文章编号:1671-8860(2007)07-0650-04　　**文献标志码**:A

基于 Hilbert 空间排列码的海量空间数据划分算法研究

王永杰[1]　孟令奎[1]　赵春宇[1]

(1　武汉大学遥感信息工程学院,武汉市珞喻路 129 号,430079)

摘　要:在深入分析了 Hilbert 空间排列码的线性映射特性后,将其应用于数据划分之中,并给出了具体的实现算法。本算法既考虑了空间目标的聚集性,又考虑了各个划分结点上数据存储量的平衡性,极大地提高了并行空间数据库的处理效率。

关键词:PC 集群; Hilbert 空间排列码; 空间数据划分

中图法分类号:P208

并行空间数据库系统是并行 GIS 和未来网格 GIS 应用的基础,是实现海量空间数据网络化存储与服务的关键技术和网格存储的重要解决方案之一,对于数字地球[1]的发展也将起到很大的推动作用。计算机集群环境下的并行空间数据库管理系统以高性能、高可用性和高扩充性为目标,充分利用多处理器平台的工作能力,多个处理机协同处理,以达到更快的数据库响应速度和分析能力。在很多情况下,并行数据库的海量空间数据分布在多个处理机中,并共同组成一个完整的数据库系统,因此,空间数据划分算法设计的好坏对并行空间数据库的处理效率有着极大的影响。对此,本文深入研究了空间数据的划分策略,并取得了较好的研究成果。

1　Hilbert 空间排列码

Hilbert 空间排列码源于经典的 Peano 曲线簇。该 Peano 曲线簇是闭合间隔单元 $I=[0,1]$ 到闭合矩形单元 $S=[0,1]^2$ 的连续映射,也是所有能够填满二维或更高维空间的连续分形曲线的总称,故又称为空间填充曲线(SFC)。目前,有多种空间填充曲线,它们的差别在于映射方法的不同[2-4]。Hilbert 空间填充曲线是一种优秀的线性映射方法,它在空间数据处理中的应用也越来越广泛。

Hilbert 曲线是递归定义的,每个曲线由 4 个同样的曲线(方向有所不同)构成,之间用短直线连接[5]。基本 Hilbert 曲线是一个大小为 2×2、阶数为 1 的栅格曲线,用 H_1 表示。为了获得 i 阶 Hilbert 曲线,将基本 Hilbert 曲线的每个顶点由 $i-1$ 阶 Hilbert 曲线所代替,同时 $i-1$ 阶 Hilbert 曲线可以进行必要的旋转或反射来适应新曲线[6]。

不同的空间填充曲线具有不同的空间聚集能力。通过对各种空间填充曲线的实验比较得知,Hilbert 空间填充曲线可以获得最佳的聚类效果[7],即 Hilbert 空间排列码的聚集性能最好。根据实验比较结果[8],一个比较好的矩形 Hilbert 空间排列码定义为:矩形 Hilbert 空间排列码设定为矩形中心点的 Hilbert 值。对于空间目标集合来说,其 Hilbert 空间排列码具有如下的空间聚集特征:① 空间目标的 Hilbert 空间排列码相邻,则空间目标是相邻的;② 空间目标相邻,则其 Hilbert 空间排列码一般是相邻的。

目前,有多种 Hilbert 码的生成算法。经典的 Hilbert 码生成算法由 Faloutsos 和 Roseman (1989)提出,它基于空间目标点所在的栅格格网的行列数的二进制位进行操作,其算法的复杂度为 $O(n_{i2})$,其中,n_i 为与空间目标点 i 所在最终栅

收稿日期:2007-04-26。

项目来源:国家 863 计划资助项目(2005AA113150);湖北省自然科学基金资助项目(2004ABA013)。

格格网行列数中较大者所对应的二进制位数。算法描述如下:① 将索引栅格格网的行列数采用二进制位交叉方法转化为对应的 Morton 码;② 以两位为一个单位,由高位开始,对生成的 Morton 码中的 10 和 11 互换;③ 由高位开始,设为 t_i,对后续各位中的 10 和 00 进行互换;④ 将结果按顺序排列,即为 Hilbert 空间排列码。

