First International Conference
Modelling and Development of Intelligent Systems
Sibiu - Romania, 22-25 October, 2009

Distribution of result sets cardinalities in heterogeneous random databases

Letiția Velcescu, Laurențiu Vasile

Abstract

In this paper, we present an extension of the concept of random database, in which the records are random vectors following a certain multidimensional probability distribution, to heterogeneous random databases, in which columns can have their own unidimensional distribution. We investigate the sizes of some relational operations results in these databases, focusing on difference, join and outer join. In this approach to random databases, we will show that the number of tuples in the results set is Poisson distributed in the cases of heterogeneous random tables with normal and exponential columns, or discrete and exponential columns, but this behavior depends on the choice of the approximation considered in the relational operations.

1 Introduction

Nowadays, every research, medical, economic or industrial field needs to store and manage large amounts of data. In many cases, these data are likely to be uncertain or to contain errors, but the problem is still to provide a good management and to be able to extract the needed information or to appropriately support the decision making, all based on such uncertain data. This is where the concept of random database has become important.

Our work mainly focuses on the behaviour of relational operations in random databases. As known so far, this type of database supposed a vision of the table as a set of random vectors, following a common multivariate distribution. In order to distinguish this concept from the one we propose, we name it homogenous random database. In this framework, previous research has been already done. We are interested in a generalization of this concept of random database and in studying the behaviour of relational operations in this context. The extended concept is that of heterogeneous random database, in which different columns of the tables can have different probability distributions.

Applying a “traditional” relational operator when working with databases that contain uncertain data will often result in an irrelevant, even empty, data set. Because of uncertainty, one should work with approximate rather than exact operations.

In our approach, we obtained samples of some specific unidimensional distributions, stored them in columns of relational tables and applied approximate relational operators on these random tables. As in the homogeneous case, we want to obtain an estimation of the distribution of the number of lines in the result set of an approximate relational operation. The technique used to confirm the likeliness of a probability distribution is based on the chi square goodness of fit test.

The paper is organized as follows: in the second part, we introduce the main database and random database concepts and results related to the homogeneous case; in the third part, we describe our extension and approach to this field; we will conclude with some considerations and perspectives of our future work in the fourth part of this paper.

2 Preliminaries

In order to introduce the definitions of some fundamental database notions, we will consider the finite domains D_1, D_2, ..., D_n, not necessarily disjoint ([6]).

Definition 1 *The cartesian product* $D_1 \times D_2 \times ... \times D_n$ *of the domains* D_1, D_2, ..., D_n *is defined by the set of the tuples* $(V_1, V_2, ..., V_n)$, *where* $V_1 \in D_1$, $V_2 \in D_2$, ..., $V_\mathrm{n} \in D_n$. *The number n defines the tuple's arrity.*

Definition 2 *A subset of the cartesian product* $D_1 \times D_2 \times ... \times D_n$ *defines a relation R on the sets* D_1, D_2, ..., D_n. *Consequently, a relation is a tuple set.*

There is an alternative definition of a relation, in the terms of a set of functions. Suppose that we associate to each domain D_i an attribute A_i.

Definition 3 *A relation R is a set* $\{f_1, f_2, ..., f_m\}$, *where* $f_i : \{A_1, A_2, ..., A_n\} \rightarrow D_1 \cup D_2 \cup ... \cup D_n$ *and* $fi(Aj) \in Dj$ *for each values of i and j.*

We can easily remark that both definitions of a relation refer to sets which are varying over time. In these sets, elements can be inserted, deleted or updated. Obviously, not the content of this set characterizes a relation, but a time-invariant element. Such an element is the relational schema, which is actually the relation's structure.

Definition 4 *The relational schema of a relation R is defined by the set of the attributes' names which correspond to the relation R. We denote the relational schema by* $R(A_1, A_2, ..., A_n)$.

The representation of a relation can be done by a table in which each line corresponds to a tuple and each column corresponds to an attribute. In other words, a column corresponds to a domain. The relational databases are perceived by the users as a set of tables.

