A Bidimensional Poisson Stochastic Process Perspective of the Result Sets Cardinalities in Random Database Queries

Letiţia Velcescu

***Abstract*—The results presented in this paper follow the author's research work in the heterogeneous random databases field. This particular type of database involves columns whose values are distributed according to different probability distributions. Usually, the data stored in a random database are likely to be uncertain, so that the operations on these data have to be modified in order to consider approximations. In this paper, we present the research which led us to the Poisson estimation of the probability distribution of the cardinality of the result set of the approximate join operation. Then, we will approach the distribution of the cardinalities values of the result set of a ε-join operation by the means of a homogeneous bidimensional Poisson process. In this context, we will prove that the cardinalities follow this type of stochastic process. Thus, the algorithms for the simulation of the Poisson bidimensional process can be applied in the simulation of the cardinalities values of the approximate join's result sets.**

***Index Terms*—Poisson distribution, poisson bidimensional stochastic process, random database, approximate join, simulation.**

I. Introduction

The databases containing random or uncertain information are important in a diversity of research fields, like bioinformatics, medicine, physics, economy, communications and notably those fields where data might be measured by sensors. Usually, besides the uncertainty, these data are characterized by their big volume, as they are stored and processed in large databases and data warehouses. In such databases, the search and identification are performed by the approximate matching of the records, instead of the classical relational operations.

In the context of queries which take place in high volumes of data, using an approximate version of the relational join, the optimization of this costly operation becomes very important. Due to the fact that, generally, a relational join operation is supposed to perform a cartesian product operation between the record sets of the relations involved in join [1], a useful approach for the optimization of this operator is to estimate the number of records resulting from each join and then order the rest of the operations adequately. This means to order the join operations in the ascending order of the cardinality of their results sets.

The first probabilistic approach to the random databases field considered relations in which the records are random vectors following a multidimensional probability distribution [2]. In the previous work, we defined the heterogeneous random databases [3] and established an estimation of the approximate join operation in this framework. Actually, the cardinality of the result of this operation is distributed poisson. In this paper, we will consider a new approach, regarding this result from the perspective of a bidimensional Poisson stochastic process, we will state and prove results in this context. This approach is important in the simulation of random databases, as there are algorithms that allow the simulation of this type of stochastic process [4].

In the second section of this article, we will present the main concepts and results we used in our research concerning the estimation of approximate join operation result set. The third section introduces the bidimensional Poisson perspective of this problem and the results we reached. The fourth section presents the algorithm for the simulation of a bidimensional Poisson stochastic process. Also, the algorithm for the simulation of a multidimensional Poisson stochastic process is presented, related to the generalization of the approximate join that we wish to realize in our future work. The article will end with a conclusion section.

Manuscript received November 5, 2012; revised January 31, 2013. This work was supported by the strategic grant POSDRU/89/1.5/S/58852, Project "Postdoctoral programme for training scientific researchers" cofinanced by the European Social Found within the Sectorial Operational Program Human Resources Development 2007-2013.

Letiţia Velcescu is with the Faculty of Mathematics and Computer Science, University of Bucharest, Romania (e-mail: letitia@ fmi.unibuc.ro).

II. The Estimation of the Approximate Join Result Set's Cardinality in Random Databases

A. Concepts

Consider a relation R in a random database. This structure can be regarded as a matrix with m rows (the tuples of the relation) and n columns (the attributes of the relation). The number n defines the arity of a tuple in the relation R [5]. Following the concepts in certain research papers in the databases field (e.g., [6]), the number of tuples in a relation is referred as the relation's cardinality.

In the research performed on random databases [2], [7], the notion of table, which is a multiset of tuples, was considered instead of the relation. In such a structure, the cardinality might be greater than the number of distinct tuples.

Consider the set of all attributes $U = \{A_1 ... A_n\}$ in the relation R. For each subset $A \subseteq U$, the projection corresponding to the j-th tuple $t(j) = t_U(j)$ is denoted by $t_A(j)$, for $j = 1, ..., m$. Each attribute A_i takes values in an associated domain D_{A_i}; thus, the tuples' values will belong to the

cartesian product $D_U := \bigtimes_{A_i \in U} D_{A_i} \lim_{x \to \infty}$.

