УДК 004.412

From Similarity to Distance: Axiom Set, Monotonic Transformations and Metric Determinacy

Sergej V. Znamenskij[*]
Ailamazyan Program Systems Institute of RAS
Peter the First Street 4, Veskovo village, Pereslavl area,Yaroslavl region, 152021
Russia

Received 18.11.2017, received in revised form 22.12.2017, accepted 20.02.2018

How to normalise similarity metric to a metric space for a clusterization? A new system of axioms describes the known generalizations of distance metrics and similarity metrics, the Pearson correlation coefficient and the cosine metrics. Equivalent definitions of order-preserving transformations of metrics (both monotonic and pivot-monotonic) are given in various terms. The metric definiteness of convex metric subspaces $\mathbb{R}^n$ and $\mathbb{Z}$ among the pivot-monotonic transformations is proved. Faster formulas for the monotonic normalization of metrics are discussed.

Keywords: metric space, similarity axioms, similarity normalization, metric determinacy, longest common subsequence.
DOI: 10.17516/1997-1397-2018-11-3-331-341.

Following [9], we will call the "distance metric" a function that satisfies the axioms of a metric space, leaving the term "metric" free for use in a broad sense as a real-valued function of two variables.

The length of the longest common subsequence of two text strings (LCS) is a commonly used similarity metric. For example, it is only natural that the line b = "BRITAIN" appears to be more similar to g_b = "GREAT BRITAIN" than to i = "IRAN" ($l(b, g_b) = 7 > 3 = l(b, i)$) and the string r_f = "RUSSIA" is more similar to r_f = "RUSSIAN FEDERATION" than to u = "USA" ($l(r, r_f) = 6 > 3 = l(r, u)$).

Usually, data clustering algorithms work in metric spaces. Known from [13, 26, 14, 17, 5, 35] formulae

$$d_1(x,y) = \frac{2l(x,y)}{|x|+|y|}, \quad d_2(x,y) = \frac{l(x,y)}{|x|+|y|-l(x,y)}, \qquad (1)$$
$$d_3(x,y) = \frac{l(x,y)^2}{|x||y|}, \quad d_4(x,y) = \frac{l(x,y)}{\sqrt{|x||y|}},$$

$$d_5(x,y) = \frac{l(x,y)}{\min(|x|,|y|)}, \qquad (2)$$

where $|x| = l(x,x)$ и $|y| = l(y,y)$, normalize LCS for use in clustering algorithms through conversion to metric or directly. Each of (1) turns the similarity order, so that "BRITAIN" becomes closer to "IRAN", and "RUSSIA" — to "USA". The turn prevents qualitative clustering.

Thought very old study of similarity metrization in psychology [29] showed the hardness of the problem, for formally specified similarity metrics including LCS, the problem came into consideration only in the last decade. Thought the subsequent studies [14] detected some LCS turn with d_2 from (1) for some data set, it did not respond the emerging issues:

— Why no right formula is known for this purpose?

[*] svz@latex.pereslavl.ru

— What is the reason for such a systematic clustering error?

— What are the practically effective formulae with minimal turn?

Some clarification was provided in [31] noting in particular that "the trivial transformations of semimetric spaces into metric ones are not suitable for efficient similarity search" and discussing known approaches, without answers to the questions above.

The construction of a metric space avoiding the turn in similarity is used to analyze experimental data for more than half a century [28]. The success of the multidimensional scaling technology based on this construction [4] caused essential progress in the development of the ordinal embedding theory [19, 2]. These recent studies have theoretically confirmed the *metric determinacy* noted in applied research: the ordinary distance metric for a domain in $\mathbb{R}^n$ is uniquely determined, up to a constant multiplier, by order comparisons.

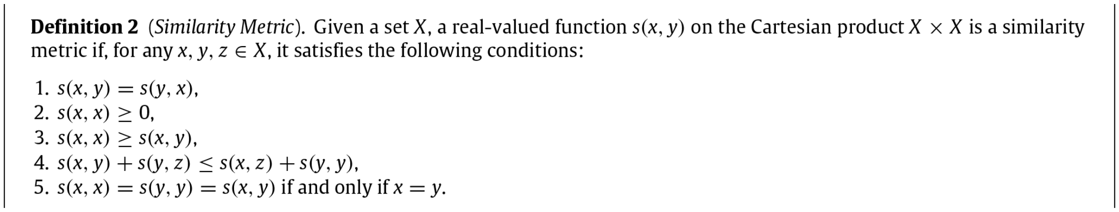
Our situation vary: the original is no longer the metric of the domain in $\mathbb{R}^n$, but the infinite-dimensional similarity metric, and not the set of compared objects should be transformed to the distance metric [9], but the similarity metric itself. We need to understand relations between the similarity metrics and distance metrics.

1. Similarity as a partial metric with minus sign

The non-negativity of distance is generally accepted. Similarity is usually evaluated with a non-negative number, so that zero means a complete lack of similarity. It is often convenient thought to consider the lack of a special similarity as zero similarity and use negative values for apparent opposites. The paper [20, 23] describes the use of Pearson correlation coefficient r as a similarity metric and its transformation to a distance metric. The cosine of the angle between vectors in Euclidean space and the distance with the minus sign $s = -d$ are also used as a metrics of similarity s.

Denote by $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, $\mathbb{R}_+ = \{x \in \mathbb{R} : \quad x > 0\}$ and $\mathbb{R}_{0+} = \{x \in \mathbb{R} : \quad x \geqslant 0\}$.

