# Why $1/(1 + d)$ Is an Effective Distance-Based Similarity Measure: Two Explanations

1st Julio C. Urenda
*Department of Mathematical Sciences and*
*Department of Computer Science*
*University of Texas at El Paso*
El Paso, Texas 79968, USA
jcurenda@utep.edu

2nd Olga Kosheleva
*Department of Teacher Education*
*University of Texas at El Paso*
El Paso, Texas 79968, USA
olgak@utep.edu

3rd Vladik Kreinovich
*Department of Computer Science*
*University of Texas at El Paso*
El Paso, Texas 79968, USA
vladik@utep.edu

***Abstract*—Most of our decisions are based on the notion of similarity: we use a decision that helped in similar situations. From this viewpoint, it is important to have, for each pair of situations or objects, a numerical value describing similarity between them. This is called a similarity measure. In some cases, the only information that we can use to estimate the similarity value is some natural distance measure $d(a, b)$. In many such situations, empirical data shows that the similarity measure $1/(1 + d)$ is very effective. In this paper, we provide two explanations for this effectiveness.**

***Index Terms*—similarity measure, distance, decision making**

## I. Formulation of the Problem

**Need for similarity measures.** Many of our decisions are based on the idea of similarity:

- if some decision was effective in similar situations,
- then it makes sense to apply a similar decision here.

If different decisions were successful in several situations which are somewhat similar to the current one, then we should select a decision corresponding to situations which are the closest to the given situation. To make this selection, we need to be able to decide which pairs of situations are more similar and which are less similar. In other words, we need to have a measure of similarity $s(a, b)$ between two situations $a$ and $b$ (or between two objects $a$ and $b$).

We rarely have an objective measure of similarity. In most cases, similarity is a subjective idea, it describes the expert's feeling. A natural way to describe this degree of similarity $s(a, b)$ is thus to ask the experts to estimate this degree on some scale. A reasonable idea is to use the interval $[0, 1]$ as this scale. This is in line with the fact that, in effect, we are asking experts to estimate to what extent the following statement is true: "$a$ and $b$ are similar". In the computer, "true" is usually represented as 1, and "false" as 0, so it is natural to represent uncertainty by a number between 0 and 1 – as, e.g., in fuzzy logic; see, e.g., [1], [3], [4], [7], [8], [15].

This way, to estimate the degree of similarity $s(a, b)$ between objects $a$ and $b$, we ask an expert to mark his/her degree of similarity between the two objects on a scale of 0 to 1, so that:

- the value $s(a, b) = 1$ means that the objects are perfectly similar, practically indistinguishable;
- the value $s(a, b) = 0$ means that the objects are completely dissimilar, i.e., that they have nothing in common; and
- values strictly between 0 and 1 describe the cases when there is some similarity, but there is some dissimilarity as well.

If experts are not comfortable providing numerical estimates of their degree of similarity, and they can only give us binary answers: similar or not similar – then we can ask several ($n$) experts this question, and if $m$ of them answer that the objects are similar, use the ratio

$$\frac{m}{n}$$

as the desired degree of similarity.

**Need for metric-based similarity measures.** In many practical situations, we have a large number of possible objects and situations, and it is not feasible to ask the experts about each possible pair. What can we do?

Often, we have a naturally defined metric $d(a, b)$ on the class $S$ of some objects, i.e., a function $d : S \times S \to [0, \infty)$ that satisfies the usual properties:

- $d(a, b) = 0$ if and only if $a = b$,
- $d(a, b) = d(b, a)$ for all $a$ and $b$, and
- $d(a, c) \le d(a, b) + d(b, c)$ for all $a$, $b$, and $c$.

This metric describes to what extent the two objects are dissimilar.

Thus, a natural idea is to estimate the desired degree of similarity $s(a, b)$ between the two objects based on this metric, as $s(a, b) = f(d(a, b))$ for some function $f(d)$.

Which function $f(d)$ should we choose?

