Generalised Prioritisation: A New Way of Combining Similarity Metrics

Alex Ferguson **Derek Bridge**
Department of Computer Science
University College, Cork
(a.ferguson@cs.ucc.ie) (d.bridge@cs.ucc.ie)

Abstract

We describe similarity metrics, which are a generalisation of similarity measures having any partial order as their result type. We then present the main result of this paper by describing generalised prioritisation, a new way of combining similarity metrics. Generalised prioritisation uses an indifference relation on the degrees of similarity. We show that the use of the indifference relation gives a (very rough) analogue of the use of weights in more traditional, numeric-valued similarity measures.

Similarity Metrics

Similarity measures are used, e.g., in the retrieval phase of case-based reasoning systems and in forming 'clusters' of objects in machine learning systems. More recently, they have been used in searching product catalogues in e-commerce systems.

In the main, similarity measures have been binary operators that, when applied to two objects of type α, return a number, usually a real from $[0, 1]$, denoting their degree of similarity. That is, their type is most usually $\alpha \to \alpha \to [0, 1]$, for any data type α.

We have elsewhere described *similarity metrics*, our generalisation of similarity measures, e.g. [3, 1]. A similarity metric is a binary operator that, when applied to two objects of type α, returns some value indicating the degree to which the two objects are similar. That is, if $\sim$ is a similarity metric, then

$$\sim :: \alpha \to \alpha \to P$$

for any data type α, and some suitable P.

All that we require of P is that it impose a relative degree of similarity, e.g. so that we can say whether objects a and b are more similar to each other than are objects c and d. We require, therefore, that P be a *partial order*,[1] $P = (S_P, \sqsubseteq_P)$.

Of course, not every function of this type is a reasonable similarity metric. To be a similarity metric, a function must also satisfy a number of similarity metric axioms. These axioms have been discussed in [1].

Examples of Similarity Metrics

Assume an estate agent has computer representations of houses for sale. And assume we can 'project out' of these representations the value of certain attributes of the houses. The number of bedrooms, *bdrms*, for example, is a value of type $\mathcal{N}$; the price, *price*, is also of type $\mathcal{N}$; the amenities of the locale, *amen*, might be a value of type $\mathcal{P}\{p, s, t\}$, the powerset of $\{p, s, t\}$, i.e. a subset of $\{p, s, t\}$ that indicates whether the house has a parking space (p), nearby shops (s) and transport connections (t).

On occasion, we might want a boolean-valued similarity measure, i.e. one which simply says whether two values are similar or not. To formulate these as similarity metrics, we construct a suitable partial order Bool as the result type of the function: the partial order is **False** $\sqsubset$ **True**[2]. The similarity metrics will then be of type $\alpha \to \alpha \to$ Bool.

An example Bool-valued similarity metric is:

$$x \sim_{bdrms} y \mathrel{\hat{=}} abs(x - y) < 2$$

[1] In our earlier work, we required P to be a complete lattice (see, e.g. [3, 1]), which is a more restrictive requirement. We thought this to be a natural restriction. However, in this paper we lift this restriction since we show similarity metrics where the result type is a partial order but not a lattice.

[2] And of course **False** $\sqsubseteq$ **False**, **True** $\sqsubseteq$ **True**.

Note that this is not transitive. A 2-bedroom and 3-bedroom house are similar (the absolute difference in the number of bedrooms is less than 2); a 3-bedroom and 4-bedroom house are also similar; but a 2-bedroom and 4-bedroom house are not similar (the absolute difference in the number of bedrooms is not less than 2).

Bool-valued similarity metrics fail to capture the intuitive notion of different degrees of similarity. This motivates numeric-valued similarity functions. These too are special cases of our framework. Any numeric set ordered by $\leq$ or $\geq$ gives the partial order we need as a result type.

A simple numeric metric is to measure the distance between the two values. For example,

$$\sim_{price} :: \mathcal{N} \to \mathcal{N} \to (\mathcal{Z}, \leq)$$

$$x \sim_{price} y \mathrel{\hat{=}} -abs(x - y)$$

By this function on house prices, an IRP80,000 house and an IRP85,000 house are similar to degree -5000 (negated absolute difference) and this is a higher degree of similarity than the similarity of an IRP80,000 and an IRP90,000 house, which are similar to degree $-10,000$.

An example numeric-valued similarity metric for amenities is given by the cardinality of the intersection of the sets of amenities:

$$\sim_{amen1} :: \mathcal{P}\{p, s, t\} \to \mathcal{P}\{p, s, t\} \to (\mathcal{N}, \leq)$$

$$x \sim_{amen1} y \mathrel{\hat{=}} |\, x \cap y \,|$$

which designates a house that has parking and nearby schools ($\{p, s\}$) to be similar to degree 1 to a house that has nearby schools and transport ($\{s, t\}$).

