applied sciences

MDPI

Article

The Duality of Similarity and Metric Spaces

Ondřej Rozinek [1,2,*] and Jan Mareš [1,3,*]

[1] Department of Process Control, Faculty of Electrical Engineering and Informatics, University of Pardubice, 530 02 Pardubice, Czech Republic
[2] CEO at Rozinet s.r.o., 533 52 Srch, Czech Republic
[3] Department of Computing and Control Engineering, Faculty of Chemical Engineering, University of Chemistry and Technology, 166 28 Prague, Czech Republic
* Correspondence: ondrej.rozinek@rozinet.net (O.R.); jan.mares@vscht.cz (J.M.)

Abstract: We introduce a new mathematical basis for similarity space. For the first time, we describe the relationship between distance and similarity from set theory. Then, we derive generally valid relations for the conversion between similarity and a metric and vice versa. We present a general solution for the normalization of a given similarity space or metric space. The derived solutions lead to many already used similarity and distance functions, and combine them into a unified theory. The Jaccard coefficient, Tanimoto coefficient, Steinhaus distance, Ruzicka similarity, Gaussian similarity, edit distance and edit similarity satisfy this relationship, which verifies our fundamental theory.

Keywords: similarity metric; similarity space; distance metric; metric space; normalized similarity metric; normalized distance metric; edit distance; edit similarity; Jaccard coefficient; Gaussian similarity

Citation: Rozinek, O.; Mareš, J. The Duality of Similarity and Metric Spaces. *Appl. Sci.* **2021**, *11*, 1910. https://doi.org/10.3390/app11041910

Received: 21 December 2020
Accepted: 18 February 2021
Published: 22 February 2021

1. Introduction

Mathematical spaces have been studied for centuries and belong to the basic mathematical theories, which are used in various real-world applications [1]. In general, a mathematical space is a set of mathematical objects with an associated structure. This structure can be specified by a number of operations on the objects of the set. These operations must satisfy certain axioms of mathematical space. The mathematical construction of metric space and similarity space are based on topological space, and a topological space is based on set theory [2]. Nowadays many research groups all over the world deal with similarity spaces in different research fields, e.g., [3,4].

For readability and to reach a broad audience, we do not treat all the mathematical circumstances and conditions in detail. Rather, we present the main concept and a way to a solution. It would take too much time to grasp all the current theories and consequently there would be no time left for innovations. We refer readers to [3,5–9] for the fundamental concepts and properties of topological spaces and metric spaces (convergence, continuity, completeness, separability, connectedness, compactness, etc.).

Similarity and dissimilarity functions are widely used in many research areas: in information retrieval, data mining, machine learning, cluster analysis and applications in database search, protein sequence comparison and many more. When a dissimilarity function is used, a distance metric is normally required. On the other hand, although similarity functions are used, there is no formally accepted definition of this concept [4]. Therefore, in this article we introduce a formal generalised mathematical theory, with all proofs. The organization of the paper is as follows—in Section 1 we briefly introduce the topic. In Sections 2 and 3 the background of metric space and partial metric space is presented. Sections 4 and 5, revised similarity space and duality theory focus on the authors' contribution in the reformulation and the replenishment of the metrics. Section 6, Application of similarity space, presents the connection between revised metrics and

selected well known coefficients. Section 7, Results, and Section 8, Conclusion, conclude the paper.

2. Metric Space

The theory of metric space is a well defined mathematical concept. Recall the formal definition of a distance metric.

Definition 1 (Metric Space [5]). *Let $\boldsymbol{X}$ be a non-empty set. Then, a function $d\colon \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ is a distance metric if for all subsets $x, y, z \in \boldsymbol{X}$, the following conditions are fulfilled:*

(D1) $d(x,y) = d(y,x)$ (symmetry),
(D2) $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality),
(D3) $d(x,y) = 0 \iff x = y$ (identity of indiscernibles).

A metric space is an ordered pair $(\boldsymbol{X}, d)$.

Given the above three axioms, we also have that $d(x,y) \geq 0$ (non-negativity) for any $x, y \in \boldsymbol{X}$. The definition from [1,4] contains this axiom redundantly, there is a derivation in [9].

3. Partial Metric Space

The concept of a partial metric space is a generalization of metric space. We think that it should be implemented as a superset of the axioms for a metric space, as well as the definition of a similarity space. We can see both as special cases of a partial metric space.

Definition 2 (Partial Metric Space). *Let $\boldsymbol{X}$ be a non-empty set. A partial metric or p-metric function $p\colon \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ is a function such that*

(P1) $p(x,y) = p(y,x)$ (symmetry),
(P2) $p(x,z) + p(y,y) \leq p(x,y) + p(y,z)$ (triangle inequality),
(P3) $p(x,x) = p(x,y) = p(y,y) \iff x = y$ (identity of indiscernibles),
(P4) $p(x,y) \geq 0$ (non-negativity),
(P5) $p(x,y) \geq p(x,x)$ (small self-distance).

A partial metric space is an ordered pair $(\boldsymbol{X}, p)$.

Our introduced partial metric space comes from the original definition [10,11]. Note that it is possible to have $p(x,x) \neq p(y,y)$. The definition of a partial metric allows for the possibility that the self-distance is non-zero. Thus, a metric space can be defined as a partial metric space in which each self-distance is zero, so $p(x,x) = p(y,y) = 0$, and hence the term $p(y,y)$ will disappear in P2. The reason for allowing non-zero self-distances first came with the definition of a similarity metric.

4. Similarity Space Revisited

Especially in the last two decades, the research and development of a formal definition of similarity metric (or similarity space) has begun. The new applications and the purpose of this article call for a general consensus and search for well-defined axiomatic systems and theoretical foundations instead of using a non-intuitive duality with the distance. Based on [1,4] we introduce a modified axiomatic system for a similarity metric, which is in agreement with the current notions but simplified so as to be a minimal axiomatic system.

Definition 3 (Similarity Space). *Given a non-empty set $\boldsymbol{X}$, a function $s\colon \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ is a similarity metric if for all subsets $x, y, z \in \boldsymbol{X}$, it satisfies the following conditions:*

(S1) $s(x,y) = s(y,x)$ (symmetry),
(S2) $s(x,z) + s(y,y) \geq s(x,y) + s(y,z)$ (triangle inequality),
(S3) $s(x,x) = s(x,y) = s(y,y) \iff x = y$ (identity of indiscernibles),
(S4) $s(x,y) \geq 0$ (non-negativity).

A similarity space is an ordered pair $(\boldsymbol{X}, s)$.

Compared to the original system, we have removed the axiom of bounded self-similarity, which can be derived from the remaining axioms.

Theorem 1 (Bounded Self-Similarity)**.** *A similarity metric satisfies* $s(x,y) \leq \frac{s(x,x)+s(y,y)}{2}$.

Proof. Appendix A.1. □

A few issues require attention. The name 'similarity metric' is an already proposed convention. Calling it a 'metric' should be understood in the sense of monotonously decreasing convex transformation of a partial metric or a distance metric that will be shown further in the next section. By this way we avoid misunderstanding.

Unlike D3, $d(x,x) = 0$, the similarity metric has an upper bound in $\frac{s(x,x)+s(y,y)}{2}$ and allows $s(x,x) \neq s(y,y)$. At first sight, this may seem counter-intuitive: x is more (or less) similar to itself than y. In spatial considerations of dissimilarity and distance, this does not arise since $d(x,x) = 0$ for all objects. Similarity depends on the set of common features, and the result is the possibility of non-identical self-similarities. If we interpret such common features as 'description lengths' or 'complexities', unequal self-similarities become quite natural, and if x has more features than y, we have $s(x,x) > s(y,y)$ [1]. For instance, having a German word x = 'Einkommensteuererklärung' (income tax return) and y = 'Steuer' (tax), then $s(x,x) \geq s(y,y)$ when counting common characters or q-grams.