在 GIS 中,对于线状和面状的空间实体,可用其中心点来计算其 Hilbert 码。将空间目标按其 Hilbert 空间排列码排序,即可将目标在一定程度上进行空间聚类,这种聚集性减少了空间数据处理所要求的磁盘操作数,加快了数据处理的速度。

2　应用 Hilbert 空间排列码进行空间数据划分

Oracle Spatial 提供的空间数据划分策略对空间数据的划分是基于空间目标的 X 坐标值或 Y 坐标值,或者 X 和 Y 坐标值进行的,即通过坐标将目标集合所在的地理区域划分成一定数量的子区域,然后把这些子区域中的空间对象存储于相应的处理结点上。这种划分策略是对空间目标集合在 X 方向或 Y 方向、或者 X 和 Y 方向上进行的一种强制划分,并没有考虑空间目标的聚集性。因此,在数据划分完毕后,原本空间上聚集的目标被随意地划分到了几个不同的处理结点中,并且,即使在同一个结点上的具有一定聚集性的空间目标也不一定是紧密存储的。此外,它也没有考虑各个结点上数据存储量的平衡性。这些都将大大降低系统处理空间数据时的性能。本文将 Hilbert 空间排列码优秀的线性映射特性应用于空间数据划分之中,克服了 Oracle Spatial 空间数据划分所存在的缺陷,并取得了良好的应用效果。

在数据挖掘中,聚类分析是一个很重要的应用领域,目前应用较广泛的一种方法是基于质心的 K-均值聚类算法。它的基本思想是:给定一个 n 个样本的数据库,将数据划分为 k 个划分($k \leqslant n$),每个划分表示一个簇(每个簇内的数据具有聚类性),同时满足:① 每个簇至少包含一个样本;② 每个样本必须属于且仅属于一个簇。在应用 Hilbert 空间排列码进行空间数据划分时,基于这种思想,笔者提出了自己的划分策略。具体算法描述如下。

1) 把空间目标按 Hilbert 空间排列码由小到大的顺序随意划分为 k 个子集合,并把每个空间目标的 Hilbert 码作为其沿某个方向上的一维坐标,然后计算每个子集合的质心。

2) 计算每个空间目标到它所在的子集合以及邻近的两个子集合的质心的直线距离,然后把该目标分配到距离它最近的子集合中。

3) 分别计算新的 k 个子集合的质心。

4) 重复步骤 2)和 3),直到 k 个子集合的质心不再发生变化或总体平方误差满足要求为止。

5) 计算各个子集合中空间目标的数据大小占总量的百分比。如果各个百分比在数值 $1/k$ 的上下浮动范围之内,则算法结束,否则转入步骤 6)。

6) 把百分比超过浮动范围上限的子集合中的空间目标按其 Hilbert 码的顺序划分为几个小集合,使每个小集合的数据大小百分比保持在$1/k$的上下浮动范围之内(最后一个小集合的数据大小百分比可较小)。然后把百分比较小的集合合并为一定数量的子集合,合并以这些子集合之间的数据大小相对均衡、每个子集合中的空间目标的 Hilbert 码尽量邻近以及最终子集合的总数等于 k 为准则。

步骤 5)中的浮动阈值根据实际情况自定,以尽量维持各个结点存储目标的数据存储量的平衡性为准则。在上述算法中,k 的值即为处理结点的个数。执行该算法,并把 k 个子集合中的空间目标分别存储到 k 个处理结点上,这样就完成了空间数据的划分与存储,使得同一个结点内的空间目标之间的聚集性最大,而结点之间的聚集性最小,同时又尽量平衡了各结点间的数据存储量。这将极大地提高并行空间数据库管理系统的处理效率。在特殊情况下,空间目标集合中可能会有一个或者一些比较离散的对象,为了获得理想的划分结果,此时可以先把它们从集合中排除,再执行以上算法。算法执行完毕,再把这些对象分配到数据大小比较小的子集合中,或者把它们存储到一个单独的结点上。

