

一种高效的海量遥感栅格数据库的空间可视化检索算法

李 峰^{①②③} 付 琨^② 尤红建^② 吴一戎^② 刘 波^④

^①(中国科学院电子学研究所 北京 100080)

^②(中国科学院西安光学精密机械研究所 西安 710068)

^③(中国科学院研究生院 北京 100039)

^④(中国科学院高技术研究与发展局 北京 100864)

摘 要 该文针对利用 GIS 现有空间查询接口进行海量遥感栅格数据库空间可视化检索效率低下的问题。在深入研究海量遥感栅格数据库的空间可视化检索特点的基础上,提出了一种直接针对关系数据库(RDBMS)存储过程的高效的海量遥感栅格数据库复杂空间可视化检索算法,并对算法进行了多个级别的性能优化。该算法可直接应用于海量遥感栅格数据库基于多边形、椭圆和线穿越等复杂空间可视化查询的应用环境。实验结果说明该算法具有稳定性和普适性。

关键词 可视化检索, 空间查询, 查询优化, 存储过程

中图分类号: TP391

文献标识码: A

文章编号: 1009-5896(2006)08-1463-05

An Efficient Spatial Visual Retrieval Algorithm on Magnanimous Remote Sensing Raster Image Data Base

Li Feng^{①②③} Fu Kun^② You Hong-jian^② Wu Yi-rong^② Liu Bo^④

^①(Institute of Electronics, CAS, Beijing 100080, China)

^②(Xi'an Institute of Optics and Precision Mechanics, CAS, Xi'an 710068, China)

^③(Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)

^④(Bureau of High-Tech Research and Development, CAS, Beijing 100864, China)

Abstract It is inefficient to visually retrieve image on magnanimous remote sensing raster image data base by some GIS spatial query interface. An efficient spatial visual retrieval algorithm aimed at directly storage procedure of Relational Data Base Management System (RDBMS) is described and optimized on performance on many levels in this paper after the character of magnanimous remote sensing raster image database spatial visual retrieval is studied in depth. This algorithm is implemented in a real application environment of magnanimous remote sensing raster image data base based on the spatial polygon query, ellipse query and line through query. The results of experiment demonstrate that this algorithm is stable and adaptable.

Key words Visual retrieval, Spatial query, Query optimization, Stored procedures

1 引言

随着遥感(RS)、地理信息系统(GIS), 和全球定位系统(GPS)的发展, 它们之间的结合日益紧密, 遥感栅格影像数据成为 GIS 的重要数据来源。然而随着遥感影像数据的增长, 如对于一个高分辨率运行型卫星遥感信息处理系统而言, 随着系统的运行, 管理的遥感栅格影像数据可能达到百 TB 的存储量级。由于数据量巨大, 结合成本的考虑, 这些海量遥感栅格数据以数据库编目的方式由关系型数据库和文件系统共同管理。直接利用现有的 GIS 软件提供的空间检索接口很难对这些海量数据进行高效检索。因此, 研究高效的空间可视化检索就成为海量遥感影像数据库的一个研究重点。

空间可视化检索是空间数据库和海量遥感影像数据库应用中最重要的一种检索方式, 它是检索满足某种空间谓词

(如相交, 包含等)的所有空间对象或遥感影像。GIS系统中的线穿越查询、多边形包含查询、椭圆和圆包含查询, 空间几何目标拓扑关系查询等都属于此类查询。而对于遥感影像数据库系统而言, 空间信息是其应用的关键要素, 基于空间信息的影像检索也是其重要的应用模式。因此, 为了有效地提高空间可视化检索效率, 遥感栅格影像的海量存储与管理、矢量栅格一体化存储管理, 维护和管理空间数据的空间存取方法(Spatial Access Method, SAM)、空间查询代价估计模型^[1], 空间查询优化^[1], 空间扩展SQL与空间信息自然语言查询接口^[2]成为空间数据库和遥感影像数据库共同面临的研究重点。

通常一颗稳定的高分辨率遥感卫星在其生命周期内, 下传的原始数据加上标准产品, 数据存储容量达到百 TB 级, 编目数据记录达到千万条量级。在这种案例条件下, 根本无法利用一种商用的空间数据管理引擎, 如 ESIR 公司的

