REGIONS OF SUFFICIENCY FOR METRICAL DATA RETRIEVAL

Vladimir Mashtalir, Konstantin Shcherbinin, Vladislav Shlyakhov, Elena Yegorova

***Abstract**: In this paper the fast metrical search for large scaled and poor structured databases using objects eliminating from the consideration without calculating the distance between them and query is considered and grounded. This search is based on the pre-calculated distances between pivot points and database objects, and triangular inequality as the base for the objects elimination without calculating distances between them. For that the sufficient and necessary conditions of the objects elimination are explored and mathematical foundation for the sufficiency and necessity regions in the metric space are given.*

***Keywords**: Metrical Retrieval, pivot point, sufficiency*

***ACM Classification Keywords**: H3.3.3 Information Search and retrieval.*

Introduction

Nowadays many application require process and retrieve data which is more and more characterized with much less structure and queries precision, as for instance multimedia or video database systems [Geetha, 2008], [Natsev, 2006]. The efficiency of such systems is mainly measured by the search speed and quality. Many ordinary retrieval techniques appear to be inadequate for modern large-scale databases [Snoek, 2008].

The need for more suitable and effective tools gave a push to develop retrieval methods based upon the distance, i.e. performing the search in the metric space [Barsi, 2011], [Kinoshenko, 2010]. Allowing the similarity queries execution this approach was used in a lot of techniques and methods of creating indexing structures. In general the approaches are divided into tree-based (M-tree algorithms, Slim-tree, Slim-Down algorithm, Multi-Way Insertion algorithm) and hash-based (D-index, Insertion and Search strategies) metric indexing [Dohnal, 2003], [Liu, 2007], [Zezula, 2006]. The first ones traverse trees and visit nodes which reside within the query region. In the best case it means the logarithmic search cost. Second ones provide a direct access to searched regions with no additional traversals of the underlying structure, but still the costs increase linearly with the growth of the dataset. It means that with the dataset growth the ability to perform the reasonable time search is rather limited.

Expansion of CBIR, CBVR, CBMR (Content Based {Image, Video, Multimedia} Retrieval) systems led to developing a wide range of algorithms using off-line processing, what allowed to narrow the search to the database objects distance matrices analysis [Liu, 2007], [Kinoshenko[1], 2010]. Still under very large data volumes using of the complete matrix becomes not effective enough and arises the need to analyze sparse matrices which practically represent inner and intra-cluster distances.

One more key aspect of increasing the search speed in the database metrical search is objects eliminating from the consideration without calculating the distance between them and query. This paradigm – eliminate what is not needed and perform the search among what is left – provides essential reduction of the potentially tolerant objects set cardinality and also allows to rank the objects based on their similarity [Zezula, 2006]. At that it is not substantial if the decision is made based on the law level features or actively using semantics. This search is based on the pivot points, the distance to them is preliminary calculated distances between the database objects and triangular inequality is the base for the objects elimination without calculating distances between them. Obviously the more a priori information is considered the bigger speed increase can be expected.

The next paper is organized as following: in section 2 the mathematical premises for metrical search sufficiency and necessity regions are defined, in section 3 and 4 sufficiency conditions analysis and the grounds of sufficiency region construction using one pivot point is conducted, section 5 describes sufficiency region construction using n pivot points.

Sufficiency and necessity regions for metric search

In some phase space Φ we shall consider a finite points configuration $K \subseteq \Phi$, for which the distance matrix $\mathcal{P}$ is known, i.e. a set of point $x_1,\dots,x_n \in \Phi$ actually defines a set of objects in the database (in this case can be video strings, characterized by segments features in time and/or space, key frames, events, etc) among which a subset of most similar to $y \in \Phi$ query will be selected. Thus, a symmetric (it is essential for the objects storing and indexing) $(n \times n)$ of-line calculated matrix is given

$$\mathcal{P} = \left(\rho(x_i, x_j)\right)_{i,j=1}^{n} \quad (1)$$

where $\rho(x_i, x_j)$ is a distance between points x_i, x_j.

For some query i.e. point $y \in \Phi$, we shall define the distance to the pivot point $x* \in K$ (we shall not discuss here the choice of one or few point like that, as well as the optimal number of those)

$$\delta = \rho(y, x*). \quad (2)$$

We shall introduce the objects similarity threshold Δ which characterizes similarity measure of points in metrical space Φ with the current point y. Thus if

$$\rho(y, x_k) > \Delta \quad (3)$$

Then point x_k so much differs from y, that it is automatically eliminated from the further analysis.

