编 者 按

数据库技术正以它特有的方式,不断地渗透和影响着社会,改变着人们处理信息和数据的行为。本期专题从数据库自调优、Web 服务、数据挖掘和地图数据库等几个不同的角度反映数据库技术的近期发展。

人民大学王珊教授等研究了数据库自管理、自调优中查询计划的自动优化问题。提出了熵相关系数的关联性自动判别新方法,该方法适应性更强、更高效、更简单。

复旦大学周傲英教授等介绍了采用“模糊 + 服务特征词 + Web 服务”的思路,给出了向量空间中进行 Web 服务发现的模糊方法。

武汉大学彭智勇教授等提出了一种基于对象代理模型的实现多表现 GIS 的新方法。其中特色技术包括对象更新迁移、跨类查询、扩展查询的等,他们实现了一个基于对象代理模型的多表现 GIS 原型,性能测试表明有很好的前景。

四川大学唐常杰教授等介绍了他们近期在社会网络分析方面的三项探索,包括虚拟社团的结构挖掘,基于六度分割和最短路径社团核心成员挖掘和基于用户属性的通信行为挖掘。

这些研究表明,“更多、更快、更好和更省”,是 DBMS 研发者的永恒目标。

文章编号:1001 - 9081(2006)09 - 2005 - 04

基于熵相关系数的关联性自动判别方法——COCA

王 珊,曹 巍,覃雄派
(中国人民大学 信息学院, 北京 100872)
(swang@ ruc. edu. cn)

摘 要:数据库自管理、自调优中查询计划的自动优化是目前的关注热点。为保证优化器估值精度,用统计学方法,给出了一种基于熵相关系数的对字段关联性的自动判别的新算法——COCA。该算法有下列特点:(1)限制少,没有卡方检验的频数限制,卡方检验只有在列联表中至少有 80% 的格子频数大于 5 的情况下才可信;(2)结果多,卡方检验(CORDS)只判断字段之间是否有关联,新方法可计算字段之间双向的关联程度。实验表明,新方法更坚固,产生更多的统计信息,可以支持后面更高效、准确地建立直方图。

关键词:查询优化; 统计信息; 关联性; 熵相关系数
中图分类号: TP311. 132 **文献标识码**:A

COCA — a new way to auto-detect association based on entropy correlated coefficients

WANG Shan, CAO Wei, QIN Xiong-pai
(*School of Information, Renmin University of China, Beijing* 100872, *China*)

Abstract: Self-managing and self-optimizing is currently a hot research field in database. To guarantee the accuracy of the estimates made by optimizer, this paper proposed a new method named COCA (entropy-COrrelated-Coefficient-based Auto-detection of association). In comparison with CORDS, COCA has the following features: (1) Fewer limitations. It overcomes the limitation that Chi-square test needs at least 80% of the cells in the contingency table have frequencies greater than 5. (2) More results. CORDS can tell the correlation between columns, while COCA can further discern the specific association degree for both directions. Experiments show that COCA is more robust and produces more statistical information, which is supportive to the creation of more effective and efficient histograms.

Key words: query optimization; statistical information; correlation; entropy correlated coefficient

0 引言

数据库自管理、自调优的研究领域中,改进查询优化是关注热点。它依赖数据库中统计信息的准确性,技术关键是:(a)判断在字段间关联性;(b)在有较强关联性的字段上创建直方图等形式的统计信息,减少因采用了不正确的数据独立性假设带来的估计值的错误。

判断字段之间是否具有较强的关联性是一个关键问题。本文提出了一种基于熵相关系数的关联性自动判别方法——COCA。这种方法的好处是利用了熵相关系数的特点,可以用

收稿日期:2006 - 07 - 14; **修订日期**:2006 - 07 - 27 **基金项目**:国家自然科学基金资助项目(60473069; 60496325)
作者简介:王珊(1944 -),江苏无锡人,教授,博士生导师,主要研究方向:高性能数据库、数据仓库、数据工程、知识工程; 曹巍(1975 -),辽宁沈阳人,博士研究生,主要研究方向:高性能数据库; 覃雄派(1971 -),广西百色人,博士研究生,主要研究方向:高性能数据库.

