

面向不同数据分布的多维直方图算法 COCA-Hist

曹 巍 王 珊 覃雄派 王秋月

(中国人民大学数据工程与知识工程教育部重点实验室 北京 100872)

摘 要 基于代价的 RDBMS 优化器需要对含有范围查询的合取谓词的结果集基数进行准确的估计,多维直方图对多维数据分布进行模拟,避免在估计结果集基数时采用数据独立性假设,造成估计误差过大,进而导致选择非优化的查询执行计划.在不同的数据分布情况下,传统的多维直方图(如 MHist-2)效果有很大不同.数据相关系数和值域密度、值域参数是准确刻画多维数据分布的有效指标,文中提出了根据不同的指标采用不同的动态优化的多维直方图算法 COCA-Hist,可以大大改善传统多维直方图在平均情况下的准确性.通过分析传统的多维直方图的最坏情况,COCA-Hist 的改进算法可以改进传统的多维直方图在最坏情况下的准确性.实验比较了 COCA-Hist 和传统的多维直方图 MHist-2 以及 GENHist 和 STHoles 的准确性和时间效率.实验显示无论在平均情况下还是在最坏情况下 COCA-Hist 的改进算法均优于传统的 MHist-2 直方图,并且 COCA-Hist 的准确性和创建时间均比 GENHist 有极大的改善,在准确性方面 COCA-Hist 较优于 STHoles,而在空间预算有限时 STHoles 的创建时间比 COCA-Hist 高两个数量级.

关键词 多维直方图;数据相关系数;值域密度;值域参数;属性值平均跨度

中图法分类号 TP311

Versatile Multidimensional Histograms for Different Data Distributions

CAO Wei WANG Shan QIN Xiong-Pai WANG Qiu-Yue

(Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872)

Abstract Traditional multidimensional histograms, which are widely used in cardinality estimation for conjunctive range query predicates in RDBMS's query optimizers, take the assumption of the existence of correlations among attributes instead of the plausible AVI assumption. But they do not further discriminate between different degrees of correlations among attributes. Based on accurate measurements of data distributions, data correlated coefficients and value domain density, the authors propose different optimal multidimensional histograms for different data distributions, COCA-Hist. Also they analyze the worst cases for traditional MHist-2 histograms and find effective ways to alleviate the situation. The authors conduct experiments to compare the accuracy and performance between COCA-Hist, and MHist-2, GENHist and STHoles. The results demonstrate that COCA-Hist histograms are superior in accuracy and performance than MHist-2 either in average case or in worst case. In the soft functional dependence situation, COCA-Hist is much better in either accuracy or building-up time by orders of magnitudes than GENHist. Under limited space budgets, COCA-Hist is one order of magnitude efficient than STHoles in building-up time. While STHoles exhibits good accuracy under sufficient space budget, in average COCA-Hist can achieve relatively better accuracy than STHoles.

Keywords multidimensional histograms; data correlated coefficients; value domain density; value domain parameter; average spread

收稿日期:2007-07-13;最终修改稿收到日期:2008-03-03. 本课题得到国家自然科学基金(60473069,60496325)和国际合作(HP Lab)项目“Large Scale Data Management”资助. 曹 巍,女,1975年生,博士研究生,研究方向为自管理自调优数据库、高性能数据库. E-mail: caowei@ruc.edu.cn. 王 珊,女,1944年生,教授,博士生导师,研究领域为高性能数据库新技术. 覃雄派,男,1971年生,博士研究生,讲师,研究方向为数据库优化技术与内存数据库技术. 王秋月,女,1974年生,博士,讲师,研究方向为数据库管理系统、XML 数据管理、XML 数据检索.

1 引言

关系数据库中的基于代价的查询优化器需要对含有范围查询的合取谓词的选择操作符 p (p 形如 $x_1 \leq X \leq x_2$ and $y_1 \leq Y \leq y_2$) 的结果集大小进行估计, X 和 Y 是数据库模式中的属性名. 最早采用的办法是假设属性 X, Y 之间相互独立, 则 $selectivity(p) = selectivity(X) \times selectivity(Y)$. 但是实际情况是数据相关性经常存在于属性之间, 数据独立性假设通常大大低估了选择率. 后果是依据不准确的选择率估计值选择的查询执行计划远远逊色于真正优化的查询执行计划.

为取代属性值独立假设 (Attribute Value Independence assumption, AVI assumption) 很多研究者提出了不同的方法, 比如采样^[1-2]、多维直方图^[3-4]、多维小波变换^[5]、概率关系模型^[6]等方法进行查询结果集大小或者选择率的估计. 其中多维直方图的方法是使用比较早而且在商用 DBMS 系统中经常采用的方法和技术.

另一方面, 相对于采用属性值独立性假设的另一个极端是笼统地采用相关性假设, 将不同程度的相关性一视同仁地对待, 这些都是比较粗略的办法.

本文研究利用两类统计信息: 数据相关系数和值域参数进一步分析刻画不同的数据分布, 根据不同数据分布特点 (数据相关性或者值域参数), 建立各自优化的多维直方图以达到比传统的多维直方图更高的准确性. 具体地说, 根据相关系数的计算结果, 可以准确地判断属性之间的相关性, 当相关系数显示两个属性之间具有一定程度的数据独立性时, 则可以直接利用 AVI 进行选择率的估计. 当值域参数显示属性间具有弱函数依赖时, 则可以用本文中提出的优化算法对弱函数依赖的情况创建优化的多维直方图.

另外通过对传统的 MHist 的最坏情况——属性值平均跨度相差过大的情况进行分析, 提出了改进 MHist 的最坏情况的算法.

实验验证了本文提出的改进的直方图算法 COCA-Hist 可以根据可靠的统计信息判断出不同相关程度的数据分布, 并进一步针对不同情况的数据分布提出优化的直方图算法, 并具有比 MHist 直方图在平均情况和最坏情况下更高的准确度. 另外本文中提出的针对弱函数依赖情况的改进算法不仅适用于两个属性的情况, 也同样适用于更高维的情况.

本文从下面几个方面对前述内容展开论述: 第 2

节是相关工作的分析和比较; 第 3 节介绍描述数据分布的重要统计信息: 数据相关系数 CF 和值域密度 $VDensity$, 值域参数 $VParam$; 第 4 节根据不同的数据分布给出优化的 COCA-Hist 多维直方图算法; 第 5 节分析传统多维直方图的最坏情况, 并给出对最坏情况的改进算法; 第 6 节给出在不同的数据分布情况和最坏情况下的 COCA-Hist 直方图算法的有效性和性能方面的实验结果.

2 相关工作

类似于文献[7]的划分方法, 多维直方图的研究可以划分为两种: 数据驱动的多维直方图和查询驱动的多维直方图.

数据驱动的多维直方图有等深多维直方图^[3,8]、基于边缘分布的 MHist 直方图^[4]、允许直方桶重叠的 GENHist 直方图^[9]等. 它们从数据出发, 依照不同的方法将数据空间划分成不同的桶, 以模拟数据分布.

等深多维直方图^[3]对数据空间的划分采用 Grid 结构, 而且划分时划分维度的顺序是事先确定好且无法变化的. 等深多维直方图对于数据分布矩阵稀疏的情况效果很不好, 因为这意味着一个桶 (bucket) 内部会出现大量的频率为 0 的单元格. 文献[8]依然采用了类似文献[3]的方法, 但是文献[8]采用了压缩 (compressed) 直方图的方法, 并讨论了如何确定单值 (singleton) 桶的个数以及优化选择划分属性的顺序和划分的 Grid 结构等问题, 给出了解决这些问题的启发性规则.