We suggest in addition having non-negativity in S4 because the similarity metric doesn't have a direction—in contrast to a vector, it is a scalar value, and so it doesn't make sense to assign to it a negative sign, similar to non-negativity in a metric space. The same principle should be valid for a similarity metric as a requirement for 'symmetric measurement'. The distance between objects remains the same when we measure a distance from another direction. The second reason follows from measure theory (see Appendix A.6), where there is a non-negativity condition $\mu(\boldsymbol{X}) \geq 0$ for a measure μ on the set $\boldsymbol{X}$.

At first glance, it is not clear how the axiom of the triangle inequality was formed, just let us refer to related Theorem 6.

Theorem 2 (Linear Transformation)**.** *Every positive linear transformation* $T_L : \mathbb{R}^+ \to \mathbb{R}^+$ *of a similarity metric is a similarity metric:*

$$s_L(x,y) = T_L(s(x,y)) = \alpha s(x,y) + \beta \tag{1}$$

where $\alpha, \beta \in \mathbb{R}$ *and* $\alpha > 0, \beta \geq 0$.

Proof. Appendix A.2. □

This theorem allows us to apply any linear standardization or re-scaling without any violations of the axioms. In statistics, there is very often used a standard score

$$X' = \frac{X - \mu_s}{\sigma_s}, \tag{2}$$

where μ_s is the mean and σ_s is the standard deviation. Another example could be taken from min-max feature scaling

$$X' = a + \frac{(X - X_{min})(b - a)}{X_{max} - X_{min}}, \tag{3}$$

where X_{min} denotes the minimum value, and X_{max} the maximum value. All values are re-scaled (normalization) to lie within the range $[a, b]$. When the parameters $a = 0$, $b = 1$ are chosen, then this is a unity-based normalization.

For instance, it should be clear that two errors in a comparison of short strings are more critical than in a comparison of long strings. Therefore, it is necessary in some circumstances to normalize the similarity metric. Until the beginning of this century, no such normalization preserving the metric axioms was known for the edit distance metric. Initially, [12] developed a normalized edit distance metric, with the range $[0, 1]$. It is obvious that for any normalized distance metric $d_n(x, y)$, there is also a normalized similarity metric $s_n(x, y) = 1 - d_n(x, y)$ satisfying Definition 3.

Because this axiomatic system is too general and valid for any unnormalized similarity metric functions, we introduce for our case a new specific axiomatic system for a normalized similarity metric in the range $[0, 1]$.

Definition 4 (Normalized Similarity Metric)**.** *A function* $s_n(x, y) \colon \boldsymbol{X} \times \boldsymbol{X} \to [0, 1] \subset \mathbb{R}$ *is a normalized similarity metric if, such that for all subsets* $x, y, z \in \boldsymbol{X}$*, it satisfies the following conditions:*

(N1) $s_n(x, y) = s_n(y, x)$ (symmetry),
(N2) $s_n(x, z) + 1 \geq s_n(x, y) + s_n(y, z)$ (triangle inequality),
(N3) $s_n(x, y) = 1 \iff x = y$ (identity of indiscernibles)
(N4) $s_n(x, y) \geq 0$ (non-negativity) .

A normalized similarity space is an ordered pair $(\boldsymbol{X}, s_n)$.

We do not relax any axiom compared to Definition 3, but we have created a stricter meaningful special case of that definition by substituting $s_n(x, x) = 1$, which is also the least upper bound $s_n(x, y)$. Due to this normalization, the self-similarity is always bounded by the same number $s(x, x) = s(y, y) = 1$. The total dissimilarity also defines the greatest lower bound $s_n(x, y) = 0$. The requirements for both limit conditions N3 and N4 thus stretch the similarity metric to its boundaries.

Theorem 3 (Self-Similarity Inequality)**.** *A normalized similarity metric satisfies* $s_n(x, y) \leq 1$.

Proof. Appendix A.3. □

With these properties, we are also connected with probability theory, where we want to ensure that the probability of the similarity is $0 \leq P(x, y) \leq 1$ as well as $0 \leq s_n(x, y) \leq 1$.

Theorem 4 (Convex Combinations)**.** *A convex combination* $T_C \colon \mathbb{R} \to \mathbb{R}$ *of normalized similarity metrics is again a normalized similarity metric:*

$$s_{n_C}(x, y) = T_C(s_n(x, y)) = \sum_{i=1}^{m} \alpha_i s_i = \alpha_1 s_1 + \alpha_2 s_2 + \ldots + \alpha_m s_m \tag{4}$$

where $\sum_{i=1}^{m} \alpha_i = 1$ *and* $0 \leq \alpha_i \leq 1$.

Proof. Appendix A.4. □

This property of convex combinations allows us to assemble different normalized similarity metrics together and obtain again a normalized similarity metric.

5. Duality of Similarity and Metric Space

The relationship between distance and similarity is not obvious, as distance derives from spatial considerations and similarity relations derive from considering common and non-common features [1]. In many cases, distance is used to measure similarity, although this is far from intuitive and it is often a non-trivial task to find such a dual notion. Let us present a transformation of distance into similarity.

Theorem 5 (Duality of Distance and Similarity [13]). *Generally, if a function* $f\colon \mathbb{R} \to [a,b] \subseteq \mathbb{R}$ *is a monotonously decreasing convex function such that* $b = f(0) > 0$ *and* $\lim_{n\to\infty} f(n) = a \geq 0$, *then*

$$s(x,y) = f(d(x,y)). \tag{5}$$

Proof. We refer further to the related proof of Corollary 4. □

The range $[a,b]$ does not have to be a closed set. We might just denote $a = \inf\{s(x,y)\}$ and $b = \sup\{s(x,y)\}$. The condition $a \geq 0$ is introduced in order to preserve the symmetry with the non-negativity of the values of the distance metric $d(x,y)$. Once we allow a convex distortion of a metric space into a similarity space, this is not necessarily an isomorphic (isometric) transformation, and so the distances between points do not have to be preserved. Most importantly, the relative 'distances' (in the meaning of the inverse of partial order) between the points are preserved in accordance with geometric terminology, for instance $d(x,z) \leq d(x,y) \implies s(x,z) \geq s(x,y)$ for any subsets x, y, z.

Theorem 6 (Triangle Inequality of Similarity). *Any decreasing monotonic convex transformation* f *of the triangular inequality of the metric* d *forms a triangular inequality of similarity* s:

$$d(x,z) \leq d(x,y) + d(y,z) \xrightarrow{f} s(x,z) + s(y,y) \geq s(x,y) + s(y,z), \tag{6}$$

Proof. Appendix A.5. □

Hence the condition of monotonously decreasing function preserves the triangle inequality. Because we measure similarities between objects, and not the distance, it is quite arguable that such a distortion would be more suitable. Moreover, many similarity metrics are not related to distance at all, but, conversely, distance is derived in many cases from similarity, for example, the passage from Jaccard similarity to the Jaccard distance [14].

The measure μ (see Appendix A.6) of the symmetric difference of two sets can be considered as a distance between sets, well known as the *distance of Fréchet–Nikodym–Aronszajn*. This distance is a particular case of the distance in the space of Lebesgue integrable functions. In fact, the distance between sets may be treated as the distance between the characteristic functions χ_x and χ_y. These characteristic functions are defined on a set $\boldsymbol{X}$ and indicate membership of an element in the subset x, respectively y. In classical set theory, its value is 1 for all elements of x and 0 for all elements of $\boldsymbol{X}$ not in x. Employing fuzzy set theory, we can give an uncertainty to the membership in the range of real values $\chi \in [0,1]$.