More conventional [0,1]-valued similarity metrics can also be defined. For example, we could normalise to [0,1] the results of any of the above numeric-valued metrics.

As a final example, we show that set-valued metrics are also possible. We use intersection again, but this time we let the intersections themselves denote the degrees of similarity:

$$\sim_{amen2} :: \mathcal{P}\{p, s, t\} \to \mathcal{P}\{p, s, t\} \to (\mathcal{P}\{p, s, t\}, \subseteq)$$

$$x \sim_{amen2} y \mathrel{\hat{=}} x \cap y$$

Now the similarity of a house with amenities $\{p, s\}$ and one with amenities $\{s, t\}$ *is* $\{s\}$. If we have two other houses with amenities $\{p, s, t\}$ and $\{s, t\}$, then these two houses are similar to degree $\{s, t\}$. When we compare the degrees of similarity, we see that our second pair of houses are more similar to each other than our first pair of houses are to each other: $\{s\} \subset \{s, t\}$.

The advantages of our framework are explained in [1]. We state them very briefly here. The first advantage is that the framework subsumes many ways of measuring similarity, e.g. boolean-valued, numeric-valued, set-valued [2], feature structure-valued [4], and linguistic-hedge-valued [1]. In subsuming so many approaches, we know that results we obtain in this framework (theorems, implementation techniques, etc.) will apply quite broadly. The second advantage is the 'naturalness' of the similarity measures that we can define: result types can be chosen to best suit the application. The third advantage is the ease with which we can combine similarity metrics. We discuss this issue more fully in the remaining sections of this paper, where we also present the main result of this paper: a new way of combining similarity metrics.

Combining Metrics

We would not usually want to compute the similarities of objects according to a single 'projected out' attribute of those objects. More usually, we will consider more than one attribute, e.g. price *and* local amenities. We will have a similarity measure for each attribute, and we will want to compute an *overall* degree of similarity by combining the results of these individual similarity measures.

It is here that there is an apparent advantage in using traditional numeric-valued similarity measures: the results of applying the individual similarity measures are easily combined, using (possibly weighted) sums, products or averages. In our similarity metric framework, it is, of course, just as easy for us to combine

two numeric-valued similarity metrics. Indeed, whenever the result types of the two individual metrics are the same (both numeric, both boolean, both set-valued, etc.), then computing an overall degree of similarity might be straightforward.

But we have to address the possibility that the two individual metrics have different result types. For example, we might wish to combine the results of $\sim_{price}$ (numeric-valued) and $\sim_{amen2}$ (set-valued). It would be regrettable at this point to have to convert $\sim_{amen2}$ into a numeric-valued function, simply in order to allow its combination with $\sim_{price}$. Were this to be necessary, we might just as well have used $\sim_{amen1}$ (numeric-valued) all along, and we would be losing the advantages of our framework. Inter-conversions of this kind are the main solution used in the literature, e.g. [5]. But, inter-conversion is not necessary.

The key insight is that our framework allows metrics to have *any* partially ordered set as a result type. There is no reason, then, why the result type should not be a partially ordered *set of pairs*. When combining numeric-valued $\sim_{price}$ and set-valued $\sim_{amen2}$, degrees of similarity will be denoted by pairs from $\mathcal{Z} \times \mathcal{P}\{p, s, t\}$. For example, an IRP80,000 house near shops and transport is similar to an IRP85,000 house with parking and near shops to degree $\langle -5000, \{s\}\rangle$: this pair (comprising values from the two individual similarity metrics) denotes their degree of similarity. An IRP80,000 house with no local amenities and an IRP90,000 house near shops are similar to degree $\langle -10,000, \{\}\rangle$.

This approach has the advantages of giving more structure and hence more explanatory power to the degrees of similarity, and of working in all possible cases, *without needing inter-conversions*.

But, of course, it is not good enough to simply say that this set of pairs is the result type of the combined metric. The result type must be partially ordered so that we can tell whether, e.g., $\langle -5000, \{s\}\rangle$ is a higher degree of similarity than $\langle -10,000, \{\}\rangle$.

The normal route to defining the ordering on the set of pairs will be to construct it from the orderings on the result types of the individual metrics. For this, we need operators that construct new partial orders from existing partial orders.

Combining Partial Orders

Let $A = (S_A, \sqsubseteq_A)$ and $B = (S_B, \sqsubseteq_B)$ be two arbitrary partial orders.

