文章编号:1001 - 2486(2004)01 - 0024 - 05

基于参数化直方图的三路互相交连接选择性估计*

张 巨,肖予钦,熊 伟,景 宁
(国防科技大学电子科学与工程学院,湖南 长沙 410073)

摘 要:估计空间算子的选择性是设计空间查询优化器的关键技术之一。选择性估计不仅能以非常小的代价给出空间算子结果集大小的近似估计,而且也可以直接用于某些仅需要近似结果的空间查询和空间分析(如数据集间的相关性评价等)处理。互相交连接是一类常见而且具有特殊性质的多路空间连接。基于对命题"两两相交的多个矩形一定有一个公共的相交区域,而且这个区域也是矩形"的证明,提出了一种可以用于三路互相交连接选择性估计的参数化直方图方法,还通过多组比较实验证明了该方法的有效性和适应性。

关键词:选择性估计;互相交连接;参数化直方图
中图分类号:TP392 文献标识码:A

Selectivity Estimation of 3-Way Clique Intersect Joins Based on Parameterized Histograms

ZHANG Ju, XIAO Yu-qin, XIONG Wei, JING Ning
(College of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Selectivity estimation is crucial in a query optimizer for choosing a good execution plan for a given query. Selectivity estimates of spatial joins can be used as responses to the specialized user queries that seek approximate figures. Spatial join selectivity can also be used for evaluating the correlation between datasets. With the proof that mutually overlapped rectangles are sharing a common region which is also a rectangle, this paper gives a selectivity estimation technique for 3-way clique intersect joins. The efficiency of our algorithm has been validated by us using synthesized and real-world datasets.

Key words: selectivity estimation; clique intersect join; parameterized histogram

空间数据库管理系统需要提供一套功能完备的空间算子,这些算子中最重要的当属空间连接,因为空间连接不仅经常被调用,而且它还是实现复杂查询谓词的基础。空间连接得到广泛重视的另一个原因是其高昂的计算和 I/O 代价。估计空间算子的选择性是设计空间查询优化器的关键技术之一。

1 相关工作

空间算子的选择性估计技术大体上可以分为以下四个类型:基于模型函数的参数化方法、基于多项式函数的曲线拟合法、采样法和基于直方图的非参数化方法。其中,直方图方法是数据库领域中应用得最广泛的类型,因为它能够适应任何形式的数据分布,而且其存储空间需求和估计误差也比较小。根据数据分布空间划分方法的不同,可以将直方图方法进一步细分为:等宽、等高、等深、MaxDiff 和 V - 最优等类型。

目前,已经提出了多种基于直方图的空间选择算子选择性估计方法,如欧拉直方图[1]、积聚密度直方图[2]、MinSkew 直方图[3] 和 SQ - 直方图[4] 等。利用欧拉直方图方法,Sun 等[5] 进一步给出了能够同时支持相交、包含和被包含连接选择性的估计方法。基于离散余弦变换[6] 和小波变换[7] 的直方图压缩算法还可以进一步降低存储空间需求。在空间连接的选择性估计方面,Faloutsos 等提出了一种利用相关分形维度上的幂律实现距离连接选择性估计的方法[8]。文献[9] 给出了一种建立在全域均匀分布假设基

* 收稿日期:2003 - 09 - 20
基金项目:国家高技术研究发展计划基金资助项目(2002AA131010,2002AA134010,2002AA134012,2002AA134020)
作者简介:张巨(1974—),男,博士生。

础上的距离连接选择性估计算式。文献[10]给出了一种通过对相交区域顶点总数的估计实现相交连接选择性估计的几何直方图(GH)方法。Mamoulis 等[11]和 C. Sun 等提出的相交连接选择性估计方法还可以用于带空间窗口选择条件的情况,他们的方法分别是以全域均匀分布假设和欧拉直方图为基础的。D. Papadias 等[12]给出了基于全域均匀分布假设的无圈、环和互相交连接的选择性计算公式,但这个公式并不适用于只需要很少统计量的选择性估计。

