Estimation of Database Unique Values

Carlo DELL'AQUILA, Ezio LEFONS, Filippo TANGORRA
Dipartimento di Informatica
Università di Bari
Via E. Orabona 4, I-70125 Bari
ITALY

Abstract: - Counts of database unique values are crucial information in query optimization. Estimating the number of the distinct values occurs frequently in database queries, due to its importance in selecting query plans. We present a nonparametric method for estimating the database distincts, and, then, the number of distinct values. The method computes few parameters which describe the distribution of distances of distinct values in the attribute value ranges. Tests have been carried out that also show the useful applicability of the method to estimate equi-join selectivity factors.

Key-Words: - Data Bases, estimation, query optimization, join selectivity.

1 Introduction

Efficient processing of complex queries over large volumes of data has an increased importance with the growing interest in datawarehousing and decision support systems.

The quantitative properties which summarize characteristics of database instances are used by query optimizers in order to determine the optimal strategy of query execution. It includes parameters such as tuple cardinality, number of distinct values of attributes, maximum, minimum and attribute distribution values. In particular, accuracy of distinct values estimation strongly impacts the query optimiser's ability to generate good plans for SQL queries.

Exact computation of the number of the distinct values is an expensive operation which requires at least one scan with sorting of the relation. Probabilistic counting methods provide the estimation of the number of distinct values and they require only to scan the relation avoiding sort or hash-based computation [1-3]. Sampling-based methods also estimate the number of the distinct values. They process only a fraction of the tuples in a relation [4-6].

In this paper, we present a nonparametric method for estimating the database distincts, and, consequently, their counts. The method computes parameters which describe the distribution of distances of distinct values in the attribute value ranges. Moreover, the method does not require *a priori* assumptions on uniformity and independence of attribute values.

In [7], we presented an analytical method for approximating the actual multivariate distribution of the attributes by means of a series of orthogonal polynomials. The method is based on the computation of a set of values, the so-called *Canonical Coefficients*, from which the main parameters of database statistics can be easily derived and efficiently computed. Paper [7] also contains an in-depth discussion of other nonparametric (both analytic and non-analytic, histograms, etc) methods in comparison with our method. Here, we extend the method to approximate attribute distinct values. Then, an application to the problem of estimating the join relational selectivity is considered. In fact, accurate estimation of join selectivity is more difficult than selection selectivity, because further information is required on matching join values and, therefore, on attribute distinct values. Different approaches have been put forward to estimate the join selectivity [8-16] but none determines matching values of join attributes using statistical profile, without accessing database instances.

The most widely investigated join operators are natural join and equijoin and we will refer to these operations using only the term "join". Our approach estimates the resulting size of the relational join by evaluating distinct values and using the estimation of the actual distribution of attributes.

2 The analytic method to approximate the data distribution

Let R be a relation of cardinality N and let X be an attribute of R. Suppose *dom*(X) = [a,b] and let x_1, x_2, ..., x_N be the occurrences of X in R.

We approximate the probability density function g(x) of the attribute X with

$$g(\mathrm{x}) = \frac{1}{\mathrm{b\text{ - }a}} \sum_{i=0}^{n} (2i+1)\, c_i\, P_i(\mathrm{x}) \qquad (1)$$

for all x∈X and for opportune n, for each i = 0,1 ,.., n, P_i(x) is the Legendre orthogonal polynomial of degree i, and coefficient c_i is the mean value of P_i(x) on the instances of X. That is,

$$c_i = \frac{1}{N} \sum_{\mathrm{x} \in \mathrm{X}} P_i(\mathrm{x})\,. \qquad (2)$$

c_0, c_1, …, c_n are computed with simple recursive formulae and they are called the *Canonical Coefficients* of X.

Legendre polynomials are defined, hence they are suitably computable, by the recurrence formula:

$$P_{i+1}(\mathrm{x}) = \alpha_i\ \mathrm{x}\ P_i(x) + (1-\alpha_i)\, P_{i-1}(\mathrm{x}), \quad \text{for } i \geq 1, \ (3)$$

where $\alpha_i = \frac{2i+1}{i+1}$ for all i, and P_0(x) = 1 and P_1(x) = x.

