DOI: 10.2478/auom-2013-0060
An. Şt. Univ. Ovidius Constanţa

VERSITA
Vol. 21(3),2013, 295–306

Distribution of Time Interval between the Modifications of Result Sets Cardinalities in Random Databases

Letiţia Velcescu

Abstract

In this paper, we propose a method to estimate the probability distribution of the time interval which ellapses between the modifications of the cardinality in a random database query's result set. This type of database is important either in modeling uncertainty or storing data whose values follow a probability distribution. The result that we introduce is important from the point of view of the database optimization, providing a useful method for an integrated module. In previous research on random databases the sizes of some relational operations results were investigated. This kind of information is rather useful in an analytical database which provides decision-making support. The result we particularly aim to present in this paper concerns the transactional random databases, addressing its specific functionality. It will be proven that the interval of time between the cardinalities changes is exponentially distributed. The proof is based on the technique of the Markovian Jelinski-Moranda model, which is used in the reliability of software programs.

1 Introduction

The huge amount of data which is accumulating in each domain of research or activity leads to different types of problems, from the storage in the corresponding data structures to the extraction of information, knowledge discovery and decision making. The different solutions that respond to the mentioned problems determined the

Key Words: Random database, Approximate query, Transactional processing, Jelinski-Moranda model.
2010 Mathematics Subject Classification: Primary 68P15; Secondary 68U20.
Received: November, 2012.
Revised: December, 2012.
Accepted: April, 2013.

apparition of two main technologies of data management, namely the on-line analytical processing (OLAP) and the on-line transactional processing (OLTP). On one hand, the OLAP databases are rather specific to data warehouses, in which information usually is not updated, but analysed in order to support decisions. On the other hand, the OLTP databases are suitable for the management of current data in the respective field.

Data that is stored in databases might originate from different sources and sometimes it is uncertain or even erronated. Even in this case, database queries have to support decisions by providing the answers that at least approximate the real information. In research, the domain of databases that store this type of uncertain or error carrying information is related closely to the random database field ([13], [14]). The importance of this type of databases is relevant in the domains who need to manage this kind of information. In this context, we can mention the fields that work with data provided by sensors.

The results introduced in this article extend the author's research work in the domain of random databases. In the previous work, the concept of heterogeneous random database was defined ([20]). In this particular type of database, where the values of the columns follow different probability distributions, relational operations were studied. Generally, random databases store data that are likely to be uncertain. Thus, the relational operations performed on these data have to be redefined in order to support approximate searches. Attention was focused on the transformation of operations from relational algebra in the context of random databases. The number of records in the queries' result sets was estimated, in terms of the probability distribution of these values ([14]). It was shown that this cardinality is Poisson distributed, but the approximation to this distribution depends on the radius of the ball in which values are accepted in the approximate queries. This estimation is important in the optimization of the queries performed in an analytical database.

Further, the approximate join operation has been approached as a multidimensional Poisson stochastic process ([21]). This result allows the algorithms that simulate this type of process to intervene in the simulation of the values which represent the cardinalities of the sets resulted from the approximate join.

This article brings as novelty the treatment of the transactional aspect of a random database, in the same perspective of queries optimization. We study the case of a database in which insertions are likely to be performed frequently. Estimating the time interval between two changes in the result set of a query can provide an important factor that might improve the database optimization process. By the means of the technique of justification the hypothesis of the Jelinski-Moranda model ([18]), we prove that this number is exponentially distributed.

In the second section of this paper we introduce the main concepts of random database theory and present a survey of the research performed so far. The third part presents our approach to the estimation of the time interval between two changes in

a query's result set. The article ends with a conclusions section that emphasizes the contributions of the presented research.

2 Relational Random Databases

In this section, we introduce the main concepts in random databases and the most important results obtained so far in this domain, namely the Poisson estimation of the probability distribution of the approximate join's cardinalities and the perspective of the approximate join as a Poisson stochastic process.

In database theory, a database is a set of relations. Each relation of the database is described by its corresponding relation schema ([1]), which usually contains the attributes names, their domains and the primary key. The implementation of a relation is a table.

Consider a relation R in a random database and its relation schema $R(U)$, where $U = \{A_1, A_2, \ldots, A_n\}$ is the set of the attributes in the relation R. The implementation of the relation R is a table denoted by $T(R)$. The number of attributes n defines the arity of a tuple in the relation R ([6]); the number of tuples (rows) in the table is referred as the table's cardinality (e.g., [17]).

