Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/070411

Simulation of Join Cardinalities in Random Databases Using Poisson Stochastic Processes

*Letiţia Velcescu**

University of Bucharest, Faculty of Mathematics and Computer Science, Academiei 14, 010014 Bucharest, Romania

Received: 6 Jan. 2013, Revised: 19 Feb. 2013, Accepted: 9 Mar. 2013
Published online: 1 Jul. 2013

Abstract: In this paper, using Poisson stochastic processes, we introduce an approach to the problem of the approximate join operation in random databases. It is shown that the cardinalities of the result sets obtained from an approximate join operation follow this type of stochastic process. Initially, we treat the case of the approximate join between two random tables, which is accomplished using a homogeneous bidimensional Poisson process. Further, we extend the obtained result to the case of the multiple join. This generalization is made through a multidimensional Poisson process. As a consequence, the algorithms that simulate these processes can also simulate the cardinalities of the sets resulting from the approximate join operation.

Keywords: Poisson multidimensional stochastic process, random database, approximate join, simulation

1 Introduction

The uncertainty or randomness of the information is relevant or even specific to a multitude of research or activity fields, including computers, physics, biology, medicine, telecommunications, electronics. Moreover, as the computer science trends are continuously changing, this type of information has a significant meaning in the context of mobile applications and service-oriented computing [1]. The importance of databases storing random or uncertain information is also given by the fact that, generally, besides the uncertainty, these data are characterized by their considerable volume. Due to its size, this information is stored and processed in large databases and data warehouses. In databases containing this type of data the search and identification are adjusted in order to support the approximate matching of the records, instead of the exact match [2,7]. The adjustment of the relational operations is needed as, in this case, an exact match query could possibly return empty result sets.

The most costly relational operation in databases is the join. The big cost of this operation is caused by the fact that, during its processing, a cartesian product between the two relations is determined [3]. In high volumes of data, the optimization of this operation is very important, as it can make the difference between a real-time processing and a prohibitive time of response to a query.

Consequently, in the context of random databases, we have to cope with the optimization of the modified, approximate join. A useful approach for the optimization of multiple joins is to estimate the number of records resulting from each join and then order the sequence of these operations according to the cardinalities of their result 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]; an estimation of the number of records resulting from the approximate join operation was established. In the previous work, we defined the heterogeneous random databases [4], in which the values of the columns follow different probability distributions, and extended the results from the homogeneous case in this framework. It was concluded that the cardinality of the result of this operation follows a Poisson probability distribution, in both cases.

In this paper, a new approach to this estimation is presented. The result of the approximate join is treated from the perspective of a bidimensional Poisson stochastic process. We state and prove the results which arrised in this context. Further, we generalize these results to the case of multiple joins, using the multidimensional Poisson stochastic process. This new approach is important in the simulation of random databases, as there

* Corresponding author e-mail: letitia@fmi.unibuc.ro

are algorithms that allow the simulation of this type of stochastic process, both in the bidimensional and multidimensional case [5].

The second section of this paper presents the main concepts in random databases and a short state of art of the research that concluded with the Poisson estimation of the probability distribution of the approximate join's result set cardinalities. The third section introduces the bidimensional and multidimensional Poisson perspective of this problem and the results we reached. The fourth section presents the algorithms for the simulation of the Poisson stochastic process, both bidimensional and multidimensional. The article ends with a conclusion section which briefly emphasizes the main contributions of the presented research.

2 Overview of the queries result's cardinality in random databases

2.1 Concepts

Consider a relation R in a random database. This structure is described by the relation schema $R(U)$, where $U = \{A_1, A_2, \ldots, A_l\}$ is the set of all attributes in the relation R; the relation is implemented by a (random) table $T(R)$. The table can be regarded as a matrix with m rows (the tuples) and l columns (the attributes of the relation). The number l defines the arity of a tuple in the relation R [6], while the number m of tuples in a relation is referred as the relation's cardinality (e.g., [8]). Due to the identification problems that might appear in random databases, in the research performed in this field [2,9] the notion of table was rather considered, instead of the relation. Because a table is a multiset of tuples, its cardinality is greater or equal to the number of distinct tuples.

