和数据规模的增长率,并且该趋势将在可预见的未来持续存在,导致超大规模数据集的 I/O、存储和分析成为制约科学与工程计算的瓶颈之一.

在大多数科学与工程应用中,数值模拟输出的数据在时间和空间维度上存在大量可压缩的冗余信息,并且可视分析对数据精度的要求往往低于数值模拟对精度的要求.因此,数据约减已成为降低 I/O 时间开销和存储成本的重要手段之一.另外,高压缩比的数据约减往往会引入不同程度的数据信息损失.因此,为实现高精度的科学可视化和可视分析,还应重点关注数据约减带来的精度损失.

1 相关工作

随着科学与工程模拟产生的数据规模的不断增长,越来越多的数据约减方法^[1]被应用于数值模拟结果数据的处理,通过数据约减或数据压缩的方式,达到减少数据量的目的.根据数据信息损失的程度,可将数据约减或数据压缩方法分为无损压缩和有损压缩.

无损压缩是指解压后的数据能够与原始数据完全一致,或仅存在浮点舍入误差.最具代表性的无损压缩技术包括基于熵的编码技术^[2]、基于字典的编码技术^[3]和基于变换的无损压缩方法^[4-5]等. bzip, zlib, gzip 等压缩工具都是基于这些技术发展起来的.经典的有损压缩方法包括基于截断的压缩方法^[6]、基于量化的压缩方法^[7-8]、基于预测的压缩方法^[9-10]和基于变换的有损压缩方法^[11-12]等.此外,原位可视化技术^[13-16]、网格约减技术^[17-19]

和时间采样技术^[20-22]在近年来被广泛用于减小可视分析数据的规模.其中,原位可视化技术在数值模拟执行过程中进行数据分析和可视化,并保存可视化图像;网格约减技术通过改变底层网格的顶点或单元数量,以达到减少数据量的目的;时间采样技术通过智能化的方式有选择性地输出用户感兴趣的时间数据片.

大多数支持完全无损压缩的方法能够减少的数据量非常有限,而有损压缩方法会造成数据精度的损失.与无损压缩相比,基于误差控制的有损压缩技术^[23-24]被认为是最佳的折中方案. Tao 等^[25]提出了一种误差可控的有损压缩技术,支持千核并行压缩 TB 量级的数据集,可将误差控制在给定的阈值范围内. Liang 等^[26]提出了一个自适应压缩框架,在数据集的不同区域动态选择改进的 Lorenzo 预测方法或优化的线性回归方法以获得最佳的压缩质量. Ainsworth 等^[27]提出了一种多重网格数据约减技术,通过误差控制确保约减后的数据集满足用户指定的容差,并将其应用于湍流模拟的数据集约减.

面向精度可控,本文提出一种基于分布式网格的数据约减 workflow,并进一步提出物理场精度可控的数据约减方法,通过网格约减的方式实现大规模结构化数据集的约减.

2 基于分布式网格的数据约减 workflow

2.1 问题描述

在高性能数值模拟中,为保证求解精度,计算网格的规模可达数十甚至上百亿.由于计算网格的规模过大,直接将定义在计算网格上的计算结果数据(物理场量)用于可视分析交互,不仅会带来庞大的 I/O 开销和数据存储开销,而且会影响可视分析交互的流畅性.

在实际应用中,可视分析对数据精度的要求往往低于数值模拟对精度的要求.因此,如果能够

① <https://www.top500.org>

在保证可视分析精度的前提下对计算结果数据进行约减, 可节省数据的存储和 I/O 的开销, 甚至提升可视分析的交互性能。

大规模数值模拟的结果数据通常定义在分布式网格上, 以一定的时间间隔输出和存储。本文面向分布式存储的大规模结构化数据集, 通过构造一套规模小于原始计算网格的轻量化背景网格, 来达到数据约减的目的。

为了保证可视分析的精度, 基于网格的数据约减应当满足精度可控的要求。本文对精度可控的定义如下。

设 u 表示定义在计算网格上的原始物理场; $\Pi_M u$ 表示物理场 u 在轻量化网格 M 上的插值结果。对于给定的误差阈值 ε , 如果满足 $\|u - \Pi_M u\|_\infty < \varepsilon$, 则称该约减方法是精度可控的。

