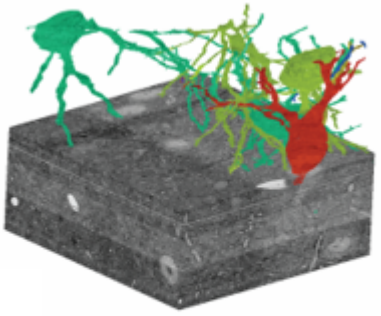
PKU Visualization Blog 北京大学可视化与可视分析博客

基于一种混合了确定性和不确定性的方法对超大规模切分体数据的剔除 (Culling for Extreme-Scale Segmentation Volumes: A Hybrid Deterministic and Probabilistic Approach)

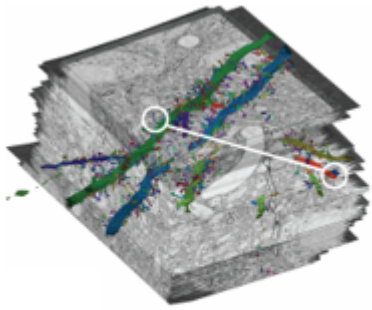
作者: Yang, Changhe 日期: 2019年5月4日

随着成像技术的快速发展，TB级别的超大规模的体数据逐渐频繁地出现在科研和产业界中。再加上近年来先进的体数据自动切分标注技术的出现，超过百万量级的密集标注的体数据已经能够生成。这样密集的标注方式，如32-bit的整数标注，又进一步增大了数据的体量。巨大的数据给对于切分标注数据的交互式可视化和可视分析带来了巨大的挑战。一方面，针对这样的数据生成高效的多分辨率层次结构十分困难，如何对于多分辨率层次结构的体数据匹配地进行标签的采样与生成需要新的技术。另一方面，对于特定的数据切块的查找和定位，十分费时费力，需要新的高效的数据组织方式来进行存储和访问。本文提出了一种新的数据管理的方式，可以有效地助力对超大规模切分体数据的剔除。

剔除（Culling）在可视化和图形学中具有较为明确的定义，具体为快速地减少对于算法输出没有影响的输入部分，在切分的标记数据中即是标注的数据。这也意味着，我们需要一个设计精良的数据组织来把这一部分数据筛除出来达到加速的效果。剔除技术针对渲染和可视分析查询的任务中都具有不可替代的作用。

