Evaluating the Size of Queries on Relational Databases with non Uniform Distribution and Stochastic Dependence

Silvio Salza, Mario Terranova
Istituto di Analisi dei Sistemi ed Informatica del CNR
Viale Manzoni, 30
I-00185 Roma, Italy

Abstract

The paper deals with the problem of evaluating how the originality of the attributes of a relation, i.e. the number of distinct values in each attribute, is affected by relational operations that reduce the cardinality of the relation. This is indeed an interesting problem in research areas such as database design and query optimization. Some authors have shown that non uniform distributions and stochastic dependence significantly affect the originality of the attributes. Therefore the models that have been proposed in the literature, based on uniformity and independence assumptions, in several situation can not be conveniently utilized. In this paper we propose a probabilistic model that overcomes the need of the uniformity and independence assumptions. The model is exact for non uniform distributions when the attributes are independent, and gives approximate results when stochastic dependence is considered. In the latter case the analytical results have been compared with a simulation, and proved to be quite accurate.

1. Introduction

Evaluating the size of the result of a relational operation is of great interest in computing the transaction execution cost for query optimization and database design, especially in a distributed environment [2]. Such estimates are usually computed from a statistical characterization of the operand relations, consisting typically in a set of parameters, such as the cardinality of the relation, and the originality (i.e. the number of distinct values) of each attribute [5,10,11].

These parameters can be usually maintained for the base relations, but must be estimated for the intermediate relations, when queries consisting in a sequence of relational operations are considered. As a typical example consider computing the execution cost of a join or semijoin. The cardinality of the result, which is proportional to the cost, depends on the originality of the join attribute in the operand relations, that are in turn usually the result of a select or semijoin [1,2,6]. Another example is the computation of the size of the result of a projection with duplicate elimination [12].

In this paper we are concerned with estimating the decrease of the originality in a selection. This extends to any relational operation (e.g. semijoin , restriction, etc.) or sequence of relational operations that reduces the cardinality of the attributes. Actually we focus on the attributes which are not involved in the select (or join) condition. The remaining attributes require a different kind of analysis, strictly based on the structure of the predicates [3,9].

We present a probabilistic model which uses a simple set of parameters to characterize the distribution of the attribute values. We think that in most cases this approach is more effective than attempting a detailed description that would be considerably more expensive and difficult to maintain. Actually a detailed information is not available in many situations, such as the optimization of compiled queries, where the literals in the predicates are not known at compilation time.

Several authors have adopted a similar approach for this or strictly related problems, but, to keep the model tractable, have made the additional assumption that the attribute values are uniformly and independently distributed [14,1,12]. These assumptions are very seldom verified in actual databases, and have been criticized because they often lead to pessimistic cost estimates [3,4].

In this paper we overcome the need for the uniformity and independence assumptions, extending the results of an earlier paper [12] to take into account the effect of non uniform distributions and statistical dependence between the attribute values. We propose a model that is exact for non uniform distributions when the attributes are independent, and gives approximate results when stochastic dependence is considered. In the latter case the analytical results have been compared with a simulation, and proved to be quite accurate. All the results confirm that both the non uniformity and dependence of the distribution of the attribute values may significantly affect the originalities of the result relation.

Notation and definitions are given in Section 2. The basic model for non uniform distributions is introduced in Section 3, where some numerical results are also given. In Section 4 the model is extended to cover the dependence between attributes and validated against a simulation.

2. Notation and Definitions

We define a *relation* A as a set of *tuples* $A = \{a^j, j = 1, \ldots, c_A\}$, where c_A indicates the *cardinality* of the relation. Each tuple of A is an ordered set of k_A values, where k_A is the *a-rity* of the relation:

$$a^j = \langle a_1^j, a_2^j, \ldots, a_{k_A}^j \rangle \qquad j = 1, \ldots, c_A \qquad (1)$$

$$a_i^j \in \mathcal{V}_i^A \qquad i = 1, \ldots, k_A; \quad j = 1, \ldots, c_A \qquad (2)$$

$$\mathcal{A}_i = \{a_i^j, j = 1, \ldots, c_A\} \qquad (3)$$

The multisets $\mathcal{A}_i$ that contain the values assumed by a given field are called *attributes*. The corresponding sets $\mathcal{V}_i^A$ are called *value-sets*, and we refer to their cardinality o_i^A as the *originality* of the attribute. The number of occurrences of a given value $v_i^A \in \mathcal{V}_i^A$ in the multiset $\mathcal{A}_i$ is called *multiplicity* of the value and is expressed by the function $m(v_i^A)$.

