

直方图下数据连接操作代价估计应用分析^①

冯凯平, 陈 衡, 冯超颖

(四川烹饪高等专科学校 信息技术系, 成都 610072)

摘 要: 直方图是一种重要的数据查询工具. 在数据库操作中查询操作需要进行代价估计, 而操作代价的估计有多种方法. 直方图估计法在所有估计法中是最容易使用和最容易维护的, 它将关系中的全部元组进行归类设桶, 对每个桶分别进行大小统计, 继而求和, 使代价统计值更接近于真实. 结合实际课题应用, 给出了将关系的连接转换为多种直方图的方法, 以及如何利用直方图进行代价估计.

关键词: 直方图; 代价估计; 元组; 等宽; 数据连接; 查询

Application Analysis of Cost Estimation for Operation of Data Connection in Histogram

FENG Kai-Ping, CHEN Heng, FENG Chao-Ying

(Information Technique Department, Sichuan Higher Institute Cuisine, Chengdu 610072, China)

Abstract: Histogram is a kind of important tool in data query. Cost estimation needed in query operation of database varies in methods, in which histogram is that is most convenient for use and maintenance. Histogram categorizes and sets barrels to all tuples in the relations, calculates the size respectively for each barrel, and then summarizes, thus making the price statistics closer to the real value. In application of the actual program, this research works out methods that convert the connection of the relations to a variety of histograms and ways to estimate the operation cost by using histograms.

Key words: histogram; cost estimation; tuple; equal width; data connection; query

1 引言

查询代价的估算是查询优化中一项不可缺少的工作. 优化操作可从多方面进行, 包括合理的投影、选择、连接、并、交、分组与聚集、消除重复等, 而各种操作中所发生的代价各不相同. 其中连接操作代价估计在各种操作代价估计中, 由于涉及多个关系、以及关系中元组的不确定性, 是最难以估计的. 人们设计了多种连接代价估计算法, 各具特点. 其中, 直方图估计法由于简洁、使用简单、易维护, 当数据连接对象的大小与查询窗口的大小相当时显示出优势^[1].

2 基于统计量收集的连接大小估计

假设有关系 $R(X, Y)$ 和 $S(Y, Z)$, 用 $T(R)$ 、 $T(S)$ 表示 R 、 S 的元组数目, 用 $V(R, X)$ 表示 R 中对于属

性 X 不同取值的种类数. 作如下假设:

a) 若 Y 是 R 、 S 的公共属性, 且 $V(R, Y) \leq V(S, Y)$, 则 R 的每个 Y 都是 S 的一个 Y .

b) 如果 Y 是 R 的属性但不是 S 的属性, 则 $V(R \bowtie S, Y) = V(R, Y)$.

根据以上假设, 可以对 $R(X, Y) \bowtie S(Y, Z)$ 的大小做如下估计^[1,2].

如果 $V(R, Y) \leq V(S, Y)$, 由于关系 S 中有 $T(S)$ 个元组, 它与 R 中元组 t 连接的期望元组数是 $T(S)/V(S)$, 因此, R 的每个元组 t 有 $V(S, Y)$ 的机会与给定的一个 S 元组相连接. 由于有 $T(R)$ 个 R 的元组, $R \bowtie S$ 的估计大小是 $T(R \bowtie S) = T(R) \cdot T(S)/V(S, Y)$.

如果 $V(R, Y) \geq V(S, Y)$, 则对应的估计是 $T(R \bowtie S) = T(R) \cdot T(S)/V(R, Y)$.

^① 基金项目:四川省高等学校重点科研项目(11ZA178)

收稿时间:2012-02-15;收到修改稿时间:2012-03-21

因此, 通常情况下用下式进行估计:

$$T(R \propto S) = T(R) \cdot T(S) / \text{Max}(V(R, Y), V(S, Y)) \quad (1)$$

实例 1 X_s (姓名, 特质, ...)是关于学生特质水平的部分属性. 假设 $T(X_s)=1200$, 特质水平分为 30 个等级, 则 $V(X_s, \text{特质})=30$.

T_i (题号, 难度, 区分度, 猜度, 综合指数...)是一个关于题库的部分属性. 设题库有 6000 道测试题, $T(T_i)=6000$, 综合指数值从 1 到 30, 分为 30 个区间, $V(T_i, \text{综合指数})=30$.

