Theoretical and Applied Informatics
ISSN 1896–5334
Vol.24 (2012), no. 1
pp. 3–22
DOI: 10.2478/v10179-012-0001-8

Asymptotically error-optimal shape of sampling zone for query selectivity estimation method based on discrete cosine transform

DARIUSZ R. AUGUSTYN

Silesian Technical University, Institute of Informatics,
16 Akademicka St., 44-100 Gliwice, Poland
e-mail:draugustyn@polsl.pl

Received 15 November 2011, Revised 20 December 2011, Accepted 16 January 2012.

Abstract: The problem of query selectivity estimation for database queries is critical for efficient query execution by database management systems. A query execution method strongly depends on early estimated size of a query result. This estimation determines a data access method used later during the query execution. The selectivity parameter is a fraction of table rows that satisfy a single-table query condition. For a selection condition of a range query where an attribute has a continuous domain, the selectivity is equivalent to a definite integral form probability density function (PDF) of attribute values distribution. For a compound selection condition based on many attributes we need a multidimensional space-efficient non-parametric estimator of multivariate PDF of attribute values distribution. A known approach based on Discrete Cosine Transform (DCT) spectrum as an representation of multidimensional PDF is considered. The energy compaction property of DCT lets omit a region of spectrum coefficients with small absolute values without significant losing an accuracy of selectivity estimation. An area of relevant spectrum coefficients is called a sampling zone. Results of experiments from previous works shows that applying the reciprocal shape of the sampling zone gives the least selectivity estimation error subject to a predetermined size of the zone. The main result of this work is a theoretical confirmation of only experimental results from previous works. The paper presents the proof of the theorem that the reciprocal shape of the sampling zone is asymptotically error-optimal. The proof is based on calculus of variations and the isoperimetric problem.

Keywords: query selectivity estimation; probability density function; discrete cosine transform; calculus of variations; isoperimetric problem

1. Introduction

The effectiveness of a database query execution is one of the main goals of any database management system (DBMS). Query processing consists of two phases: a prepare phase and an execution one. During the prepare phase a cost-based query optimizer

module (CBO) chooses the best query execution method. CBO obtains various methods of query execution (with different methods of data access) so-called access paths. Using some cost function CBO estimates a cost of a query evaluating mainly in terms of number I/O operations for every access path. CBO chooses the best access path with the least cost value.

Choice of the method depends on an expected size of data satisfying the query condition. This size should be estimated before the query execution. For this reason – the early estimation of query result set size - the selectivity factor parameter was introduced in query processing. The selectivity values for given queries can be obtained using statistical data describing table attribute values distributions stored and maintained in a database system catalog.

The selectivity for a simple single-table query could be defined as a number of table rows satisfying the query condition divided by a number of all table rows. Selectivity values belong to an interval [0, 1]. The selectivity can be also considered as a probability of drawing a sample row satisfying the selection condition from set of all table rows.

The selectivity for a range query $Q_1\,(a_1 < X < b_1)$ (so-called a window-query) with a simple selection condition based on a one attribute X (where X is a table column with continuous domain) and a given probability density function f_x can be obtained as follows:

$$sel_1(Q_1) = \int_{a_1}^{b_1} f_x\,(x)\,dx. \tag{1}$$

A nonparametric estimator of probability density function (PDF) is required for accurate selectivity calculation. Histograms is commonly used as estimators of PDF [7]. Most of DBMS e.g. Oracle, MS SQLServer, IBM DB2, Sybase Adaptive Server, PostgreSQL support histogram-based selectivity calculation methods. Methods using equi-width (equi-depth) histograms are usually applied.

The problem of selectivity estimation becomes more difficult for complex query selection conditions based on several table attributes. For example the selectivity of some 2-dimensional query $Q_2\,(a_1 < X < b_1 \wedge a_2 < Y < b_2)$ based on X and Y table attributes can be obtained as follows:

$$sel_2\,(Q_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_{xy}(x,y)dxdy \tag{2}$$

where f_{xy} is a PDF of joint distribution of values of X and Y. As we can see in (2) an estimator of bivariate PDF is needed.

This can be extended for more than 2 dimensions when query selection condition is based on many attributes. For continuous attribute domains the selectivity value of a range query is a value of definite integral of multivariate PDF. Hence the problem of space-efficient nonparametric estimator of multivariate PDF for high dimensions. Mul-

tidimensional histograms are commonly too much space-consuming representation of joint-distribution of attribute values.

Many approaches to the multidimensional distribution representation problem have been known since years. The most simple one with no-storing any multidimensional estimator, is based on AVI rule (an attribute value independence assumption [11]). According to the AVI assumption a selectivity for a composite condition is a product of simple component condition selectivities. This method bases on the probability multiplication rule for independent events. The AVI rule usage results in an inaccuracy of obtained values of a query selectivity estimator for correlated data. Despite of this obvious disadvantage the AVI rule is very often used in DBMS optimizers because of its simplicity.

There are many advanced techniques of representing multidimensional distribution suitable for selectivity estimating e.g.: multidimensional kernel estimator [6, 12], bivariate spline [8], PHASED [11], MHIST [11], GENHIST [6], STHoles [1], STHoles+ [4], Bayesian Network [5], Discrete Cosine Transform [9], Cosine Series [13], Discrete Wavelets Transform [3] and many others.

Those theoretical approaches may results ib important practical applications. User-defined specific methods of selectivity estimation may be implemented in some advanced commercial DBMS. For example Oracle DBMS provides ODCIStats module (Oracle Data Cartridge Interface Statistics) [10]. It supports creating domain-specific extensions for the cost-based query optimizer module. This enables to create/update/delete user-defined "statistics" (specific representations of attribute values distributions) and define adequate effective user-defined selectivity estimation functions. This lets to implement in Java and transparently integrate with DBMS any mentioned method of selectivity estimation based on a multidimensional attribute values distribution.

