the similarity metric

Garry Morrison
garry@semantic-db.org

July 28, 2016

So, let's give a nicely formatted version of my similarity metric. There are several permutations, but I call all of them simm. Let's start with the general definition:

$$simm(w, f, g) = \frac{wf + wg - wfg}{2 \cdot max(wf, wg)} \tag{1}$$

where w is a weight, and f and g are "patterns". In the discrete case we have:

$$wf = \sum_i |w_i f_i| \tag{2}$$

$$wg = \sum_i |w_i g_i| \tag{3}$$

$$wfg = \sum_i |w_i f_i - w_i g_i| \tag{4}$$

In the continous case we have:

$$wf = \int dx\, |w(x) f(x)| \tag{5}$$

$$wg = \int dx\, |w(x) g(x)| \tag{6}$$

$$wfg = \int dx\, |w(x) f(x) - w(x) g(x)| \tag{7}$$

The next version we have is called the rescaled simm. In this case the shape of the pattern is important, but not the amplitude. In practice, this is the one we use most of the time. It is derived from the general simm by rescaling f and g so that $wf = wg = 1$. Here is the discrete rescaled simm:

$$simm(w, f, g) = 1 - \frac{1}{2} \sum_i \left| \frac{w_i f_i}{\sum_j |w_j f_j|} - \frac{w_i g_i}{\sum_j |w_j g_j|} \right| \tag{8}$$

And here is the continuous, rescaled simm:

$$simm(w, f, g) = 1 - \frac{1}{2} \int dx \left| \frac{w(x)f(x)}{\int ds\, |w(s)f(s)|} - \frac{w(x)g(x)}{\int ds\, |w(s)g(s)|} \right| \tag{9}$$

Next, the unweighted, rescaled simm follows in the obvious way by setting $w = 1$:

$$simm(f, g) = 1 - \frac{1}{2} \sum_i \left| \frac{f_i}{\sum_j |f_j|} - \frac{g_i}{\sum_j |g_j|} \right| \tag{10}$$

$$simm(f, g) = 1 - \frac{1}{2} \int dx \left| \frac{f(x)}{\int ds\, |f(s)|} - \frac{g(x)}{\int ds\, |g(s)|} \right| \tag{11}$$

The final permutation of simm is, if $w_i, f_i, g_i \geq 0$ for all i, then we have this simplification of the unscaled discrete simm:

$$simm(w, f, g) = \sum_i \frac{w_i \cdot min(f_i, g_i)}{max(wf, wg)} \tag{12}$$

And then an extension of this to more than 2 “patterns”:

$$simm(w, f_1, f_2, \cdots, f_p) = \sum_i \frac{w_i \cdot min(f_{1i}, f_{2i}, \cdots, f_{pi})}{max(wf_1, wf_2, \cdots, wf_p)} \tag{13}$$

where:

$$wf_k = \sum_i |w_i f_{ki}| \tag{14}$$

Now, let’s derive the general case for p “patterns”, in full ugly detail, though we could use the $p = 2$ case to guess the general form. To do this we need to move to complex numbers, and p’th roots of unity. Let’s define a short cut notation for them:

$$j_{pk} = e^{2\pi i k/p} \tag{15}$$

and note this identity:

$$\sum_{k=1}^{p} j_{pk} = 0 \tag{16}$$

Next, to make the derivation cleaner, define these guys:

$$wf^p = \int dx \left| \sum_{k=1}^{p} j_{pk} w(x) f_k(x) \right| \tag{17}$$

$$wf_k = \int dx \, |w(x) f_k(x)| \tag{18}$$

$$A = \left| \sum_{k=1}^{p} j_{pk} w f_k \right| \tag{19}$$

which has the key property:

$$0 \leq wf^p \leq \sum_{k=1}^{p} wf_k \tag{20}$$

$$0 \leq wf^p + A \leq \sum_{k=1}^{p} wf_k + A \tag{21}$$

$$\tag{22}$$

which attains the lower bound when all $f_k(x)$ are equal, by way of (16), and attains the upper bound when all $f_k(x)$ are disjoint. Now, normalize that:

$$0 \leq \frac{wf^p + A}{\sum_{k=1}^{p} wf_k + A} \leq 1 \tag{23}$$

And invert so that we have 0 when disjoint, and 1 when all $f_k(x)$ are equal:

$$0 \leq 1 - \left(\frac{wf^p + A}{\sum_{k=1}^{p} wf_k + A} \right) \leq 1 \tag{24}$$