查询特质水平与试题综合指数接近的所有考生所被抽取的试题:

Select a.姓名, b.题号

From X_s as a, T_i as b

Where $ABS(a.\text{特质}-b.\text{综合指数}) \leq 0.5$ (2)

由式(1)可知其连接规模的估计值大约是:

$$T(X_s \propto T_i) = T(X_s) \cdot T(T_i) / \text{Max}(30, 30) = 1200 \times 6000 / 30 = 240000.$$

3 基于直方图的连接大小的估计

以上统计法是针对整个关系进行统计估计. 很多情况下会带来较大误差. 为此可将所有元组划分成小块, 对每个小块分别进行统计法估计, 再将所有的小块估计进行求和.

使用直方图估计连接的大小比前一种统计法更准确. 尤其当连接属性的值显式地出现在参加连接的两个关系的直方图上时, 则可准确地知道结果中有多少元组将具有该值. 对于那些非显式出现在一个或两个关系直方图上的连接属性的值, 仍按统计法进行估计.

3.1 最频值型直方图

基本原理是, 找出各个关系中出现最频繁的值以及这些值出现的次数, 再找出次频繁出现的值以及它们出现的次数, 以此类推, 最后记录其他出现频率较低的值及其次数.

对于关系 $R(Y, A)$, 如果 $V(R, A)$ 不是太大, 则该直方图由具有属性 A 的每个值的元组的数目(或一部分)组成. 如果这个属性存在大量的不同的值, 则重点记录最常出现的值, 而其他值则分组统计^[1-3].

实例 2 假设关系 X_s 、 T_i 有如表 1 所示的频值表. 有一个 3 个最频值以及它的数目. 欲计算连接 X_s (姓名, 特质) $\propto$ T_i (题号, 综合指数)操作后的元组数目. 其中,

特质和综合指数为连接属性.

令 b ="特质"、 c ="综合指数". 则 $X_s.b$ 和 $T_i.c$ 的直方图如表 1、图 1.

表 1 $X_s.b$ 与 $T_i.c$ 频值表

$X_s.b$		$T_i.c$	
特质	频数	综合指数	频数
18	200	17.6-18.5	320
19	180	18.6-19.5	280
20	170	20.6-21.5	290
其他值	650	其他值	2110

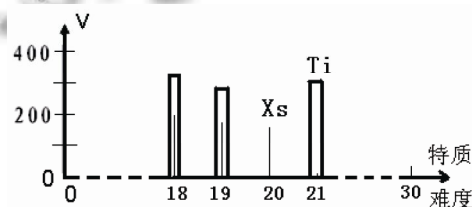


图 1 $X_s.b$ 与 $T_i.c$ 直方图

在 X_s 的 1200 个元组(1200 名考生)中, 有 200 个 b 值为 18, 180 个 b 值为 19, 170 个 b 值为 20. 此外, 还有 650 个元组的 b 值不为 18、19 和 20, 并且其余值中的任一值出现次数不超过 170. 已知 $V(X_s, b)=30$, 即 X_s 的未知 b 值的 650 个元组被分成 27 个值, 每个值平均有 24 个元组.

关系 T_i 共有 6000 个元组(6000 道题目). 17.6-18.5 的值出现了 320 次, 18.6-19.5 出现了 280 次, 20.6-21.5 出现了 290 次. 已知 $V(T_i, c)=30$, 即 T_i 的未知 c 值的 2110 个元组被分成 27 个值, 每个值平均有 78 个元组.

b 值的 18 与之相适应的 c 值 17.6-18.5、19 与之相适应的 c 值 18.6-19.5 在两个关系中显式出现, 因此, 可以计算 X_s 的 200 个 $b=18$ 的元组与 T_i 的具有对应 c 值的 320 个元组的连接结果有 64000 个元组. 类似地, 180 个 $b=19$ 的 X_s 元组与 c 对应的 280 个 T_i 元组的连接结果有 50400 个元组.

对其余元组连接数的估计则较为复杂.

作如下假设: 在具有较小值集的关系 X_s 中出现的每个值将在关系 T_i 的值集中出现.

