维普资讯 http://www.cqvip.com

空间查询代价模型*)

郭 平 陈海珠
(重庆大学计算机学院 重庆400044)

摘 要 空间数据固有的复杂性使空间数据查询效率成为了空间信息处理和应用的关键与瓶颈,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。本文分析比较了几种主要的查询代价估算模型,指出了它们的使用范围与存在的问题,最后展望了空间查询代价模型研究的发展趋势。

关键词 空间查询优化,代价估算,代价模型

Cost Models of Spatial Queries

GUO Ping CHEN Hai-Zhu
(College of Computer Science, Chongqing University, Chongqing 400044)

Abstract Because of spatial data 's inherent complexity, the efficiency of spatial queries has been the key and bottleneck in processing and application of spatial information. Query optimizations based on cost estimation are an important method to improve the efficiency of spatial queries. This paper analyzes and compares some of the main cost models of the cost estimation, and points out their application fields and existing problems. In the end of this paper, we briefly discuss the future research directions of the cost model of spatial queries.

Keywords Spatial query optimization, Cost estimation, Cost model

1 背景

在空间数据库中既存储了属性数据又存储了空间数据,这些数据特别是空间数据数量庞大、结构复杂、操作代价巨大。因此,空间数据查询的效率成为了空间数据库性能的瓶颈,空间查询优化势必成为空间数据应用的难点和突破点。而现有的关系数据库查询优化技术不能完全适用于空间数据,空间查询优化必须充分考虑空间数据自身的特点。目前,对空间查询优化的研究在空间索引技术、查询算法优化、查询代价估算等分支领域都有了一定的进展。

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。查询代价估算指按照给定的代价指标,根据查询操作的处理特性、操作间的相互关系、操作数据的统计信息等估算查询操作的执行代价。代价模型是利用可供选择的查询执行计划的代价估算结果,以选择代价较小的执行计划作为执行策略。

关系数据库中的查询代价估算研究已较为成熟,并在一些实际数据库系统中得到了应用,例如 SQLServer、Oracle 等著名数据库系统中均有关于数据的审计信息以供估算查询代价使用。相对而言,空间数据库的查询代价估算还处于研究阶段,实际应用也比较少。本文研究了几类目前主要的空间查询代价优化模型,对它们的性能进行了评价,最后简要地讨论了空间查询代价模型研究的发展方向。

2 空间查询与查询代价

2.1 空间查询

空间选择和空间连接是空间查询中最常用的两种操作。在实际 GIS 应用中,用户在一张地图中给出一个查询窗口,查找与之相交的空间对象以及关于这些对象的属性数据,这一操作被称为空间选择(Spatial Selection),地图称为空间数据集。空间对象包括空间的点、线、面。空间选择也称为范围查询(Range Queries)。类似地,判断两张地图上的空间对象是否相交,可将其中一张地图上的对象作为查询窗口在另一张上做空间选择操作,对所有的结果求并集,从而得到查询结果,这一个过程称为空间连接(Spatial Join)。其它的空间查询操作大都在这两种操作的基础上进行演化。

与关系数据库查询不同,空间查询对空间数据库中的空间数据和属性数据都进行了操作。由于空间数据结构复杂并具有多维性等特点,在空间查询中判断空间数据是否满足空间查询条件(例如,空间对象是否与查询窗口相交)是一个复杂的问题。

2.2 空间查询代价

为提高空间查询的效率,空间查询处理一般分为两步:过滤和求精[17]。过滤步中,首先对空间对象进行近似描述,如以对象的最小边界矩形 MBR。然后,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对象候选集。求精步是对候选集中的空间对象按查询要求进行进一步的计算处理(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。

由此,空间查询代价包括过滤和求精两步的代价。过滤步获得的候选集的大小直接影响求精步的代价。一般说来,通过引入和改进空间数据索引方法以及使用直方图估计等方法可以降低过滤步的代价同时减少候选集的大小;降低求精步代价主要从减少 I/O 操作代价等方面考虑。