MHist-p 方法就是针对等深多维直方图的缺陷提出的改进算法, 它可以动态地选择待划分的区域和划分属性, 这样比文献[3]中事先确定的方法要灵活和准确得多; 但是它并没有解决一个桶 (bucket) 内部会出现大量的频率为 0 的单元格的问题, 即属性值分布稀疏的问题.

另外文献[10]专门讨论了属性值稀疏的情况, 提出了改进的直方图算法. 但是文献[10]仅限于一个属性的情况, 并且基于一维等宽直方图实现.

GENHist 直方图允许直方图的桶相交, 这样可以为更多的数据区域提供数据密度的计算, 为范围查询提供较高的准确性, 但是它的缺点是需要至少 5~10 次对数据空间的遍历, 并且输入参数较多, 而这些输入参数设置的好坏对直方图的准确性很有影响^[11].

查询驱动的直方图根据查询返回的准确结果

生成多维直方图, 比如允许桶之间的包含关系的 STHoles 直方图^[11]等。

查询驱动的直方图需要有一个用训练集“预热”的阶段, 在这个预热阶段直方图的准确性是较低的, 但是一旦经过预热阶段后, 查询驱动的直方图会为预热范围内数据空间的范围查询合取谓词提供非常准确的结果估计。但是在数据倾斜度较高的情况下, STHoles 直方图的准确性仍然低于 MHist^[9]。

STHoles 直方图和 GENHist 直方图允许直方图的桶按照任意的方式(比如桶之间可以重叠、包含等)划分数据分布空间, 但是这样会导致直方图为每个桶存放更多的位置信息, 在给定空间的条件下会影响直方图的准确性。文献[12]提出了压缩桶的位置信息的 STHoles 改进算法 STHoles+, 牺牲桶的精确位置信息获得更多数量的桶。

多维直方图是假设数据属性之间具有相关性的, 但是对哪些属性间具有相关性、相关的程度等问题还有专门的研究。

文献[5, 13]针对高维数据集利用马尔可夫图模型找出具有内部相关性和外部独立性的属性组, 并在此基础上提出了为高维数据创建的直方图。CORDS^[7]采用卡方检验的方法基于二维数据列联表判断两个属性之间的相关性。但是卡方检验的方法无法适用于数据分布稀疏的情况, 而且涉及到统计检验, 系统实现起来代价较高。文献[14]针对 CORDS 方法的局限, 提出了用熵相关系数进行二维数据相关性检验的方法。

我们发现, 对于不同的数据分布, 传统的直方图准确性是有差异的。本文要研究的是传统的 MHist-2 直方图的改进, 根据数据分布的不同, 采用不同的经过优化的直方图方法, 提高直方图的准确性, 尤其是改善最坏情况下的准确性。新的针对不同数据分布统计特征的多维直方图称为 COCA-Hist。

COCA-Hist 判断数据分布特征, 依据数据分布特征建立优化的多维直方图, 其算法流程如图 1 所示。

3 数据分布统计信息的描述

和 CORDS^[7]类似, 本文以二维数据分布的情况为例。原因有二: (1) 通常情况下维数越高, 数据联合分布矩阵越稀疏, 数据相关程度越高, 本文中弱函数依赖情况下的直方图算法同样可以适用高维情况; (2) 二维数据分布的情况可以直接使用熵相关系数对属性间的相关性进行刻画; 高维数据分布也可以转化为多个二维数据分布的情况来处理。当然高维数据分布, 其属性或属性组之间的相关性有很多组合的情况, 也可以用专门的统计学模型比如对数线形模型进行刻画。

基本定义: 对于有 n 个元组的关系 R , 其中两个属性 X 和 Y , 属性 X 在 R 中出现的不同值为 $(x_1, x_2, \dots, x_{|X|})$ 且 $x_{i_1} \leq x_{i_2}$ 若 $i_1 < i_2$; 属性 Y 在 R 中出现的不同值为 $(y_1, y_2, \dots, y_{|Y|})$ 且 $y_{j_1} \leq y_{j_2}$ 若 $j_1 < j_2$ 。属性 X 和 Y 的联合数据分布矩阵定义为二元组的集合

$$\Gamma_{X,Y} = \{((x_1, y_1), f_{11}), ((x_1, y_2), f_{12}), ((x_1, y_3), f_{13}), \dots, ((x_i, y_j), f_{ij}), \dots, ((x_{|X|}, y_{|Y|}), f_{|X||Y|})\},$$

其中, f_{ij} 是组合值 (x_i, y_j) 在 R 中的出现次数, $1 \leq i \leq |X|, 1 \leq j \leq |Y|$ 。

属性 X 的边缘数据分布定义为 $\Gamma_X = \{(x_1, f_{1.}), (x_2, f_{2.}), \dots, (x_i, f_{i.}), \dots, (x_{|X|}, f_{|X|.})\}$, $f_{i.}$ 为属性值 x_i 在 R 中的出现次数, $1 \leq i \leq |X|$ 。属性 Y 的边缘数据分布为 $\Gamma_Y = \{(y_1, f_{.1}), (y_2, f_{.2}), \dots, (y_j, f_{.j}), \dots, (y_{|Y|}, f_{.|Y|})\}$, $f_{.j}$ 为属性值 y_j 在 R 中出现的次数, $1 \leq j \leq |Y|$ 。属性 X 的跨度定义为 $S_i = x_{i+1} - x_i$, 当 $i < |X|$ 时, $S_i = 1$, 当 $i = |X|$ 时, 面积 A_i 定义为 $A_i = S_i \times f_{i.}$ 。

进一步地, 在 $\Gamma_{X,Y}$ 上建立的二维直方图是对 $\Gamma_{X,Y}$ 的近似, 定义为七元组 $(MINX, MINY, MAXX, MAXY, V\#X, V\#Y, FREQ)$ 的集合。其中 $MINX, MINY, MAXX, MAXY$ 是桶覆盖的矩形区域的顶点, $V\#X(V\#Y)$ 分别记录在该区域中属性 $X(Y)$ 的

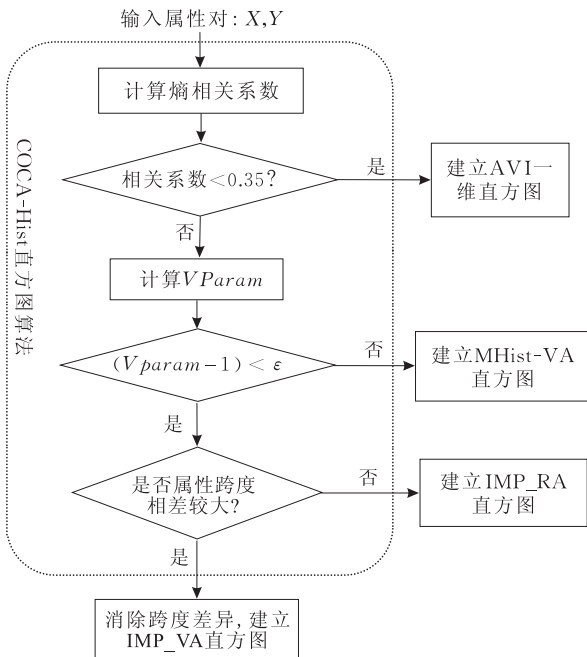


图 1 COCA-Hist 的处理流程

不同值个数, $FREQ$ 记录在该区域中的元组个数. 那么在一个桶内部, 假设这 $FREQ$ 个表中元组均匀分布在 $V \# X * V \# Y$ 个组合值上.

下面我们给出在联合数据分布矩阵基础上刻画数据分布的统计信息.