Theorem 7 (Distance between Two Objects). *Let* x, y *be subsets of set* $\mathbf{X}$. *The symmetric difference between two objects is a distance metric.*

$$d(x,y) = \mu(x \triangle y) = \int |\chi_x - \chi_y| d\mu, \tag{7}$$

where $x \triangle y = (x \cup y) \setminus (x \cap y)$ *is the symmetric difference.*

Proof. Appendix A.7. □

For better illustration, let us suppose two Lebesgue measurable sets A, B. Let us imagine that these sets are described by non-negative real-valued functions f, g in Cartesian system $\mathbb{R}^1$ or $\mathbb{R}^2$ (A corresponds to f and B corresponds to g) [15].

The shaded gray area at the top of Figure 1 essentially shows the distance between objects. Then, we can calculate the area between functions f and g that corresponds area between sets A and B. We can compute the areas in the regions $d(A,B) = \mu(A \triangle B) = \int\int |f(x,y) - g(x,y)| dxdy$ and the functions $d(f(x), g(x)) = \mu(f \triangle g) = \int |f(x) - g(x)| dx$.

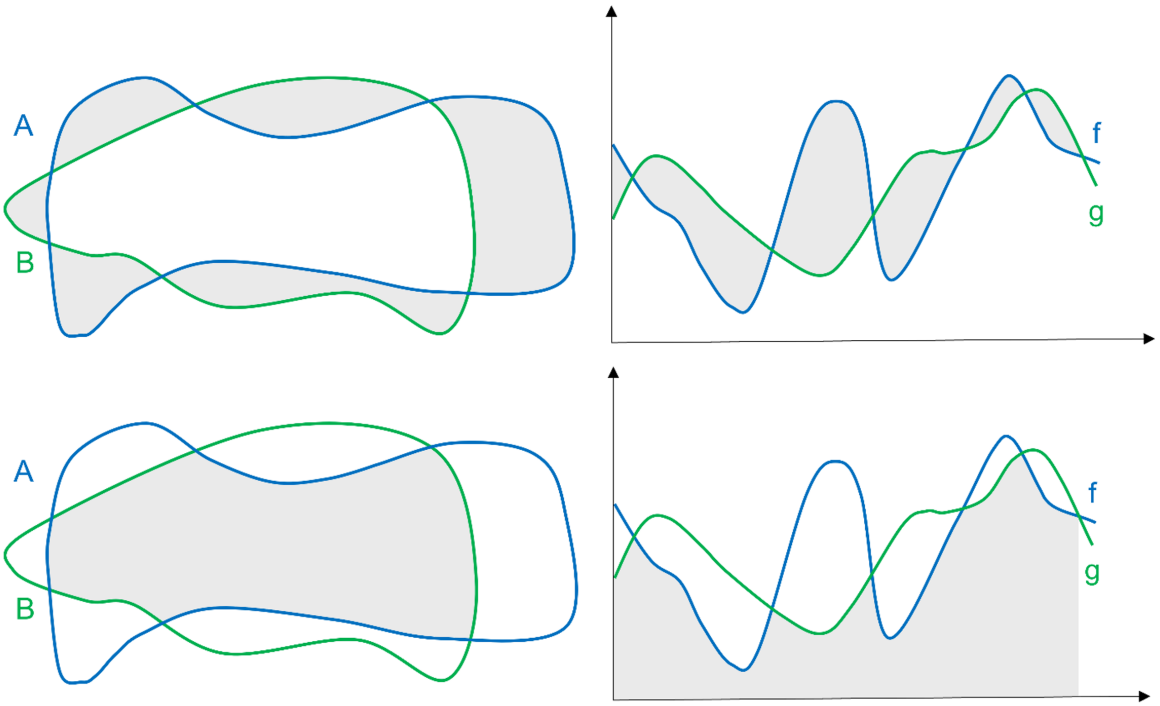

Figure 1. Top—symmetric difference (gray area); bottom—intersection (gray area); left—sets A and B; right—sub-graphs of f and g (inspired by [15]).

Conversely, the shaded gray area at the bottom Figure 1 within overlapping regions A and B and under both graphs f and g represent the similarity between those objects. Analogically, we may deduce a calculation $s(A,B) = \mu(A \cap B) = \int \int \min\{f(x,y), g(x,y)\} dxdy$ and $s(f(x), g(x)) = \mu(f \cap g) = \int \min\{f(x), g(x)\} dx$.

This fundamental observation allows us to create a bridge between set theory and topology, such as the theories of metric spaces and similarity spaces. From the definition of the similarity $s(x,y)$ it can be deduced that the number of features shared between two objects x and y is given by their intersection $\mu(x \cap y)$. The idea behind definition is very simple, direct and intuitive too, assuming that a similarity metric is a measure $s(x,y) = \mu(x \cap y)$.

Theorem 8 (Similarity of Two Objects)**.** *The intersection of two objects represented by subsets x and y is a similarity metric*

$$s(x,y) = \mu(x \cap y) = \int \min\{\chi_x, \chi_y\} d\mu = \frac{\mu(x)+\mu(y)-\mu(x \triangle y)}{2} = \frac{\mu(x)+\mu(y)-d(x,y)}{2}, \quad (8)$$

Proof. Appendix A.8. □

Now we can generalize our knowledge using the similarity axioms.

Corollary 1 (Similarity of Two Objects using Duality)**.** *The similarity metric of two objects given by subsets $x, y \in \mathbf{X}$ is expressed*

$$s(x,y) = \frac{s(x,x) + s(y,y) - d(x,y)}{2}, \quad (9)$$

Proof. Appendix A.9. □

As a result from the proof, self-similarity is equivalent to a measure on set $\mu(x)$, e.g., cardinality of a countable set, $s(x,x) = |x|$, respectively $s(y,y) = |y|$. We can go back to the distance metric from the similarity metric, too.

Corollary 2 (Distance between Two Objects using Duality). *The distance metric applied to two objects defined by subsets $x, y \in \mathbf{X}$ is given by*

$$d(x,y) = s(x,x) + s(y,y) - 2s(x,y), \tag{10}$$

Proof. Expressing $d(x,y)$ from Corollary 1. □

Corollary 3 (Total Dissimilarity using Duality). *The total dissimilarity between two objects is given*

$$s(x,y) = \mu(x \cap y) = 0 \iff \mu(x \triangle y) = \mu(x) + \mu(y) \iff d(x,y) = s(x,x) + s(y,y), \tag{11}$$

Proof. Appendix A.10. □

Total dissimilarity should mean that there are no features shared between the two objects. In set theory, this is equivalent to being a pair of disjoint sets.

Corollary 4 (Duality of Axiomatic Systems). *Consider a similarity space $(\boldsymbol{X}, s)$ and a metric space $(\boldsymbol{X}, d)$. We can define a similarity s on $\boldsymbol{X}$, dual to metric d, vice versa a distance metric d on $\boldsymbol{X}$, dual to similarity s, as follows:*

$$\begin{aligned} s(x,y) &= f(d(x,y)) = \frac{s(x,x) + s(y,y) - d(x,y)}{2} \text{ by Corollary 1} \\ d(x,y) &= f^{-1}(d(x,y)) = s(x,x) + s(y,y) - 2s(x,y) \text{ by Corollary 2,} \end{aligned} \tag{12}$$

Proof. Appendix A.11. □

Comparing similarity axiom system with the partial metrics from Definition 2, we can see the relation $p(x,y) = f^{-1}(d(x,y)) = s(x,x) + s(y,y) - 2s(x,y)$ with dependency on Corollary 2 and Corollary 4 which differs from the source [16].