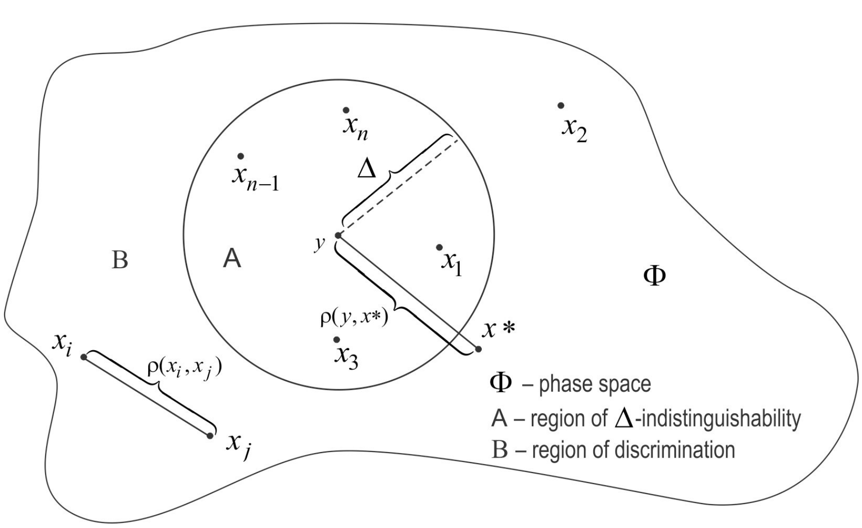


Fig. 1. Geometrical interpretation of metrical search conditions

Based on the stated above, we shall aim to create a procedure of defining the closest points of the K configuration to the given point $y \in \Phi$. At that we consider a distance matrix $\mathcal{P}$, formed according to (1), similarity measure Δ and known distance (2) between current point y and pivot point $x* \in K$. It should be emphasized that this task is first of all conditioned with the complexity of computing $\rho(y, x_k)$ distances, what under large cardinalities of configurations $K \subseteq \Phi$ can lead to intolerable time expenditures. It is clear, that the search result (if exists, as not all points satisfy (3)) can be the closest according the given metric point or m points ordered ascending on the distance to the query. In other words, the result of (Δ, k)- search on query $y \in \Phi$ are elements of set $K^m = \{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\} \subseteq K$, $0 < m \le k \le n$, for which

$$\begin{aligned} &\forall x_{i_j} \in K^m, \forall x \in K \setminus K^m, \forall y \in \Phi \quad \rho(y, x_{i_j}) \le \Delta, \ \Delta \ge 0, \\ &\rho(y, x_{i_j}) \le \rho(y, x), \rho(y, x_{i_j}) \le \rho(y, x_{i_{j+1}}), j = \overline{1, m-1}. \end{aligned} \tag{4}$$

Let us now analyze sufficiency and necessity regions for search (4). We shall fix two points of given configuration $x_1, x_2 \in K$. Then naturally constructing the procedure of choosing the closest among two objects in the database without calculating the distance between then and query y allows to solve the stated task in the most general case. At that it is assumed that analysis (choice of the closest one) of any pairs of points of set K (total number of variants for the analysis is C_n^2) in sense of time and material costs in view of measuring the distance $\rho(y, x_1)$ and $\rho(y, x_2)$ is quite insignificant value. Such an assumption is common for the video frames processing tasks as searching for most similar frames or their sequences to the query commonly needs to be carried out within rough time restrictions, and distance calculations in a phase space Φ is usually computationally expensive. Let us introduce the next definitions.

Definition 1. A sufficiency region S we shall call a region from $\mathbb{R}^n$ with following properties:

– a function $\varphi : K \times K \to \mathbb{R}^n$ exists, i.e. $\varphi(x_1, x_2) = (\alpha_1, \alpha_2, \ldots, \alpha_n)$;

– for this function is true that if $\varphi(x_1, x_2) \in S$, then it is sufficient to find $min\{\rho(y, x_1), \rho(y, x_2)\}$.

Shall we point out that the values of distances $\rho(y, x_1), \rho(y, x_2)$ are not known, but we can indicate the minimal one. Namely this property gives the main premise for constructing the fast-operating elimination procedures.

Definition 2. The region of necessity N we shall call a region from $\mathbb{R}^n$ which has the following properties:

– a function $\psi : K \times K \to \mathbb{R}^n$ exists, i.e. $\psi(x_1, x_2) = (\alpha_1, \alpha_2, \ldots, \alpha_n)$;

– for this function is true that if one can indicate a proper inequality from the set

$$\begin{cases} \rho(y, x_1) \ge \rho(y, x_2); \\ \rho(y, x_2) \ge \rho(y, x_1), \end{cases}$$

then condition $\psi(x_1, x_2) \in N$ should be fulfilled.

The core of these definitions is that dimensionality of the sufficiency and necessity is defined by the fact how fully we use the input information (the more fully we use it the larger dimension we shall get). From the other hand if $\varphi(x_1, x_2)$ is in region S it is enough to choose the point most close to query y. On the contrary, if we know the point most close to y, it is necessary that membership $\psi(x_1, x_2) \in N$ is fulfilled. It is clear that knowing of the sufficiency region in general allows to solve the given task. And ability to choose the closest point from an arbitrary pair allows to sequentially find the most close one from the K set when following for instance the next procedure:

– choose the closest to query y point from x' and x'';

– suppose that it is point x', then we eliminate x'' and move to the next step;

– choose the next (is matching has not finished) point x''', and then the closest one from x' and x''' and go back to the previous step.

Finally we end up with the point from K set which is the closest one to the query. At that the matching operations number would be equal to $card(K) - 1$. All said above can be formulated in next statement.

Proposition 1. Knowledge of sufficiency region S for arbitrary pair of point $x_1, x_2 \in K$ allows to find the point closest to query y.

Now let us analyze the necessity region. For that we should first consider some properties of the sufficiency and necessity regions.

Regions S and N are parted into two subregions S', S'' (N', N''). Region S' satisfies the situation when

$$\rho(y, x_1) \geq \rho(y, x_2) \tag{5}$$

and S'' is for

$$\rho(y, x_2) \geq \rho(y, x_1). \tag{6}$$

Then $S = S' \cup S''$ and $S_1 \cap S'' = \varnothing$ except for the situation when $\rho(y, x_1) = \rho(y, x_2)$. It is clear that similar property is also fulfilled for N.