观察直方图, X_s 的其余 28 个 b 值中存在 20, 出现了 170 次. 并且假定 X_s 中有与 c 中 20.6-21.5 对应的

21, 因为它是 T_i 中的最频值之一, 并且在 X_s 中出现 24 次, 其连接元组数估计值是 $24 \times 290 = 6960$. 另外, T_i 中必定有与 20 对应的 19.6-20.5, 估计在 T_i 中出现 78 次, 这些估计是分别通过假设该值是它所属关系的直方图中“其他值”中的一个而获知的, 其估计元组数是 $78 \times 170 = 13260$.

在两个关系中还有其他 27 个 b 和 27 个 c 值, 估计它们每一个值在 X_s 中出现 24 次, 在 T_i 中出现 78 次. 因此 27 种值中的每一个连接增加的元组数是 $24 \times 78 = 1872$. 因而结果大小的估计是:

$$64000 + 50400 + 6960 + 13260 + 27 \times 1872 = 185164$$

比用统计值收集法估算的结果 240000 更精确一些.

3.2 等宽型直方图

将数据分布 S 划分成 n 个不相交的等宽子集 $A_i(i=1,2,\dots,n)$, 使其满足 $S=A_1 \cup A_2 \cup \dots \cup A_n$. 在平面直角坐标系中, 直方图横轴表示每个子集的取值范围 v , 纵轴表示数据位于该子集内的频数^[1].

使用等宽直方图, 两个关系的连接属性有相同的带宽, 可以先估计相应带宽上连接的大小, 并对这些估计值求和^[4,5].

实例 3 表 2、表 3 分别为低特质和高特质某一级别学生平均成绩, 分析低特质与高特质平均成绩相当时的重合度.

此查询计划是对关系 $Low(Level, AvgSco)$ 与 $Hig(Level, AvgSco)$ 在平均成绩上做等值连接^[6,7], 然后投影到两个关系的 $Level$ 上. 查询为:

Select Low.Level, Hig.Level

From Low as a, Hig as b

Where a.AvgSco=b.AvgSco (3)

通过式(3)连接后, $T(Low \bowtie Hig) = 20$.

将表 2、表 3 设计成等宽型直方图, 每 10 分设置一个宽度, 带宽为 10(桶), 如表 4. 一般而言, 如果两个连接属性有相同的带宽集(此时也有可能整个带宽内为空)的等宽直方图, 则可以通过估计相应带宽的每一对连接的大小并求和来估计连接的大小.

表 4 中, 只有三行是两个关系均不为 0, 其中带宽 61-70 的连接大小为 $(6 \times 6) / 10 = 3.6$ 个元组、带宽 71-80 的连接大小为 $(7 \times 6) / 10 = 4.2$ 个元组、带宽 81-90 则为 $(4 \times 11) / 10 = 4.4$ 个元组.

最终的连接大小为 $T(Low \bowtie Hig) = 3.6 + 4.2 + 4.4 = 12.2$. 与真实值较为接近.

表 2 低特质成绩表

Low (Level, AvgSco)

Level	AvgSco	Level	AvgSco
1	-	16	67
2	-	17	68
3	-	18	67
4	-	19	69
5	-	20	71
6	-	21	74
7	-	22	77
8	-	23	76
9	40	24	76
10	45	25	77
11	53	26	78
12	53	27	81
13	56	28	82
14	62	29	82
15	65	30	86

表 3 高特质成绩表

Hig (Level, AvgSco)

Level	AvgSco	Level	AvgSco
31	69	46	82
32	68	47	86
33	69	48	86
34	69	49	86
35	70	50	88
36	71	51	88
37	70	52	90
38	77	53	91
39	76	54	90
40	78	55	-
41	78	56	-
42	80	57	-
43	81	58	-
44	81	59	-
45	82	60	-

表 4 成绩带宽表

成绩范围	T(Low)	T(Hig)
31-40	1	0
41-50	1	0
51-60	3	0
61-70	6	6
71-80	7	6
81-90	4	11
91-100	0	1

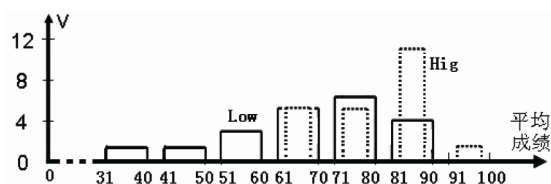


图 2 成绩直方图

采用统计法进行估计, 首先需确定两个表的 V 值, 其中 $V(\text{Low}, \text{AvgSco})=22$, $V(\text{Hig}, \text{AvgSco})=24$. $T(\text{Low})=30$ 、 $T(\text{Hig})=30$.

