引文格式: 张小虎,钟耳顺,王少华,等. 多尺度空间格网数据的索引编码研究[J]. 测绘通报,2014(7):35-38. DOI: 10.13474/j.cnki.11-2246.2014.0220

多尺度空间格网数据的索引编码研究

张小虎[1,2],钟耳顺[1],王少华[1,2],张　珣[1,2]

(1. 中国科学院地理科学与资源研究所,北京 100101; 2. 中国科学院大学,北京 100049)

Research on Index and Code for Multi-scale Grid Data

ZHANG Xiaohu, ZHONG Ershun, WANG Shaohua, ZHANG Xun

摘要: 针对多尺度格网数据访问效率问题,设计了一种基于格网空间索引的多尺度格网数据索引编码体系。该索引编码有着较高的编码和访问效率,能较好地满足多尺度格网数据分析时的效率需求。

关键词: 多尺度; 格网数据; 索引编码

中图分类号: P208　　**文献标识码:** B　　**文章编号:** 0494-0911(2014)07-0035-04

一、引　言

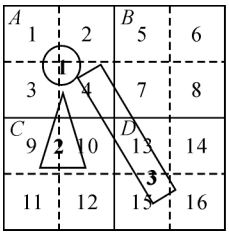
随着地理信息系统应用的日益深入,不同的应用越来越需要不同尺度的格网数据[1-3]。研究表明,规则格网数据能够较好地满足多尺度地理数据空间表达的需求,并一定程度上解决可塑面积单元问题(MAUP)对原始数据的影响[3]。在实际应用中,格网数据会根据具体的尺度规则,生成多个不同尺度数据集合 $L=\{l_1,l_2,\cdots,l_n\}$,并将 L 存储在空间文件或空间数据库中。然而,不管以栅格数据格式还是矢量数据格式,L 中各个子层$\{l_1,l_2,\cdots,l_n\}$概念上存在空间联系,而存储及索引上彼此独立。这种传统的"一库多版本技术"并未将多尺度格网数据的空间拓扑关系、尺度关系表现出来[4],不但降低了多尺度格网数据集的逻辑一致性,而且对后续的多尺度分析带来巨大的时间消耗。多尺度格网数据之间的运算操作涉及大规模的地理查询操作,直接依靠已有的空间查询和拓扑操作,且每次查询将直接影响数据运算效率。因此,必须建立一套有效的数据索引体系,减少格网数据的查询和拓扑操作,使得多尺度格网数据的多尺度分析得以实现。本文的研究目的就是为多尺度格网数据集 L 建立一种有效的索引方式,使得各子层之间的空间关系得以明确,并采用合理的编码将这种关系固定存储起来,使得多尺度格网的空间运算高效实施,各个尺度的格网数据可以高效访问其他尺度的相关格网数据。

二、格网索引及多尺度格网统计数据特征

空间数据索引技术并不是一个新兴的研究问题,其在空间数据库技术中得到了广泛的研究,并形成 3 类空间索引方法,即基于点区域划分的索引、基于面区域划分的索引和基于体区域划分的索引[5-6]。空间索引是采用一定的顺序在不同的空间划分区域内搜索查找地理实体,从而加快空间查询[7]。面区域划分的索引方法主要有 3 种,即格网索引、区域四叉树索引和 R 树索引。根据多尺度格网数据的空间特征,传统的格网索引比较合适用多尺度格网数据建立索引。

1. 格网索引

格网索引是一种基于空间填充曲线(space-filling curve)的索引方式[8]。其基本方法是将二维正方形平面划分成 $m\times n$ 的格网,并利用一种空间填充曲线建立起这些格网的一维索引(如图 1 所示)。典型的空间填充曲线有 Z-curve 和 Hilbert-curve,其对应的一维索引编码为 Morton 码和 Hilbert 码[9]。其中 Hibert 编码可以使得原二维的平面形状不必限制为正方形,以拓展格网索引的使用范围[10]。



索引1	
A	1,2,3
B	3
C	2,3
D	3

索引2			
1	1	10	2,3
2	1	13	3
3	1	15	3
4	1,2,3	16	3
7	3		
9	2		

图 1　基于 Z-curve 的二级格网索引

格网索引的性能依赖于格网和对象的大小,以及对象密度之间的关系[11]。对于空间密度变化大的数据,通常需要建立多层的索引格网,以保证性能

收稿日期: 2013-03-12
基金项目: 国家科技支撑计划(2011BAH06B03)
作者简介: 张小虎(1986—),男,江苏宝应人,博士生,研究方向为格网空间数据多尺度分析。