2　互相交连接的特性

一个建立在数据集 DS_1, DS_2,…, DS_n 上的多路空间连接对应于一个连接图 $Q(V, E)$,其中,$V = \{DS_1, DS_2, \cdots, DS_n\}$,$E = \{e_{i,j}\}$,$e_{i,j} = (DS_i, DS_j)$ 表示 DS_i 与 DS_j 间的连接条件[9,12]。J. K. Min 等[13]根据连接图特征将空间连接划分为四个子类:团(完全图)连接、有圈连接、环连接和无圈(树形)连接。对应于一个完全连接图的团相交连接的目的是要在 DS_1, DS_2,…, DS_n 上找出所有符合两两相交条件的多元组 $< s_1, s_2, \cdots, s_n >$,其中,$s_i$ 是 DS_i 中的一个对象。团相交连接通常又称为互相交连接。Papadias 等[12]发现了互相交连接的一个典型特征:两两相交的矩形一定有一个公共的矩形区域。下面,给出这个特征的严格证明。

引理 1　如果 $R_1 \cap R_2 \cap \cdots \cap R_n = A$ 且 $A \neq \Phi$,则有:① 对 $\forall i,j \in \{1,\cdots,n\}$,有 $A \subseteq \bigcap_i R_j$ 并且②A 是一个左下角坐标和右上角坐标分别为 $(\max_{1\leqslant i\leqslant n}\{x_i^{(1)}\}, \max_{1\leqslant i\leqslant n}\{y_i^{(1)}\})$ 和 $(\min_{1\leqslant i\leqslant n}\{x_i^{(2)}\}, \min_{1\leqslant i\leqslant n}\{y_i^{(2)}\})$ 的矩形。其中,$(x_i^{(1)}, y_i^{(1)})$ 和 $(x_i^{(2)}, y_i^{(2)})$ 表示 R_i 的左下角坐标和右上角坐标。

证明　结论①是很直观的,下面给出结论②的一个证明过程。令 $(x_i^{(1)}, y_i^{(1)})$ 和 $(x_i^{(2)}, y_i^{(2)})$ 表示 R_i 的左下角和右上角坐标,则有 $x_i^{(1)} < x_i^{(2)}$ 且 $y_i^{(1)} < y_i^{(2)}$。因为 $R_i \cap R_j \supseteq A$ 且 $A \neq \Phi$,所以有 $x_i^{(1)} \leqslant x_j^{(2)}$、$x_j^{(1)} \leqslant x_i^{(2)}$、$y_i^{(1)} \leqslant y_j^{(2)}$ 和 $y_j^{(1)} \leqslant y_i^{(2)}$,即对 $\forall i,j \in \{1,\cdots,n\}$ 且 $i \neq j$ 的情况,都有 $x_i^{(1)} \leqslant x_j^{(2)}$ 且 $y_i^{(1)} \leqslant y_j^{(2)}$ 成立。因此,通过令 $x^{(1)} = \max_{1\leqslant i\leqslant n}\{x_i^{(1)}\}$、$y^{(1)} = \max_{1\leqslant i\leqslant n}\{y_i^{(1)}\}$、$x^{(2)} = \max_{1\leqslant i\leqslant n}\{x_i^{(2)}\}$ 和 $y^{(2)} = \max_{1\leqslant i\leqslant n}\{y_i^{(2)}\}$ 构造出的矩形区域 R 一定是被矩形 R_1, R_2, …, R_n 所共享且最大的区域。□

定理 1　$R_1 \cap R_2 \cap \cdots \cap R_n \neq \Phi$(其中,$R_1$, R_2, …, R_n 表示 n 个矩形),当且仅当在 $\forall i,j \in \{1,\cdots,n\}$ 且 $i \neq j$ 的情况下有 $R_i \cap R_j \neq \Phi$ 成立。

证明　J.K. Min 等[13]已经给出了 $n = 3$ 情况的证明。假设定理 1 在有 n 个数据集($n \geqslant 3$)参与连接的情况下成立,对于有 $n+1$ 个数据集参与连接的情况,只要证明 $R_1 \cap R_2 \cap \cdots \cap R_{n+1} \neq \Phi \Leftrightarrow R_1 \cap R_2 \cap \cdots \cap R_n \neq \Phi$ 且对 $\forall i \in \{1,\cdots,n\}$ 有 $R_i \cap R_{n+1} \neq \Phi$ 即可。令 $R = R_1 \cap R_2 \cap \cdots \cap R_{n-1}$,则 $R_1 \cap R_2 \cap \cdots \cap R_{n+1} \neq \Phi \Leftrightarrow R \cap R_n \neq \Phi$、$R_n \cap R_{n+1} \neq \Phi$ 且 $R \cap R_{n+1} \neq \Phi$。由于 $R \cap R_{n+1} \neq \Phi$ 隐含了 $\forall i \in \{1,\cdots,n-1\}$ 有 $R_i \cap R_{n+1} \neq \Phi$,故 $R \cap R_n \cap R_{n+1} \neq \Phi \Leftrightarrow \forall i,j \in \{1,\cdots,n+1\}$,$i \neq j$,$R_i \cap R_j \neq \Phi$ 成立。□