This paper focuses on the theoretical analysis of some aspects of a known unconventional method of selectivity estimation based on Discrete Cosine Transform (DCT) proposed by Lee L., Deok-Hwan K., Chin-Wan Ch. in [9]. Authors proposed a selectivity obtaining method based on DCT spectrum calculated for a histogram of attribute values frequencies. One of the most important advantages of their method is possibility of selectivity calculation directly from a spectrum (without inverse transform calculation). Because of DCT energy compaction property a significant part of spectrum is commonly concentrated in a compact area located near the space-origin of spectrum domain (Fig. 1). DCT-coefficients out of this area, with small absolute values, can be omitted without significant loosing of selectivity estimation accuracy. This area was called sampling zone. Lee at al. proposed some types of sampling zone shapes: spherical, triangular, reciprocal, rectangular (Fig. 2). They showed experimentally that reciprocal zone shape is optimal subject to a predetermined size of the sampling zone i.e. the experimentally obtained mean relative selectivity estimator error for given set of sample attribute value distributions is the least for the reciprocal zone.

The previous work is based on experimental results. This paper presents the rigorous proof that a reciprocal shape of sampling zone is asymptotically error-optimal for predetermined size of the sampling zone. This theorem was proven under some restrictions: 2-dimensional case, a high resolution of spectrum (the asymptotic analysis), the assumed definition of selectivity estimation error (44), the assumed form of spectrum (62). The proof is based on Lagrange multipliers method, calculus of variations and isoperimetric problem (known in optimal control theory). A genuine part of the paper is included in sections 3∼11.

This work is organized as follows: Section 2 is devoted to describe the DCT-based method of selectivity estimation. Section 3 introduces the assumed definition of a selectivity estimation error. Section 4 defines the mean selectivity based on DCT spectrum (for 1-dimensional distribution). Section 5 defines the approximate mean selectivity based on DCT spectrum (for 2-dimensional distribution) using a sampling zone. Section 6 introduces the asymptotic approximate mean selectivity (for large values of DCT spectrum size). Section 7 presents the assumed definition of spectrum (were the energy compaction property has no meanings). In section 8 we define the criterion of error-optimal sampling zone that bases on difference between the asymptotic approximate mean selectivity (for a sampling zone) and the asymptotic mean selectivity (for a full spectrum). Section 9 presents the method of finding a shape of optimal sampling zone using methods of calculus of variations and Lagrange multipliers (the section introduces an adequate Euler-Lagrange equation). Finally in section 10 we obtain a reciprocal function as asymptotically error-optimal shape of sampling zone.

2. Discrete Cosine Transform and selectivity estimation method – the theoretical background

This section describes the energy compaction property of discrete cosine transform. It also introduces the DCT-based method of query selectivity estimation [9] where DCT spectrum represents attribute values distribution. The section presents shapes of sampling zones (regions of relevant spectrum coefficient) used to obtain compressed space-efficient DCT-based representation of distribution.

2.1. Discrete Cosine Transform and Energy Compaction Property

Well-known Discrete Cosine Transform is useful in an image and signal processing (especially in a compression domain).

1-dimensional DCT can be defined as:

$$g\left(u\right)=\sqrt{\frac{2}{N}}k_u\sum_{n=0}^{N-1}f\left(n\right)\cos\left(\frac{(2n+1)u\pi}{2N}\right) \tag{3}$$

and

$$k_u = \begin{cases} \frac{1}{\sqrt{2}} \text{ for } u = 0 \\ 1 \text{ for } u \neq 0 \end{cases} \tag{4}$$

where series $G = (g(u))$ is a DCT spectrum of a signal $F = (f(n))$ for u and n = 0,1,..., N - 1.
In DCT-based selectivity estimation method, F will be a vector of frequencies of attribute values, an estimator of 1-dimensional PDF, a series of values of equi-width histogram.

1-dimensional inverse transform (IDCT) is defined as follows:

$$f(n) = \sqrt{\frac{2}{N}} \sum_{u=0}^{N-1} k_u g(u) \cos\left(\frac{(2n+1)\, u\pi}{2N}\right). \tag{5}$$

2-dimesional DCT can be defined as follows:

$$g(u,v) = \sqrt{\frac{2}{N}} k_u \sum_{n=0}^{N-1} \left\{ \sqrt{\frac{2}{N}} k_v \sum_{m=0}^{M-1} f(n,m) \cos\left(\frac{(2m+1)v\pi}{2M}\right) \right\} \cos\left(\frac{(2n+1)u\pi}{2N}\right) \tag{6}$$

and

$$k_v = \begin{cases} \frac{1}{\sqrt{2}} \text{ for } v = 0 \\ 1 \text{ for } v \neq 0 \end{cases} \tag{7}$$

where $G = (g(u,v))$ is a 2-dimensional DCT spectrum of $F = (f(n,m))$ for u and $n = 0,1, \ldots, N$ - 1 and v and $m = 0,1, \ldots, M-1$. In DCT-based selectivity estimation method, F will be a $N \times M$ matrix of frequencies, an estimator of 2-dimensional PDF.

2-dimensional inverse transform is defined as follows:

$$f(n,m) = \sqrt{\frac{2}{N}} \sum_{u=0}^{N-1} k_u \left[\sqrt{\frac{2}{M}} \sum_{v=0}^{M-1} k_v g(u,v) \cos\left(\frac{(2m+1)v\pi}{2M}\right) \right] \cos\left(\frac{(2n+1)\, u\pi}{2N}\right). \tag{8}$$

The definition of DCT and IDCT can be extended for high dimensions.

For correlated data in F coefficients $g(u,v)$ with significant absolute values are located near space-origin of $U \times V$. This well-known energy compaction property of DCT is illustrated for 2-dimesional case on Fig. 1. An exemplary bivariate PDF of distribution based on 4 Gaussian clusters (Fig. 1.a) and corresponding 100×100 DCT spectrum (Fig. 1.b) are presented. Different grayscales of areas on Fig.1.b show regions grouping spectrum coefficients that absolute values are within intervals: [0, 0.01], (0.01, 0.1], (0.1, 1], (1,10], (10, ∞). The white area on Fig.1.b means the region of coefficients with absolute value near or equal zero.