The approximation of the cumulative distribution function G(x) of g(x) is

$$G(\mathrm{x}) = \int_{\mathrm{a}}^{\mathrm{x}} g(y)\mathrm{d}y = \frac{1}{\mathrm{b}-\mathrm{a}} \sum_{i=0}^{n} c_i \left(P_{i+1}(\mathrm{x}) - P_{i-1}(\mathrm{x})\right) \ (4)$$

Let I = [x_1,x_2] ⊆ [a,b] be a generic query-range of X. We denote with *count*(x;I) or N × *percent*(x;I) the number of tuples of R whose x values belong to interval I. *count*(x;I) can be approximated by N ×· (G(x_2)-G(x_1)).

We have called coefficients c_0, c_1 ,.., c_n the *Canonical Coefficients* (CC) of the attribute X because they contain the information needed to represent the distribution of the attribute X. Based on formulas (2) and (3), the calculation of the canonical coefficients of X only requires a (sequential) scanning of the attribute X.

3 Distinct value estimation

The method for estimating the distinct values of attribute X does not make any hypotheses on the spacing distribution of distinct values or their join equivalence. Distinct values are approximated using canonical coefficients themselves and, on the basis of the additive property, they can be easily updated [7].

Supposing that the distinct values for X are X={x_1, x_2, ..., x_d}, the canonical coefficients $(d_i)_{0 \leq i \leq n}$ up to degree n which contain information on how distinct values are spaced are computed in the following way:

$$d_i = \frac{1}{d \times dns} \sum_{j=1}^{d} \sum_{k=1}^{dns} P_i(y_{jk}) \qquad i=0,1,...,n \qquad (5)$$

where y_{jk} are dns random values in the interval I_j =]x_j-δ/2, x_j+δ/2[, with δ = (b-a)/N.

We assume that $\bar{x}_j$ is an approximation of a distinct value, say, x_j, if it verifies, in the interval $\bar{I}_j$ = [$\bar{x}_j$ -δ/2, $\bar{x}_j$+δ/2], the condition $count(x;\bar{I}_j) \cong dns$ or, equivalently, if it veriefies for $\varepsilon > 0$ that

$$\left| count(x;\bar{I}_j) \text{ - } dns \right| \leq \varepsilon\,. \qquad (6)$$

The set of approximated distinct values is

$$\bar{X} = \{\, \bar{x}_j = \mathrm{a}+(i\text{-}1)\delta \mid \left| count(x;\bar{I}_j) \text{ - } dns \right| \leq \varepsilon \ \text{ and } \ i = 2,..., N\} \qquad (7)$$

$\bar{X}$ contains distinct values estimated using an approximation degree n a density dns and an approximation error

$$\varepsilon = k \times dn\mathrm{s} \qquad (0<k<1) \qquad (8)$$

Condition (6), which allows us to consider $\bar{x}_j$ as an approximation of a distinct value, is called *distinct condition* and ε is the approximation error of the cumulative density function.

The canonical coefficients $(d_i)_{0<i\leq n}$ are used for calculating the *count* function in (6) and are different from those which approximate the distribution of attribute X.

4 Join selectivity estimation

In this section, we provide an approach to estimate the cardinality of the join of two relations, using the analytical model described in the previous section.

Let T be the result of the join of relations R and S over the respective attributes X and Y. The canonical formula giving cardinality of T is

$$card(\mathrm{T}) = j\rho \times card(\mathrm{R}) \times card(\mathrm{S}), \qquad (9)$$

where $j\rho$ is the join selectivity factor and represents the fraction of tuples of the Cartesian product of R and S, for which attribute values of X and Y are equivalent. We want to estimate *card*(T) or equivalently we want to provide an estimation for $j\rho$.

