北京大学学报(自然科学版) 第 54 卷 第 2 期 2018 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 2 (Mar. 2018)
doi: 10.13209/j.0479-8023.2017.147

多尺度空间填充曲线空间连续性研究

翟卫欣[1] 陈波[2] 童晓冲[3] 程承旗[2,†]

1. 北京大学遥感与地理信息系统研究所, 北京 100871; 2. 北京大学工学院空天信息工程研究中心, 北京 100871;
3. 信息工程大学地理空间信息空间学院, 郑州 450001; † 通信作者, E-mail: ccq@pku.edu.cn

摘要 将二维 Hilbert 编码和 Z 编码拓展到以尺度维作为第三维的三维填充曲线: 多尺度 Hilbert 曲线和 Z 曲线。在多尺度数据条件下, 这两种曲线能够提高空间填充曲线的空间连续性, 适应多尺度的需求。依托四叉树模型, 将多尺度的 Hilbert 曲线与按照相同思路设计的多尺度 Z 曲线进行两类对比试验, 验证了多尺度 Hilbert 曲线相对于 Z 曲线在空间连续性方面的优势, 提高的比例在 15%~30%之间。

关键词 多尺度; Hilbert曲线; 空间连续性

中图分类号 P208

Research on Continuity of Multi-Scale Space-Filling Curves

ZHAI Weixin[1], CHEN Bo[2], TONG Xiaochong[3], CHENG Chengqi[2,†]

1. Institute of Remote Sensing and Geographic Information System, Peking University, Beijing 100871; 2. Aerospace Information Engineering Research Center, Peking University, Beijing 100871; 3. Institute of Survey and Mapping, Information Engineering University, Zhengzhou 450001; † Corresponding author, E-mail: ccq@pku.edu.cn

Abstract Multi-scale two-dimensional Hilbert curve is constructed, and specially the scale dimension is treated as the third dimension. The new structure embodies the multi-level characteristics and overcomes the drawback of Z sequence coding pattern, thus improving the continuity of the curve and advancing the spatial retrieval efficiency. The authors conducted two kinds of experiments based on the quad-tree model to compare the retrieval efficiency of Hilbert curve and Z curve. The consequence indicates that the multi-scale Hilbert curve performs better than Z curve, and the improvement on different data distributions vary from 15% to 30%.

Key words multi-scale; Hilbert curve; spatial continuity

随着空间数据获取方式的不断进步, 多源、多类型的空间数据急剧增长, 但空间数据的组织仍面临许多实际问题, 尤其是空间数据编码的问题, 给实际应用带来极大的挑战。地球剖分理论提出一套空间数据的统一编码体系, 为解决此类问题提供了可行的解决方案[1–2]。

在地球剖分的体系内, 建立合理的空间索引结构能够优化空间数据库的各类操作。在目前空间数据量高速增长的情况下, 对于空间索引效率的提高尤为重要。二维空间中的四叉树索引是一种常见的空间索引结构, 1974 年由 Finkel 等[3]提出, 四叉树的基本思想是, 将地理信息的规划分为不同层次的树结构, 将每一级的全空间等分为 4 个子空间, 直至达到最高层级为止。传统的四叉树索引的编码以二维的 Z 曲线为基础[4–5], 也有学者提出线性四叉树等编码方式, 并应用于空间库索引结构, 提高了查询效率[6]。

空间填充曲线的空间连续性用于描述在空间上连续的地理实体被空间填充曲线转化为一维数值编码后编码的连续性状况。对于同样的地理实体, 转化后的编码间隔数越小, 该空间填充曲线的空间连续性越好。

国家科技重大专项(11-Y20A02-9001-16/17, 30-Y20A01-9003-16/17, 30-Y30B13-9003-14/16)和公益性行业(测绘地理信息)科研专项(201512020)资助

收稿日期: 2016–12–12; 修回日期: 2017–10–17; 网络出版日期: 2017–10–23

Z 曲线的连续性较差，在四叉树索引结构中编码的连续性不强，在一定程度上影响了查询效率。作为另一种空间填充曲线，Hilbert 曲线同样可应用到四叉树索引中。一些学者在单一尺度的 Hilbert 曲线与 Z 曲线的空间连接性效率的研究中，证明了 Hilbert 曲线的优越性[7–8]。

