基于直方图的空间查询选择率估计研究

朱焰炉[1,2] 程昌秀[1] 陈荣国[1] 颜 勋[1]
(中国科学院地理科学与资源研究所 北京 100101)[1] (中国科学院研究生院 北京 100049)[2]

摘 要 空间查询优化是空间数据库中的关键问题之一,以查询代价估算为基础的查询优化技术是提高查询效率的一种重要方法,而估算代价的主要问题是估算查询结果(选择率)的大小。针对空间数据库中最常用的两种查询——空间选择和空间连接,阐述了几种主要用于查询选择率估计的直方图算法,并对各算法的优缺点做了分析,最后对空间查询选择率估计的研究方向进行了展望。

关键词 空间查询,直方图,选择率估计,空间查询优化
中图法分类号 TP311.131,TP208 **文献标识码** A

Selectivity Estimation for Spatial Query Based on Histogram

ZHU Yan-lu[1,2] CHENG Chang-xiu[1] CHEN Rong-guo[1] YAN Xun[1]
(Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China)[1]
(Graduate University of Chinese Academy of Sciences, Beijing 100049, China)[2]

Abstract Spatial query optimization is one of the key topic in spatial database. Query optimization technology based on query cost estimation is an important method to improve the efficiency of queries. But the key problem of query cost estimation is to estimate the size of query results(i. e. selectivity). This paper focused on the two queries operations: spatial selection and spatial join, which are most commonly used in spatial database. The paper expatiated some histogram algorithms for selectivity estimation of spatial queries, and analyzed their advantages and disadvantages. In the end of this paper, we discussed the future research directions of the selectivity estimation for spatial queries.

Keywords Spatial query, Histogram, Selectivity estimation, Spatial query optimization

1 引言

近年来,空间数据库在许多应用中变得越来越重要,如地理信息系统(GIS)、计算机辅助设计(CAD)、图像处理、多媒体系统等[1,2]。由于空间查询能够处理空间数据库中的多维数据和满足广泛的应用类型,因此对空间查询的研究近年来颇受关注。这是因为空间数据量庞大,数据结构复杂,操作代价昂贵,所以空间查询的优化势必成为空间应用的难点和突破点[3]。而关系数据库中的查询优化代价模型并不能完全适用于空间数据,空间数据库的查询代价估计模型尚处于初步研究阶段。优化空间查询能提高空间数据库的性能,对空间数据库的应用具有重要意义。

空间选择(Spatial Selection)和空间连接(Spatial Join)是空间查询中最重要也是代价(I/O和CPU代价)比较昂贵的两种操作[4,5],有关它们的查询优化引起了广泛关注。但在空间查询优化领域,至今还没有提出一套完整的优化系统设计方案[6]。目前,基于查询代价估算的代价模型是最常用的一种查询优化方法[7]。估算代价的主要问题是估算查询结果(选择率)的大小[8]。

关系数据库系统中用于估算选择率的算法主要有以下3种:采样法(Sampling)[9-11]、参数法(Parametric technique)[12-14]和直方图(Histogram)[15-17]。采样法是在原数据集上随机选择一些样本,对样本进行查询,得到该样本的查询结果大小,由此来估计在原始数据集上的查询结果大小[11]。其缺点在于:(1)它一般只对均匀分布的数据对象有效;对于非均匀分布的数据,要想得到较高的估算精度,需要足够多的样本数;(2)它必须在查询处理时进行,时间开销大[8,18]。参数法是用代数多项式或其他含参数分布(如均匀、正态、泊松和Z分布)等近似地表示数据的实际分布[8]。参数法的主要缺点在于:很难找到用于描述数据实际分布的函数,而估算结果又依赖于所用函数,故参数法估算结果往往不够准确。直方图是许多商用关系数据库系统(DB2, Informix, Ingres, Sybase)中最常用的一种估算查询结果大小的方法[8]。在空间查询选择率估计众多算法中,直方图也是最常用的方法[1,2,20,22,23,31]。

本文比较系统地研究了空间数据库中用于空间查询选择率估计的各类算法,尤其是空间直方图算法,并对其优缺点进行了分析,最后对基于直方图的空间查询选择率估计算法进

到稿日期:2010-01-27 返修日期:2010-04-26 本文受中科院知识创新工程重要方向项目(kzcx2-yw-304),国家863计划项目(2007AA120401,2007BAH16B03),所自主创新项目(09V90220ZZ)资助。