Let $X=\{x_1, x_2, ..., x_{d_X}\}$and $Y=\{y_1, y_2, ..., y_{d_Y}\}$ ordered sets of the distinct values of attributes X and Y, we have that

$$card(T) = \sum_{x \in X \cap Y} count(x;\mathrm{R}) \times count(x;\mathrm{S}) \,, \qquad (10)$$

where *count(x;*R*)* and *count(x;*S*)* are respectively the number of tuples *r* of R and *s* of S such that *r*[X] = *x* =*s*[Y]. It can be easily shown that

$$card(T) = \sum_{x \in X}\sum_{y \in Y} \left[count(x;\mathrm{R}) \times count(y;\mathrm{S})\right]_{x=y} =$$

$$\sum_{i=1}^{d_X}\sum_{j=1}^{d_Y} \left[count(x_i;\mathrm{R}) \times count(y_j;\mathrm{S})\right]_{x_i=y_j} \qquad (11)$$

The notation introduced in (11) indicates that factors $count(x_i;\mathrm{R}) \times count(y_j;\mathrm{S})$ are considered for summary only when the condition $x_i= y_j$ is verified. Supposing that *X* and *Y* are known, the cardinality of the join can be estimated using the following estimations for *count(x*$_i$;R*)* and *count(y*$_j$;S*)*:

$$count(x_i;\mathrm{R}) \cong count(x;\ \mathrm{I}_{x_i}) \quad \text{and}$$
$$count(y_j;S) \cong count(y;\mathrm{I}_{y_j}), \qquad (12)$$

where intervals I_{x_i} and I_{y_j} are defined as follows:

$$I_{x_i} = \left[\frac{x_{i-1}+x_i}{2}, \frac{x_i+x_{i+1}}{2}\right] \quad \text{and}$$
$$I_{y_j} = \left[\frac{y_{j-1}+y_j}{2}, \frac{y_j+y_{j+1}}{2}\right]. \qquad (13)$$

So the estimation for the selectivity factor can be obtained

$$j\rho \cong \sum_{i=1}^{d_X}\sum_{j=1}^{d_Y} \left[percent(x;\mathrm{I}_{x_i})percent(y;\mathrm{I}_{y_j})\right]_{x_i=y_j} .(14)$$

It can be observed that the estimation for *jρ* can be obtained only if distinct values for X and Y are known. However, these are often unknown and many researchers suppose that the differences between two adjacent domain values are approximately equal [13]. This assumption defines the semantics to be assigned to a join operation and holds when attribute values are integer or decimal, with a maximum of *r* decimal digits. In these cases the number of possible distinct values are respectively *d* = b-a+1 and *d* ≤ ((b-a)/10^{-r})+1. In [11] the number of distinct values *d* for join range [a,b], where a=*max*(a$_x$,a$_y$) and b=*min*(b$_x$,b$_y$), is estimated as follows:

$$d = \begin{cases} min\left(d_X \dfrac{\mathrm{b}-\mathrm{a}}{\mathrm{b}_X - \mathrm{a}_X}, d_Y \dfrac{\mathrm{b}-\mathrm{a}}{\mathrm{b}_Y - \mathrm{a}_Y}\right) & \text{if } \mathrm{a} < \mathrm{b} \\ 1 & \text{if } \mathrm{a} = \mathrm{b} \\ 0 & \text{otherwise} \end{cases}$$

Moreover, we outline an estimation for join range distinct values can be obtained using d_X and d_Y, and these can be obtained only by accessing attributes X and Y. The join range is then divided into a number of intervals with amplitude Δ=(b-a)/*d*, then the estimated join selectivity factor is

$$j\rho = \frac{1}{\Delta}\sum_{i=1}^{d}(F(x_i) - F(x_{i-1}))(G(x_i) - G(x_{i-1})) \qquad (15)$$

in the hypotheses that X and Y are uniformly distributed in each of the d intervals of amplitude Δ. In (15) F and G are cumulative distribution functions respectively of attribute X and Y, which are evaluated using equal-width histograms of X and Y.

We propose an approach for estimating distinct values and for applying formula (11), which is not applicable when sets of estimated distinct values $\overline{X}$ and $\overline{Y}$, are available instead of sets of actual distinct values *X* and *Y*. In these cases couples $(\overline{x}_i, \overline{y}_j) \in \overline{X} \times \overline{Y}$ rarely exists such that $\overline{x}_i = \overline{y}_j$, therefore we consider the following approximation:

$$j\overline{\rho} \cong \sum_{i=1}^{d_{\overline{X}}}\sum_{j=1}^{d_{\overline{Y}}} \left[percent(\overline{x}_i, I_{\overline{x}_i}) \times percent(\overline{y}_j, I_{\overline{y}_j})\right]_{\overline{x}_i \cong \overline{y}_j} \qquad (16)$$