Definition 5 *The degree of a relation is represented by the number of its attributes. The cardinality of a relation is given by the number of its tuples.*

Generally, a table is a representation of a relation, but it is important to mention that there is an important difference between these notions: a table is a sequence of records, contrary to a relation, which is a set of records. This means that the tuples of a relation must be distinct, whereas those of a table can be not.

2.1 Relational operations

The relational operations are performed by the operators of the relational model, which includes the relational algebra. The operators of the relational algebra are either the usual set operators (*union*, *intersect*, *product*, *difference*) or some specific relational operators (*project*, *select*, *join*, *division*). The usual set operators, except *product*, require that the operands had the same type.

The relations in the database are subjected to operations. The result of a relational operation is a new relation.

Definition 6 *The intersection of two relations R and S is the relation composed by the set of the tuples which belong both to R and S.*

Definition 7 *The difference between two relations R and S is the relation composed by the set of the tuples which belong to R but do not belong to S.*

The join operator allows the information retrieval from more correlated relations. The required condition in order to apply the join operator is that the tuples are similar.

Definition 8 *The join between two relations R and S is a binary operation whose result is a new relation in which each tuple is a combination of a tuple in the first relation and a tuple in the second one, satisfying a given join condition.*

The *join* operator composes projection, selection and cartesian product. Generally, the cartesian product is built, some tuples are eliminated by selection and some attributes are eliminated by projection. When dealing with large tables, the cartesian product is a costly operation and consequently, so is the join. In some cases, when evaluating a join operation, even if the volume of the intermediate cartesian product is significant, the result of the join is quite small. Therefore, when joining more than two tables, it is relevant to know the sizes of the joins, in order to perform these operations in the proper order. This is the reason why the work related to the estimation of the number of lines in the result set of each operation is important, in the context of database optimization, when joining multiple tables or query's result sets.

There are several types of the *join* operation, such as the *equi-join* and the *outer join*. The *equi-join* requires that the values of the specified attributes are equal.

From definition 8, it results that a *join* operation will lose a tuple which belongs to a relation if there is no tuple in the other relation such that the *join* condition be satisfied. In order to keep such tuples in the result set, we use the *outer join* operation. This one combines the tuples in the two relations for which the correlation conditions are satisfied, without losing the other tuples. This operator assigns *null* values to the attributes that exist in a tuple of one of the input relations, but does not exist in the second relation. There are three types of *outer join* operators: *left*, *right* and *full*. They keep in the result each tuple of the relation in the left, right, respectively in both relations.

2.2 Approximate relational operations in random databases

In random databases, the operations above find a homologous in the approximate operations. As stated in the introduction, one would get an irrelevant result if using the exact match in the cases when uncertainty arises.

In this section, we will describe what some of the relational operations mentioned above become in the context of approximation in the random databases. In order to define the approximate version of these operations, we consider a distance $d(x, y)$ between the elements in the domains D_A and D_B, which are the projections on the attributes in A and B of the domains D_{U_1} and D_{U_2} and which are assumed to be subsets of a metric space where the distance d is defined ([7]). An example of such a distance is the Hamming distance, given by the number of different join attributes in the two tuples. We denote by $B_\tau\left(x\right)$ the ball with the centre in x having the radius ε.

We define the following ε-operations: *ε-difference*, *ε-equi-join* and *ε-outer-join*. For two relations R and S, we denote these approximate operations by $difference_\varepsilon(R, S)$, $join_\varepsilon(R, S)$ and $outer\text{-}join_\varepsilon(R, S)$, respectively.