3.1 熵相关系数

文献[7]和文献[14]分别提出了根据属性对之间的相关程度, 为数据库选择需建立二维直方图的属性对的方法 CORDS 和 COCA. COCA 运用的熵相关系数形为^[15]:

$$\left\{ \begin{aligned} \rho'_{Y \rightarrow X} &= - \frac{\sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} f_{ij} \ln(n f_{ij} / f_{i.} f_{.j})}{\sum_{i=1}^{|X|} f_{i.} \ln(f_{i.} / n)} \\ \rho'_{X \rightarrow Y} &= - \frac{\sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} f_{ij} \ln(n f_{ij} / f_{i.} f_{.j})}{\sum_{j=1}^{|Y|} f_{.j} \ln(f_{.j} / n)} \end{aligned} \right. \quad (1)$$

熵相关系数是统计学上用从分散性减少来导出关联性的度量方法, 是从熵这一统计量导出的检验^[14-15], 它用来刻画属性 X 对属性 Y 的依赖程度和属性 Y 对属性 X 的依赖程度. 相关系数的取值范围为 $[0, 1]$. 当 $\rho'_{Y \rightarrow X}$ 和 $\rho'_{X \rightarrow Y}$ 均为 0 时, 表示属性 X 和属性 Y 之间完全无关(即属性值独立), 当 $\rho'_{Y \rightarrow X}$ 或 $\rho'_{X \rightarrow Y}$ 为 1 时, 代表属性 X 和属性 Y 的函数依赖关系. 相关系数越大, 代表属性间的依赖程度越大.

如图 2 所示的数据分布, 计算相关系数得到 $\rho'_{Y \rightarrow X} = 0.21$, $\rho'_{X \rightarrow Y} = 0.19$.

	$y_1(1)$	$y_2(5)$	$y_3(6)$	$y_4(8)$	$y_5(9)$	$f_{.j}$
$x_1(10)$	2	2	0	1	3	8
$x_2(11)$	2	4	4	0	3	13
$x_3(13)$	0	4	1	3	0	8
$x_4(19)$	4	4	2	1	0	11
$f_{i.}$	8	14	7	5	6	40

图 2 一个数据分布区域

3.2 值域密度和值域参数

(1) 值域密度和值域参数的定义

值域密度 $VDensity = \mathbf{I}_{X,Y}$ 中 $f_{ij} \neq 0$ 的二元组数 / $(|X| \times |Y|)$. $VDensity$ 的取值范围为 $[1/\min(|X|, |Y|), 1]$, 它描述的是联合数据分布矩阵的稀疏程度. 当 $VDensity$ 取最小值 $1/\min(|X|, |Y|)$ 时, 是联合分布矩阵最稀疏的情况, 此时一定是函数依赖的情况即属性 X 函数决定属性 Y , 如果 $|X| > |Y|$. 当 $VDensity$ 取上确界 1 时, 属性 X 和

属性 Y 的所有值的组合均在 R 中出现过, 但是属性 X 和属性 Y 具体相关程度却要视频率值 f_{ij} 的分布而定.

图 2 所示的数据分布, 其值域密度 $VDensity = 15/(4 \times 5) = 0.75$.

我们也可以如同文献[7]中所示, 采用值域参数 $VParam = \mathbf{I}_{X,Y}$ 中 $f_{ij} \neq 0$ 的二元组数 / $\max(|X|, |Y|)$ 来判断是否属于弱函数依赖的情况: 如果 $(VParam - 1) < \epsilon$, 则属于弱函数依赖, 如果 $VParam$ 取最小值 1, 则是函数依赖的情况. $VParam$ 的取值范围是 $[1, \min(|X|, |Y|)]$, 最大值 $\min(|X|, |Y|)$ 对应完全稠密的联合数据分布矩阵.

(2) 值域密度和值域参数的关系

值域密度 $VDensity$ 和值域参数 $VParam$ 都能够反映出在联合数据分布矩阵中是否存在弱函数依赖或者函数依赖的情况. 我们在 COCA-Hist 直方图算法中采用值域参数判断弱函数依赖的情况, 即当 $(VParam - 1) < \epsilon$ 时, 按照对弱函数依赖的优化算法处理.

(3) 熵相关系数与值域密度和值域参数

熵相关系数、值域密度和值域参数都是在联合数据分布矩阵上计算的, 但是从熵相关系数的计算公式可以看到熵相关系数的值是由值域分布和组合值的频率分布共同决定的, 因此它是准确描述属性间相关性的指标.

值域密度和值域参数仅描述组合值在值域上是否出现的分布状况, 它并不刻画出现多少次的频率信息, 因此仅可用来粗略反映某种特殊的数据分布, 比如弱函数依赖或者函数依赖的情况.

但是通过大量的测试数据集, 我们发现熵相关系数和值域参数两个统计信息也具有一定的相关性: 值域参数低则两个相关系数中一定有一个较高; 但值域参数高时, 两个相关系数可能均较低, 也可能较高. 相关系数均非常低(比如小于 0.01)时, 值域参数一定不会很低, 当相关系数较高时, 值域参数可能很高也可能很低.

为更好地说明这两个统计信息之间的相关性, 我们用表 1 刻画可能存在的不同组合情况.

在表 1 中值域参数和相关系数分别存在 4 种不同的取值情况, 说明如下:

(1) 值域参数的取值情况

- 1: 值域参数的最小值;
- 较低: 当 $1 < VParam < 1 + \epsilon$ 时;
- 较高: 当 $1 + \epsilon \leq VParam < \min(|X|, |Y|)$ 时;
- $\min(|X|, |Y|)$: 值域参数的最大值.

- (2) 相关系数的取值情况
- 0: 此时 $\rho'_{Y \rightarrow X}$ 和 $\rho'_{X \rightarrow Y}$ 均为 0;
- 较低: 当 $\max(\rho'_{Y \rightarrow X}, \rho'_{X \rightarrow Y}) < Threshold_1$ 时;
- 较高: 当 $Threshold_1 \leq \max(\rho'_{Y \rightarrow X}, \rho'_{X \rightarrow Y}) < 1$ 时;
- 1: 当 $\rho'_{Y \rightarrow X}$ 和 $\rho'_{X \rightarrow Y}$ 至少有一个为 1 时.

表 1 中 \surd 代表具有这种特征的数据分布存在, \times 代表具有这种特征的数据分布不存在. 例如, 当属性 X 和属性 Y 完全相互独立时, 其联合数据分布矩阵一定是稠密的, 即 $VParam = \min(|X|, |Y|)$ ($VDensity = 1$); 而当相关系数接近于 0 (较低) 时, 联合分布矩阵不可能非常稀疏 ($VParam = \text{'较低'}$).

表 1 相关系数和值域参数之间的相关性

值域参数	相关系数			
	0	较低	较高	1
1	\times	\times	\times	\surd
较低	\times	\times	\surd	\times
较高	\times	\surd	\surd	\times
$\min(X , Y)$	\surd	\surd	\surd	\times

基于上述分析, 我们将这两个统计信息结合起来刻画不同情况的数据分布, 为不同的数据分布情况分析传统的基于数据联合分布矩阵建立的多维直方图 MHist-2 的质量, 并提出优化的多维直方图 COCA-Hist 方法.

并且依照表 1 的分析, 我们在实验阶段生成了表 1 中所有 \surd 代表的数据分布, 来比较 COCA-Hist 和其它多维直方图算法的有效性和效率, 具体请见第 6 节.

4 优化的多维直方图

文献[4]中提出的 MHist-2 直方图从 P 开始, P 是将联合数据分布矩阵进行划分后得到的分区的集合, 初始时 P 里只有一个元素, 即联合数据分布矩阵 $\mathbf{F}_{X,Y}$, 创建方法如下:

(1) 为 P 中每一个分区, 根据其各属性的边缘分布, 找出最需要被进一步划分的分区 Γ' 及划分属性 Z (可以是属性 X 或属性 Y). 寻找的依据是不同的划分约束, 比如对于 $V\text{-Optimal}$ 的划分约束, 即找出边缘分布具有源参数的最大方差的分​​区及其对应属性; 对于 $MaxDiff$ 的划分约束, 即找出含有最大相邻源参数之差的分​​区及其对应属性.