Let us assume now that it is a priori known which of the inequalities (5) or (6) is fulfilled. Then if (5) is fulfilled, it should be $\psi(x_1, x_2) \in N'$, and if (6) is fulfilled – $\psi(x_1, x_2) \in N''$.

Suppose that

$$\psi(x_1, x_2) \notin N' \tag{7}$$

then (5) is not fulfilled and so the (6) is true. In other words from $\psi(x_1, x_2) \notin N'$ follows condition (6), i.e. it is sufficient for (6) or $\psi(x_1, x_2) \notin N' \subset S''$.

From the other hand if (7) is true then (6) is fulfilled and so $\psi(x_1, x_2) \in N''$. Thus for the regions exists an enclosures chain: $N' \subset \psi(x_1, x_2) \notin N' \subset S''$, i.e. $N'' \subset S''$. But D'' is a sufficiency region of inequality (6), from which follows the sufficiency region N'', and so $S'' \subset N''$. From here it follows that $S'' = N''$. Similarly we can show that $S' = N'$. Therefore $N = S$. We can formulate that as following.

Proposition 2. Under the metrical search based on alternatives (5), (6) the sufficiency region coincides the necessity region.

Thus, the conditions which characterize regions of sufficiency and necessity are necessary and sufficient. From proposition 2 it follows that if one can construct the necessity region, it fully solves the given task. But any sufficiency subregion is also of the same kind, thus the sufficiency region can be narrowed, which is not possible for the necessity region. Practically we state that the maximal sufficiency region coincide the necessity region. At that regions of sufficiency and necessity should satisfy the $S \subset N$ enclosure and only $S_{max} = N$.

When stating the above we certainly assume that S', S'' are the maximal regions of sufficiency for fulfillment of inequalities (5) and (6) accordingly.

It should be particularly noted that when solving the non-alternative search tasks the regions S and N do not have to satisfy the Proposition 2.

Summarizing the above we can introduce some generalized concept of aggregated phase space.

Definition 3. Cumulative phase space $\widetilde{\Phi}$ we shall call a couple $\widetilde{\Phi} = \langle \Phi, U \rangle$, where Φ is a phase space to which belong the current point y and points configuration K, and U is a set of input data used under the objects elimination.

All so far obtained results are true also for the cumulated phase space $\widetilde{\Phi}$. It relates to the fact that considering the conditions of sufficiency and necessity we did not take into consideration the initial data set (1) – (3) as set

$$U = \{\mathcal{P}, \delta, \Delta\}. \quad (8)$$

Further we shall consider finding the sufficiency regions for pairs of points under different concrete realizations of U conditions.

Sufficiency conditions analysis when having one pivot point

We shall start from the situation when conditions have form (8) and configuration K consist of 3 points $K = \{x*, x_1, x_2\}$, where the pivot point $x*$ is separated and for it $\rho(y, x*) = \delta$ is known. Now we can concretize the elements of U as

$$\mathcal{P} = \begin{pmatrix} \rho(x*, x*) & \rho(x*, x_1) & \rho(x*, x_2) \\ \rho(x*, x_1) & \rho(x_1, x_1) & \rho(x_1, x_2) \\ \rho(x*, x_2) & \rho(x_1, x_2) & \rho(x_2, x_2) \end{pmatrix}, \rho(y, x*) = \delta > 0 \quad (9)$$

and point for which the condition of similarity degree condition $\rho(y, x*) < \Delta$ is fulfilled are considered.

We should notice that taken into consideration points configuration K of phase space Φ automatically satisfies the last condition as in contrary the closest point search algorithm is obvious or leads to the confluent situation. Indeed if at least one point of configuration $K = \{x*, x_1, x_2\} \in \Phi$ does not lie in the neighborhood $U_\Delta(y)$ of radius Δ of the current point $y \in \Phi$, then other variants are possible.

Let $x* \notin U_\Delta(y)$. Then the chosen pivot point in the configuration conditionally is “not quite similar” to the current one. So in this case the pivot point choice is not successful and another point should be chosen

Assume that one of the points x_1 or x_2 does not belong to neighborhood $U_\Delta(y)$. If it is x_1, then obviously x_2 is the closest one. If both of then do not belong, these is no sense of searching the closest one.

It has to be mentioned that as opposed to the pivot point $x*$, it is not possible to define membership of configuration K points in neighborhood $U_\Delta(y)$, as distances $\rho(y, x_1)$ and $\rho(y, x_2)$ are not known. But it is meant that when searching for the sufficiency region this condition should be automatically fulfilled. If there is not need in similarity thresholds or there is not possible to choose them, it is assumed that $\Delta = \infty$, and according to (4) one search k points ranged according to their similarity degree.

Let us denote $\rho(x*, x_1) = \alpha, \rho(x*, x_2) = \beta$ and $\rho(x_1, x_2) = \gamma$. Then distance matrix $\mathcal{P}$ with consideration of (9) and metric properties takes form

$$\mathcal{P} = \begin{pmatrix} 0 & \alpha & \beta \\ \alpha & 0 & \gamma \\ \beta & \gamma & 0 \end{pmatrix}. \tag{10}$$

Based on the distance matrix $\mathcal{P}$ (actually assemblage of numbers (α, β, γ)) the sufficiency region can be constructed in $\mathbb{R}^3$. Using the definition from Section 2 we have function $\varphi : K \times K \to \mathbb{R}^3$ defined by the distance matrix $\mathcal{P}$ and equal to

$$\varphi(x_1, x_2) = (\alpha, \beta, \gamma) \in \mathbb{R}^3 . \tag{11}$$

Thus consideration of the metric properties, and first of all the triangular inequality creates the premises for sufficiency region constructing.