2.2 数据约减 workflow

针对高性能数值模拟产生的大规模时变数据集, 本文提出如图 1 所示的基于分布式网格的数据约减 workflow, 将数据约减环节内嵌至数值模拟流程中, 在线执行数据约减操作, 并将约减后的数据输出到可视化文件中。

图 1 数据约减 workflow

数据约减 workflow 由 3 个部分组成。

(1) 数据产生。数值模拟程序在每个迭代步产生分布式存储的计算结果数据(物理场量), 数据定

义在分布式存储的计算网格上。

(2) 数据约减。针对每个时间步产生的物理场数据, 采用约减网格的方式实现数据量的约减。约减过程中需考虑轻量化网格生成、数据通信和数据插值等环节的并行效率。

(3) 数据输出。将约减后的轻量化数据以高效的存储形式并行输出到支持分布式并行读写的可视化文件中。

整个数据约减流程完成后, 采用可视分析工具读取数据文件, 实现计算结果数据的可视分析。

3 精度可控的结构化数据集约减方法

3.1 整体流程

基于第 2 节提出的数据约减 workflow, 本文面向大规模结构化数据集, 提出一种物理场精度可控的数据约减方法, 在保证物理场精度的前提下, 通过约减网格的方式生成一套轻量化的背景网格, 并将定义在原始计算网格上的物理场映射到轻量化网格上, 从而实现计算结果数据的约减。

为实现计算结果数据的高效存储和流畅交互显示, 采用多层嵌套的自适应结构网格作为轻量化的背景网格, 自适应网格的局部分辨率由物理场的空间分布特征和给定的误差阈值决定。数据约减并行算法基于并行编程框架 JASMIN^①[28]实现。目标是能够无缝地对接基于 JASMIN 框架开发的数值模拟程序, 在每个数据输出时间步完成数据约减操作, 并将约减后的数据集并行输出到可视化文件。

数据约减的并行算法流程如下。

输入。误差阈值 ε , 网格区域 $[x_{low}, x_{up}]$, 初始网格层的单元数 N_x, N_y, N_z , 最大网格层数 L 。

输出。约减后的数据文件(包含多层嵌套的自适应结构网格以及定义在网格上的物理量)。

Step1. 轻量化网格的初始网格层构造。根据输入的网格区域 $[x_{low}, x_{up}]$ 和单元数 N_x, N_y, N_z , 构造一个低分辨率的结构化初始网格层; 之后, 基于 JASMIN 框架将其分解成若干个网格片, 并分配给不同的处理器。

Step2. 数据收集。采用网格片映射技术, 针对初始网格层的单个网格片, 收集其覆盖区域的原始网格片及其物理场数据。

Step3. 局部分辨率计算。根据物理场数据的分布特征和误差阈值 ε , 计算原始网格片的单元中心点或网

① <http://www.caep-scns.ac.cn/JASMIN.php>

格节点处的局部分辨率。

Step4. 层次化自适应网格生成. 根据 Step3 得到的局部分辨率, 在轻量化网格初始网格层的基础上逐层加密网格, 使得物理场变化剧烈的区域网格分辨率较高, 变化平缓的区域网格分辨率较低。

Step5. 网格层聚类. 为了减少各网格层的结构化网格碎片, 采用 Box 聚类算法^[29]对网格单元进行聚类。

Step6. 数据插值. 通过插值算法将原始计算网络的物理量映射到轻量化的背景网格上, 得到轻量化的计算结果数据。

Step7. 数据输出. 采用层次化的数据存储格式将约减后的数据输出到支持分布式读写的数据文件中。

3.2 实现细节

3.2.1 网格片映射技术

原始计算网格和轻量化网格都由若干网格片组成, 这些网格片被分配到不同的处理器. 每个处理器独立管理被分配的网格片和定义在网格片上的物理场数据. 为了获取原始计算网格上的数据, 需要建立计算网格片到轻量化网格片的映射关系, 并获取每个轻量化网格片所覆盖区域范围内的原始计算结果数据。