Without any loss of generality we consider a selection that generates a relation B from a relation A. Let us assume that the attribute $\mathcal{A}_h$ is not involved in the operation (i.e. in the selection condition), and let $\mathcal{B}_h$ be the corresponding attribute in the result relation.

Our goal is to compute the originality o_h^B of the attribute $\mathcal{B}_h$ in the result relation. This is a function of the originality of the corresponding attribute $\mathcal{A}_h$, of the cardinality c_A of the operand relation, and of the *selectivity factor* σ of the operation, defined as the fraction of tuples of A that are retained by the selection. More formally we define the *compression factor* as:

$$F_h(o_h^A, c_A, \sigma) = \frac{o_h^B}{o_h^A} \qquad (4)$$

Although we refer through the paper to the simple case of a selection, the results directly apply also to any other operation, or sequence of operations, possibly involving more than one relation, if the appropriate value of the selectivity factor is considered. For instance if B is the result of a join between A and C, for the computation of the compression factor of an attribute $\mathcal{B}_h$ originating from A, σ must be taken as the fraction of the tuples in A that are successfully joined.

3. Non uniform distributions

In this section we analyze how the distribution of the attribute values affects the compression factor. In doing so we initially make an independence assumption, stating that the attribute $\mathcal{A}_h$ that we are considering is stochastically independent from any other attribute $\mathcal{A}_k$ involved in the selection, i.e.:

$$Prob\{a_h^j = x, a_k^j = y\} =$$

$$Prob\{a_h^j = x\} Prob\{a_k^j = y\} \qquad (5)$$

$$x \in \mathcal{V}_h^A, y \in \mathcal{V}_k^A$$

As a consequence of this all the elements of the multiset $\mathcal{A}_h$ have the same probability to be retained in $\mathcal{B}_h$, and then the compression factor does not depend on the structure of the select condition, but only on the selectivity of the operation, i.e. on the reduction of the relation cardinality.

As far as the distribution of the attribute values in $\mathcal{A}_h$ is concerned, we assume that every value $v_k^A \in \mathcal{V}_h^A$ has at least one occurrence in the multiset $\mathcal{A}_h$. The remaining $c_A - o_h^A$ elements of the multiset are distributed in $\mathcal{V}_h^A$, according to a non uniform distribution $\{\pi(v_k^A), v_k^A \in \mathcal{V}_k^A\}$. Therefore the multiplicity of each value $v_h^A \in \mathcal{V}_h^A$ has a binomial distribution:

$$Prob\{m(v_h^A) = n\} = B(n-1, c_A - o_h^A; \pi(v_h^A)) = \qquad (6)$$

$$\binom{c_A - o_h^A}{n-1} \pi(v_h^A)^{n-1} (1 - \pi(v_h^A))^{c_A - o_h^A - n + 1}$$

with an expected value of:

$$E[m(v_h^A)] = 1 + (c_A - o_h^A)\pi(v_h^A) \qquad (7)$$

The expectation of the originality of the attribute in the result relation can be expressed as:

$$E[o_h^B] = \sum_{v_h^A \in \mathcal{V}_h^A} E[\delta_h^B(v_h^A)] \qquad (8)$$

where the random function δ_h^B is defined as:

$$\delta_h^B = \begin{cases} 1 & \text{if } v_h^A \in \mathcal{V}_h^B \\ 0 & \text{otherwise} \end{cases} \qquad (9)$$