Theorem 9 (Rozinek Similarity). *Rozinek similarity is a normalized similarity metric*

$$R(x,y) = \frac{\mu(x) + \mu(y) - \mu(x \triangle y)}{\mu(x) + \mu(y) + \mu(x \triangle y)} = \frac{\mu(x) + \mu(y) - d(x,y)}{\mu(x) + \mu(y) + d(x,y)}, \tag{13}$$

Proof. Appendix A.12. □

Theorem 10 (Generalized Rozinek Similarity). *Generalized Rozinek similarity is a normalized similarity metric*

$$R_{GS}(x,y) = \frac{s(x,x) + s(y,y) - d(x,y)}{s(x,x) + s(y,y) + d(x,y)}, \tag{14}$$

Proof. Appendix A.13. □

As has been proved, this similarity forms the bridge between Jaccard similarity (see Theorem 13) and similarity metrics derived from distances. From this equation one can deduce that
$\mu(x \cup y) = \frac{s(x,x)+s(y,y)+d(x,y)}{2}$.

We can again return from a normalized similarity metric to a normalized similarity distance.

Theorem 11 (Generalized Rozinek Normalized Distance). *Generalized Rozinek normalized distance is the following normalized distance metric*

$$R_{GD_n} = \frac{2d(x,y)}{s(x,x) + s(y,y) + d(x,y)}, \tag{15}$$

Proof. Appendix A.14. □

Theorem 12 (Generalized Rozinek distance)**.** *Generalized Rozinek distance is the distance metric*

$$R_{GD}(x,y) = \frac{s(x,x) + s(y,y) - s(x,x)s(x,y) - s(y,y)s(x,y)}{s(x,y) + 1}, \tag{16}$$

Proof. Appendix A.15. □

6. Applications of Similarity Spaces

We will show how this is connected to some already well-known coefficients from previous developments of similarity space theory.

Theorem 13 (Jaccard Similarity)**.** *Jaccard similarity is a normalized similarity metric*

$$J_S(x,y) = \frac{\mu(x \cap y)}{\mu(x \cup y)}, \tag{17}$$

Proof. Appendix A.16. □

The Jaccard similarity is a fundamental similarity measure on sets. Whenever it is used, it is called mainly an index or a coefficient, but it is never called a proper similarity metric. Note that the nonexistence of a mathematical foundation on similarity metrics imposes the necessity of transforming the Jaccard index into the Jaccard distance $J_D(x,y) = 1 - J_S(x,y)$ and then verifying the triangle inequality $J_D(x,z) \leq J_D(x,y) + J_D(y,z)$ for that distance [14].

Theorem 14 (Jaccard Distance)**.** *The Jaccard distance is a normalized distance metric*

$$J_D(x,y) = 1 - \frac{\mu(x \cap y)}{\mu(x \cup y)} = \frac{\mu(x \triangle y)}{\mu(x \cup y)} \tag{18}$$

Proof. [14]. □

Theorem 15 (Tanimoto Coefficient)**.** *The Tanimoto coefficient is a generalized Rozinek similarity*

$$R_{GS}(x,y) = S(x,y) = \frac{s(x,y)}{s(x,x) + s(y,y) - s(x,y)}, \tag{19}$$

Proof. Appendix A.17. □

Theorem 16 (Steinhaus Distance [17])**.** *Steinhaus distance is a generalized Rozinek normalized distance*

$$R_{GS}(x,y) = \sigma_\mu(f,g) = \frac{\int |f(x) - g(x)| d\mu(x)}{\int \max\{|f(x)|, |g(x)|, |f(x) - g(x)|\} d\mu(x)}, \tag{20}$$

Proof. Appendix A.18. □

Theorem 17 (Ruzicka Similarity)**.** *Ruzicka similarity [3] (generalized Jaccard similarity [18]) is a generalized Rozinek similarity*

$$R_{GS}(x,y) = J_G(x,y) = \frac{\sum_k \min\{x_k, y_k\}}{\sum_k \max\{x_k, y_k\}}, \tag{21}$$

Proof. Appendix A.19. □

The distance derived from the Ruzicka similarity $d_n(x,y) = 1 - J_G(x,y)$ is known by the name 'Soergel distance' [3].

Theorem 18 (Gaussian Similarity)**.** *Gaussian similarity is a similarity metric*

$$s(x,y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-y)^2}{2\sigma^2}\right\} \approx \exp\{-(x-y)^2\} = \exp\{-d(x,y)^2\}, \tag{22}$$

Proof. Appendix A.20. □

Gaussian similarity is relevant to the natural human and animal perception of similarity based on psychological research [19], where it is shown that a stimulus decays exponentially with the distance. Numerous experiments have provided empirical observations of learned responses to some measure of different stimuli. As the independent variable of a physical measure of the difference between two stimuli, there have been chosen, for example, the difference in wavelengths of light, frequencies of tones or angular orientations of shapes.

Theorem 19 (Rozinek Natural Distance)**.** *The Rozinek natural distance is a distance metric*

$$R_{ND}(x,y) = \sigma\sqrt{-2\ln\left(\sigma\sqrt{2\pi}s(x,y)\right)} \approx \sqrt{-\ln(s(x,y))}, \tag{23}$$

Proof. Expressing $d(x,y)$ from Theorem 18. □

This distance is derived from Gaussian similarity and describes an inverse problem of how human and animal perception treats a distance depending on the similarity. In addition, there is a limit $\lim_{s(x,y)\to 0^+} R_{ND}(x,y) = +\infty$.

In cases where the similarity measurement is only dependent on the distance and Jaccard-like similarities cannot be used directly, for example, for an edit distance—also called the *k difference problem* [20]—our similarity metric is very appropriate. We can see an analogy between the k difference problem and the symmetric difference set $x \triangle y$ in set theory.

Theorem 20 (Normalized Edit Similarity)**.** *The normalized edit similarity is a Rozinek similarity over the alphabet* Σ

$$s_n(x,y) = \frac{|x| + |y| - d(x,y)}{|x| + |y| + d(x,y)}, \tag{24}$$

where $d(x,y)$ *is an edit distance.*

Proof. Appendix A.21. □

The normalized edit similarity is suitable for conversion from Levenshtein distance or the normalization of the longest common subsequence (LCS). The procedure is shown in proofs.

Definition 5 (Disjoint Strings)**.** *Let* Σ *be a finite alphabet, and let* Σ^* *denote the set of all finite strings over* Σ*. Given* x *and* x^d *any two strings in* Σ^**, we say that they are disjoint strings if they have no character or symbol in common*

$$x \cap x^d = \emptyset. \tag{25}$$

From the meaning of similarity, we should measure the amount of features shared by two objects. If they have no common features, they are disjoint from each other in their features. Indeed, between two strings, there are no common features if they have no

common symbols or letters, hence this is the maximum possible dissimilarity (or, dually, the minimum possible similarity).

Theorem 21 (Total Dissimilarity of Strings)**.** *The total dissimilarity of strings x and y over the alphabet Σ means*

$$s(x, x^d) = 0, \tag{26}$$

Proof. Trivial. □

The derived property of total dissimilarity results from the property of the set of the alphabet Σ, which contains distinguishable different objects, namely its letters or symbols. We should always have a textual similarity of zero if two strings x and y have no character or symbol in common. Take, for example, $x = "abc"$ and $y = "def"$—it doesn't make much sense for them to have a positive string similarity $s(x, y) > 0$.