原始计算网格与轻量化网格的数据传递通过 JASMIN 框架中的网格片映射技术实现, 如图 2 所示. 首先, 根据每个目标网格片的形状筛选出与该目标数据片重叠的所有源网格片. 然后, 将源网格片上(原始计算网格片)的数据进行合并和复制, 并传输给目标网格片(轻量化网格片). 为了保证后续逐层加密的轻量化网格在网格片交界处的连续性, 需要将轻量化网格片向外扩展一层影像区, 并存储相邻网格片上的数据. 整个过程涉及的网格片映射、数据复制和通信都通过调用 JASMIN 框架完成。

图 2 计算网格与轻量化网格的数据传递

3.2.2 局部分辨率计算

对于给定的误差阈值 ε , 根据物理场误差约束条件 $\|u - \Pi_M u\|_\infty < \varepsilon$, 计算轻量化网格需要满足的局部分辨率。

对于定义在网格单元上的物理场, 在可视化绘图时通常采用最邻近点插值算子进行数据插值,

即 $\Pi_M u$ 为最邻近点插值得到的物理场. 根据误差约束条件, 轻量化网格的局部分辨率计算公式为

$$S(u(\mathbf{x})) = \max \left\{ \min \left\{ \min_{k=1}^d \left\{ \frac{2\varepsilon}{\partial_k u(\mathbf{x}) \cdot d} \right\}, h_{\max} \right\}, h_{\min} \right\} \quad (1)$$

其中, d 表示维数; $\partial_k u(\mathbf{x})$ 表示物理场 u 在点 $\mathbf{x}$ 处的第 k 个方向的梯度。

对于定义在网格节点上的物理场, 在可视化绘图时通常采用线性插值算子进行数据插值, 即 $\Pi_M u$ 为线性插值得到的物理场. 根据误差约束条件, 轻量化网格的局部分辨率计算公式为

$$S(u(\mathbf{x})) = \max \left\{ \min \left\{ \frac{2(d+1)^2 \varepsilon}{d^2 \lambda_{\max}(u(\mathbf{x}))}, h_{\max} \right\}, h_{\min} \right\} \quad (2)$$

其中, d 表示维数; $\lambda_{\max}(u(\mathbf{x}))$ 表示物理场 u 在点 $\mathbf{x}$ 处的 Hessian 矩阵的最大特征值。

在式(1)和式(2)中, $h_{\max}$ 为轻量化网格初始网格层的单元尺寸; $h_{\min} = h_{\max} / 2^{L-1}$ 为轻量化网格最细网格层的单元尺寸. $h_{\min}$ 的设置通常不小于原始计算网格的网格尺寸. 在轻量化网格片所覆盖的区域内, 根据从计算网格上获取的物理场数据, 计算每个单元中心点(对于单元量)或网格节点(对于节点量) $\mathbf{p}_i$ 处的局部尺寸 s_i , 形成局部分辨率数据集 $\{(\mathbf{p}_i, s_i)\}_{i=1}^n$. 局部分辨率计算仅与周围一圈的相邻中心点或网格节点的物理量有关. 根据网格片影像区中存储的相邻网格片物理量, 可对每个网格片单独计算局部分辨率, 并且能够保证计算结果在网格片交界面上的连续性。

3.2.3 层次化自适应轻量化网格生成

轻量化网格的自适应加密过程如下。

Step1. 细分初始网格层, 形成 L 个网格层, 网格层编号为 $0 \sim L-1$. 相邻网格层之间的细化率为 $1:2$.

Step2. 根据特征源点数据集 $\{(\mathbf{p}_i, s_i)\}_{i=1}^n$, 对各网格层上的网格单元进行标记, 被标记为 0 的单元表示无效单元, 被标记为 1 的单元表示有效单元. 以第 l 层网格 ($0 \leq l \leq L-1$) 为例, 网格层上所有单元的初始标记为 0. 对于点 $\mathbf{p}_i$, 如果 $s_i < h_l = h_{\max} / 2^l$, 将点 $\mathbf{p}_i$ 附近的网格单元标记为 1.