由于空间数据固有的复杂性使精确计算空间查询的代价非常困难,同时这种计算本身的代价也比较大,因此,常用的

*)本研究得到国家自然科学基金项目(编号:50378093)和重庆大学基础及应用基础研究项目(编号:717411038)共同资助。郭 平 副教授,研究方向:DW&DM,定性推理,GIS。陈海珠 硕士研究生,研究方向:空间查询优化,GIS。

维普资讯 http://www.cqvip.com

方法是对空间查询的代价进行估算，建立代价模型，实现查询优化。

3 空间查询代价模型

如前所述，空间选择和空间连接是空间查询中最常用的两种操作。在这两种操作代价估算结果的基础上，可估算出其它查询操作的代价。而且，空间查询中对属性数据的操作代价可使用关系数据库查询代价模型进行估算。因此，大多数空间查询代价模型都是估算二维空间里空间选择和空间连接这两种查询中的空间数据操作的代价。为叙述方便，将空间对象分布的空间称为数据空间。

3.1 基于索引结构的代价模型

早期的空间查询优化研究集中在高效空间索引研究上，其目的是通过改进索引效率来提高查询效率。空间数据库的研究经过近三十年的发展，涌现了大量的空间索引结构[10]。不同的应用采用不同的索引结构，如何评价这些索引结构对空间数据库系统性能的影响进而提高优化的效率，将比提出新的索引结构更有意义。

商用空间数据库大多采用 R 树及其变形树作为索引结构，因此有不少基于 R 树的代价模型。Faloutsos 最早提出了评价 R 树性能的代价模型[16]。随后，Kamel 和 Faloutsos 在假定查询窗口大小固定且数据均匀分布的前提下提出了基于 R 树的选择代价模型[11]。文[12]提出的空间选择代价模型不要求查询窗口大小固定。文[13]提出了空间连接处理算法，它以其中一个数据集中的对象作为查询窗口，同步遍历建立在两个数据集上的 R 树，在此基础上建立了代价模型。选择哪一个数据集对象作为查询窗口对估算结果影响颇大。文[12,13]提出的代价模型通过估算访问磁盘的次数来反映查询代价。文[14]则从 I/O 角度出发，先在无 I/O 缓冲区情况下处理连接的 I/O 代价，再结合缓冲区存在情况下产生的页面失效率来估算基于 R 树的连接处理代价。经实验，文[11～13]提出的代价模型的估算结果与实际代价之间的平均误差率均小于 15%，它们均不受数据分布特征的影响。前述的代价模型均假设 R 树是预先建立好的，文[14]提出了可动态构建的索引结构--种子树(Seed -Tree)，文[15]给出了其代价模型。

由以上所列举的几种代价模型可看出，它们均与索引结构紧密结合，从索引结构的特点出发来估算查询代价。除种子树[14]外，这些代价模型不适于在未建立索引的数据集上估算查询代价。

3.2 基于直方图的代价模型

直方图是许多商用数据库系统中最常用的一种估算查询结果大小的方法[5]。近年来，基于直方图的空间代价模型成为空间查询优化研究的一个热点。其基本思想是：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，由此估算查询代价。这些记录单元称为桶，桶的集合称为直方图。

Acharye 和 Poosala 构建了基于二元划分的 MinSkew 直方图[6]。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。如何把一个子空间划分为更小的子空间(是二等分还是其它的划分方式)是一个复杂的问题。

与 MinSkew 直方图不同，CD[7]直方图将数据空间划分为大小相等的单元格(即子空间)，并且查询窗口 Q 的边界与数据空间的划分线相重合。CD 直方图在每个桶内分别统计了落在该桶对应区域内的空间对象的四个顶点数目，使用 CD 公式：