In turn the expectation of δ_h^B can be expressed through the conditional probabilities:

$$E[\delta_h^B(v_h^A)] = Prob\{v_h^A \in \mathcal{V}_h^B\} = \qquad (10)$$

$$\sum_{n=1}^{c_A - o_h^A + 1} Prob\{v_h^A \in \mathcal{V}_h^B \mid m(v_h^A) = n\} Prob\{m(v_h^A) = n\}$$

Therefore the problem is reduced to the computation of the conditional probabilities in the second member of (10). These, because of the independence assumption (5), do not depend on the particular value v_h^A, but only on the multiplicity n, and can be computed as the probability that all the n occurrences of v_h^A in the multiset $\mathcal{A}_h$ are in $\mathcal{A}_h - \mathcal{B}_h$, i.e. are discarded by the selection:

$$Prob\{v_h^A \in \mathcal{V}_h^B \mid m(v_h^A) = n\} =$$

$$1 - Prob\{v_h^A \notin \mathcal{V}_h^B \mid m(v_h^A) = n\} = \qquad (11)$$

$$\begin{cases} 1 - \dfrac{\binom{c_A - c_B}{n}}{\binom{c_A}{n}} & c_A - c_B \geq n \\ 1 & \text{otherwise} \end{cases}$$

The (9), (10) and (11) allow to compute the expectation of the originality o_h^B in the general case of non uniform distribution of the attribute values, and assuming stochastic independence among the attributes, under the very reasonable assumption that the multiplicity of the values has a binomial distribution.

A simpler expression can be found assuming a deterministic distribution for the multiplicity $m(v_h^A)$. In this case our formula can be shown to be equival ent to the one proposed by Yao for uniform distribution of the values [14], and to its generalization for non uniform distributions by [8]. However these formulas, that were originally introduced for a different problem, can be used only when the multiplicity has an integer value.

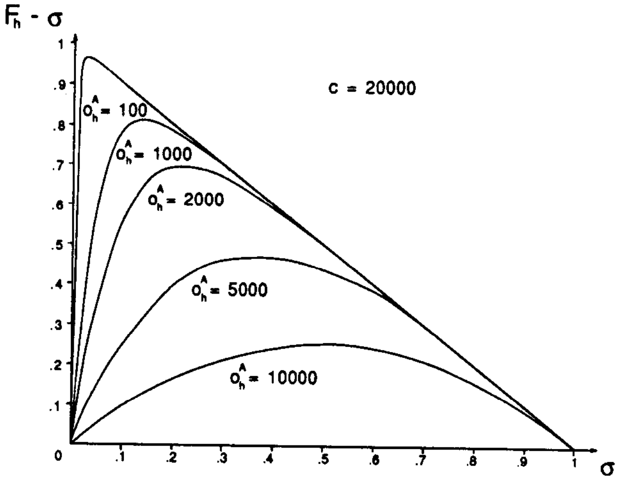

Figure 1: Compression factor with uniform distribution

A good choice for the distribution of the attributes is the Zipf distribution [15], that has been utilized by several authors [7,9] to fit measured data:

$$\pi(v_h^A) = C\, ord(v_h^A)^{-z} \qquad C = \frac{1}{\sum_{i=1}^{o_h^A} i^{-z}} \qquad (12)$$

where $ord(v_h^A)$ refers to the ordering of the set $\mathcal{V}_h^A$ based on the relative frequency in the multiset $\mathcal{A}_h$. The parameter z is a non-negative constant and allows to fit the skewness of the distribution.

The behavior of the compression factor is shown in Figure 1, where the difference $F_h - \sigma$ is plotted as a function of the selectivity factor σ, for several values of the originality. The figure clearly shows that the compression factor is always larger than the selectivity factor, i.e. the originality always decreases less than the cardinality. Moreover the difference decreases with the average multiplicity, i.e. is smaller for larger values of the originality.