(2) 找到这样的分区 Γ' 及划分属性 Z 后, 对分区 Γ' 沿属性 Z 的划分点将 Γ' 划分成 2 个更小的分区, 新产生的分区用来替换 P 中的 Γ' .

如此反复进行上述 2 个步骤, 直到 P 中的分区

个数等于给定空间可容纳的二维直方图的桶的个数. 此时将 P 中的每一个分区转换为二维直方图中的桶.

4.1 对该算法的评价

MHist-2 直方图未考虑不同的数据分布. 在数据联合分布矩阵稀疏的情况下准确率较低, 在数据近似相互独立的情况下不如 AVI 假设的方法. 这一点在后面的实验部分会加以说明.

另外正如文献[11]中的实验指出, MHist-2 直方图在采用 $MaxDiff(V, A)$ 类型的一维直方图刻画边缘分布时, 当不同的属性其属性值平均跨度相差很远时, 容易偏向平均跨度较高的属性维度, 造成不断按照该维度划分的情况^[11]. 对这个问题的改进, 我们在第 5 节展开论述, 在本节我们假设属性 X 和属性 Y 的平均跨度均属于同一个数量级.

我们采取的对策是: 在创建直方图时将不同数据分布的因素考虑进去, 分析不同数据分布的情况, 为不同的数据分布提出不同的优化直方图算法.

4.2 不同的数据分布

根据第 3 节中给出的两类统计信息, 我们将其结合起来分析不同的数据分布的情况.

(1) 数据近似相互独立

它是指满足 $\max(\rho'_{Y \rightarrow X}, \rho'_{X \rightarrow Y}) < Threshold_1$ 的情况. 根据计算出来的相关系数可以准确判定属性 X 和 Y 是否接近相互独立. 如果是, 则可以用 AVI 的方法, 分别为属性 X 和 Y 建立边缘分布的一维直方图.

经过实验我们发现, 在数据近似相互独立的情况下, 使用 AVI 方法进行的结果集基数估计误差比 MHist-2 方法的误差要小得多. 我们方法的优势在于可以根据计算出来的相关系数的大小准确地判断属性间相互独立的状况. 实验检验结果表明, 可将相关系数低于 0.35 的情况均视为属性间相互独立.

(2) 弱函数依赖的情况

它是指 $VParam < 1 + \epsilon$ 的情况. 由观察我们发现此时分布矩阵中有很多 f_{ij} 为 0 的单元. 如果在 MHist-2 多维直方图的创建过程中一个频率非 0 的桶覆盖很多的 0 单元格, 则会导致该桶的准确性下降; 如果生成一个频率为 0 的直方桶, 则可以将其丢弃, 不必为它分配桶空间.

联合分布矩阵最稀疏的情况对应于属性 X 和属性 Y 之间的函数依赖.

经过实验发现在此种数据分布的情况下, 传统的 MHist-2 直方图误差很大, 原因在于数据分布矩阵中“空格子”(f_{ij} 为 0 的单元格)的比例很高. 针对

这种情况,我们采用了三种途径改进传统的 MHist-2 直方图的质量,具体在 4.3 节中介绍.改进的算法在 $\epsilon \leq 1$ 时比传统的 MHist-2 方法准确度提高了 3 倍.

(3) 其它的情况

它是指数据的相关性既没有达到近似相互独立,数据分布矩阵又不够稀疏,无法用前述两种优化的直方图解决的情况.此时我们可以用传统的 MHist-2 方法达到较高的估计准确度.

我们分别生成了这几种不同类型的数据分布矩阵.为不同的数据分布矩阵建立了不同的优化的多维直方图,来验证不同的直方图在不同数据分布下的质量好坏.

4.3 改进的直方图算法 COCA-Hist

我们从以下几个方面对弱函数依赖情况下原有的 MHist-2 直方图算法进行改进:

(1) 改进数据分布矩阵中待划分区域的选择指标

如文献[4]所示,采用 $MaxDiff(V, A)$ 和 $MaxDiff(V, F)$ 作为建立 MHist 直方图时边缘数据分布的划分方法,能生成准确度很高的多维直方图,而且建立方法比较简单容易.但是作为待划分区域的选择依据时,我们采用边缘分布的规范化最大相邻源参数之差替换原有算法中的最大相邻源参数之差来作为 P 中待选择区域和相应维度的分数,从中选出分数最大的区域和维度作为待划分区域和划分属性.

具体规范化的做法是将原来的边缘分布中的相邻频率或者面积之差的最大值除以该边缘分布中的频率或者面积的最大值.

下面用图示说明:取前面图 2 中所示的数据分布矩阵 T_{XY} 的一个区域,属性 X 在这个区域有 4 个不同的取值,属性 Y 有 5 个,分别用 $()$ 标出,边缘分布用黑框标出.

以属性 Y 的边缘分布为例,基于 $MaxDiff(V, F)$ 的原有方法如图 3 所示.

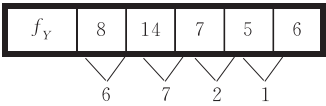


图 3 $MaxDiff(V, A)$ 方法下的分数计算

$MaxDiff(V, F)$ 方法中最大相邻源参数之差即最大相邻频率之差,在图 2 中为 7. 所以该区域在属性 Y 情况下的分数为 7. 我们的方法将分数用规范化的最大相邻源参数计算,此例中得到 7 后,再除以边缘分布中的最大频率值 14,所以用改进后的方

法,该区域在属性 Y 的情况分数为 $7/14 = 0.5$. 每一个区域在每一个属性上都计算分数,分数最大的区域和属性就是待划分区域和划分属性.

属性 Y 的 $MaxDiff(V, A)$ 方法的分数计算如图 4 所示.

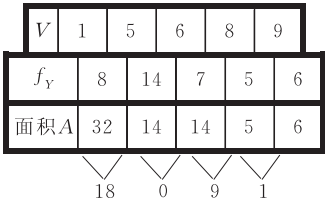


图 4 $MaxDiff(V, A)$ 方法下的分数计算

当面积作为源参数时,我们用规范化后的最大面积之差作为分数,计算方法为最大面积之差/最大面积,即 $18/32 = 0.56$ 作为分数代替原来的分数 18.

我们把原来以及改进后的基于 $MaxDiff(V, F)$ 和 $MaxDiff(V, A)$ 的 MHist-2 直方图算法分别简化表示为 MHist_VF, MHist_VA, MHist_RF^①, MHist_RA^①. 用规范化的相对指标选择待划分区域和划分属性,可以避免对某一维度的偏好,实验证明这种方法在弱函数依赖的情况下比传统的 MHist_VF, MHist_VA 直方图质量高. 这种方法也可以称为“规范法”.

(2) 避免在划分数据分布区域时产生频率为 0 的空桶

通过实验我们发现,在数据分布矩阵稀疏的情况下,各种多维直方图算法都很容易产生大量空区域(该区域中没有任何组合值落入). 这些空区域既不会被进一步划分,又不能生成对应的直方桶(在有限的直方图空间中,对应全空区域的桶无疑不提供任何统计信息并且还占用空间). 改进的算法将检查创建过程中是否产生空区域,如果产生空区域则将其丢弃,将空间分配给更具有统计意义的其它区域. 这种方法可称为“舍 0 法”.