3.2.4 网格层聚类

由于各网格层上的有效单元是非规则分布的, 为了实现有效的网格单元的结构化存储, 采用 Box 聚类算法对网格单元进行聚类, 即在各网格层构造若干个互不相交的 Box, 使其覆盖网格层上所有有效单元. 本文基于聚类算法^[29]进行结构化网

格聚类, 网格层聚类示意图如图 3 所示. 该算法为了避免产生过多的网格碎片, 允许 Box 中存在少量无效单元, 并通过参数控制无效单元的比例. 无效单元越多, 越有利于减少碎片(Box)的数量, 同时也会增加网格的数据量. 本文将无效单元的最大占比设置为 20%, 调用聚类算法对网格单元进行聚类, 实现结构化的网格数据存储.

图 3 结构网格规则化 Box 聚类示意图

网格层聚类操作从最细的网格层 $l = L - 1$ 开始执行, 当第 l 层网格聚类完成之后, 需更新第 $l - 1$ 层的有效网格单元, 使第 l 层的所有 Box 覆盖的区域都位于第 $l - 1$ 层有效网格区域的内部, 从而得到多层嵌套的结构化自适应网格.

3.2.5 数据插值

基于多层嵌套的自适应轻量化网格, 通过插值算法将计算网格的物理量映射到轻量化网格上, 获得每个网格层的网格单元或节点的物理量, 从而得到轻量化的可视分析数据. 对于单元中心量, 采用最邻近点插值算子进行插值; 对于节点量, 采用线性插值算子进行插值.

3.2.6 层次化数据存储

采用层次化的数据结构对多层嵌套的自适应轻量化网格及定义在网格上的数据进行存储、组织和管理. 图 4a 所示为 3 层嵌套的自适应结构网格, 图 4b 所示为各个网格层包含的网格单元, 以及细网格层在相邻粗网格层覆盖的计算区域.

图 4 二维多层嵌套结构化网格的层次化数据存储

层次化的数据结构包含多个网格层, 由粗到细依次为第 0, 1, ..., $L - 1$ 个网格层. 每个网格层由若干 Box 形状的网格片组成. 单个网格片的几何位置由左下角和右上角坐标确定, 网格片各方

向的单元数记为 (n_x, n_y, n_z) , 定义在网格片上数据则按照网格点或网格单元的索引 (i, j, k) 顺序存储.

层次化的数据结构还包括对相邻网格层之间各网格片嵌套关系(父子关系)的描述. 可视化绘图时, 网格片嵌套关系的描述用于保证重叠数据片的正确绘制. 对于一个网格片, 相邻细网格层中与其相交的网格片称为该网格片的子网格片; 相邻粗网格层中与其相交的网格片称为该网格片的父网格片. 网格片嵌套关系的核心数据结构包含: 子网格片数据, 用于记录子网格片个数以及所有子网格片的全局索引; 父网格片数据, 用于记录父网格片个数以及所有父网格片的全局索引.

最终, 多层嵌套的自适应轻量化网格和定义在网格上的数据以统一的 HDF5 数据格式输出到支持分布式并行读写的可视化文件中.

4 应用效果

本文的数据约减方法已成功地应用于复杂电磁环境瞬态脉冲照射下的电磁响应模拟, 通过对接复杂电磁环境时域全波模拟程序 JEMS-FDTD^[30], 实现了大规模结构化时变数据集的并行约减和可视化数据并行输出.

下面通过 2 个算例验证本文方法对大规模结构化数据集的数据约减能力.

(1) 喇叭模型电磁传播模拟算例. 计算规模为 $512 \times 512 \times 512$, 约 1.34 亿网格单元, 单时刻存储空间约为 2 102 MB, 在误差阈值 ε 分别为 0.01, 0.05, 0.10 的约束下对模拟结果数据进行约减.

(2) 无人机模型平面波辐照模拟算例. 计算规模为 $9200 \times 5200 \times 2400$, 约 1150 亿网格单元, 单时刻存储空间约为 1 030 GB, 在误差阈值 $\varepsilon = 0.10$ 的约束下对模拟结果数据进行约减.

针对上述算例, 分别从数据约减性能、可视化效率和可视化绘图误差 3 个方面评估本文方法.