ArcSDE, 完成遥感栅格数据的存储管理, 也无法高效地对海量遥感栅格数据库进行空间可视化检索。这是因为通过 ArcSDE 进行空间可视化检索时, ArcSDE 提供的空间查询接口首先检索出满足某种空间拓扑关系(圆包含、多边形包含、线穿越等)的结果记录集, 再将这些结果整合, 最后作为客户端查询结果记录集返回客户端。该过程出现了多个数据库表的连接操作、检索结果记录集的重组操作, 在表数据记录巨大的情况下, 造成了检索效率的下降。若不通过 ArcSDE 提供的空间检索接口, 而是直接通过 RDBMS 进行空间检索时, 对于矩形检索容易实现, 而对于多边形, 圆等形状的检索均无法通过构造一条 SQL 语句来实现。

为此, 论文基于计算几何分析方法提出了一种海量遥感栅格数据库复杂的空间可视化检索算法。算法通过 RDBMS 的存储过程实现, 既解决了无法通过构造一条 SQL 语句实现复杂空间可视化检索问题, 同时又减少了查询数据的网络传输频度, 极大地提高了检索的效率。基于该算法思路的空间可视化检索软件已经应用在某大型运行卫星地面应用系统。

2 算法设计与实现

空间可视化检索主要是利用可视化的检索方式解决以下两类问题: (1) 空间几何对象的定位问题; (2) 空间几何范围的查询问题。其中, 空间几何对象的定位问题就是指用来表示空间几何对象的点, 矩形, 多边形或圆的定位查询; 空间几何范围的查询包括线穿越查询、多边形或圆的包含查询, 邻近查询等。

多边形分为凸多边形与任意多边形, 这里所说的任意多边形是指简单多边形, 即边之间除顶点外不相交的多边形, 隐含多边形每个顶点的度数为 2 这一条件。根据计算几何的理论可知, 任意简单多边形都存在三角剖分, 而三角形可以看作是特殊的凸多边形, 因此任意多边形的检索问题可以转换为剖分后的凸多边形的检索问题, 进而我们提出了一种通用的凸多边形包含查询算法。

2.1 算法分析

假设有任何凸多边形 S 和任一目标点 $T(x_0, y_0)$, 如图 1 所示。在 S 上任取相邻两点 $P(x_i, y_i)$ 和 $Q(x_{i+1}, y_{i+1})$, 有两个向量 $\mathbf{V} = \mathbf{PQ}$ 和向量 $\mathbf{V}' = \mathbf{PT}$, 我们可以得出这样的结论:

(1) 若按顺时针方向遍历凸多边形所有的点, 向量 $\mathbf{V}$ 和 $\mathbf{V}'$ 之间的夹角 θ 满足 $0^\circ \leq \theta \leq 180^\circ$, 即目标点 T 始终在向量 $\mathbf{V}$ 的右方或在向量上, 则有点 T 在凸多边形 S 内, 且有不等式 $(y_{i+1} - y_i)(x_0 - x_i) - (y_0 - y_i)(x_{i+1} - x_i) \geq 0$ 成立。

(2) 若按逆时针方向遍历凸多边形所有的点, 向量 $\mathbf{V}'$ 和 $\mathbf{V}$ 之间的夹角 θ 满足 $0^\circ \leq \theta \leq 180^\circ$, 即目标点 T 始终在向量 $\mathbf{V}$ 的左方或在向量上, 则有点 T 在凸多边形 S 内, 且有不等式 $(y_{i+1} - y_i)(x_0 - x_i) - (y_0 - y_i)(x_{i+1} - x_i) \geq 0$ 成立。

证明(1) 对于向量 $\mathbf{V}$ 有直线方程:

$$y = \frac{(y_{i+1} - y_i)(x - x_i)}{(x_{i+1} - x_i)} + y_i \quad (1)$$

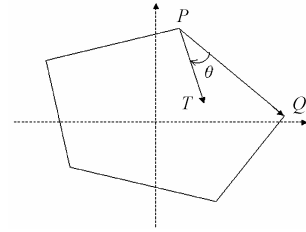


图1 点在多边形内示意图

Fig.1 Illustration of point in polygon

若 $x_{i+1} > x_i$, 根据以上的假设目标点 T 始终在向量 $\mathbf{V}$ 的右方或在向量上, 可以得到以下的不等式:

$$\frac{(y_{i+1} - y_i)(x_0 - x_i)}{(x_{i+1} - x_i)} \geq y_0 - y_i \quad (2)$$

又因为 $x_{i+1} > x_i$ 成立, 则可以得到以下不等式:

$$(y_{i+1} - y_i)(x_0 - x_i) - (y_0 - y_i)(x_{i+1} - x_i) \geq 0 \quad (3)$$

若 $x_{i+1} < x_i$, 根据以上的假设可得

$$\frac{(y_{i+1} - y_i)(x_0 - x_i)}{(x_{i+1} - x_i)} \leq y_0 - y_i \quad (4)$$

又因为 $x_{i+1} < x_i$ 成立, 则有式(3)成立。

若 $x_{i+1} = x_i$, $y_{i+1} > y_i$, 根据目标点 T 始终在向量 $\mathbf{V}$ 的右方或在向量上的假设, 可得 $x_0 \geq x_{i+1} = x_i$, 则有式(3)成立。