When needed, we will use the usual notations for the database specific operations [5]. Consider $T = (t_{U_1}(j), j = 1,..., m_1)$ and $S = (s_{U_2}(j), j = 1,..., m_2)$ two tables with the domains D_{U_1} and, respectively, D_{U_2}. We denote by $|C|$ the cardinality of a finite set C. In the case of the equi-join operation, all the comparisons are equalities and we write $T \bowtie_{A=B} S$, where $|A| = |B|$, $A \subseteq U_1$, $B \subseteq U_2$. The result of this operation is a table that contains concatenated tuples from T and S such that $t_A(i) = s_B(j)$, for $i = 1,\ldots,m_1$ and $j = 1,\ldots,m_2$. The attributes having the same name in the two tables are qualified by the table's name.

The approximate matching problems in random databases have been studied for the equi-join operation, whose definition was modified as follows. Consider $d(x, y)$ a distance between two elements $x \in D_A$ and $y \in D_B$, where D_A and D_B are the projections of D_{U_1} and D_{U_2} on the attribute sets A and B, $|A| = |B|$, $A \subseteq U_1$ and, respectively, $B \subseteq U_2$. Suppose that D_A and D_B are subsets of a metric space on which the distance d is defined.

Definition 1.[2] The values $x \in D_A$ and $y \in D_B$ are ε-close, $\varepsilon \geq 0$, if $d(x, y) \leq \varepsilon$.

Consider the approximate join operation, denoted by $T \bowtie_{A \approx B} S$, whose result is composed of the ε-close tuples according to the given distance. For the particular case $\varepsilon = 0$, we obtain the equi-join operation.

Definition 2. Consider the tables T and S. The ε-join operation is defined as follows:

$$join_\varepsilon(R, S) = \{(x, y) \in R \times S \mid d(x_A, y_B) \leq \varepsilon\} \quad (1)$$

Denote by $N_\varepsilon = N_\varepsilon(join_\varepsilon(R,S))$ the cardinality of the result of the ε-join operation.

B. Probability Distribution of the Approximate Join's Cardinalities

In order to estimate the distribution of the values N_ε, we considered the concept of heterogeneous random table, in which different subsets of columns can follow different probability distributions. In this context, we proposed two methods of estimation [3].

The first method consisted of the generation of histograms for the join result of the considered random tables and then we applied the χ^2 test [8] in order to determine if the cardinalities' distribution is Poisson, as suggested by the obtained histograms.

The second method to justify that the number of records in the result set obtained in a ε-operation on random tables follows a Poisson distribution was accomplished by the means of a Poisson approximation using the Stein-Chen method [9]. This method starts from a distribution $\mathcal{P}$ which we want to approximate by a simpler distribution $\mathcal{Q}$. The simplicity refers to the possibility to simulate the respective random variables. The proof of the Poisson estimation of the cardinalities distribution was realized using concepts as entropy [10] and coincidence probabilities; the difference between the actual probability and the Poisson one was measured by the total variation distance [2].

Both approaches lead to the conclusion that the values N_ε are distributed Poisson of parameter λ, where λ is the mean value of N_ε.

III. A Bidimensional Poisson Approach to the Approximate Join in Random Databases

In this section, we will approach the distribution of the cardinalities values of the result set of a ε-join operation by the means of a homogeneous bidimensional Poisson process.

Consider the random tables T, S, whose attributes A and, respectively, B are involved in the ε-join condition. Naturally, these attributes have the same domain of values D. Suppose that the values of A and B follow a unidimensional probability distribution on the domain D. The type of the distribution is the same for each attribute and their parameters are equal.

Without loss of generality, consider that the domains in which the attributes A and B take values are the intervals $[0, L]$, respectively $[0, M]$. Consequently, the records in the ε-join operation result will correspond to points in the rectangle $\mathcal{D} = [0, L] \times [0, M]$.

We will define the bidimensional Poisson process of intensity λ and we will show that the number of result records in the ε-join operation for sets formed by a single attribute follow a process of this type.

Definition 3. ([11]) A process which consists of random points in a bidimensional plane is a Bidimensional Poisson process of intensity λ if the following conditions are satisfied:

1) The number of points which appear in any region of area C is distributed Poisson of parameter λC.
2) The number of points which appear in disjoint regions are independent.

We will consider that the points in the rectangle $\mathcal{D}$ are uniformly distributed.

From the results we mentioned in section II.B, we know that the number of points $N_\varepsilon = N_{\mathcal{D},\varepsilon}$ in the rectangle $\mathcal{D}$ is distributed Poisson (λ), with the parameter λ specified before as the mean value of N_ε. Let D be the area of the rectangle $\mathcal{D}$. From the uniformity hypothesis mentioned above, we can consider that the number of points in a rectangle of area C, included in the rectangle $\mathcal{D}$, is distributed Poisson of parameter $\lambda \cdot \frac{C}{D}$.