于数据较稀疏的情况,能提供双向的数据依赖关系的信息,而且实现简单直接,可为下一步创建二维直方图提供有益的指导信息。

1 相关工作

多个字段之间的关联性判别问题,一个显著特点就是很大程度上借助了统计学特别是多元统计分析中对列联表的关联性分析的思想和方法。比如 CORDS[1] 就是利用了二维列联表中常用的卡方检验的方法,而 DB-Histograms[2] 采用的是相对较新的对数线形模型的方法。CORDS 的多字段关联性判断方法的核心是“成对判断”的基于列联表的算法,而 DB-Histograms则基于对数线形模型,直接在对数线形模型的模型空间中搜索能描述多个字段间相关情况的某一模型。

CORDS 利用统计学中卡方检验的方法自动地发现数据之间存在的关联性或者非严格的函数依赖关系。但是从经验上一般认为,为了使卡方检验的结果准确可靠,列联表应该有至少 80% 的格子频数大于 5[1,4]。这样在数据比较稀疏的情况下卡方检验就不适用于检验关联性。但是在现实数据中,常常会出现比较稀疏的情况。为解决此问题,CORDS 采用了数据分块方法,并区分数据是否倾斜。

CORDS 方法的另一个不足是,卡方检验是对称的,只能判断两个字段 A 和 B 之间是否有关联性,但是在存在关联性的情况下,无法进一步判断字段 A 对字段 B 的依赖程度以及字段 B 对字段 A 的依赖程度。

2 COCA

针对 CORDS 方法的核心——卡方检验在现实数据中的问题,我们提出了一种基于熵相关系数的关联性自动检验方法 COCA。该方法与 CORDS 方法的相同点在于它们均成对地检验数据库中的字段,而且都是基于采样的方法。文献[1]的实验表明,样本容量为几千时,不管数据库的尺寸有多大,基于卡方的关联性检验可以得到比较令人满意的结果。例如 1000 ~ 2000元组的样本容量得到的 CORDS 准确度与更大的样本容量很接近。在实际应用 COCA 时,我们也采用了随机采样的方法,而且通过实验验证了不同的样本大小对 COCA 的准确度和性能的影响。

2.1 熵相关系数

根据文献[3],列联表的独立性检验可以采用熵相关系数来进行,熵相关系数的公式为:

$$\begin{cases} r_{j\to i} = \dfrac{-\sum_{i=1}^{d1}\sum_{j=1}^{d2} n_{ij} Ln(n n_{ij}/n_{i.}\, n_{.j})}{\sum_{i=1}^{d1} n_{i.} Ln(n_{i.}/n)} \\ r_{i\to j} = \dfrac{-\sum_{i=1}^{d1}\sum_{j=1}^{d2} n_{ij} Ln(n n_{ij}/n_{i.}\, n_{.j})}{\sum_{j=1}^{d2} n_{.j} Ln(n_{.j}/n)} \end{cases}$$

其中,列联表中的两个因素分别对应数据库中的两个字段 A 和 B,行方向上的因素对应字段 A,列方向上的因素对应字段 B。列联表中因素的多个水平对应数据库中字段 A 和字段 B 的取值。字段 A 中不同值的个数为 $d1$,字段 B 中不同值的个数为 $d2$(另外,A 和 B 两个字段不同的组合值个数用 $|AB|$ 表示);n_{ij} 代表列联表中对应的格子频数,$n_{i.}$ 和 $n_{.j}$ 代表对应的边缘频数。两个相关系数 $r_{j\to i}$ 和 $r_{i\to j}$ 分别表示了字段 A 依赖于字段 B 和字段 B 依赖于字段 A 的双向的关联程度。

2.2 熵相关系数的原理

这是一种“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[3]。熵相关系数不同于刻画线性相关关系的皮尔逊相关系数,它不仅限于线性相关关系,更确切地说,这种相关系数描述的是更广泛的关联性,是字段 A 的取值依赖于字段 B 的取值的程度(反之亦然)。因此熵相关系数的取值范围是[0,1],而皮尔逊相关系数的取值范围是[-1,1],因为后者区分了两个变量之间线性正相关和线性负相关的不同情况。熵相关系数也可以这样理解:它表示当一个字段取定一个值时,另一个字段的取值以多大概率被确定。当熵相关系数为 0 时,代表这两个字段是相互独立的;当熵相关系数为 1 时,代表这两个字段满足函数依赖关系。