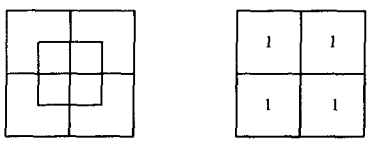
朱焰炉 男,硕士生,主要研究方向为空间查询优化,E-mail: zhuyl@lreis.ac.cn;程昌秀 女,副研究员,主要研究方向为空间数据库等技术;
陈荣国 男,研究员,主要研究方向为GIS&RS、空间数据库;颜 勋 男,博士生,主要研究方向为空间查询优化。

行了总结和展望。

2 空间数据库中查询选择率估计算法

传统关系数据库中的直方图算法不能适用于空间数据。原因在于:空间数据对象在大小和形状上是有区别的,而关系数据库中选择率估计与空间对象的大小和形状基本上是没有关系的,即传统的用于关系数据库的查询方法因不能保持空间数据对象的空间近似性而不适用于空间查询[19,14]。

建立传统直方图的方法如下:对数据空间(空间数据对象分布的整个空间范围称作数据空间)划分格网,给每个格网分配一个记录单元;对于每个空间数据对象,如果它与一个格网相交,就将该记录单元里的数值增加 1。该方法的一个严重缺陷是对线和多边形数据存在重复计数问题。例如,图 1(a)中间的矩形区域代表一个跨越了 4 个单元格的空间数据对象,(b)是其对应的直方图。可以看出,单个对象被重复计算 4 次,估计值超过 300%。点对象要么落入格网内,要么就在格网外,它不存在重复计数问题。因此,传统直方图只能用于点状空间数据,不能用于线状或多边形等其他类型空间数据。



(a) 一个空间数据对象　(b) 与 (a) 对应的传统直方图

图 1　传统直方图的重复计数问题

为了对空间数据的查询选择率进行有效的估计,国内外研究者提出了空间直方图算法。空间直方图的基本思想是:采用某种策略将数据空间划分为数个子空间,一个记录单元对应一个子空间;在记录单元中统计落在其对应子空间内的空间数据对象的数目;用相应的计算公式对这些统计值进行计算,得到查询结果集大小的估算值。这些记录单元称为桶,桶的集合称为直方图[7]。

根据使用技术的不同,空间直方图可以分为两大类:(1)基于数据分割技术的直方图;(2)基于单元格密度技术的直方图[20]。

2.1 基于数据分割技术的空间直方图

基于数据分割技术的空间直方图是依据空间数据特点来进行查询选择率估计的直方图方法。MinSkew 直方图[14]和 SQ 直方图[21]均属于此类。它们将相似的空间数据对象分到一个桶里,根据查询窗口来估计具有相离(disjoint)和相交(intersection)拓扑关系空间数据对象的数目。

2.1.1 MinSkew 直方图

Acharya 等人于 1999 年提出了基于空间二次划分(binary space partitionings,BSPs)的 MinSkew 直方图。MinSkew 直方图的创建过程是:首先,直方图中只有一个桶,它对应于整个数据空间,然后按照最大程度减少空间数据倾斜(spatial skew,即每个桶内空间数据对象的数目不均匀)的原则从直方图中选出计数值最大的桶,将其对应的子空间划分成两个子空间,并把该桶分裂成两个子桶。重复这一过程,直到桶的数目达到指定的数目为止[7]。MinSkew 由于是空间数据集的密度直方图,因此能有效地将多边形数据转换为点数据。

MinSkew 的主要缺点是:如果一个空间数据对象跨越多个直方图的桶,MinSkew 算法会对每个桶内都计数,这样就造成了重复计数问题[31]。此外,如何把一个子空间划分为更小的子空间(是二等分还是其他的划分方式)是一个复杂的问题[7,31]。

2.1.2 SQ 直方图

Aboulnaga 等人于 2000 年提出了 SQ 直方图(Structural Quadtree Histogram)[21],即四叉树直方图。SQ 直方图是为多边形数据集提出的,其基本思想是根据空间数据对象的最小外包矩形(MBR),将相似的空间对象划分到一个桶中。每个桶中存储空间数据对象的数目、平均宽度和平均高度以及桶的边界信息,在每个桶内假设空间对象是均匀分布的。它首先根据四叉树来构建具有层次结构的桶,空间数据对象根据它们的中心位置和面积分配给相应的桶,将相似的桶合并起来以达到桶数目的要求。

SQ 直方图的缺点在于:建立直方图时,多边形数据往往会跨越几个桶,存在重复计数问题;当同一子空间内的数据特征(即空间数据对象的大小、顶点数目等)差异太大时,SQ 直方图的性能并不理想[7];SQ 直方图所用到的四叉树结构、空间数据对象在桶内均匀分布的假设等在实际应用中并不多见。

