估算查询结果大小的直方图方法之研究*

吴胜利

(中国科学院软件研究所 北京 100080)

摘要 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.从实用的观点来看,过去已提出的一些直方图方法有局限性,主要是它们不能保证估算值的准确程度.本文将提出两种新的直方图方法,它们不仅使用方便,而且可以保证所有的估算值均在给定的误差范围内.此外,本文还探讨了不同的数据分布对直方图的影响,通过运用一些重要的参数刻画数据分布,用以帮助生成效果较佳的直方图.

关键词 数据库系统,查询优化,查询代价的估算,直方图.

中图法分类号 TP311.56

在数据库系统的实现中,查询优化器一般都是基于代价的.查询优化器比较各种不同执行方案的代价,从中选出代价最小者执行.估算代价的主要问题是估算查询结果的大小,对查询结果大小的估算准确与否直接影响到相应方案代价估算的准确性,从而对查询优化器的质量有实质性的影响.因此,该问题引起许多研究人员的关注.目前已有不少论文发表,并有一些方法问世.其中主要可分为以下4种:采样法(Sampling)[1,2]、曲线拟合(Curve Fitting)[3]、含参数学分布(Parameterized Mathematical Distribution)[4]和直方图(Histogram).[5~9]采样法无需储存任何信息,只是在查询优化时随机选取一些数据样本,并执行真正的查询来估算查询结果的大小.只要样本数足够多,用这种办法所估算的结果很准确,缺点是必须在查询处理时进行,时间开销大,所以实际应用并不多.曲线拟合用代数多项式近似地表示数据的实际分布,而含参数学分布用含参数学分布(如均匀、正态、泊松和Z分布)近似地表示数据的实际分布.上述两种方法的缺点是估算结果不够准确.直方图方法概念直观、实现简单,是所有方法中最常用的,已在一些商用关系数据库系统(如DB2,Informix,Ingress,Sybase)中应用.

直方图方法又可细分为等宽(Equal-width)[5]、等高(Equal-height)[6]、变宽(Viriable-width)[7]、偏向两端(End-biased)[8]等各种类型.Viswanath Poosala等人在文献[9]中给出了一种直方图分类法,并引入了一些新的直方图,如压缩(Compressec)和最大差异(Maxdiff)方法,它们都是特殊的变宽直方图.已有的讨论部分地解决了这样一个问题,即对于一些随机(或特定)的数据分布,各种方法在统计意义上孰优孰劣.由于效果较好的方法一般使用和维护比较麻烦,所以要在追求好的效果和使用方便两者之间进行折衷.两种极端的情况是,等宽方法使用最方便,但效果最差;最佳直方图的生成和维护是最困难的,如其生成就是一个NP-完全问题.[8]

从实用的观点来看,过去提出的一些方法是有局限的,它们中哪一种也不能保证估计值的准确程度,而这对于查询处理来说是至关重要的.保证估计值的准确度应作为生成直方图的一个前提.为此,我们在本文中提出两种新的直方图方法,即嵌套等宽直方图和限定误差的变宽直方图.此外,本文还将探讨不同的数据分布对直方图的影响,并通过运用一些重要的参数刻画数据分布的特征,用以帮助生成效果较佳的直方图.

1 基本概念与直方图定义

下面以关系数据库为例,但所讨论的直方图方法同样适用于其他种类的数据库系统,如面向对象数据库等,只要在该数据库系统中支持基于集合的选择操作.

在关系数据库系统中,每一个关系框架包括若干属性的定义,关系实例包括若干个元组.后面的讨论均假定属性的值域是整数或实数集合.

设R是一关系框架,A是R中一取整数或实数为值的属性,即A的值域为[min..max],min和max均为实常数

* 本文研究得到国家自然科学基金资助.作者吴胜利,1963年生,博士,副研究员,主要研究领域为数据库,信息系统.面向对象技术.