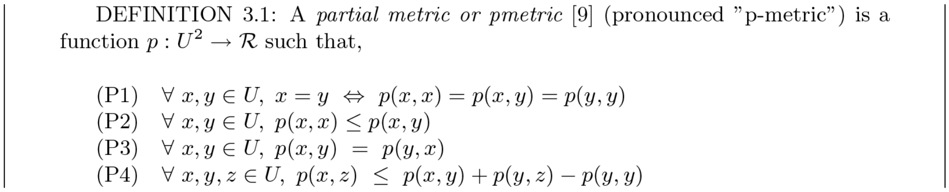
Attempts to construct the axiom set for similarity metrics and dissimilarity metrics [32] showed the triangle inequality from the metric space definition to be not suitable for similarity metrics [33]. New form of the triangle inequality for the similarity metric in [9] reflects the similarity as a measure of coinciding content (i.e., the power of a set of common characteristics, the length of the longest common subsequence, the amount of general information etc.):



Definition 2 (*Similarity Metric*). Given a set X, a real-valued function $s(x, y)$ on the Cartesian product $X \times X$ is a similarity metric if, for any $x, y, z \in X$, it satisfies the following conditions:

1. $s(x, y) = s(y, x)$,
2. $s(x, x) \geq 0$,
3. $s(x, x) \geq s(x, y)$,
4. $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$,
5. $s(x, x) = s(y, y) = s(x, y)$ if and only if $x = y$.

Compare this system of axioms with the system of axioms of partial metrics [22]. Partiall metric space is a generalization of metric space in which the elements can have non-zero dimensions.



DEFINITION 3.1: A *partial metric or pmetric* [9] (pronounced "p-metric") is a function $p : U^2 \to \mathcal{R}$ such that,

(P1) $\forall\ x, y \in U,\ x = y \ \Leftrightarrow\ p(x,x) = p(x,y) = p(y,y)$

(P2) $\forall\ x, y \in U,\ p(x,x) \leq p(x,y)$

(P3) $\forall\ x, y \in U,\ p(x,y)\ =\ p(y,x)$

(P4) $\forall\ x, y, z \in U,\ p(x,z)\ \leq\ p(x,y) + p(y,z) - p(y,y)$

Adding the axiom $\forall_x \in U \quad p(x, x) = 0$ makes this system of axioms equivalent to the usual system of axioms of a metric space.

If we set $s(x,y) = -p(x,y)$, then we see that axioms (P1), (P2), (P3), and (P4) are exactly the axioms 5, 3, 1, and 4. The remaining axiom 2 and similar later additions in the partial metric definition (i.e. [6]) reflect just the natural desire to avoid negative numbers. It should better to move it from axiom set to a set for metric values, usually either $V = \mathbb{R}_{0+}$ or $V = [0,1]$.

The second letter may be uniquely associated with each of the suitable axiom:

$\forall_{x\in U} \quad p(x,x) = 0$ — shotness, thinness;
3(P2) — direction,small self-distances [6], self-similarity [14];
5(P1) — coincidence [25], nondegenerate [15], identity of indiscernibles [9], T0 separation [11], strict positiveness [8];
4(P4) — triangle inequality;
1(P3) — symmetry.

Let U be an arbitrary set, $V \subset \overline{\mathbb{R}}$. For distance $f = d : U \times U \to V$ or for similarity $f = s : U \times U \mapsto V$ under the exception of the axiom of symmetry, the axiom of the direction becomes more complicated and the full list of axioms takes the form:

$$\begin{array}{lll}
\text{(h)} & \forall_{x\in U} & f(x,x) = 0, \\
\text{(i)} & \forall_{x,y\in U} & s(x,y) \leqslant \min(s(x,x), s(y,y)) \;\;\Big|\;\; d(x,y) \geqslant \max(d(x,x), d(y,y)), \\
\text{(o)} & \forall_{x,y\in U} & f(x,y) = f(x,x) = f(y,y) \implies y = x, \\
\text{(r)} & \forall_{x,y,z\in U} & s(x,z) + s(y,y) \geqslant s(x,y) + s(y,z) \;\;\Big|\;\; d(x,z) + d(y,y) \leqslant d(x,y) + d(y,z), \\
\text{(y)} & \forall_{x,y\in U} & f(x,y) = f(y,x).
\end{array}$$

Definition 1. *Let U an arbitrary set, $V \subset \overline{\mathbb{R}}$ and $\mathcal{U} = V^{U\times U}$ the set of all functions of two variables U with values in V and $\mathcal{A} = \{h,i,o,r,y\}$. For any $\mathcal{B} \subset \mathcal{A}$ denote $\mathrm{Sim{:}}\mathcal{B}(U,V) \subset \mathcal{U}$ the subset consisting of all functions $s \in \mathcal{U}$ that satisfies all the axioms of $\mathcal{B}$ and $\mathrm{Dist{:}}\mathcal{B}(U,V) \subset \mathcal{U}$ the subset consisting of all functions $d \in \mathcal{U}$, satisfying all the axioms of $\mathcal{B}$.*

Corollary 1.

1. p is a partial metric on U in the sense of [22] if and only if $p \in \mathrm{Dist{:}iory}(U,\mathbb{R})$.

2. p is a partial metric on U in the sense of [6] if and only if $p \in \mathrm{Dist{:}iory}(U,\mathbb{R}_{0+})$.

3. d is a prameric [3] U if and only if $d \in \mathrm{Dist{:}hi}(U,\mathbb{R})$.

4. d is a semi-metric [1] U if and only if $d \in \mathrm{Dist{:}hioy}(U,\mathbb{R})$.