3　三路互相交连接选择性估计

两个矩形的相交区域也是一个矩形,N. An 等[10]将这个矩形相交区间的四个顶点称为交叉点(intersecting points)。交叉点有两种情况:(1)一个矩形的顶点落入另一 1 个矩形;(2)一个矩形的水平边与另一个矩形的垂直边相交。如果能够估算出在两个数据集上执行相交连接所产生的交叉点的数目(IP),那么 IP 被 4 除的结果即可以作为空间相交连接结果集大小的估计。为了估计两个数据集上矩形对象的交叉点数,N.An 等[14]提出了几何直方图(GH)方法。GH 首先将数据集 DS_k 归一化的数据空间划分为若干个形状相同的网格(桶);随后,在每个网格 $C(i,j)$ 中记录如下参数:$C_k(i;j)$:落入网格 $C(i,j)$ 中的 MBR 顶点的数目;$O_k(i;j)$:落入 $C(i, j)$ 的 MBR 区域的面积之和与 $C(i, j)$ 的面积的比值;$H_k(i;j)$:落入 $C(i, j)$ 的 MBR 的水平线长度之和与 $C(i,j)$ 的宽度的比值;$V_k(i;j)$:落入 $C(i, j)$ 的 MBR 的垂直线长度之和与 $C(i, j)$ 的高度的比值。N. An 等[14]证明,在每个网格内部提供均匀分布假设的前提下,IP 的数学期望为:

$$IP = \sum [C_1(i,j) \times O_2(i,j) + C_2(i,j) \times O_1(i,j) + H_1(i,j) \times V_2(i,j) + H_2(i,j) \times V_1(i,j)] \tag{1}$$

根据前面给出的定理1,3路互相交连接的选择性也可以通过估计满足连接条件的三元组 $< s_1, s_2, s_3 >$ 所共同占有的矩形区域的顶点数获得。如果 $A = R_1 \cap R_2 \cap R_3$,则 A 的4个顶点或者是某个矩形同时落入另外两个矩形的顶点,或者是某个矩形的水平边与另一矩形的垂直边落入第三个矩形的交点。我们区分下面两种情况以估计 DS_1, DS_2, DS_3 间互相交连接的结果集大小:

- 交叉点是来自某个数据集中某个矩形的顶点

我们用图1所示的矩形 a,b 和 c 分别表示来自数据集 DS_1,DS_2 和 DS_3 的矩形对象,用 CW 和 CH 表示当前网格 $C(i, j)$ 的宽和高。图中的灰色区域 I_b 和斜线区域 I_c 代表矩形 b 和 c 落入网格 $C(i, j)$ 的部分,用 hb,hc 和 vb,vc 分别表示 I_b 和 I_c 的宽和高,则顶点 P_a 同时落入 I_b 和 I_c 的概率可以用 $p[p\_a\ in(I\_b \cap I\_c)]$ 表示。令 $(I\_b \cap I\_c) = A$,根据均匀分布假设,有 $p[P\_a\ in(I\_b \cap I\_c)] = \frac{area(A)}{CW \times CH}$,即顶点 P_a 同时落入 I_b 和 I_c 的概率问题转化为对 A 的面积估计。