本文通讯联系人:吴胜利,北京 100080,中国科学院软件研究所

本文1997-01-21收到原稿,1997-05-14收到修改稿

(或整常数),它们分别为 A 中可能出现的最小值和最大值.直方图方法将[min..max]分为若干个区间(亦称之为直方或桶),然后统计 A 属性值落在这些区间内的元组个数.当对 R 进行 A 属性上的范围查询 $a_i \leqslant t.A \leqslant a_j$(即求 R 关系中 A 属性值在 a_i 与 a_j 之间的元组个数)时,可根据桶中的数值估算该查询结果所含的元组个数.与每一个 a_i(min$\leqslant a_i \leqslant$max)相对应,存在一个 v_i 表示 R 关系实例中 A 属性取值'a_i'的元组个数.对所有的 v_i 显然有 $v_i \geqslant 0$ 成立.v_i 称作 a_i 的频数.A 的数据分布是二元组集合 $T=\{(a_1,v_1),(a_2,v_2),\ldots,(a_n,v_n)\}$.$a_1=\min$,$a_n=\max$.在关系 R 中满足 $t.A \leqslant a_i$ 的元组称作 a_i 的累积频数,用 c_i 表示,A 的累积数据分布是二元组集合 $\mathbf{T^c}=\{(a_1,c_1),(a_2,c_2),\ldots,(a_n,c_n)\}$,其中 $c_j=\sum_{i=1}^{j} v_i (j=1,2,\ldots,n)$.设 A 的值域为 **A**,**A** 上的直方图定义如下.

定义 1(直方图).定义于 **A** 上的直方图 **H** 是一个四元组 $(as_i,at_i,rai_i,att_i)(1\leqslant i\leqslant m)$ 的集合.其中 $[as_i,at_i]$(min$\leqslant as_i \leqslant at_i \leqslant$max)表示一个区间,$as_i$ 和 at_i 分别是该区间的起点和终点.att_i 表示落入该区间的元组总个数,rai_i 表示该区间中不同的域值个数.下面 3 个条件是 **H** 必须满足的.

(1) $[as_1,at_1]\cap[as_2,at_2]\cap\ldots\cap[as_m,at_m]=\varnothing$

(2) $[as_1,at_1]\cup[as_2,at_2]\cup\ldots\cup[as_m,at_m]=\mathbf{A}$

(3) $\sum_{i=1}^{m} att_i=Sum$,其中 Sum 是 R 关系实例所含元组的个数.

如果 **A** 是一个整数域,则 rai_i 可省略.

一些直方图,如文献[8]中提出的偏向两端直方图,不能被该定义所包含,因为我们限定直方图中每一个桶与 **A** 中一个子区间相对应.有理当相信,不符合该定义的直方图实现效率低,实用价值不大,因此本文未作讨论.

例 1:设关系 R 含有 100 个元组,其属性 A 的值域是(1,2,...,9).直方图 **H** 含有 3 个桶 h_1(1,3,10),h_2(4,6,30),h_3(7,9,60).现查询 A 属性值落在 4～8 区间内的元组,则估算值为 $30+60\times 2/3=70$.其中对每个桶,均假定所有的元组对各个属性值而言是均匀分布的.

2 两种新的直方图方法

本节引入两种新的直方图方法.对于等值查询,用这两种方法所给出的结果大小估算值满足任意给定的准确度要求.下面分别进行讨论.