$$B_{ll}(x_b,y_b)-B_{lr}(x_a,y_b)-B_{ul}(x_b,y_a)+B_{ur}(x_a,y_a)$$

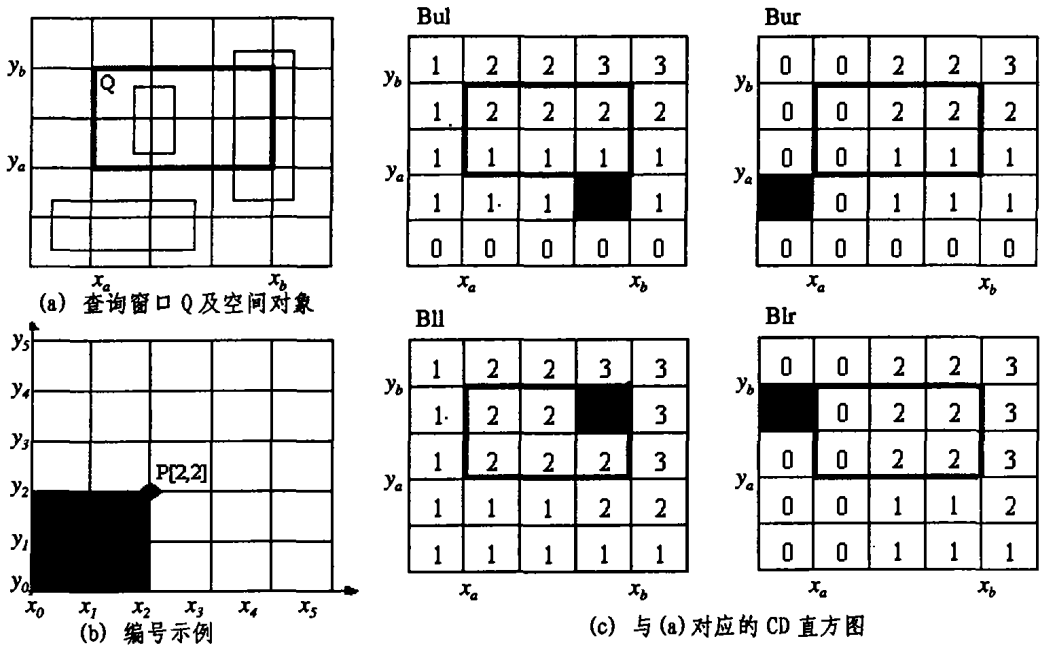


图1 查询窗口 Q 及空间对象、对应的 CD 直方图以及编号示例

可以获得与查询窗口 Q 相交的空间对象数目。其中，B_{ul} 表示各桶对对象左上角顶点的统计结果，类似地可以描述 B_{lr}、B_{ul}、B_{ur} 的统计内容。这里以图1为例说明 CD 直方图的统计方法。为便于问题的说明，对划分对象空间的水平线和垂直线以及对象空间的边界进行编号，如图1(b)所示。$P[i,j]$ 表示 x_i 与 y_j 的交点，$R[i,j]$ 表示以 $P[0,0]$ 为左下角顶点、$P[i,j]$ 为右上角顶点的区域。$B_{ul}[i,j]$ 表示落在区域 $R[i,j]$ 内对象的左上角顶点的数目。B_{ur}、B_{ll}、B_{lr} 等符号意义以及统计方法类

维普资讯 http://www.cqvip.com

似。由 CD 计算公式计算得：

$$B_{ll}(x_b,y_b)-B_{lr}(x_a,y_b)-B_{ul}(x_b,y_a)+B_{ur}(x_a,y_a)$$
$$=3-0-1+0=2$$

即与查询窗口 Q 相交的对象数目为2。

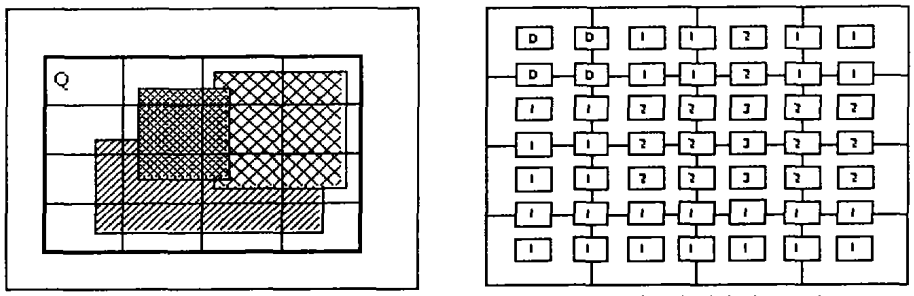
EH[1~3]直方图则以图形学中的欧拉公式为基础。它只需要对查询窗口进行均匀划分。通过统计与单元格相交的对象数目，使用欧拉公式：