Denote:

$$\lambda' = \frac{\lambda}{D}, \quad (2)$$

so that the number of points in a rectangle of area C is distributed Poisson of parameter $\lambda' C$. Consequently, the conditions of the definition 3 are satisfied and we can state the following result:

Proposition 1. The cardinality of a ε-join operation between the attributes A and B of the tables T, respectively S, with A and B following the same probability distribution, forms a homogeneous bidimensional Poisson process of

parameter λ' given in formula (2).

A consequence of the proposition 1 is the establishment of a relation between *m*, ε and *D*. Suppose that the table S is decomposed in *m* tables S_i, $1 \le i \le m$, each having a single record, on which one can find the value B_i of the attribute *B*. Thus, the join operation is decomposed in *m* join operations between the tables S_i, $1 \le i \le m$, and the table *T*. Denote by N_ε^i the number of records obtained after the join between S_i and *T*, $1 \le i \le m$.

From the previous considerations, we know that N_ε^i is a Poisson random variable of parameter λ_i. We will suppose that the sets of result records obtained in the *m* join operations are disjoint. Then, the following relation takes place:

$$\sum_{i=1}^{m} N_\varepsilon^i = N_\varepsilon \tag{3}$$

From the property of the sum of Poisson variables, it is known that:

$$\sum_{i=1}^{m} N_\varepsilon^i \sim Poisson(\sum_{i=1}^{m} \lambda_i) \tag{4}$$

From the relations (3), (4) and because N_ε is distributed Poisson of parameter λ, it results that:

$$\sum_{i=1}^{m} \lambda_i = \lambda \tag{5}$$

From proposition 1, we obtain that:

$$\lambda_i = \lambda' \cdot mas(\mathcal{B}_\varepsilon(B_i)) \tag{6}$$

From the relations (5) and (6), we obtain the following result:

Proposition 2. Consider the ε-join operation between the attributes *A* and *B* of the random tables *T*, respectively *S*, and B_i, $1 \le i \le m$, the distinct values of the attribute *B*. Then:

$$\sum_{i=1}^{m} mas(\mathcal{B}_\varepsilon(B_i)) = \frac{\lambda}{\lambda'} \tag{7}$$

Because $mas(\mathcal{B}_\varepsilon(B_i))$ depends on ε and $\frac{\lambda}{\lambda'} = D$, proposition 2 provides the connection between *m*, ε and *D*. If we take into consideration the standard Lebesgue measure, then:

$$mas(\mathcal{B}_\varepsilon(B_i)) = \varepsilon^3, \tag{8}$$

so the relation (7) becomes:

$$m\varepsilon^3 = D \tag{9}$$

Taking into consideration the result from proposition 2, we can use methods of simulation of the cardinalities of these processes, based on the methods of simulation of the bidimensional Poisson processes.

IV. Simulation of a Bidimensional Poisson Process

In the unidimensional case, it is known that the distances between the random points of the Poisson process of parameter λt, $t \in [0, \infty)$, are distributed exponentially of parameter λ. From this observation, it results the following algorithm (SIMPO) for the simulation of a *k* points trajectory of the homogeneous unidimensional Poisson process of intensity λ [4].

The algorithm SIMPO produces the sequence $T_1, T_2, ..., T_k$ which is a trajectory of the process *Poisson*(λ) on the interval $[0, \infty)$. In the second step of the algorithm, the simulation of the variable *E* can be done by the means of classical simulation methods, such as the inverse method or the rejection method [12].

Algorithm Simpo: Simulation of a Poisson Process

Input: λ, *k*

Step 1. $i := 0$; $T := 0$;

Step 2. Repeat

Generate $E \sim Exp(1)$;

$i := i + 1$; $T_i := T_i + E / \lambda$;

Until $i = k$.

An algorithm for the simulation of a bidimensional Poisson process, derived from the previous algorithm, on the rectangle $A = [0, t] \times [0, 1]$, with the intensity λ, is the following [4]:

Algorithm Simbpo: Simulation of A Bidimensional Poisson Process

Input: *k*

Step 1. Generate $T_1, T_2, ..., T_k$ a Poisson trajectory on [0, *t*];

Step 2. Generate $U_1, U_2, ..., U_k$ indepedant random variables, uniformly distributed on the interval [0, 1].

Output: The points $(U_1, T_1), (U_2, T_2), ..., (U_k, T_k)$, which determine a uniform Poisson process on *A*.