2.1 嵌套等宽方法

在所有种类的直方图中,等宽直方图的使用和维护最为方便,但有时它们所产生的估算值不够准确.我们引入嵌套等宽直方图的意图是使它能保持等宽直方图使用方便的优点,同时又要克服等宽直方图估算不够准确的缺点.具体做法是将 A 的值域 **A** 分成 $p(p\geqslant 1)$ 个相等的区间,每个区间分配给一个基桶,对于那些频数变化不大、已符合规定准确度要求的基桶,则到此为止;对那些频数变化较大、尚不符合规定准确度要求的基桶,再将它们化分成 q 个等宽子桶.对于不符合要求的子桶还可再作进一步划分,直至所有的基桶(或子桶)均满足准确度要求.在使用时,最上层的基桶是长久保留的,因为数据是动态变化的,所以子桶要根据需要作相应的调整.如桶的最大嵌套深度为 r,最底层的子桶含域值数为 k,则直方图所表示的域值数 $v=k\times p\times q^{-1}$.若 **A** 中所含的值个数 $|\mathbf{A}|=v$,则不需特殊处理.如 $|\mathbf{A}|<v$,则需为 **A** 添加 d 个虚假值,使得 $|\mathbf{A}|+d=v$,形成 p 个符合条件的基桶,但在最后一些基桶中可能含有虚假的域值,需要特殊处理.如 $|\mathbf{A}|>v$,则需要增加 p 或 q 或 r 的值.

文献[10]为嵌套等宽方法设计了数据结构,并给出了初始化等实现算法.如 R 中含有的元组个数为 Sum,则在求出数据分布 **T** 的前提下,初始化可在 $O(Sum)$ 时间内完成.等值查询和范围查询均可在 $O(p+q+r)$ 时间内完成.

从整体上说,不能完全排除重组操作.但只要进行一些必要的维护,重组的周期就能大大延长.如及时修改,则可保持每个基(子)桶中的 v_i 值都是准确的.再则在进行范围查询时,可乘机对嵌套等宽直方图结构进行局部重组.

2.2 限定误差的变宽直方图

定义 2(限定误差的变宽直方图).设 $h_i=(as_i,at_i,rai_i,att_i)(1\leqslant i\leqslant m)$ 是直方图 **H** 中任一直方,对任意的 $a_1,a_2,as_i\leqslant a_1\leqslant at_i,as_i\leqslant a_2\leqslant at_i,a_1\neq a_2$.令 v_1 是 a_1 的频数,v_2 是 a_2 的频数.Const 是一给定的常数,有 $|v_1-v_2|\leqslant$Const,则称 **H** 是限定误差为 Const 的变宽直方图.

定理 1.如采用限定误差为 Ccnst 的变宽直方图,则对任一等值查询 $t.A=a_i$ 的结果大小的估计误差不会超过 Const.

定理的证明可直接从限定误差的变宽直方图的定义推出.

在得到二元组集合 $T=\{(a_1,v_1),(a_2,v_2),\ldots,(a_n,v_n)\}$的基础上,限定误差的变宽直方图可由一遍扫描 T 中成员而得.在生成每一个直方时,记住所遇到的最大值 Max 和最小值 Min(当开始一个新直方时,将 v_i 当作 Max 和 Min 的值),并将新的 v_i 值与 Max 比较,若 $|v_i-\text{Min}|>\text{Const}$ 或 $|v_i-\text{Max}|>\text{Const}$,则结束本直方.与该 v_i 值相对应的 a_i 不包含在该直方内.若 $v_i>\text{Max}$,则将 v_i 作为新的 Max 值.若 $v_i<\text{Min}$,则将 v_i 作为新的 Min 值.具体过程由下述算法 1 描述:

算法 1. 由二元组集合 T 生成限定误差为 Const 的变宽直方图

输入:$T=\{(a_1,v_1),(a_2,v_2),\ldots,(a_n,v_n)\}$

输出:限定误差为 Const 的变宽直方图 $H=\{h_1,h_2,\ldots,h_m\}$

方法:

(1) $j=0$; $New=$ 'T';

(2) for $i=1$ to n do

　　if $New=$ 'T'

　　then {$j=j+1$; $low=v_i$; $high=v_i$; $as_j=a_i$; $New=$ 'F'}

　　else {if $|low-v_i|>\text{Const}\vee high-v_i|>Const$

　　then {$at_j=a_{i-1}$; $New=$ 'T'; $i=i-1$}

　　else {if $v_i<low$ then $low=v_i$;

　　if $v_i>high$ then $high=v_i$;

　　$att_j=att_j+v_i$}

算法 1 可在 $O(n)$时间内完成,n 为 R 关系中在 A 属性上的所有域值个数.