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

来估算与查询窗口相交的空间对象数目 $Selectivity(Q)$。图2给出了 EH 直方图的示例。由欧拉公式有：

$$Selectivity(Q)=\sum_{0\leqslant k\leqslant 2}(-1)^k F_k(Q)=10-31+24=3$$

即与查询窗口 Q 相交的对象数目为3。



(a) 查询窗口 Q、空间对象　　(b) 与(a)对应的 EH 图

图2　查询窗口 Q、空间对象以及对应的 EH 图

CD 和 EH 两种直方图均不受数据分布的影响，而且在空间选择代价估算中能获得较精确的结果。EH 直方图的计算量要比 CD 直方图的少，其已在 SJGS 选择因子的估算[1]、空间对象分布等研究[2,3]中得到了应用。

在估算两个数据集空间连接的选择因子中，GH[9]的精度比 PH[8]的高。这两种两直方图均假定空间对象均匀分布在整个数据空间或局部数据空间，空间对象的分布特征对这两种代价模型的精确性影响很大。由于它们的统计基于几何概率，即以空间对象边长与单元格边长之比作为单元格内两对象相交的概率，故当这个比值很小时，估算错误率极大。

以上所述的估算模型都是基于 MBR(最小边界矩形)的，也就是说，估算得到的结果只能反映在过滤这一步的执行代价，而忽略了求精步的执行代价。文[6]提出了可用于估算求精步执行代价的代价模型——SQ 直方图。它用四叉树分解技术来划分对象分布的空间。每一个子空间对应四叉树的一个节点，直方图的一个桶对应于该树的一个非空节点，桶中记录了在该节点上的对象的信息(如多边形对象的大小、顶点数目等)。从其划分策略来看，当同一子空间内的数据特征(即空间对象的大小、顶点数目等)差异太大时，SQ 直方图的性能并不理想。此外，它允许子空间交叠，那么自然会出现同一个对象出现在多个子空间的情况，即重复计算问题[1,9]。

由上所述，基于直方图的代价模型适用范围更广，它不要求在数据集上建立索引，且构造简单。在选择查询估算中，它可达到较高的准确度。而在连接查询估算中，其准确度易受到数据分布特征的影响。由构造方法决定了一些直方图无法避免重复计算问题。如何提高连接估算准确性、克服重复计算问题是直方图代价模型研究面临的一个主要问题。

3.3　其它的代价模型

除上述两类常用的空间查询代价模型外，还有其它的一些代价估算方法。

采样技术是较早使用的一种代价估算方法。其基本思想是，在原始数据集上选择一个样本，在其上进行查询得到该查询的执行代价，由此来估算出在于原始数据集上的查询执行代价。该方法在数据均匀分布的情况下可得到较理想的估算结果。但在实际应用中，样本的不稳定性限制了这一方法的应用范围。一直以来仍有学者对之进行研究，力图扩大其使用范围并取得了一些成果[4,18]。

基于分形的代价估算方法[19]则不需要考虑数据分布问题。利用数据之间的自相似的特点概括出一定的分布规律，Manfred Schroeder 定义了 Hausdoff 分形维 D_0[20]，文[19]对之进行概化得到了相关分形维 D_2并应用到可执行查询窗口具有不同形状的空间选择代价估算中。Christos Faloutsos[21]等人发现，在两个数据集上进行空间连接操作，对于满足连接条件的空间对象对，将间距在某个距离内的对象对的数目取对数，所得到的值与该距离长度的对数成线性关系，由此可快速地估算出整个结果集的大小(其时间复杂度为 O(1))。文[19]的相关分形维 D_2是这一规律的特例。这两种方法值只适用于点集数据，不利于推广到使用二维数据的实际应用中。

结论与展望　顾名思义，代价估算只是查询代价大小的不精确值，倘若为了求精确值而花费了过多的系统开销，是得不偿失，有悖于优化查询的初衷。因此，代价模型应该是在不花费过多的系统资源的基础上尽可能准确地估算查询代价，从而得到最优的执行计划。

