文章编号 :1001 - 2486(2004)06 - 0082 - 04

一种有效的距离连接选择度估计方法*

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

摘 要 :距离连接在空间数据库中有着广泛的应用,而距离连接的选择度估计是优化距离查询的基础。通过综合分析和比较了现有的选择度估计技术,提出了一种利用米诃夫斯基和与直方图进行距离连接选择度估计的新方法。实验结果证明该种方法能够有效地进行距离连接选择度估计。

关键词 :距离连接 ;选择度估计 ;米诃夫斯基和 ;直方图
中图分类号 :TP392 文献标识码 :A

An Efficient Selectivity Estimation for Distance Joins

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

Abstract :Distance join is widely used in spatial database. Selectivity estimation for distance join is the basis of optimizing the query of the distance. Incorporating the existing selectivity estimation techniques, a new selectivity estimation method for the distance join based on Minkowski sum and histogram is proposed. The experimental results show that the method is efficient for selectivity estimation for the distance joins.

Key words :distance join ;selectivity estimation ;minkowski sum ;histogram

距离连接在空间数据库中有着广泛的应用。例如 :查询‘ 找出距离河流最近且人口超过五百万的城市 ”;“ 距离商店最近的货栈 ’等。Worboys[1] 对空间对象间距离计算的有效方法进行了深入的探讨。涉及区域的距离定义是建立在特定语义基础上的,如两个区域间的距离可以定义为两个区域的重心间的距离(重心距)、两个区域上的点间距离的最小值(最小距离)、两个区域上的点间最大距离的最小值(最小最大距离)、两个区域上的点间最小距离的最大值(最大最小距离)等。特定的应用可以根据需要定义相应距离语义,不同的语义类型也可能出现在同一个应用中。空间距离关系查询计算通常在 CPU 和I/O 代价很高的空间连接操作中完成,空间连接选择度估计技术能够以小的计算代价给出空间连接结果集大小的近似估计,根据估计值就可以在查询优化器中构造出好的查询计划,尽量减少连接时的操作代价。

直方图选择度估计方法能适应各种数据分布,而且存储空间需求和表示误差也比较小,是数据库领域中应用最为广泛的空间连接选择度估计方法。使用直方图方法需要解决完全空间划分所造成的重复计数或不使用完全空间划分方法。相关工作有 :SQ 直方图[2] 用数据划分代替空间划分来避免了重复计数问题 ;MinSkew 直方图[3] 可以支持任意形状的查询窗口 ;Jin 等[4] 提出了一种可以集成数据压缩算法的积聚密度直方图技术 ;文献[5] 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度估计的几何直方图方法 ;基于小波变换的空间直方图压缩算法[6] 主要用于降低存储空间需求。在现有空间连接选择度估计方法的基础上,本文基于直方图方法,引入米诃夫斯基和(Minkowski Sum)进行距离连接选择度估计。该新方法的有效性将通过实验予以评估。

* 收稿日期 :2004 - 06 - 08
基金项目 :国家 863 高技术资助项目(2002AA131010,2002AA134010)
作者简介 :熊伟(1976—),男,博士生。

1　缓冲区距离连接

在对象索引层面上,可以将距离连接划分为“点—点”连接、“点—区域”连接和“区域—区域”连接三种类型(“点—区域”、“线—线”和“线—区域”连接可以看成是“区域—区域”连接的特例,这里不再进行探讨)。对于“区域—区域”连接,如果连接距离相对于区域对象的边长平均值不是很大,则称为缓冲区连接 如果连接距离比区域对象的边长平均值大很多,那么“区域—区域”连接可以作为“点—点”连接看待。首先,针对缓冲区连接展开探讨。

多维范围查询的选择度估计经常采用米诃夫斯基和,通过 MBB(Minimum Bounding Box)将范围查询转换为等效的点查询。假设归一化二维空间均匀分布,将 Y 的 MBR(Minimum Bounding Rectangle)高度和宽度分别扩展后,X、Y 对象的 MBR 相交概率可以转换为 X 的 MBR 中心点落入到这个扩展区域的概率,这个概率等于该区域的面积,如图 1 所示。令 $LLC_{i,j}$和 $URC_{i,j}$分别表示对象包围框 i 的左下角和右上角在 j 轴上的坐标($0 \leqslant i < N, 0 \leqslant j < d$)。其中,$N$ 是 R 树 R_1 所检索的空间对象的数目,d 是度量空间的维度。用同样的方法定义另一棵 R 树 R_2 中对象 MBB 的左下角 $LLC_{k,j}$和右上角 $URC_{k,j}$,其中,$0 \leqslant k < M$,且 M 是 R_2 中对象的数目,则 R_1 与 R_2 间忽略数据空间边界影响的相交连接选择度估计可以通过式(1)给出(这里假设至少其中一棵索引树上的对象大小和中心点坐标是符合均匀分布假设的),其中,$X_{i,j} = URC_{i,j} - LLC_{i,j}$,$Y_{k,j} = URC_{k,j} - LLC_{k,j}$。