5. (U,d) is a metric space if and only if when $d \in \mathrm{Dist{:}hiory}(U,\mathbb{R})$.

6. d is a quasi-metric [36, 18] if and only if when $d \in \mathrm{Dist{:}hior}(U,\mathbb{R})$.

7. d is a pseudo-metric [18] if and only if when $d \in \mathrm{Dist{:}hiry}(U,\mathbb{R})$.

8. d is a pseudo-quasi-metric (p-q-metric) [18] if and only if when $d \in \mathrm{Dist{:}hir}(U,\mathbb{R})$.

9. s is a similarity metric on U if and only if when $s \in \mathrm{Sim{:}hiory}(U,\mathbb{R}_{0+})$.

10. $\mathrm{Sim{:}i}(U,V) \cap \mathrm{Dist{:}i}(U,V)$ *consists of constants* $\mathrm{Sim{:}io}(U,V) \cap \mathrm{Dist{:}io}(U,V) = \varnothing$ *for nontrivial U and V.*

11. The Pearson correlation coefficient r and the cosine of the angle between the vectors belong to $\mathrm{Sim{:}iory}(\mathbb{R}^n,\mathbb{R})$.

12. $\forall_{\mathcal{B}\subset\mathcal{A}} \quad s \in \mathrm{Sim{:}}\mathcal{B}(U,V) \iff (-s) \in \mathrm{Dist{:}}\mathcal{B}(U,V)$.

13. $\forall_{\mathcal{B}\subset\mathcal{C}\subset\mathcal{A}} \quad \mathrm{Sim{:}}\mathcal{C}(U,V) \subset \mathrm{Sim{:}}\mathcal{B}(U,V)$ *and* $\mathrm{Dist{:}}\mathcal{C}(U,V) \subset \mathrm{Dist{:}}\mathcal{B}(U,V)$.

We see that the cone of partial metrics on U with values in R is a mirror reflection of the cone of similarity metrics and that the volumes of these concepts are reflected by the scheme in Fig. 1, in which a metric reversion $d = -s$ looks as a central symmetry.

Definition 2. *We call the metric $s \in \mathrm{Sim{:}i}(U,V)$ to be a* LCS-like similarity metric *if it satisfies* align-base *axiom about the existence of* common part*:*

(l) $\quad s(x,y) = \sup\{s(z,z)\colon s(z,z) = s(x,z) = s(z,y),\, z \in U\}$.

and to be a Tversky similariy metric if common part always unique:

(v) $\quad \forall_{x,y\in U} \exists_{z\in U} \qquad s(x,y) = s(z,z) = s(x,z) = s(z,y).$

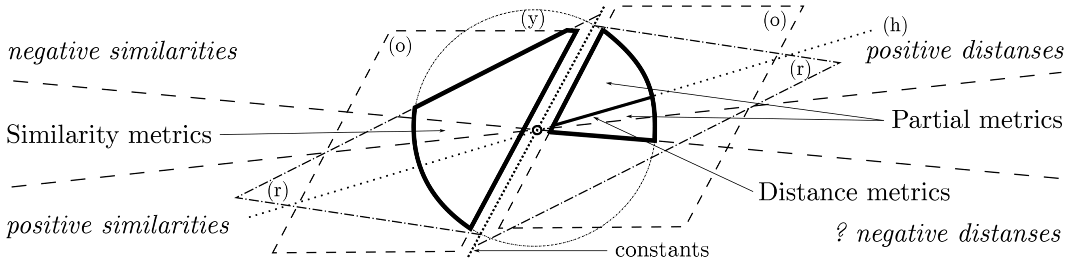


Fig. 1. Relations of distance and similarity metrics satisfying the direction axiom (i)

An important for the data analysis possibility of visual representation of the hierarchy of proximity with the tree phylogenetic tree or evolutionary tree [12] is provided by replacing with a stronger inequality, the *a<u>d</u>ditive inequality* (*same as the four points inequality*)

$$\text{(d)}\quad \forall_{x,y,u,v\in U} \qquad s(x,y)+s(u,v) \geqslant \min(s(x,u)+s(y,v), s(x,v)+s(u,y))$$
$$\Big|\quad d(x,y)+d(u,v) \leqslant \max(d(x,u)+d(y,v), d(x,v)+d(u,y)),$$

or even more powerful *the <u>u</u>ltrametric inequality*

$$\text{(u)}\quad \forall_{x,y,z\in U} \qquad s(x,z) \geqslant \min(s(x,y), s(y,z)) \quad \Big| \quad d(x,z) \leqslant \max(d(x,y), d(y,z)).$$

All the definitions above remains valid for the extension $\mathcal{A} = \{\text{d,i,h,o,l,r,u,v,y}\}$.

Further investigation of this axiom system appears in [37].

2. Monotonic transformations

The axiom of direction allows clusterization algorithms to use closed and open balls with center at a fixed point a:

$$\overline{B}(a,r) = \{x \in U : \, d(a,x) - d(a,a) \leqslant r\}, \qquad B(a,r) = \{x \in U : \, d(a,x) - d(a,a) < r\}$$

for similarity and respectively

$$\overline{B}(a,r) = \{x \in U : \, s(a,a) - s(a,x) \leqslant r\}, \qquad B(a,r) = \{x \in U : \, s(a,a) - s(a,x) < r\}$$

for distance. In contrast to the usual formulae for the balls, these proposed in [27] (cited in [16]) guarantee the emptiness of balls with negative radii and the belonging of the center to balls with positive radii.