7. Results

We have proposed a formal generalized mathematical theory of space similarity and similarity functions. General relations for converting a metric to a similarity were derived and general solutions for the normalization of a given similarity space or metric space were introduced. All proofs are attached as appendices. The highlights of the presented concepts are as follows:

- Development of a new revised theory of similarity space.
- The main contribution is a direct explanation and unified theory of the duality between a similarity space (similarity coefficients) and a metric space. Similarity spaces are as important as metric spaces, and should be used wherever similarity measurements are used, avoiding the confused notion of a dual to the distance.
- New Rozinek similarities and distances, using the duality between similarity spaces and metric spaces, have been derived on the basis of set theory. In principle, they are equivalent to a measure of set intersection or Jaccard similarity. This point of view has a general application in transforming any distance metric into a similarity metric, and back to distance metric.

8. Conclusions

Similarity functions are used in different areas of research, from data mining to protein sequence comparison. This paper introduced a generalized mathematical theory of similarity space, which leads to many already used similarity and distance functions.

The main novelty of the approach is an explanation of unified theory between similarity and metric spaces. From this unified theory, it is possible to derive all the widely used functions, such as the Jaccard coefficient, Tanimoto coefficient, Steinhaus distance, Ruzicka similarity, Soergel distance, Gaussian similarity, edit distance and edit similarity.

Moreover, we introduced new Rozinek similarity metrics and distance metrics based on set intersection, Jaccard-like coefficients, the Gaussian function and edit similarity. The novelty and benefit of Rozinek metrics is an easy way to transform any distance metric into a similarity metric, and vice versa.

Future Work

In our future work, we will mainly focus on:

- Development of a novel definition of a space based on elementary mathematical particles from set theory. It is possible to imagine them as "basic" particles that are indivisible in the space and the space is built on them. Therefore, we should be able to describe some relationships with a wider area of applications.
- Verification selected conjectures about elementary space particles.
- Development applications to approximate string search in a cloud environment and databases for customers worldwide.

Author Contributions: O.R.: Development of revised similarity metrics. J.M.: Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by an SGS grant from the Faculty of Electrical Engineering and Informatics, University of Pardubice, Czech Republic. This support is very gratefully acknowledged.

Acknowledgments: We would like to thank to the software technology company Rozinet s.r.o. (www.rozinet.net, accessed on 21/02/2021) without this support, this article would not have been possible to create by providing technical resources, data and space for research. Further we would like to express to Pavel Kříž for his help in technical aspects of mathematical language.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Proof of Theorem 1

Assuming $z = x$,

$$\begin{aligned} &s(x,x) + s(y,y) \geq s(x,y) + s(y,x) \text{ by triangle inequality} \\ &s(x,x) + s(y,y) \geq 2s(x,y) \text{ by symmetry} \\ &s(x,y) \leq \frac{s(x,x) + s(y,y)}{2} \end{aligned} \tag{A1}$$

Appendix A.2. Proof of Theorem 2

Let $\bar{s}(x,y)$ be a positive linear transformation of $s(x,y)$ such that $\bar{s}(x,y) = \alpha s(x,y) + \beta$ for $\alpha > 0$ and $\beta \geq 0$.

S1. By symmetry by multiplication α and adding β

$s(x,y) = s(y,x) \implies \alpha s(x,y) + \beta = \alpha s(y,x) + \beta \implies \bar{s}(x,y) = \bar{s}(y,x)$

S2. By the triangle inequality, we obtain

$$\begin{aligned} &s(x,z) + s(y,y) \geq s(x,y) + s(y,z) \\ &\alpha(s(x,z) + s(y,y)) + 2\beta \geq \alpha(s(x,y) + s(y,z)) + 2\beta \\ &\bar{s}(x,z) + \bar{s}(y,y) \geq \bar{s}(x,y) + \bar{s}(y,z) \end{aligned} \tag{A2}$$

By multiplication by α and adding 2β, the proof is complete. We proceed similarly in cases S3 and S4.

Appendix A.3. Proof of Theorem 3

By identity of indiscernibles (N3) is an upper bound given

$$s_n(x,y) = s_n(y,y) = s_n(x,x) = 1 \tag{A3}$$

and Theorem 1 implies

$$s_n(x,y) \leq \frac{s_n(x,x) + s_n(y,y)}{2} = 1 \tag{A4}$$

Appendix A.4. Proof of Theorem 4

Just we continue to prove for each axiom, assuming $\bar{s}(x,y) = \sum_{i=1}^{m} \alpha_i s_i(x_i, y_i)$,

N1. Obviously,

$$\begin{aligned} &\sum_{i=1}^{m} \alpha_i s_i(x_i, y_i) = \sum_{i=1}^{m} \alpha_i s_i(y_i, x_i) \\ &\bar{s}(x,y) = \bar{s}(y,x) \end{aligned} \tag{A5}$$

Similarly, N2, N3 and N4 are trivial.

Appendix A.5. Proof of Theorem 6

A real-valued function $f(d)$ is said to be convex over the interval $[a,b] \in \mathbb{R}$ if for any $d_1, d_2 \in [a,b]$ and any $\lambda \in [0,1]$ we have that

$$\lambda f(d_1) + (1-\lambda) f(d_2) \geq f(\lambda d_1 + (1-\lambda) d_2) \tag{A6}$$

The validity of the triangle inequality $s(x,z) + s(y,y) \geq s(x,y) + s(y,z)$ can be proven from the dual notion of distance $s(x,y) = f(d(x,y))$ by applying $d(x,z) \leq d(x,y) + d(y,z)$ and considering possible cases as follows [13].

Case 1: $d(x,z) \leq d(x,y)$

Thus we get $f(d(x,z)) \geq f(d(x,y))$. As $0 \leq d(y,z)$, we have $f(0) = f(d(y,y)) \geq f(d(y,z))$. We sum both expressions

$$f(d(y,y)) + f(d(x,z)) \geq f(d(x,y)) + f(d(y,z)) \tag{A7}$$

So the claim is proven.

Case 2: $d(x,z) \leq d(y,z)$

The reasoning is analogous to the above, just flipping x and z.

Case 3: $d(x,z) > d(x,y) \wedge d(x,z) > d(y,z)$

As a metric is assumed, $d(x,z) \leq d(x,y) + d(y,z)$. Hence

$$1 \leq \frac{d(y,z)}{d(x,z)} + \frac{d(x,y)}{d(x,z)} \implies 0 \leq 1 - \frac{d(y,z)}{d(x,z)} \leq \frac{d(x,y)}{d(x,z)} \leq 1 \tag{A8}$$

Let us pick any $\lambda \in \left[0, \frac{d(x,y)}{d(x,z)}\right]$ such that $1 - \frac{d(y,z)}{d(x,z)} \leq \lambda \leq \frac{d(x,y)}{d(x,z)}$. Obviously $0 \leq \lambda \leq 1$. We see immediately that $\lambda d(x,z) \leq d(x,y)$ and $(1-\lambda) d(x,z) \leq d(y,z)$. From the definition of convexity, we have that

$$(1-\lambda) f(0) + \lambda f(d(x,z)) \geq f((1-\lambda)0 + \lambda d(x,z)) \tag{A9}$$

$$f(\lambda d(x,z)) \geq f(d(x,y)) \tag{A10}$$

with the last inequality being due to the fact that f is monotonic decreasing. Similarly

$$\lambda f(0) + (1-\lambda) f(d(x,z)) \geq f(\lambda 0 + (1-\lambda) d(x,z)) \tag{A11}$$

$$f((1-\lambda) d(x,z)) \geq f(d(y,z)) \tag{A12}$$

By summing all inequalities (Appendix A9–A12) by transitivity $\geq$ we get

$$f(0) + f(d(x,z)) \geq f(d(x,y)) + f(d(y,z)) \tag{A13}$$

so obviously the triangle inequality holds here too.