定理 2. 由算法 1 所形成的限定误差为 Const 的变宽直方图,就直方图所具有的直方数而言,是最优的.

换句话说,我们不能找到另一个限定误差为 Const 的变宽直方图,其直方数少于用算法 1 生成的直方图的直方数.

证明:用反证法.假定存在这样一个直方图 $\mathbf{H}_1$,它含有的直方图较用算法 1 所生成的直方图 $\mathbf{H}_2$ 要少.这样,在 $\mathbf{H}_2$ 中,一定有某个直方 h_j,它的起点与 $\mathbf{H}_1$ 的任一个直方起点不相同.设 h_j 的起点落在 $\mathbf{H}_1$ 中的 h_i' 中,如图 1 所示.则(1) h_j 的终点 at_j 不可能大于 h_{i-1}'的终点 at_{i+1}',即 $at_j\leqslant at_{i+1}'$,同样有 $at_{i+1}\leqslant at_{i+2}',\ldots$,

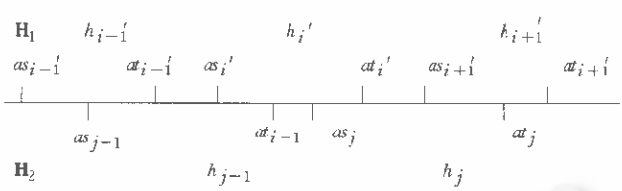

图1　$\mathbf{H}_1$ 和 $\mathbf{H}_2$ 的图示情形

这样在 $\mathbf{H}_2$ 中从 h_j 开始所含的直方数不会少于 $\mathbf{H}_1$ 中从 h_{i+1}'开始所含的直方数.(2) 再看 h_j 的左边,h_{j-1}的终点 $at_{j-1}\geqslant as_i'$,所以 $as_{j-1}>as_{i-1}',\ldots,as_1>as_1'$.则 $\mathbf{H}_2$ 对于小于 as_1'的值至少要定义一个直方,这样在 $\mathbf{H}_2$ 中从 h_1 开始至 h_{i-1} 为止所含的直方数不会少于 $\mathbf{H}_1$ 中从 h_1'开始至 h_{i-1}'为止所含的直方数加 1.综合以上两点,$\mathbf{H}_2$ 所含的直方数不少于 $\mathbf{H}_1$ 所含的直方数,与假设矛盾,故定理成立. □

3　数据分布

要有效地构造和使用直方图方法,了解具体的数据分布特性具有很好的指导作用.

常见的描述数据分布的参数有最大值 $\text{Max}=\text{Max}(v_1,v_2,\ldots,v_n)$,最小值 $\text{Min}=\text{Min}(v_1,v_2,\ldots,v_n)$,平均值 $v=\frac{1}{n}\sum_{i=1}^{n}v_i$,方差 $E^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(v_i-v)^2$.

本节再引入两个对定义直方图有用的参数.

①平均值差 $V_{\text{diff}}=\frac{1}{(n-1)}\sum_{i=2}^{n}|v_i-v_{i-1}|$.$V_{\text{diff}}$定义两两相邻的 V_i 值之间的平均差异.

②误差不超过 Const 的平均域值个数 $V_{\text{buc}}=\frac{1}{(n-1)}\sum_{i=1}^{n}buc(a_i)$,其中 $buc(a_i)$是从位置 a_i 开始,误差不超过 Const 的域值个数.$buc(a_i)$定义为:对二元组集合 $T=\{(a_1,v_1),(a_2,v_2),\ldots,(a_n,v_n)\}$和任意 $k,l(1\leqslant i\leqslant k\leqslant l\leqslant j\leqslant n)$,有 $|v_k-v_l|\leqslant$Const 且至少存在一个 $a_m(1\leqslant i\leqslant m\leqslant j)$,使 $|v_m-v_{j+1}|>$Const(当 $j=n$ 时无此要求),则 $buc(a_i)=j-i+1$.