Each attribute A_i, for $i = \overline{1,l}$, has an associated domain of values D_{A_i}; thus, the tuples' values will belong to the cartesian product $D_U := D_{A_1} \times D_{A_2} \times \ldots \times D_{A_l}$. For each subset $A \subseteq U$, the projection corresponding to the j-th tuple t_j is denoted by $pr_A(t_j)$, for $j = \overline{1,m}$. Consider $T = \left\{t_j \middle| j = \overline{1,m_1}\right\}$ and $S = \left\{s_j \middle| j = \overline{1,m_2}\right\}$ two tables with the attribute sets U_1 and U_2 and the domains D_{U_1} and D_{U_2}, respectively [6]. The equi-join operation between T and S, denoted by $T \bowtie S$, results in a new table whose records are determined based on equality tests between the corresponding values in the attributes sets A and B, where $A \subseteq U_1$, $B \subseteq U_2$ and $|A| = |B|$. More precisely, the result set of this operation contains combined tuples t_i, s_j from T and S, respectively, such that $pr_A(t_i) = pr_B(s_j)$, $1 \le i \le m_1$ and $1 \le j \le m_2$.

Due to the approximate matching problems in random databases, the equi-join operation [2,4] was replaced by the approximate join, denoted by $T \bowtie_{A \approx B} S$. In order to define this operation, we have to consider a distance d between the elements in D_A and 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. Suppose that D_A and D_B are subsets of a metric space on which the distance d is defined. In the case of numeric attributes, the Euclidean distance can be considered as d.

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

The result of the approximate join operation is composed of the ε-close tuples according to the given distance. For the particular case $\varepsilon = 0$, we obtain the equi-join operation. In what follows, we consider A and B fixed and we will omit them in the approximate (ε-) join's notation.

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

$$T \bowtie_\varepsilon S = \{(x,y) \in T \times S | d(x_A, y_B) \le \varepsilon\}$$

The random variable $N_\varepsilon = |T \bowtie_\varepsilon S|$, specifying the cardinality of the result set of the ε-join defined above, has been a debated subject of research [2,7,9], mainly in terms of probability distribution's estimation.

2.2 Probability distribution of the ε-join's cardinalities

At the beginning of the research concerning the cardinality N_ε, the random tables consisted of records which followed the same (possibly multidimensional) probability distribution [2]. Further, we considered the concept of heterogeneous random table, in which different subsets of columns can follow different probability distributions. In this context, in [4] we proposed two methods of estimation.

Firstly, we generated the histograms for the cardinalities of the join result between two random tables. The shape of the histograms suggested that the values N_ε follow a Poisson distribution, thus we applied the χ^2 test [10] in order to determine that more accurately. Besides, it could be noticed that there was a threshold of ε up to which the Poisson distribution was followed by N_ε.

Secondly, we proved in a sounder manner that the number of records in the result set obtained in an ε-operation on random tables follows a Poisson distribution. The result actually extended the main result in [2]. The proof was obtained through a Poisson approximation using the Stein-Chen method [11], which approximates a probability distribution $\mathscr{P}$ by a simpler distribution $\mathscr{Q}$, easier to define and to use in simulations. Also, the proof of the Poisson estimation of the cardinalities distribution uses concepts as entropy [12] and coincidence probabilities [13]. The difference between the actual probability and the Poisson one was measured by the total variation distance [2,14].

Following the previous research for the homogeneous case and the approaches described above for the heterogeneous one, we can state that the values N_ε are

Poisson distributed of parameter λ, where $\lambda = \mathbf{E}(N_\varepsilon)$, the mean value of N_ε.

3 A multidimensional Poisson process approach to the ε-join in random databases

In what follows, we study the cardinalities of an ε-join operation's result set through the perspective of a homogeneous bidimensional Poisson process. The result that we present will be further generalized to the multidimensional case.