**Natural properties of the transformation $f(d)$.** The degree of similarity must satisfy the following natural properties:

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology. It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

- the degree of similarity $s(a,b)$ should attain its largest value $s(a,b)=1$ if the objects are identical, i.e., if $d(a,b)=0$; thus, we must have $f(0)=1$;
- the larger the distance between the objects, the smaller the similarity between them; thus, the function $f(d)$ should be strictly decreasing: if $d<d'$, then we should have
$$f(d)>f(d');$$
- in the limit, when the objects are as far away from each other as possible, the resulting degree of similarity should be close to 0; in other words, as $d\to\infty$, we should have
$$f(d)\to 0.$$

There are many functions $f(d)$ that satisfy these three properties. Which one should we choose?

**Empirical fact: an efficient transformation.** In many practical applications, the following transformation leads to a reasonable description of similarity – that fits expert opinions well (see, e.g., [11]):

$$f(d)=\frac{1}{1+d}. \qquad (1)$$

A natural question is: why this transformation works well?

**What we do in this paper.** In this paper, we provide two explanations of this empirical success. The fact that two different explanations lead to the same formula increases our confidence in both explanations.

## II. First Explanation

**Need to make expert estimates more accurate.** When the degree of similarity comes from a poll of $n$ experts, we only get $n+1$ possible degrees:

$$0,\frac{1}{n},\frac{2}{n},\ldots,\frac{n-1}{n},1.$$

When $n$ is small, these values provide a rather crude description of the actual degree of similarity.

Thus, a natural way to increase the accuracy of the estimate is to ask more experts. This is similar to statistics, where we can estimate the probability of an event by taking the ratio $m/n$ between the general number of situations $n$ and the number of cases $m$ in which this event was observed. In statistics, the larger the sample size $n$, the more accurate this estimation of the probability; see, e.g., [12].

**Resulting problem.** To make our estimate more accurate, we ask the more knowledgeable experts. So, at first, we asked $n$ top experts. Then, to increase the accuracy, we ask $n'$ additional experts. These additional experts may be intimidated by the opinion of the top experts. This intimidation may be described in two ways:

- Additional experts may be unwilling to say anything: if top experts are disagreeing, who are we to voice our humble opinions? In this case, out of $n+n'$ experts, we still have the same number $m$ of experts who answer that the objects $a$ and $b$ are similar. Thus, instead of the original degree of similarity
$$s=\frac{m}{n},$$
we have a new degree
$$s'=\frac{m}{n+n'}.$$
One can easily see that the new degree $s'$ can be obtained from the original degree by a transformation
$$s'=c_1\cdot s, \qquad (2)$$
where we denoted
$$c_1\overset{\text{def}}{=}\frac{n}{n+n'}.$$
- Alternatively, additional experts can simply side with the majority. We are looking for cases when there *is* a similarity – in this case, we can use this similarity to make a decision – so let us consider the case when originally, the majority of experts believed that the objects are similar. In this case, now we have $m+n'$ experts who answer that the given objects $a$ and $b$ are similar. Thus, instead of the original degree of similarity
$$s=\frac{m}{n},$$
we have a new degree
$$s'=\frac{m+n'}{n+n'}.$$
One can easily see that the new degree $s'$ can be obtained from the original degree by a transformation
$$s'=c_1\cdot s+c_2, \qquad (3)$$
where
$$c_2\overset{\text{def}}{=}\frac{n'}{n+n'}.$$

In both cases, we have linear transformations (2) and (3) between different scales, i.e., linear functions $s'=g(s)$.

**This is similar to measurements in general.** This possibility of a linear transformation between different scales is similar to the fact that in measurements:

- we can select a different measuring unit, and
- for some quantities like time or temperature, we can select a different starting point;

see, e.g., [9]. Here:

- When we use a measuring unit which is $c_1$ times smaller, than all numerical values get multiplied by $c_1$:
$$x\mapsto c_1\cdot x.$$
For example, wehn we replace meters with centimeters, then 1.7 m becomes 170 cm.
- When we use a starting point which is $c_2$ units earlier than the original one, then this value $c_2$ is added to all numerical values:
$$x\mapsto x+c_2.$$

If we change both the measuring unit and the starting point, then we get a general linear transformation

$$c \mapsto c_1 \cdot x + c_2.$$

In measurements, we often also have nonlinear transformations. For example:

- The energy of an earthquake can be measured either by its energy, or by the logarithm of its energy – which is the usual Richter scale.
- Similarly, the energy of a signal can be measured in the usual energy units, or in decibels, which is the logarithmic scale.