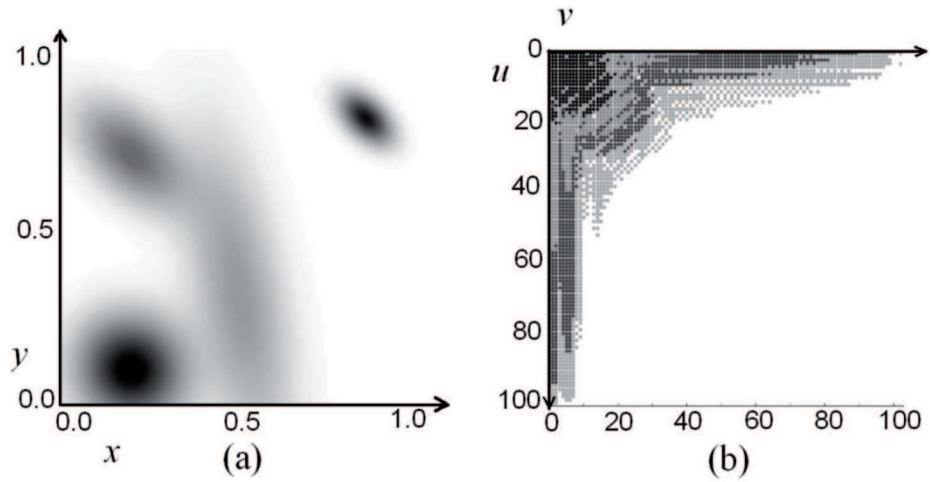

Fig. 1. (a) – bivariate probability density function of an exemplary distribution based on 4 Gaussian clusters. (b) – corresponding DCT spectrum with regions of coefficients with comparable absolute values.

2.2. DCT-based selectivity estimation method

DCT spectrum can be used as representation of distribution of attribute values. The method of selectivity calculation based on DCT spectrum proposed by Lee at al. in [9] will be explained below.

For 1-dimensional case i.e. the simple query $Q_1\,(a < X < b)$, let's assume some normalization that domain of X is [0, 1]. Hence $a_1 \in [0, 1)$ and $b_1 \in (a_1, 1]$. Domain of X is divided into N partition:

$$x_i{=}\frac{2i{+}1}{2N}, i{=}\ 0,\ 1, \ldots, N{-}1. \tag{9}$$

Distribution of X values described by PDF f_x can by expressed using (5) as follows:

$$f_x\,(x_n) = f\,(n){=}\sqrt{\frac{2}{N}}\sum_{u=0}^{N-1} k_u g\,(u)\cos\left(\frac{(2n{+}1)\,u\pi}{2N}\right). \tag{10}$$

Using (9) and (10) the selectivity for Q_1 based on 1-dimensional spectrum can be obtained:

$$sel_1\,(Q_1) = \int_{a_1}^{b_1} f_x\,(x)\,dx{=}\sqrt{\frac{2}{N}}\sum_{u=0}^{N-1}\left[k_u g\,(u)\int_{a_1}^{b_1}\cos\,(u\pi x)\,dx\right]. \tag{11}$$

For 2-dimensional case i.e. the query $Q_2\,(a_1 < X < b_1 \wedge a_2 < Y < b_2)$ let's assume the normalization that domain of $X{\times}\ Y$ is $[0, 1]^2$. Hence a_1 and $a_2 \in [0, 1)$, $b_1 \in (a_1, 1]$

and $b_2 \in (a_2, 1]$. Space $X \times Y$ is divided into $N \times M$ partition by a set of pairs (x_i, y_j):

$$x_i = \frac{2i+1}{2N}, y_j = \frac{2j+1}{2M},\ i = 0,\ 1, \ldots, N-1,\ j = 0,\ 1, \ldots, M-1. \tag{12}$$

Joint distribution of $X \times Y$ values described by bivariate PDF f_{xy} can by expressed using (8) as follows:

$$\begin{gathered} f_{xy}(x_n, y_m) = f(n,m) = \\ = \sqrt{\tfrac{2}{N}} \sum_{u=0}^{N-1} k_u \left[\sqrt{\tfrac{2}{M}} \sum_{v=0}^{M-1} k_v g(u,v) \cos\left(\tfrac{(2m+1)v\pi}{2M}\right) \right] \cos\left(\tfrac{(2n+1)u\pi}{2N}\right). \end{gathered} \tag{13}$$

Using (12) and (13) the selectivity for Q_2 based on 2-dimensional spectrum can be obtained:

$$\begin{gathered} sel_2(Q_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_{xy}(x,y)\, dxdy = \\ = \int_{a_1}^{b_1} \sqrt{\tfrac{2}{N}} \sum_{u=0}^{N-1} k_u \left\{ \int_{a_2}^{b_2} \sqrt{\tfrac{2}{M}} \sum_{v=0}^{M-1} k_v g(u,v) \cos(yv\pi)\, dy \right\} \cos(xu\pi)\, dx, \end{gathered} \tag{14}$$

$$sel_2(Q_2) = \sqrt{\frac{2}{N}} \sqrt{\frac{2}{M}} \sum_{u=0}^{N-1} \sum_{v=0}^{M-1} k_u k_v g(u,v) \int_{a_1}^{b_1} \cos(u\pi x)\, dx \int_{a_2}^{b_2} \cos(v\pi y)\, dy. \tag{15}$$

As we can see in (10) and (15) the selectivity can be calculated directly from spectrum coefficients g (without reconstruction of f by IDCT).

This selectivity calculating method can be extended for more than 2 dimensions [9].