Definition 9 *The ε-difference between two relations R and S is a relation containing the following set of tuples:*

$$difference_\varepsilon(R, S) = \{x \in R \mid \neg\exists y \in S, x \in B_\varepsilon(y)\} \tag{1}$$

Definition 10 *The ε-join between two relations R and S is a relation containing the following set of tuples:*

$$join_\varepsilon(\mathrm{R}, \mathrm{S}) = \{(x, y) \in R \times S \mid d(x_A, y_B) \le \varepsilon\} \tag{2}$$

Definition 11 *The ε-outer-join between two relations R and S is a relation containing the following set of tuples:*

$$\textit{outer-join}_\varepsilon(R, S) = \text{join}_\varepsilon(R, S) \cup \{(x, y) \in R \times S \mid \neg\exists y \in S, x \in B_\varepsilon(y)\} \tag{3}$$

The number of lines in the result set of each of these operations is denoted by $N_\varepsilon(\textit{difference}_\varepsilon(R,S))$, $N_\varepsilon(\textit{join}_\varepsilon(R,S))$, respectively $N_\varepsilon(\textit{outer-join}_\varepsilon(R,S))$ ([8]). In the cases when it is clear what these values refer, we denote them by N_ε for simplicity.

The approximative match problems have been already studied for the *equi-join* operation, in the case of the homogeneous random databases.

Definition 12 *Two tuples $x \in D_A$ and $y \in D_B$ are ε-close, with ε ≥ 0, if d(x, y) ≤ ε.*

As one can remark, the *ε-equi-join* operation's result set contains the ε-close tuples according to the given distance. For the particular case ε = 0, we get the usual *equi-join* operation.

2.3 Previous work

The definition of the number N_ε of lines in the join's result has one other significant importance, concerning the constraints in the database, as shown in the previous research on the random databases ([7]). The keys and functional dependencies represent the constraints in a database. The concept of key generalizes to that of functional dependency, which specifies relations between two distinct sets of attributes, meaning that the values of the first set determine the values of the second set of attributes. The cardinality of the set of constraints is extremely important in the database design. A model in which this cardinality is exponential depending on the number of attributes becomes impossible to manage ([2]).

Definition 13 *A minimal set of attributes whose values uniquely identify a tuple in a relation represents a key for that relation.*

Consequently, a key of a relation *R* is a set of attributes *K*, such that ([6]):

i) for each tuples t_1, t_2 of *R*, we have $t_1(K) \neq t_2(K)$;

ii) there is no proper subset of *K* having the property i).

Extending the notion of key to the one of *ε*-key, a set of attributes *K* is an *ε*-key if there are no *ε*-close tuples $t_i(K)$, $i = 1,\ldots,m$. Such a property is denoted by $R \models_\varepsilon K$. For the particular case *ε* = 0, one gets the usual definition of a key.

The set of all attributes of a relation composes a key, but the keys are better as their attributes set is smaller. We have that $N_\varepsilon(\textit{join}_\varepsilon(R,R))$ is the number of *ε*-close tuples in *R*, for a specified attributes join set *A*. The distribution of $N_\varepsilon(\textit{join}_\varepsilon(R,R))$ defines the capacity of the set *A* to distinguish the tuples in the relation *R*. Thus, the attributes set *A* is an *ε*-key if and only if $N_\varepsilon(\textit{join}_\varepsilon(R,R)) = m$.

Initially, the problem of characterization of the most relevant properties of the constraints has been addressed in the worst case ([1]). Then, the problem has been studied in the average case for a general class of probabilistic models ([7]). In this second approach, the entropy of the records distribution was used in order to explain the properties of the constraints.

A recent research direction concerning the random databases uses, as main tools, the Poisson approximation and the Rényi *ε*-entropy. This type of entropy is introduced as a generalization of the Rényi entropy for the discrete distributions.

In the stochastic models which were considered in the random databases modeling so far, it is assumed that, in a random table *T*, the tuples are random vectors with n elements, independent and identically distributed, with a common probability distribution *P*.