最优。在格网索引中,当用户进行空间查询时,首先计算被查对象所在的空间格网,使用格网快速定位到所选的空间对象。格网索引方法是对空间对象最为直观简单的索引方法,其对应的算法也比较简单,可以实现对象的快速目标查询,并且这是一种典型的以空间换时间的索引方法,数据冗余大,但并不影响格网索引的高效而带来的广泛应用。

需要指出的是任何索引对查询、访问的操作时间的优化都是相对的,不合适的索引不但不能加快查询速度,反而会使操作更为复杂耗时。因此,建立索引必须针对数据本身特征。本文将根据多尺度格网统计数据自身特征,建立符合多尺度格网数据特征的索引编码体系,满足多尺度格网数据访问的效率需求。

2. 多尺度格网数据的两种形式及索引编码需求

多尺度格网数据是一种在统一的数据管理系统下存储的不同尺度的空间数据,该数据索引有其特殊的需求,即所构建的索引能使格网数据实现不同尺度层数据的快速切换及各尺度对应格网单元快速互访问。这种索引不需要满足优化空间查询等基本地图操作需求,因为这些格网数据录入到空间数据库时,已有的空间数据库索引技术已经满足了这些基本要求。然而,多尺度格网数据多尺度分析时需要反复查询目标格网对应的其他尺度下的相应格网及其邻近格网。利用已有的空间索引技术需要反复采用空间查询及空间拓扑操作,极大地降低了空间分析的效率,尤其当格网数据的尺度规模和空间规模都非常大时,这种影响将更为明显。因此,需要一种建立一种专门的索引及其编码,满足多尺度格网数据多尺度分析的效率需求。

根据不同的应用需求,多尺度格网统计数据各层格网主要有两种不同的拓扑关系,根据这两种关系,多尺度格网统计数据主要表现成两种形式(如图 2 所示)。第 1 种形式为不同尺度的格网的尺度缩放因子成整数倍数关系,大尺度格网完整包含小尺度格网,不同层的数据完整覆盖,不交叉;第 2 种形式为不同尺度格网的尺度缩放因子不成整数倍数关系,大尺度的格网不能完整包含小尺度格网,不同层的数据存在交叉。在后续的多尺度空间分析时,这两种形式的数据相互访问时存在明显差异,主要表现在:

1) 第 1 种格网小尺度格网仅仅对应一个大尺度格网数据,各尺度间数据可以按照树结构进行相互访问,并可根据树形结构推测其邻接格网单元。

2) 第 2 种格网小尺度对应多个大尺度格网数据,各尺度间数据不能按照简单的空间关系建立树形关系,各尺度直接的相互访问比较复杂。

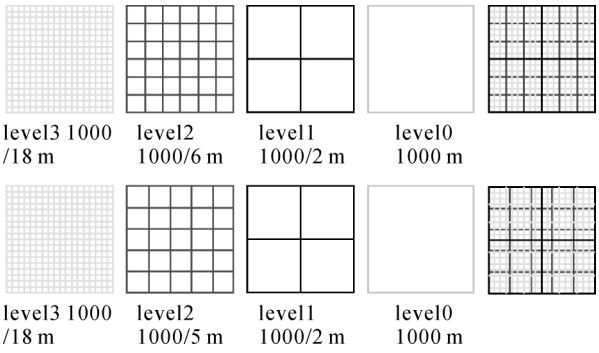


图 2 多尺度格网数据的两种表现形式

在实际应用中,根据不同的需求,格网尺度可能是人为指定的,但是往往是规则的(常用 1 km、500 m、250 m、100 m、50 m)。格网数据往往表现为上述第 1 种形式;或者是第 2 种形式中大部分数据呈现第 1 种数据形式,只有少数层缩放因子不符合整倍数关系(如 1 km、500 m、250 m、100 m、50 m 中只有 100 m 不符合第 1 种数据形式要求的整倍数关系)。

因此,多尺度格网索引及其编码的基本策略是在建立第 1 种形式格网数据索引编码的基础上,解决第 2 种形式数据的特殊问题。

三、多尺度格网数据的索引编码

1. 多尺度格网统计数据的多尺度编码

本文提出一种基于树形结构和格网空间索引的多尺度格网索引编码体系,解决了多尺度格网统计数据多尺度分析时的效率问题。这种索引编码较好地实现了多尺度数据多尺度特征,其基本方法为(如图 3 所示)。