Proposition 1 (equivalence of left-monotonic relatedness definitions)**.** *The following conditions on $s_1, s_2 \in \text{Sim:i}(U,V)$ are equivalent:*

1. *The (non)strict inequalities between distances to the third point coincide:*

$$\forall_{x,y,z\in U} \quad s_1(x,z) < s_1(y,z) \iff s_2(x,z) < s_2(y,z); \tag{3}$$
$$\forall_{x,y,z\in U} \quad s_1(x,z) \geqslant s_1(y,z) \iff s_2(x,z) \geqslant s_2(y,z). \tag{4}$$

2. *Open or closed balls with any given center differ only in radii:*

$$\forall_{\{j,j\}=\{1,2\}} \forall_{a\in X} \forall_{r>0} \exists_{r'>0} \quad B_{s_i}(a,r) = B_{s_j}(a,r'); \tag{5}$$
$$\forall_{\{j,j\}=\{1,2\}} \forall_{a\in X} \forall_{r>0} \exists_{r'>0} \quad \overline{B}_{s_i}(a,r) = \overline{B}_{s_j}(a,r'). \tag{6}$$

3. *Open or closed halph-spaces separating arbitrary $a, b \in U$ coincide:*

$$\{x \in X : s_1(a,x) < s_1(b,x)\} = \{x \in X : s_2(a,x) < s_2(b,x)\}; \tag{7}$$
$$\{x \in X : s_1(a,x) \geqslant s_1(b,x)\} = \{x \in X : s_2(a,x) \geqslant s_2(b,x)\}. \tag{8}$$

Symmetric to this statement can be obtained by permuting the arguments of each s_i in Proposition 1. We call s_i to be pivot-monotonically related if they are both left-monotonically related, and right-monotonically related. The pivot-monotonic relatedness means exactly the preservation of the SimOrder (pre)order relation considered in [30] and subsequent studies.

The transformation of metrics we call *pivot-monotonic* if the image of any metric is pivot-monotonically related with its preimage (SimOrder preserving).

Proposition 2 (equivalence of monotonic relatedness [13] definitions)**.** *The following conditions on $s_1, s_2 \in \mathrm{Sim{:}i}(U,V)$ are equivalent:*

1. *The (non)strict inequalities between distances coincide:*

$$\forall_{x,y,u,v \in U} \quad s_1(x,y) < s_1(u,v) \iff s_2(x,y) < s_2(u,v); \tag{9}$$
$$\forall_{x,y,u,v \in U} \quad s_1(x,y) \geqslant s_1(u,v) \iff s_2(x,y) \geqslant s_2(u,v). \tag{10}$$

2. *Open or closed balls differ only in radii and inequalities signs always coincide:*

$$\forall_{\{i,j\}=\{1,2\}} \forall_{a_1,a_2 \in B} \forall_{r_1>r_2>0} \exists_{r'_1>r'_2>0} \forall_{k=1,2} \quad B_{s_i}(a_k, r_k) = B_{s_j}(a_k, r'_k); \tag{11}$$
$$\forall_{\{i,j\}=\{1,2\}} \forall_{a_1,a_2 \in B} \forall_{r_1>r_2>0} \exists_{r'_1>r'_2>0} \forall_{k=1,2} \quad \overline{B}_{s_i}(a_k, r_k) = \overline{B}_{s_j}(a_k, r'_k). \tag{12}$$

The transformation of metrics is called *monotonic* [13] if the image of any metric is monotonically related with its preimage.

Proof of propositions 1 and 2.

(9) $\iff$ (10) is trivial.

(9) $\iff$ (11). Let $r'_k = \inf\limits_{x \notin B_{f_i}(a_k, r_k)} s_j(a_k, x)$. Then

$$\begin{aligned} x \in B_{s_j}(a_k, r'_k) &\iff \forall y \notin B_{s_i}(a_k, r_k) \quad s_j(a_k,x) < s_j(a_k,y) \\ &\iff (s_i(a_k,y) \geqslant r_k \implies s_j(a_k,x) < s_j(a_k,y)) \\ &\iff (s_i(a_k,y) \geqslant r_k \implies s_i(a_k,x) < s_i(a_k,y)). \end{aligned}$$

The last condition obviously holds under $s_i(a,x) < r_k$ and fails under $s_i(a,x) \geqslant r_k$.

(9) $\iff$ (12). Let $r' = \sup\limits_{x \in \overline{B}_{s_i}(a_k, r_k)} s_j(a_k, x)$. Then

$$\begin{aligned} x \notin \overline{B}_{s_j}(a_k, r'_k) &\iff \forall y \in \overline{B}_{s_i}(a_k, r_k) \quad s_j(a_k,x) > s_j(a_k,y) \\ &\iff (s_i(a_k,y) \leqslant r_k \implies s_j(a_k,x) > s_j(a_k,y)) \\ &\iff (s_i(a_k,y) \leqslant r_k \implies s_i(a_k,x) > s_i(a_k,y)). \end{aligned}$$

The last condition obviously holds under $s_i(a,x) > r_k$ and fails under $s_i(a,x) \leqslant r_k$.