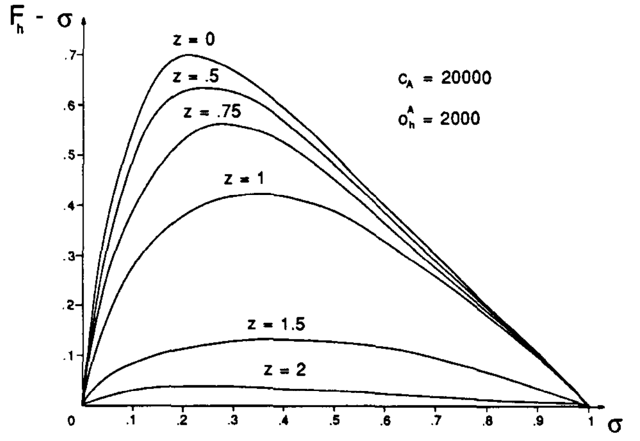

Figure 2: Compression factor with non uniform distribution

Figure 2 depicts the effect of non uniform distributions of the attribute values. For fixed values of the cardinality and the originality, several Zipf distributions of different skewness are considered. The picture shows that the non-uniformity has a moderate effect for low values of the selectivity factor, at least when z is less than .5. To have an idea of the skewness of the distribution, consider that, for the values of c_A and o_h^A in the figure, when $z = .5$ the ratio between the maximum and the minimum average multiplicity is about 40.

Figure 3 shows how the compression factor decreases when z increases. For low values of z the compression factor keeps close to the value for the uniform distribution. The relative errors are reported in Table 1. Note that for $z < .5$ the error is below 10%.

For larger values of z, F_h sharply decreases and approaches a limit value F_h^* that corresponds to a completely skewed distribution. In such cases a better approximation for the compression factor is given by F_h^*. This can be computed considering the extreme situation when a single element of $\mathcal{V}_h^A$ has multiplicity $c^A - o_h^A + 1$, and the remaining $o_h^A - 1$ elements have just one occurrence. Therefore, according to the (9), (10) and (11), F_h^* can be expressed as:

$$F_h^* = Prob\{v_h^A \in \mathcal{V}_h^B \mid m(v_h^A) = c^A - o_h^A + 1\}+$$

$$(o_h^A - 1)Prob\{v_h^A \in \mathcal{V}_h^B \mid m(v_h^A) = 1\} \approx \qquad (13)$$

$$1 + (o_h^A - 1)(1 - \frac{c_A - c_B}{c_A}) = 1 + (o_h^A - 1)\sigma$$

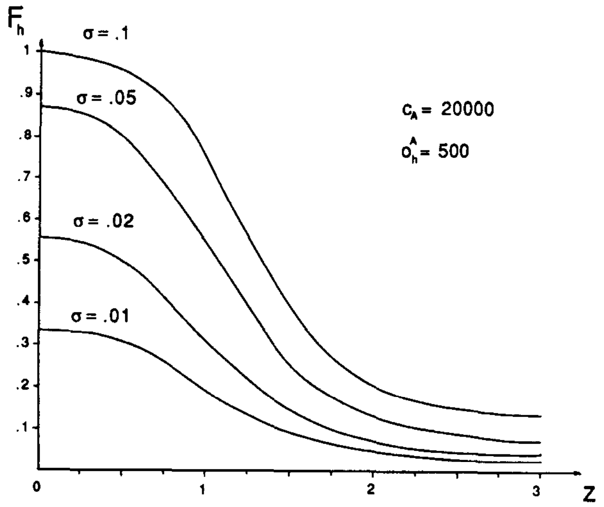

Figure 3: Effect of the non uniformity

	$\sigma = .01$	$\sigma = .02$	$\sigma = .05$	$\sigma = .10$
$z = .25$	.01	.02	.02	.01
$z = .50$	.08	.10	.08	.03
$z = 1.0$	.40	.63	.37	.24
$z = 1.5$	.72	.82	.71	.61
$z = 2.0$	.86	.87	.85	.79
$z = 3.0$	.93	.93	.92	.87

$c_A = 20000$ $\qquad$ $o_h^A = 500$

Table 1: Relative error for non uniform distributions

4. Stochastic dependence between attributes

We now consider the case of stochastic dependence between the attributes involved in the selection and the attribute we are considering for the compression factor. This always produces a decrease in the compression factor, and therefore leads to overestimate the originality if the independence assumption is made [4,13].