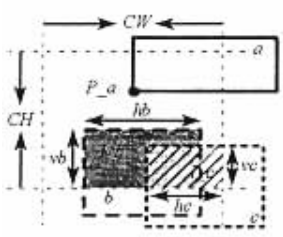

图1 几何直方图中的统计量

Fig.1 Description of parameters in geometric histogram

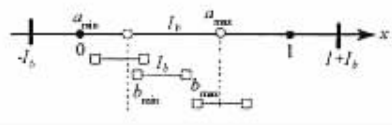

图2 [0,1] 空间上两条线段相交的情况

Fig.2 Intersect cases between segments in [0, 1] interval

考虑一维单位空间 U 的情况(图2),U 中的线段 I_a,I_b 相交区域 $I_a \cap I_b$ 长度* 的期望可以用 $E[\max\{0,(\min\{a_{\max}, b_{\max}\} - \max\{a_{\min}, b_{\min}\})\}]$ 表示,其中 $a_{\min}/a_{\max}$, $b_{\min}/b_{\max}$ 分别表示 I_a,I_b 的左/右端点坐标。假设 $I_a < I_b$,且令 I_b 为区间 $[-I_b, 1 + I_b]$ 上的均匀分布,则 $I_a \cap I_b$ 在 $[-I_b, 1 + I_b]$ 上的数学期望为 $E_{[-I_b,1+I_b]}[I_a \cap I_b] = \left(\int_0^{I_b} x\mathrm{d}x + \int_0^{I_a - I_b} I_b\mathrm{d}x + \int_0^{I_b} x\mathrm{d}x\right) \Big/ (1 + I_b) = I_a \times I_b/(1 + I_b)$。

直观地:

$$E_{[-I_b,1+I_b]}[I_a \cap I_b] = E_{[0,1]}[I_a \cap I_b] \times \frac{1 - I_b}{1 + I_b} + E_{[-I_b,0]}[(I_a \cap I_b) \mid b_{\min} < 0] \times \frac{I_b}{1 + I_b} + E_{[1,1+I_b]}[(I_a \cap I_b) \mid b_{\max} > 1] \times \frac{I_b}{1 + I_b} \tag{2}$$

因为 $E_{[-I_b,0]}[(I_a \cap I_b) \mid b_{\min} < 0] = E_{[1,1+I_b]}[(I_a \cap I_b) \mid b_{\max} > 1]$,上式可重写为:

$$E_{[0,1]}[I_a \cap I_b] = (I_a \times I_b - 2I_b \times E_{[-I_b,0]}[(I_a \cap I_b) \mid b_{\min} < 0])/(1 - I_b) \tag{2'}$$

(1) 当 $I_a + I_b < 1$ 时,$E_{[-I_b,0]}[(I_a \cap I_b) \mid b_{\min} < 0] = \int_0^{I_b} \frac{\int_0^x y\mathrm{d}y}{1 - I_a}\mathrm{d}x \Big/ I_b = \frac{I_b^2}{6(1 - I_a)}$。从而,$E_{[0,1]}[I_a \cap I_b] = \left[I_a \times I_b - \frac{I_b^3}{3(1 - I_a)}\right] \Big/ (1 - I_b)$。又因为 $\frac{I_b^3}{3(1 - I_a)} < \frac{I_b^2}{3} < \frac{I_b \times I_a}{3}$,所以有 $E_{[0,1]}[I_a \cap I_b] \approx I_a \times I_b/(1 - I_b)$。

(2) 当 $I_a + I_b \geqslant 1$ 时,$E_{[-I_b,0]}[(I_a \cap I_b) \mid b_{\min} < 0] = \left[\int_0^{1-I_a} \frac{\int_0^x y\mathrm{d}y}{1 - I_a}\mathrm{d}x + \int_{-I_a}^{I_b} \frac{\int_0^{1-I_a} (x - y)\mathrm{d}y}{1 - I_a}\mathrm{d}x\right] \Big/ I_b$

* 在没有歧义的情况下,线段 a 和 a 的长度都用 I_a 表示。

$= \frac{(1-I_a)^2}{6I_b} + \frac{1}{2}(I_a + I_b - 1)$。忽略$\frac{(1-I_a)^2}{6I_b}$对$\mathrm{E}_{[0,1]}[I_a \cap I_b]$的影响后,有$\mathrm{E}_{[0,1]}[I_a \cap I_b] \approx I_b$。

综合上述两种情况,可以得出下面的估计公式:

$$\mathrm{E}_{[0,1]}[I_a \cap I_b] = \begin{cases} \min\{I_a, I_b\} & I_a + I_b \geqslant 1 \\ I_a \times I_b/(1 - \min\{I_a, I_b\}) & I_a + I_b < 1 \end{cases} \tag{3}$$