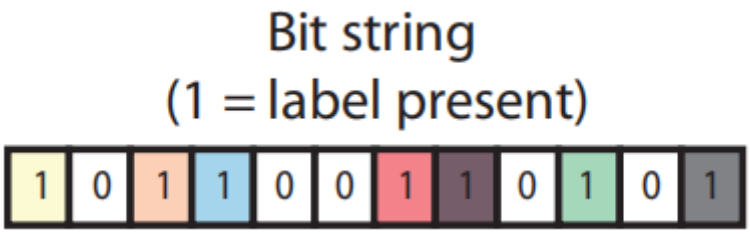


Empty Space Skipping Based Volume Rendering

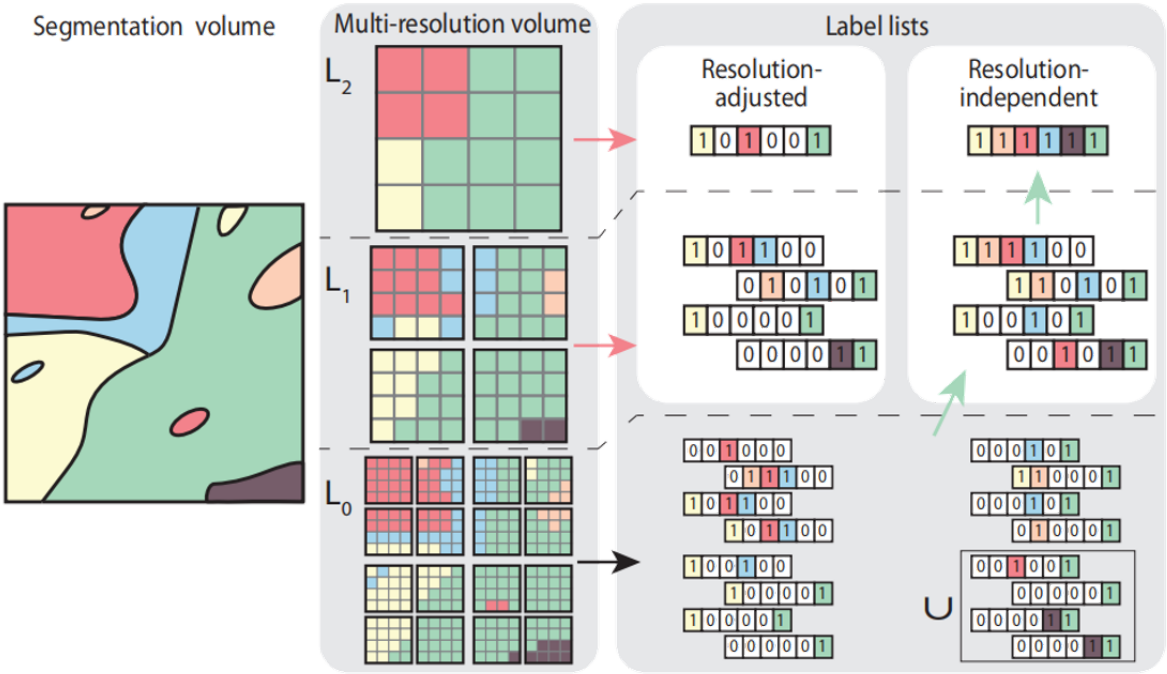


Spatial Queries of Segments

标注数据在体数据不同的数据空间区域，提供了空间内的标注ID的集合，通常可以表达为比特串的形式，即若空间内存在该类别的标注，则比特串中对应位置设置为1，反之为0。本文采用一种标注树（Label list tree）的结构，可以有效地减少存储的空间占用，提升查询的效率。标注树是一个对于标签数据的层次化描述，每一个节点代表体数据的一块空间区域，存储该区域内的标注信息。本文对于标注的比特串采用了多样的数据表达形式，包含多分辨率的标注（Multi-resolution label lists ）和数据自适应的标注（Data-adaptive label lists）。数据自适应的标注数据表达又具体为通过混合确定性和不确定性的方式。



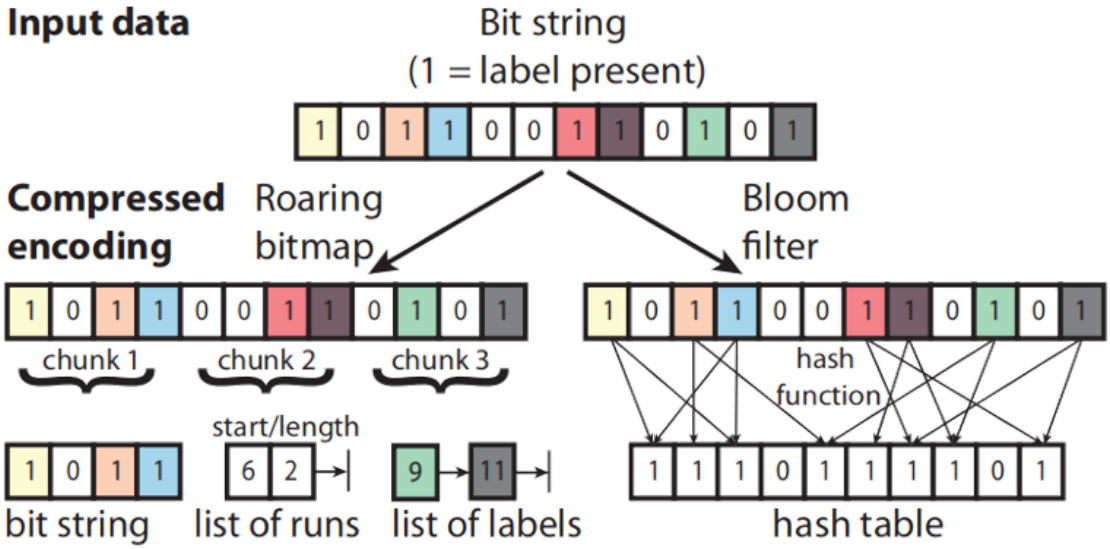
多分辨率的标注层次结构的生成针对不同的可视化和可视分析任务具体分为两种，即分辨率依赖的方式（Resolution-adjusted）和分辨率独立的方式（Hierarchy generation）。分辨率依赖的方式适应于体渲染，因为体渲染中仅仅关注于产生在屏幕上的可视化本身，因而可以仅仅关注于产生的低分辨率的中间数据。分辨率独立的方式适应于精准的空间查询，因为其必须依赖于原始数据本身。分辨率依赖的方式的标注层次结构直接由对应中间的体数据产生，分辨率独立的方式的标注层次结构则由原始分辨率的信息，采用并运算循序渐进的生成低分辨率的结果。