2.2 基于单元格密度技术的空间直方图

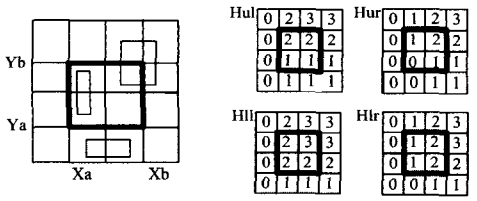
基于单元格密度技术的空间直方图是指通过对整个数据空间划分格网单元来进行查询选择率估计的方法。目前主要有 CD 直方图[1]、Euler 及其扩展直方图[19,22,23]、MP 直方图[24]、PH 直方图[25]和 GH 直方图[18]。其中,CD 与 Euler 直方图是为了解决空间选择查询的选择率估计而设计的;MP,PH 和 GH 直方图是针对空间连接查询的选择率估计而提出来的。

空间选择是指在一个空间数据集上给定一个查询窗口,查找与之相交的空间数据对象。其中,空间数据对象包括点、线、多边形等空间数据。空间选择也称为窗口查询(Window Query)或范围查询(Range Query)。空间连接是指两个数据集基于一个空间谓词进行连接操作[26],主要用来根据空间属性组合两个或多个数据集的空间数据对象。

2.2.1 CD(Cumulative Density)直方图

Jin 等人于 2000 年提出了 CD 直方图(Cumulative Density Histogram)[1]。CD 直方图将数据空间划分为大小相等的格网单元,用 4 个子直方图分别与空间数据对象 MBR 的 4 个角点对应,子直方图中的一个桶存储落入该桶中对应角点的数目。

在图 2 中,假设所有的空间数据对象均是用 MBR 近似表示的。建立 4 个直方图 H_{ll},H_{lr},H_{ul},H_{ur},每个直方图的大小是 N,且每个桶对应一个格网单元。H_{ll} 中的一个桶记录了落入该桶中空间数据对象的左下角顶点数目;H_{lr},H_{ul},H_{ur} 分别表示右下角、左上角、右上角顶点的数目。为了提高查询效率,所有直方图都是累积的,即一个桶 $H(i,j)$ 存储了区域 $(0,0,i,j)$ 中顶点的数目。对于一个查询窗口 (X_a,Y_a,X_b,Y_b),与窗口相交的空间对象数目可以按式(1)计算。这样,图 2 中计算结果为 3−0−1+0=2,即与查询窗口相交的空间数据对象数目为 2。



(a) 空间数据对象与查询窗口　(b) 与(a)对应CD直方图

图 2　CD 直方图

$$N_{\text{intersection}} = H_{ll}(X_b, Y_b) - H_{lr}(X_a - 1, Y_b) - H_{ul}(X_b, Y_a - 1) + H_{ur}(X_a - 1, Y_a - 1) \quad (1)$$

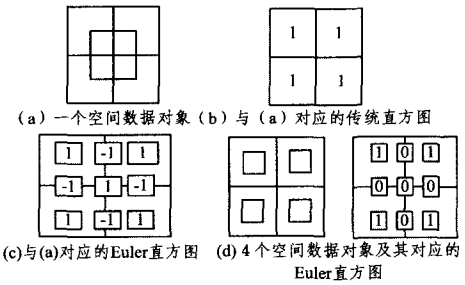
CD 直方图能精确地返回与查询窗口相交的空间数据对象数目，能够避免重复计数问题。

CD 直方图的缺点：它由于需要为每个格网存储 4 个变量，因此需要一定的存储空间，而且需要一些时间来计算累积信息。

2.2.2　Euler 直方图及其扩展

(1)Euler 直方图

为了解决重复计数问题，Beigel Richard，Tanin Egemen 于 1998 年提出了基于欧拉原理的直方图，故被人称为 Euler 直方图[23]。一般的直方图算法只为格网单元分配桶，而欧拉直方图不仅为格网单元(grid cell)分配桶，而且为格网的边(edge)和顶点(vertex)分配桶。图 3(a)对应的 Euler 直方图为图 3(c)。



(a) 一个空间数据对象　(b) 与(a)对应的传统直方图

(c)与(a)对应的Euler直方图　(d) 4 个空间数据对象及其对应的Euler直方图

图 3　传统直方图与欧拉直方图

在二维空间中，Euler 直方图的创建过程如下[20]：

首先将包含整个空间数据对象的数据空间分为 $n_1 \times n_2$ 个相等的格网单元，直方图就是建立在 $n_1 \times n_2$ 格网单元上的。

在 $n_1 \times n_2$ 格网上建立欧拉直方图，需要用 $(2n_1-1) \times (2n_2-1)$ 个桶来保存信息，它们分别与查询窗口内部的格网顶点、格网边、格网单元对应。各个桶内信息计算如下：