假设当前直方图网格$C(i, j)$中有来自DS_2中矩形的水平边$h_{2,1}, h_{2,2}, \cdots, h_{2,m}$和来自$DS_3$的水平边$h_{3,1}, h_{3,2}, \cdots, h_{3,n}$,则$\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{E}_{[0,1]}[h_{2,i} \cap h_{3,j}]$是$C(i, j)$中$DS_2$和$DS_3$的水平边相交区域长度总和的期望值。令$C(i, j)$中$DS_2$和$DS_3$的水平边平均长度分别为$I_2$和$I_3$,即$I_2 = \sum_{i=1}^{m} h_{2,i}/m$,$I_3 = \sum_{j=1}^{n} h_{3,j}/n$;且令$\hat{I}_2 = \sum_{i}[h_{2,i}/(1-h_{2,i})]/m$,$\hat{I}_3 = \sum_{j}[h_{3,j}/(1-h_{3,j})]/n$。则$\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{E}_{[0,1]}[h_{2,i} \cap h_{3,j}]$可以进一步简化为:

$$\begin{cases} mnI_k & I_2 + I_1 \geqslant 1 \text{ 且 } I_k \leqslant I_l;k,l \in \{2,3\} \\ mnI_l \times \hat{I}_k & I_2 + I_1 < 1 \text{ 且 } I_k \leqslant I_l;k,l \in [2,3] \end{cases} \tag{4}$$

考虑到A在x和y轴上投影的分布是独立的,有$\mathrm{E}[area(A)] = \mathrm{E}[A_x] \times \mathrm{E}[A_y]$。

综上所述,只要在每个数据集所对应的直方图网格中记录落入其中的水平边总数、垂直边总数、水平边长度平均值、垂直边长度平均值以及水平边长和垂直边长对应于算子$x/(1-x)$的平均值,就可以实现"交叉点是来自某个数据集中某个矩形的顶点"情况的点数估计。

• 交叉点是来自不同数据集的两个矩形的水平边与垂直边的交点

在二维空间$CW \times CH$中,一个长度为v的垂直线与一个长度为h的水平线相交叉的概率为$\frac{h \times v}{CH \times CV}$[14],其交叉点进一步落入来自第三个数据集中某个矩形c的概率为$\frac{h \times v}{CH \times CV} \times area(c)$。因而,当前网格$C(i, j)$中"交叉点是来自不同数据集的两个矩形的水平边与垂直边的交点"情况的点数估计值为$\sum_{1 \leqslant i,j \leqslant 3}\left[(H_i \times V_j + H_j \times V_i) \times \prod_{1 \leqslant k \leqslant n; k \neq i, k \neq j} A_k\right]$,其中,$H_i$表示$DS_i$中的矩形落在当前网格$C(i, j)$的水平边长和,$V_i$为垂直边长和,$A_i$为面积和。为此,还需要为每个直方图网格追加一个"面积和"参数。在直方图的所有网格上对上述两种情况的点数估计值求和,并将总和除以4,即可得到三路互相交连接的选择性估计值。

4　实验结果分析

我们通过两组数据集对本文提出的选择性估计方法予以评估。第一组来自Census TIGER®2000的真实数据,我们选择了华盛顿特区的公路、街区和典型地物作为验证选择性的三组对象集(如图3所示)。三个数据集的对象总数分别为15 143、5 666和611。图4是不同直方图粒度下选择性估计值与真值间的比较实验结果。第二组是合成的矩形对象总数均为1 000且边长均值为0.15的三组归一化空间上均匀分布的正方形数据集。图5给出了第二组数据在不同直方图粒度下选择性估计值与真值间的比较实验结果。

实验结果显示:① 我们提出的参数化直方图方法对数据集的空间分布具有很好的适应能力;② 过高的直方图粒度并不能给选择性估计的结果精度带来太多好处。需要声明的一点是,所讨论的选择性估计仅仅考虑了空间连接的过滤步骤,这个步骤所产生结果集是真实结果集的超集。