$$Sel_{\text{int}}(R_1, R_2) = \sum_{k=0}^{M-1} \left\{ \sum_{i=0}^{M-1} \left(\prod_{j=0}^{d} (X_{i,j} + Y_{k,j}) \right) \right\} \tag{1}$$

将米诃夫斯基和应用于缓冲区连接估计,对于二维空间情况,在进行缓冲区连接时,可以将一个对象的 MBR 放大成图 2 中阴影部分所示的米诃夫斯基和形式,其中 r 表示连接的归一化距离。根据该区域的特征,R_1 与 R_2 间忽略数据空间边界影响的缓冲区连接选择度估计可以通过下面的等式给出:

$$Sel_{buffer}(R_1, R_2, r) = \sum_{k=0}^{M-1} \left\{ \sum_{i=0}^{N-1} \left[\prod_{j=0}^{1} (X_{i,j} + Y_{k,j} + 2r) - (4r^2 - \pi r^2) \right] \right\} \tag{2}$$

$$= \sum_{k=0}^{M-1} \left\{ \sum_{i=0}^{N-1} \left[\prod_{j=0}^{1} (X_{i,j} + Y_{k,j}) \right] \right\} + 2r \times \sum_{k=0}^{M-1} \left\{ \sum_{i=0}^{N-1} \left[\sum_{j=0}^{1} (X_{i,j} + Y_{k,j}) \right] \right\} + NM\pi r^2 \tag{3}$$

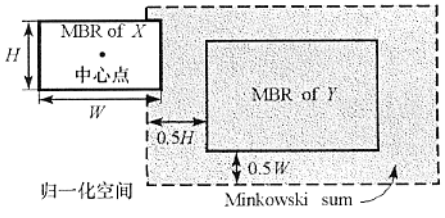

图 1　米诃夫斯基和示意图

Fig. 1　Minkowski sum

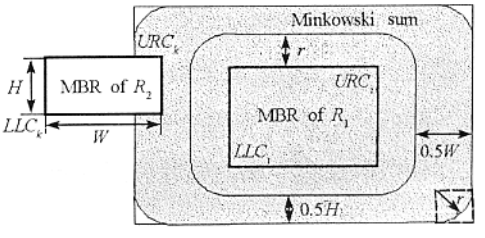

图 2　缓冲区连接示意图

Fig. 2　Buffer join

式(3)中右面和式的第一项恰好是 R_1 与 R_2 间相交连接的选择度估计公式,而第二项则可以进一步重写为 $2Mr \times \sum_{i=0}^{N-1} \left(\sum_{j=0}^{1} X_{i,j} \right) + 2Nr \times \sum_{k=0}^{M-1} \left(\sum_{j=0}^{1} Y_{k,j} \right)$,其中,$\sum_{i=0}^{N-1} \left(\sum_{j=0}^{1} X_{i,j} \right)$是 R_1 中对象 MBR 的周长和,而 $\sum_{k=0}^{M-1} \left(\sum_{j=0}^{1} Y_{k,j} \right)$则是 R_2 中对象 MBR 的周长和。从式(3)可以看出,$Sel_{buffer}(R_1, R_2, r)$ 可以通过现有的相交连接选择度估计方法和几个简单的统计量(数据集的基数和对象 MBR 的平均周长)计算出来,它是 r 的二次函数。因为相交连接的选择性可以用精度更高的直方图方法予以估计,所以式(3)在距离 r 相对于对象 MBR 边长而言比较小的情况下的选择度估计也可以达到比较高的精度。必须说明的是:首先,选择度估计公式(3)假定至少一个空间关系是符合均匀分布的;其次,式(3)没有考虑数据空间边界的影响,考虑边界影响的情况要复杂许多;我们希望在后续工作中对这一问题予以探讨。