The proofs of (3) $\iff$ (5) and (3) $\iff$ (6) are the same as the above just without indexes k. The equivalences (3) $\iff$ (4), (3) $\iff$ (7), and (4) $\iff$ (8) are trivial. □

Corollary 2. *Left-monotonic relatedness $s, t \in \mathrm{Sim{:}i}(U,V)$ implies*

$$\forall_{x,y,z,\in U} \quad s(x,z) = s(y,z) \iff t(x,z) = t(y,z), \tag{13}$$

and monotonic relatedness implies

$$\forall_{x,y,u,v \in U} \quad s(x,y) = s(u,v) \iff t(x,y) = t(u,v). \tag{14}$$

3. Convexity of metric and monotonic determinacy

Corollary 3. *For the metrics $f_1, f_2 \in \mathrm{Sim{:}i}(U, V)$ to be monotonically connected, it is necessary and sufficient that there exist a strictly increasing function ϕ defined on the image of f_1, that*

$$f_2(x, y) = \phi(f_1(x, y)) \quad \forall x, y \in U. \tag{15}$$

Let's denote $f(U \times U)$ the set of all values of f.

Corollary 4. *For the metrics $f_1, f_2 \in \mathrm{Sim{:}i}(U, V)$ to be pivot-monotonically connected, it is necessary and sufficient that there exist a strictly increasing over second argument function $\varphi_p : U \times f_1(U \times U) \to V$, that*

$$f_2(x, y) = \varphi_p(x, f_1(x, y)) \quad \forall x, y \in U. \tag{16}$$

We call a metric transformation to be *monotone* if the image of any metric is monotonically related to its preimage.

The monotone transformation of metrics by the formula (15) is named in [30] SP-modification, and the function ϕ in (15) is SP-modifier. We restrict our discussion to the case $V = \mathbb{R}$, since $V \subset \mathbb{R}$ is mainly used, from which ϕ can be extended to all $\mathbb{R}$. The set of all SP-modifiers forms a partially ordered group (po-group) $G = \mathrm{Aut}(\overline{\mathbb{R}})$ of isomorphisms of linear ordered set $\overline{\mathbb{R}}$. The group operation in it is a superposition of functions ϕ and the unit 1_S is the function $\phi(x) = x$. and the positive cone $\mathcal{P}$ consists of all concave functions on $\overline{\mathbb{R}}$ that are different from linear functions. The cone $\mathcal{P}$ accurately characterizes those monotonic metric transformations that always preserve the triangle inequality [10] but differs from multiplication to a constant.

A partially ordered group G defines on $\mathcal{U}$ a relation of strict partial order $\succ_{\mathcal{P}}$ by the rule $t \succ_{\mathcal{P}} s \overset{\text{def}}{\iff} \exists_{\phi\in\mathcal{P}} \forall_{x,y\in U} \quad t(x, y) = \phi(s(x, y))$. If t and s here are distance metrics, then t is named in [30] *triangle-generating modification* or TG-modification of metric s.

The strict partial order $\succ_{\mathcal{P}}$ induces the preorder $t \succcurlyeq_{\mathcal{P}} s \overset{\text{def}}{\iff} (s \succ_{\mathcal{P}} u \implies t \succ_{\mathcal{P}} u)$ and the equivalence relation $t \sim_{\mathcal{P}} s \overset{\text{def}}{\iff} (t \succcurlyeq_{\mathcal{P}} s) \,\&\, (s \succcurlyeq_{\mathcal{P}} t)$, that mean coincidence up to an affine transformation $t(x, y) = cs(x, y) + b$ with the appropriate constants $c > 0$ and b. The constant term b disappears if the axiom of shortness (h) is satisfied.

Note that d_2 in (1) is monotonically related to the Levenshtein distance, and that the examples of geographical names cited at the beginning of the article clearly indicate the particularity of the Levenshtein formula, commonly used in data cleansing applications [21, 34] in comparison with the LCS similarity metric or other similarity metric [38].

Clear statement of our problem requires a criterion for the optimality of the distance metric. For this purpose [7] suggest to use the *intrinsic dimensionality* calculated through the mathematical expectation μ_d and the variance σ_d^2 of prametric d by the formula $\mathrm{IDim}_\mu(d) = \dfrac{\mu_d^2}{2\sigma_d^2}$. Comparative review of other definitions for intrinsic dimensionality with computer evaluation can be found in [24].

Let $\mathbb{N}_k = \{0, \ldots, k + 1\}$ be an integer segment with the metric $d_{\mathbb{N}_k}(m, n) = |n - m|$. It seems intuitively plausible that the metric $d_{\mathbb{N}_k}$ has the smallest internal dimension among all pivot-monotonically equivalent metrics.

None of the approaches to determining the intrinsic dimension considered in [24] succeeded to prove or disprove this assertion. Most of them assume the assignment of a probability measure μ on $U \times U$, which is quite natural for a data set. On $\mathbb{N}_k$ one can use the uniform probability measure.

As an alternative, consider the notion of convexity that allocates this space. The convexity of a metric space is intuitively associated with the presence of a segment with arbitrary ends. The convexity of a metric space is intuitively associated with the presence of the segment free ends. Usually the definition of convexity requires the existence of midpoint of a segment and the consistent application of this definition gives a dense on the segment set of points. To discrete metric spaces such a definition is inapplicable and an alternative is required.

Definition 3. *We call the metric $d \in \mathrm{Dist{:}i}(U, V)$ or $s = -d$ to be convex if for any $x, z \in U$ and $t \in V$ inside $(d(x, x), d(x, z))$ there exists $y \in U$, for which $d(x, y) = t$ and the triangle inequality for x, y, z turns to equality.*

Proposition 3. *Let $d \in \mathrm{Dist{:}iy}(U, \mathbb{N}_k)$ is convex and $d(x, y) = k$. Then there exists an isometric inclusion $\psi : \mathbb{N}_k \to U$ with the ends $\psi(0) = x$ u $\psi(k) = y$.*

Proof. We use induction on k. For $k = 1$, by assumption, there exist $(x, y) \in U$ such that $d(x, y) = 1$. Assuming $\psi(0) = x$ and $\psi(1) = y$, we obtain the desired isometry.