2.3. Shapes of sampling zone

Basing on the definition (15) the selectivity estimator can be formulated as follows:

$$sel_2(Q_2) = \sqrt{\frac{2}{N}} \sqrt{\frac{2}{M}} \sum_{(u,v) \epsilon U \times V} k_u k_v g(u,v) \int_{a_1}^{b_1} \cos(u\pi x)\, dx \int_{a_2}^{b_2} \cos(v\pi y)\, dy. \tag{16}$$

Approximated selectivity estimator based on a sampling zone Z was proposed in [9] as follows:

$$\hat{sel}_2(Q_2, Z) = \sqrt{\frac{2}{N}} \sqrt{\frac{2}{M}} \sum_{(u,v) \epsilon Z} k_u k_v g(u,v) \int_{a_1}^{b_1} \cos(u\pi x)\, dx \int_{a_2}^{b_2} \cos(v\pi y)\, dy \tag{17}$$

where $Z \subset U \times V$. The approximate estimator is calculated using only coefficients from sampling zone Z. A sampling zone contains significant DCT spectrum coefficients (coefficients with big absolute values). (u, v) pairs belonging to a sampling zone should rather concentrate near space origin because of mentioned before DCT energy compaction property (Fig.1.b).

Some types of sampling zone proposed in [9] were presented on Fig. 2. Blue regions are sampling zones and white areas mean regions of omitted coefficients. B parameter determinates a depth of spectrum cutting off.

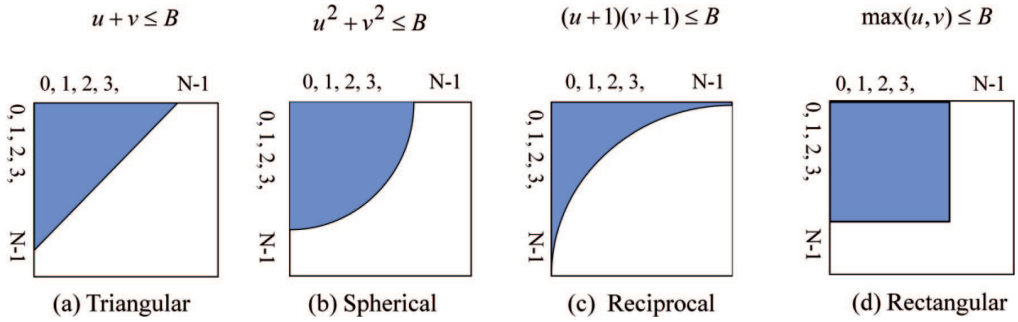

Fig. 2. Geometrical zonal sampling in 2-dimensional$U\times$ V space, proposed by Lee at al. in [9]

Using the definition of relative selectivity estimation error:

$$RelErr(Q,Z)=\frac{\left|sel(Q)-\hat{sel}(Q,Z)\right|}{sel(Q)} \tag{18}$$

basing on set of sample attribute value distributions and set of sample range queries, Lee at al. experimentally showed that for predetermined size of sampling zone the reciprocal sampling zone (Fig. 2.c) is rather error-optimal.

In next sections we will present a theoretical proof of the hypothesis that the reciprocal shape of sampling zone is error-optimal for high resolutions of DCT spectrum.

3. Selectivity estimation error definition

To take into account all possible range query bounds we introduce *msel* – the mean selectivity estimator.

For 1-dimensional case *msel* is defined as follows:

$$msel_1=\frac{\int\int_{(a_1,b_1)\in D} sel_1\left(Q\left(a_1<X<b_1\right)\right)da_1db_1}{\int\int_{(a_1,b_1)\in D} da_1db_1} \tag{19}$$

where $D=\{(a_1,b_1):a_1\in[0,1)\wedge b_1\in(a_1,1]\}$.

This $msel_1$ calculation method based on the assumption that values of query bounds a_1 and b_1 are randomly chosen from intervals defined above according to a uniform distribution.

Using (1) and (19) the 1-dimensional mean selectivity estimator can be defined as follows:

$$msel_1 = \frac{\int_0^1 [\int_{a_1}^1 \left[\int_{a_1}^{b_1} f_x(x)\,dx\right] db_1]da_1}{\int_0^1 [\int_{a_1}^1 db_1]da_1} = \frac{\int_0^1 [\int_{a_1}^1 \left[\int_{a_1}^{b_1} f_x(x)\,dx\right] db_1]da_1}{0.5}. \quad (20)$$

For 2-dimensional case *msel* is defined as follows:

$$msel_2 = \frac{\int_0^1 \left[\int_{a_2}^1 \left[\int_0^1 \left[\int_{a_1}^1 sel_2(Q_2)db_1\right] da_1\right] db_2\right] da_2}{\int\limits_0^1 \int\limits_{a_2}^1 \int\limits_0^1 \int_{a_1}^1 db_1 da_1 db_2 da_2}. \quad (21)$$

Using (2) and (21) the 2-dimensional mean selectivity estimator can be defined as follows:

$$msel_2 = \frac{\int_0^1 \left[\int_{a_2}^1 \left[\int_0^1 \left[\int_{a_1}^1 \left[\int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f_{xy}(x,y)\,dy\right] dx\right] db_1\right] da_1\right] db_2\right] da_2}{0.25}. \quad (22)$$

The definition (22) based on the assumption that query bounds have random uniform distributions and they belong to intervals :$a_1 \in [0,1)$, $b_1 \in (a_1,1]$, $a_2 \in [0,1)$, $b_2 \in (a_2,1]$.

For large values of size in each dimension, *amsel* the asymptotic mean selectivity estimator is introduced below.

For 1-dimensional case when N is large the *amsel* is defined as follows:

$$amsel_1 = msel_1|_{N \gg 1}\,. \quad (23)$$

For 2-dimensional when N and M are large the *amsel* is defined as follows:

$$amsel_2 = msel_2|_{N \gg 1 \wedge M \gg 1}\,. \quad (24)$$

The absolute selectivity estimation error for any query can be defined as follows:

$$Err(Q,Z) = \left|sel(Q) - \hat{sel}(Q,Z)\right|. \quad (25)$$

The error of mean selectivity estimation defined as follows:

$$MErr(Z) = \left|msel - \hat{msel}(\mathrm{Z})\right|. \quad (26)$$

The definition of $\hat{msel}(Z)$ – the approximate mean selectivity estimator (basing only on coefficients from sampling zone Z) will be introduced in sections 4 and 5.

The error of asymptotic mean selectivity estimation defined as follows:

$$AMErr(Z)=\left|amsel-\hat{amsel}(\mathrm{Z})\right|. \tag{27}$$

The definition of $\hat{amsel}(Z)$ – the asymptotic approximate mean selectivity estimator (for a high resolution of spectrum and basing only on coefficients from sampling zone Z) will be introduced in section 6.