2 “点—点”距离连接

地理信息系统中用户定义的“商业规则”往往包含两个点集距离连接的条件，下面开始讨论面向“点—点”距离连接的选择度估计问题。Faloutsos 等[7]提出的利用相关分形维度上的幂律实现距离连接选择度估计的优势是只需要存储很少的数据就可以实现距离连接的选择度估计，它更适用于点集自连接的选择度估计。但对于不同点集，则只有当参与连接的两个数据集具有类似的分布特征时才是有效的。本文给出一种基于直方图的“点—点”距离连接选择度估计方法。该方法区分距离指标与直方图网格边长的不同比例情形，分别给出了长距离连接和短距离连接情形下的选择度估计公式。这两个公式的前提假设是参与连接的数据集具有一定程度的自相似性，这通常是可以满足的[8]。对于连接距离相对于直方图网格的边长而言不是很大的情况（如图 3 所示），可以将其中任意一个直方图（如 R_2）的网格扩充为一个足以容纳该网格与以距离 r 为半径的圆的米诃夫斯基和的区域，并且假设这个新扩充的区域上的对象与扩充前的网格具有相同的均匀分布特征（即自相似性假设），则这个区域中的点数 $m=\dfrac{n_2\times Xarea}{Area}$，其中，$Area$ 表示直方图网格的面积，$Xarea$ 表示这个扩充区域的面积，n_2 表示直方图网格中的对象数目。

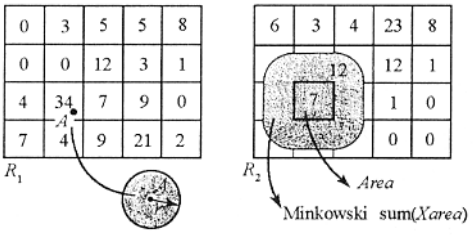

图 3 近距离“点—点”连接的情况

Fig. 3 Point-point close distance join

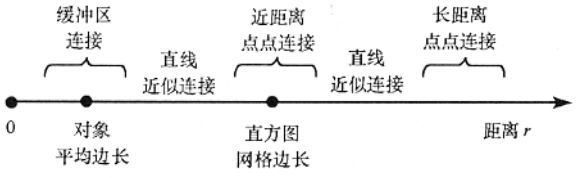

图 4 距离连接选择度估计的方案选择策略

Fig. 4 Policy of selectivity estimation for distance join

直方图 R_1 的相应网格中的某个点对象与 R_2 中点对象距离连接的结果即以该点对象为圆心、r 为半径的圆在 $Xarea$ 中盖住的点对象，这样，直方图 R_1 的相应网格中的所有点对象与直方图 R_2 中点对象连接所产生的结果集大小 j 则可以通过下面的估计式给出：

$$j=n_1\times\pi r^2\times\frac{m}{Xarea}=n_1\times\frac{\pi r^2}{Xarea}\times\frac{n_2\times Xarea}{Area}=\frac{n_1\times n_2\times\pi r^2}{Area} \tag{4}$$

其中，n_1 表示直方图 R_1 的相应网格中的对象数目。根据式(4)，数据集 R_1 和 R_2 上针对距离 r 的距离连接选择性可以通过下面的估计公式给出：

$$Sel_{dist}(R_1,R_2,r)=\frac{\pi r^2}{Area}\times\sum_{k=1}^{N}n_{1,k}\times n_{2,k} \tag{5}$$

其中，N 表示直方图的网格数目，$n_{1,k}$ 和 $n_{2,k}$ 分别表示直方图 R_1 和 R_2 上第 k 个网格中的对象数目。对于连接距离相对于直方图的网格边长而言比较大的情况，一个可行的方案是将直方图网格抽象为一个处在网格中心的加权点（其权值是该网格中的对象计数）。两个数据集的长距离连接选择度估计可以通过以一个直方图上的加权点为中心，以 r 为半径的圆所覆盖的另一个数据集上的加权点权值之和给出。长距离连接结果集大小 $j=\sum_{i=1}^{N}\left(n_{1,i}\times\sum_{\|P_{2,k}-P_{1,i}\|\leqslant r}n_{2,k}\right)$，其中，$N$ 表示直方图的网格数目，$n_{1,i}$ 和 $n_{2,k}$ 分别表示直方图 R_1 和 R_2 上第 i 和 k 个网格中的对象数目，$\|P_{2,k}-P_{1,i}\|$ 表示 R_2 中第 k 个加权点 $P_{2,k}$ 到 R_1 中第 i 个网格中心点 $P_{1,i}$ 的距离，该值只需要经过常数时间的比较。对于长距离连接而言，数据空间边界的影响是不能忽略的，但是本文方法对数据空间边界的影响具有天然的适应性。估计值的准确性跟直方图划分的粒度有很大关系，越细的划分可以使直方图网格加权点越能代表该网格中的对象集，同时网格边长的影响也可以忽略。