当前的空间数据除数据量激增外，其空间尺度的巨大差异性也为数据管理提出了更高的要求[9–11]，例如需要同时对海洋(10^3 km 级)、湖泊(km 级)、池塘(m 级)进行管理时，空间数据库尺度的不一致会带来效率的低下。传统的 Hilbert 曲线和 Z 曲线并不具备多层次的特征。本文将 Hilbert 曲线和 Z 曲线与四叉树索引相结合，并考虑空间尺度维的需求，构造一种按照 Hilbert 方式编码的二维空间数据多尺度编码方式。这种结构既能够适应海量的空间数据多尺度的管理需求，又具备较强的连续性特征。本文提出的多尺度 Hilbert 曲线和 Z 曲线是将二维 Hilbert 编码和 Z 编码拓展到以尺度维作为第三维的填充曲线，在多尺度的数据条件下，提高了空间填充曲线的空间连续性。

1 Hilbert 曲线

Hilbert 曲线属于 Peano 曲线族。Peano 曲线族是闭合间隔单元 $I=[0, 1]$到闭合单元 $S=[0, 1]^n$ 的连续映射[7–8]，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，又称为空间填充曲线。Hilbert 曲线已在图像存储[12]和检索[13]、空间数据库索引等领域得到成功的应用[14–17]。

Hilbert 曲线有多种定义方式，例如递归定义[12,18]、浮点数字节定义[7,19]和 L 系统定义[20]等。最常见的是递归定义：每个曲线由 4 个同样的子曲线(方向有所不同)构成，子曲线之间用短直线连接。以二维 3 级 Hilbert 曲线为例，如图 1 所示。

尽管空间填充曲线均能实现从高维空间到一维空间的一一映射(或者相反)，但是不同的空间填充曲线的空间聚集能力并不相同。Faloutsos 等[8]通过对各种空间填充曲线的大量实验得到结论：Hilbert 空间填充曲线可以获得最佳的聚类效果，即 Hilbert 编码的聚集性能最好。

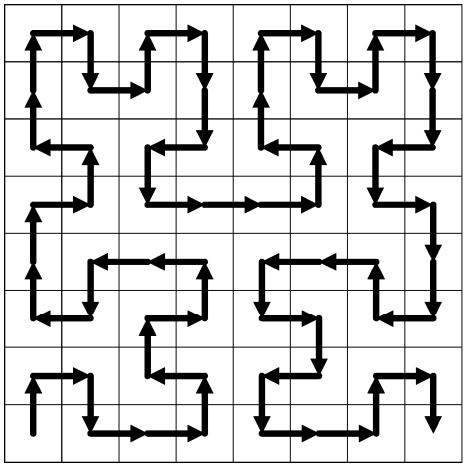

图 1 二维 3 级 Hilbert 曲线
Fig. 1 2-D Hilbert curve on level 3

2 多尺度 Hilbert 曲线

2.1 曲线结构

本文提出的多尺度 Hilbert 曲线构造了多层次的曲线结构。该结构中每一层均为二维 Hilbert 曲线，相邻高层级的曲线由相邻低层级的曲线平移或旋转得到。如图 2 所示，从 0 级到 1 级，是将一个网格等分成 4 个小网格，第 1 级中的 Hilbert 曲线依次从左上角的网格 1 出发，往右到右上角的网格 2，再往下到右下角的网格 3，再往左到左下角的网格 4，这是第一次遍历。再如，从 1 级到 2 级，高层级的网格中，右上角的网格和右下角的网格与相邻低层级的 Hilbert 曲线相同，左上角的网格由相邻低层级的 Hilbert 曲线顺时针旋转 90°得到，左下角的网格由相邻低层级的 Hilbert 曲线逆时针旋转 90°得到。更高层级的 Hilbert 曲线也均是相邻低层级的 Hilbert 曲线以其开口方向为基准，经过规则的变换得到。如果对每一层级的 4 个小网格均继续上述过程，往下划分，反复进行，最终得到一条可

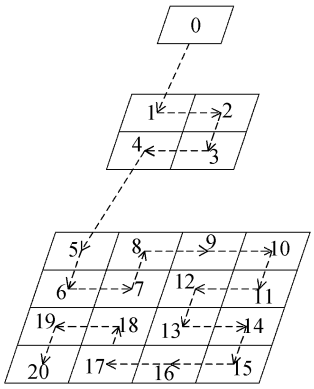

图 2 多尺度 Hilbert 曲线
Fig. 2 Multi-dimensional Hilbert curve

以填满整个网格的多尺度的曲线。第 1 层有 1 个网格，第 2 层有 4 个网格，第 3 层有 16 个网格，……，第 n 层有 4^{n-1} 个网格。每一层均采取 Hilbert 曲线的编码方式，既能适应空间对象的尺度不同对空间填充曲线的要求，又具备较强的连续性。