应用本算法对空间数据进行划分,使得后续进行的并行空间操作的执行效率得到了很大的提高。例如,进行空间范围查询时,通过查询区域内的栅格格网的范围,得到对应的 Hilbert 空间排列码,以此作为地址码范围进行分区(处理结点)排除,系统性能得到了提高(极大地减小了空间数据搜索量,减少了空间索引量)。由于 Hilbert 空间排列码优秀的线性映射特性,所以空间数据经划分之后,空间中邻近的(聚集的)目标在各个处理结点上也是紧密存储的,从而减少了磁盘的访问时间,提高了数据处理效率。

3 实验分析

本实验是应用以上算法对某空间对象集合进行数据划分。图1是划分后的实验结果图(对于面状对象,取其中心点进行聚类划分)。在实验中,给定了4个处理结点。图中同一颜色的对象集合对应一个处理结点,每个结点对应的Hilbert码的数值范围如图例所示。表1列出了各个处理结点所对应的Hilbert码的数值范围和数据存储量百分比,可以看出,每个结点上的数据存储量是比较平衡的。可见,划分结果比较理想,达到了预期的效果。

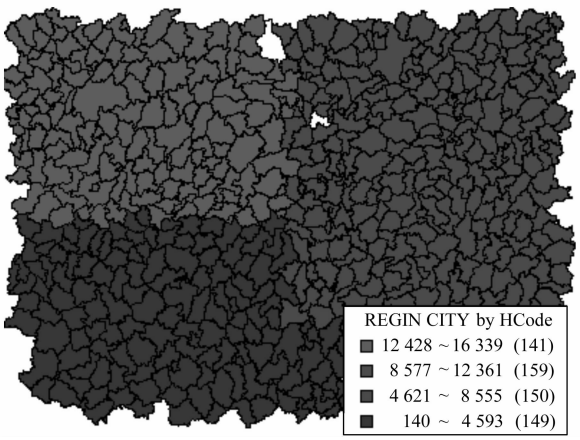

图1 应用Hilbert码进行空间数据划分后的实验结果图

Fig. 1 Result Map of Spatial Partitioning of Data

表1 各个处理结点对应的Hilbert码范围和数据存储量百分比

Tab. 1 Range of Hilbert Codes and Percentage of Data Storage on Each Processing Node

	结点1	结点2	结点3	结点4
Hilbert码范围	140～4 593	4 621～8 555	8 577～12 361	12 428～16 339
数据存储量百分比/%	23.9	21.8	28.1	26.2

表2是本文算法同使用Oracle Spatial提供的空间数据划分算法在对以上数据集进行范围查询实验时的响应速度对照表。其中的查询范围框比例是指查询范围框大小与整个数据集范围框之比;并行加速比是衡量并行系统性能的一个重要技术指标,本文把它定义为使用Oracle Spatial划分算法时的查询响应时间与使用本文算法时的查询响应时间之比。本实验中,同样大小的查询范围框各进行了30次查询,查询起始点随机选取,查询响应时间是30次查询的平均时间。

表2 两种划分算法下的响应速度对照表

Tab. 2 Comparison of Response Speed

查询范围框比例	并行加速比
1/15	0.98
1/5	1.29
2/5	1.26
1/2	1.17

从表2中的并行加速比可以发现,从总体上,本文的数据划分算法所对应的系统响应时间少于采用Oracle Spatial数据划分算法的系统响应时间,其主要原因是:本文算法使得空间上邻近(聚集)的目标位于同一个或几个结点(磁盘)上,并且空间中邻近的目标在各个结点上也是紧密存储的,这样可以减少磁盘的访问时间;各个结点上的数据存储量比较均衡,从而提高了数据处理效率,而Oracle Spatial所采用的划分策略则不具备这样的优势。此外,当查询范围框比例为1/15时,本文的划分算法所对应的系统响应时间多于采用Oracle Spatial划分算法时的系统响应时间。这是因为当查询范围较小时,系统一般只需要访问一个磁盘,而本文算法需要根据查询区域内的栅格格网的范围计算对应的Hilbert空间排列码,以此作为地址码范围进行分区(结点)排除,这将带来一定的系统开销。不过,从此时的并行加速比可知,这并没有较大地影响系统的数据处理效率。