Sufficiency region construction under one pivot point

Let us assume that we have a situation according to the schema on the Figure 1 and function $\varphi(x_1, x_2)$ given by equality (11). Then sufficiency region in $\mathbb{R}^3$ under parameters (α, β, γ) is defined as following:

– restrictions introduced by virtue of α, β, γ being distances, i.e.

$$\alpha, \beta, \gamma > 0 ; \tag{12}$$

– restrictions which follow from the triangular inequality according to Figure 2.

$$\begin{cases} |\rho - \alpha| \le \rho(y, x_1) \le \rho + \alpha; \\ |\rho - \beta| \le \rho(y, x_2) \le \rho + \beta \end{cases} \tag{13}$$

are inequalities corresponding to the triangular with current point y, but there is a set of inequalities corresponding to the configuration K:

$$\begin{cases} \alpha + \beta \ge \gamma; \\ \alpha + \gamma \ge \beta; \\ \gamma + \beta \ge \alpha; \end{cases} \tag{14}$$

– condition which follows from the similarity criterion, i.e. restrictions of configuration $K \in U_\Delta(y)$ or

$$\begin{cases} \rho(y, x*) = \rho \le \Delta; \\ \rho(y, x_1) \le \Delta; \\ \rho(y, x_2) \le \Delta. \end{cases} \tag{15}$$

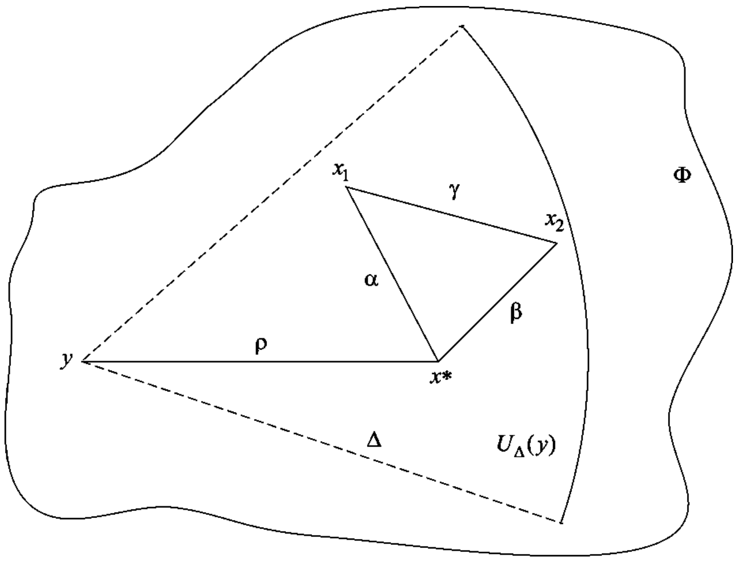


Fig. 2. Geometrical interpretation of search conditions in $\mathbb{R}^3$

In the result it is easy to understand that the sufficiency region will be a region where at least one of the lower bounds of inequalities (13) ranks over the upper bound of the same inequalities. Formally it can be written as

$$|\rho-\alpha| \geq \rho+\beta \tag{16}$$

or

$$|\rho-\beta| \geq \rho+\alpha \ . \tag{17}$$

One can see that from (16) follows $\rho(y,x_1) \geq \rho(y,x_2)$, i.e. point x_2 is closer to current point y. And vice versa, from (17) it follows that $\rho(y,x_2) \geq \rho(y,x_1)$, i.e. point x_1 is closer to current point y.

Taking into account all inequalities (12) – (15) the sufficiency region in $\mathbb{R}^3$ satisfies the conditions set

$$\begin{cases} \alpha,\beta,\gamma > 0; \rho,\Delta > 0; \\ \left[\begin{array}{l} |\rho-\alpha| \geq \rho+\beta; \\ |\rho-\beta| \geq \rho+\alpha; \end{array}\right. \\ \alpha+\beta \geq \gamma; \\ \alpha+\gamma \geq \beta; \\ \gamma+\beta \geq \alpha; \\ \rho \leq \Delta; \\ \rho+\alpha \leq \Delta; \\ \rho+\beta \leq \Delta. \end{cases} \tag{18}$$

One should note that the last three inequalities of system (18) provide the membership of configuration $K \in U_\Delta(y)$. Indeed, $\rho \leq \Delta$ signifies that $x* \in U_\Delta(y)$. And if $\rho+\alpha \leq \Delta$, then from (13) it follows

$$\rho(y,x_1) \leq \rho+\alpha \leq \Delta$$

or $x_1 \in U_\Delta(y)$. Similarly from $\rho + \beta \le \Delta$ follows that $x_2 \in U_\Delta(y)$.

Thus the next statement is proved.

Proposition 3*.* If elements α, β, γ of distance matrix (10) satisfies the system of conditions (9), then the choice of the closest point of configuration K to the query y can be made without calculating the distance to it.