Suppose that the assertion is proved for $k = n$. By hypothesis, for $k = n + 1$ there exist $(x, y) \in U$ such that $d(x, y) = k$. Using the convexity of the metric with $t = n$, we obtain $m \in U$ for which $d(x, m) = t$ and $d(m, y) = d(x, y) - t$. Applying the induction hypothesis to the points x, m of an open ball of radius k with the center at x, we obtain a map that remains to be extened by the equality $\psi(k) = y$. In this case, the equalities $d(\psi(i), y) = ki$ follow from the triangle inequalities $d(\psi(0), \psi(i)) + d(\psi(i), y) \leqslant d(x, y) = k$ and $d(\psi(i), n) + d(\psi(n), y) = n - i + 1 \leqslant$ $\leqslant d(\psi(i), y)$. □

Theorem 1. *Suppose that two convex pseudometrics are monotonically connected and the set of values of one of them is an arithmetic progression or is closed with respect to addition or is closed with respect to the calculation of the half-sum. Then they differ by multiplication by a constant.*

Note that the multiplication of a metric by a constant does not change its intrinsic dimension and convexity.

Proof. The case of an arithmetic progression follows directly from the Proposition 3. Let $0 <$ $< x_1 = f(a_1, b_1) < x_2 = f(a_2, b_2)$. It is necessary to show that $\dfrac{x_1}{x_2} = \dfrac{\varphi(x_1)}{\varphi(x_2)}$.

For any $k \in \mathbb{N}$ let $n_k \leqslant k\dfrac{\varphi(x_1)}{\varphi(x_2)} < n_k + 1$. Let's consider the equivalent inequality:

$$n_k\varphi(x_2) \leqslant k\varphi(x_1) < (n_k + 1)\varphi(x_2). \tag{17}$$

1. Closedness with respect to the operation of addition allows us to apply the Proposition 3 with an arbitrarily large k to the functions $\dfrac{f(x, y)}{x_1}$, $g(x, y)$ as well as to functions $\dfrac{f(x, y)}{x_2}$, $g(x, y)$ and obtain

$$\varphi(n_k x_2) \leqslant \varphi(k x_1) < \varphi((n_k + 1)x_2).$$

Applying the monotonicy of φ, we get $n_k x_2 \leqslant k x_1 < (n_k + 1)x_2$ and it remains to pass to the limit in $\dfrac{n_k}{k} < \dfrac{x_1}{x_2} < \dfrac{n_k + 1}{k}$.

2. Closedness with respect to the mean calculation also allows us to combine two sufficiently long arithmetic progressions of the values, constructing them by midpoint selections.

Fix $2^m > r + n_k + k$ and using the middle point selections construct in $f(X \times X)$ two grids of size 2^m with the endpoints $x_{10} = 2^{-m}x_1$ and $x_{20} = 2^{-m}x_2$ respectively. Then (17) gives

$$2^{-m}n_k\varphi(x_2) \leqslant 2^{-m}k\varphi(x_1) < 2^{-m}(n_k + 1)\varphi(x_2).$$

In contrast to the proof of the previous case, here we have to apply the Proposition 3 to each part also in the opposite direction with a 2^m grid to get

$$\varphi(2^{-m}n_k x_2) \leqslant \varphi(2^{-m}kx_1) < \varphi(2^{-m}(n_k+1)x_2).$$

This and the monotonicity of φ imply the equality $n_k x_2 \leqslant kx_1 < (n_k+1)x_2$ and it remains to pass to the limit in $\dfrac{n_k}{k} < \dfrac{x_1}{x_2} < \dfrac{n_k+1}{k}$. □

Corollary 5 (monotonic determinacy of metrics)**.** *The statement "If the monotonous transformation of the distance metric is convex, then it is multiplication by a suitable positive constant" is true for each of the following metric spaces: $\mathbb{Z}^n$, $\mathbb{N}$, $\mathbb{N}_k$, an arbitrary convex subset of $\mathbb{R}^n$.*

4. Monotonic normalization

Proposition 4. *Let the metric of similarity $s \in \mathrm{Sim{:}iy}(U,\mathbb{R})$ satisfies $\exists_{x,y,z\in U} \quad s(x,y) > s(z,z)$. Then no metric $d \in \mathrm{Dist{:}hi}(U,\mathbb{R}_{0+})$ can be monotonically related to s.*

Proof. $0 = d(z,z) > d(x,y)$. □

If we change the zero values, the situation will change:

Proposition 5. *Let the metric of similarity $s \in \mathrm{Sim{:}iy}(U,\mathbb{R}_{0+})$. Then the formula*

$$d(x,y) = \frac{1}{2} + \frac{1}{2+s(x,y)}, \qquad x \neq y \tag{18}$$

defines the distance metric $d \in \mathrm{Dist{:}hiory}(U,[0,1])$, which satisfies the condition (9) *of monotonic relatedness to s for all $x \neq y, u \neq v \in U$.*

Unfortunately, Theorem 3 in [8] state that none current acceleration technology for the nearest neighbor search will be effective for this metric since all non-zero values are between 1 and 2. The formulae (1)–(2) for LCS also narrow the range of nonzero values, which usually leads to greater intrinsic dimension and consequently to smaller efficiency. The formula

$$d(x,y) = 1 - \frac{s(x,y)}{M}, \qquad x \neq y \tag{19}$$

also monotonically transforms to $\mathrm{Dist{:}hiory}(U,[0,1])$ and defines the distance metric if M is large enough. It may be possible to decrease the intrinsic dimension selecting smaller M.