若 $x_{i+1} = x_i$, $y_{i+1} < y_i$, 根据目标点 T 始终在向量 $\mathbf{V}$ 的右方或在向量上的假设, 可得 $x_0 \leq x_{i+1} = x_i$, 则有式(3)成立。

同理(2)得证。

证毕

2.2 算法伪代码实现

初始化标记变量 b_Flag 为 TRUE;

FOR 凸多边形 S 的每一个点

按顺时针方向或逆时针方向依次获取该点 P 坐标 (x_i, y_i) 和相邻点 Q 坐标 (x_{i+1}, y_{i+1}) ;

取得目标点 T 坐标 (x, y) ;

计算 P 点和 Q 点的坐标之差分别为: $d_x = x_{i+1} - x_i$,

$d_y = y_{i+1} - y_i$;

计算 T 点和 Q 点的坐标之差分别为: $d'_x = x - x_i$,

$d'_y = y - y_i$;

IF $(d_y * d'_x - d_x * d'_y < 0)$ THEN

点不在凸多边形内, 退出循环, $b_Flag =$

FALSE;

END IF

END FOR

IF $(b_Flag = \text{TRUE})$ THEN

目标点 T 在凸多边形 S 内;

END IF

2.3 算法复杂度分析

假设完成一次乘法运算的开销为常数 c_1 , 完成一次减法运算的开销为 c_2 , 则一次循环内的运算总开销为 $c = 2c_1 + 5c_2$, 该算法的最佳情况是第一次循环就可以判断出目标点不在多边形内, 这种情况总的运算开销为 c , 最坏的情况是

目标点在多边形内,这种情况总的运算开销为 $M \times c$, 其中 M 是多边形的边数。因此,该算法的平均运算开销为 $M \times c / 2$ 。

假设数据库表的记录数为 N , 从数据库表中依次获取每一条记录, 根据该记录中目标点的空间信息, 判断该目标点是否在多边形范围内, 该过程总的平均开销为 $N \times M \times c / 2$, 由于在用户完成多边形的输入后, M 为定值, 无法通过程序来改变它, 因此要提高算法的运行效率, 就需要减小 N 值, 或在 N 很大的情况下减小 c 值。我们接下来对算法优化的工作就有对 N 值进行调整的方面。

3 空间可视化检索的优化策略

对于海量遥感影像数据库系统而言, 性能是影响系统运行的关键因素。而决定系统运行性能的因素是来自于多个方面的, 如硬件存储设备、主机处理能力、数据库运行环境、前端应用程序开发等, 而这些方面的优化又相互影响, 并会造成系统价格和维护代价的提高, 需要达到一种平衡。因此, 对于系统性能的调优是一个层次状的不断调整的过程。其中, 硬件存储设备的优化包括采用高速光纤硬盘或 SCSI 硬盘代替 IDE 硬盘、实现磁盘 RAID 技术等; 主机处理能力的优化包括扩展 CPU 数量, 扩充内存容量等; 数据库运行环境的优化包括升级为并行处理模式, 如 ORACLE 和 DB2 都有相应的并行数据库版本, 创建索引, 对巨型表采用分区策略等。本文主要针对基于 ORACLE 数据库运行环境的调优, 使得调优之后海量遥感影像数据库空间检索的效率得到提高, 主要包括如下几种优化策略。