In the previous research, three types of random databases were considered ([8]): uniform, if *P* is a uniform multivariate distribution; normal, if *P* is a multivariate normal distribution; Bernoulli, if the attributes A_i are independent and identically distributed, with a common univariate discrete distribution. Two particular cases of the latter database type are the standard Bernoulli database, where the attributes are uniformly distributed, and the conventional Bernoulli databases, where the domains of the attributes have the property $|D_i| = 2$.

3 Cardinality of ε-operations result sets in heterogeneous random databases

The technique used to estimate the distribution of the N_ε values is the chi square test of goodness of fit ([5]). Before that, we generated the histograms for the frequency of the number of lines in the results set. These histograms indicate the possibility that the N_ε values are Poisson distributed.

Our approach considered both cases: homogenous random tables, denoting the concept existing so far, and heterogeneous random tables, which represent an extension we propose. We recall that a heterogeneous random table is a table in which different columns can have different unidimensional distributions.

3.1 Cardinality of ε-operations on heterogeneous tables

For the heterogeneous case, we considered random tables in which we used either samples of two continuous distributions, namely the normal $N(0, 1)$ and exponential *Exp*(1), or samples of a discrete and a continuous distribution, namely *Binomial*(200, 0.5) or *Geometric*(0.5) and exponential *Exp*(1).

We considered two random relations $R(A_1, A_2)$ and $S(B_1, B_2)$, with at most two attributes each, differently distributed and containing 1000 lines. In Table 1 we show the types of relations we worked with.

We made the implementations of these tables and performed the queries representing the relational operations in the *Oracle 11g* DBMS. Afterwards, we realized the histograms concerning the cardinality of the ε-operations.

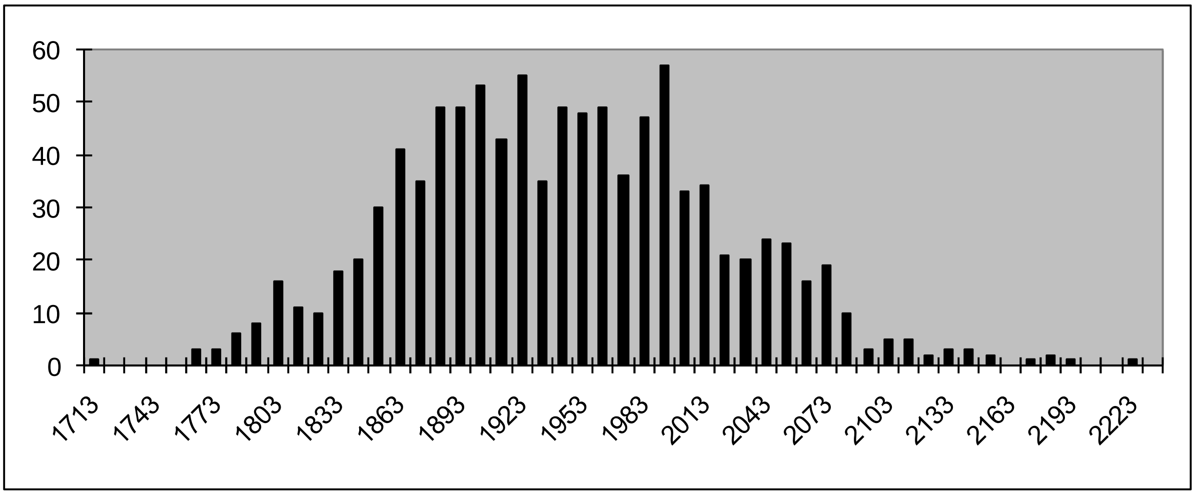

Figure 1. *Histogram for the* 0.05*-join operation between two sets of columns, each with a Binomial*(200, 0.5) *column and an Exp*(1) *one.*

As an example, Figure 1 shows the histogram for the ε-join in the heterogeneous case, for a *Binomial*(200, 0.5) column and an *Exp*(1) one, with $\varepsilon = 0.05$. In Figure 2, we have the histogram for the similar distributions, but $\varepsilon = 0.01$.