We give here an extension of the model proposed in the previous section, that allows to analyze and quantify the effect of the dependence. We restrict our presentation to the case of two attributes, the first one $\mathcal{A}_k$ being involved in the selection and the second one $\mathcal{A}_h$ being investigated for the originality.

As in actual databases one may find many different kind of dependence, it is hard to give a general model to cover all the cases. However it does not make sense to include in the model a very detailed representation of the dependence, as this would require more information on the database than it is usually available.

We propose a simple way to model the dependence, that assumes that in every tuple, for a given value of the first attribute, the value of the second attribute must belong to a proper subset of the corresponding value set.

More precisely, for every value $v_k^A \in \mathcal{V}_k^A$, we define an *associated set* $\mathcal{S}(v_k^A) \subset \mathcal{V}_h^A$ such that for every tuple in the relation:

$$a^j = \langle \ldots, a_k^j, \ldots, a_h^j, \ldots \rangle \qquad a_h^j \in \mathcal{S}(a_k^j) \tag{14}$$

Moreover we assume an uniform distribution of the values for the attribute $\mathcal{A}_k$, and an uniform distribution inside the subsets $\mathcal{S}(v_k^A)$:

$$Prob\{a_h^j = x \mid a_k^j = z\} = Prob\{a_h^j = y \mid a_k^j = z\} \tag{15}$$

$$x, y \in \mathcal{S}(z), z \in \mathcal{V}_k^A$$

A measure of the dependence is given by the ratio between the cardinality of the subset $\mathcal{S}(v_k^A)$ and the originality of $\mathcal{A}_h$, which we assume to be the same for every $v_k^A \in \mathcal{V}_k^A$:

$$\rho = \frac{card(\mathcal{S}(v_k^A))}{card(\mathcal{V}_h^A)} \qquad \frac{1}{o_h^A} \leq \rho \leq 1 \tag{16}$$

It is evident that the dependence is maximum when ρ, that we call *spreading factor*, attains its minimum value. Such kind of dependence can be regarded as a stochastic relaxation of the functional dependence. Although very simple, it can effectively take into account several kind of real world situations. A further advantage is that the model requires to estimate just one parameter.

For example considering the relation *BRANCHES:* ⟨*BANK#*, *BRANCH#*, *CITY#*,...⟩, an estimate of ρ between the attributes *BANK#* and *CITY#* can be given as the average percentage of cities in which a given bank has a branch.

Assuming this kind of stochastic dependence we can derive an approximate model for the computation of the compression factor as an extension of the one presented in Section 3.

Let us consider a selection on $\mathcal{A}_k$ and the set of values $\mathcal{V}_k^B \subset \mathcal{V}_k^A$ that are retained by the operation. Observe that, because of the dependence, especially for low values of the spreading factor ρ and of the selectivity σ, the number of values of the attribute $\mathcal{A}_h$ that can *potentially* be retained by the selection may be considerably lower than o_h^A. More precisely we define a set $\overline{\mathcal{V}}_h^A \subseteq \mathcal{V}_h^A$ of *eligible values*, that we shall call *virtual value set*:

$$\overline{\mathcal{V}}_h^A = \{\overline{v}_h^A \mid \exists x \in \mathcal{V}_h^B, \overline{v}_h^A \in \mathcal{S}(x)\} \tag{17}$$

Accordingly we define a *virtual relation* as the set of the *eligible tuples* of A, i.e. having the value a_h^j belonging to $\overline{\mathcal{V}}_h^A$:

$$\overline{A} = \{\overline{a}^j \mid \overline{a}^j \in A, \overline{a}_h^j \in \overline{\mathcal{V}}_h^A\} \tag{18}$$

As an approximation, to take into account the dependence, we may consider that the selection takes place only on the eligible tuples of A, and then apply the model of Section 3 to the virtual relation. To do this we consider a modified set of parameters, that we shall call *virtual parameters*:

$$\overline{o}_h^A = card(\overline{\mathcal{V}}_h^A) \tag{19}$$

$$\overline{c}_A = card(\overline{A}) \tag{20}$$

$$\overline{\sigma} = \frac{\sigma c_A}{\overline{c}_A} \tag{21}$$

Hence the *virtual originality* $\overline{o}_h^A$ is defined as the expected number of distinct values of $\mathcal{A}_h$ that, according to the constraint (14), can be associated to the values of $\mathcal{A}_k$ retained by the selection. Similarly the *virtual cardinality* in (20) represents the number of tuples in A that have in the h-th field values belonging to the virtual value set. Finally the *virtual selectivity* σ is adjusted to produce the proper cardinality for the result relation.