(3) 在一个数据分布区域内部改进划分方法

在数据分布矩阵稀疏的情况下,一个区域即使不是全空区域,可能在其内部也有若干相当大的空区域. 如图 5 所示.

该区域由划分产生,虽然非全空区域,但图中三个阴影部分是频率值全部为 0 的空区域,这样的空区域对于查询范围可能落在其中或者部分相交的联合选择谓词来说误差较大.

我们提出的改进算法 3 可以将这样的空区域识

① R 在这里表示相对的(Relative).

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	f_X
x_1	0	0	0	0	0	0	0	0	0
x_2	0	0	0	0	0	9	0	0	9
x_3	0	0	0	5	0	0	0	0	5
x_4	0	0	0	0	0	0	0	0	0
x_5	0	0	0	0	0	0	0	3	3
f_Y	0	0	0	5	0	9	0	3	17

图 5 改进数据分布区域内部的划分方法

别出来并将其丢弃,划分出分布更为紧凑的非空区域来,在图中即两个非相邻的白色背景的矩形区域.因此方法 3 也可称为“挤榨法”(squeezing).

具体算法是在为每个区域的每个边缘分布计算分数时,检查在边缘分布中有没有一串连续的 0,如果有,则可以判定存在空区域;丢弃该空区域并产生可能的重新划分(当空区域位于边缘分布的中间而不是首尾时会产生新的划分).

方法 3 可以和 MHist_VF, MHist_VA, MHist_RF, MHist_RA 方法中的任何一种方法结合起来,在不同方法得到的区域中将全为 0 的连续区域挤出去.因此表示为 IMP_VF, IMP_VA, IMP_RF, IMP_RA.

这种改进方法的一个问题是当算法进行到最后,对最后一个待划分区域如此划分时,有可能划分出来的区域总数超过了给定空间能容纳的桶的个数.通过实验发现,大多情况下,改进方法的直方桶个数一容许的直方桶个数 ≤ 5 .对 PC 机的 VC 编译环境下,不超过 150 个字节.

还可以设置并调整参数,控制对边缘分布上连续“0”串的长度检查:例如对某个区域属性 B 上的边缘分布,只有当连续的“0”串的长度 $\geq |B| \times Empty_factor$ 时,才将空区域择出并丢弃.通常 $Empty_factor$ 越小(比如 0.01),直方桶的个数越容易越界,要使改进后的直方桶个数严格控制在可允许范围内,可根据 $|X|$ 或 $|Y|$ 的大小将 $Empty_factor$ 设置为 0.1 或 0.15.更合理的参数设置方法有待今后进一步深入研究.

“挤榨法”对于稀疏的数据分布非常有效.比如传统的 MHist_VA 方法没有区别对待稀疏数据分布的情况.对于图 5 所示的数据分布矩阵,在给定 buckets 个数为 3 的情况下,传统的 MHist_VA 方法很可能生成这样的含有 3 个 bucket 的直方图: $\{((x_1, y_1), (x_5, y_5), 5), ((x_1, y_6), (x_5, y_6), 9), ((x_1, y_7), (x_5, y_8), 3)\}$.我们发现,这 3 个 bucket 中,都各自只含有一个非 0 单元格, bucket 覆盖的

区域中却含有很多 0 单元格.在这种结构上估算谓词 $P: x_1 \leq X \leq x_4 \text{ AND } y_1 \leq Y \leq y_3$ 的选择集大小为 2.4(因为假设不同值在 bucket 的值域范围内均匀分布,如果假设 x_i 和 y_j 也是均匀的话,结果集大小应为 $5 \times (3 \times 4) / 25 = 2.4$),而实际上这个谓词的结果集大小应该为 0.

如果采用“挤榨法”,会产生 3 个精确的 bucket: $\{((x_3, y_4), (x_3, y_4), 5), (x_2, y_6), (x_2, y_6), 9), (x_5, y_8), (x_5, y_8), 3)\}$.在这个直方图基础上,谓词 $P: x_1 \leq X \leq x_4 \text{ AND } y_1 \leq Y \leq y_3$ 的选择集大小为 0,即真实的结果集大小.

“舍 0 法”和“挤榨法”都可以归为“Empty region aware”的方法,旨在减少空单元格对直方图准确性的负面影响,提高弱函数依赖时的直方图准确度.方法 1(“规范法”)也在数据接近弱函数依赖时比较有效,后面我们用实验分别说明改进的多维直方图的效果.

在进行实验时,我们有一组实验比较采用“舍 0 法”和不采用“舍 0 法”对各种类型的直方图产生的频率为 0 的桶的个数的比较.因为各种方法的直方图均会产生一定数目的空桶,对所有类型的直方图我们都采用了“舍 0 法”进行优化.

通过实验我们发现在数据具有弱函数依赖的情况下 MHist_RF 和 MHist_RA 的效果要好过传统的 MHist_VF 和 MHist_VA 方法很多,因此我们进一步地将 IMP_RA 用在弱函数依赖的 COCA-Hist 直方图创建过程中.

结合 4.2 节和 4.3 节的内容,我们提出的基于不同的数据分布的改进直方图(COCA-Hist)创建算法框架如图 6 所示.

算法. COCA-Hist.
输入: 关系 R 、属性 X 和属性 Y 及其联合数据分布矩阵、空间限制 S 和浮动空间大小 $\Delta (\Delta \leq 200B)$
输出: 总体上优化的二维直方图 COCA-Hist
算法描述:
if $\max(\rho'_{Y \rightarrow X}, \rho'_{X \rightarrow Y}) < Threshold_1$ then
 分别在属性 X 和属性 Y 的边缘分布上建立一维直方图,一维直方桶个数按照 $|X| : |Y|$ 的比例来划分;
else if $(VParam - 1) < \epsilon$
 按照 IMP_RA 算法,利用方法 1~3,创建优化的直方图
else
 利用优化方法 2,创建 MHist_VA 类型的直方图

图 6 COCA-Hist 直方图算法

5 对最坏情况的改进

如文献[11]提出,基于 $MaxDiff(V, A)$ 的

MHist-2 直方图在最坏情况下会一直沿着同一个维度划分数据分布矩阵. 通过对这种最坏情况的模拟和分析, 我们发现了这种情况产生的原因, 提出了改进最坏情况的解决方法, 并用实验验证了在最坏情况下我们提出的方法的有效性.

我们发现当两个属性的属性值平均跨度相差较大(比如大于 100 倍), 基于 $MaxDiff(V, A)$ 的 MHist-2 直方图的最坏情况很容易发生. 比如工资和年龄这两个属性, 年龄的值域范围是 20~60, 其相邻属性值的跨度不会超过 5; 但是工资字段如果是 500 元为一个级别, 则相邻属性值的跨度最少为 500.

之所以说很容易发生而非确定发生, 还要看具体的频率分布. 因为我们知道, $MaxDiff(V, A)$ 中的面积 $A_i = S_i \times f_i$. S_i 即跨度, f_i 为边缘频数. 只有在数据相关程度非常高即弱函数依赖时, 属性跨度的差别才确定地突出显示出来.

判别属性值跨度差别的方法有很多, 我们可以用上下界检验法或者属性值平均跨度方法检验跨度的差别. 比如属性值平均跨度检验法可表示为, 若

$$\max\left(\frac{1}{|X|-1} \sum_{i=1}^{|X|-1} S_i^X, \frac{1}{|Y|-1} \sum_{j=1}^{|Y|-1} S_j^Y\right) / \min\left(\frac{1}{|X|-1} \sum_{i=1}^{|X|-1} S_i^X, \frac{1}{|Y|-1} \sum_{j=1}^{|Y|-1} S_j^Y\right) \geq 100,$$

则两个属性有显著的跨度差异.