4.1 数据约减性能评估

喇叭算例和无人机算例的数据压缩比和数据数据处理时间分别如表 1 和表 2 所示. 其中, 喇叭算例使用了 128 个 CPU 核, 无人机算例使用了 28 800 个 CPU 核, 测试环境为自主国产 CPU 超算系统. 在表 1 和表 2 中, 原始数据只统计数据输出时间, 压缩数据统计数据约减和数据输出的总时间. 对于喇叭数据集, 取误差阈值 $\varepsilon = 0.10$ 时, 可将网格单元数量从 1.34 亿约减至 277 万, 单元压缩比达到 48.5. 对于无人机数据集, 取误差阈值 $\varepsilon = 0.10$ 时,

可将网格单元数量从 1 150 亿约减至 2.59 亿, 数据压缩比达到 443.3. 对于千亿网格规模的无人机数据集, 能够大幅缩减数据的规模, 并减少数据输出时间. 最终, 数据约减和轻量化数据输出的总时间 (160.3 s) 比原始数据输出时间 (277.8 s) 减少了 42%.

表 1 喇叭数据集数据约减性能统计

数据类型	误差 阈值	单时刻数据 存储空间/MB	存储 压缩比	$10^{-6} \times$ 单元数	单元 压缩比	时间/s
原始数据		2 102		134.22		0.57
	0.01	170	12.4	16.57	8.1	1.66
约减数据	0.05	65	32.3	4.69	28.6	1.24
	0.10	47	44.7	2.77	48.5	1.11

表 2 无人机数据集数据约减性能统计

数据类型	误差 阈值	单时刻数据 存储空间/GB	存储 压缩比	$10^{-8} \times$ 单元数	单元压 缩比	时间/s
原始数据		1030.0		1148.16		277.8
约减数据	0.10	3.1	332.3	2.59	443.3	160.3

喇叭算例和无人机算例执行数据约减产生的层次化自适应轻量化网格分别如图 5 和图 6 所示. 可以看出, 约减之后的网格在物理场变化剧烈的区域具有较高的网格分辨率, 物理场变化平缓的区域则采用较低的网格分辨率.

图 5 喇叭算例的层次化多分辨率网格

图 6 无人机算例的层次化多分辨率网格

4.2 可视化时间统计结果

采用可视分析工具 TeraVAP^①测试数据轻量化之后数据读入和可视化绘图的性能提升情况. 喇叭数据集和无人机数据集的数据读入和绘图时间统计结果如表 3 所示. 其中, 喇叭数据集使用单核读入和绘制, 与原始数据相比, 约减数据的读入和绘制时间得到了小幅提升. 无人机数据集使用 64 核并行读入和绘制, 与原始数据集相比, 约减后的数据集读入时间减少了 91%, 体绘制时间减少了 90%. 总体而言, 数据约减有利于提升可视分析的效率.

表 3 可视化数据读入及绘图时间统计

数据集	误差 阈值	时间/s			
		读入	伪彩色图 绘制	半透明 绘制	体绘制
喇叭	原始数据	32.07	1.88	9.37	12.93
	0.01	24.02	1.02	5.57	12.56
	约减数据	0.05	21.69	1.13	5.05
	0.10	18.25	1.38	3.77	10.92
无人机	原始数据	208.99	6.33	59.65	125.63
	约减数据	0.10	18.59	7.06	44.61

4.3 可视化绘图误差统计结果

基于喇叭数据集原始数据绘制的电磁强度伪彩色切片图像和基于约减数据绘制的图像如图 7 所示, 无人机数据集的原始图像和约减数据绘图如图 8 所示. 分别采用结构相似性度量 (structural similarity, SSIM)、归一化相关度量 (normalized cross correlation, NCC) 和峰值信噪比 (peak signal-to-noise ratio, PSNR) 评估图像之间的相似度, 统计结果如表 4 所示. 统计结果显示, 基于约减数据绘制的图像与基于原始数据绘制的图像间的 SSIM 和 NCC 指标都大于 0.99, PSNR 指标都大于 30 dB, 具有良好的相似度, 满足可视分析的分辨率要求.