数据自适应的标注方式采用混合确定性和不确定性的方式，来压缩标注数据的比特串。确定性的方式为Roaring bitmaps的方式，基本思路为分箱存储。通过将32位的标注的比特串分解为前后16位，前者作为块(chunks)进行索引，后者存储在不同的容器之中。本文采用三种不同的编码方式，分别为非压缩的比特串（Uncompressed bit string）、排序的数组（Sorted array）、和游程编码（Run length encoding）。非压缩的比特串适应于相对的密集编码，排序的数组则相对适应于稀疏编码，在集合中元素小于4k时采用。将标注的索引存储为链表的形式。游程编码由起始点和长度的组成，适应于具有特点的比特串，仅当存储代价小于前两者时才会采用。在进行插入和检索时，可以通过对块进行二分查找，再定位到特定的容器。这是一种无损的编码方式，可以在多项式时间内完成数据的随机访问。

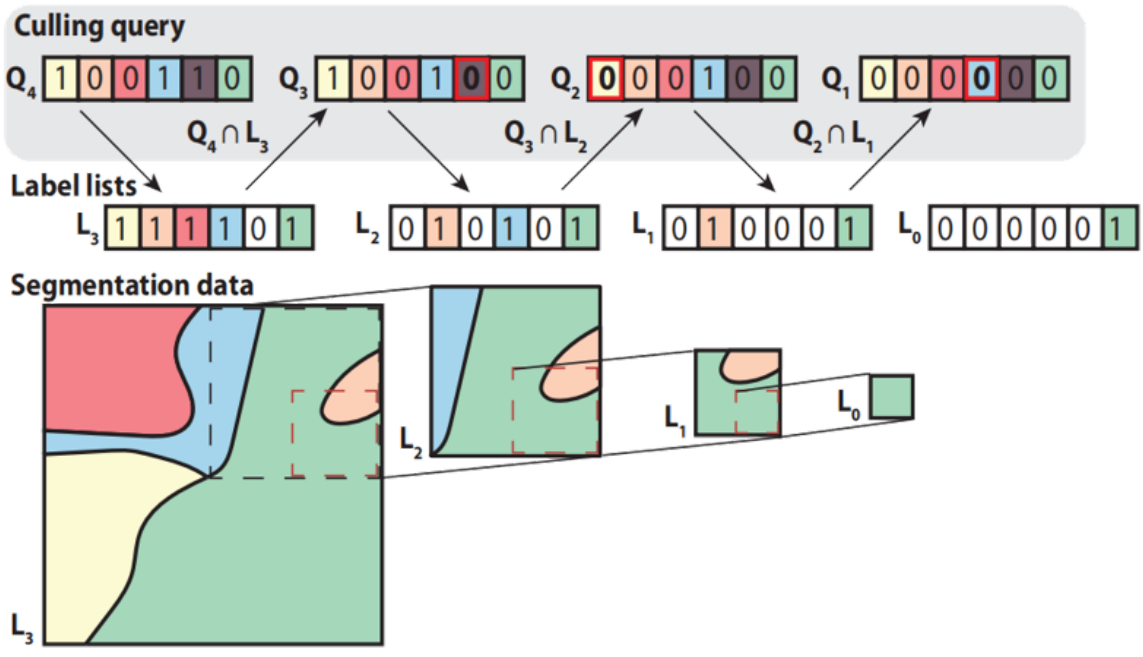
非确定性的方式为布隆选择器（Bloom filters ）的方式，其基本思想是对于每一个标注类别进行哈希映射，能够快速确定一个元素是否处于一个集合之中。插入元素是通过预先定义的m个哈希函数，将哈希表中由插入元素映射生成的m个对应位置由0设为1，而查找时则判断由待查找元素映射生成的m个对应位置是否均为1。若有一个位置不为1，则该元素不在集合中，否则有可能在集合中。布隆选择器可以以常数的时间判断一