(1)范围限定策略 对于大数据量的巨型数据表而言, 全表扫描查询是一项极为耗时的查询操作, 尽量避免全表扫描可以提高查询效率。在空间可视化检索中, 用户提供查询的多边形, 系统完成对包含于多边形中的目标的可视化检索, 其中检索的基表中所有空间对象覆盖了特定坐标系下的很宽的经纬度范围, 而用户提供的查询多边形覆盖的经纬度范围有限。本文首先计算该多边形的最小边界矩形(Minimum Bounding Rectangle, MBR)^[3,4], 接着在该最小边界矩形所包含的空间对象的范围内应用我们的检索算法, 大大地提高了检索效率。其中, 最小边界矩形中包含的数据表记录是全表记录的子集。设表数据记录数为 N , 多边形的最小外接矩形包含的记录数为 M , 则 M/N 的取值在 $[0, 1]$ 之间, 若 M/N 的值越小, 查询的性能改进越明显。

(2)索引策略 索引是 RDBMS 的一种重要模式对象, 它是一种供数据库在数据表中快速查询记录的数据结构, 建立了表中某列或多列数据与物理存储位置的关联, 从而提高了数据检索的效率。然而, 索引并不是越多越好, 只有根据实际应用建立适当的索引, 才能提高整个系统的性能。因为索引需要物理存储空间的开销, 在数据插入、更新、删除之后会有重建索引的开销。

空间索引就是存储空间数据时根据空间对象的位置和

形状或空间对象之间的某种空间关系, 按一定顺序排列的一种数据结构, 其中包含空间对象的概要信息, 如对象的标识, 外接矩形及指向空间对象实体的指针等。 R 树^[5,6]是一种广泛应用于空间数据存取的索引结构, 它是 B 树在多维空间上的扩展, 也是一种可动态调整的动态平衡树。为了改善 R 树的性能, 还提出了 R^+ 树、 R^* 树、压缩 R 树等, 形成了一个 R 树族。同时, 广泛应用于空间查询的还有四叉树结构和 $k-d$ 树结构^[7]。其中 ORACLE Spatial 就是通过对定义的空间对象建立 R 树和四叉树索引, 从而提高空间检索的效率。本文所提出的算法是在 RDBMS 的基础上, 不涉及空间数据引擎或中间件, 所以我们采用的索引策略就是针对 SQL 语句中条件查询经常用到的字段而建立的索引。

(3) 多边形三角剖分策略 根据计算几何的方法, 任意多边形可以被剖分成若干三角形, 因此基于多边形的包含查询就可以转换为多个三角形的包含查询, 前者的查询结果是后者查询结果的合并。这样可以利用并行处理同时创建多个数据库连接会话, 多个会话完成基于三角形的查询, 最终将这些并行查询的结果合并返回给用户。并行查询需要提供支持并行查询的硬件、操作系统, 数据库系统级别设置等条件, 如多 CPU, 磁盘 RAID 技术, 多线程应用程序等。

(4) 分页策略 对于海量遥感影像数据库应用而言, 空间可视化检索经常会出现检索结果是一个很大的数据集的情况, 而这会对数据的 I/O 和网络传输带来很大的压力, 同时还会影响系统查询的效率。我们提出的算法采用了数据分页技术, 即对查询结果数据集根据与用户的交互进行分割, 每次返回结果数据集的一部分, 待用户提出翻页的请求时再从查询结果数据集中获取下一页的数据, 从而提高的查询效率。

(5) 网格化策略 首先对完成空间查询的空间坐标系进行逻辑上的网格划分, 将查询条件限制在与多边形相交的那些网格中, 再利用以上算法查询出包含于多边形的目标点。设多边形的面积 S , 空间坐标系的经度范围为 $[-180^\circ, 180^\circ]$, 纬度范围为 $[-90^\circ, 90^\circ]$, 面积为 A , S/A 越小, 过滤的数据越多, 查询的效率就越高。如图 2 所示。

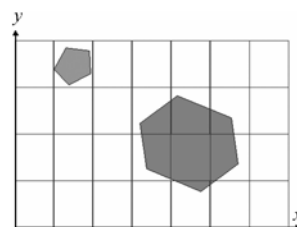


图2 网格化查询优化示意图

Fig.2 Illustration of grid query optimization

4 实验结果与分析

我们将基于 ArcSDE 查询接口实现空间可视化检索的方法和基于本文提出的算法进行了比较。其中, 前者客户端利

用 VB 开发可视化检索程序, 用户加载全球地图并在地图上勾画多边形, 圆或直线作为可视化检索条件, 并将它们的空间参考信息作为 ArcSDE 所提供的空间查询接口的参数, 由 ArcSDE 检索出满足包含、穿越等空间拓扑关系的栅格对象, 以 ADO 记录集的形式返回给客户端可视化检索程序, 后台数据库是 Oracle 9i; 后者客户端利用 VC.NET 可视化检索程序, 用户加载全球地图并在地图上勾画多边形、圆或直线作为可视化检索条件, 并将它们的空间参考信息作为 PL/SQL 实现的存储过程的参数, 由本文提出的算法检索出包含、穿越等空间拓扑关系的栅格对象, 以 ADO 记录集的形式返回给客户端可视化检索程序, 后台数据库是 Oracle 9i。