2.2　曲线性质

多尺度的 Hilbert 曲线具备普通二维 Hilbert 曲线的递归性质，即每一层的曲线均可由相邻较低层级的曲线进行旋转变换得到。相应地，每一个网格均对应其相邻较高层级的 4 个网格，并且这种对应方式是金字塔形的，即空间上具备一致性。编码为 code_i 的网格对应的 4 个网格分别的编码集合为 $\{4\times\text{code}_i+1, 4\times\text{code}_i+2, 4\times\text{code}_i+3, 4\times\text{code}_i+4\}$。以图 2 为例，编码为 1 的网格向其较低层级对应编码为 0 的网格，向其较高层级对应编码为 5、6、7、8 的网格。与多尺度的 Z 曲线结构相同，多尺度的 Hilbert 曲线各层级之间的对应关系同样是金字塔形状。

对于以多尺度 Hilbert 曲线作为编码的 n 层四叉树结构，共包含 $(4^n-1)/3$ 个不同尺度的网格编码。每一层内部网格对应的编码范围是 $[(4^{n-1}-1)/3, (4^n-4)/3]$，层级内部的各个网格之间的编码按照单一尺度的 Hilbert 曲线进行排列。按照从低到高层级的排列，每一个网格的编码均为 Base(n)+Hilbert(2^n)的形式，即之前所有较低层级的网格之和加上在本层的 Hilbert 编码[21]。在以多尺度的 Z 曲线和 Hilbert 曲线作为空间索引进行检索时，其编码连续性也有较大的区别。

3　分析方法

为了将多尺度的 Z 曲线和 Hilbert 曲线进行比较，本文设计了两组实验来验证两种曲线的连续性。1) 空间–编码方法：如果空间目标是相邻的，比较其对应的曲线编码是否相邻；2) 编码–空间方法：如果曲线编码是相邻的，比较其对应的空间目标是否相邻。

3.1　空间–编码方法

求得每一个网格邻域内的编码与之邻近的网格个数，将所有个数累加，得到的值越大，表明空间连续性越好。

C 为编码的集合，V 为网格的集合，定义两个映射 F_1 和 F_2。

$$F_1: C \to V, \tag{1}$$

$$F_2: V \to C, \tag{2}$$

F_1 与 F_2 均为满射，且两者互逆。

$$f_{\text{near\_}v}(v_i, v_j) = \begin{cases} 1, & \text{distance}(v_i, v_j) = 0, \\ 0, & \text{其他。} \end{cases} \tag{3}$$

$f_{\text{near\_}v}$ 的返回值为 1，说明两个网格是相邻的。

$$f_{\text{near\_}c}(\text{code}_i, \text{code}_j) = f_{\text{near\_}v}(F_1(\text{code}_i), F_1(\text{code}_j))。 \tag{4}$$

$f_{\text{near\_}c}$ 的返回值为 1，说明两个编码对应的网格是相邻的。

$$\text{TCount}_1 = \sum_{i=1}^{N} \sum_{\text{dt}_{\min}}^{\text{dt}_{\max}} \sum_{j=\text{code}_i-\text{dt}}^{\text{code}_i+\text{dt}} f_{\text{near\_}c}(\text{code}_i, \text{code}_j)。 \tag{5}$$

定义编码距离为两编码的数值之差的绝对值，dt 代表编码距离阈值，其中 $\text{dt}_{\max}$ 和 $\text{dt}_{\min}$ 是选取的编码距离阈值的上下限。

利用这种方法，定义的网格邻域有 9 邻域和 33 邻域两种。9 邻域包括：该网格的 1 个父网格，该网格同层级的前、后、左、右 4 个网格，该网格的 4 个子网格。33 邻域包括：该网格的 1 个父网格，该网格的父网格同层级的前、后、左、右、左前、左后、右前、右后 8 个网格，该网格同层级的前、后、左、右、左前、左后、右前、右后 8 个网格、该网格的 4 个子网格以及这 4 个子网格外围的 12 个网格。

以图 3 为例，在 Z 曲线采用的 Z 序编码中编码为 2 的网格的 9 邻域共包括：编码为 0, 1, 3, 9, 10, 11, 12 的 7 个网格(缺少两个是因为编码为 4 的网格在本层级的边缘位置，缺少两个相邻网格)。如果编码距离小于编码距离阈值，则表明两编码在数值上相近。仍以图 3 为例，当阈值为 3 时，编码为 4 的网格的 9 邻域中共有 0, 1, 3 这 3 个编码，均在编码距离阈值范围内。本文中编码阈值的上下限分别是 1 和 20。