个元素是否在集合中，但其具有可能假阳性的特点。它十分适应于我们的方法，即在庞大的全集中查找较小规模的情形。



本文同时采用了差异编码的方式（Delta encoding），来进一步紧致地压缩数据存储的空间。具体来说，差异编码的方式不去编码一个节点标注信息本身，而是通过编码其与父节点的标注信息的差异。通过计算父子节点间汉明距离和子节点标注集合的势可以估计采用差异编码是否能够压缩存储。

进行全局查询的方式为在待标注的体数据中查找特征的感兴趣区域，具体为查找带有某些标注种类的体数据区域并返回。由本文定义的层次结构，我们可以发现其具有反单调性，即父节点中若没有某些标注类别，则字节点中也没有，这可以视为我们进行快速终止和剪枝的判断条件。同样地，本文利用该性质还逐渐减少查询的内容。当发现某一个查询类别在某一节点没有找到时，则删除该查询中对于该标注类别的部分，逐渐减少查询请求，加速整个过程。对于上述确定性和不确定性两种编码方式的选用，本文采用一个基于阈值的选择，查询的集合元素个数大于某一阈值的采用Roaring bitmaps的方式，反之采用布隆选择器。



在应用方面，本文将该数据管理的策略集成到了体渲染和体数据查找的两个系统之中，有效地提升了存储和运行效率。

总的来说，本工作提出一种针对层次化的、带标注的体数据的行之有效的数据管理策略。本工作实则并非局限于科学可视化，许多时空查找、轨迹查找均可借鉴，值得一读。

参考文献

Beyer J , Mohammed H , Agus M , et al. Culling for Extreme-Scale Segmentation Volumes: A Hybrid Deterministic and Probabilistic Approach[J]. 2019.

论文报告　数据管理, 科学可视化

← 大分子轨迹可视化（Visualization of Large Molecular Trajectories)

用于交通数据预测的深度时空3维卷积神经网络 (Deep Spatial–Temporal 3D Convolutional Neural Networks for Traffic Data Forecasting) →

评论关闭。

RSS 订阅

功能

登录
文章RSS
评论RSS
WordPress.org

链接

北京大学可视化与可视分析研究小组主页 – PKU Vis Home Page
北京大学可视化研究维基 – PKU Vis WIKI

分类目录

应用
新闻
未分类
活动
研究
论文报告

标签

ChinaVis graph interaction PacificVis pacificvis2019 pvis2016 不确定性 主题模型 交互 交互设计 人机交互 会议 体可视化 体绘制 信息可视化 动态图可视化