A1	A2	B1	B2
Bin(200, 0.5) or Geom(0.5)	-	Bin(200, 0.5) or Geom(0.5)	
Exp(1)	*N*(0, 1)	Exp(1)	*N*(0, 1)
Bin(200, 0.5) or Geom(0.5)	Exp(1)	Bin(200, 0.5) or Geom(0.5)	Exp(1)

Table 1 *The distribution of the columns in the random tables.*

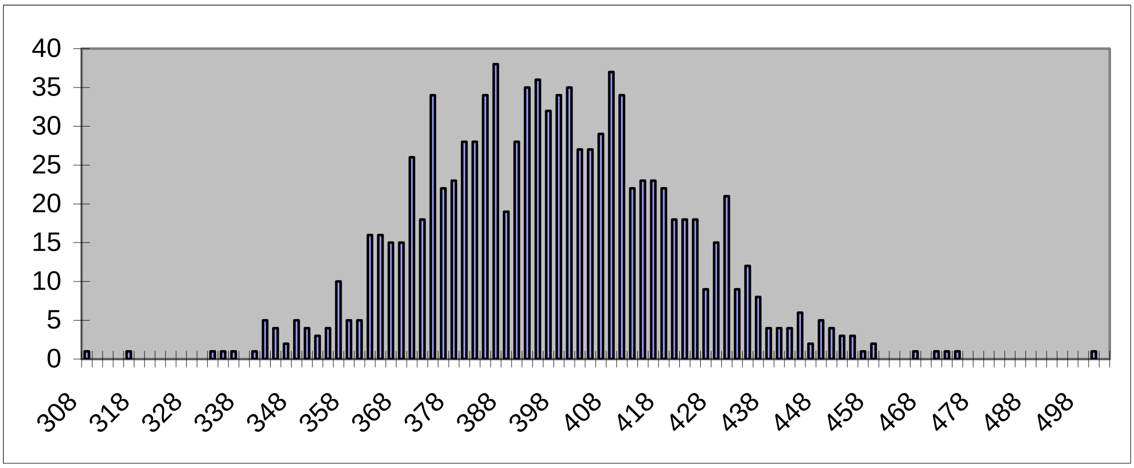

Figure 2. *Histogram for the* 0.01*-join operation between two sets of columns, each with a Binomial*(200, 0.5) *column and an Exp*(1) *one.*

From these histograms, one could observe that it is possible that the random variable N_ε follow a Poisson process on the line. This result has been already stated for homogeneous random databases in [8]. In our homogeneous cases we considered, we observed that the property that the number of lines in the *ε-join* operation has a Poisson distribution depends on the value of ε. This means there is a threshold up to which this distribution is followed.

3.2 The Chi square test of goodness of fit for the distribution of the result of the heterogeneous ε-operations

Consider *N* experiments, where *N* is the number of points on the line for which $P(N=n)=\frac{\lambda^n}{n!}\cdot e^{-\lambda},\ \lambda>0$. The parameter λ of the corresponding Poisson distribution, also called the intensity of the Poisson process, is *E*(*N*). The parameter λ is constant, which determines that the Poisson process is homogenous.

In order to estimate the parameter λ, we consider a sample $\nu_1,...\nu_N$ of the homogenous uniform Poisson process, for a given ε, and we take the following estimation of the Poisson parameter:

$$\hat{\lambda}_N=\frac{1}{N}\sum_{i=1}^{N}\nu_i \tag{4}$$

The chi square test of goodness of fit is the technique used in order to show that the values of the random variable N_ε can be estimated by a homogenous uniform Poisson process.