Appendix A.6. Measure Space Definition

A measurable space is a set $\mathbf{X}$ and σ-ring $\mathbf{S}$ of subsets of $\mathbf{X}$ with the property that $\bigcup \mathbf{S} = \mathbf{X}$. A measure is an extended real valued, non-negative, and countably additive set function μ, defined on a σ-ring $\mathbf{S}$, and such that $\mu(0) = 0$. An ordered triple $(\mathbf{X}, \mathbf{S}, \mu)$ is called a measure space.

The meaning of this definition lies in the abstraction of measurement on countable set given by cardinality or on Lebesgue measurable set. For more details we refer the readers to the source [21].

Appendix A.7. Proof of Theorem 7

We must show that a distance equals the symmetric difference of two sets $d(x,y) = \mu(x \triangle y)$ [21–23]. If μ is a σ-finite measure on a σ-ring $\mathbf{S}$, this function is pseudometric on $\mathbf{S}$ (D1 and D2 must be satisfied), assuming $x, y, z \in \mathbf{S}$

$$\begin{aligned} d(x,z) &= \mu(x \triangle z) = \mu(z \triangle x) = \mu((x \triangle y) \triangle (y \triangle z)) \\ &\leq \mu((x \triangle y) \cup (y \triangle z)) \\ &\leq \mu(x \triangle y) + \mu(y \triangle z) \\ &= d(x,y) + d(y,z) \end{aligned} \tag{A14}$$

The relation (D3) $x \sim y \iff d(x,y) = 0$ is an equivalence relation on $\mathbf{S}$, so d becomes a metric on the set $\mathbf{S}$. We need also to prove sequential continuity with Cauchy sequence, i.e., $\{x_n\}_{n \in \mathbb{N}_0}$,

$$\lim_{n \to \infty} d(x_n, x) = 0 \implies \lim_{n \to \infty} |\mu(x_n) - \mu(x)| = 0 \tag{A15}$$

This implies that

$$\begin{aligned} d(x,y) &= |\mu(x) - \mu(y)| = |(\mu(x \setminus y) + \mu(x \cap y)) - (\mu(x \cap y) + \mu(y \setminus x))| \\ &= |\mu(x \setminus y) - \mu(y \setminus x)| \\ &\leq |\mu(x \setminus y)| + |\mu(y \setminus x)| = \mu(x \setminus y) + \mu(y \setminus x) = \mu(x \setminus y) \cup \mu(y \setminus x) \\ &= \mu(x \triangle y) = \int |\chi_x - \chi_y| d\mu \end{aligned} \tag{A16}$$

We call d the symmetric difference metric. The symmetric difference between two sets can be considered a measure of how 'far apart' they are.

Appendix A.8. Proof of Theorem 8

At first, we will prove the relation for the intersection of the two objects.

$$\begin{aligned} s(x,y) = \mu(x \cap y) &= \mu(x) + \mu(y) - \mu(x \cup y) = \frac{2(\mu(x) + \mu(y) - \mu(x \cup y))}{2} \\ &= \frac{\mu(x) + \mu(x) + \mu(y) + \mu(y) - (\mu(x) + \mu(y) - \mu(x \cap y)) - \mu(x \cup y)}{2} \\ &= \frac{\mu(x) + \mu(y) + \mu(x \cap y) - \mu(x \cup y)}{2} = \frac{\mu(x) + \mu(y) - \mu(x \triangle y)}{2} \end{aligned} \tag{A17}$$

The conditions S1, S3 and S4 are trivial. We show only S2. Since $y \supseteq (x \cap y) \cup (z \cap y)$, we have

$$\mu(y) \geq \mu(x \cap y) + \mu(z \cap y) - \mu(x \cap z \cap y), \tag{A18}$$

and, consequently,

$$\mu(x \cap z) + \mu(y) \geq \mu(x \cap z \cap y) + \mu(y) \geq \mu(x \cap y) + \mu(z \cap y). \tag{A19}$$

This yields the desired triangle inequality.

Appendix A.9. Proof of Corollary 1

The self-similarity could be derived from Theorem 8

$$s(x,x) = \frac{\mu(x) + \mu(x) - d(x,x)}{2} = \frac{2\mu(x)}{2} = \mu(x) \tag{A20}$$

Similarly we obtain $s(y,y) = \mu(y)$. We substitute these terms into

$$s(x,y) = \frac{\mu(x) + \mu(y) - d(x,y)}{2} = \frac{s(x,x) + s(y,y) - d(x,y)}{2} \tag{A21}$$

Appendix A.10. Proof of Corollary 3

Let x, y be disjoint subsets of a set $\boldsymbol{X}$. Let be total dissimilarity given by expression $s(x,y) = \mu(x \cap y) = 0$, thus satisfying

$$\begin{aligned} d(x,y) &= \mu(x \triangle y) = \mu((x \setminus y) \cup (y \setminus x)) = \mu(x \cup y) = \mu(x) + \mu(y) \\ &= s(x,x) + s(y,y) \end{aligned} \tag{A22}$$

Appendix A.11. Proof of Corollary 4

Let us show you the duality on this case $d(x,y) = f^{-1}(d(x,y)) = s(x,x) + s(y,y) - 2s(x,y)$ by applying Corollary 2

$D1 \xrightarrow{f^{-1}} S1$

$$\begin{aligned} &d(x,y) = d(y,x) \\ &s(x,x) + s(y,y) - 2s(x,y) = s(x,x) + s(y,y) - 2s(y,x) \\ &s(x,y) = s(y,x) \end{aligned} \tag{A23}$$

$D2 \xrightarrow{f^{-1}} S2$

$$\begin{aligned} &d(x,z) \leq d(x,y) + d(y,z) \\ &s(x,x) + s(z,z) - 2s(x,z) \leq s(x,x) + s(y,y) - 2s(x,y) + s(y,y) + s(z,z) - 2s(y,z) \\ &- 2s(x,z) \leq -2s(x,y) + 2s(y,y) - 2s(y,z) \\ &s(x,z) + s(y,y) \geq s(x,y) + s(y,z) \end{aligned} \tag{A24}$$

so we receive S2 from the Definition 3 and the triangle inequality is proven.

$D3 \xrightarrow{f^{-1}} S3$

$$\begin{aligned} &d(x,y) = 0 \Longrightarrow x = y \\ &s(x,x) + s(y,y) - 2s(x,y) = 0 \Longrightarrow x = y \\ &s(x,y) = s(y,y) = s(x,x) \Longrightarrow x = y \end{aligned} \tag{A25}$$

$D3 \xrightarrow{f^{-1}} S4$

Since $d(x,y) = 0 \Longleftrightarrow x = y$ is bounded by zero at the same axiom by that reason we should add S4 as explained previously for Definition 3.

Similarly, on the opposite we proceed by applying Corollary 1 to transform $s(x,y) = f(d(x,y))$.