3.1 The ε-join of two tables from the perspective of a bidimensional Poisson process

In our setting, we consider the random tables T, S, $|T| = m_1$, $|S| = m_2$, whose attributes A and B, respectively, are supposed to link the two tables in an ε-join condition. These attributes have compatible domains of values D_A and D_B, respectively. As these attributes should have a similar meaning, we suppose that their values follow the same type of unidimensional probability distribution on the domains D_A, D_B. This probability distribution has the same parameters for both attributes.

Without loss of generality, we consider that the domains D_A and D_B are the intervals $[0,K]$, respectively $[0,L]$, where $K > 0$, $L > 0$. In this case, the result of the ε-join operation can be represented by points in the rectangle $\mathscr{D} = [0,K] \times [0,L]$.

As it can be noticed, in this setting the sets of join attributes have a single element. Further, we introduce the bidimensional Poisson process of intensity λ and we show that the number of records in the ε-join operation's result follow a process of this type.

Definition 3. [15] A process which consists of random points in the bidimensional plane is a bidimensional Poisson process of intensity λ if the following conditions are satisfied:

1. The number of points in any region of area Γ is Poisson distributed of parameter $\lambda\Gamma$.
2. The numbers of points in disjoint regions correspond to independent random variables.

We will consider that the points in the rectangle $\mathscr{D}$ are uniformly distributed. From the results we mentioned in section 2.2, the number of points $N_\varepsilon = N_{\mathscr{D},\varepsilon}$ in the rectangle $\mathscr{D}$ is Poisson(λ) distributed, with the parameter λ specified before as the mean value of N_ε. Let Δ be the area of the rectangle $\mathscr{D}$. From the uniformity hypothesis mentioned above, we can consider that the number of points in a rectangle of area Γ, included in the rectangle $\mathscr{D}$, is Poisson distributed of parameter $\lambda \cdot \frac{\Gamma}{\Delta}$.

Denote:

$$\lambda' = \frac{\lambda}{\Delta}. \tag{1}$$

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

Proposition 1. The cardinality of an ε-join operation between the tables T and S, based on the attributes A and B, respectively, with A and B following the same probability distribution, forms a homogeneous bidimensional Poisson process of parameter λ' given in Eq. 1.

A consequence of the proposition 1 is the statement of a relation between m_2, ε and Δ. Suppose that the table S is decomposed into m_2 tables S_j, $1 \le j \le m_2$, each having a single record s_j and $pr_B(s_j) = B_j$. Thus, the join operation is decomposed into m_2 join operations between the tables S_j, $1 \le j \le m_2$, and the table T.

Denote by N_ε^j the number of records obtained in the join between S_j and T, $1 \le j \le m_2$. From the previous considerations, we know that N_ε^j is a Poisson random variable of parameter λ_j. We will suppose that the sets of result records obtained in the m_2 join operations are disjoint. Then, the following relation holds:

$$\sum_{j=1}^{m_2} N_\varepsilon^j = N_\varepsilon \tag{2}$$

From the property of the sum of Poisson variables [16], it is known that:

$$\sum_{j=1}^{m_2} N_\varepsilon^j \sim \text{Poisson}\left(\sum_{j=1}^{m_2} \lambda_j\right) \tag{3}$$

From the Eq. 2, 3 and because N_ε is Poisson distributed of parameter λ, it results that:

$$\sum_{j=1}^{m_2} \lambda_j = \lambda \tag{4}$$

Denote $\mathscr{B}_\varepsilon(B_j) = \{x \in D_B | d(x,B_j) \le \varepsilon\}$ and consider $mes(\mathscr{B}_\varepsilon(B_j))$ the measure of $\mathscr{B}_\varepsilon(B_j)$. Then, from proposition 1, we obtain that:

$$\lambda_i = \lambda' \cdot mes(\mathscr{B}_\varepsilon(B_j)) \tag{5}$$

From the Eq. 4 and 5, we obtain the following result:

Proposition 2. Consider the ε-join operation between the attributes A and B of the random tables T and S, respectively, and B_i, $1 \le i \le m_2$, the values of attribute B. Then:

$$\sum_{i=1}^{m_2} mes(\mathscr{B}_\varepsilon(B_i)) = \frac{\lambda}{\lambda'} \tag{6}$$

Because $mes(\mathscr{B}_\varepsilon(B_j))$ depends on ε and $\frac{\lambda}{\lambda'} = \Delta$, proposition 2 provides the connection between m_2, ε and Δ. If we take into consideration the standard Lebesgue measure, then:

$$\text{mes}(\mathscr{B}_\varepsilon(B_i)) = 4\varepsilon^2 \tag{7}$$

so the Eq. 6 becomes:

$$4m_2\varepsilon^2 = \Delta \tag{8}$$

We recall that in the previous research [2,4] it has been seen that the Poisson probability distribution is followed up to a threshold of ε, which was not determined. Using the Poisson perspective of the ε-join, this value may be determined. The relation stated in proposition 2 allows to obtain a value for ε which ensures that N_ε is Poisson distributed.

Taking into consideration the result from proposition 1, we can use methods of simulation of the cardinalities of operations, based on the methods of simulation of the bidimensional Poisson processes, which are presented in section 4.

3.2 Generalization to the ε-join of n tables

In what follows, the results in section 3.1 are extended to the case of a multiple join. Consider the random tables $T_1, T_2, \ldots, T_n$, $n \geq 2$, and the corresponding attributes $A_1, A_2, \ldots, A_n$ which participate in the ε-join operation. As in the bidimensional case, suppose that the domain of each attribute A_j, $1 \leq j \leq n$, is $[0, K_j]$, $K_j > 0$.

We consider the n-dimensional cube $\mathscr{C} = [0,K_1] \times [0,K_2] \times \ldots \times [0,K_n]$, $K_j > 0$ for each $j = \overline{1,n}$. Similar to the reasoning in the bidimensional case, the points in this cube are uniformly distributed and $N_\varepsilon = N_{\mathscr{C},\varepsilon} \sim \text{Poisson}(\lambda)$, with $\lambda = \mathbf{E}(N_\varepsilon)$.

Proceeding in an inductive manner, we find that the parameter λ' of the Poisson process is

$$\lambda' = \frac{\lambda}{\text{vol}(\mathscr{C})}. \tag{9}$$

Denote by

$$T' = T_1 \bowtie_\varepsilon T_2 \bowtie_\varepsilon \ldots \bowtie_\varepsilon T_{n-1} \tag{10}$$

and consider $\mathscr{C}'$ the corresponding $(n-1)$-dimensional cube. As the join operation is closed and associative, it implies that

$$T_1 \bowtie_\varepsilon T_2 \bowtie_\varepsilon \ldots \bowtie_\varepsilon T_n = T' \bowtie_\varepsilon T_n \tag{11}$$

In the inductive step of the proof of Eq. 9, we suppose that the parameter λ' of the Poisson process corresponding to the ε-join which produces the table T' is

$$\lambda' = \frac{\lambda}{\text{vol}(\mathscr{C}')} \tag{12}$$

From the right-side member of Eq. 11 we obtain that:

$$\lambda' = \frac{\lambda}{\text{vol}(\mathscr{C}')} \cdot \frac{1}{K_n} = \frac{\lambda}{\text{vol}(\mathscr{C})} \tag{13}$$

The above considerations justify the following result:

Proposition 3. The cardinality of an ε-join operation between the tables $T_1, T_2, \ldots, T_n$, based on the attributes $A_1, A_2, \ldots, A_n$, $n > 2$, where A_j, $j = \overline{1,n}$ follow the same probability distribution, forms a homogeneous multidimensional Poisson process of parameter λ' given in Eq. 9.