- 如果一个空间对象与格网单元相交，则桶中格网单元对应的数值增加 1；
- 如果一个空间对象包含一个格网顶点，则桶中格网顶点对应的数值增加 1；
- 如果一条格网边与一个空间对象相交，则桶中格网边对应的数值减少 1。

对于一个查询窗口 Q，它的选择率可以用式(2)来计算：

$$\text{Selectivity}(Q) = \sum_{k=0}^{d} (-1)^k F_k(Q) \quad (2)$$

式中，$F_k(Q)$ 表示 Q 内的 k 维 $(0 \leqslant k \leqslant d)$ 空间数据对象的数目，如 0 维是点、1 维是线、2 维是多边形。

图 3(c)中的选择率为

$(-1)^0 \times 1 + (-1)^1 \times (-1-1-1-1) + (-1)^2 \times (1+1+1+1) = 1$

通过比较图 3(b)和图 3(c)对应的直方图可以看出，Euler 直方图能避免重复计数问题。为了便于比较，一个拥有 4 个小空间数据对象的数据集及其对应的欧拉直方图如图 3(d)所示。可以看出，对格网的边和顶点都建立桶，能区分较大空间数据对象跨越多个格网单元和较小空间数据对象位于每个格网内的情况，而这一点是传统直方图做不到的。

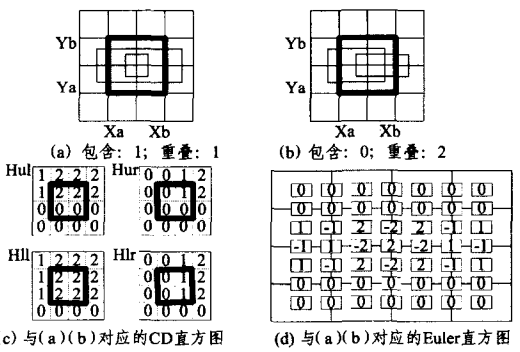
(2)闭欧拉直方图

文献[19]中指出，当空间数据对象的边界与查询窗口分割线部分重合时，用欧拉公式计算会出现统计错误，即欧拉直方图的边界问题。产生错误的原因是：欧拉直方图对格网顶点、格网边和格网单元分别计数，由于格网单元不包含边界、边不包含端点，因此在对格网单元和格网边对应的桶计数时这些桶是开的，即边界问题。为避免该问题，陈海珠等提出了闭欧拉直方图统计方法，即在原来欧拉直方图基础上修改空间数据对象边界与查询窗口分割线重合时相应桶的统计方法，将格网单元和格网边对应的桶视为闭的。闭欧拉直方图在文献[19]中有详细介绍。

(3)S-EulerApprox 和 EulerApprox 直方图

CD 直方图和 Euler 直方图只能为相离和相交拓扑关系提供精确的解决方法，但它们却不能处理更精细的关系，如包含(contains)、重叠(overlap)等[2]。

图 4 中，有两种不同的情况(a)和(b)，在查询窗口(Xa，Xb，Ya，Yb)处，包含关系在第一种情况下应该为 1，在第二种情况下应该为 0。但是，CD 和 Euler 直方图算法却对这两种情况建立了相同的直方图[22]。



(a) 包含：1；重叠：1　(b) 包含：0；重叠：2

(c) 与(a)(b)对应的CD直方图　(d) 与(a)(b)对应的Euler直方图

图 4　CD 和 Euler 直方图的缺陷

空间查询的选择性将依具体操作而定，例如常用的相离(disjoint)、外接(meet)、相等(equal)、重叠(overlap)、包含(contains)、被包含(contained)、内接于(covers)、内接(covered)等 8 种拓扑关系[27]，如图 5 所示，它们的选择性就有很大差异。一般地，相离的选择性最低，外接、重叠、被包含、内接于的选择性较低，而相等、内接、包含的选择性则较高[3]。因此，在估计空间查询的选择率时需要考虑空间数据间的拓扑关系。

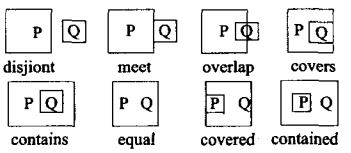


图 5　两个空间数据对象间的空间关系

两个空间数据对象间的空间关系可以参考 9 交模型[29]，

Sun(2002)等人认为有4种拓扑关系——相离、重叠、包含、被包含很重要。他们在欧拉直方图的基础上，提出了求解这4种空间拓扑关系的近似算法——S-EulerApprox和EulerApprox[22]。

在近似算法中，均假设相等关系数目为0，即认为查询窗口与空间数据对象的MBR间不存在相等关系。在实际应用中，查询窗口与空间数据对象的MBR存在相等关系的数目最多为1，故该假设是合理的。