As the considered choice situation is limited by $\mathbb{R}^3$, there is a possibility to give a geometrical interpretation of the system (9) or explicitly point out the sufficiency region in the $\mathbb{R}^3$ space as a region restrained by planes. We shall start with the inequality of type

$$\begin{cases} |\rho - \alpha| \ge \rho + \beta; \\ |\rho - \beta| \ge \rho + \alpha \end{cases} \tag{19}$$

Note that this set of conditions is the most essential one, as it allows to choose the closest point (otherwise the algorithm of choosing can be terminated). Practically below we shall construct a sufficiency region projection on the plane of variables (α, β) or plane $\gamma = 0$. System (19) can be put in form of

$$\begin{cases} \begin{cases} \rho - \alpha \ge \rho + \beta; \\ \rho - \alpha \le -\rho - \beta; \end{cases} \\ \begin{cases} \rho - \beta \ge \rho + \alpha; \\ \rho - \beta \le -\rho - \alpha. \end{cases} \end{cases} \tag{20}$$

We shall consider the first part of system (20), i.e. union of first two inequalities. It is easy to see that the first one $\rho - \alpha \ge \rho + \beta$ leads to inequality $\alpha + \beta \le 0$ and taking into account $\alpha, \beta > 0$ is an empty set. It means that now we have only the second one $\rho - \alpha \le -\rho - \beta$, which is equivalent to inequality $\alpha - \beta \ge 2\rho$, which defines the region of the first quadrant, restricted by lines $\beta = 0$ and $\beta = \alpha - 2\rho$, what is shown as a hatched part on Figure 3.a.

In similar way the second two unity conditions are shown as hatched region of Figure 3.b., and Figure 3c shows all inequalities of (20) in $\mathbb{R}^2$ variant. The $\mathbb{R}^3$ variant is shown on Figure 4.

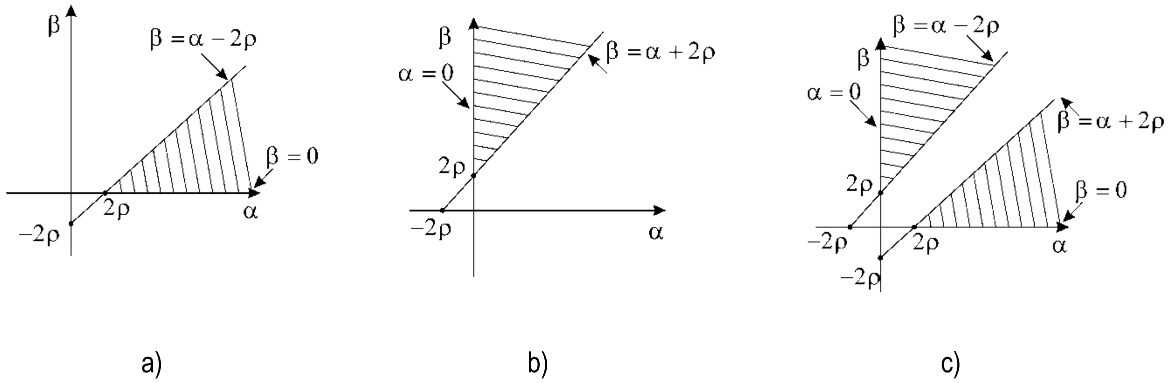


Fig. 3. Graphical interpretation of the a) first pair of inequalities (20); second pair of inequalities (20); all inequalities (20) in $\mathbb{R}^2$

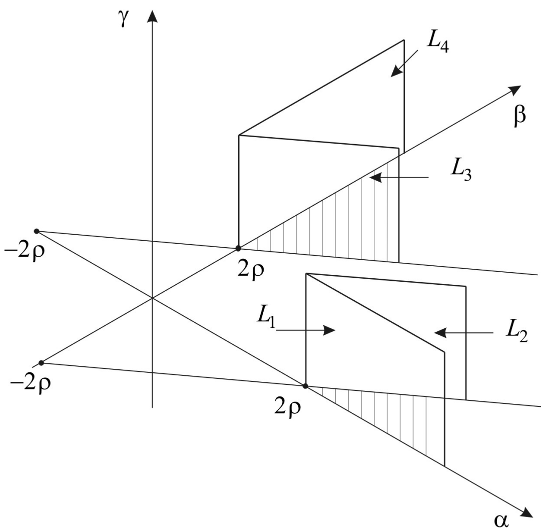


Fig 4. Graphical interpretation of inequalities (20) in $\mathbb{R}^3$

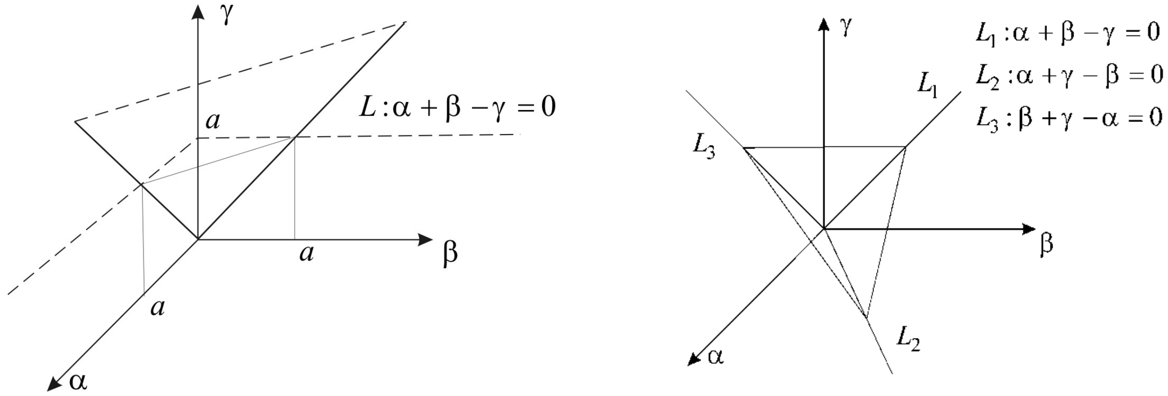
Here the region corresponding to (20), is restrained below by plane $\gamma = 0$ and side planes L_1, L_2, L_3, L_4, the inequalities of which have form

$$\begin{aligned} &L_1 : \beta = 0, \\ &L_2 : \alpha - \beta - 2\rho = 0, \\ &L_3 : \alpha - \beta + 2\rho = 0, \\ &L_4 : \alpha = 0. \end{aligned}$$