4 结 语

基于递归Hilbert空间填充曲线的空间排列码具有良好的目标聚集特征,在空间数据的存储与处理领域极具应用潜力。本文算法既考虑了空间目标的聚集性,又考虑了各个划分结点上数据存储量的平衡性,避免了并行处理空间数据时单个或者极少数结点"过热"的情况,克服了Oracle Spatial空间数据划分所固有的一些缺陷。并结合具体实验取得了良好的数据划分结果,为后续对海量空间数据进行高效的并行运算作好了数据准备。随着网格技术的兴起,本算法也可以为网格环境下海量空间数据存储、组织与管理中的数据划分关键技术提供有意义的参考。

参 考 文 献

[1] 李德仁. 信息高速公路、空间数据基础设施与数字地球[J]. 测绘学报,1999,28(1):1-5

[2] 曹小林,莫则尧. 一种基于实测的高维动态负载平衡方法[J]. 计算机学报,2005,28(9):1 440-1 446

[3] Tetsuo A, Desh R, Thomas R, et al. Space-Filling Curves and Their Use in the Design of Geometric

Data Structures[J]. Theoretical Computer Science, 1997, 181(1): 3-15

[4] Lawder J K, King P J H. Using Space-Filling Curves for Multi-dimensional Indexing[C]. The 17th British National Conference on Databases: Advances in Databases, London, 2000

[5] Breinholt G, Schierz C. Algorithm 781: Generating Hilbert's Space-Filling Curve by Recursion[J]. ACM Transactions on Mathematical Software, 1998, 24(2): 184-189

[6] Kamel I, Faloutsos C. Hilbert *R*-tree: an Improved *R*-tree Using Fractals[C]. The 1994 International Conference on VLDB, Morgan Kaufmann, 1994

[7] Faloutsos C, Roseman S. Fractals for Secondary Key Retrieval[C]. The 8th ACM SIGACT-SIGMOD-SIGART Symposium on Principle of Database System, New York, 1989

[8] Kamel I, Faloutsos C. On Packing *R*-trees[C]. The 2nd International Conference on Information and Knowledge Management, New York, 1993

第一作者简介:王永杰,博士生。主要研究方向为 GIS、网络集群理论与技术、并行空间数据库。

E-mail:yjw1018@163.com

Spatial Partitioning of Massive Data Based on Hilbert Spatial Ordering Code

WANG Yongjie[1]　*MENG Lingkui*[1]　*ZHAO Chunyu*[1]

(1　School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: The excellent linear mapping characteristics of Hilbert spatial ordering code is studied and applied to spatial partitioning of data, and a concrete algorithm is given. In this algorithm, the clustering performance of spatial objects is taken into account, and the balance of data storage on each processing node is also taken into account, which greatly improves the processing efficiency of parallel spatial database.

Key words: PC cluster; Hilbert spatial ordering code; spatial partitioning of data

About the first author: WANG Yongjie, Ph. D candidate, majors in GIS, parallel spatial database, etc.

E-mail: yjw1018@163.com

(上接第 632 页)

Compression and Optimization of the Line Features Based on Wavelet Analysis

WANG Yuhai[1]　*ZHU Changqing*[2]

(1　Institute of Science, Information and Engineering University, 66 Middle Longhai Road, Zhengzhou 450052, China)

(2　Institute of Surveying and Mapping, Information and Engineering University, 66 Middle Longhai Road, Zhengzhou 450052, China)

Abstract: It is important to compress and optimize the line features in the studying of the terrain environment simultaion, cartographic generalization and GIS. Based on the theory of wavelet analysis, the algorithm of Douglas and curvature analysis, the compression and optimization of the line features and the self-adaptive model are studied. The results of some experiments show that the proposed approach maintaines not only high compressing ratio but also the characteristics of the line features very well.

Key words: wavelet analysis; Douglas algorithm; curvature analysis algorithm; line features

About the first author: WANG Yuhai, lecturer, Ph. D candidate, majors in interests are integration of GPS, GIS and GSM, spatial data processing and wavelet analysis.

E-mail: wyh7121@163.com