3.2　编码–空间方法

用滑动窗口在最高层级的网格进行遍历，再比较每次与滑动窗口有交集的各个层级内网格曲线编码的连续段个数，将所有的个数累加之后，得到的值越小，表明空间连续性越好。

$$\text{TCount}_2 = \sum_{\text{window}_i} \text{IntervalNum}(\text{window}_i)。 \tag{6}$$

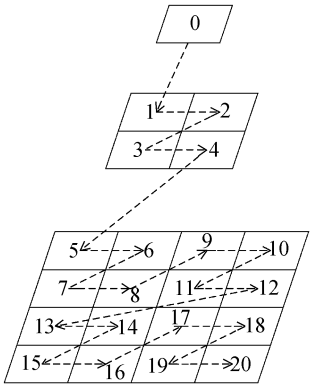

图 3 多尺度 Z 序曲线
Fig. 3 Multi-dimensional Z curve

以图 2 和 3 为例，当检索范围为 2×2，且在最高层级网格中对应的是图 3 中的 14, 12, 17, 18 或图 2 中的 12，11，13 和 14 时，那么第 1 层级的一个网格和第 2 层级的两个网格也都与之有交集。最终，图 3 中与该窗口有交集的网格的编码包括 0，2，4，11，12，17 和 18，图 2 中与该滑动窗口有交集的网格的编码包括 0, 2, 3, 11, 12, 13 和 14。图 3 中的连续段数为 5 个: 0, 2, 4, 11~12 和 17~18; 图 2 中的连续段数为 3 个: 0, 2~3 和 11~14。在这个示例中，多尺度 Hilbert 曲线的段数更少，连续性更强。

4 对比实验

本实验构造了层数为 3，4，5，6，7，8 和 9 的多尺度的 Z 曲线和 Hilbert 曲线，并分别利用空间–编码方法和编码–空间方法进行对比，不同尺度的面片的编码方式如前所述。

4.1 空间–编码方法实验

表 1 为 33 邻域和 9 邻域的多尺度 Z 曲线与 Hilbert 曲线相比较得到的总相邻个数 $TCount_1$ 之差。表 1 中“提高比例”是两模型 $TCount_1$ 之差与阈值取最大值时 Z 曲线总邻接之和的比值。由此可见，多尺度 Hilbert 曲线的连续性比多尺度 Z 曲线显著提高，并且随着层数的增加，提高的比例不断加大。此优势在 33 邻域中表现得比 9 邻域中更明显。

在层数较小时，由于有较多的网格处于整个四叉树的边缘位置(在最顶层或最底层，或某一层的边界位置)，因此整体结构的典型特征不明显。随着层数的增加，处于边缘位置的网格数量所占比例降低，Hilbert 曲线的优势更加显著地体现出来，提高比例也有收敛的趋势。

表 1 $TCount_1$ 的提高
Table 1 Increase of label $TCount_1$

层数	9 邻域	提高比例/%	33 邻域	提高比例/%
3	–8	–7.3	11	5.0
4	90	23.3	192	29.2
5	606	41.6	1054	47.7
6	2934	50.8	5108	60.4
7	12474	53.9	22026	65.8
8	50898	55.0	90304	67.5
9	204822	55.2	364118	68.1

4.2 编码–空间方法实验

表 2 为 2×2 窗口和 3×3 窗口的多尺度 Hilbert 曲线与 Z 曲线相比较得到的总段数与数量变化的百分比。表 2 表明，相对于 Z 曲线，多尺度 Hilbert 曲线更能增加一定空间范围内编码的连续性，但随着层数增加，这个优势有一定程度的减小，大窗口(3×3)的优势比小窗口(2×2)更明显。

层数的增加使得多尺度 Hilbert 曲线和 Z 曲线均由于编码的跨级跳跃而间断更多，这是不可避免的，并且层数越大，类似的影响越大。相对于 Z 曲线，Hilbert 曲线更大的优势在于单一层级内部，因此多尺度 Hilbert 曲线相对于 Z 曲线的优势被削弱了。窗口的增大会导致连续性的差异被扩大，与 Faloutsos 等[8]的实验结果一致。