Notation $\overline{x}_i \cong \overline{y}_j$ establishes that *percent* values are considered for summary only if $\overline{x}_i$ and $\overline{\mathrm{y}}_j$ satisfy both the following conditions

$$\left|\overline{x}_i - \overline{y}_j\right| \le min\left\{\left|\overline{x}_{i-1} - \overline{y}_j\right|, \left|\overline{x}_{i+1} - \overline{y}_j\right|\right\} \qquad (17)$$

and

$$\left|\, \overline{x}_i - \overline{y}_j \right| \le min\left(\frac{\mathrm{s}_{\overline{X}}}{2}, \frac{\mathrm{s}_{\overline{Y}}}{2}\right), \qquad (18)$$

where $s_{\overline{X}} = \frac{\sum_{i=2}^{d_{\overline{X}}}(\overline{x}_i - \overline{x}_{i-1})}{d_{\overline{X}} - 1}$ and $s_{\overline{Y}} = \frac{\sum_{i=2}^{d_{\overline{Y}}}(\overline{y}_i - \overline{y}_{i-1})}{d_{\overline{Y}} - 1}$

are the mean distance between approximated distinct values of X and Y respectively and intervals $I_{\overline{x}_i}$ and $I_{\overline{y}_j}$ are defined according to (13). Conditions (17) and (18) specify whether two estimated distinct values are to be or are not to be considered joining values and we refer to these conditions as *join equivalence* conditions.

5 Experimental results

We present experimental results of the analytic method to estimate the attribute distinct values, and its performance when applied to the estimation problem of the join selectivity factor.

We have performed experiments on a real database considering two relations R(A1,A2,A3) and S(B1, B2, B3, B4). The values of attributes for these relations are respectively integer (with equally-spaced distinct values) and real. Their features are reported in the table 1.

Attribute Name	Attribute Type	*Card*(X)	Distincts	a	b
A1	integer	20000	26	1	52
A2	integer	20000	16	0	31
A3	integer	20000	16	0	31
B1	real	26568	80	0.1	18.0
B2	real	28730	287	2.0	600.0
B3	real	28730	323	1.4	881.0
B4	real	28363	175	0.17	2.33

Table 1. - Features of attributes in relations R and *S*

We have considered only real databases because our method maintains any distribution form of data and of distinct values separatly. Null values of the attributes have not been considered.

5.1 Distinct Value Estimation

The aim of the experiment is to observe the accuracy of the estimations for the actual distinct values and for the number of distinct values. In fact, the number of distinct values is estimated not always equal to the number of effective distinct values.

The results obtained have been analyzed taking into consideration the percentage error ε and the number of random values *dns* distributed around each distinct value. For measuring errors in estimating the distinct value number *d*, we have adopted the metric:

$$M = \frac{\left|d - \overline{d}\right|}{d} \times 100 . \qquad (19)$$

In the first series of experiments we have considered $\overline{d} = card\left(\left\{x \in \overline{X} \mid x \in X\right\}\right)$, that is the number of distinct values which have been correctly estimated. In these experiments we denote the metric *M* with M_e.

In the second series of experiments $\overline{d} = card\left(\overline{X}\right)$ is the number of values which satisfy the distinct condition; in other words an estimation for *d*, *M* in these experiments has been denoted with M_d.

We have computed the average of M_d and M_e for approximation degree *n*=5,6,...33 and for the following values of k in (8) *k*=0.05, 0.1, 0.15, ..., 0.3.

We have observed that when *k* and, consequently, ε grow we obtain good estimations for actual distinct values. Pessimistic estimations for the number of distinct values have been obtained because the number of distinct values is always overestimated.

The average of M_d and M_e , independently of density *dns,* is generally constant and decreases with growing values of approximation degree *n.* So, we have used the low percentage error ε, density *dns*, and an approximation degree *n* = 27. Generally better estimations have been obtained for integer attributes than real ones; however for all the attribute, we have observed that when *k* grows, (see, formula (8)), M_e converges and M_d diverges.