Proposition 6. *Let the similarity metric $s \in \mathrm{Sim{:}iry}(U,\mathbb{R}_{0+})$. Then the formula $d(x,y) = s(x,x) - s(x,y)$ defines p-q-metric $d \in \mathrm{Dist{:}hir}(U,\mathbb{R}_{0+})$ to be left-monotonically related with s.*

However, to make it at least pseudometric, we need some symmetrized function $\phi(s,t)$. This function should be convex if s is convex to preserve the triangle inequality and should be close to linear to avoid high intrinsic dimension.

Uzing a segmnet $[\varepsilon,\lambda] \subset \mathbb{R}_0$ containing all possible positive value of s on the data set, it is possible to limit values of d by the $[0,1]$ segment:

$$d(x,y) = \lambda^{-1}\phi(s(x,x),s(y,y)) - s(x,y). \tag{20}$$

Among the admissible functions, there are the largest

$$d(x,y) = \frac{s(x,x)+s(y,y)-2s(x,y)}{2\lambda}, \tag{21}$$

intermediate

$$d_p(x,y) = \lambda^{-1}\sqrt[p]{(s(x,x))^p + (s(y,y))^p} - s(x,y), \tag{22}$$

$$d(x,y) = \lambda^{-1}\sqrt{s(x,x)s(y,y)} - s(x,y). \tag{23}$$

and the smallest

$$d(x,y) = \frac{\varepsilon + (\min(s(x,x), s(y,y)) - s(x,y))}{\lambda + \varepsilon}, \qquad x \neq y. \tag{24}$$

To localize the turns, in any arbitrary triplet $x, y, z \in U$, we rename the vertices so that $s(x,x) \leqslant s(y,y) \leqslant s(z,z)$. The pivot-monotonicity is violated if $\phi(s(y,y), s(z,z)) - s(y,z) < \phi(s(x,x), s(y,y)) - s(x,y)$ for $s(x,y) < s(y,z)$. For d_8 this means $2(s(y,z) - s(x,y)) > s(y,y) - s(x,x)$, and the Levenshtein metric d_6 turns more often: $2(s(y,z) - s(x,y)) > s(z,z) - s(x,x)$.

In each of these cases, rather natural restrictions on s provides the convexity of the metric so that Theorem 1 confirm the quality of the metric. The possible normalization formulae with some generic restrictions on usage are shown in Tab. 1.

Table 1. Expected performance of normalization formulae for convex similarities

Formula	Turns	Speed	Recommended applications
(18)	never	slow	none
(2)	rare	slow	none
(1)	less rare	slow	none
(19)	never	?	cluster analisys of medium size data
(21)	rare	faster	focused only on objects of high similarity (such as correcting typographical errors)
(22),(23)	more rare	faster	?
(24)	most rare	faster	common use

This work was performed under financial support from the Government, represented by the Ministry of Education and Science of the Russian Federation (Project ID RFMEFI60716X0153).

References

[1] C.Alexander, Semi-developable space and quotient images of metric spaces, *Pacific J. Math*, **37**(1971), 277–293.

[2] E.Arias-Castro, Some theory for ordinal embedding, arXiv:1501.02861 [math.ST], 2015.

[3] A.V.Arkhangel'skii, L.S.Pontryagin, General Topology I: Basic Concepts and Constructions Dimension Theory, Springer, 1990.

[4] I.Borg, P.J.Groenen, Modern multidimensional scaling: Theory and applications, Springer, 2005.

[5] S.Budalakoti, R.Akella, A.N.Srivastava, E.Turkov, Anomaly Detection in Large Sets of High-Dimensional Symbol Sequences, NASA/TM-2006-214553, September, 2006.

[6] M.Bukatin, R.Kopperman, S.Matthews, H.Pajoohesh, Partial metric spaces, *Amer. Math. Monthly*, **116**(2009), no. 8, 708–718.

[7] E.Chávez, G.Navarro, A Probabilistic Spell for the Curse of Dimensionality. ALENEX'01, LNCS 2153, Springer, (2001), 147–160.

[8] E.Chávez, G.Navarro, R.Baeza-Yates, J.L.Marroquín, Searching in metric spaces, *ACM Computing Surveys*, **33**(2001), no. 3, 273–321.

[9] S.Chen, B.Ma, K.Zhang, On the similarity metric and the distance metric, *Theoretical Computer Science*, **410**(2009), no. 24-25, 2365–2376.

[10] P.Corazza, Introduction to metric-preserving functions, *American Mathematical Monthly*, **104**(1999), no. 4, 309–323.

[11] M.M.Deza, E.Deza, Encyclopedia of Distances, Springer, 2009.

[12] A.J.Dobson, Unrooted Trees for Numerical Taxonomy, *Journal of Applied Probability*, **11**(1974), no. 1, 32–42.

[13] N. J. P. van Eck, L.Waltman, How to Normalize Co-Occurrence Data? An Analysis of Some Well-Known Similarity Measures (No. ERS-2009-001-LIS), ERIM report series research in management Erasmus Research Institute of Management, Erasmus Research Institute of Management, 2009. Retrieved from http://hdl.handle.net/1765/14528.

[14] C.H.Elzinga, M.Studer, Normalization of Distance and Similarity in Sequence Analysis, LaCOSA II, Lausanne, June 8-10, 2016, 445–468.