5 结论

随着地球剖分的空间数据管理方法不断进步，空间索引的地位显得更加重要。作为空间数据索引，Z 曲线和 Hilbert 曲线广泛地应用于多类空间数据库中，但当前的空间填充曲线不适用于多尺度数据管理的需求。本文以四叉树结构为依据，对 Hilbert 曲线和 Z 曲线的效率进行对比。传统的 Z 曲线连续性较差，在进行空间数据检索时效率有一定程度的降低。本文提出的两种曲线(多尺度 Hilbert 曲线和多尺度 Z 曲线)在海量多尺度数据存在的条件下，能够适应数据分布情况的要求，保持较好的空间连续性。本文结果表明，从空间连续性的两种定义来看，多尺度 Hilbert 曲线均比 Z 曲线空间连续性更强。

表 2　$TCount_2$ 的降低

Table 2　Decrease of label $TCount_2$

层数	2×2 窗口			3×3 窗口		
	多尺度 Hilbert 曲线	多尺度 Z 曲线	降低比例/%	多尺度 Hilbert 曲线	多尺度 Z 曲线	降低比例/%
3	29	34	14.7	13	19	31.6
4	236	278	15.1	216	292	26.0
5	1391	1618	14.0	1512	1980	23.6
6	7122	8166	12.8	8220	10292	20.1
7	33941	38410	11.6	39928	49900	20.0
8	155416	173902	10.6	183204	224956	18.6
9	694811	770002	9.8	814832	985676	17.3

参考文献

[1] 金安, 程承旗. 基于全球剖分网格的空间数据编码方法. 测绘科学技术学报, 2013, 30(3): 284–287

[2] 程承旗, 任伏虎, 濮国梁, 等. 空间信息剖分组织导论. 北京: 科学出版社, 2012

[3] Finkel R A, Bentley J L. Quad trees a data structure for retrieval on composite keys. Acta Informatica, 1974, 4(1): 1–9

[4] Xu H. An approximate nearest neighbor query algorithm based on Hilbert curve // International Conference on Internet Computing & Information Services (ICICIS). Hong Kong, 2011: 514–517

[5] To Q C, Dang T K, Küng J. A Hilbert-based framework for preserving privacy in location-based services. International Journal of Intelligent Information and Database Systems, 2013, 7(2): 113–134

[6] 何雄. 空间数据库引擎关键技术研究[D]. 北京: 中国科学院计算技术研究所, 2006

[7] Butz A R. Convergence with Hilbert's space filling curve. Journal of Computer and System Sciences, 1969, 3(2): 128–146

[8] Faloutsos C, Roseman S. Fractals for secondary key retrieval // Proceedings of the Eighth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Philadelphia, 1989: 247–252

[9] 何雄, 方金云, 唐志敏. 基于 ORDB 的空间数据库的研究与实现. 计算机工程, 2005, 31(2): 42–44

[10] Christopher D L. Exploring spatial scale in geography. New York: Wiley & Sons, 2014

[11] Kitchin R. Big data and human geography. Opportunities, challenges and risks. Dialogues in Human Geography, 2013, 3(3): 262–267

[12] Schrack G, Stocco L. Generation of spatial orders and space-filling curves. IEEE Transactions on Image Processing, 2015, 24(6): 1791–1800

[13] Song Z, Roussopoulos N. Using Hilbert curve in image storing and retrieving. Information Systems, 2002, 27(8): 523–536

[14] Kamel I, Faloutsos C. Hilbert R-tree: an improved R-tree using fractals // Proc 20th Int Conf on Very Large Databases. Santiago, 1994: 500–509

[15] Böhm C, Berchtold S, Keim D A. Searching in high-dimensional spaces: index structures for improving the performance of multimedia databases. ACM Computing Surveys (CSUR), 2001, 33(3): 322–373

[16] Lu F, Zhou C H. A GIS spatial indexing approach based on hilbert ordering code. Journal of Computer Aided Design & Computer Graphics, 2001, 13(5): 424–429

[17] He Z, Owen A B. Extensible grids: uniform sampling on a space-filling curve. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016, 78(4): 917–931

[18] Breinholt G, Schierz C. Algorithm 781: generating Hilbert's space-filling curve by recursion. ACM Transactions on Mathematical Software (TOMS), 1998, 24(2): 184–189

[19] Kamata S, Eason R O, Bandou Y. A new algorithm for *N*-dimensional Hilbert scanning. IEEE Transactions on Image Processing, 1999, 8(7): 964–973

[20] Fracchia F D, Lindenmayer A, Prusinkiewicz P, et al. Synthesis of space-filling curves on the square grid [D]. Regina: University of Regina, 1989

[21] 周艳, 朱庆, 张叶廷. 基于 Hilbert 曲线层次分解的空间数据划分方法. 地理与地理信息科学, 2007, 23(4): 13–17