我们针对弱函数依赖的情况, 消减属性的跨度差异. 具体做法是选出属性值平均跨度大的属性, 假设为属性 Y , 选择一个单调递增函数 f , 将属性 B 的每一个跨度值 S_j^Y 转换为 $\hat{S}_j^Y = f(S_j^Y)$, 使转换后的 \hat{S}_j^Y 尽量落在和属性 X 的平均跨度值 \bar{S}^X 相当的范围内. 在创建多维直方图的过程中, 根据 $MaxDiff(V, A)$ 的方法进行划分时, 在属性 Y 的维度上计算面积时, 我们采用 \hat{S}_j^Y 代替原来的 S_j^Y 计算面积, 并据此选择待划分区域和划分属性.

采用了这种方法生成的 COCA-Hist 多维直方图可以避免发生只在一维上划分数据分布矩阵的情况, 使划分属性的选择在两个属性之间实现公平.

如实验显示, 若 $0 < (VParam - 1) < 1$ 时, 改进属性值跨度算法是有效的, 而且在修正属性值跨度之后的 IMP_VA 算法具有最好的准确度; 当 $1 < (VParam - 1) < 2$ 时, 改进属性值跨度算法是有效的, 但是此时 MHist_VF 算法始终具有最好的准确度; 其它情况下改进属性值跨度不会带来好处, 并且传统的 MHist_VA 直方图具有较好的准确度.

6 实 验

我们的实验生成了大量属于不同分布情况的数据, 验证 4.2 节、4.3 节和第 5 节提出的算法的有效性.

6.1 生成的实验数据

我们尝试不同的值域大小(值域范围中等: $|X|$ 和 $|Y|$ 为 800/500; 或者值域范围较大: $|X|$ 和 $|Y|$ 为 4000/2000 或 6000/4000 等), 通过调整不同的参数, 生成了表 1 中由 \checkmark 代表的不同情况的联合数据分布矩阵. 另外我们也尝试了当两个属性的属性值跨度相近和相差比较大(两个属性的属性值平均跨度相差 100 倍和 10000 倍不等)的不同情况.

6.2 评价的查询负载

我们生成了不同的查询负载检验不同直方图的准确性: 生成具有一定规律的联合范围查询的负载集合 W , 针对 W 中的每一个形如 $p: x_1 \leq X \leq x_2$ and $y_1 \leq Y \leq y_2$ 的查询(点查询可以看作是范围查询的特例当 $x_1 = x_2$ 且 $y_1 = y_2$ 时), 计算用不同的直方图 $Hist_i$ 进行估算的结果集大小 $est(p)$, 再根据 p 的真实结果大小 $act(p)$, 计算出绝对误差:

$$Err_{abs}(W, Hist_i) = \sum_{p \in W} |est(p) - act(p)|.$$

为了便于比较多种直方图针对不同的查询结果集的准确率, 我们采用相对误差:

$$Err_{rel}(W, Hist_i) = Err_{abs}(W, Hist_i) / \sum_{p \in W} act(p).$$

为了使实验的结果尽可能客观, 我们生成的查询负载集合通常容量较大(2000, 3000, 20000 个查询). 并且每一个查询的 x_1 和 y_1 都在值域范围内随机产生, 不同维度上查询范围的跨度 $x_2 - x_1$ 和 $y_2 - y_1$ 也都是随机生成的. 这样做的目的是使生成的查询负载尽可能模拟各种不同情况下的即席查询(ad-hoc queries).

6.3 实验平台

实验环境是 PC 机, Intel Pentium 4 CPU, 2.8GHz, 1GB 内存, Windows 2000 操作系统, 编译器 Microsoft VC++ 2005.

6.4 实验结果

(1) 近似数据独立的实验

此时, MHist_VA 具有较高的准确度; 进一步我们比较在近似相互独立的不同程度下, 采用 AVI 方法的 COCA-Hist 和 MHist_VA 方法的绝对误差的比值, 图 7 显示给定 80KB 空间时直方图准确性

的比较,但是随着给定空间的增大(比如 120KB),这种比值会进一步缩小,也就是说 COCA-Hist 方法的准确度会比 MHist_VA 提高更多的倍数.

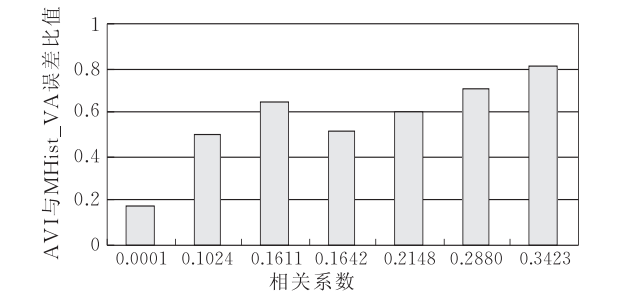


图 7 数据近似相互独立时的准确度比较

图 7 中纵坐标表示 AVI 方法得到的查询误差与 MHist_VA 方法得到的查询误差的比值. 这个比值小于 1,说明 AVI 方法的估算准确度高于 MHist_VA 方法.

这组实验也是我们初步将 $Threshold_l$ 的值设为 0.35 的依据. 我们看到在相关系数 <0.35 的 7 组数据中, AVI 假设仍然是适用的,它的准确度比 MHist_VA 要高得多,而且创建一维直方图比创建二维直方图的代价要小得多(见表 2).

以数据中属性 X 和 Y 的不同值个数为 4000 和 2000 的规模为例,在存在近似数据独立性的情况下不同类型直方图的创建时间见表 2(单位:ms).

表 2 属性间相互独立情况下直方图创建时间比较	
直方图	相关系数
	0.000024
MHist_VF	13531
MHist_VA	14969
MHist_RF	10828
MHist_RA	8859
MHist_Imp	156

其它规模的数据情况下,不同直方图创建时间的比较趋势类似,可以看出,如果有充分的证据说明数据之间具有相互独立性,采用 AVI 假设不仅可以获得较好的估计准确率,而且还会避免创建二维直方图的代价.

(2) 弱函数依赖的实验

根据条件 $(VParam-1)<\epsilon$,此时我们设 $\epsilon=1$,且假设两个属性之间的属性值平均跨度大体相当. 这组实验验证 4.3 节的改进算法,显示 IMP_RA 和 MHist_VA 的绝对误差的比值,如图 8 所示.

我们从大量实验数据中选择 8 组数据说明弱函数依赖的情况. 我们提出的 COCA-Hist 直方图改进算法在全部数据中均比传统的 MHist_VA 算法优

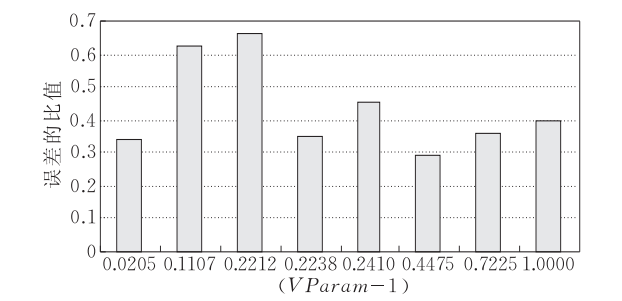


图 8 弱函数依赖时准确度的比较

越,准确率平均提高至 2.5 倍以上,最多情况改进了 3.5 倍.

以数据中属性 X 和 Y 的不同值个数为 4000 和 2000 的规模为例,在存在弱函数依赖的情况下不同类型直方图的创建时间见表 3(单位:ms).

表 3 弱函数依赖情况下直方图创建时间比较					
VParam	创建时间/ms				
	MHist_VF	MHist_VA	MHist_RF	MHist_RA	IMP_VA
0.1	5438	10781	3734	4235	2031
0.5	4781	11563	3734	4250	2360
0.75	4469	10828	3562	3781	2594

通过实验我们发现,在不同的数据规模的情况下,当数据的分布呈现弱函数依赖时,均具有相同的趋势——IMP_RA 方法创建多维直方图的时间最短.