无论使用哪一种代价模型作为查询优化的策略，都要考虑空间对象在空间数据库中的表示问题。通常以 MBR 作为二维空间对象的近似描述。但是，很多情况下这种不精确描述影响了估算的准确性。目前，已有研究者采用近似精度更高的描述形式来表示空间对象(如五角形、直角多边形等)，并在此基础上进行查询优化的研究。这样的研究已成为空间查询优化研究的趋势。如何将这些近似精度更高的表示数据应用到代价模型中提高估算的准确度，将会成为今后研究的一大趋势。

空间查询优化已成为空间信息处理的一个研究热点，代价估算是其中的一种主要的优化方法。本文对空间查询代价估算中的主要代价模型进行了综述，并分析了各种代价模型存在的问题。从中我们可以看出，其理论研究仍有待完善，其实际应用还是很少的。如何解决这些问题，找到一种可应用到实际的代价估算模型对提高空间数据库的性能具有重要的意义。

参 考 文 献

1　Sun C, Agrawal D, El Abbadi A. Selectivity Estimation for Spatial Joins with Geometric Selections. In: Proc. of EDBT, 2002

2　Sun C, Agrawal D, El Abbadi A. Exploring spatial datasets with histograms. In: Proc. of IEEE ICDE, 2002

(下转第80页)

维普资讯 http://www.cqvip.com

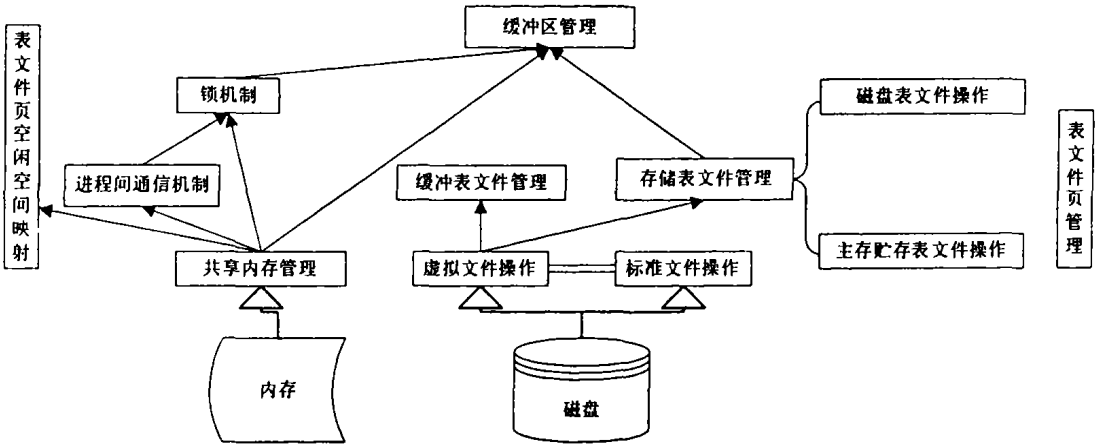


图7 PostgreSQL 磁盘管理体系结构

结论 通过对 PostgreSQL 存储管理实现机制的研究我们可以得出如下结论：

·PostgreSQL 存储文件系统的操作本身不涉及多个进程的并发访问，所以它的操作必须严格串行进行；

·后端可以并发访问的地方是共享文件缓冲区，它是存储表数据的内存映像，所以必须对它的访问操作进行精确的并发控制。PostgreSQL 采取两步走：(1)对于缓冲区本身的访问由于涉及的时间相对短，这部分的并发控制依靠轻量级锁来实现；(2)在替换缓冲区时要把缓冲区的内容刷新到磁盘文件中，由于 PostgreSQL 磁盘文件系统本身没有并发控制，因此也要在缓冲区层面来实现临界区的并发访问；再者，由于我们更新的是表文件的一个页面，花费的时间也不会太长，这部分的控制也依靠轻量级锁机制来实现；