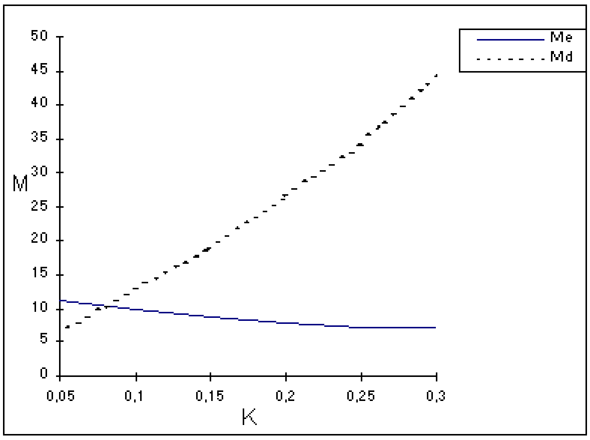

Fig. 1 – Mean percentage error for attribute B1

Figure 1 shows the averages of M_d and M_e for approximation degrees *n*=5, 6, …, 33. It points out that a value 0,05 is adequate for the k factor in (8).

5.2 Join Selectivity Estimation

In this section we illustrate and discuss the experimental results obtained for the join selectivity estimation. The purpose of the experiments is to compare analytic method performances by varying the approach used in estimating attribute distinct values and their cardinalities.

In estimating the join selectivity we have used three approaches. In the first approach (the DD method), no hypotheses have been made on the number of distinct values and their spacing in the join range [a,b]. In this case distinct values for the joining attributes X and Y have been estimated using the method discussed in Section 3. According to the experimental results, we have considered *dns*=50 and $\varepsilon = 0.05 \times dns$ while the join selectivity factor has been estimated using (16).

In the other approaches the number of distinct values for each attribute has been estimated using the method presented in [11] and the join range has been divided into equal-width intervals. In other words, attributes have been considered equally-spaced or values in each interval have been considered join equivalent.

In the second approach (the EW method), formula (15) has been used in estimating the join selectivity factor, where the functions F and G have been estimated using (4).

The last approach (the CC method) determines $j\rho' = \Delta^* j\rho$, where Δ is that defined in Section 2 and $j\rho$ is obtained as shown in (15).

The entries in Table 2 indicate the join attributes, the actual number of distinct values *d* of join range [a,b], the number of distinct values d_{ew} estimated using the method proposed in [11], the number of distinct values d_{cc} estimated using the method illustrated in Section 2, the mean percentage error is been measured using the metric $M = \frac{|j\rho - \overline{j\rho}|}{j\rho}$; where $j\rho$ and $\overline{j\rho}$ are respectively the join selectivity factor and its estimation relative to methods DD, EW and CC.

In estimating distinct values with method DD, the approximation degree $n = 27$ has been used. For each method, in estimating the cumulative distribution function by canonical coefficients, we have used approximation degrees 5 to 33. However, for the sake of readability, here we only report the mean percentage error measured for n_{cc} =8,13,27 using metric *M*.

X join Y	*d*	d_{ew}	d_{cc}	n_{cc}	*Mean Percentage Error of M*		
		(n=27)			**DD**	**EW**	**CC**
A1 A1	26	29	25	8	8.03	3.89	3.89
				13	8.02	3.85	3.85
				27	7.89	3.74	3.74
A2 A2	16	18	15	8	1.04	3.12	3.12
				13	0.81	2.80	2.8
				27	0.93	2.95	2.95
A2 A3	16	18	15	8	1.98	0.42	0.42
				13	1.70	0.04	0.04
				27	1.58	0.08	0.08
B1 B1	80	85	179	8	22.09	695.94	80.35
				13	10.15	821.94	108.90
				27	3.09	888.55	123.99
B2 B2	286	298	285	8	0.90	13.85	80.76
				13	8.23	4.62	100.14
				27	5.26	1.36	106.96
B2 B3	253	285	219	8	13.87	34.85	77.98
				13	8.97	14.82	132.69
				27	5.01	2.11	167.41
B4 B4	175	184	174	8	12.34	5586.7	12.50
				13	5.84	5586.7	21.91
				27	4.49	5586.7	23.59

Table 2. - Mean percentage errors for join selectivity estimation.

Methods EW and CC give better results than DD only for the integer attributes for which the distinct values are equally-spaced; however, for this type of attribute, method DD shows very slight errors which are in any case acceptable. On the other hand, the results of EW and CC are pessimistic when we consider the attributes whose distinct values are not equally-spaced. For these attributes, the mean percentage error is lower for method DD than the others, which improves reliability when the approximation degree of canonical coefficients grows.