表 4 伪彩色切片图的相似度比对结果

数据集	误差阈值	SSIM	NCC	PNSR/dB
喇叭	0.01	0.998 67	0.999 99	56.166
	0.05	0.995 04	0.999 76	43.392
	0.10	0.992 17	0.993 07	38.071
无人机	0.10	0.999 17	0.995 76	47.083

① <http://www.caep-scns.ac.cn/TeraVAP.php>

图 7 喇叭算例原始数据和约减数据的伪彩色切片

图 8 无人机算例原始数据和约减数据的伪彩色切片

5 结 语

面向精度可控, 本文提出一种基于分布式网格的数据约减 workflow, 并进一步提出物理场精度可控的数据约减方法, 通过减少网格规模的方式实现大规模结构化数据集的约减. 本文的贡献在于: (1) 实现了精度可控的数据约减, 能够将数据的精度损失控制在给定的误差范围内; (2) 实现了数据约减并行算法, 能够无缝对接基于 JASMIN 框架研发的数值模拟程序, 在上万 CPU 核上并行约减千亿网格规模的结构化数据集; (3) 采用多层嵌套的自适应结构网格作为数据的背景网格, 实现了约减数据的高效存储, 大幅提升了数据 I/O 和后处理可视分析的效率.

本文提出的数据约减方法适用于单变量结构化数据集, 后续工作将对现有方法进行优化和扩展, 使其适用于多变量数据集和非结构化数据集.

此外, 本文采用低阶插值算子进行数据插值, 后续将探索高精度的数据插值算法, 进一步减少数据约减的误差.

参考文献(References):

- [1] Li S, Marsaglia N, Garth C, *et al.* Data reduction techniques for simulation, visualization and data analysis[J]. Computer Graphics Forum, 2018, 37(6): 422-447
- [2] Witten I H, Neal R M, Cleary J G. Arithmetic coding for data compression[J]. Communications of the ACM, 1987, 30(6): 520-540
- [3] Ziv J, Lempel A. A universal algorithm for sequential data compression[J]. IEEE Transactions on Information Theory, 1977, 23(3): 337-343
- [4] Strang G, Nguyen T. Wavelets and filter banks[M]. Wellesley: Wellesley-Cambridge Press, 1996
- [5] Ahmed N, Natarajan T, Rao K R. Discrete cosine transform[J]. IEEE Transactions on Computers, 1974, C-23(1): 90-93
- [6] Gong Z H, Rogers T, Jenkins J, *et al.* MLOC: multi-level layout optimization framework for compressed scientific data exploration with heterogeneous access patterns[C] //Proceedings of the 41st International Conference on Parallel Processing. Los Alamitos: IEEE Computer Society Press, 2012: 239-248
- [7] Iverson J, Kamath C, Karypis G. Fast and effective lossy compression algorithms for scientific datasets[C] //Proceedings of European Conference on Parallel Processing. Heidelberg: Springer, 2012: 843-856
- [8] Gersho A, Gray R M. Vector quantization and signal compression[M]. Heidelberg: Springer, 1992
- [9] Lakshminarasimhan S, Shah N, Ethier S, *et al.* Compressing the incompressible with ISABELA: in-situ reduction of spatio-temporal data[C] //Proceedings of European Conference on Parallel Processing. Heidelberg: Springer, 2011: 366-379
- [10] Di S, Cappello F. Fast error-bounded lossy HPC data compression with SZ[C] //Proceedings of the IEEE International Parallel and Distributed Processing Symposium. Los Alamitos: IEEE Computer Society Press, 2016: 730-739
- [11] Norton A, Clyne J. The VAPOR visualization application[M] //Bethel E, Childs H, Hanson C. High Performance Visualization. Boca Raton: CRC Press, 2012: 415-428
- [12] Fowler J E. QccPack: an open-source software library for quantization, compression, and coding[C] //Proceedings of Data Compression Conference. Los Alamitos: IEEE Computer Society Press, 2000: 554
- [13] Woodring J, Ahrens J, Figg J, *et al.* In-situ sampling of a large-scale particle simulation for interactive visualization and analysis[C] //Proceedings of the 13th Eurographics/IEEE-VGTC Conference on Visualization. Chichester: The Eurographics Association & John Wiley & Sons, Ltd., 2011: 1151-1160
- [14] Salloum M, Bennett J C, Pinar A, *et al.* Enabling adaptive scientific workflows via trigger detection[C] //Proceedings of the 1st Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization. New York: ACM