·磁盘管理辅助事务管理的思想体现在：在写磁盘块时，总是先写事务日志后写数据；

·由于 PostgreSQL 基本运行环境是单机环境(允许多 CPU)，因此它的锁机制实现的基点是：运行环境有：(1)共享内存；(2)统一的时间定位机制。

PostgreSQL 的存储管理机制是建立在支持平台的文件系统上的，它的实现效率、可靠性等在很大程度上取决于支持的文件系统。如果把文件系统的某些功能直接由数据库存储管理来实现而不要通过调用平台的文件系统接口，那么存储系统的性能可望得到提高。

参 考 文 献

1 PostgreSQL Development Group. PostgreSQL V- 7. 3. 4 source codes. PostgreSQL website http:www. Postgresql. org,2003

2 PostgreSQL Development Group. PostgreSQL V- 7. 3. 4 Documentation. PostgreSQL website http:ww-w. postgresql. org,2003

3 Stonebraker M. The design of the post- gres stroge system. EECS Department Univers-ity of California Berkely,Ca,94720,1987

4 Garcia-Molina H, Ullman J, Widom J. Database System Implementation. Prentice Hall,2001

5 Ullman J, Widom J. A First Course in Dat-abase sytems. Prentice-Hall,1997

6 何伟平. PostgreSQL 的昨天今天和明天——自由软件数据库 PostgreSQL 简介 v2. 0. PostgreSQL 中文网站. http://www.pgsqldb. org,2003

(上接第67页)

3 Liu Q,Yuan Y,Lin X. Multi-resolution Algorithms for Building Spatial Histograms. In:Proc. Fourteenth Australasian Database Conf. (ADC2003), Pages145～151

4 Wu Y, Agrawal D, El Abbadi A. Query Estimation by Adaptive Sampling. IEEE ICDE, 2002

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

6 Aboulnaga A S,Naughton J F. Accurate estimation of the cost of spatial selections. In:ICDE'00, Proc. of the 16th Intl. Conf. on Data Engineering, March 2000. 123～134

7 Jin J,An N,Sivasubramaniam A. Analyzing range queries on spatial data. In:ICDE'00, Proc. of the 16th Intl. Conf. on Data Engineering, March 2000. 525～534

8 Aref W,Samet H. A Cost Model for Query Optimization Using R-Trees. In:Proc. of ACM GIS, Gaithersburg, Maryland, Nov. 1994. 60～67

9 An N,Yang Z Y, Sivasubramaniam A. Selectivity Estimation for Spatial Joins. In:ICDE'01, Proc. of the 17th Intl. Conf. on Data Engineering, April,2001. 368～375

10 梁中,孙小燕,谭勇桂. 空间索引技术-回顾与展望. 计算机工程与应用,2002,24:197～199

11 Kamel I,Faloutsos C. On Packing R-tree. In:Proc. of the CIKM, 1993. 490～499

12 Theodoridis Y,Sellis T. A Model for the Prediction of R-tree Performance. In: Proc. 15th ACM PODS Symposium, 1996. 161～171

13 Theodoridis Y, Stefanakis E, Sellis T. Cost Models for Join Queries in Spatial Databases. IEEE ICDE, 1998

14 Lo M L,Ravishankar C V. Spatial Joins Using Seeded Trees. In: Proc. of ACM SIGMOD Conf. 1994

15 Lo M L,Ravishankar C V. Spatial Hash-Joins. In:Proc. of ACM SIGMOD Conf. 1996

16 Faloutsos C, Sellis T,Roussopoulos N. Analysis of Object Oriented Spatial Access Methods. In:Proc. of the ACM SIGMOD/PODS Int. Conf. on Principle Of Data, 1987

17 Orenstein J A. Spatial Query Processing in An Object-Oriented Database System. In:Proc. ACM SIGMOD Int. Conf. on Management of Data, May,1986. 326～336

18 Wu Y L. Query Result Estimation in Database:[Doctoral Dissertation]. University of California Santa Barbara,2001

19 Belussi A, Faloutsos C. Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension. VLDB 1995. 299～310

20 Schroeder M. Fractals,Chaos,Power Laws:Minutes From an Infinite Paradise. W. H. Freeman and Company, New York, 1991

21 Faloutsos C, Seeger B, Traina A, Traina C Jr. Spatial join selectivity using power laws. In: Proc. of the 2000 ACM-SIGMOD Conf. Dallas,TX,May 2000. 177～188