(3) 改进最坏情况的实验

当属性值跨度相差较大时,我们的改进算法对于弱函数依赖的情况有较大改进. 下面我们给出 $(VParam-1)<1$ 情况下改进算法的实验结果,如图 9 所示.

图 9 中可以看出,在弱函数依赖的情况下,COCA-Hist 采用的修正属性值跨度的 IMP_VA 方法的准确度比传统的 MHist_VA 方法平均提高 6.76 倍,最好的情况是改进了 11 倍多.

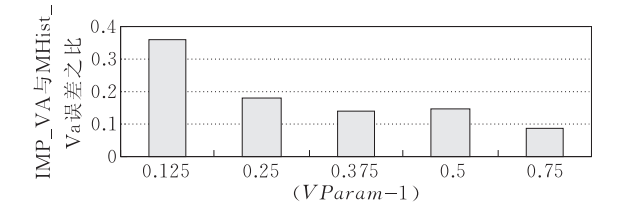


图 9 COCA-Hist 对最坏情况的改进

修正了属性值跨度的差异后,COCA-Hist 采用针对弱函数依赖的挤榨法,得到了更准确的估计方法. 同时 COCA-Hist 方法的时间代价与其它直方图算法比起来是最低的. 以数据中属性 X 和 Y 的不同值个数为 4000 和 2000 的规模为例,具体见表 4(单位:ms),其它规模的数据具有同样的趋势.

表 4 在属性值跨度差异较大情况下
各种直方图创建时间的比较

VParam	创建时间/ms				
	MHist_VF	MHist_VA	MHist_RF	MHist_RA	IMP_VA
0.125	6578	5218	3860	4094	1765
0.5	4703	4828	3500	3985	2000
0.75	4547	4875	3500	4219	2203

值得指出的是,实验显示在属性间值域跨度相差较大的情况下,相关系数显示属性间近似独立的情况时,采用 AVI 方法并不比 MHist_VA 方法有明显的优势.原因在于属性间跨度相差很大时,MHist_VA 可以更好地支持各种组合范围查询,特别是当查询结果集大小为 0 时,通常 MHist_VA 方法可以准确地估计,即此时 MHist_VA 的 false positive 较低.实验显示当相关系数较低(小于 0.35)时 AVI 方法比 MHist_VA 误差率只高出几个百分点.篇幅所限这部分实验结果略去.

(4) 与多维直方图 GENHist 和 STHoles 的比较

前面我们比较了在近似数据相互独立和弱函数依赖的分布情况下 COCA-Hist 与传统的 MHist-2 多维直方图算法的比较,并比较了不同的联合分布下值域分布最坏情况的改进.在得出 COCA-Hist 是更智能更优化的多维直方图的结论之前,我们还有几组模拟实验比较 COCA-Hist 和著名的 GENHist 及 STHoles 进行有效性和性能方面的比较.

GENHist^[9]的提出是针对实数数域上的联合分布,这种分布的特点是一维属性值很少有重复,这种分布情况对应于具有函数依赖或者弱函数依赖的联合数据分布矩阵.因此我们比较了在函数依赖和弱函数依赖情况下 COCA-Hist 与 GENHist 和 STHoles 直方图的效果(图 10)与性能(图 11).

这组实验是在规模为 4000×2000 的联合数据分布矩阵中完成的,并且给定的直方图预算空间是 80KB,GENHist 中初始划分参数 ξ 为 8,STHoles 采用了 1000 个随机产生的训练查询,经过训练,在给定空间预算下生成 STHoles 直方图.

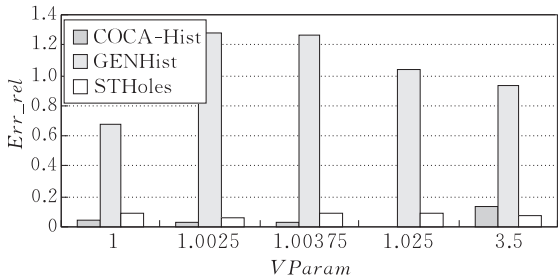


图 10 COCA-Hist, GENHist 与 STHoles 准确率比较

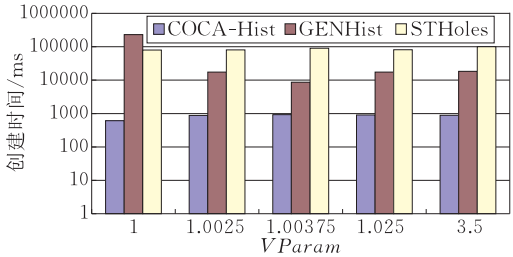


图 11 COCA-Hist, GENHist 与 STHoles 创建时间比较

三种直方图均由 10000 个随机的查询检验直方图的准确性.

从准确率来看,在弱函数依赖的情况下,COCA-Hist 采用的挤榨法非常有效地排除了数据分布矩阵中“0”单元格的干扰,有效提高了多维直方图的准确度,并且准确率总体上高于 STHoles,大大高于 GENHist.

从算法性能来看,COCA-Hist 是无论时间效率还是空间效率都非常高效. GENHist 由于需要多次遍历联合分布矩阵,时间效率比 COCA-Hist 要低一个数量级;而 STHoles 算法需要动态地维护树结构,并且为了保证直方图的准确性,不可能采用较少的训练查询集来创建直方图,另外当直方图中直方桶的个数超过空间预算时,还需要进行繁琐的合并(merge)操作,这不仅花费时间,而且为了合并而计算代价(penalty)也需要维护与直方图同构的一个树,因此 STHoles 算法的时间效率比 COCA-Hist 低两个数量级.

但从模拟的效果来看,STHoles 的一个特点就是准确率在不同的充足空间预算下比较稳定,而创建时间却抖动很大. 我们有一个实验在预算 80KB 和 100KB 的不同情况下,STHoles 的相对误差率变化很小,分别为 0.061 和 0.064,但是创建时间却发生了数量级的变化,从 112953ms 下降到 828ms. 这是因为空间预算增加后,STHoles 可以避免很多耗时的 merge 操作. 而同样的空间预算改变,却使 COCA_Hist 的相对误差率从 0.139 下降到 0.057,但是创建时间却从 1359ms 提高到 2109ms,这是因为更多的直方桶需要更多的时间来生成. 如图 12 所示.

对于其它情况的数据分布,我们也分别做了实验. 在近似函数依赖和其它情况下,我们都得到了类似的结论:COCA-Hist 能够在很短的时间内创建出准确率总体优于 STHoles 的多维直方图,而无论在时间效率还是在准确率方面,COCA-Hist 都远优于 GENHist. 限于篇幅,其它实验结果略去.