根据以上流程, 测试了两种方法的检索效率, 并对实验结果进行了详细的分析, 实验结果表明本文提出的算法对于海量遥感栅格数据库空间可视化检索性能更优。

一个高分辨率运行型卫星遥感信息处理系统中, 典型的空间可视化检索模式是在一个全球范围的地图底图上, 用户通过可视化检索界面提供的工具构造多边形, 圆或直线, 系统实现从海量遥感栅格数据库中检索出包含于多边形内或圆内的遥感栅格对象, 或检索出线穿越的遥感栅格对象。我们可以将这些多边形, 圆, 直线看作是用户提供的空间可视化检索条件, 如图 3 所示。

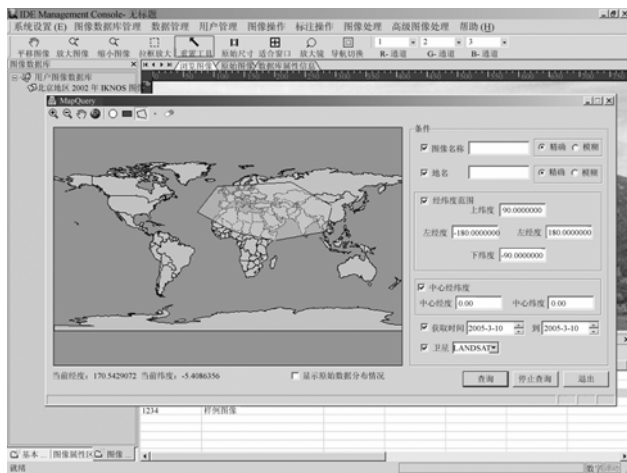


图 3 遥感图像数据库空间可视化检索系统界面

Fig.3 The interface of spatial visual query system of remote sensing image database

该系统中的遥感栅格对象覆盖了全球范围, 数量达到了千万条记录量级。使用基于 ArcSDE 的空间可视化检索接口, 随着满足检索条件的返回结果记录集数量增加, 检索的时间也随之增加, 实验结果如表 1 所示。

使用本文的算法, 检索时间不随着满足检索条件的结果记录集数量增加而增加, 同时也不会随着多变形的边数增加或多边形的覆盖范围增加而增加, 而是基本维持在某个固定值上, 实验结果如表 2, 表 3 和表 4 所示。

当栅格数据表的记录数增加时, 检索时间也随之增加, 基本满足正比例关系, 如表 5 所示。

表 1 基于 ArcSDE 的检索时间统计表

Tab.1 Statistics of query time based on ArcSDE

总记录数	1,000,000				
返回结果记录数	2,000	4,000	10,000	20,000	40,000
检索时间(s)	125	153	219	442	573

表 2 基于本文算法的三角形检索时间统计表

Tab.2 Statistics of triangle query time based on algorithm in this paper

总记录数	1,000,000				
返回结果记录数	2,000	4,000	10,000	20,000	40,000
检索时间(s)	118.73	119.21	120.34	121.71	124.02

表 3 基于本文算法的矩形检索时间统计表

Tab.3 Statistics of rectangle query time based on algorithm in this paper

总记录数	1,000,000				
返回结果记录数	2,000	4,000	10,000	20,000	40,000
检索时间(s)	119.17	120.01	120.98	122.32	125.61

表 4 基于本文算法的五边形检索时间统计表

Table 4 Statistics of pentagon query time based on algorithm in this paper

总记录数	1,000,000				
返回结果记录数	2,000	4,000	10,000	20,000	40,000
检索时间(s)	122.73	123.98	124.15	125.47	126.08

表 5 基于本文算法的检索时间统计表

Tab.5 Statistic of query time based on algorithm in this paper

总记录数	6,000	10,000	30,000	60,000	10,000
检索时间(s)	1.09	2.04	8.9	15.23	24.95

由于基于 ArcSDE 的空间可视化检索需要完成多个属性表与栅格数据表的连接操作, 当需要返回的结果记录集越多, 连接表的数量就越大, 因此随着返回的结果记录数的增加, 检索时间就会增加。本文算法针对栅格数据表遍历所有的记录, 判断是否满足空间可视化检索的条件, 因此当栅格数据表中的总记录数固定时, 检索时间基本是个定值, 而随着总记录的增加, 检索时间也会增加。两种检索方式的比较如图 4 所示, 本文算法的多种多边形检索方式的比较如图 5 所示, 本文算法总记录数与检索时间的关系如图 6 所示。其中, 我们实验系统的软、硬件环境配置如下:

主机: Pentium 4 (2.4G, 1G);

数据库: Oracle 9i Release 2;

空间数据引擎: ArcSDE 9.0;

实验平台 1 工具: VB, ADO, ArcEngine, ArcSDE, 基于 ArcSDE 查询接口的实验平台;

实验平台 2 工具: TOAD, PL/SQL, ORACLE 9i, 即本

文算法的实验平台;

总记录数: 1,000,000。

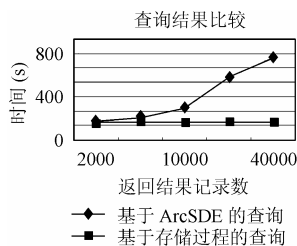


图 4 两种检索结果比较图

Fig.4 Comparing chart of two kinds of query result

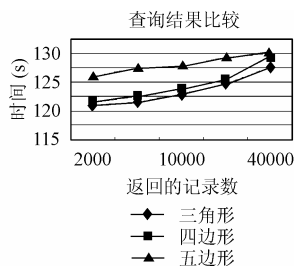


图 5 本文算法多边形检索比较

Fig.5 Comparing chart of polygon query based on the algorithm in this paper

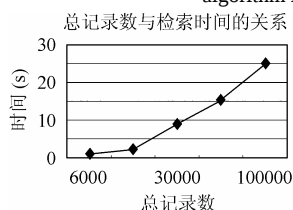


图 6 检索时间与本文算法总记录数关系图

Fig.6 The relationship chart between query time and record based on the algorithm in this paper

5 结束语

本文通过传统的关系型数据库存储过程实现了海量栅格数据库空间可视化检索的功能,并对算法进行了多级优化,在千万条级数据记录的情况下,与 ArcSDE 提供的空间检索接口相比,极大地提高了检索效率,解决了无法通过简单的 SQL 语句实现复杂的空间可视化检索的问题,对空间可视化检索方面的研究有着重要的参考价值。实验结果说明算法实现的效率具有稳定性和普适性,解决了海量遥感栅格数据库复杂的空间可视化检索的难题。对凸多边形之外的任意

多边形空间可视化检索的研究将是我们下一步的工作重点。

参 考 文 献

- [1] 方裕, 楚放. 空间查询优化. 中国图像图形学报, 2001, 6(4): 307 - 314.
- [2] 刘宇, 朱仲英, 施颂椒. 派生索引空间连接查询. 上海交通大学学报, 2001, 35(2): 275 - 278.
- [3] 李立言, 秦小麟. 空间数据库中连接运算的处理与优化. 中国图像图形学报, 2003, 8(7): 732 - 737.
- [4] 马林兵, 龚健雅. 空间信息自然语言查询接口的研究与应用. 武汉大学学报信息科学版, 2003, 28(3): 301 - 305.
- [5] Guttman A. R-tree: a dynamic index structure for spatial search. Proc ACM SIGMOD Int Conf on Management of Data[C]. Boston, MA, 1984: 47 - 57.
- [6] Brinkhoff T, Kriegel H P, Seeger B. Efficient processing of spatial joins using R-tree[A]. In: Processings of the 1993 Association for Computing Machinery Special Interest Group, International Conference on Management of Data[C]. Washington DC, USA, 1993: 237 - 246.
- [7] Orenstein J A. Spatial query processing in an object-oriented database system. Proc. ACM SIGMOD Int. Conf. on Management of Data[C]. Washington DC, 1986: 181 - 190.

李 峰: 男, 1977 年生, 博士生, 研究方向为图像数据库、基于内容的图像检索、遥感图像处理.

付 琨: 男, 1974 年生, 博士后, 从事遥感图像处理、情报解译等研究.

尤红建: 男, 1969 年生, 博士, 副研究员, 从事遥感图像处理.

吴一戎: 男, 1963 年生, 研究员, 博士生导师, 从事合成孔径雷达数据处理.

刘 波: 男, 1957 年生, 研究员, 博士生导师, 从事图像处理研究.