最后给出三种距离连接估计方案的选择策略，与距离 r 之间的关系如图 4 所示。这三种方案通常并不能覆盖 r 的整个值域，在两种方案之间的模糊区域可以利用一条连接两条估计曲线的端点的

方法予以近似。

3　实验结果分析

通过一组数据对提出的距离连接选择性公式进行了评估,这组数据是来自 Census TIGER® 2000 的纽约公路和典型地物 MBR 数据集(见图 5),两个数据集的对象数分别为 8356、354。

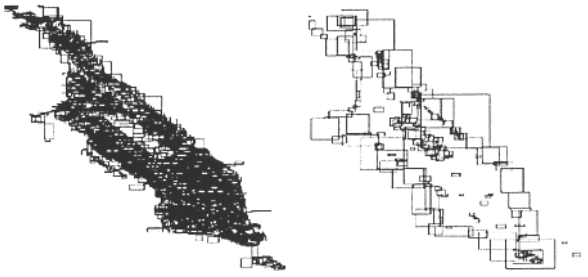

图 5　Census TIGER® 2000 的纽约公路和典型地物 MBR 数据集

Fig. 5　Typical geo-set of New York from Census TIGER® 2000

图 6 给出了在不同粒度下几何直方图(GH)对相交连接的选择度估计结果与真值的对比。结果显示:当 GH 的划分达到一定粒度时,GH 的选择度估计误差可以控制在 15% 以内;直方图划分得过细并不能为选择度估计带来什么好处。图 7 给出了缓冲区连接结果与真值的对比,采用的是 10×10 直方图网格。结果显示:在距离相对于对象 MBR 的平均边长而言比较小(< 1)时,本文给出的缓冲区连接选择度估计方法可以取得比较好的估计效果。图 8 显示了不同直方图粒度下 $S = \left(\sum_{k=1}^{N} n_{1,k} \times n_{2,k}\right)/Area$ 的变化趋势。理论上说,该值应该随着直方图粒度的细分而趋于一个稳定的常量,图 8 确实呈现了这种趋势。图 9 给出了点集近距离连接选择度估计的比较实验结果,其中,直方图网格的粒度为 10×10。这组实验数据显示:距离在 0 ~ 2 倍网格边长间变化时,式(5)的估计效果都比较令人满意。长距离连接在实际应用中是比较少见的,本文给出了估计公式,并根据公式推导过程分析了影响估计准确性的因素,因此不再用实验加以验证。