To compute the virtual originality we consider the probability that a value v_h^A belongs to the associated set of a given value v_k^A of $\mathcal{A}_k$, that, according to the (16), is given by:

$$Prob\{v_h^A \in \mathcal{S}(v_k^A)\} = \rho \qquad (22)$$

As the number of distinct values of $\mathcal{A}_k$ that are retained in the selection is σo_k^A, the probability that v_h^A is discarded during the operation can be computed as:

$$Prob\{v_h^A \notin \mathcal{V}_h^B\} = (1 - \rho)^{\sigma o_k^A} \qquad (23)$$

Therefore the virtual originality is given by:

$$\bar{o}_h^A = o_h^A(1 - Prob\{v_h^A \notin \mathcal{V}_h^B\}) = \qquad (24)$$
$$o_h^A(1 - (1 - \rho)^{\sigma o_k^A})$$

Similarly, because of the uniformity of the distribution of the attribute values, the cardinality of the virtual relation is reduced by the same factor and then:

$$\bar{c}^A = c^A(1 - (1 - \rho)^{\sigma o_k^A}), \qquad (25)$$

Figure 4 shows how the dependence affects the compression factor. Observe how F_h sharply decreases for low values of the spreading factor ρ. As the selectivity increases the effect of the dependence is less important and can be disregarded unless ρ is very low.

	$\rho = .05$	$\rho = .1$	$\rho = .3$	$\rho = .5$	$\rho = .8$
$\sigma = .01$	.001	.014	.019	.038	.028
$\sigma = .02$	.011	.004	.027	.078	.046
$\sigma = .05$	.016	.007	.063	.092	.024
$\sigma = .10$	.001	.002	.049	.026	.001

$c_A = 20000 \qquad o_h^A = 500 \qquad o_k^A = 100$

Table 2: Relative error of the approximate model

The approximate model has been successfully validated against a simulation for several values of the parameters. Sample data are reported in Table 2 that gives the relative error of the approximate model. Most of the time the error is inside the confidence intervals of the simulation, which have a relative width of less than 5%.

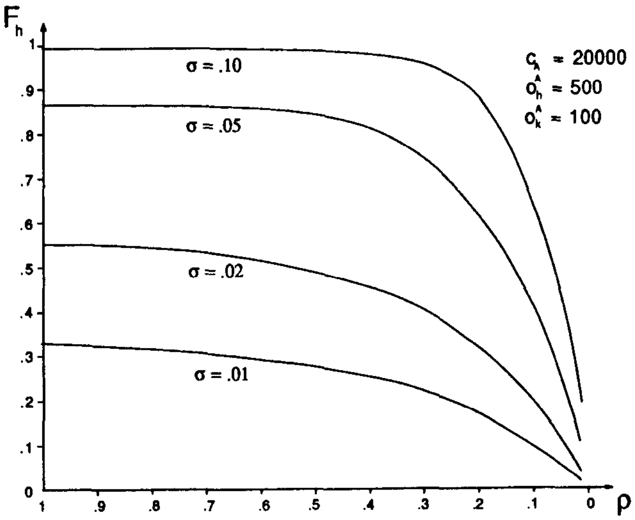

Figure 4: Effect of stochastic dependence

	$\rho = .05$	$\rho = .1$	$\rho = .3$	$\rho = .5$	$\rho = .8$
$\sigma = .01$	5.60	2.37	0.49	0.20	0.05
$\sigma = .02$	4.68	1.90	0.36	0.12	0.02
$\sigma = .05$	2.83	1.00	0.14	0.02	0.00
$\sigma = .10$	1.45	0.51	0.03	0.00	0.00
$\sigma = .20$	0.56	0.10	0.00	0.00	0.00
$\sigma = .50$	0.08	0.01	0.00	0.00	0.00