Table 3 reports the average of M of the join estimation on relations R and S based on EW, CC and DD methods.

	EW	CC	DD
M	4028,55	1377,42	36,99

Table 3. - The average of the metric *M*.

6 Conclusion

The possibility to include parameters in the database profile, suitable to provide estimates of attribute distinct values, can improve greatly the performance of the profile database in estimating the selectivity factor of relational operations. Since the traditional assumption of equal-spaced distinct values, in the absence of other information, is not realistic for many of the actual databases, experimental results encourage further research in this field (on relational projection and inclusion of null values, for example). The performance of the statistical profile is highly unsatisfactory when this assumption does not hold true, as our experiments of join estimates have shown.

References

[1] K.Whang, B.T.Vander-Zanden and H.M.Taylor, A linear-time probabilistic counting algorithm for data base applications, *ACM Trans. Database Systems,* Vol. 15, No. 2, 1990, pp. 208-229.

[2] Astrahan M.M., Schkolnick M., and Whang K., Approximating the number of unique values of attribute without sorting, *Information Systems,* Vol. 12, No. (1), 1987, pp.11-15.

[3] P. Flajolet and G.N. Martin, Probabilistic counting algorithms for data base applications, *J. Computer Sys. Sci.* Vol. 31, 1985, pp. 182-209.

[4] P.B. Gibbons, Distinct sampling for higly-accurate answers to distinct values queries and event reports, *Proc. 27th Int. Conf. Very Large Data Bases*, Roma, Italy, 2001, pp. 541-550.

[5] M. Charikar, S. Chaudhuri, R. Motwanni and V. Narasayya, Towards estimation error guarantees for Distinct values, *Proc. of ACM PODS*, 2000, pp. 268-279.

[6] P.J. Haas, J.F. Naughton, S. Seshadri, and L. Stokes, Sampling-based estimation of the number of distinct values of an attribute, *Proc. 21th Int. Conf. Very Large Data Bases*, Zurich, Swizwrland, 1995, pp. 311-322.

[7] E. Lefons, A. Merico, and F. Tangorra, Analytical profile estimation in database systems, *Information Systems,* Vol. 20, No. 1, 1995, pp. 1-20.

[8] D.A. Bell, D.H.O. Ling, and S. McClean, Pragmatic estimation of join sizes and attribute correlations, *Proc. 5th IEEE Data Engineering Conf.*, Los Angeles, 1989, pp. 76-84.

[9] Y. E. Ioannidis and S. Christodoulakis, Optimal histograms for limiting worst-case error propagation in the size of the join results, *ACM Trans. Database Systems,* Vol 18 No. 4, 1993, pp. 709-748.

[10] M.V. Mannino, P. Chu, and T. Sager, Statistical profile estimation in database systems, *ACM Computing Surveys*, *20*(3), 1988, 191-221.

[11] T. Mostardi, Estimating the size of relational SPθJ operation: an analytical approach, *Information Systems,* Vol.15, No 5, 1990, pp. 591-601.

[12] J. K. Mullin, Estimating the size of a relational join, *Information Systems*, Vol. 18, No. 3, 1993, pp. 189-196.

[13] W. Sun, Y. Ling, N. Rishe, and Y. Deng, An instant and accurate size estimation method for joins and selection in retrieval-intensive environment, *Proc. ACM SIGMOD Int. Conf. Management of Data*, Washington, DC, 1993, pp. 79-88.

[14] C.M. Chen and N. Roussopoulos, Adaptive selectivity estimation using query feedback, *Proc. ACM SIGMOD Int. Conf. Management of Data*, Minneapolis, Minnesota, 1994, pp. 161-172.

[15] Y E. Ioannidis, V. Poosala, Balancing histogram optimality and practicality for query result size estimation, *Proc. ACM SIGMOD Int. Conf. Management of Data*, San Jose, CA, 1995, pp. 233-244.

[16] P. J. Haas, J.F.Naughton, S. Seshadri, A.N. Swami, Selectivity and cost estimation for joins based on random sampling, *J. Computer Sys. Sci.* *52*(3), 1996, pp. 550-569.