在S-EulerApprox算法中，假设被包含关系的空间数据对象数目为0，即认为查询窗口足够大，不会被包含在空间数据对象内。S-EulerApprox的准确度与两个因素有关：包含查询窗口的空间数据对象的数量、空间数据对象穿越查询窗口的数量。当数据集中较小的空间数据对象占大多数时，该算法精度较高。

当数据集中有大量的较大空间数据对象，或者查询窗口相当小时，S-EulerApprox算法的假设就不再有效，即被包含关系的空间数据对象数目不再为0，于是就设计了EulerApprox算法。但是EulerApprox在计算空间数据对象的外部与查询窗口内部相交的空间数据对象数目时存在较大误差，导致它估算特别大的查询窗口时性能不好。

2.2.3 MP直方图

据现有的资料显示，MP直方图[24]是第一个提出解决带有几何选择的空间连接查询的选择率估计算法。MP直方图的基本假设是：与格网单元相交的空间数据对象不仅是均匀分布的，而且要有相似的宽度和高度。

为了估计两个空间数据对象数据集间的空间连接查询的选择率，MP直方图需要为每个数据集保留3个参数：空间数据对象数目n、空间数据对象平均高度$\bar{h}$和平均宽度$\bar{w}$，则空间连接查询的选择率可用式(3)算得：

$$N=n_1\times n_2\times\min(1,(\bar{h_1}+\bar{h_2}))\times\min(1,(\bar{w_1}+\bar{w_2})) \quad (3)$$

在MP直方图中，一个空间数据对象跨越多个桶时，它只记录该对象中心所在位置的桶的信息，不分割该对象。这样，在精细格网粒度下，效果非常差。一般只有空间数据对象的大小与查询窗口的大小相当时，MP直方图的误差才比较小。

2.2.4 PH(Parametric Histogram)直方图

在估计两个数据集空间连接的选择率时，可以将一个数据集当作基本的数据集合(内层遍历集合)，另一个数据集作为查询窗口集合(外层遍历集合)。这样，窗口查询(空间选择)的选择率估计的总和就是这两个数据集空间连接选择率的估计值[3,4]。

PH直方图[25]是将整个数据空间划分为格网单元，认为数据均匀分布的假设在每个单元内成立。基本思想是将跨越多个格网单元的空间数据对象的最小外包矩形分解为更小的矩形，然后在适当的单元格内处理这些矩形。

在此，需要为每个格网单元保留如下的信息：

D_k：一个数据集；A：整个给定数据空间的面积；N_k：数据集D_k中所有数据项的数目；C_k：数据覆盖范围，即数据集D_k中每个数据项的面积之和与A的比率；W_k：数据集D_k中所有数据项的平均宽度；H_k：数据集D_k中所有数据项的平均高度。

这样，两个数据集D_1和D_2间的空间连接的选择率估计可以按式(4)算得：

$$\mathrm{Size}_{1\_2}=N_1\times C_2+C_1\times N_2+N_1\times N_2\times\frac{W_1\times H_2+W_2\times H_1}{A}$$

$$\mathrm{Selectivity}_{1\_2}=\frac{\mathrm{Size}_{1\_2}}{N_1\times N_2} \quad (4)$$

显然，格网单元等级越精细，PH重复计数问题越严重。

2.2.5 GH(Geometric Histogram)直方图

GH直方图[18]是基于对两个相交的矩形总会产生4个相交的点的观察，空间连接的选择率能通过首先估计两个数据集间相交点的数目，然后除以4得到。

GH建立直方图的方法同PH。对每个格网单元cell(i,j)，需要记录以下信息：(a)$V_k(i,j)$表示有多少条MBRs的垂直边通过cell(i,j)；(b)$H_k(i,j)$表示有多少条MBRs的水平边通过cell(i,j)；(c)$I_k(i,j)$表示有多少个MBRs与cell(i,j)相交；(d)$C_k(i,j)$表示有多少个MBRs的角点落入cell(i,j)。

这样，两个数据集a和b空间连接产生的相交点数目可以用式(5)估计：

$$N_{a\_b}=\sum(C_a(i,j)\times I_b(i,j)+I_a(i,j)\times C_b(i,j)+V_a(i,j)\times H_b(i,j)+H_a(i,j)\times V_b(i,j)) \quad (5)$$

式中，前两项是计算两个MBR的边彼此相交的相交点；后两项是计算一个MBR的角点落入另一个MBR中的相交点。

一般地，GH估计误差随着格网粒度的增加而不断减小。GH直方图的精确度依赖于空间对象的高和宽比查询窗口的高和宽相对要小的假设，即如果数据集中是一些较小的空间数据对象，GH估算精度较高。反之，数据集中以较大空间数据对象为主，GH估算精度较低。