Now consider the part of system (18) in form

$$\begin{cases} \alpha + \beta \geq \gamma \\ \alpha + \gamma \geq \beta \\ \gamma + \beta \geq \alpha \end{cases} \tag{21}$$

It is clear to see that in the first inequality its boundary is plane $\alpha + \beta - \gamma = 0$, and region satisfying condition $\alpha + \beta \geq \gamma$ is under plane L (Figure 5a). Naturally the same will correspond (regarding to other axes) to the two other inequalities of system (21). In whole the searched region can be geometrically represented (Figure 5b) as inner part of an infinite regular triangular pyramid.



a) b)

Fig. 5. Geometrical interpretation of the a) first inequality of system (21); b) inequalities system (21)

Finally consider the last three inequalities of system (18)

$$\begin{cases} \rho \le \Delta; \\ \rho + \alpha \le \Delta; \\ \rho + \beta \le \Delta. \end{cases}$$

First of it is practically the condition which should be fulfilled for all parameters of system (18) (in this system α, β, γ are variables in regard to which the sufficiency region is constructed, and ρ and Δ are the system's parameters), so the sufficiency region would not be empty. And the other two inequalities from geometrical point of view define the inner part of an infinite parallelepiped refined by planes $\alpha = 0, \beta = 0, \gamma = 0, \alpha = \Delta - \rho, \beta = \Delta - \rho$ (Figure 6).

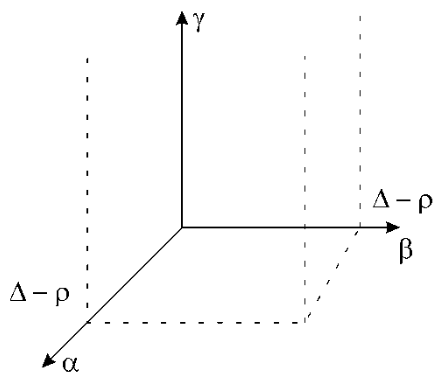


Fig. 6. Geometrical interpretation of threshold restrictions in (18)

Result of intersection of regions shown on Figures 5 and 6 gives us the sufficiency region. At that in order of the sufficiency region not to be empty inequality $\Delta - \rho > 2\rho$ or $\Delta > 3\rho$ should be fulfilled. It finally means that parameters of system (18) ρ and Δ should satisfy more strict condition than $\rho \le \Delta$, namely $\rho < \Delta/3$.

Sufficiency region construction on n pivot points

We shall consider the most general situation when the procedure of defining the closest configuration point is carried out by using n pivot points. Under our approach it means that configuration K consists of $n + 2$ points, as to the pivots 2 compared points are added, and taking into account the current point y it can be represented as shown on Figure 7, notations of which will be explained below.

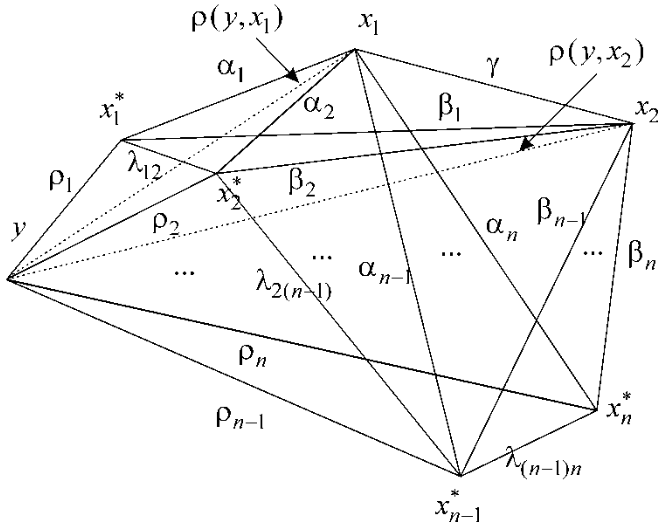


Fig. 7. Geometrical interpretation of search conditions with *n* pivot points

The notations are based on the following: 1) compared distances are denoted as $\rho(y, x_1)$ and $\rho(y, x_2)$; 2) $\rho(x_1, x_2) = \gamma$; 3) $\rho(x_i^*, x_1) = \alpha_i$, $\rho(x_i^*, x_2) = \beta_i$ $i = 1,...,n$ 4) $\rho(y, x_i^*) = \rho_i$; 5) $\rho(x_i^*, x_j^*) = \lambda_{ij}, i \neq j, \lambda_{ij} = \lambda_{ji}$. Here the distance matrix $\mathcal{P}$ will have the order $(n+2)\times(n+2)$, as configuration K contains $n+2$ points and has a form

$$\mathcal{P} = \begin{pmatrix} 0 & \gamma & \alpha_1 & \dots & \alpha_i & \dots & \alpha_n \\ \gamma & 0 & \beta_1 & \dots & \beta_i & \dots & \beta_n \\ \alpha_1 & \beta_1 & 0 & \dots & \lambda_{1i} & \dots & \lambda_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_i & \beta_i & \lambda_{i1} & \dots & 0 & \dots & \lambda_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_n & \beta_n & \lambda_{n1} & \dots & \lambda_{ni} & \dots & 0 \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_1^* \\ \dots \\ x_i^* \\ \dots \\ x_n^* \end{matrix} . \quad (22)$$
$$\begin{matrix} x_1 & x_2 & x_1^* & \dots & x_i^* & \dots & x_n^* \end{matrix}$$

From here it follows that matrix (22) contains $(n+2)^2 - (n+2)/2$ or $r = \frac{(n+2)(n+1)}{2}$ parameters, what is $\psi(x_1, x_2) \in \mathbb{R}^r$, where we should construct the sufficiency region.