The error definition from (27) will be used to find error-optimal shape of sampling zone in section 8.

4. DCT-based mean selectivity estimator for 1-dimensional case

From (11) we can obtain:

$$sel_1\left(Q_1\right)=\sqrt{\frac{2}{N}}\sum\nolimits_{u=0,\ldots,N-1}\left[k_u g(u)\frac{\sin\left(u\pi b_1\right)-\sin(u\pi a_1)}{\pi u}\right]. \tag{28}$$

Using:

$$\lim\nolimits_{\mathrm{u}\to 0}\frac{\sin(u)}{u}=1=>\lim\nolimits_{\mathrm{u}\to 0}\frac{\sin\left(u\pi b_1\right)-\sin(u\pi a_1)}{\pi u}=b_1-a_1 \tag{29}$$

and

$$\begin{gathered}\sum\nolimits_{u=0,\ldots,N-1}\left[k_u g(u)\frac{\sin(u\pi b_1)-\sin(u\pi a_1)}{\pi u}\right]= \\ =k_u g(u)\left.\frac{\sin(u\pi b_1)-\sin(u\pi a_1)}{\pi u}\right|_{u=0}+\sum\nolimits_{u=1,\ldots,N-1}\left[k_u g(u)\frac{\sin(u\pi b_1)-\sin(u\pi a_1)}{\pi u}\right]\end{gathered} \tag{30}$$

(specially handling case where u is equal 0) we can obtain the selectivity estimator as follows:

$$sel_1\left(Q_1\right)=\sqrt{\frac{2}{N}}\left[k_0 g\left(0\right)\left(b_1-a_1\right)+\sum\nolimits_{u=1,\ldots,N-1}\left[k_u g(u)\frac{\sin\left(u\pi b_1\right)-\sin(u\pi a_1)}{\pi u}\right]\right]. \tag{31}$$

Using (20) and (31) we can obtain $msel_1$ – the mean selectivity estimator based on DCT spectrum coefficients:

$$\begin{gathered}msel_1=\frac{1}{0.5}\sqrt{\frac{2}{N}}k_0 g\left(0\right)\int_0^1\left[\int_{a_1}^1\left(b_1-a_1\right)db_1\right]da_1+ \\ +\sum_{u=1}^{N-1}k_u g(u)\int_0^1\left[\int_{a_1}^1\frac{\sin(u\pi b_1)-\sin(u\pi a_1)}{\pi u}db_1\right]da_1.\end{gathered} \tag{32}$$

After calculating the first definite integral (which is equal 1/6) and placing concrete values of k_u using (4) we can obtain from (32):

$$msel_1=\frac{1}{3}\sqrt{\frac{1}{N}}g\left(0\right)+2\sqrt{\frac{2}{N}}\sum_{u=1}^{N-1}g(u)\frac{1}{\pi u}\int_0^1\left[\int_{a_1}^1\left[\sin\left(u\pi b_1\right)-\sin(u\pi a_1)\right]db_1\right]da_1. \tag{33}$$

The definite integral in (33) can be expressed as follows:

$$\int_0^1 [\int_{a_1}^1 [sin\,(u\pi b_1) - \sin(u\pi a_1)]db_1]da_1 = -\frac{\cos(u\pi)}{u\pi} + \frac{2\sin(u\pi)}{u^2\pi^2} - \frac{1}{u\pi}. \quad (34)$$

The formula (34) is defined for discrete values of $u = 1, \ldots, N-1$, thus it can be evaluated as follows:

$$-\frac{\cos(u\pi)}{u\pi} + \frac{2\sin(u\pi)}{u^2\pi^2} - \frac{1}{u\pi} = \begin{cases} -\frac{2}{\pi u} \wedge u = 2, 4, ... \\ 0 \quad \wedge u = 1, 3, \ldots \end{cases}. \quad (35)$$

Using (33) and (35) the 1-dimensional mean selectivity estimator can be obtained as follows:

$$msel_1 = \frac{1}{3}\sqrt{\frac{1}{N}}\mathrm{g}\,(0) - 2\sqrt{\frac{2}{N}}\sum\nolimits_{\mathrm{u}=2,4,..} g(u)\frac{2}{\pi^2 u^2}, \quad (36)$$

$$msel_1 = \sqrt{\frac{1}{N}}\left[A_0 g\,(0) + A_1 \sum\nolimits_{u=2,4,...} g(u)\frac{1}{u^2}\right] \quad (37)$$

where A_0, A_1 are some constants.

5. DCT-based approximate mean selectivity estimator for 2-dimensional case

Specially handling cases where v or u is equal to 0 (similarly to (28)-(31)), the 2-dimensional selectivity estimator can be obtained as follows:

$$sel_2\,(Q_2) = \sqrt{\frac{2}{N}}\sqrt{\frac{2}{M}}\{k_0 k_0 g\,(0,0)\,(b_1 - a_1)\,(b_2 - a_2) +$$

$$+ \sum\nolimits_{u=1,...,N-1} [k_u k_0 g(u,0)\frac{\sin\,(u\pi b_1) - \sin\,(u\pi a_1)}{\pi u}\,(b_2 - a_2)\,] +$$

$$+ \sum\nolimits_{v=1,...,M-1} [k_0 k_u g\,(0,v)\,(b_1 - a_1)\,\frac{\sin\,(v\pi b_2) - \sin\,(v\pi a_2)}{\pi v}] +$$

$$+ \sum_{\substack{u=1,\ldots,N-1 \\ v=1,\ldots,M-1}} [k_u k_v g(u,v)\frac{\sin\,(u\pi b_1) - \sin\,(u\pi a_1)}{\pi u}\frac{\sin\,(v\pi b_2) - \sin\,(v\pi a_2)}{\pi v}]\}. \quad (38)$$

Similarly to (31)-(37) forms applied for 1-dimensional case, we can find *msel*$_2$ – the 2-dimensional mean selectivity estimator from (22) and (38):

$$msel_2 = \sqrt{\tfrac{1}{N}}\sqrt{\tfrac{1}{M}}\{A_{00}g(0,0) + A_{10}\textstyle\sum_{u=2,4,\ldots} g(u,0)\,\frac{1}{u^2} + \\ + A_{01}\textstyle\sum_{v=2,4,\ldots} g(0,v)\,\frac{1}{v^2} + A_{11}\sum_{\substack{u=2,4,\ldots \\ v=2,4,\ldots}} g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}\} \tag{39}$$

where A_{00}, A_{10}, A_{01}, A_{11} are some constants.