虽然GH直方图也存在重复计数问题，但是研究发现，PH的估计精确度没有GH高。目前它是空间连接选择率估计的最佳算法。

3 空间直方图算法小结

本文所述各种空间直方图算法的空间数据对象都是用其MBR近似描述的。空间查询处理的过程一般分为两个步骤：过滤和求精[29]。过滤步中，常用MBR近似描述空间数据对象，在近似描述的基础上进行查询操作，以获得满足查询条件的空间数据对象候选集。求精步是对候选集中的空间数据对象按查询要求做进一步的计算处理(包括几何计算和属性值计算)，以获得满足查询条件的最终结果[7]。因此，由这些直方图算法得到的查询选择率估计结果能反映且也只能反映过滤步的代价。但SQ直方图除外，它考虑了求精步的代价。

MinSkew直方图和SQ直方图是为了有效地解决分割数据空间来适应任意查询窗口的问题。MinSkew直方图是数据集的密度直方图，能有效地将多边形数据转换为点数据；SQ直方图是针对多边形数据提出来的。它们的共同点是都存在重复计数问题，估计效果并不是很理想。区别主要在于数据空间的划分方式不同，前者是二次划分，后者是用4叉树。此外，SQ直方图基于MBR得到的查询选择率估计值只能作为实际查询选择率的上限。

CD直方图和Euler直方图能准确估计给定的查询窗口与空间数据对象相交的数目，即空间选择的选择率估计。研究发现，CD直方图一般只对使用MBR近似表示的空间对象具有较高的估计精度，这主要是因为它用4个直方图分别存储空间对象的4个角点信息，如果用三角形或五边形等描述

空间数据对象，其准确性就会降低。Euler 直方图当给定查询窗口与直方图格网对齐时，能保证结果的正确性。但是，CD，Euler 直方图只能处理极其简单的空间拓扑关系（相离与相交这两种）的空间选择查询选择率估计。陈海珠针对 Euler 直方图存在的“边界问题”提出了闭 Euler 直方图的概念。Sun 等人提出了解决更加精细的空间关系算法，将相交关系分解为重叠、包含和被包含关系，在 Euler 直方图的基础上，他们提出了近似算法来计算这 3 种较为详细的拓扑关系以及相离关系的选择率估计。而目前尚未出现用 CD 直方图处理这些详细关系的算法，作者认为一个主要原因在于 CD 直方图的存储代价较大（需要存储 4 个子直方图来统计累加信息）。

可以直接用于空间连接查询的选择率估计的空间直方图算法有 MP，PH，GH 等。前人从格网粒度和查询窗口大小方面对这三者进行了研究，发现 GH 直方图是三者中最稳定的。在大多数情况下，GH 估计误差随着格网粒度的增加而不断减小，而 MP 在所有的格网层次上都有较高的误差。另外，查询窗口的大小改变时，对选择率估计精度也有影响。大致的趋势是：随着查询窗口面积的增加，估计精度随之提高。当查询窗口较小时，GH 的精度会较低，这是因为 GH 估计的是相交点的数目，较小的查询窗口意味着大量的相交点落在窗口外，使得估算不够准确。MP 由于其严格的假设条件，导致其应用不多。PH 存在重复计数问题。

结束语 空间查询的选择率估计是为空间查询优化服务的，直方图是空间查询选择率估计的最常用方法。在选择直方图算法时，应该让其满足 3 个特性：(1)估计具有较小的误差；(2)创建、使用和维护直方图的代价小；(3)能用于多种不同查询的选择率估计问题。这样，空间数据库系统才能在花费较少资源的基础上获得准确度较高的查询估计结果，从而选取最优的执行策略。

本文所讨论的各种直方图算法，都是以 MBR 作为空间数据对象的近似描述，查询窗口也都是矩形的。虽然一般能满足查询处理过程中过滤步骤的基本要求，但是有时 MBR 表示的空间拓扑关系与实际空间对象间的空间拓扑关系存在不一致的问题[30]，这样就影响了估计的精确程度。因此，用更为精确的描述形式来表示空间数据对象、研究非矩形查询窗口的选择率估计，将是空间查询优化研究的发展趋势之一。此外，避免重复计数问题是直方图研究的一个主要问题。

由于不同拓扑关系的选择性是有很大差异的，因此研究更为精细拓扑关系的查询选择率估计的精确算法是很有必要的。

空间查询优化已成为空间数据库应用的一个研究热点，选择率估计是其中一个重要方面。目前有关空间查询选择率估计的理论研究仍有待完善，在空间查询优化领域还面临诸多挑战。

参 考 文 献

[1] Jin J, An N, Sivasubramaniam A. Analyzing Range Queries on Spatial Data[C]// Proceedings of the 16th International Conference on Data Engineering. 2000: 525-534