Consider that the cardinality of the table T_n is m_n, $m_n > 0$, and the values of the join attribute A_n are A_n^j, $j = \overline{1,m_n}$. Generalizing proposition 2 above, we state the following:

Proposition 4. Consider the ε-join operation between the attributes A' and A_n of the random tables T', respectively T_n, and A_n^j, $1 \leq j \leq m_n$, the values of attribute A_n. Then:

$$\sum_{j=1}^{m_n} mes\left(\mathscr{B}_\varepsilon\left(A_n^j\right)\right) = \text{vol}(\mathscr{C}). \tag{14}$$

Similarly to the result stated in the bidimensional case, $mes\left(\mathscr{B}_\varepsilon\left(A_n^j\right)\right)$ depends on ε and $\frac{\lambda}{\lambda'} = \text{vol}(\mathscr{C})$; thus, proposition 4 provides the connection between m_n, ε and $\text{vol}(\mathscr{C})$. Again, for the standard Lebesgue measure, this relation is described by the formula:

$$2^n m_n \varepsilon^n = \text{vol}(\mathscr{C}) \tag{15}$$

The relation in Eq. 15 allows to determine the threshold value of ε for which the probability distribution Poisson is followed in the case of the multiple ε-join between n tables.

4 Simulation of bidimensional and multidimensional Poisson stochastic processes

The distances between the random points of the unidimensional Poisson process of parameter λt, $t \in [0, \infty)$, are distributed exponentially of parameter λ [17]. As a consequence, the simulation of a p points trajectory of the homogeneous unidimensional Poisson process of intensity λ [5] can be done with the algorithm 1. This algorithm (1) outputs the sequence $T_1, T_2, \ldots, T_p$ which is a trajectory of the process *Poisson*(λ) on the interval $[0, \infty)$. The iterative step of the algorithm uses the standard exponential variable that can be simulated through usual simulation methods, such as the inverse or the rejection method [18].

In the bidimensional case, the simulation of the Poisson process with the intensity λ on the rectangle $\mathscr{R} = [0,t] \times [0,1]$ can be realized through the algorithm 2 [5], which actually extends the previous (unidimensional) method.

In the algorithm 2, the integer p is a selection value which results from the simulation of the random variable Poisson distributed of parameter λt.

In section 3.2 we generalized the approximate join operation to n tables. In this case, the simulation of a

Algorithm 1 Simulate a Poisson process

Require: λ, p
Ensure: the Poisson trajectory $T_1, T_2, \ldots, T_p$
1: $T_0 \Leftarrow 0$
2: **for** $i = 1$ **to** p **do**
3: Generate $E \sim \mathrm{Exp}(1)$
4: $T_i \Leftarrow T_{i-1} + \frac{E}{\lambda}$
5: **end for**

Algorithm 2 Simulate a bidimensional uniform Poisson process

Require: λ
Ensure: the 2-dimensional Poisson trajectory $(U_1, T_1), (U_2, T_2), \ldots, (U_p, T_p)$
1: Generate $T_1, T_2, \ldots, T_p$ a Poisson trajectory on $[0, t]$
2: **for** $i = 1$ **to** p **do**
3: Generate $U_i \sim \mathscr{U}([0, 1])$
4: **end for**

multidimensional Poisson process is needed. The corresponding algorithm derives further from the previous algorithm 2; thus, one obtains an algorithm for simulating a uniform Poisson process of intensity λ on the n-dimensional cube $\mathscr{C} = [0, K_1] \times [0, K_2] \times \ldots [0, K_n]$.

Denote

$$V = \prod_{i=2}^{n} K_i \tag{16}$$

the volume of the $(n-1)$-dimensional cube $\mathscr{C}' = [0, T_2] \times \ldots [0, T_n]$. The algorithm 3 provides a method for the simulation of this type of Poisson process. The vectors obtained as output of the algorithm 3 are an implementation of the Poisson process of intensity λ on the n-dimensional cube $\mathscr{C}$. The points $Q_i = (X_i, U_i)$, $1 \leq i \leq p$, determine a uniform Poisson process of parameter λ on $\mathscr{C}$.