In some applications, more complex transformations are used as well.

Similarly to this, we can potentially envision non-linear transformation between different scales of degree of similarity. What form can these transformations have?

**What are possible nonlinear transformations?** Let us analyze what are reasonable transformations in general.

First of all, all linear transformations are reasonable. Also:

- If a transformation from one scale to another is reasonable, then an inverse transformation is also reasonable.
- If we have two reasonable transformations, then applying them one after another – i.e., performing a superposition of these transformations – should also lead to a reasonable transformation.

Thus, the class of all reasonable transformations should be closed under taking the inverse and under taking the superposition. In mathematics, such classes are called *transformation groups*.

Finally, our goal is to use this information in computer-aided decision making. In each computer, we can only store finitely many values, so it makes sense to limit ourselves to classes of transformations which are determined by finitely many parameters. Such transformation groups are called *finite-dimensional*.

So, the question of which transformations are reasonable can be reformulated as a question of what are the possible finite-dimensional transformation groups that contain all linear transformations. A general description of such groups was conjectured by Norbert Wiener, the father of Cybernetics, in [14]. This conjecture was proved in [2], [13]. In particular, in the 1-D case, when we confider functions of one variables, the conclusion is that all the transformations from each such group must be fractionally linear, i.e., have the form

$$g(x) = \frac{A \cdot x + B}{1 + C \cdot x}; \qquad (4)$$

(see also [6] for the 1-D proof).

**Let us apply this conclusion to our case.** Both the similarity measure $s(a, b) = f(d(a, b))$ and the original metric $d(a, b)$ describe the similarity between the two objects $a$ and $b$. Thus, we can consider similarity and metric as representing the same quantity in two different scales. So, based on what we have concluded, the transformation $f(d)$ between these two scales must be fractionally-linear, i.e., must have the form

$$f(d) = \frac{A \cdot d + B}{1 + C \cdot d}, \qquad (5)$$

for some $A$, $B$, and $C$.

To find the values of these three parameters, let us recall the above-mentioned properties of the function $f(d)$:

- that $f(0) = 1$,
- that $f(d) \to 0$ as $d \to \infty$, and
- that $f(d)$ is a decreasing function of $d$.

Substituting $d = 0$ into the formula (5) and equating the result to 1, we conclude that $B = 1$, so

$$f(d) = \frac{A \cdot d + 1}{1 + C \cdot d}. \qquad (6)$$

For $d \to \infty$, this expression tends to

$$\frac{A}{C}.$$

Thus, the fact that this limit should be equal to 0 means that

$$\frac{A}{C} = 0,$$

i.e., that $A = 0$. Thus, the desired nonlinear transformation has the form

$$f(d) = \frac{1}{1 + C \cdot d}. \qquad (7)$$

The requirement that the function $f(d)$ is decreasing leads to

$$C > 0.$$

**From "almost exactly" to "exactly".** The formula (7) is almost exactly the formula (1). To get exactly the formula (1), let us take into account that the distance $d(a, b)$ can also be described by using different measuring units.

- If for distance, we select a measuring unit which is $C$ times smaller than the original one,
- then the new numerical values of the distance take the form $d' = C \cdot d$.

If we describe the distance in these new units, then the formula (7) takes exactly the desired form (1); to be more precise, the form

$$f(d') = \frac{1}{1 + d'}.$$

Thus, we have indeed explained the emergence of the empirical formula (1) – it is the only formula corresponding to natural requirements.

## III. Second Explanation

**Main idea behind the second explanation.** In the first explanation, we focused on analyzing what is the actual dependence between the distance and the similarity. In this explanation, we kind of ignored the fact that similarity usually comes from people marking a value on the interval $[0, 1]$.

However, in reality, such markings are very uncertain. There is a well-known "seven plus minus two law" (see, e.g., [5],

[10]), according to which, in particular, when we do such types of markings, we, in effect, only distinguish between 5 to 9 different values. Thus, the accuracy with which we mark the similarity value ranges from 11% (corresponding to 9 classes on the interval $[0,1]$) to 20% (corresponding to 5 classes on this interval). This inaccuracy can be easily checked: if we ask people to mark the same thing again, they may use somewhat different values (within this accuracy).