Appendix A.12. Proof of Theorem 9

We can proceed according to Appendix A.16 for $J_S(x,y)$ and we will now show the equivalence with the Jaccard similarity as follows.

$$\begin{aligned} J_S(x,y) &= \frac{\mu(x\cap y)}{\mu(x\cup y)} = \frac{2\mu(x\cap y)}{2(\mu(x)+\mu(y)-\mu(x\cap y))} \\ &= \frac{\mu(x)+\mu(y)-\mu(x)-\mu(y)+\mu(x\cap y)+\mu(x\cap y)}{\mu(x)+\mu(y)+\mu(x)+\mu(y)-\mu(x\cap y)-\mu(x\cap y)} \\ &= \frac{\mu(x)+\mu(y)-\mu(x\cup y)+\mu(x\cap y)}{\mu(x)+\mu(y)+\mu(x\cup y)-\mu(x\cap y)} \\ &= \frac{\mu(x)+\mu(y)-\mu((x\cup y)-(x\cap y))}{\mu(x)+\mu(y)+\mu((x\cup y)-(x\cap y))} \\ &= \frac{\mu(x)+\mu(y)-\mu(x\triangle y)}{\mu(x)+\mu(y)+\mu(x\triangle y)} = R(x,y) \end{aligned} \tag{A26}$$

Appendix A.13. Proof of Theorem 10

At first we proceed in proving $R(x,y)$ being a normalized similarity metric in Appendix A.12 then we substitute into equation of Theorem 9

$$\begin{aligned} R(x,y) &= \frac{\mu(x)+\mu(y)-d(x,y)}{\mu(x)+\mu(y)+d(x,y)} \\ &= \frac{\frac{\mu(x)+\mu(x)-d(x,x)}{2}+\frac{\mu(y)+\mu(y)-d(y,y)}{2}-d(x,y)}{\frac{\mu(x)+\mu(x)-d(x,x)}{2}+\frac{\mu(y)+\mu(y)-d(y,y)}{2}+d(x,y)} \\ &= \frac{s(x,x)+s(y,y)-d(x,y)}{s(x,x)+s(y,y)+d(x,y)} = R_{GS}(x,y) \end{aligned} \tag{A27}$$

Appendix A.14. Proof of Theorem 11

We express a direct relationship between the Jaccard distance (Theorem 14) and the generalized Rozinek normalized distance

$$\begin{aligned} J_D(x,y) &= \frac{\mu(x\triangle y)}{\mu(x\cup y)} = \frac{d(x,y)}{\mu(x\cup y)} = \frac{d(x,y)}{\frac{\mu(x)+\mu(y)+d(x,y)}{2}} \\ &= \frac{2d(x,y)}{\mu(x)+\mu(y)+d(x,y)} = \frac{2d(x,y)}{s(x,x)+s(y,y)+d(x,y)} \\ &= R_{GD_n}(x,y) \end{aligned} \tag{A28}$$

Obviously, conditions D1 and D3 are satisfied. We refer to [14] for D2.

Appendix A.15. Proof of Theorem 12

From Theorem 10 we can write $d(x,y)$ as follows

$$\begin{aligned} & R_{GS}(x,y) = s(x,y) = \frac{s(x,x)+s(y,y)-d(x,y)}{s(x,x)+s(y,y)+d(x,y)} \\ \Longrightarrow\ & s(x,y)(s(x,x)+s(y,y)+d(x,y)) = s(x,x)+s(y,y)-d(x,y) \\ \Longrightarrow\ & s(x,x)s(x,y)+s(y,y)s(x,y)+d(x,y)s(x,y)+d(x,y) = s(x,x)+s(y,y) \\ \Longrightarrow\ & d(x,y)(s(x,y)+1) = s(x,x)+s(y,y)-s(x,x)s(x,y)-s(y,y)s(x,y) \\ \Longrightarrow\ & d(x,y) = \frac{s(x,x)+s(y,y)-s(x,x)s(x,y)-s(y,y)s(x,y)}{s(x,y)+1} = R_{GD} \end{aligned} \tag{A29}$$

From the previously proven theorems, it is obvious that $d(x,y) = \mu(x \triangle y)$ satisfies the axioms for being a distance metric.

Appendix A.16. Proof of Theorem 13

N1. Trivial.

N2. Since we know that $s_n(x,y) = 1 - d_n(x,y)$, we modify theorem 3 of [14] in the dual form of Jaccard similarity instead of Jaccard distance. Then, for all sets $x, y, z \in \boldsymbol{X}$, from Definition 4 one has

$$J_S(x,z) + 1 \geq J_S(x,y) + J_S(y,z) \tag{A30}$$

Let f be a nonnegative, monotone, modular set function on $\boldsymbol{X}$. Say that a set x is a null set if $f(x) = 0$. Observe that if at least one of the sets is a null set, then the inequality is satisfied. So, it is enough to show the equivalent inequality

$$\frac{f(x \cap z)}{f(x \cup z)} + 1 \geq \frac{f(x \cap y)}{f(x \cup y)} + \frac{f(y \cap z)}{f(y \cup z)} \tag{A31}$$

for arbitrary non-null sets $x, y, z \subseteq \boldsymbol{X}$. For more details of proof we refer the readers to [14].

N3. If $x = y \iff \mu(x \cap y) = \mu(x \cup y) \iff J_S(x,y) = 1$.

N4. Let x^d be any disjoint set to x. Then $\mu(x \cap x^d) = \emptyset \iff J_S(x,y) = 0$.

Appendix A.17. Proof of Theorem 15

Substituting $s(x,y)$ into the Tanimoto coefficient we get

$$\begin{aligned} S(x,y) &= \frac{s(x,y)}{s(x,x) + s(y,y) - s(x,y)} = \frac{\frac{s(x,x)+s(y,y)-d(x,y)}{2}}{s(x,x) + s(y,y) - (\frac{s(x,x)+s(y,y)-d(x,y)}{2})} \\ &= \frac{s(x,x) + s(y,y) - d(x,y)}{2} \frac{2}{s(x,x) + s(y,y) + d(x,y)} \\ &= R_{GS}(x,y) \end{aligned} \tag{A32}$$

Appendix A.18. Proof of Theorem 16

Let A, B be Lebesgue measurable sets. Then we can write [17]

$$\begin{aligned} \mu(A \triangle B) &= \int |\chi_A(x) - \chi_B(x)| d\mu(x) = \int |f(x) - g(x)| d\mu(x) \\ \mu(A \cup B) &= \int \max\{\chi_A(x), \chi_B(x)\} d\mu(x) = \int \max\{|f(x)|, |g(x)|, |f(x) - g(x)|\} \end{aligned} \tag{A33}$$

We obtain, by substitution of $\mu(A \triangle B)$ and $\mu(A \cup B)$,

$$\begin{aligned} J_D(x,y) &= \frac{\mu(A \triangle B)}{\mu(A \cup B)} = \frac{\int |f(x) - g(x)| d\mu(x)}{\int \max\{|f(x)|, |g(x)|, |f(x) - g(x)|\}} \\ &= \frac{2d(x,y)}{s(x,x) + s(y,y) + d(x,y)} \end{aligned} \tag{A34}$$

Appendix A.19. Proof of Theorem 17

For the proof, we apply the characteric functions χ_x, χ_y and non-negative real valued functions $f(x), g(x)$:

$$\begin{aligned} J_S(x,y) &= \frac{\mu(x \cap y)}{\mu(x \cup y)} = \frac{\int \min\{\chi_x, \chi_y\} d\mu(x)}{\int \max\{\chi_x, \chi_y\} d\mu(x)} = \frac{\int \min\{f(x), g(x)\} d\mu(x)}{\int \max\{f(x), g(x)\} d\mu(x)} \\ &= \lim_{\delta x \to 0} \frac{\sum \min\{f(x), g(x)\} \delta x}{\sum \max\{f(x), g(x)\} \delta x} = J_G(x,y) \end{aligned} \tag{A35}$$

In the last step, we discretized the continous functions where δ may be chosen as a sampling length. The relation of $J_S(x,y)$ to $R_{GS}(x,y)$ is derived in Appendix A.13.