$c_A = 20000 \qquad o_h^A = 500 \qquad o_k^A = 100$

Table 3: Relative error for stochastic dependence

These results prove that the approximation is very good for small values of ρ, i.e. that the model can be conveniently used when there is a strong dependence. These are indeed the cases of practical interest, in which the effect of the dependence cannot be disregarded. Although larger deviations may occur for large values of the spreading factor the model has proved to be reasonably accurate for a large range of parameter values.

To have an idea of the improvement given by the approximate model, consider the data in Table 3 that reports the relative error introduced by the independence assumption. Note that for low selectivities and strong dependence the error may be larger than 500%.

5. Conclusions

In this paper we have extended a probabilistic model presented in an earlier paper for computing the originality of the attributes of the result of a relational operation. A new methodology is proposed to take into account the effect of the non uniformity of the distribution of the attribute values, and of the dependence between attributes.

The model has been used to investigate how these two factors affect the originality. Numerical results show that the non uniformity must be considered only for very skewed distributions. On the contrary the stochastic dependence has to be taken into account, especially for low values of the selectivity factor. Therefore the model, although approximate, may significantly improve the estimates of the originality in several practical cases. A further step would be to extend the model to consider the simultaneous effect of non uniformity and dependence.

References

[1] P.A. Bernstein et al. Query processing in a system for distributed databases (SDD-1). *ACM Trans. on Database Syst.*, 6(4):602–625, 1981.

[2] S. Ceri and G. Pelagatti. *Distributed Databases: Priciples and Systems.* McGraw-Hill, New York, 1984.

[3] S. Christodulakis. Estimating record selectivities. *Information Systems*, 8(2):105–115, 1983.

[4] S. Christodulakis. Implications of certain assumptions in database performance evaluation. *ACM Trans. on Database Syst.*, 9(2):163–186, 1984.

[5] P. Griffiths Selinger et al. Access path selection in a relational database management system. In *ACM SIGMOD International Conf. on Management of Data*, pages 23–34, 1979.

[6] A.R. Hevner and S. B. Yao. Query processing in distributed database systems. *IEEE Transactions on Software Engineering*, 5(3):177–187, 1979.

[7] L. Kerschberg, P.L. Ting, and S. B. Yao. Query optimisation in star computer networks. *ACM Trans. on Database Syst.*, 7(4):678–711, 1982.

[8] W. S. Luk. On estimating block accesses in database organisations. *Communications of the ACM*, 26(11):945–947, 1983.

[9] C.A. Lynch. Selectivity estimation and query optimization in large databases with highly skewed distributions of column values. In *Fourteenth International Conference on Very Large Data Bases, Los Angeles*, pages 240–251, 1988.

[10] P. Richard. Evaluation of the size of a query expressed in relational algebra. In *ACM SIGMOD International Conf. on Management of Data*, pages 155–163, 1981.

[11] S. Salza and M. Terranova. Database workload modeling. In F. Cesarini and S. Salza, editors, *Database Machine Performance: Modeling Methodologies and Evaluation Strategies, Lecture Notes in Computer Science 257*, chapter 4, pages 50–94, Springer-Verlag, Berlin, 1987.

[12] S. Salza and M. Terranova. Evaluating the cardinality of the result of relational operations: a probabilistic approach. In *6-th Advanced Database Symposium, Tokyo*, pages 223–231, 1986.

[13] B.T. Vander Zanden, H.M. Taylor, and D. Bitton. Estimating block accesses when attributes are correlated. In *Twelfth International Conference on Very Large Data Bases, Kyoto*, pages 119–127, 1986.

[14] S. B. Yao. Approximating block accesses in database organisations. *Communications of the ACM*, 20(4):260–261, 1977.

[15] G.K. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison Wesley Publ. Co., Reading, Massachussetts, 1949.