In this respect, we take k the number of distinct values $\nu_1,\ldots,\nu_k$ in the sample of N_ε and $f_1,\ldots,f_k$ the frequencies corresponding to each value. Obviously, we have that $f_1+\ldots+f_k=N$. We compute the following theoretic probabilities:

$$p_j = P(N_\varepsilon = \nu_j) = \frac{\hat{\lambda}_N^{\nu_j}}{\nu_j!}\cdot e^{-\hat{\lambda}_N} \tag{5}$$

Based on the above values, we can we determine the following statistics:

$$\chi_c^2 = \sum_{j=1}^{k}\frac{(f_j - N\cdot p_j)^2}{N\cdot p_j} \tag{6}$$

Because the parameter λ_N has been estimated before, it implies that this statistics has the distribution χ^2 with $k-2$ degrees of freedom.

The last step of this test is to determine if there is a level of significance $\alpha \le 0.05$ such that $P(\chi_c^2 \ge \chi_{k-1,\alpha}^2)=\alpha$. In the case when the inequality $\chi_c^2 < \chi_{k-1,\alpha}^2$ is true, we can state that the values ν_i do not differ significantly from a Poisson distribution. Otherwise, the possibility that the values ν_i follow a Poisson distribution is rejected.

4 Conclusions and perspectives

We showed that it is possible to extend the concept of random tables to more general cases, in which one table can host columns of different probability or mass distributions. For the two heterogeneous random table cases above, with a binomial and an exponentially distributed column, we considered the values 0.05, 0.005 and 0.001 for ε. In this case, we remarked that the chi square test of goodness of fit does not pass for larger values of ε (e.g. 0.05), but it passes successfully for smaller values (e.g. 0.01 or 0.005). The same remarks are available for the case when the binomial distribution was replaced by a geometric one.

We also performed the test in the case of the homogeneous random tables with normal distribution or exponential distribution. Here, the chi square test of goodness of fit does not pass for a larger value of ε (e.g. 10^{-3}); for $\varepsilon = 10^{-4}$ the test fails, but very closely, and for smaller values of ε (e.g. 10^{-5}), the test passes, so we can state that the cardinality is Poisson distributed.

These conclusions remain true in the case of the difference or outer-join operations.

Concerning the perspectives of our work, one direction would be to determine accurately the threshold up to which the ε-operations cardinalities remain Poisson distributed. Another direction is to find the dependency between the value of ε, the sample size and the acceptance of the Poisson distribution. We also intend to extend the meaning of heterogeneous databases to the case in which some combinations of attributes can follow multidimensional distributions.

References

[1] S. Abiteboul, R. Hull and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] A. Gupta, Y. Sagiv, J.D. Ullman and J. Widom. Efficient and complete tests for database integrity constraint checking. *PPCP 1994*, LNCS vol. 874, Springer, 1994.

[3] N. Johnson, A. Kemp and S. Kotz. *Univariate Discrete Distributions*. 3rd edition, John Wiley & Sons, 2005.

[4] W. Martinez and A. Martinez. *Computational Statistics Handbook with Matlab*. Chapman&Hall, 2002.

[5] Gh. Mihoc and V. Craiu. *Tratat de statistică matematică*. vol. 2, Editura Academiei RPR, Bucuresti, 1977.

[6] I. Popescu and L. Velcescu. *Proiectarea bazelor de date.* Editura Universității din București, 2008.

[7] O. Seleznjev and B. Thalheim. Average Case Analysis in Database Problems. *Methodology and Computing in Applied Probability*, Kluwer Academic Publishers, 2003.

[8] O. Seleznjev and B. Thalheim. Random Databases with Approximate Record Matching. *Methodol. Comput. Appl. Probab.*. Springer Verlag, 2008.

[9] I. Văduva. *Modele de simulare*. Editura Universităţii din Bucureşti, 2004.

LETIŢIA VELCESCU
University of Bucharest
Faculty of Mathematics and Informatics
Department of Informatics
14, Academiei Street, 010014 Bucharest
ROMANIA
E-mail: letitia@fmi.unibuc.ro

LAURENŢIU VASILE
University of Bucharest
Faculty of Mathematics and Informatics
Department of Informatics
14, Academiei Street, 010014 Bucharest
ROMANIA
E-mail: vsl@fmi.unibuc.ro