In the previous algorithm, the value *k* is an integer value of selection of the random variable distributed Poisson of parameter λt.

In our future work, we will generalize the approximate join operation to *n* tables. In this case, a multidimensional Poisson process will be needed in the simulation of the results' cardinalities. The algorithm needed in this case can be built similarly to the previous SIMBPO; thus, one obtains an algorithm for simulating a uniform Poisson process of intensity λ on the *n*-dimensional interval $I = [0, T_1] \times [0, T_2] \times ... \times [0, T_n]$.

Denote $V_0 = \prod_{i=2}^{n} T_i$ the volume of the $(n-1)$-dimensional interval $I_1 = [0, T_2] \times ... \times [0, T_n]$, then we obtain the algorithm SIMPOMD above for simulating this type of Poisson process.

The vectors obtained as output of the following algorithm are an implementation of the Poisson process of intensity λ on the *k*-dimensional interval *I*. The points $Q_i = (X_{1i}, P_i)$, $1 \le i \le k$, determine a uniform Poisson process of parameter λ on *I*.

Algorithm Simpomd: Simulation of a Bidimensional Poisson Process

Input: k, λ

Step 1. Generate the sequence $0 < X_{11} < X_{12} < ... < X_{1k}$, which is a unidimensional Poisson process of parameter λV_0 on the interval [0, T_1], as follows:

1.1 Initialize $t := 0$; $k := 0$;

1.2 Repeat

Generate $E \sim Exp(1)$

$$k := k + 1;\ t := t + \frac{E}{\lambda};\ X_{1k} := \frac{t}{V_0};$$

Until $X_{1k} \geq T_1$;

Step 2. Generate $P_1, P_2, ..., P_k$ independent points, uniformly distributed on I_1.

Output: $(X_{11}, P_1), ..., (X_{1k}, P_k)$.

V. Conclusion

This article presented an approach to the problem of the estimation of the result set cardinality's probability distribution in heterogeneous random databases' approximate join operation. This problem was previously studied from the point of view of the estimation of the probability distribution of the result sets' cardinalities. With the approach presented in this article, using the algorithms from the fourth section, we provided a method of simulating the cardinalities values in the approximate join operations.

References

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

[2] O. Seleznjev and B. Thalheim, "Random databases with approximate record matching," *Methodology and Computing in Applied Probabability*, Springer, vol. 12, no. 1, pp. 63-89, 2008.

[3] L. Velcescu and L. Vasile, "Relational operators in heterogeneous random databases," *IEEE Proceedings of the 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (SYNASC), IEEE Computer Society Press, 2009, pp. 407-413.

[4] F. Suter, I. Văduva, and B. Alexe, "On simulation of poisson processes to be used for analyzing a bivariate scan statistic," Scientific Annuals of the University, Al. I. Cuza, Romania, Tom XV, pp. 23-35, 2006.

[5] M. Kifer, A. Bernstein, and P. Lewis, *Database Systems, an Application-Oriented Approach*, Addison Wesley, 2005.

[6] B. Thalheim, "Konzepte des datenbank-entwurfs," *Entwicklungsten denzen bei Datenbanksystemen*, 1991, pp. 1-48.

[7] G. O. H. Katona, "Random Databases with Correlated Data," *Conceptual Modelling and Its Theoretical Foundations*, Lecture Notes in Computer Science, vol. 7260, Springer-Verlag, pp. 29–35, 2012.

[8] G. H. Mihoc and V. Craiu, *Tratat de statistică matematică*, vol. 2, Academy of Romania Publishing House, Bucharest, 1977.

[9] A. D. Barbour, L. Holst, and S. Janson, *Poisson approximation*, Clarendon, Oxford, 1992.

[10] O. Onicescu and V. Ștefănescu, *Elements of Informational Statistics with Applications* (in Romanian), Technical Publishing House, Bucharest, 1979.

[11] S. Ross, *Simulation*, Academic Press, San Diego, London, 1997.

[12] I. Văduva, *Simulation Models* (in Romanian), University of Bucharest Publishing House, Bucharest, 2005.

Letitia Velcescu was born in Bucharest, Romania. She got her bachelor degree in Computer Science, the Master of Science degree in Applied Computer Science and the PhD in Mathematics (2010) at the Faculty of Mathematics and Informatics of the University of Bucharest. Then, she pursued an academic career in the same institution. In present, she is a Lecturer in the Department of Informatics. Her main research interests cover database theory, algorithms and data structures, modeling and simulation, theory of probabilities and statistics.