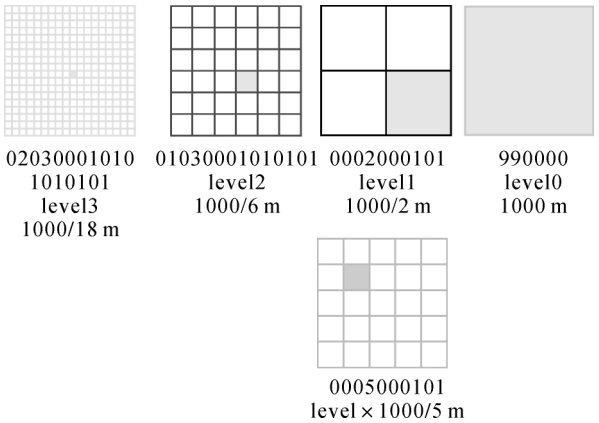


图 3 多尺度格网统计数据的索引编码示例

1) 编码规则: 格网编码分为 3 个部分,即尺度代码,尺度因子代码,格网位置代码。如“99” “02” “00” “00”,其中第 1 个数值是上级格网尺度代码,“99”表示无上一层尺度的格网; 第 2 个数值表示与上个尺度的缩放因子,“02”表示该尺度格网是上个尺度格网 1∶2 划分而来; 第 3 个数值是上个尺度对应格网的顺序编码; 第 4 个数值为该格网顺序编码。

2) 按顺序从大尺度分别建立下级格网编码,编码以十进制行列顺序编码。如 level 1,根据其行列值对其采用二维十进制编码,如第 1 行第 1 列赋值为“001001”,并记录该格网的上级格网及其尺度因子、对应格网编码,最终表示为“0002” “00” “001001”。

3) 如果尺度数据为第 2 种形式数据(如图 2 数据中有 level x) ,先建立其他尺度格网的索引编码。对 level x 数据,根据其尺度因子选择最临近的满足整倍数关系的上尺度(大尺度) 格网(level 0) ,根据该数据建立其格网编码。

4) 当加入新的尺度格网数据时,如果该尺度数据与原有数据的尺度缩放因子满足整倍数关系,需要根据步骤 1) —2) 重新建立索引编码; 当不满足需求时,根据步骤 3) 加入编码。

5) 当需要删除某个尺度的格网数据时,如果该尺度数据与原有数据的尺度缩放因子满足整倍数关系,删除该数据同时需要按步骤 1) —3) 重建索引编码; 当不满足关系时,直接删除该数据。

6) 按照上述编码,根据大尺度格网索引小尺度格网、编码从小到大的顺序写入索引文件,便于后续使用。

上述编码规则适用于任何情况下的多尺度格网数据。然而实际应用中,多尺度格网尺度规则的制定是有弹性的,可以根据需求建立等级尺度缩放的多尺度格网,这不但不影响实际多尺度分析的效果,而且索引编码的效率将大为增加。如采用等级为 2 的缩放因子,使得各尺度间满足四叉树的索引模式,进而提高后续分析的效率。

2. 多尺度格网数据索引编码应用效率分析

本文中多尺度格网数据索引体系的设计实际上是利用大尺度格网作为小尺度格网建立空间索引的格网参考,从而使得多尺度格网数据不同尺度间相互联系,并加快各尺度数据层之间的相互访问。没有这种索引机制的多尺度格网数据进行多尺度分析时需要额外的空间查询操作。本文通过对比多尺度格网数据建立索引前后的多尺度分析操作的时间及多尺度格网索引编码建立时间来分析多尺度格网统计数据索引编码的应用效率。

检验索引编码的编码效率的计算机环境为: CPU: Intel(R) Core(TM) 2 Quad CPUQ9550@ 2.83 GHz; 内存: 4.00 GB(1067 HZ) /3.37 GB 可用; 操作系统: Microsoft Windows 7 专业版 (32 位) ; 编程语言: python2.6; 地理数据操作基本库: arcpy; 数据格式: Esri shapefile 文件型空间数据格式。

试验为 4 个不同研究范围建立了 4 个不同尺度的格网。不同规模的格网统计数据索引编码时间如表 1、图 4 所示。

表 1　不同规模多尺度格网数据索引编码时间　　s

规模	尺度/km²		
	1	100	10 000
1 km×1 km	0.060 2	0.124 8	4.878 4
500 m×500 m	0.083 6	0.272 8	18.937 6
250 m×250 m	0.075 5	0.826 4	76.089 8
125 m×125 m	0.109 0	3.086 1	310.424 7