[15] D.J.Greenhoe, Properties of distance spaces with power triangle inequalities, 2016. https://doi.org/10.7287/peerj.preprints.2055v1.

[16] R.Heckmann, Approximation of metric spaces by partial metric spaces, InformatikBerichte 96-04, Technische Universitat Braunschweig, 1996, Workshop Domains II.

[17] A.Islam, D.Inkpen, Semantic text similarity using corpus-based word similarity and string similarity *ACM Transactions on Knowledge Discovery from Data*, **2**(2008), no. 2, 1–25.

[18] J.C.Kelly, Bitopological spaces, *Proc. London Math. Soc.*, **13**(1963), no. 3, 71–89.

[19] M.Kleindessner, U. von Luxburg, Uniqueness of Ordinal Embedding JMLR: Workshop and Conference Proceedings, vol. 35, 1–28, 2014.

[20] L.Leydesdorff, L.Vaughan, Co-occurrence matrices and their applications in information science: extending ACA to the web environment, *Journal of the American Society for Information Science and Technology*, **57**(2006), no. 12, 1616–1628.

[21] S.Lim, Cleansing Noisy City Names in Spatial Data Mining. 2010 International Conference on Information Science and Applications (ICISA), 2010.

[22] S.G.Matthews, Partial metric topology, in: Proc. 8th Summer Conference on General Topology and Applications, *Ann. New York Acad. Sci.*, **728**(1994), 183–197.

[23] E.Mêgnigbêto, Controversies arising from which similarity measures can be used in co-citation analysis, *Malaysian Journal of Library* & *Information Science*, **18**(2013), no. 2, 25–31.

[24] G.Navarro, R.Paredes, N.Reyes, C.Bustos, An empirical evaluation of intrinsic dimension estimators, *Information Systems*, **64**(2017), 206–218.

[25] V.W.Niemytzki, On the "third axiom of metric space" , *Trans. Amer. Math. Soc.*, **29**(1927), 507–513.

[26] K.Nyirarugira, T.Kim, Stratified gesture recognition using the normalized longest common subsequence with rough sets, In Signal Processing: Image Communication, Vol. 30, 2015, 178-189, ISSN 0923-5965. https://doi.org/10.1016/j.image.2014.10.008.

[27] S.J. O'Neill, Two topologies are better than one, Technical report, University of Warwick, April 1995.

[28] R.N.Shepard, The analysis of proximities: Multidimensional scaling with an unknown distance function. I, *Psychometrika*, **27**(1962), 125–140.

[29] R.N.Shepard, Representation of structure in similarity data: Problems and prospects. *Psychometrika*, **39**(1974), no. 4, 373–422.

[30] T.Skopal, On fast non-metric similarity search by metric access methods, In Proc. 10th International Conference on Extending Database Technology (EDBT'06), LNCS 3896, Springer, 2006, 718–736.

[31] T.Skopal, B.Bustos, On nonmetric similarity search problems in complex domains, *ACM Computing Surveys*, **43**(2011), no. 4, Article 34.

[32] A.Tversky, Features of similarity, *Psychological Review*, **84**(1977), 327–352.

[33] A.Tversky, I.Gati, Similarity, separability and the triangle inequality, *Psychological Review*, **89**(1982), 123–154.

[34] A.Ugon, T.Nicolas, M.Richard, P.Guerin, P.Chansard, C.Demoor, L.Toubiana, (2015). A new approach for cleansing geographical dataset using Levenshtein distance, prior knowledge and contextual information, Studies in health technology and informatics, Vol. 210, 227–229. 10.3233/978-1-61499-512-8-227.

[35] M.Vlachos, G.Kollios, D.Gunopulos, Discovering similar multidimensional trajectories, In Proceedings of the International Conference on Data Engineering, ICDE '02, San Jose, CA, USA, IEEE Computer Society Press, 2002, 673–684.

[36] W.A.Wilson, On quasi-metric spaces, *Am. J. Math.*, **53**(1931), 675–684.

[37] S.V.Znamenskij, Models and axioms for similarity metrics *Programmnye systemy: Theoriya i prilozheniya*, **8**(2017), no. 4, 247–357 (in Russian).

[38] S.Znamenskii, V.Dyachenko, An Alternative Model of the Strings Similarity, Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), CEUR Workshop Proceedings (CEUR-WS.org), 177–183 (in Russian).

От сходства к метрике: система аксиом, монотонные преобразования и метрическая определенность

Сергей В. Знаменский
Институт программных систем им. А. К. Айламазяна РАН
ул. Петра I, 4а, с. Веськово, Ярославская обл., Переславский район, 152021
Россия

Исследуется сохранение порядка преобразованиями произвольной метрики (сходства или расстояния) в метрическое или полуметрическое пространство. Вводится система аксиом, по-новому объединяющая известные обобщения метрик расстояния и метрик сходства, коэффициент корреляции Пирсона и косинус угла между векторами. Сохраняющие порядок (как монотонные, так и стержнево-монотонные) преобразования метрик эквивалентно определяются в различных терминах. Метрическая определенность среди стержнево-монотонных преобразований выпуклых метрических подпространств $\mathbb{R}^n$ и $\mathbb{Z}$ доказывается при условии выпуклости метрики расстояния. Обсуждаются формулы ускоренной монотонной нормализации метрик сходства.

Ключевые слова: метрическое пространство, аксиомы сходства, нормализация сходства, метрическая определённость, длиннейшая общая подпоследовательность