例 2:$T=\{(1,1),(2,2),(3,5),(4,8),(5,7),(6,9),(7,8),(8,8),(9,10)\}$,要求误差不超过 3,则 $buc(1)=2$,$buc(2)=2,\ldots,V_{\text{buc}}=(2+2+3+6+5+4+3+2+1)\times 1/9=3.111$.

4 实 验

本节中各项实验均采用下述方法产生随机的数据分布.属性的值域范围为(1,2,...,n),n 在 1 000～2 000 之间不等,频数的范围在 0～1 000 之间.随机产生一个 0～1000 范围内的随机数作属性值'1'的频数 v_1,然后按下式生成 $v_2,\ldots,v_n$.

$$v_{i+1}=\begin{cases}(v_i+d_i), & \text{if } 0<(v_i+d_i)<1\ 000\\ 0, & \text{if } (v_i+d_i)\leqslant 0\\ 1\ 000, & \text{if } (v_i+d_i)\geqslant 1\ 000\end{cases}$$

其中 d_i 是随机生成的在[$-k$,k]范围内的随机数,k 选择 20,40,60,80,100,150,200 几个数.最后再随机地将 20～70%的频数值清为 0.

4.1 几种直方图方法的性能比较

我们对限制误差的变宽方法与等宽(Equi-width)、等高(Equi-depth)、压缩(Compressed)和最大差异(Maxdiff)几种方法进行了比较.由于限制误差的变宽直方图与其他几种直方图生成方法有异.所以先确定误差 Const,然后生成限制误差为 Const 的变宽直方图,假设其桶的个数为 m,再用等宽、等高、压缩和最大差异诸方法生成桶个数为 m 的直方图.对每一组直方图,运行 40 个查询,其中 20 个是等值查询,20 个是范围查询.对于一个查询的错误率用下式计算.

$$E=\frac{|S_q-S_q'|}{n}$$

其中 S_q 和 S_q' 分别表示查询结果的实际和估计大小,n 为元组的总个数.

上述各种直方图方法的错误率见表 1.

表 1 几种直方图的错误率比较

直方图	等值		范围	
	平均(%)	最坏(%)	平均(%)	最坏(%)
等宽	15.43	59.76	16.18	23.56
等深	8.53	42.36	9.97	18.70
压缩	2.58	9.23	4.93	8.93
最大差异	2.55	8.19	4.72	8.61
限制误差的变宽	2.52	4.01	4.71	6.33

从表中可看出,无论是等值查询还是范围查询,限制误差的变宽方法比其他方法都要好一些,特别是最坏性能明显优于其他方法.

4.2 嵌套等宽方法与等宽方法的比较

设属性的值域范围为 1～1 024.嵌套等宽方法采用 64 个基桶,每个基(子)桶含两个子桶,最大嵌套层数为 5 层.

对任一等值查询,其结果的估计值与准确值的绝对误差不超过特定误差,比较满足上述条件的嵌套等宽直方图和等宽直方图两者对存储的要求.共比较了 32 000 对,其结果是,两者对存储的要求平均比值为 0.78:1,极端情况分别为 28:1 和 0.37:1.它表明要满足给定的条件,嵌套等宽方法比等宽方法一般所需的存储空间要少.当然,此时用等宽直方图估计查询结果的大小比相应的嵌套等宽方法具有更少的平均错误率,但最坏情形是一样的.

4.3 对经验公式的验证

经验公式

$$V_{buc}=\left\lceil\frac{Sum}{t}\right\rceil \tag{1}$$

其中 $t=\mathrm{Min}(Sum,1+V_{diff}/\mathrm{Const})\times\mathrm{Const}/V_{diff}$.