The 2-dimensional approximate mean selectivity estimator based only on Z is defined as follows:

$$\hat{msel}_2(Z) = \sqrt{\frac{1}{N}}\sqrt{\frac{1}{M}}\{A_{00}g(0,0) + A_{10}\sum_{u=2,4,\ldots} g(u,0)\,\frac{1}{u^2} + \\ + A_{01}\sum_{v=2,4,\ldots} g(0,v)\,\frac{1}{v^2} + A_{11}\sum_{\substack{u=2,4,\ldots \\ v=2,4,\ldots \\ (u,v)\in Z}} g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}\}. \tag{40}$$

The formula $\hat{msel}_2(U \times V) = msel_2$ is always true.

Taking into account an assumption for boundaries of the sampling zone Z:

$$(u,0)\in Z => \max(u) = N-1 \text{ and } (0,v)\in Z => \max(v) = M-1, \tag{41}$$

only the last addend in (40) is relevant (i.e. only the last addend in (40) is different from the last addend in (39)).

6. DCT-based asymptotic approximate mean selectivity estimator for 2-dimensional case

For large values of N and M ($N \gg 1 \wedge M \gg 1$) we can assume that u and v become continuous variables and sums in (39) can be approximated by definite integrals. Then the 2-dimensional asymptotic mean selectivity estimator is defined as follows:

$$amsel_2 = \sqrt{\tfrac{1}{N}}\sqrt{\tfrac{1}{M}}\{A_{00}g(0,0) + B_{10}\textstyle\int_{1\le u\le N-1} g(u,0)\,\frac{1}{u^2}du + \\ + B_{01}\textstyle\int_{1\le v\le M-1} g(0,v)\,\frac{1}{v^2}dv + B_{11}\int\limits_{\substack{(u,v)\in U\times V \\ \wedge\, u\ge 1 \\ \wedge\, v\ge 1}} g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv\}. \tag{42}$$

where B_{10}, B_{01}, B_{11} are some constants.

Similarly, for large N and M we can obtain from (40) the 2-dimensional asymptotic approximate mean selectivity estimator:

$$\hat{amsel}_2(Z) = \sqrt{\frac{1}{N}}\sqrt{\frac{1}{M}}\{A_{00}g\,(0,0) + B_{10}\int_{1\le u\le N-1} g\,(u,0)\,\frac{1}{u^2}du +$$

$$+ B_{01}\int_{1\le v\le M-1} g\,(0,v)\,\frac{1}{v^2}dv + B_{11}\int\limits_{\substack{(u,v)\in Z\\ \wedge\, u\ge 1\\ \wedge\, v\ge 1}}\int g\,(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv\}. \quad (43)$$

7. Assumption for DCT spectrum

Till now there was no assumptions for a distribution of attribute values. Let's assume the flat DCT spectrum in our further consideration. This is the worst case for the described method of DCT-based selectivity estimation because there is no energy compaction for this spectrum. This assumption is equivalent to formula:

$$g\,(u,v) \equiv C = \text{const} \wedge C > 0. \quad (44)$$

8. The criterion of error-optimal sampling zone

The asymptotically optimal shape of Z is obtained for the least value of error of asymptotic mean selectivity estimator defined in (27). The criterion of finding the best shape of 2-dimensional sampling zone Z is defined as follows:

$$\inf_Z\ [AMErr(Z)] = \inf_Z\ \left|amsel_2 - \hat{amsel}_2(Z)\right| \quad (45)$$

$$\wedge\, \text{size}\,(\text{Z}) = K \quad (46)$$

where K is predetermined size of set Z. A space needed for storing DCT coefficients in memory is linearly depended on K.

Using (42) and (43) in (45) we obtain:

$$\inf_{Z} \left| \left\{ \sqrt{\frac{1}{NM}} B_{11} \int\limits_{\substack{(u,v)\in U\times V \\ \wedge u\geq 1 \\ \wedge v\geq 1}} \int g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv - \sqrt{\frac{1}{NM}} B_{11} \int\limits_{\substack{(u,v)\in Z \\ \wedge u\geq 1 \\ \wedge v\geq 1}} \int g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv \right\} \right|. \tag{47}$$

The expression $\sqrt{\frac{1}{NM}} B_{11} \int\limits_{\substack{(u,v)\in U\times V \\ \wedge u\geq 1 \\ \wedge v\geq 1}} \int g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv$ doesn't depend on Z so (47) can be reformulated:

$$\inf_{Z} \left| \left\{ C_0 - \sqrt{\frac{1}{NM}} B_{11} \int\limits_{\substack{(u,v)\in Z \\ \wedge u\geq 1 \\ \wedge v\geq 1}} \int g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv \right\} \right| \tag{48}$$

where C_0 is some constant.

Because $g(u, v)$ >0 what results from (44), the formula (48) is equivalent to:

$$\sup_{Z} \left\{ +\sqrt{\frac{1}{NM}} B_{11} \int\limits_{\substack{(u,v)\in Z \\ \wedge u\geq 1 \\ \wedge v\geq 1}} \int g(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv \right\}. \tag{49}$$

(49) can be simplified as:

$$\sup_{Z} \left\{ \int \int_{\substack{(u,v) \in Z \\ \wedge u \geq 1 \\ \wedge v \geq 1}} g(u,v) \frac{1}{u^2} \frac{1}{v^2} du dv \right\}. \tag{50}$$

Because some pairs $(uv) \in Z$ where $u = 0$ or $v = 0$, don't belong to domain of definite integral in (50) we introduce set $Z^* = \{(u, v) \in Z \wedge u \geq 1 \wedge v \geq 1\}$. Z^* is a subregion of the optimal Z satisfying (50), i.e. $Z^* \subset Z$. A constant K^* denotes a size of Z^* where $K^* < K$.