With such imprecise values, it makes sense not to seek exact matching of the dependence $s = f(d)$, but rather to look for functions which are the fastest to compute – as have mentioned, from the very beginning, the ultimate goal of assigning similarity values is to make decisions, and often, we need to make decision as soon as possible. So, the question becomes: of all the functions $f(d)$ that satisfy the above three conditions, which ones are the fastest to compute?

**Which functions are the fastest to compute?** In the computer, the only directly hardware supported operations are arithmetic operations: addition, substraction, multiplication, and division. Everything else is computed as a sequence of such arithmetic operations, for which the operands are:

- either constants,
- or the input values,
- or the results of previous arithmetic operations.

For example, when we ask a computer to compute the values $\exp(x)$, what the computer will actually compute is the sum of the first few terms of the Taylor series for this functions:

$$\exp(x) \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^k}{k!}.$$

From this viewpoint, the computation time of each computation is, crudely speaking, proportional to the number of arithmetic operations that constitute these computations. So, the fastest computations are the ones that use the smallest number of such arithmetic operations.

**Computing $f(d)$ must include division.** Let us first explain that computing the function $f(d)$ must include division. Indeed, if this computation only included addition, subtraction, and multiplication, then we would compute a polynomial, and polynomials do not tend to 0 as $d \to \infty$. Thus, at least one arithmetic operation must be division.

**Can we have just one division?** Can we just have one division? Not really. In this case, when we start with $d$ and constants, the only things we can get by division are

$$\frac{C}{d}, \quad \frac{d}{C}, \text{ and } \frac{d}{d} = 1.$$

The first two expressions do not satisfy the property $f(0) = 1$, the third expression is not decreasing to 0 as $d$ increases. Thus, we cannot have only one arithmetic operation, we must have at least one more arithmetic operation.

**Which functions $f(d)$ can be computed in two computational steps?** The expression (1) requires two arithmetic operations:

- first, we add 1 and $d$, and
- then, we divide 1 by $1 + d$.

So, this is clearly one of the fastest-to-compute functions $f(d)$. Let us analyze what other functions $f(d)$ satisfying all three requirements we can compute by using two arithmetic operations – one of which is division.

**What if we perform division first: first step.** If we perform division first, we get

$$\frac{C}{d} \text{ or } \frac{d}{C}.$$

**What if we first compute $C/d$.** If we start with the first of these options, then on the next step, as a second input to the second arithmetic operation, we can have a constant or the original value $d$. Thus, we have the following options:

- If the second operation is addition or subtraction, we get
$$\frac{C}{d} + C' \text{ or } \frac{C}{d} \pm d.$$
None of these expressions satisfies the condition $f(0) = 1$.
- If the second operation is multiplication, we get
$$\frac{C}{d} \cdot C' = \frac{C \cdot C'}{d} \text{ or } \frac{C}{d} \cdot d = C.$$
Here, we do not get any new functions.
- If we second operation is division, then we get:
$$\frac{\frac{C}{d}}{C'} = \frac{C/C'}{d}, \quad \frac{C'}{\frac{C}{d}} = \frac{C'}{C} \cdot d,$$
$$\frac{\frac{C}{d}}{d} = \frac{C}{d^2}, \quad \frac{d}{\frac{C}{d}} = \frac{1}{C} \cdot d^2.$$
The first and third expressions do not satisfy the requirement that $f(0) = 1$, and the second and fourth are polynomials – and we have already mentioned that the transformation $f(d)$ cannot be a polynomial.

**What if we first compute $d/C$.** If we start with the second of these options, i.e., if we first compute

$$\frac{d}{C} = \frac{1}{C} \cdot d,$$

, then on the next step, as a second input to the second arithmetic operation, we can have a constant or the original value $d$. If the second operation is addition, subtraction, or multiplication, we get a polynomial, and we have already mentioned that that the function $f(d)$ cannot be a polynomial. This, the only possible option is when the second arithmetic operation is division. In this case, we get the following options:

$$\frac{\frac{1}{C} \cdot d}{C'} = \frac{C}{C'} \cdot d, \quad \frac{C'}{\frac{1}{C} \cdot d} = \frac{C \cdot C'}{d},$$

$$\frac{\frac{1}{C}\cdot d}{d}=\frac{1}{C}, \quad \frac{d}{\frac{1}{C}\cdot d}=C.$$

In the first case we get a polynomial. In the second case, we do not satisfy the requirement that $f(0)=1$, and in the third and fourth cases, we get constants. So, none of these options lead to functions $f(d)$ that satisfy all three requirements.