其连接估计值为: $30 \times 30 / \max(22, 24) = 38$.

结果与真实值之间有较大差距. 在表 4 中, 7 个带宽中两个关系的 V 值各有 0 值, 仅有 3 个共同有值. 优化器只知道每个关系分别有 30 个元组且分布于 1-100 之间, 致使统计结果产生较大误差. 此时直方图估算法显示出优势.

3.3 嵌套等宽型直方图

等宽型直方图最容易使用和维护, 但其精度稍差^[5]. 为此可以加以一定的改进. 具体做法是:

将图 2 中的某一个桶(如 71-80 桶)或若干个桶再进行划分(即嵌套), 比如划分成 71-75 和 76-80 两个子桶, 此时, 对于 71-75 子桶, 观察表 2 和表 3, 其连接大小为 $2 \times 1/5 = 0.4$, 而 76-80 子桶的连接大小为 $5 \times 5/5 = 5$, 最终结果为:

$$T(\text{Low} \cap \text{Hig}) = 3.6 + (0.4 + 5) + 4.4 = 13.4$$

精度得到了提高. 总体上说, 如果将整个值域(假设为 A)分成 $p(p \geq 1)$ 个相等的区间, 则为等宽型直方图, 每个区间称为一个桶(基桶). 如果欲进行桶嵌套, 对于那些频数变化不大且已符合准确度要求的基桶则不需变动. 对那些频数变化较大、尚不符合准确度要求的基桶, 再将它们进一步分成 q 个等宽子桶, 可不断地往下分, 直至所有的基桶、子桶均满足准确度要求为止. 在进行估算时, 基桶的带宽不再改变, 而子桶因为数据是动态变化的, 则根据需要应作相应的调整. 假设桶的最大嵌套层数为 r , 最底层的子桶含域值数为 k , 则直方图所表示的最大域值数 $v = k \times p \times q^{r-1}$. 经计算后, 若 A 中所含的值个数 $|A| = v$, 此时因无重复元组而不需特殊处理; 如 $|A| < v$, 则需为 A 添加 d 个虚假值, 使得 $|A| + d = v$, 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 $|A| > v$, 则需要增加 p 或 q 或 r 的值^[5].

4 结语

本文通过实例应用分析, 重点阐述了基于直方图的数据查询优化过程中的连接操作代价估计方法. 事实上, 连接操作代价的估计是十分复杂的, 没有一种准确的方法进行精确估计. 直方图估算法简单、明了而实用, 在等宽型数据结构下通过桶嵌套, 连接规模的估计与实际操作情况更为接近.

参考文献

- 1 朱焰炉, 程昌秀, 陈荣国, 颜勋. 基于直方图的空间查询选择率估计研究. 计算机科学, 2010, 12, 125-129.
- 2 Garcia-Molina H, Ullman JD, Widom J. Database System Implementation. Palo Alto, California: Stanford University, 2007. 274-277.
- 3 姜旭宝, 李光耀. 基于变宽直方图的无线传感器网络异常数据检测算法. 计算机应用, 2011, 31(3): 694-697.
- 4 曹巍, 王珊, 覃雄派, 王秋月. 面向不同数据分布的多维直方图算法 COCA-Hist. 计算机学报, 2008, 31(6): 1013-1023.
- 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): 286-289.
- 6 曲卫民, 孙乐, 孙玉芳. XML 数据查询中值匹配查询代价估计算法. 软件学报, 2005, 16(4): 561-569.
- 7 张晓琳, 戴华忠. 面向对象 XML 数据查询的代价估计研究. 计算机工程与应用, 2007, 43(18): 181-183.
- 8 刘云生, 彭楚冀, 廖国琼. 一种实时数据库查询执行方法的设计. 计算机应用, 2005, 25(2): 279-282.
- 9 樊新华. 关系数据库的查询优化技术. 计算机与数字工程, 2009, 37(12): 188-192.
- 10 徐丽萍, 金雄兵, 赵小松. 并行数据库查询优化技术研究. 华中科技大学学报, 2006, 34(3): 11-13.