2.3 COCA 方法与 CORDS 方法的比较

a) 双向依赖关系的度量

COCA 采用的是基于熵相关系数的关联性判别方法,它优于卡方检验的一个显著之处在于它可以提供双向的依赖关系度量信息,比卡方检验提供了更多信息,这些信息在后面创建二维直方图时具有较重要的指导意义。

b) 实现简单

卡方检验在实际使用中为了保证正确可靠会有一些限制,如要求至少 80% 以上的格子频数大于 5,CORDS 因此对于不符合这样要求的稀疏数据采取了分块统计的方法,而且还考虑了数据倾斜的情况,这样会使算法的复杂度增加。另一方面,根据统计检验的经验,卡方检验需要查找事先计算好的卡方分布表,但是现实数据会使卡方分布的自由度非常大,即便在自由度的增大使卡方分布越来越趋于对称,或者只需要粗糙的卡方分析时,卡方分布表的表示和存储仍然是一个问题。文献[1]并未给出 CORDS 算法中卡方检验的具体实现。我们提出的 COCA 方法没有对数据稀疏或者稠密的要求,不限制频数的取值范围,另外 COCA 无需事先存放任何标准的数据分布表,这些都是由熵相关系数这一便捷的样本统计量决定。

c) 指标的同一性

CORDS 中关联性的判断和非严格的函数依赖分成两个独立的步骤完成,依据的是不同的指标,关联性的判断依照的是均方列联的样本统计值;非严格的函数依赖依照样本中字段不同值的个数与不同的联合值个数的比值,比如字段 A 非严格函数决定字段 B(字段 B 非严格函数依赖于 A),则 $d1/|AB|$ 的比值接近于 1。

在 COCA 中独立性、相关性、非严格的函数依赖以及函数依赖均由熵相关系数这一个统计量体现,如果 k 个字段对的熵相关系数均为 1 或者接近于 1 的纯小数,则 COCA 推荐的结果会偏重于函数依赖的情况。因此,COCA 中为解决这一问题,采用如下方法:如果字段 A 严格或者非严格函数决定字段 B,而字段 B 对字段 A 的关联程度相对较低,则参考字段 B 的不同值个数 $d2$,即如果 $d2 < d'$(考虑存储空间的因素,d' 为一个阈值)则字段对 A 和 B 的二维直方图可以转化为 $d2$ 个一维直方图,每一个一维直方图对应一个字段 B 的取值,这样 COCA 就不会建议在二维分布比较偏斜的情况下为字段对 A 和 B 创建二维直方图了。

当然我们在判断非严格的函数依赖时也可以采用类似于 CORDS 的方法,用 $d1/|AB|$(或者 $d2/|AB|$)是否接近于 1 来判断,这样就可以避免熵相关系数的复杂计算,但是这样将非严格函数依赖与关联性严格区分开来,无法用同一的指标

来衡量。实验显示,虽然对数操作的实现较复杂,但在1000 ~ 3000 个元组的样本大小的情况下,COCA 计算熵相关系数的时间可以控制在 16ms 之内。

3　算法实现

3.1　候选字段集 ICS 的生成算法描述

在候选集的生成方面,COCA 与 CORDS 的最大不同在于 COCA 是面向工作负载的。COCA 分析构成工作负载的每一个 SQL 语句,对于查询涉及的每个表 T_i,条件子句中出现的在 T_i 的字段 $C^i_1, C^i_2, \cdots, C^i_k$ 上的谓词若以合取形式 $p^i_1 \wedge p^i_2 \wedge \cdots \wedge p^i_k$ 出现,则 $C^i_1, C^i_2, \cdots, C^i_k$ 放入 *ICS*(Interesting Columns Set)中;对于出现主/外键之间的连接 $T_i.C^i_P = T_j.C^j_F$,若 *ICS* 中没有表 T_j 的外键 C^j_F,则将其加入到 *ICS* 中。