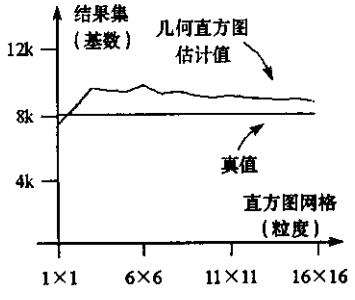

图 6　不同粒度下的 GH 估计效果

Fig. 6　GH estimation of different granularity

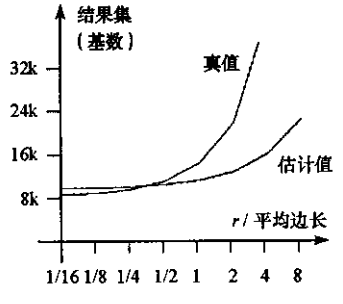

图 7　面向 MBR 的 DJ 选择度估计

Fig. 7　Selectivity estimation for DJ based on MBR

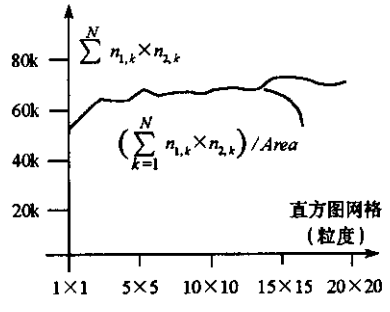

图 8　不同粒度下 S 的趋势

Fig. 8　Trend of S for different granularity

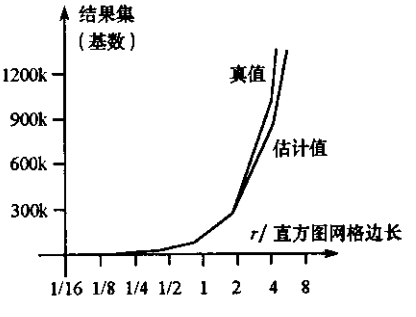

图 9　近距离点集 DJ 选择度估计效果

Fig. 9　Estimation for point-set close DJ

(下转第 90 页)

等。通信互联网所需的通信带宽由组成网络化作战系统的节点数、雷达跟踪的目标数、跟踪数据的更新率、目标数据量等因素决定。通信互联网应该具有大容量、高速率、高质量、较高的抗干扰能力以及安全保密性等特点。

考虑要地防空反导作战,防空导弹网络化作战系统通信互联网的骨干网可以采用有线光缆。采用有线光缆不仅容量大,而且可靠性高和抗干扰能力强。针对要地防空,在系统建设过程中,可以根据可能的作战地域和位置预设通信端口。战时,各节点之间的通信主要依靠光纤骨干网。根据作战任务和要求,各节点在要地周围实施机动,并根据所在的位置选择附近的端口入网。节点内各系统或部分的通信则可以采用无线局域网。因为,节点内各部分可以是物理上分散的,但彼此之间的距离不会太大(距离大的发射任务可以由其他发射节点完成),而且节点内系统之间的通信数据量不大,所以可以采用无线局域网结构。如指挥控制节点与跟踪制导节点通过光缆网通信,发射节点内各发射车与发射控制系统的通信则可以采用无线局域网,这样保证了节点各部分之间是分布结构,提高了系统的生存能力。

3 结束语

本文研究了由跟踪制导网、指挥控制网和拦截兵器网组成的防空导弹网络化作战 C^4ISR 系统体系结构。对于防空导弹网络化作战来说,该体系结构只是系统结构的一种高层抽象。要真正实现防空导弹的网络化作战,还有很多问题要进行深入研究,如统一战场态势的生成与共享,通信互联网的拓扑结构和容量,通信网络的管理和控制,作战指挥的协同等。

参 考 文 献:

[1] Alberts D S, Garstka G G. Understanding Information Age Warfare[M]. ISBN 1 - 893723 - 04 - 6, 2000.

[2] Alberts D S,et al. Network Centric Warfare Developing and Leveraging Information Superiorty[M]. ISBN 1 - 57906 - 019 - 6, 2000.

[3] Applied Physics Lab. The Cooperative Engagement Capability[J]. Johns Hopkins Technical Digest, 1995,16(4): 377 - 396.

[4] Department of Defense Report to Congress,Network Centric Warfare[R]. Http://www. c3i. osd. mil/NCW/ ncw _ main. pdf.

(上接第 85 页)

参 考 文 献:

[1] Worboys W F. GIS: A Computing Perspective[M]. Taylor & Francis, Ltd. ,ISBN 0 - 7484 - 0065 - 6, 1995.

[2] Aboulnaga A, Naughton J F. Accurate Estimation of the Cost of Spatial Selections[A]. In Proc. of ICDE, 2000.

[3] Acharya S, Poosala V, Ramaswamy S. Selectivity Estimation in Spatial Databases[A]. In Proc. of ACM SIGMOD, 1999.

[4] Jin J, An N, Sivasubramaniam A. Analyzing Range Queries on Spatial Data[A]. In Proc. of ICDE, 2000.

[5] An N, Yang Z, Sivasubramaniam A. Selectivity Estimation for Spatial Joins[A]. In Proc. of ICDE, 2000.

[6] Wang M, Vitter J S, Lim L, et al. Wavelet-based Cost Estimation for Spatial Queries[A]. In the Proc. of the 7th SSTD, 2001.

[7] Faloutsos C, Seeger B, Traina A,et al. Spatial Join Selectivity Using Power Laws[A]. In Proc. of ACM SIGMOD, 2000.

[8] Belussi A, Faloutsos C. Estimating the Selectivity for Spatial Queries Using the Correlation's Fractal Dimension[A]. In Proc. of 21st VLDB, 1995.