The values of each attribute A_i, for $i \in \{1, 2, \ldots, n\}$, belong to an associated domain of values D_{A_i}. Consequently, the tuples' values will belong to the cross product $D_U = D_{A_1} \times D_{A_2} \times \ldots \times D_{A_n}$. The projection corresponding to the i-th tuple t_i of $T(R)$ on a subset of attributes $A \subseteq U$ is denoted by $pr_A(t_i)$, for $i \in \{1, 2, \ldots, |T(R)|\}$. Consider the relations $R_1(U_1)$ and $R_2(U_2)$ implemented by the corresponding tables $T = \{t_i | 1 \leq i \leq m_1\}$ and $S = \{s_i | 1 \leq i \leq m_2\}$, respectively ([6]).

The most important and frequent relational operation when dealing with databases is the equi-join. At the same time, this operation is costly, so its analysis constitutes an important subject in queries optimization ([14]). The equi-join between the tables T and S based on the attributes sets A and B, respectively, where $A \subseteq U_1$, $B \subseteq U_2$ and $|A| = |B|$, is denoted by $T \bowtie S$. This operation's result is a table whose records satisfy the equalities between the corresponding values of the join attributes, i.e. the resulting table contains combined tuples t_i, s_j from T and S, respectively, such that $pr_A(t_i) = pr_B(s_j)$, $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$.

Because of the uncertainty of information in random databases, the equi-join operation ([14], [20]) was replaced by the approximate join, denoted by $T \bowtie_{A \approx B} S$. This operation was defined considering a distance d between the elements in subsets 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, respectively. We also consider that D_A and D_B are subsets of a metric space on which the distance d is defined. For instance, if the join attributes are numeric, then d might be the Euclidean distance.

The values $x \in D_A$ and $y \in D_B$ are ϵ-close ([14]), $\epsilon \geq 0$, if $d(x, y) \leq \epsilon$.

The table resulting from the approximate join operation contains the ϵ-close tuples according to the given distance. From now on, we consider that the join attributes A and B are fixed and we will omit them in the ϵ-join's notation. In the particular case $\epsilon = 0$, the equi-join operation is obtained.

Definition 1. The ϵ-join operation between two random tables T and S is defined by the following records set:

$$T \bowtie_\epsilon S = \{(x, y) \in T \times S | d(x_A, y_B) \leq \epsilon\} \tag{1}$$

The random variable $N_\epsilon = |T \bowtie_\epsilon S|$, that denotes the cardinality of the result set of the ϵ-join defined above, has been studied in the previous research ([14], [13], [5], [20]) from the perspective of its probability distribution. In the next subsections, we introduce briefly the main approaches and results concerning this problem.

2.1 Estimation of the probability distribution of the ϵ-join's cardinalities

In the first researches regarding the probability distribution of the cardinalities N_ϵ, the random tables consisted of records which followed the same multidimensional probability distribution. In this framework, it was proved that the random variable N_ϵ is Poisson distributed ([14]). Then, we defined the concept of heterogeneous random table in which different subsets of columns can follow different probability distributions. Two methods of estimation have been proposed for this type of random table in ([20]).

Initially, we generated the histograms for the cardinalities of the approximate join result between two random tables. The analysis of these histograms indicated that the values N_ϵ are Poisson distributed. In order to validate this assumption, we applied the χ^2 test of goodness of fit ([8]). Following this approach, the conclusion that we reached was that the cardinalities are Poisson distributed. Nevertheless, the existence of a threshold of ϵ up to which the Poisson distribution was followed by N_ϵ could be noticed.

Further, 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 [14]. The proof was obtained through a Poisson approximation using the Stein-Chen method ([2]), which approximates a probability distribution $\mathcal{P}$ by a simpler distribution $\mathcal{Q}$, easier to define and to use in simulations. Also, the proof of the Poisson estimation of the cardinalities distribution uses concepts as entropy ([10]) and coincidence probabilities ([4]). The difference between the actual probability and the Poisson one was measured by the total variation distance ([14], [15]).

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 $\lambda = \mathbf{E}(N_\epsilon)$ is the mean value of N_ϵ.