- Press, 2015: 41-45
- [15] Ahrens J, Jourdain S, O'Leary P, *et al.* An image-based approach to extreme scale in situ visualization and analysis[C] // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Los Alamitos: IEEE Computer Society Press, 2014: 424-434
- [16] Cao Y, Mo Z Y, Ai Z W, *et al.* In-situ visualization infrastructure for large scale simulations with structured meshes[C] // Proceedings of the 11th International Conferences on Computer Graphics, Visualization, Computer Vision and Image Processing. Lisbon: IADIS Press, 2017: 139-146
- [17] Renze K J, Oliver J H. Generalized unstructured decimation[J]. IEEE Computer Graphics and Applications, 1996, 16(6): 24-32
- [18] Trotts I J, Hamann B, Joy K I. Simplification of tetrahedral meshes with error bounds[J]. IEEE Transactions on Visualization and Computer Graphics, 1999, 5(3): 224-237
- [19] Clyne J, Mininni P, Norton A, *et al.* Interactive desktop analysis of high resolution simulations: application to turbulent plume dynamics and current sheet formation[J]. New Journal of Physics, 2007, 9: 301
- [20] Tong X, Lee T Y, Shen H W. Salient time steps selection from large scale time-varying data sets with dynamic time warping[C] // Proceedings of the IEEE Symposium on Large Data Analysis and Visualization. Los Alamitos: IEEE Computer Society Press, 2012: 49-56
- [21] Myers K, Lawrence E, Fugate M, *et al.* Partitioning a large simulation as it runs[J]. Technometrics, 2016, 58(3): 329-340
- [22] Ling J, Kegelmeyer W P, Aditya K, *et al.* Using feature importance metrics to detect events of interest in scientific computing applications[C] // Proceedings of the 7th IEEE Symposium on Large Data Analysis and Visualization. Los Alamitos: IEEE Computer Society Press, 2017: 55-63
- [23] Tao D W, Di S, Chen Z Z, *et al.* Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization[C] // Proceedings of the IEEE International Parallel and Distributed Processing Symposium. Los Alamitos: IEEE Computer Society Press, 2017: 1129-1139
- [24] Lindstrom P. Fixed-rate compressed floating-point arrays[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2674-2683
- [25] Tao D W, Di S, Chen Z Z, *et al.* In-depth exploration of single-snapshot lossy compression techniques for N -body simulations[C] // Proceedings of the IEEE International Conference on Big Data. Los Alamitos: IEEE Computer Society Press, 2017: 486-493
- [26] Liang X, Di S, Tao D W, *et al.* Error-controlled lossy compression optimized for high compression ratios of scientific datasets[C] // Proceedings of the IEEE International Conference on Big Data. Los Alamitos: IEEE Computer Society Press, 2018: 438-447
- [27] Ainsworth M, Tugluk O, Whitney B, *et al.* Multilevel techniques for compression and reduction of scientific data—the multivariate case[J]. SIAM Journal on Scientific Computing, 2019, 41(2): A1278-A1303
- [28] Mo Z Y, Zhang A Q, Cao X L, *et al.* JASMIN: a parallel software infrastructure for scientific computing[J]. Frontiers of Computer Science in China, 2010, 4: 480-488
- [29] Berger M, Rigoutsos I. An algorithm for point clustering and grid generation[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(5): 1278-1286
- [30] Li Hanyu, Zhou Haijing, Liao Cheng. Application of JEMS-FDTD in complicated electromagnetic environment study[J]. High Power Laser and Particle Beams, 2014, 26(7): 205-208(in Chinese)
(李瀚宇, 周海京, 廖成. 时域全波电磁计算程序 JEMS-FDTD 在复杂电磁环境研究中的应用[J]. 强激光与粒子束, 2014, 26(7): 205-208)