在生成 *ICS* 的过程中我们采用了如下裁减规则:

1)参照系统字典中的表和字段的统计信息,一些小表(元组数 $n < n_{threshold}$)无需创建二维直方图,不同值个数太少的字段($d < d_{threshold}$)和不同值个数太多的字段($|n-d|/n < \varepsilon$)也无需参与二维直方图的创建。

2)数据类型的限制与 CORDS 的限制相同,除了整数、长度在 100 以内的字符串类型[1],以及其他具有分类特征的字段外,其他类型字段暂时不能参加 COCA 方法的关联性自动判别。

在工作负载中的 SQL 语句分析结束之后,生成候选字段的集合 *ICS* 为 COCA 采用熵相关系数进行自动的关联性判别之用。

3.2　候选字段集生成算法

算法:候选字段集生成 *ICSGen*

输入:工作负载,用 *W* 表示

输出:候选字段集 *ICS*

For each SQL statement in *W* do {
  Analyze against which tables it poses the query;
  For each such table T_i with $n >= n_{threshold}$ do {
    If it has conjunctive predicate $p^i_1 \wedge p^i_2 \wedge \ldots \wedge p^i_k$
      Add $C^i_1, C^i_2, \ldots, C^i_k$ into *ICS*, pruning out ineligible columns according to 1) and 2) in Section 3.(1);
    If T_i participates joins not by its primary keys
      Add all its participating columns into *ICS* with the same pruning rules applied; }}

3.3　COCA 关联性自动检测算法

算法:关联性自动检测

输入:候选字段集 *ICS*,空间限制 *SC*

输出:推荐需要建立直方图的字段对 *RCP*(Recommended Column Pair)

Sampling: For each table T_i showing up in *ICS* do
  Sampling T_i to get a sample S_i over T_i;
Calculating: In *ICS* for each pair of columns (denoted as C^i_A and C^i_B) from the same table T_i
Calculate correlated coefficients CC^i_{AB} and CC^i_{BA} from sample S_i;
RCP generating: By $CC^x_{\{AB\}} = \max\{CC^x_{AB}, CC^x_{BA}\}$, sort all the $CC_{\{,\}}$'s over all tables and all column pairs in descending order;
For each element $CC^x_{\{AB\}}$ do
If $((1 - CC^x_{\{AB\}}) < \varepsilon$ AND $|CC^x_{AB} - CC^x_{BA}| > \Delta_{threshold}$ AND the weakly determining column has number of distinct values $d2 < d'$) then
  Add the column pair into RCP as candidates for $d2$ one-dimensional histograms attaching to each one of the $d2$ distinct values;
Else
  Add the column pair into RCP (column pair set) in companion with correlated coefficient values to be suggested as candidates for two-dimensional histograms;
End If
Until space constraint SC is reached.

4　实验结果

4.1　熵相关系数的有效性

我们目前做了三组实验,第一组实验显示,熵相关系数的计算没有卡方检验的限制条件,特别是在数据稀疏的情况下,或者大多数格子频数比较小的情况下,熵相关系数的检验方法比 CORDS 采用了"数据分块"方法处理过的卡方检验更有效,这组实验是在较小的数据量上比较 COCA 和 CORDS 两种方法的有效性。实验结果总结如表 1。

表 1　COCA 与 CORDS 的有效性比较

	*d*1	*d*2	|*AB*|	稀疏否 (|*AB*|/(*d*1**d*2)<15%)
数据集 1[4]	2	2	4	否
数据集 2[4]	4	3	12	否
数据集 3	9	5	10	是
数据集 4	20	9	167	否
数据集 4'	20	5	21	是
数据集 5	30	109	120	是

	n	χ^2 自由度	χ^2 统计量(卡方)	CCF1 ($A \rightarrow B$)	CCF2 ($B \rightarrow A$)
数据集 1[4]	6200	1	4.82	0.00095	0.00056
数据集 2[4]	580	6	19.63	0.025	0.013
数据集 3	25	32	94.79	0.96	0.74
数据集 4	1882	152	664.6	0.087	0.063
数据集 4'	1882	76	5422.34	0.96	0.44
数据集 5	137	3132	3748.00	0.67	0.96