The criterion in (50) and (46) can be formulated using Z^* as finding the optimal shape of Z^* satisfying (51) subject to the constraint (52):

$$\sup_{Z^*} \{ \int \int_{(u,v) \in Z^*} g(u,v) \frac{1}{u^2} \frac{1}{v^2} du dv \tag{51}$$

$$\wedge \operatorname{size}(Z^*) = K^*. \tag{52}$$

Let's define a function $v(u)$ which defines the shape of Z^* . An example of the function $v(u)$ is shown on Fig. 3. The interval $[1, u_{max}]$ is a domain of function $v(u)$. The $v(u)$ satisfies the condition: $\underset{u \in [1, u_{max}]}{\forall} v(u) \geq 0$.

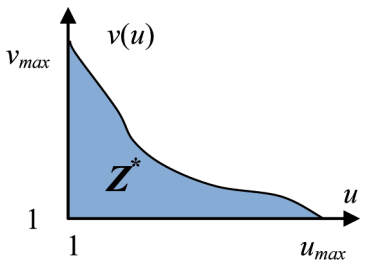

Fig. 3. Example of sampling zone Z^*

The boundary conditions for Z^* are:

$$v(1) = v_{max} \wedge v_{max} \leq M-1 \wedge v(u_{max}) = 1 \wedge u_{max} \leq N-1 \tag{53}$$

(concrete values of v_{max} and u_{max} have no meanings for the proof).

Size of $Z*$can be obtained as follows:

$$\text{size}\,(Z^*) = \int_1^{u_{max}} v\,(u)\,du{=}K^*. \tag{54}$$

The maximized cost functional J can be defined as follows:

$$J{=}\int\limits_{(u,v)\in Z^*}\int g\,(u,v)\,\frac{1}{u^2}\frac{1}{v^2}dudv{=}\int_1^{u_{max}}[\int_1^{v(u)} g(u,v)\frac{1}{u^2v^2}dv]du. \tag{55}$$

The problem of finding optimal $v(u)$ can be solved basing on the classic isoperimetric problem [2] by finding the extremum of J defined in (55) subject to constraints form (54).

9. The method of finding the shape of optimal sampling zone

The problem of finding optimal shape of Z^* i.e. $v(u)$ will be solved by using the method of Lagrange multipliers [2].

The formula (54) can be rewritten as:

$$\int_1^{u_{max}} F_1\left(u,v,v^{'}\right)du{=}K^*{=}\,\text{const} \tag{56}$$

where

$$F_1\left(u,v,v^{'}\right){=}v. \tag{57}$$

The definition of J from (55) can be rewritten as:

$$J{=}\int_1^{u_{max}} F_0\left(u,v,v^{'}\right)du \tag{58}$$

where

$$F_0\left(u,v,v^{'}\right) = \int_1^{v(u)} g(u,v)\frac{1}{u^2v^2}dv. \tag{59}$$

Let's define Lagrange function as follows:

$$L\,(v) = \int_1^{u_{max}} [F_0\left(u,v,v^{'}\right) + \lambda F_1(u,v,v^{'})]du \tag{60}$$

where λ is a Lagrange multiplier.

The Euler's equation required for obtaining the extremum of L is defined as follows:

$$\frac{\partial F_0}{\partial v} - \frac{d}{du}\left(\frac{\partial F_0}{\partial v'}\right) + \lambda\left[\frac{\partial F_1}{\partial v} - \frac{d}{du}\left(\frac{\partial F_1}{\partial v'}\right)\right] = 0. \tag{61}$$

Neither F_0 nor F_1 nor depends on v' so the Euler's equation (61) can be simplified as:

$$\frac{\partial F_0}{\partial v}+\lambda\frac{\partial F_1}{\partial v}=0. \tag{62}$$

Solving the equation (62) allows to find the optimal $v(u)$.

10. Obtaining asymptotically error-optimal shape of sampling zone

Referencing to (59) and (44) we find:

$$F_0\left(u,v,v^{'}\right)=C\int\limits_1^v\frac{1}{u^2v^2}dv=C\frac{1}{u^2}\left(1-\frac{1}{v}\right). \tag{63}$$

Placing $\frac{\partial F_0}{\partial v}=\frac{2C}{u^2v^2}$ and $\frac{\partial F_1}{\partial v}=1$ in (62) we obtain:

$$\frac{2C}{u^2v^2}+\lambda=0, \tag{64}$$

$$u^2v^2=-\frac{2C}{\lambda}=\text{const}. \tag{65}$$

Let's define the some constant:

$$C_1=-\frac{2C}{\lambda} \tag{66}$$

Referencing (65) and (66) and both u and v are positive we can find:

$$uv=\sqrt{C}_1=\text{const}. \tag{67}$$

Hence the reciprocal shape of sampling zone was found as:

$$v\left(u\right)=\frac{\text{const}}{u}. \tag{68}$$

The last equation finishes the presented proof. It is the main goal of this work. This is the confirmation of experimental results from the previous work that for high resolution of spectrum the reciprocal shape of sampling zone (Fig. 2.c) is error-optimal.

11. Conclusions

Effective database query processing requires an estimation of size of a query result before this query is actually executed. This is needed for choosing the best data access method. The early estimation of the query result size is performed by the DBMS cost-query optimizer. The optimizer obtains so-called selectivity factor – the fraction of table

rows satisfying a query condition. Selectivity calculation methods require an estimator of probability density function (PDF) of distribution of table attribute values. For a range query condition based on attributes with continuous domain the selectivity can be obtained as a value of definite integral of PDF (e.g. (2)). A space-efficient multidimensional non-parametric estimator of multivariate PDF is required for obtaining selectivity for complex query conditions involving many attributes.