And tidy:

$$0 \leq \frac{\sum_{k=1}^{p} wf_k - wf^p}{\sum_{k=1}^{p} wf_k + A} \leq 1 \tag{25}$$

Apply one more identity:

$$\sum_{k=1}^{p} x_k + \left| \sum_{k=1}^{p} j_{pk} x_k \right| \leq p \cdot max(x_1, x_2, \cdots, x_p) \tag{26}$$

Resulting in:

$$0 \leq \frac{\sum_{k=1}^{p} wf_k - wf^p}{p \cdot max(wf_1, wf_2, \cdots, wf_p)} \leq 1 \tag{27}$$

So there we have it. The p pattern version of simm, which we can clearly see has the same structure as (1), and reduces to (1) when $p = 2$:

$$simm(w, f_1, f_2, \cdots, f_p, p) = \frac{\sum_{k=1}^{p} wf_k - wf^p}{p \cdot max(wf_1, wf_2, \cdots, wf_p)} \tag{28}$$

$$wf_k = \int dx \, |w(x) f_k(x)| \tag{29}$$

$$wf^p = \int dx \left| \sum_{k=1}^{p} j_{pk} w(x) f_k(x) \right| \tag{30}$$

We obtain the discrete version by swapping the integral with a sum:

$$wf_k = \sum_i |w_i f_{ki}| \tag{31}$$

$$wf^p = \sum_i \left| \sum_{k=1}^{p} j_{pk} w_i f_{ki} \right| \tag{32}$$

Now, the p pattern version of rescaled simm obtained by mapping:

$$w(x) f_k(x) \Rightarrow \frac{w(x) f_k(x)}{\int ds \, |w(s) f_k(s)|} \tag{33}$$

$$w_i f_{ki} \Rightarrow \frac{w_i f_{ki}}{\sum_j |w_j f_{kj}|} \tag{34}$$

effectively setting:

$$wf_k = 1 \tag{35}$$

$$\sum_{k=1}^{p} wf_k = p \tag{36}$$

$$p \cdot max(wf_1, wf_2, \cdots, wf_p) = p \tag{37}$$

resulting in:

$$simm(w, f_1, f_2, \cdots, f_p, p) = 1 - \frac{1}{p} \int dx \left| \sum_{k=1}^{p} j_{pk} \frac{w(x) f_k(x)}{\int ds \, |w(s) f_k(s)|} \right| \tag{38}$$

$$simm(w, f_1, f_2, \cdots, f_p, p) = 1 - \frac{1}{p} \sum_i \left| \sum_{k=1}^{p} j_{pk} \frac{w_i f_{ki}}{\sum_j |w_j f_{kj}|} \right| \tag{39}$$

Next, we need to observe some symmetries of $simm(w, f_1, f_2, \cdots, f_p, p)$. Global symmetry:

$$w(x) \to s_1 \cdot w(x) \tag{40}$$

$$f_k(x) \to s_2 \cdot f_k(x) \tag{41}$$

Global symmetry of the rescaled simm:

$$f_k(x) \to s_{2k} \cdot f_k(x) \tag{42}$$

Local symmetry:

$$w(x) \to w(x) \cdot s(x) \tag{43}$$

$$f_k(x) \to s(x)^{-1} \cdot f_k(x) \tag{44}$$

Providing that $s_1, s_2, s_{2k}, s(x) \neq 0$

Finally, simm() is somewhat stable to changes in the order of the patterns f_k, though not identical. This is due to the wf^p term in (28). The other 2 terms are order independent.

Anyway, that's it. Pick and choose which variation of simm you need depending on what you are trying to do. Though we haven't mentioned the superposition versions of simm, which follow from (12) and (13).