在这 5 个数据集中,数据集 1 和数据集 2 直接从列联表[4]开始计算 *V* 和 CCF 等统计量,数据集 3 和数据集 5 是来自现实世界中的数据,数据集 4 和数据集 4′是同一生成的数据集中不同的列对。对数据比较稀疏的情况若使用 CORDS 的数据分块的方法,得到的结果如表 2 所示。

表 2　CORDS 进行数据分块后与 COCA 比较

	*d*1'	*d*2'	|*AB*|	稀疏否 (|*AB*|/(*d*1**d*2)<15%)
数据集 3	3	2	5	否
数据集 4'	5	5	10	否
数据集 5	6	8	21	否

	n	χ^2 自由度	χ^2 统计量(卡方)	CCF1 ($A \rightarrow B$)	CCF2 ($B \rightarrow A$)
数据集 3	25	2	10.58	0.39	0.29
数据集 4'	1882	16	3803.62	0.96	0.44
数据集 5	137	35	287.26	0.45	0.58

此处 *d*1′ 和 *d*2′ 分别代表两个字段的值域划分成的区间数。

实验评价:从卡方计算结果可以看出,单纯从 χ^2 统计量的取值很难看出两个字段是否独立,必须配合查找 χ^2 分布表才可以实现统计检验;现实世界中的数据会使卡方分布的自由度 $(d1-1)*(d2-1)$ 变化很大,这使直接用卡方检验难于实现;CORDS 的数据分块处理会使数据稀疏的情况得到了

改善,并且将卡方分布的自由度减小,使 χ^2 检验成为可行,但是在数据分块的过程中会丢失一些可能比较重要的信息,比如数据集 3 和数据集 5 中,原来由熵相关系数所示的比较显著的数据相关关系在分块处理后,变得不显著了。

4.2 大数据量情况下样本大小的影响

第二组实验检验在大数据量的情况下 COCA 和 CORDS 的有效性,以及不同的样本大小对相关性判别的影响。在这里我们的实验采用类似 CORDS 的方法,考察了完全独立、具有一定的相关性,以及具有函数依赖的不同数据分布情况,事先生成具有一定相关性的数据,分别在数据和样本上运行 COCA,分析比较不同的结果。通过选择不同数量级的 $d1$ 和 $d2$, 我们事先生成具有一定相关性的数据,在数据上直接计算熵相关系数,实验结果总结如表 3。

表 3 大数据量并且数据稀疏情况下 COCA 的实验数据

	$d1$	$d2$	$\|AB\|$	稀疏否 $(\|AB\|/(d1*d2)<15\%)$
数据集 I	49475	97420	538818	是
数据集 II	4997	20000	154656	是
数据集 III	475	200	1800	是

	n	CCF1($A\dashrightarrow B$)	CCF2($B\dashrightarrow A$)	运行时间/ms
数据集 I	10^6	0.76	0.83	111984
数据集 II	10^6	0.63	0.74	7141
数据集 III	10^6	0.72	0.61	4469

这个实验结果说明 COCA 即使在数据量非常大并且数据分布非常稀疏(例如数据集 I)的情况下,也能在可接受的时间内准确地计算出两个相关系数, 对字段 A 和字段 B 的关联性给出正确判断。而在这样的情况下,CORDS 是根本无法直接进行卡方检验的,只能依赖数据分块才可以进行计算,但是如前所述数据分块却可能带来信息的丢失,导致计算的结果与实际情况有所偏差。

我们还分别考察了对数据集 I、数据集 II 和数据集 III 采用随机抽样计算熵相关系数的准确性,因为实验结果大体相近,现将对数据集 III 进行不同的采样结果总结如表 4。

表 4 关联性适中,不同样本大小对结果的影响

样本大小	$d1^s$	$d2^s$	$\|AB\|^s$	稀疏否
1000	298	164	1800	是
3000	399	193	1053	是
5000	437	198	1249	是
10000	462	200	1491	是

样本大小	CCF1($A\dashrightarrow B$)	CCF2($B\dashrightarrow A$)	运行时间/ms
1000	0.81	0.71	15
3000	0.76	0.66	16
5000	0.75	0.64	16
10000	0.74	0.63	31