Appendix A.20. Proof of Theorem 18

The properties S1, S3 and S4 are trivial. S2 is satisfied as follows

$$\begin{aligned}
&\left(1-\exp\{-d(x,y)^2\}\right)\left(1-\exp\{-d(y,z)^2\}\right) \geq 0\\
&\exp\{-d(x,y)^2-d(y,z)^2\}+1 \geq \exp\{-d(x,y)^2\}+\exp\{-d(y,z)^2\}\\
&\exp\{-d(x,z)^2\}+1 \geq \exp\{-d(x,y)^2\}+\exp\{-d(y,z)^2\}\\
&s(x,z)+s(y,y) \geq s(x,y)+s(y,z)
\end{aligned} \tag{A36}$$

Appendix A.21. Proof of Theorem 20

Case 1: Levenshtein similarity

$$\begin{aligned}
R(x,y) &= \frac{\mu(x)+\mu(y)-d(x,y)}{\mu(x)+\mu(y)+d(x,y)} = \frac{|x|+|y|-d(x,y)}{|x|+|y|+d(x,y)} = 1-\frac{2d(x,y)}{|x|+|y|+d(x,y)}\\
&= 1-d_{N-GLD}(x,y)
\end{aligned} \tag{A37}$$

where $d_{N-GLD}(x,y)$ is a normalized generalized Levenshtein distance of the form

$$d_{N-GLD}(x,y) = \frac{2d(x,y)}{\alpha(|x|+|y|)+d(x,y)} \tag{A38}$$

where $\alpha = 1$ is the minimum cost of insertion and deletion costs and $d(x,y)$ is an edit distance [24]. This proof has been inspired by [1,12] where it is further proved that d_{N-GLD} is a normalized distance metric. Hence by the duality between normalized similarity metrics and normalized distance metrics, $s_n(x,y) = 1-d_n(x,y)$ is proven.

Case 2: Longest Common Subsequence (LCS)

We obtain the same results when we normalize the LCS. Let l be the length of the LCS [25]

$$l(x,y) = \frac{1}{2}(|x|+|y|-d_{LCS}(x,y)) \tag{A39}$$

where l(x,y) satisfies the similarity axioms from Definition 3 and d_{LCS} denotes the edit distance based on unit insertion and deletation cost [1]. Now we turn our attention to normalizing similarity through evaluating a generalized Tanimoto's coefficient [1,26]

$$S(x,y) = \frac{s(x,y)}{s(x,x)+s(y,y)-s(x,y)} \tag{A40}$$

The $s(x,y)$ is interpreted as a count of common features, while $S(x,y)$ express this count as a fraction of the total number of features of x and y. We set $s(x,y) = l(x,y)$ and hence obtain

$$S(x,y) = \frac{l(x,y)}{|x|+|y|-l(x,y)} \tag{A41}$$

Since $l(x,x) = |x|$ and $l(x,y) = |y|$, we elaborate the above expressions to

$$S(x,y) = \frac{|x|+|y|-d_{OM}(x,y)}{|x|+|y|+d_{OM}(x,y)} \tag{A42}$$

where d_{OM} is an edit distance (for details see [1]). Hence we have proved that also $S(x,y) = R(x,y)$.

References

1. Elzinga, H.S.; Studer, M. Normalization of Distance and Similarity in Sequence Analysis. In Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, Switzerland, 8–10 June 2016; pp. 445–468.
2. Sutherland, W.A. *Introduction to Metric & Topological Spaces*; Oxford University Press: Oxford, UK, 2009.
3. Deza, M.M.; Deza, E. *Encyclopedia of Distances*, 4th ed.; Springer: Berlin, Germany, 2016.
4. Chen, S.; Ma, B.; Zhang, K. On the Similarity Metric and the Distance Metric. *Theor. Comput. Sci.* **2009**, *410*, 2365–2376. [CrossRef] [CrossRef]
5. Muscat, J. *Functional Analysis: An Introduction to Metric Spaces, Hilbert Spaces, and Banach Algebras*; Springer: Berlin, Germany, 2014.
6. Alabiso, C.; Weiss, I. *A Primer on Hilbert Space Theory: Linear Spaces, Topological Spaces, Metric Spaces, Normed Spaces, and Topological Groups*; Springer: Berlin, Germany, 2015.
7. Garling, D.J.H. *A Course in Mathematical Analysis Volume 2: Metric and Topological Spaces, Functions of a Vector Variable*; Cambridge University Press: Cambridge, UK, 2013.
8. Searcóid, M.Ó. *Metric Spaces*; Springer: Berlin, Germany, 2007.
9. Zorich, A.V. Metric Spaces. In *Mathematical Analysis II*, 2nd ed.; Springer: Berlin, Germany, 2016; pp. 1–8.
10. Matthews, G.S. Partial Metric Topology. *Ann. N. Y. Acad. Sci.* **1994**, *728*, 183–197. [CrossRef]
11. Bukatin, M.; Kopperman, R.; Matthews, S.; Pajoohesh, H. Partial Metric Spaces. *Am. Math. Mon.* **2009**, *116*, 708–718. [CrossRef]
12. Yujian, L; Bo, L. A Normalized Levenshtein Distance Metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1091–1095. [CrossRef] [PubMed]
13. Wierzchoń, T.S.; Kłopotek, M.A. Measures of Similarity/Dissimilarity. In *Modern Algorithms of Cluster Analysis*; Springer: Berlin, Germany, 2018; pp. 16–26.
14. Kosub, S. A note on the triangle inequality for the Jaccard distance. *Pattern Recognit. Lett.* **2019**, *120*, 36–38. [CrossRef]
15. Jourlin, M. Metrics Based on Logarithmic Laws. In *Logarithmic Image Processing: Theory and Applications*; Academic Press: Cambridge, MA, USA, 2016.
16. Znamenskij, V.S. From Similarity to Distance: Axiom Set, Monotonic Transformations and Metric Determinacy. *J. Sib. Fed. Univ.* **2018**, *11*, 331–341.
17. Marczewski, E.; Steinhaus, H. On a certain distance of sets and the corresponding distance of functions. *Colloq. Math.* **1958**, *6*, 319–327. [CrossRef]
18. Wu, W.; Li, B.; Chen, L.; Zhan, C.; Yu, P.S. Improved Consistent Weighted Sampling Revisited. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 2332–2345. [CrossRef]
19. Shepard, R.N. Toward a universal law of generalization for psychological science. *Am. Assoc. Adv. Sci.* **1987**, *237*, 1317–1323. [CrossRef] [PubMed]
20. Jokinen, P.; Ukkonen, E. Two algorithms for approximate string matching in static texts. In Proceedings of the International Symposium on Mathematical Foundations of Computer Science, Kazimierz Dolny, Poland, 9–13 September 1991; pp. 240–248. [CrossRef]
21. Halmos, P.R. *Measure Theory*; Springer: Berlin, Germany, 1950.
22. Conway, B.J. Elements of Measure Theory. In *A Course in Abstract Analysis*; American Mathematical Society: Providence, RI, USA, 2010; pp. 90–91.
23. Bell, J. The Symmetric Difference Metric. Department of Mathematics, University of Toronto. **2015**, unpublished. [CrossRef]
24. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710. [CrossRef]
25. Wagner, R.A.; Fisher, M.J. The String-to-String Correction Problem. *J. ACM* **1974**, *21*, 168–173. [CrossRef]
26. Duda, R.O.; Hart, P.E. Stork. In *Pattern Classification*; Wiley: Chichester, UK, 2001.