[2] Liu Qing, Yuan Yidong, Lin Xuemin. Multi-resolution Algorithms for Building Spatial Histograms[C]// Proceedings of the 14th Australasian Database Conference. 2003: 145-151

[3] 方裕，楚放. 空间查询优化[J]. 中国图象图形学报，2001，6(4)：307-314

[4] Papadias D, Sellis T, Theodoridis Y, et al. Topological relations in the world of minimum bounding rectangles: A study with R-trees[C]// Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. 1995: 92-103

[5] Theodoridis Y, Stefanak is E, Sellis T. Cost models for join queries in spatial databases[C]// Proceedings of the 14th International Conference on Data Engineering (ICDE'98). 1998: 476-483

[6] 蒋苏蓉，石青青，黄志良. 空间查询优化[J]. 计算机工程与应用，2004：188-190

[7] 郭平，陈海珠. 空间查询代价模型[J]. 计算机科学，2004，31(12)：65-67

[8] 吴胜利. 估算查询结果大小的直方图方法之研究[J]. 软件学报，1998，9(4)：285-289

[9] Lipton R J, Naughton J F, Schneider D A. Practical selectivity estimation through adaptive sampling[C]// Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data. 1990: 1-11

[10] Hass P J, Swami A N. Sequential sampling procedures for query size estimation[C]// Proceedings of the 1992 ACM SIGMOD International Conference on Management of Data. 1992: 341-350

[11] Harangsri B, Shepherd J, Ngu A. Selectivity estimation for joins using systematic sampling[C]// Proceedings of the 8th International Workshop on Database and Expert System Applications. 1997: 384-389

[12] Chen C M, Roussopoulos N. Adaptive selectivity estimation using query feedback[C]// Proceedings of ACM SIGMOD Conference. 1994: 161-172

[13] Belussi A, Faloutsos C. Estimating the selectivity of spatial queries using the correlation' fractal dimension[C]// Proceedings of the 21st International Conference on Very Large Data Bases. 1995: 299-310

[14] Acharya S, Poosala V, Ramaswamy S. Selectivity Estimation in Spatial Databases[C]// Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. 1999: 13-24

[15] Kooi R P. The optimization of queries in relational databases [D]. Case Western Reserve University, Sept. 1980

[16] Poosala V, Haas P J, Ioannidis Y E, et al. Improved histograms for selectivity estimation of range predicates[C]// Proceedings of ACM SIGMOD Conference. 1996: 294-305

[17] Viswanath P. Histogram-based estimation techniques in databases[D]. University of Wisconsin-Madison, 1997

[18] An N, Yang Z Y, Sivasubramaniam A. Selectivity estimation for spatial joins[C]// Proceedings of the 17th International Conference on Data Engineering. 2001: 368-375

[19] 陈海珠. 空间查询优化研究[D]. 重庆：重庆大学，2004

[20] Liu Qing, Lin Xuemin, Yuan Yidong. Summarizing Spatial Relations-A Hybrid Histogram[C]// Web Technologies Research and Development-APWeb2005 7th Asia-Pacific Web Conference. 2005: 464-476

[21] Aboulnaga A, Naughton J F. Accurate estimation of the cost of spatial selections[C]// Proceedings of the 16th International Conference on Data Engineering. 2000: 123-134

[22] Sun C, Bandi N, Agrawal D, et al. Exploring spatial datasets with histograms[J]. Distributed and Parallel Databases, 2006, 20: 57-88

（下转第 148 页）

表 2　Mashroom 和 Pumsh 的特征

数据库	项数	事务总数	平均长度
Mashroom	120	8124	23
Pumsb	7117	49046	50

图 5 显示 Mashroom 在较低支持度下各算法效率的变化情况。支持度从 14%到 4%，随着支持度的降低，满足最小支持度的频繁项集的长度和个数快速增加，各算法花费的时间也快速增加。其中 Apriori 算法变化最大，花费时间最多，Hash-BFI 较 BitTableFI 所花费的时间要少。

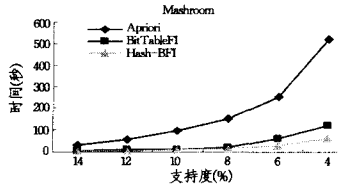


图 5　低支持度下算法性能比较

如图 6 所示，对包含大量项数测试集的 Pumsb 在较高支持度下进行测试，支持度从 50%到 30%。随着支持度的降低，Hash-BFI 算法所花费的时间缓慢增长，花费的时间明显比 BitTableFI 和 Apriori 要少。通过实验和理论分析，本文算法引入散列函数，在处理像 Pumsb 这样包含大量项数的测试集时，大大减少了候选 2-项集产生的个数，突破了 BitTableFI 和 Apriori 算法瓶颈，充分显示本文算法的时间优越性。然而我们也发现，随着测试集的增大和支持度的降低，这几个算法时间花费明显增大。

