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

Julio Urenda[1], Olga Kosheleva[2],
and Vladik Kreinovich[3]
[1]Department of Mathematical Sciences
[1]Department of Teacher Education
[2]Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA
jcurenda@utep.edu, olgak@utep.edu,
vladik@utep.edu

1. 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.
- Suppose that we know, for each i from 1 to n, that a decision d_i was successful in situation s_i.
- It can be a control decision, a medical decision, a financial decision, etc.
- Then, to select a decision d in a new situation s, we should use the following natural rules:
  - if s is similar to s_1, then d should be similar to d_1;
  - if s is similar to s_2, then d should be similar to d_2;
  - $\ldots$
  - if s is similar to s_n, then d should be similar to d_n.

2. Need for similarity measures (cont-d)

- These rules use an imprecise ("fuzzy") natural-language word "similar".
- Such natural-language words are ubiquitous.
- To transform such rules into a precise decision making strategy, Lotfi Zadeh invented fuzzy methodology.
- In this methodology, each imprecise property can be described by assigning:
  - to each possible object,
  - the degree to which, according to the expert, this object satisfies this property.
- In our case, we can ask the expert:
  - for each pair of situations (or pair of decisions) a and b
  - to estimate to what extent the following statement is true: "a and b are similar".

3. Need for similarity measures (cont-d)

- In a computer, "true" is usually represented as 1, and "false" as 0.
- So it is natural to represent an intermediate degree of confidence by a number between 0 and 1.
- 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.
- 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.
- Values strictly between 0 and 1 describe the cases when there is some similarity, but there is some dissimilarity as well.

4. Need for similarity measures (cont-d)

- Sometimes, experts are not comfortable providing numerical estimates of their degree of similarity.
- They can only give us binary answers: similar or not similar.
- Then we can ask several (n) experts this question.
- If m of them answer that the objects are similar, use the ratio $\frac{m}{n}$ as the desired degree of similarity.

5. Need for metric-based similarity measures

- In many practical cases, we have a large number of possible objects and situations.
- In such cases, 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.
- In other words, we have 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.

6. Need for metric-based similarity measures (cont-d)

- Thus, a natural idea is to estimate the desired degree of similarity $s(a,b)$ between the two objects based on this metric, as:

$$s(a,b)=f(d(a,b)) \text{ for some function } f(d).$$

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

7. Natural properties of the transformation $f(d)$

- The degree of similarity must satisfy the following natural properties.
- The degree of similarity $s(a,b)$ should attain its largest value $s(a,b) = 1$ if the objects are identical (under given representation).
- So, 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?

8. Empirical fact: an efficient transformation

- In many practical applications, the following function leads to reasonable similarity-based decisions

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

- A natural question is: why this functions works well?
- In this talk, 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.

9. Towards the first explanation

- 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.

10. 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 $s = \dfrac{m}{n}$, we have a new degree $s' = \dfrac{m}{n + n'}$.

11. Resulting problem (cont-d)

- One can easily see that the new degree s' can be obtained from the original degree by a transformation

$$s' = c_1 \cdot s, \text{ where } c_1 \stackrel{\text{def}}{=} \frac{n}{n+n'}.$$

- 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 $s = \dfrac{m}{n}$, we have a new degree $s' = \dfrac{m+n'}{n+n'}$.

12. Resulting problem (cont-d)

- One can easily see that the new degree s' can be obtained from the original degree by a transformation $s' = c_1 \cdot s + c_2$, where $c_2 \stackrel{\text{def}}{=} \dfrac{n'}{n + n'}$.
- In both cases, we have linear transformations between different scales, i.e., linear functions $s' = g(s)$.

13. 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.
- 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, when 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 $c \mapsto c_1 \cdot x + c_2$.

14. This is similar to measurements in general (cont-d)

- In measurements, we often also have nonlinear transformations.
- 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?

15. What are possible nonlinear transformations?

- Let us analyze what are reasonable transformations in general.
- First of all, all linear transformations are reasonable.
- 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.

16. What are possible nonlinear transformations (cont-d)

- 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.

17. What are possible nonlinear transformations (cont-d)

- This conjecture was proved in the 1960s.
- In particular, for functions of one variables, all the transformations from each such group must be fractionally linear:

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

18. 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:

$$f(d) = \frac{A \cdot d + B}{1 + C \cdot d} \text{ for some } A, B, \text{ and } C.$$

- 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.

19. Let us apply this conclusion to our case (cont-d)

- Substituting $d = 0$ into the general formula and equating the result to 1, we conclude that $B = 1$, so $f(d) = \dfrac{A \cdot d + 1}{1 + C \cdot d}$.
- For $d \to \infty$, this expression tends to $\dfrac{A}{C}$.
- Thus, the fact that this limit should be equal to 0 means that $\dfrac{A}{C} = 0$, i.e., that $A = 0$.
- Thus, the desired nonlinear transformation has the form

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

- The requirement that the function $f(d)$ is decreasing leads to $C > 0$.

20. From "almost exactly" to "exactly".

- This is almost exactly the desired formula.
- 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 above formula takes exactly the desired form $f(d') = \dfrac{1}{1 + d'}$.
- Thus, we have indeed explained the emergence of the empirical formula – it is the only formula corresponding to natural requirements.

21. 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 ignored the fact that similarity usually comes from people marking a value on the interval $[0, 1]$.
- In reality, such markings are very uncertain.
- There is a well-known "seven plus minus two law" 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 observed.

22. Main idea behind the second explanation (cont-d)

- 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.
- Indeed, as we have mentioned, the ultimate goal of assigning similarity values is to make decisions.
- 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?

23. Which functions are the fastest to compute?

- In a computer, the only directly hardware supported operations are arithmetic ones: 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:

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

24. Which functions are the fastest to compute (cont-d)

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

25. 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.
- Polynomials do not tend to 0 as $d \to \infty$.
- Thus, at least one arithmetic operation must be division.

26. 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.

27. Which functions $f(d)$ can be computed in two computational steps?

- The empirical expression 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 $f(d)$.
- What other functions $f(d)$ satisfying all three requirements we can compute in two arithmetic operations – one of which is division.

28. What if we perform division first

- If we perform division first, we get
$$\frac{C}{d} \text{ or } \frac{d}{C}.$$
- 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
$$\frac{C}{d} \cdot C' = \frac{C \cdot C'}{d} \text{ or } \frac{C}{d} \cdot d = C.$$

29. What if we perform division first (cont-d)

- 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 $f(0) = 1$.
- The second and fourth are polynomials – and we have already mentioned that the transformation $f(d)$ cannot be a polynomial.

30. What if we first compute d/C

- What if first compute d/C?
- 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.
- We have already mentioned that the function $f(d)$ cannot be a polynomial.
- This, the only possible option is when the second arithmetic operation is division.

31. What if we first compute d/C (cont-d)

- In this case, we get the following options:

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

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

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

32. What if division is the second arithmetic operation

- The cases when division is the first arithmetic operation do not lead to a function $f(d)$ that satisfies all three conditions.
- So, we need to 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 $\dfrac{C'}{d+C}$.

33. What if division is the 2nd arithmetic operation (cont-d)

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

34. What if division is the 2nd arithmetic operation (cont-d)

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

35. Summarizing

- We have considered all possible options; so
  - 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 $1/(1 + C \cdot d)$.
- We showed that these functions are, in effect, equivalent to the desired formula $1/(1 + d)$.
- Thus, we get the second explanation of the effectiveness of the empirical formula – that this function is the fastest to compute.

36. Acknowledgments

- 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).
- It was also supported 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.