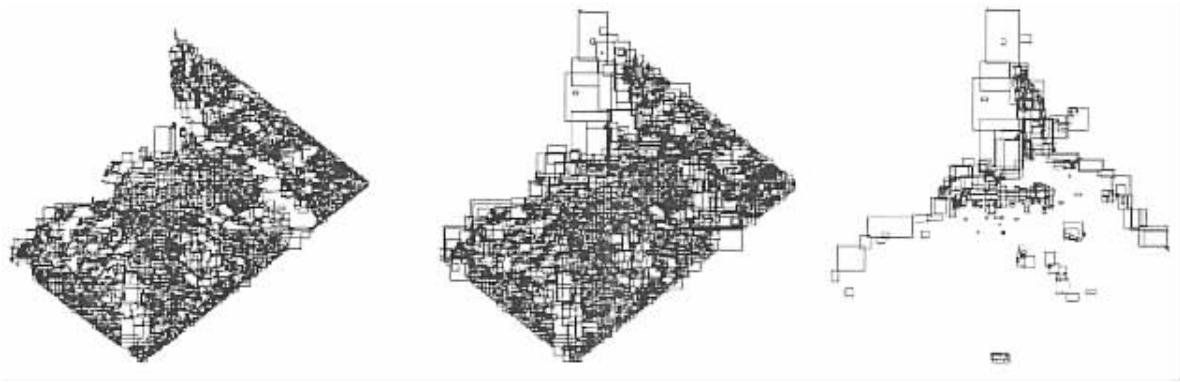

图3 来自 Census TIGER®2000 的华盛顿特区公路、街区和典型地物对象集

Fig.3 Three typical geo-sets of Washington D.C. comes from Census TIGER®2000

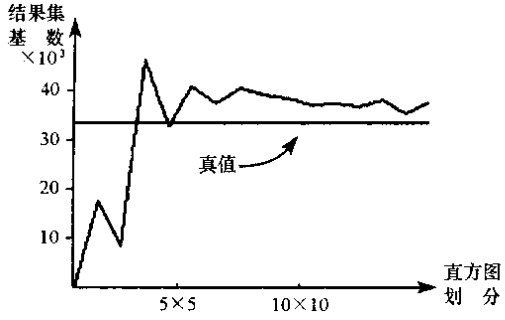

图4 第一组数据的选择性估计比较实验结果

Fig.4 Estimation-capability tests with real-world datasets

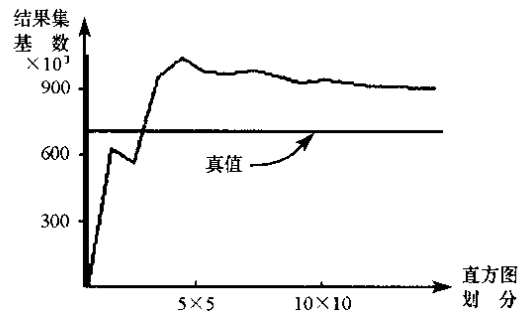

图5 第二组数据的选择性估计比较实验结果

Fig.5 Estimation-capability tests with simulated datasets

参考文献：

[1] Beigel R,et al. The Geometry of Browsing[C]. In Proc. of Latin American Symp. on Theoretical Informatics, 1998.

[2] Jin J,et al. Analyzing Range Queries on Spatial Data[C]. In Proc. of ICDE, 2000.

[3] Acharya S,et al. Selectivity Estimation in Spatial Databases[C]. In Proc. of ACM SIGMOD, 1999.

[4] Aboulnaga A,et al. Accurate Estimation of the Cost of Spatial Selections[C]. In Proc. of ICDE, 2000.

[5] Sun C, et al. Exploring Spatial Datasets with Histograms[R]. Technical Report, University of California, Santa Barbara, 2001.

[6] Lee J H,et al. Multi-dimensional Selectivity Estimation Using Compressed Histogram Information[C]. In Proc. of ACM SIGMOD, 1999.

[7] Wang M,et al. Wavelet-based Cost Estimation for Spatial Queries[C]. In the Proc. of the 7th SSTD, 2001.

[8] Faloutsos C,et al. Spatial Join Selectivity Using Power Laws[C]. In Proc. of SIGMOD, 2000.

[9] **张巨. 基于空间规划鉴别网络的空间关系约束关键技术研究[D]. 国防科技大学,**2003.

[10] An N,Sivasubramaniam A,et al. Selectivity Estimation for Spatial Joins[C]. In Proc. of ICDE, 2000.

[11] Mamoulis N,et al. Selectivity Estimation of Complex Spatial Queries[C].In Proc. of SSTD, 2001.

[12] Papadias D,et al. Processing and Optimization of Multi-way Spatial Join Using R-tree[C]. In Proc. of PODS, 1999.

[13] Min J K,et al. The Multi-way Spatial Join Selectivity for the Ring Join Graph[R]. Tech. Report of Department of Computer Science, KAIST, 2002.

[14] An N. Accessing Spatial Information in Resource-constrained and Resource-rich Environments[D]. Ph. D Thesis, CS, the Pennsylvania State University, 2001.