可通过规定误差 Const、元组个数 Sum 和平均步值差 V_{diff} 的值估算满足该条件的变宽直方图的直方数 V_{buc}.不管用何种直方图,Const,Sum 和 V_{diff} 都是较容易得到或估算出的参数,而 V_{buc} 的值会随规定误差 Const 和元组个数 Sum 的变化而变化.该经验公式所得到的 V_{buc} 估计值可对直方图中设置多少个直方提供有用的信息.例如,采用等宽直方图,可将直方数设置成估计值 V_{buc} 的 2～3 倍.如采用等高直方图,可将直方图设置成估计值 V_{buc} 的 1.5～2 倍较妥当.

实验共生成了 5 000 个数据分布,通过式(1)计算出 V_{buc} 的估算值,并通过生成规定误差为 Const 的变宽直方图,

算出V_{buc}的准确值.结果平均估计错误率

$$E=\frac{1}{5\ 000}\sum_{i=1}^{5\ 000}|V_{buc}(i,\text{估计值})-V_{buc}(i,\text{准确值})|/V_{buc}(i,\text{准确值})$$

为14.2%.其中$V_{buc}(i,\text{估计值})$为第i个数据分布中的V_{buc}估计值,$V_{buc}(i,\text{准确值})$为第i个数据分布中V_{buc}的准确值.

5 结束语

本文从实用的观点出发,讨论了两种新的直方图方法,即嵌套等宽直方图和限制误差的变宽直方图.和已有的一些方法相比,实现简单且能保证估计值的准确程度,具有较高的实用价值.此外,本文还对数据分布引入了一些新的参数以支持直方图应用.其中第4.3节引入的经验公式较为有用.进一步可探讨的问题为:①可对限制误差的变宽直方图作改进以保证估算范围查询结果大小的精度.②更广泛地研究数据分布与直方图的关系,为生成简单、有效的直方图提供指导手段.

参考文献

1 Hass P J, Swami A N. Sequential sampling procedures for query size estimation. In: Proceedings of ACM SIGMOD Conference, 1992

2 Lipton R J, Naughton J F, Schneider D A. Practical selectivity estimation through adaptive sampling. In: Proceedings of ACM SIGMOD Conference, 1990

3 Chen C M, Roussopoulos N. Adaptive selectivity estimation using query feedback. In: Proceedings of ACM SIGMOD Conference, 1994

4 Federowicz J. Database estimation evaluation using multiple regression techniques. In: Proceedings of ACM SIGMOD Conference, 1984

5 Merett T H and Otoo E. Distribution models of relations. In: Proceedings of the 5th VLDB Conference. Rio de Janeiro, Brazil, Oct. 1979

6 Piatetsky-Shapiro G, Connell C. Accurate estimation of the number or tuples satisfying a condition. In: Proceedings of ACM SIGMOD Conference, 1984

7 Muthuswamy B, Kerschberg L. A ddsm for relational query optimization. In: Proceedings of ACM SIGMOD Annual Conference, Denver, USA, Oct. 1985

8 Ioannidis Y E, Poosala V. Balancing histogram optimality and practicality for query result size estimation. In: Proceedings of ACM SIGMOD Conference, 1995

9 Poosala V, Ioannidis Y E. Improved histograms for selectivity estimation of range predicates. In: Proceedings of ACM SIGMOD Conference, 1996

10 吴胜利.面向对象数据库系统的查询优化[博士论文].南京:东南大学,1996
(Wu Sheng-li. Query optimization in object-oriented database systems[Ph. D. Thesis]. Nanjing: Southeast University, 1996)

Histogram Method for Size Estimation of Query Result

WU Sheng-li

(*Institute of Software The Chinese Academy of Sciences Beijing* 100080)

Abstract Histogram is the commonest method for the size estimation of query result in many commercial database systems. Several histogram methods presented in the past have certain limitations in practicability due to their lack of guarantee of the accuracy of the estimation. In this paper, the author presents two kinds of new histogram methods, which are easy to use and can guarantee the accuracy of the estimation. Otherwise, the effect of different data distributions to histograms is discussed in the paper, and some important parameters of data distribution are introduced to help produce better histograms.

Key words Database systems, query optimization, estimation of query cost, histogram.