As in the previous situation, the main part of parameters restrictions which provide the sufficiency region constructing are those which contain unknown compared distances $\rho(y, x_1)$ and $\rho(y, x_2)$, and if in case with one pivot point there were two of them (12), then for n pivots their number will be $2n$ and they will take form

$$\begin{cases} |\rho_1-\alpha_1| \le \rho(y,x_1) \le \rho_1+\alpha_1; \\ |\rho_1-\beta_1| \le \rho(y,x_2) \le \rho_1+\beta_1; \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ |\rho_i-\alpha_i| \le \rho(y,x_1) \le \rho_i+\alpha_i; \\ |\rho_i-\beta_i| \le \rho(y,x_2) \le \rho_i+\beta_i; \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ |\rho_n-\alpha_n| \le \rho(y,x_1) \le \rho_n+\alpha_n; \\ |\rho_n-\beta_n| \le \rho(y,x_2) \le \rho_n+\beta_n. \end{cases}$$

Having grouped them in a different way, namely

$$\{|\rho_i-\alpha_i| \le \rho(y,x_1) \le \rho_i+\alpha_i \ , \tag{23}$$

where $i=\overline{1,n}$ and

$$\{|\rho_i-\beta_i| \le \rho(y,x_2) \le \rho_i+\beta_i \tag{24}$$

Where $i=\overline{1,n}$, then we can make a choice of the closest point among x_1 and x_2 to the current point y and define the main part of the sufficiency region if at least one of the inequality system (3) lower bounds exceeds at least one of the inequality system (4) upper bounds and vice versa. To be more precise it means that the sufficiency region has a projection in coordinate subspace of parameters $\{\alpha_1,\dots,\alpha_n,\beta_1,\dots,\beta_n\}$ of the main space $\mathbb{R}^r$, yet without considering other restrictions. This projection is defined by union of two inequalities systems: the first one define x_1 as the closest point and has a form

$$\{\rho_i+\alpha_i \le |\rho_j-\beta_j| \tag{25}$$

where the union is fulfilled for all parameters with repetition from the index set $\{1,2,\dots,n\}$, т.е. $i,j=\overline{1,n}$. The number of all these pairs is equal to n^2 .

Similar system in form

$$\{\rho_j+\beta_j \le |\rho_i-\alpha_i|$$

provides the choice of x_2 as the closest one. In the result the sufficiency region projection (without the restrictions which shall be defined later in the coordinate space $\{\alpha_1,\dots,\alpha_n,\beta_1,\dots,\beta_n\}$) is a union of $2n^2$ inequalities which have a form

$$\begin{cases} \rho_i+\alpha_i \le |\rho_j-\beta_j|; \\ \rho_j+\beta_j \le |\rho_i-\alpha_i| \end{cases} \tag{26}$$

where $i,j=\overline{1,n}$.

Let us now consider other restrictions which arise when using n pivot points. First of all it is related to the triangulars which describe one or two pivot points and one or two of the compared ones. Formally there is a set of pivot points $\{x_1^*,\dots,x_n^*\}$ and set of compared ones $\{x_1,x_2\}$, then it is clear that number of the triangulars with vertexes from these two sets can be found as next: if there is one pivot point, their number is n, namely $\Delta(x_i^*, x_1, x_2)$, where $i = \overline{1,n}$; if there are two pivot points, then the number is $2C_n^2$, as from the set $\{x_1, x_2\}$ we can choose one point by two means, and from set $\{x_1^*,\ldots,x_n^*\}$ choosing two points is by C_n^2 choosing means. Thus we get triangulars of type $\Delta(x_i^*, x_j^*, x_k)$, where $i \neq j$, $i, j = \overline{1,n}$ and $k = 1,2$.

From the triangular inequality to the first type of triangulars $\Delta(x_i^*, x_1, x_2)$ correspond n restrictions of type

$$\begin{cases}\alpha_i + \beta_i \geq \gamma;\\ \alpha_i + \gamma \geq \beta_i;\\ \gamma + \beta_i \geq \alpha_i,\end{cases} \tag{27}$$

($i = \overline{1,n}$), to the second type of triangulars $\Delta(x_i^*, x_j^*, x_k)$ correspond C_n^2 of restrictions of type

$$\begin{cases}\alpha_i + \alpha_j \geq \lambda_{ij};\\ \alpha_i + \lambda_{ij} \geq \alpha_j;\\ \lambda_{ij} + \alpha_j \geq \alpha_i,\end{cases} \tag{28}$$

when $k = 1$, i.e. $\Delta(x_i^*, x_j^*, x_1)$, and when $k = 2$, i.e. $\Delta(x_i^*, x_j^*, x_2)$ restrictions will be as following

$$\begin{cases}\beta_i + \beta_j \geq \lambda_{ij};\\ \beta_i + \lambda_{ij} \geq \beta_j;\\ \lambda_{ij} + \beta_j \geq \beta_i,\end{cases} \tag{29}$$

at that $i, j = \overline{1,n}$ and $i \neq j$ for systems (27) and (28).