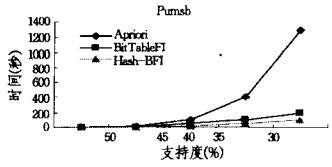


图 6　高支持度下算法性能比较

结束语　针对 BitTableFI 算法的不足提出 Hash-BFI 算法，在处理非常花费资源的二项集上引入散列函数，BitTableFI 算法没有对候选项集进行剪枝。本文算法对候选项集进行有效的剪枝，并且完全通过 AND，OR 运算产生候选项集和计算候选项集支持度。然而，实验过程中也发现了一些问题，由于 Hash-BFI 是基于宽度比较算法，虽然在稀疏测试集上有较大优势，但是在低支持度下处理稠密测试集用时还是比较多。文献[7,8]把位表和深度探索方法相结合，取得了很好的效果，我们下一步要在这方面做一些工作。

参 考 文 献

[1] Agrawal R, Imilienski T, Swami A. Mining association rules between sets of items in large databases[C]// Proceedings of the ACMSIGMOD International Conference on Management of Data. Washington, DC, 1993: 207-216

[2] Han J W, Pei J, Yin Y W, et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach[J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53-87

[3] Tsay Y J, Chiang J Y. CBAR: an efficient method for mining association rules[J]. Knowledge-Based Systems, 2005, 18(2/3): 99-105

[4] Dong Jie, Han Min. BitTableFI: An efficient mining frequent itemsets algorithm[J]. Knowledge-Based Systems, 2007, 20: 329-335

[5] Burdick D, Calimlim M, Gehrke J, et al. MAFIA: a maximal frequent itemset algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1490-1504

[6] Park J S, Chen M S, Yu P S. Using a hash-based method with transaction trimming for mining association rules[J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9(5): 813-825

[7] Song Wei, Yang Bing-ru, Xu Zhang-yan. Index-BitTableFI: An improved algorithm for mining frequent itemsets[J]. Knowledge-Based Systems, 2008, 21: 507-513

[8] Duemong F, Preechaveerakul L, Vanichayobon S. FIAST: A novel algorithm for mining frequent itemsets[C]// International Conference on Future Computer and Communication. 2009: 140-144

[9] 柴华昕，王勇. Apriori 挖掘频繁项目集算法的改进[J]. 计算机工程与应用，2007，43(24)：158-171

[10] 朱玉全，杨鹤标，孙蕾. 数据挖掘技术[M]. 南京：东南大学出版社，2006

(上接第 129 页)

[23] Beigel R, Tanin E. The geometry of browsing[C]// Proceedings of the Third Latin Ameriacn Symposium on Theoretical Informatics. 1998: 331-340

[24] Mamoulis N, Papadias D. Selectivity estimation of complex spatial queries[C]// Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Database. 2001: 155-174

[25] Aref W, Samet H. A Cost Model for Query Optimization Using R-Trees[C]// Proceedings of the Second ACM Workshop on Advances in Geographic Information Systems (ACM-GIS). 1994: 60-67

[26] Shekhar S, Chawla S. Spatial Databse: A Tour [M]. Prentice Hall, 2003

[27] Date C J. An introduction to database systems[M]. New York: Addison-Wesley Publishing Company, 1994

[28] Egenhofer M J, Herring J R. Categorizing binary topological relations between regions, lines, and points in geographic databases[C]// Egenhofer M J, Mark D M, Herring J R. editors. The 9-Intersection: Formalism and Its Use for Natural-Language Spatial Predicates. National Center for Geographic Information and Analysis, Report 94-1, 1994: 13-17

[29] Brinkhoff T, Horn H, Kriegel H P, et al. A Storage and Access Architecture for Efficient Query Processing in Spatial Databse Ssystems[C]// Proceedings of 3rd International Symposium on Advaces in Spatial Databases. 1993: 357-376

[30] 陈琳，杜友福，王元珍. MRR：基于 MBR 的空间关系模型[J]. 计算机工程与应用，2002：76-78

[31] Sun C, Agrawal D, Abbadi A El. Selectivity Estimation for Spatial Joins with Geometric Selections[C]// Proceedings of the 8th International Conference on Extending Database Technology: Advances in Database Technology. 2002: 609-626