Among many methods of a representation of multidimensional distribution there is the space-efficient unconventional one, based on Discrete Cosine Transform (DCT) proposed in [9]. In this method the selectivity estimator is calculated using only relevant coefficients of multidimensional DCT spectrum representing a joint distribution of attribute values. Only DCT coefficients belonging to a so-called sampling zone are used in the selectivity estimation. Hence the selectivity calculation is based on a lossy compressed DCT representation of distribution. Lee at al. considered several types of sampling zones (e.g. Fig .2). They experimentally showed that a sampling zone with a reciprocal shape gives the least relative error value for given predetermined size of this sampling zone (Fig. 2.c). The sampling zone‘s size is a critical parameter because either a size of space needed for storing DCT-based representation of distribution or a complexity of the algorithm of the selectivity estimations linearly depends on it.

This paper is devoted to the proof of the theorem that the reciprocal shape of sampling zone is asymptotically error-optimal i.e. for high resolutions of DCT spectrum applying the reciprocal zone gives the least value of selectivity estimation error subject to predetermined size of the sampling zone. This proof was presented in sections 3$\sim$10.

The main contribution of this work is the theoretical confirmation of some only experimental results from the previous work. The proof was carried out under some restrictions: 2-dimensional case, a flat DCT spectrum, an assumed definition of selectivity estuation error. The proof is based on calculus of variations, method of Lagrange multipliers and isoperimetric problem.

References

1. N. Bruno, S. Chaudhuri, L. Gravano: *STHoles: a multidimensional workload-aware histogram*, Proc. of ACM SIGMOD Int. Conf. on Management of Data 30(2), ACM, New York (2001) 211-222.

2. B. Brunt: *The Calculus of Variations*, Series Universitext. Springer-Verlag (2004).

3. K. Chakrabarti, M. Garofalakis, R. Rastogi, K. Shim: *Approximate query processing using wavelets*, The VLDB Journal 10(2-3), Springer-Verlag, New York (2001) 199-223.

4. D. Fuchsa, Z. Zhen Heb, B. S. Lee: *Compressed histograms with arbitrary bucket layouts for selectivity estimation*, Information Sciences. Volume 177, Issue 3, 1 (2007) 680-702.

5. L. Getoor, B. Taskar, D. Koller: *Selectivity estimation using probabilistic models*, Proc. of ACM SIGMOD Int. Conf. on Management of Data 30(2), ACM, New York (2001) 461-472.
6. D. Gunopulos, G. Kollios, V. J. Tsortas, C. Domeniconi: *Selectivity estimator for multidimensional range queries over real attributes*, The VLDB Journal 14(2), Springer-Verlag, New York (2005) 137-154.
7. Y. Ioannidis: *The History of Histograms (abridged)*, Proc. of VLDB Conference (2003).
8. F. Korn, T. Johnson, H. V. Jagadish: *Range Selectivity Estimation for Continuous Attributes*, In Proc. International Conference on Scientific and Statistical Database Management (1999) 244-253.
9. L. Lee, K. Deok-Hwan, Ch. Chin-Wan: *Multi-dimensional selectivity estimation using compressed histogram estimation information*, Proc. of ACM SIGMOD Int. Conf. on Management of Data, ACM, Philadelphia (1999) 205-214.
10. Oracle 10g documentation, Using Extensible Optimizer Page http://download.oracle.com/docs/cd/B14117_01/appdev.101/b10800/dciextopt.htm
11. V. Possala, Y. E. Ioannidis: *Selectivity estimation without the attribute value independence assumption*, Proc. of the 23rd Int. Conf. on Very Large Databases, The VLDB Journal, Athens (1997) 486-495.
12. D. W. Scott, S. R. Sain: *Multi-dimensional Density Estimator*, Handbook of Statistics 24 North-Holland Publishing Co, Amsterdam (2004).
13. F. Yan, W-C. Hou, Z. Jiang, C. Luo, Q. Zhu: *Selectivity estimation of range queries based on data density approximation via cosine series*, Data & Knowledge Engineering 63(3), SienceDirect (2007) 855-878.

Asymptotycznie optymalny kształt strefy próbkowania w metodzie szacowania selektywności zapytań, opartej na dyskretnej transformacie kosinusowej

Streszczenie

Szacowanie selektywności zapytań jest krytyczne dla efektywnej realizacji zapytań w systemach zarządzania bazami danych. Sposób realizacji zapytania zależy od wstępnego oszacowania rozmiaru danych spełniających kryteria zapytania. Takie oszacowanie pozwala wybrać metodę dostępu do danych użytą później podczas realizacji zapytania. Selektywność dla zapytań jednotablicowych to stosunek liczby wierszy spełniających kryteria zapytania do liczby wszystkich wierszy tablicy. Dla zakresowych warunków zapytania, określonych na atrybutach z ciągłą dziedziną, selektywność jest całką oznaczoną z funkcji gęstości prawdopodobieństwa (PDF), określającej rozkład wartości tego atrybutu. Dla złożonych warunków zapytania, opartych na kilku atrybutach, istnieje potrzeba użycia nieparametrycznego estymatora wielowymiarowej PDF, którego reprezentacja powinna być oszczędna pod względem zajętości pamięci. Jedno

ze znanych podejść do konstrukcji takiego estymatora oparte jest na dyskretnej transformacie kosinusowej (DCT) – tzn. widmie z histogramu wielowymiarowego. Własność kompakcji energii pozwala na pominięcie nieznaczących współczynników widma DCT bez istotnej utraty oszacowania selektywności. Obszar znaczących współczynników widma nazywany jest strefą próbkowania. Wyniki prac eksperymentalnych innych autorów wskazują, że dla zadanego rozmiaru reprezentacji widma, optymalną strefą próbkowania (kształtem strefy o najmniejszym błędzie oszacowania selektywności) jest tzw. strefa odwrotnie proporcjonalna. Głównym wynikiem tego opracowania jest teoretyczne potwierdzenie tych eksperymentów. Artykuł przedstawia dowód twierdzenia o asymptotycznej optymalności strefy odwrotnie proporcjonalnej dla przypadku dwuwymiarowego. Dowód opiera się na elementach rachunku wariacyjnego i zagadnieniu izoperymetrycznym.