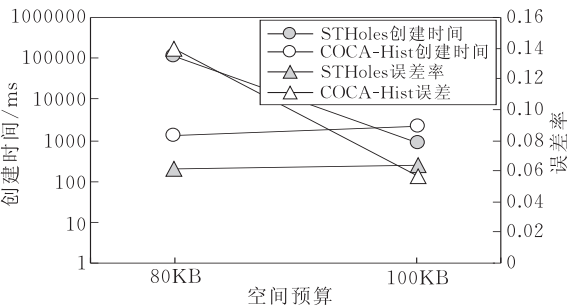


图 12 不同的空间预算对误差率和创建时间的影响

通过实验比较,我们得出结论,STHoles 的优势在于直接利用查询的结果,可以无需事先知道任何关于数据分布的信息,创建出准确的直方图,但是训练查询的时间代价却不可忽略;COCA-Hist 是数据驱动的多维直方图创建方法,通过重要的统计信息描述不同数据分布特征,优化地采用不同的创建多维直方图的算法,可以在很短的时间内创建出更准确的多维直方图。

7 总结与未来工作

本文提出了针对传统的 MHist-2 直方图进行改进的算法 COCA-Hist. 选择 MHist-2 直方图进行改进的原因在于它直接基于数据分布矩阵创建多维直方图,可以和其它的统计信息比如相关系数、值域密度等很自然地结合起来考察不同数据分布情况下的准确度,另一方面 MaxDiff 族的直方图在一些著名的商用 DBMS(比如 SQL Server)中广泛使用。

本文针对数据近似相互独立和弱函数依赖的情况下给出了改进的直方图 COCA-Hist 算法,实验显示 COCA-Hist 直方图算法比传统的 MHist-2 多维直方图在平均水平上具有更高的准确度,并极大改观了 MHist-2 多维直方图的最坏情况。与同类的多维直方图 GENHist 和 STHoles 相比,COCA-Hist 的时间效率与准确率大大高于 GENHist,并且总体准确率略好于 STHoles. GENHist 在初始划分参数 ξ 为 8 时的时间效率高于 STHoles,但是当 ξ 增大时需要反复遍历联合数据分布矩阵,时间效率又远低于 STHoles。

未来还有一些细致的工作需要进一步完善,比如自动地选择单调递增函数 f ,使转换后的属性值跨度差别尽可能小。另外可以结合 Compressed 直方图的方法对 COCA-Hist 进行改进,利用单值桶类型改善 COCA-Hist 的空间利用。例如在 4.3 节的例子中,用“挤榨法”为图 4 所示的分布矩阵建立改进后的直方图,我们发现在数据分布矩阵稀疏的情况

下,常常会有一些桶(bucket)从一个矩形区域缩成了一个点(单值桶),这样存储的信息容量就会减少一半,如果能够动态地将单值桶节省下来的存储空间回收,可以在给定的空间预算中创建更多的桶(bucket),从而提高改进后多维直方图的准确性。

参 考 文 献

[1] Gibbons P B, Matias Y, Poosala V. Fast incremental maintenance of approximate histograms//Proceedings of the 23rd International Conference on Very Large Data Bases. Athens, Greece, 1997; 466-475

[2] Ganguly S, Gibbons P B, Matias Y, Silberschatz A. Bifocal sampling for skew-resistant join size estimation//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Montreal, Quebec, Canada, 1996; 271-281

[3] Muralikrishna M, DeWitt David J. Equi-depth histograms for estimating selectivity factors for multi-dimensional queries//Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data. Chicago, Illinois, USA, 1988; 28-36

[4] Poosala V, Ioannidis Y. Selectivity estimation without the attribute value independence assumption//Proceedings of the 23rd International Conference on Very Large Data Bases. Athens, Greece, 1997; 486-495

[5] Lim L, Wang M, Vitter J S. SASH: A self-adaptive histogram set for dynamically changing workloads//Proceedings of the 29th International Conference on Very Large Data Bases. Berlin, Germany, 2003; 369-380

[6] Getoor L, Taskar B, Koller D. Selectivity estimation using probabilistic models//Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. Santa Barbara, USA, 2001; 461-472

[7] Ilyas I F, Markl V et al. CORDS: Automatic discovery of correlations and soft functional dependencies//Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. Paris, France, 2004; 647-658

[8] Pham H T A, Sevcik K C. Structure choices for two-dimensional histogram construction//Proceedings of the 2004 Conference of the Centre for Advanced Studies on Collaborative Research. Markham, Ontario, Canada, 2004; 13-27

[9] Gunopulos D, Kollios G, Tsotras V, Domeniconi C. Approximating multi-dimensional aggregate range queries over real attributes//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas, Texas, USA, 2000; 463-474

[10] Oommen B J, Chen J. A new histogram method for sparse attributes: The averaged rectangular attribute cardinality map//Proceedings of the 1st International Symposium on Information and Communication Technologies. Dublin, Ireland, 2003; 119-125

[11] Bruno N, Chaudhuri S, Gravano L. STHoles: A multidimensional workload-aware histogram//Proceedings of the

2001 ACM SIGMOD International Conference on Management of Data. Santa Barbara, USA, 2001; 211-222

[12] Fuchs D, He Z, Lee B S. Compressed histograms with arbitrary bucket layouts for selectivity estimation. *Information Sciences: An International Journal*, 2007, 177(3): 680-702

[13] Deshpande A, Garofalakis M. Independence is good: Dependency-based histogram synopses for high-dimensional data//*Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*. Santa Barbara, USA, 2001; 199-210

[14] Wang Shan, Cao Wei, Qin Xiong-Pai. COCA—A new way

to auto-detect association based on entropy correlated coefficients. *Computer Applications*, 2006, 26(9): 2005-2008(in Chinese)

(王珊, 曹巍, 覃雄派. 基于熵相关系数的关联性自动判别方法 COCA. *计算机应用*, 2006, 26(9): 2005-2008)

[15] Zhang Yao-Ting et al. *Statistical Analysis of Qualitative Materials*. Guangxi: Guangxi Normal University Press, 1991(in Chinese)

(张尧庭等编著. 定性资料的统计分析. 广西: 广西师范大学出版社, 1991)



CAO Wei, born in 1975, Ph. D. candidate. Her research interests include self-managing and self-tuning data bases, and high performance databases.

WANG Shan, born in 1944, professor, Ph. D. supervisor. Her research interests include new technologies of high performance databases.

QIN Xiong-Pai, born in 1971, Ph. D. candidate, lecturer. His research interests include database optimization technologies and main memory databases.

WANG Qiu-Yue, born in 1974, Ph. D. , lecturer. Her research interests include database management systems, XML data management and XML data retrieval.

Background

Multidimensional histograms are most influential and important for relational database query optimizers to accurately estimate the cardinality of the results of query predicates. In self-managing and self-tuning databases, many efforts have been put into improving the quality of query execution plans produced by query optimizers. So the statistics which are heavily relied on by query optimizers are key techniques of self-managing databases.

In the database research community, many excellent kinds of histograms have been proposed to replace the fallible uniform distribution and attribute value independent assumptions. They start from what histograms should be like, rectangular shape, frequency approximation, value approximation, partitioning constraints or even more simply, just the queries, etc. to build the histograms.

On the other hand, self-managing databases gather and analysis useful statistical information to reveal the hidden characteristics behind the data, so that they could adjust themselves to automatically reacting to the changing context.

The authors found the famous histograms seldom take statistical characteristics of the underlying data into account when building up their own histograms. And they tried to combine both the above two sides of efforts to produce more accurate multidimensional histograms with less cost.

As the authors showed in the paper, COCA-Hist, as the

first to use such statistical information as correlation coefficients and value densities to describe multidimensional data distributions, exhibits superior quality and performance to those traditional multidimensional histograms. It is a data-driven method, and computing the statistical information can be done on the fly when the multidimensional data distribution matrix is being built thus inducing negligible burden.

The research is partially supported by the National Natural Science Foundation of China under grant No. 60473069 and No. 60496325. Also it is partially supported by international cooperation program with HP Labs in the project of Large Scale Data Management. All the granted projects aim to explore and find novel, effective and efficient ways to manage large scale of data under both the rapid improvements of hardware and the more and more challenging demands of database applications.

Their team has made much progress and achievements in the research direction. They set up a platform to test the quality of different multidimensional histograms. Also they have published some papers to share the ideas of using correlation coefficients to accurately and easily describe the different degrees of correlations.

The research in this paper is the important part to improve the quality of relation query optimizers which, in turn, improve the performance of the whole database systems.