It should be mentioned that starting with three pivot points additional restrictions arise. They are related to the triangulars which vertexes contain only pivot points (in case that we have at least three pivot points). The number of such triangulars is C_n^3 and in our notations they have a form $\Delta(x_i^*, x_j^*, x_l^*)$ and correspond to the inequalities systems

$$\begin{cases}\lambda_{ij} + \lambda_{il} \geq \lambda_{jl};\\ \lambda_{ij} + \lambda_{jl} \geq \lambda_{il};\\ \lambda_{jl} + \lambda_{il} \geq \lambda_{ij},\end{cases} \tag{30}$$

where $i, j, l = \overline{1,n}$ и $i \neq j, i \neq l, l \neq j$.

Finally the last type of restrictions is connected with the triabgular inequality when one of the triangular's vertexes is the current point y, and two others are pivots x_i, x_j, where $i \neq j$ and $i, j = \overline{1,n}$. These are triangulars of type $\Delta(y, x_i^*, x_j^*)$. The number of those is C_n^2, and they correspond to the inequalities system

$$\begin{cases}\rho_i + \rho_j \geq \lambda_{ij};\\ \rho_i + \lambda_{ij} \geq \rho_j;\\ \lambda_{ij} + \rho_j \geq \rho_i,\end{cases} \tag{31}$$

where $i \neq j$ and $i, j = \overline{1, n}$.

But only one parameter λ_{ij} of the distance matrix takes part in these systems set, so it is expedient to substitute system (31) with the following one

$$\left\{ \left| \rho_i - \rho_j \right| \leq \lambda_{ij} \leq \rho_i + \rho_j, \right. \tag{32}$$

where $i \neq j$ and $i, j = \overline{1, n}$.

Thus we can formulate a statement as for the type of the sufficiency region in the case of n pivots.

Proposition 4. If elements of distance matrix (22) удовлетворяют системам условий (26) – (30), (32), all are positive (situated in the first quadrant) and belong to Δ-neighborhood of query y, i.e. satisfy the restrictions

$$\begin{cases} \{\rho_i + \alpha_i < \Delta; \\ \{\rho_i + \beta_i < \Delta \end{cases} \tag{32}$$

where $0 < \rho_i < \Delta$, then it is enough to choose the closest point of configuration K to the query y without calculating the distance to it.

Conclusion

The sufficient conditions for the database search based on the pre-calculated distances between pivot points and database objects, and triangular inequality as the base for the objects elimination without calculating distances between them are obtained and grounded. There is still a problem of choosing the pivot points and their quantity which was not a part of the current exploration.

Acknowledgements

*The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (*www.ithea.org*) and the Association of Developers and Users of Intelligent Systems ADUIS (*www.aduis.com.ua*).*

Bibliography

[Barsi, 2011] R. Barsi, T. Hassner, L. Zelnik-Manor. Approximate Nearest Subspace Search. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No 2, 2011.

[Dohnal, 2003] V. Dohnal, C. Gennaro, P. Savino, P. Zezula. D-index: Distance Searching Index for Metric Data Sets. In: Multimedia Tools and Applications, Vol. 21, No 1, 2003.

[Geetha, 2008] P. Geetha, V. Narayanan. A Survey of Content-Based Video Retrieval. In: Journal of Computer Science, Vol.4, No 6, 2008.

[Kinoshenko, 2010] D. Kinoshenko, V. Mashtalir, V. Shlyakhov, E. Yegorova. Nested Partitions Properties for Spatial Content Image Retrieval. In: International Journal of Digital Library Systems, Vol. 1, No 3, 2010.

[Kinoshenko1, 2010] D. Kinoshenko, V. Mashtalir, E. Yegorova. Block-Diagonal Forms of Distance Matrices for Partition Based Image Retrieval. In: Pattern Recognition, Recent Advances, A. Herout (ed). Vucovar: In-Thech, 2010.

[Liu, 2007] Y.Liu, D. Zhanga, G. Lua, W.-Y. Ma. A Survey of Content-Based Image Retrieval with High-Level Semantics. In:

Pattern Recognition. Vol. 40, No. 1, 2007.

[Natsev, 2006] A. (P). Natsev. Multimodal Search for Effective Video Retrieval. In: Image and Video Retrieval. Berlin-Heidelberg: Springer-Verlag, Lecture Notes in Computer Science, Vol. 4071, 2006.

[Snoek, 2008] C.G.M. Snoek, M. Worring. Concept-Based Video Retrieval. In: Foundations and Trends in Information Retrieval, Vol. 2, No. 4. 2008.

[Zezula, 2006] A.P. Zezula, G. Amato, V. Dohnal, M. Batko. Similarity Search. The Metric Space Approach. In: Advances in: Database Systems, Vol.32. N.Y.: Springer Science+Business Media, Inc, 2006.

Authors' Information

Vladimir Mashtalir *– Professor, Dean of Computer Science Faculty, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine, email: mashtalir@kture.kharkov.ua*

Major Fields of Scientific Research: Pattern recognition, information retrieval

Konstantin Shcherbinin *– Ph.D. student, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine, email: shcherbinin@kture.kharkov.ua*

Major Fields of Scientific Research: Video segmentation and retrieval

Vladislav Shlyakhov *– Leading Researcher, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine.*

Major Fields of Scientific Research: Artificial intelligence, multialgebraic systems

Elena Yegorova *– Senior Researcher, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine, email: elena.yegorova@gmail.com*

Major Fields of Scientific Research: Image segmentation and retrieval