**What if division is the second arithmetic operation.** Since the cases when division is the first arithmetic operation do not lead to a function $f(d)$ that satisfies all three conditions, we need to consider the remaining cases when we perform division only as a second arithmetic operation. In this case, the first arithmetic operation is addition, subtraction, or multiplication. Thus, as a result of the first arithmetic operation, we get $d+C$, $C-d$, or $C\cdot d$.

When the first arithmetic operation results in $d+C$, we have $d$, constants, and $d+C$. Thus, we have the following division options:

- The first option is
$$\frac{C'}{d+C}.$$
The requirement that $f(0)=1$ leads to $C'=C$, so this expression is equal to
$$\frac{C}{d+C}=\frac{1}{1+C^{-1}\cdot d}.$$
This is exactly the expression (7) that, as we have shown, is equivalent to (1) after an appropriate re-scaling of distance.
- The second option is
$$\frac{d}{d+C}$$
which does not satisfy the condition $f(0)=1$.
- The third option is
$$\frac{d+C}{C'}=\frac{1}{C'}\cdot d+\frac{C}{C'}.$$
This is a polynomial, so it cannot satisfy all three conditions.
- The fourth option is
$$\frac{d+C}{d}=1+\frac{C}{d}.$$
This option does not satisfy the condition $f(0)=1$.

When the first arithmetic operation is substraction, the conclusions are similar.

When first operation results in $C\cdot d$, we have $d$, constants, and $C\cdot d$. Thus, we have the following division options:

- The first option is
$$\frac{C'}{C\cdot d}=\frac{C''}{d}, \text{ where } C''\stackrel{\text{def}}{=}\frac{C'}{C}.$$
So, in this case, we do not get a new function
- The second option is
$$\frac{d}{C\cdot d}=\frac{1}{C},$$
a constant function which is not decreasing.
- The third option is
$$\frac{C\cdot d}{C'}=\frac{C}{C'}\cdot d,$$
a polynomial.
- The fourth option is
$$\frac{C\cdot d}{d}=C,$$
a constant.

**Summarizing.** By considering all possible options, we conclude that the out of all functions $f(d)$ that satisfy all three requirements, the only functions that can be computed the fastest – in two arithmetic steps – are the functions of type (7), and these functions are, in effect, equivalent to the desired formula (1). Thus, we get the second explanation of the effectiveness of the empirical formula (1) – that this function is the fastest to compute.

## Acknowledgments

The authors are greatly thankful to Janusz Kacprzyk for his encouragement.

## References

[1] R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.

[2] V. M. Guillemin and S. Sternberg, "An algebraic model of transitive differential geometry", *Bulletin of American Mathematical Society*, 1964, Vol. 70, No. 1, pp. 16–47.

[3] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[4] J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, Springer, Cham, Switzerland, 2017.

[5] G. A. Miller, "The magical number seven plus or minus two: some limits on our capacity for processing information", *Psychological Review*, 1956, Vol. 63, No. 2, pp. 81–97.

[6] H. T. Nguyen and V. Kreinovich, *Applications of Continuous Mathematics to Computer Science*, Kluwer, Dordrecht, 1997.

[7] H. T. Nguyen, C. L. Walker, and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2019.

[8] V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*, Kluwer, Boston, Dordrecht, 1999.

[9] S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, New York, 2005.

[10] S. K. Reed, *Cognition: Theories and Application*, SAGE Publications, Thousand Oaks, California, 2022.

[11] T. Segaran, *Programming Collective Intelligence: Building Smart Web 2.0 Applications*, O'Reilly, Sebastopol, California, 2007.

[12] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[13] I. M. Singer and S. Sternberg, "Infinite groups of Lie and Cartan, Part 1", *Journal d'Analyse Mathematique*, 1965, Vol. XV, pp. 1–113.

[14] N. Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine*, 3rd edition, MIT Press, Cambridge, Massachusetts, 1962.

[15] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.