2.2 Simulation of the ϵ-join in random databases using Poisson stochastic processes

In this section, we present an overview of the ϵ-join problem using the approach of a homogeneous bidimensional Poisson process, which will be further generalized to the multidimensional case.

2.2.1 The bidimensional case

In the research introduced in [21], we considered the random tables T, S, with the ϵ-join attributes A and B, respectively. The domains of these attributes are D_A and D_B, respectively, and they are supposed to be compatible. Since these attributes should have a similar meaning in the relation schemas, we also supposed that their values follow the same type of unidimensional probability distribution on the domains D_A, D_B and this probability distribution has the same parameters for both attributes.

In our framework, we considered that the sets of join attributes have a single element and the domains D_A and D_B are the intervals $[0, K]$ and $[0, L]$, respectively, where $K > 0$, $L > 0$. In this case, the result of the ϵ-join operation can be represented by points in the rectangle $\mathcal{D} = [0, K] \times [0, L]$.

Definition 2. ([11]) 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 distributed Poisson of parameter $\lambda\Gamma$.
2. The numbers of points in disjoint regions correspond to independent random variables.

From the results we mentioned in section 2.1, the number of points N_ϵ in the rectangle $\mathcal{D}$ 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 $\mathcal{D}$.

Denote:

$$\lambda' = \frac{\lambda}{\Delta}. \tag{2}$$

It was proven ([21]) that the number of records in the ϵ-join operation's result follow a bidimensional Poisson process; we could state the following:

Proposition 1. The cardinality of a ϵ-join operation between the tables T, respectively S, based on the attributes A and B, with A and B following the same probability distribution, forms a homogeneous bidimensional Poisson process of parameter λ' given in Eq. 2.

Taking into consideration the result from proposition 1, the methods of simulation of the bidimensional Poisson processes can be used ([16]). A consequence of the proposition 1 is the statement of a relation between $|S|$, ϵ and Δ.

For each value B_j of the attribute B, denote $\mathcal{B}_\epsilon\left(B_j\right) = \left\{x \in D_B | d(x, B_j) \leq \epsilon\right\}$ and consider $mes\left(\mathcal{B}_\epsilon\left(B_j\right)\right)$ the measure of $\mathcal{B}_\epsilon\left(B_j\right)$. Then, the following result could be obtained:

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

$$\sum_{i=1}^{|S|} mes\left(\mathcal{B}_\epsilon\left(B_i\right)\right) = \frac{\lambda}{\lambda'} \tag{3}$$

In the relation obtained in proposition 2, the standard Lebesgue measure can be considered.

In the research concerning the results presented in section 2.1 ([14], [20]) it could be noticed that the Poisson probability distribution is followed up to a threshold of ϵ. The Poisson perspective of the ϵ-join allows to determine this value, using the relation stated in proposition 2.

2.2.2 Generalization to the n-dimensional case

The results presented in section 2.2.1 could be 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 ϵ-join attributes $A_1, A_2, \ldots, A_n$. Suppose that the domain of each attribute A_j, $1 \leq j \leq n$, is $\left[0, K_j\right]$, $K_j > 0$.

We consider the n-dimensional cube $\mathcal{C} = [0, K_1] \times [0, K_2] \times \ldots \times [0, K_n]$, $K_j > 0$ for each $j \in \{1, 2, \ldots, n\}$. Similar to the bidimensional case, N_ϵ is Poisson (λ) distributed, where $\lambda = \mathbf{E}\left(N_\epsilon\right)$.

The value of the parameter λ' of the Poisson process can be found by induction:

$$\lambda' = \frac{\lambda}{\text{vol}\left(\mathcal{C}\right)}. \tag{4}$$

Further, the proposition 1 in the previous subsection was generalized ([21]) as:

Proposition 3. The cardinality of a 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, $j \in \{1, 2, \ldots, n\}$ follow the same probability distribution, forms a homogeneous multidimensional Poisson process of parameter λ' given in Eq. 4.

Also, a result that allows to find the threshold of ϵ could be extended in this case, following the reasoning in Proposition 2.

3 Estimation of the time interval between cardinalities changes

In time, the contents of transational databases is submitted to frequent modifications. We will consider the problem of insertions in the transactional random databases, in

the context of the estimation of time interval between the modifications of the join operation's result set cardinality. The technique that we will use is the one of the justification of some hypothesis in the Jelinski-Moranda model ([18]). We will transpose these justifications in the random databases model, when the temporal aspect of database modification and queries launched on them is taken into consideration.