Algorithm 3 Simulate a multidimensional Poisson process

Require: λ
Ensure: the n-dimensional Poisson trajectory $(X_1, U_1), (X_2, U_2), \ldots, (X_p, U_p)$
1: $t \Leftarrow 0; p \Leftarrow 0$
2: **repeat**
3: Generate $E \sim \mathrm{Exp}(1)$
4: $p \Leftarrow p + 1; t \Leftarrow t + \frac{E}{\lambda}$
5: $X_p \Leftarrow \frac{t}{V}$
6: **until** $X_p \geq T_1$
7: **for** $i = 1$ **to** k **do**
8: Generate $U_i \sim \mathscr{U}(I')$
9: **end for**

5 Conclusion

This article introduced an approach to the problem of the approximate join operation in random databases, from the perspective of the Poisson stochastic processes. This problem was previously studied from the point of view of the estimation of the probability distribution of the result sets' cardinalities.

There are two main benefits which derive from the advances presented in this article. The first contribution consists of the fact that this approach provides a method for the simulation of the cardinalities of the results of the approximate join operations. Also, this article proposes a way to compute the threshold of the values of ε up to which the Poisson distribution of the cardinalities values, proved in the previous researches in the domain, is followed.

Acknowledgement

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.

References

[1] D. Zhang and X. Zhang, A New Service-Aware Computing Approach for Mobile Application with Uncertainty, Appl. Math. Inf. Sci. **6**, 9-21 (2012).

[2] O. Seleznjev and B. Thalheim, Random Databases with Approximate Record Matching, Methodology and Computing in Applied Probability **12**, Springer, 63-89 (2008).

[3] S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases, Addison-Wesley (1995).

[4] 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, 407-413 (2009).

[5] F. Suter, I. Vaduva, 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 **XV**, 23-35 (2006).

[6] M. Kifer, A. Bernstein, P. Lewis, Database Systems. An Application-Oriented Approach, Addison Wesley (2005).

[7] O. Seleznjev and B. Thalheim, Average Case Analysis in Database Problems, Methodology and Computing in Applied Probability **5**, Springer-Verlag (2003).

[8] B. Thalheim, Konzepte des Datenbank-Entwurfs, Entwicklungsten denzen bei Datenbanksystemen, 1-48 (1991).

[9] G.O.H. Katona, Random Databases with Correlated Data, Conceptual Modelling and Its Theoretical Foundations, Lecture Notes in Computer Science **7260**, Springer-Verlag, 29-35 (2012).
[10] Gh. Mihoc and V. Craiu, Tratat de statistica matematica, vol. 2, Academy of Romania Publishing House, Bucharest (1977).
[11] A.D. Barbour, L. Holst, S. Janson, Poisson approximation, Clarendon, Oxford (1992).
[12] O. Onicescu and V. Stefanescu, Elements of Informational Statistics with Applications (in Romanian), Technical Publishing House, Bucharest (1979).
[13] S. Karlin and J. Gregor, Coincidence probabilities, Pacific Journal of Matematics **9**, 1141-1164 (1959).
[14] T. Steerneman, On the Total Variation and Hellinger Distance between Signed Measures; an Application to Product Measures, Proceedings of the American Mathematical Society **88**, 684-688 (1983).
[15] S. Ross, Simulation, Academic Press, San Diego, London (1997).
[16] N.L. Johnson, S. Kotz, A.W. Kemp, Univariate Discrete Distributions, John Wiley & Sons (2005).
[17] S. Ross, Stochastic Processes, John Wiley & Sons (1995).
[18] I. Vaduva, Simulation Models (in Romanian), University of Bucharest Publishing House, Bucharest (2005).

Letiţia Velcescu is a Lecturer in the Department of Computer Science of the Faculty of Mathematics and Computer Science of the University of Bucharest. She received the PhD degree in Mathematics while pursuing an academic career in the same institution. Her main research interests cover database theory, algorithms and data structures, modeling and simulation, theory of probabilities and statistics. She published research papers in international journals and she participated at scientific conferences.