在实验中我们发现,在数据关联性适中的情况下,小样本会使熵相关系数增大,而显然函数依赖的判别不会受样本大小的影响。我们事先生成具有函数依赖关联的数据,并随机抽取不同大小的样本,结果表 5 所示。

这一组实验充分说明,当样本达到 3 000 ~ 5 000 的规模时,计算出的熵相关系数的误差比较小,可以通过控制样本大小的方法使误差控制在可接受的范围内。同样地,我们也生成了相互独立的数据,实验发现采样会影响数据独立性的判断,但是我们仍然可以通过控制随机样本容量,将关联性检验的 false positive 比率降低。

表 5 函数依赖的情况下,不同样本大小的实验数据

大小	$d1$	$d2$	$\|AB\|$	稀疏否
10^6(原始数据)	500	2000	2000	是
1000(样本)	426	762	762	是
3000(样本)	497	1529	1529	是
5000(样本)	500	1803	1803	是
10000(样本)	500	1975	1975	是

大小	CCF1($A\dashrightarrow B$)	CCF2($B\dashrightarrow A$)	运行时间/ms
10^6(原始数据)	0.82	1	4641
1000(样本)	0.90	1	16
3000(样本)	0.85	1	16
5000(样本)	0.84	1	31
10000(样本)	0.83	1	47

4.3 COCA 的性能

第三组实验是关于熵相关系数计算性能的实验,在大数据量的情况下,运行时间值参见表 3 ~ 表 5。我们在一台 Genuine Intel Core Duo CPU T2300, 1.66GHz 的 PC 机上运行整个实验,内存容量为 512MB,操作系统为 Windows XP,编译器为 GCC; 实验结果显示的时间中包括数据分布的列联表生成和相关系数的计算等步骤,实验表明,在 1 000 ~ 3 000 个元组的样本大小情况下,COCA 为一对字段检验关联性所花的平均时间为仅为 16ms。说明 COCA 采用的基于熵相关系数的关联性检验方法比 CORDS 采用的基于卡方检验的方法具有更好的统计特性并且更可行,虽然熵相关系数的计算基于对数函数,要复杂一些,但是在现代计算机的强大的 CPU 能力支持下,其运行的性能还是可接受的。

5 结语

卡方检验和基于熵相关系数的检验是检验字段间关联性的两种不同的方法,我们提出 COCA 是为解决卡方检验方法在适用条件和在计算机上实现的限制。COCA 基于的熵相关系数是关联性的直接度量,而卡方检验需要对照事先存储的卡方分布表。在现实数据分布中自由度变化很大的情况下,卡方检验在具体实现时要采取应对的措施,熵相关系数取值范围在 0 和 1 之间,具有作为同一的度量关联性尺度的天然优良特性,而且还提供字段之间双向关联程度的度量,对创建更准确高效的直方图很有价值。

COCA 采用了分析工作负载生成候选字段集的方法,按照文献[1]的分类方法,是属于查询—数据混合驱动的方法,CORDS 则面向全部数据生成候选字段对的集合,属于数据驱动的方法。

研究工作中我们发现,熵相关系数虽然能够胜任在数据稀疏的情况下关联性判别,在数据稀疏的情况下创建高效的直方图是一个值得更深入研究的课题,而前期获得的熵相关系数具有重要的参考意义。未来要考虑 COCA 支持更多种数据类型,并进一步实验分析不同的数据库模式及其工作负载对 COCA 的有效性和性能的影响。

参考文献:

[1] ILYAS IF, MARKL V, HAAS PJ, *et al*. CORDS: Automatic Discovery of Correlations and Soft Functional Dependencies[A]. Proceedings 2004 ACM SIGMOD, 2004.

[2] DESHPANDE A, GAROFALAKIS M. Independence is Good: Dependency-Based Histogram Synopses for High-Dimensional Data [A]. Proceedings of ACM SIGMOD, 2001.

[3] 张尧庭. 定性资料的统计分析[M]. 桂林: 广西师范大学出版社, 1991. 22, 198 - 203.

[4] 周兆麟, 李毓芝. 数理统计学[M]. 北京: 中国统计出版社, 1987.