3.1 Preliminaries

The Jelinski-Moranda model ([9]) is a Markovian model for software program's reliability. This model supposes that the number of errors existing at a given moment is a Markov process. A software program is considered as a reparable system, in which the errors occurence leads to the interruption of the execution and program debug, in order to eliminate the errors. In the program, there is an initial number of errors that decreases while the program is debugged. The Jelinski-Moranda model starts from some hypothesis that can be justified based on the particularities of a software product.

The main hypothesis of the model are (e.g., [7]):

1. The number of initial errors in the system is finite and fixed.
2. The errors are of the same type.
3. The repair of errors is done immediately and perfectly.
4. The detection of error is done independently of each other.
5. The intervals of time between errors is exponentially distributed; the parameter of the distribution is proportional to the number of remaining errors.
6. The hazard rate remains constant over the interval between error occurrences.

Generally, the Jelinski-Moranda model is characterized in terms of the probability distribution of the intervals of time between the appearance of errors. We describe briefly the main points of this characterization ([18], [7]).

Consider n random variables $X_1, X_2, \ldots, X_n$ independant and identically distributed, whose distribution is exponential of parameter λ, $\lambda > 0$. The density function of the variable X_i, $i \in \{1, 2, \ldots, n\}$ is:

$$f_i(x_i) = \lambda e^{-\lambda x_i}. \tag{5}$$

Denote by $X_{(i)}$ the statistic of order i and suppose that $X_{(0)} = 0$. Then, the joint distribution of the order statistics is:

$$f_{(1,2,\ldots,n)}(x_1, x_2, \ldots, x_n) = n!\lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}. \tag{6}$$

On one hand, the intervals of time between errors are given by $T_i = X_{(i)} - X_{(i-1)}$ for $i \in \{1, 2, \ldots, n\}$. Consequently, it implies that:

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} (n - i + 1)\, T_i. \tag{7}$$

On the other hand, from Eq. 6 it can be noticed that:

$$f_{(1,2,\ldots,n)}(x_1, x_2, \ldots, x_n) = f(t_1, t_2, \ldots, t_n) = n!\lambda^n \cdot e^{-\lambda \sum_{i=1}^{n}(n-i+1)t_i}. \tag{8}$$

Further, it can be noticed that $f(t_1, t_2, \ldots, t_n)$ is the joint density of the variables $T_1, T_2, \ldots, T_n$, and the marginal density function of T_i is:

$$f_i(t_i) = (n - i + 1)\, \lambda \cdot e^{-\lambda t_i (n-i+1)}. \tag{9}$$

From Eq. 8, 9, it results immediately that:

$$f(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} f_i(t_i). \tag{10}$$

Consequently, the time intervals between the errors are independent random variables, exponentially distributed of parameter $(n - i + 1)\, \lambda$, for $i \in \{1, 2, \ldots, n\}$.

Due to this exponential distribution, the hazard rate after the detection of the i-th error is constant, for $i \in \{1, 2, \ldots, n\}$:

$$h_i(t) = (n - i + 1)\, \lambda. \tag{11}$$

To conclude, $(T_n)_{n \geq 0}$ is a Markovian process with independant increments due to the fact that the time intervals are independant.

3.2 Applying the Jelinski-Moranda model in transactional random databases

Suppose that we have two random tables T_1 and T_2 which are frequently submitted to approximate join operations. We consider that in one of the tables frequent insertions of different data volumes take place, followed by the re-evaluation of the ϵ-join operation. In this situation, the following question arises: can we estimate the probability distribution of time which passes between two subsequent modifications of the result set? Similar to the justification of the Jelinski-Moranda model ([18]), we can consider: M the set of records in the cross product, $M^* \subset M$ the set of records in the result set of the join operation, and τ the length of the time interval between two subsequent modifications of the result set.

Generally, we can suppose that the commit operations of the insertions in the database's tables appear rarely, so we can consider that these operations, interpreted as events, form a Poisson(ω), where ω is the intensity of the insertion operations.