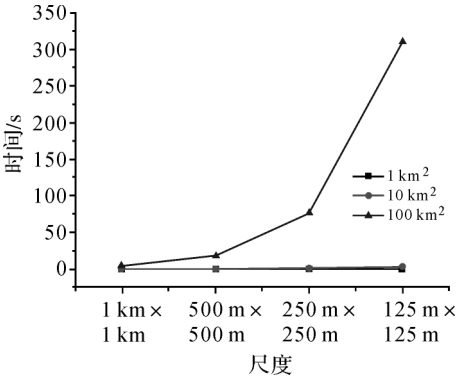


图 4　不同规模多尺度格网数据索引编码时间

对于大规模的格网编码索引编码效率能够满足需求,格网统计数据的索引编码效率与尺度、规模关系密切,同一规模下索引编码时间与尺度因子成对数关系。

同时,为了评价本文提出的多尺度格网索引对多尺度格网统计数据分析效率的改善效果,设计了统一的多尺度数据互访问操作,即通过相邻尺度两层格网的相互访问,计算所有尺度格网某个属性的总和。试验分别对表 1 中 10 km² 的原始格网、具有一般格网索引的格网(Esri ArcGIS shapefile 的空间索引) 和具有本文设计的多尺度格网索引的格网进行了上述操作,其耗时见表 2。原始格网的多尺度运算操作十分耗时,效率低下; 具有一般格网空间索引有效地提高了多尺度数据互访问的效率; 本文所设计的多尺度格网索引也较好地提高了多尺度数据

不同尺度间互访问的效率，并且比一般格网空间索引好，效率提高了 30%。

表 2 多尺度格网数据实验耗时 s

规模尺度	100 km²		
	原始格网	一般格网空间索引	多尺度格网索引
Level 0: 1 km×1 km	0.028 2	0.018 1	0.015 7
Level 1: 500 m×500 m	10.538 6	8.698 5	6.156 2
Level 2: 250 m×250 m	65.767 2	33.931 3	25.274 7
Level 3: 125 m×125 m	658.841 5	135.957 3	109.243 1

四、结论与讨论

本文针对多尺度格网设计了一种基于格网索引的索引编码体系。该方法有着较高的编码效率和访问效率，较好地解决了多尺度格网数据尺度间访问的效率问题。同时需要指出的是：①任何索引对查询、访问的操作时间的优化都是相对的，不合适的索引不但不能加快查询速度，反而会使得操作更为复杂耗时。因此，建立索引必须针对数据本身特征，这也是本文设计的索引比一般格网索引效率更优的根本原因和基础。②索引对原始数据访问的优化，并不一定降低空间查询、访问及运算的时间。根据索引选择合适的数据查询方法、制定优化的访问策略、设计高效的运算算法对多尺度格网数据的分析效率都是至关重要的。

参考文献：

[1] 狄琳. 建立新国家地理格网服务地理国情监测初探[J]. 测绘通报，2011(11)：1-2.

[2] 左伟，张桂兰，万必文，等. 中尺度生态评价研究中格网空间尺度的选择与确定[J]. 测绘学报，2003，32(3)：267-271.

[3] GEHLKE C K，BIEHL K. Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material[J]. Journal of the American Statistical Association，1934(29)：169-170.

[4] 王艳慧，李小娟，宫辉力. 地理要素多尺度表达的基本问题[J]. 中国科学 E 辑：技术科学，2006(36)：38-44.

[5] 阎超德，赵学胜. GIS 空间索引方法述评[J]. 地理与地理信息科学，2004(4)：23-28，39.

[6] 郑坤，朱良峰，吴信才，等. 3D GIS 空间索引技术研究[J]. 地理与地理信息科学，2006(4)：35-39.

[7] 吴敏君. GIS 空间索引技术的研究[D]. 镇江：江苏大学，2006.

[8] PEANO G. Sur Une Courbe，Qui Remplit Toute Une Aire Plane[J]. Mathematische Annalen，1890(36)：157-160.

[9] HILBERT D. Ueber Die Stetige Abbildung Einer Line Auf Ein Flächenstück[J]. Mathematische Annalen，1891(38)：459-460.

[10] HAMILTON C H，RAU-CHAPLIN A. Compact Hilbert Indices：Space-filling Curves for Domains with Unequal Side Lengths[J]. Information Processing Letters，2008(105)：155-163.

[11] LONGLEY P. Geographic Information Systems and Science(2nd ed)[M]. West Sussex：John Wiley & Sons Inc，2005：229-234.

[12] 杨族桥，郭庆胜，牛冀平，等. DEM 多尺度表达与地形结构线提取研究[J]. 测绘学报，2005，34(2)：134-137.