The distribution function of the time interval τ between the modifications of the result set is $F(t) = P(\tau < t)$. The probability not to have data that enter in the join operation's result in the interval $[0, t]$ is:

$$\overline{F}(t) = 1 - F(t) = \sum_{j=0}^{\infty} \left(\frac{e^{-\omega t}(\omega t)^j}{j!} \right) \left(\frac{M - M^*}{M} \right)^j \tag{12}$$

The preceding formula has the following justification: the first factor of the sum represents the probability to insert j records in the time interval $[0, t)$ according to the Poisson(ω) distribution, where ω is the intensity of the insert operations, whereas the second factor represents the probability that this data do not enter in the join operation's result. Performing the sum represents a mediation concerning the random values of j.

We denote by

$$\lambda = \frac{M^*}{M} \cdot \omega \tag{13}$$

the intensity of occurence of new records in the join operation. In this case, the formula 12 becomes:

$$\overline{F}(t) = e^{-\omega t} \sum_{j=0}^{\infty} \frac{\omega t^j}{j!} \cdot \left(\frac{M - M^*}{M} \right)^j = e^{-\omega t} \sum_{j=0}^{\infty} \frac{\omega t^j}{j!} \cdot \left(1 - \frac{\lambda}{\omega} \right)^j = \tag{14}$$

$$= e^{-\omega t} \sum_{j=0}^{\infty} \frac{\left[\omega t \cdot \left(1 - \frac{\lambda}{\omega} \right) \right]^j}{j!}$$

Further, we obtain that:

$$\overline{F}(t) = e^{-\omega t} \cdot e^{\omega t \cdot \left(1 - \frac{\lambda}{\omega} \right)} = e^{-\lambda t} \tag{15}$$

As a consequence, we can state that τ is exponentially distributed, of parameter λ.

The preceding considerations ensure the proof of the following result:

Theorem 1. The probability distribution of the time intervals between the insertions that determine the modification of the random databases operations' result sets is exponential of parameter λ, defined in Eq. 13.

4 Conclusions

In this article the attention is paid to the approximate join problem, specific to random databases. Starting from the important classification of databases in two main classes, analytical and transactional, we first presented the results that mainly apply

in the optimization of the random databases which are rather analytical. Then, the attention was focused on the transactional case. The most original point of this paper is the approach of the probability distribution of the time interval between two subsequent modifications of the result set of a query. In this respect, we used the Jelinski-Moranda model. This contribution provides important information to the optimization module integrated in a random database, as it can avoid recomputations of queries result sets.

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

[2] A.D. Barbour, L. Holst, S. Janson, Poisson approximation, Clarendon, Oxford (1992).

[3] N.L. Johnson, S. Kotz, A.W. Kemp, Univariate Discrete Distributions, John Wiley & Sons (2005).

[4] S. Karlin and J. Gregor, Coincidence probabilities, Pacific Journal of Matematics **9**, 1141-1164 (1959).

[5] 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).

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

[7] M.R. Lyu, Handbook of Software Reliability Engineering, McGraw-Hill and IEEE Computer Society (1996).

[8] Gh. Mihoc, V. Craiu, Tratat de statistica matematica, vol. 2, Academy of Romania Publishing House, Bucharest (1977).

[9] Z. Jelinski, P. Moranda, Software reliability research, Statistical Computer Performance Evaluation, Academic Press, 465-497 (1972).

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

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

[12] S. Ross, Stochastic Processes, John Wiley & Sons (1995).

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

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

[15] 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).

[16] 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).

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

[18] I. Vaduva, Program's Reliability (in Romanian), University of Bucharest Publishing House, Bucharest (2003).

[19] I. Vaduva, Simulation Models (in Romanian), University of Bucharest Publishing House, Bucharest (2005).

[20] L. Velcescu, 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).

[21] L. Velcescu, Simulation of Join Cardinalities in Random Databases Using Poisson Stochastic Processes, Applied Mathematics & Information Sciences **7**, No. 4 (to appear), Natural Sciences Publishing Corporation (2013).

[22] D. Zhang, X. Zhang, A New Service-Aware Computing Approach for Mobile Application with Uncertainty, Applied Mathematics & Information Sciences **6**, Natural Sciences Publishing Corporation, 9-21 (2012).

Letiţia VELCESCU,
Department of Informatics,
Faculty of Mathematics and Informatics,
University of Bucharest,
14 Academiei, 010014 Bucharest, Romania.
Email: letitia@fmi.unibuc.ro