Product

Define the product of A and B, $A \times B$ as follows:

$$A \times B \mathrel{\hat{=}} (\{\langle x, y\rangle \mid x \in S_A, y \in S_B\}, \sqsubseteq_{A\times B})$$

where

$$\langle x_1, x_2\rangle \sqsubseteq_{A\times B} \langle y_1, y_2\rangle \mathrel{\hat{=}} x_1 \sqsubseteq_A y_1 \wedge x_2 \sqsubseteq_B y_2$$

Here, both individual functions contribute equally to the overall degree of similarity. For example, houses that are similar to degree $\langle -5000, \{s\}\rangle$ are not as similar as houses that are similar to degree $\langle -2000, \{s, t\}\rangle$ (because $-5000 < -2000$ and $\{s\} \subset \{s, t\}$), whereas houses that are similar to degree $\langle -5000, \{p, s, t\}\rangle$ are neither more nor less similar to houses that are similar to degree $\langle -2000, \{s, t\}\rangle$ (because they are less similar on one attribute, $-5000 < -2000$, but more similar on the other, $\{p, s, t\} \supset \{s, t\}$).

Strict prioritisation

Define the strict prioritisation of A over B, $A \gg B$ as follows:

$$A \gg B \mathrel{\hat{=}} (\{\langle x, y\rangle \mid x \in S_A, y \in S_B\}, \sqsubseteq_{A\gg B})$$

where

$$\langle x_1, x_2\rangle \sqsubseteq_{A\gg B} \langle y_1, y_2\rangle \mathrel{\hat{=}} x_1 \sqsubset_A y_1 \vee (x_1 = y_1 \wedge x_2 \sqsubseteq_B y_2)$$

Here, the first individual function counts more to the overall degree of similarity than the second. Indeed, this is a lexicographic ordering: the ordering is based on the first ordering, but the second ordering may resolve ties in the first

ordering. For example, houses that are similar to degree $\langle -5000, \{s\}\rangle$ are not as similar as houses that are similar to degree $\langle -2000, \{\}\rangle$: $-5000 < -2000$ and so the second element of the pair is irrelevant. However, houses that are similar to degree $\langle -5000, \{s\}\rangle$ are not as similar as houses that are similar to degree $\langle -5000, \{s, t\}\rangle$: they tie on the first element of the pair but the tie is resolved by the second element, $\{s\} \subset \{s, t\}$.

Indifference Between Different Degrees of Similarity

Product and strict prioritisation represent extremes in the way we combine two partial orders: the partial orders either both count equally, or one takes precedence over the other.

To have available only these extremes is undesirable. More traditional numeric-valued similarity measures offer the possibility of using *weighted* combinations: by choosing different amounts by which one is weighted more highly than the other, we can influence the extent to which that similarity measure contributes more highly to the overall measure than the others. Were we to adopt the use of weights, however, we would have again to resort to conversion of all metrics to ones with numeric result types, and we regard this as undesirable.

The solution turns out to be to introduce the idea of a 'level of indifference' to the result type of a metric. Then, in a prioritisation, the second metric will take effect not when the first metric reports equal degrees of similarity but when it reports degrees of similarity about which we are indifferent. The details of this new prioritisation definition are given in the next section. Before that, we say what we mean by indifference between degrees of similarity.

Consider the similarity metric $\sim_{price}$, from above. Even if the price metric is nominally the most important criterion, a prospective buyer is unlikely to regard small differences in the degrees of similarity of two houses according to this metric as being decisive in the choice of which match is 'best', if other criteria also apply. For example, it would probably be unwise to regard an IRP80,000 house and an IRP80,001 house (similar to degree -1) to be decisively more similar than an IRP80,000 house and an IRP80,002 house (similar to a strictly lower degree, -2). Yet, by our definition, strict prioritisation of this price metric over other criteria would make such small differences decisive. Better in the case of small differences (where 'small differences' for a house price metric might be differences as large as, say, 2500) would be to let the other similarity criteria contribute to the ranking.

In order to do this, we must change the signatures of similarity metrics. We augment the result type with a boolean-valued binary function that will be used to indicate whether we are indifferent between two degrees of similarity:

$$\sim :: \alpha \to \alpha \to (P, \simeq_P)$$

We need to formalise this notion of indifference between degrees of similarity ($\simeq_P$). The relations that formalise this notion will generally *not* be equivalence relations. To use an equivalence relation would have the same effect as simply using a less discriminating result type in the similarity metric; there would still be cases of small differences in the similarity metric that were having decisive effects.

For example, suppose we wished to declare, for the price metric, similarities within 2500 to be equivalent:

$$\sim_{price} :: \mathcal{N} \to \mathcal{N} \to ((\mathcal{Z}, \leq), R_{price})$$

$$x \sim_{price} y \mathrel{\hat{=}} -abs(x - y)$$

$$s_1 R_{price} s_2 \mathrel{\hat{=}} s_1 \; div \; 2500 = s_2 \; div \; 2500$$

Here R_{price} is indeed an equivalence relation. But this is, in essence, the same as simply using $=$ as the indifference relation with prior integer division by 2500:[3]

$$\sim_{price} :: \mathcal{N} \to \mathcal{N} \to ((\mathcal{Z}, \leq), =)$$

[3] It is also the same, in this case, as using integer division on the argument types:

$$x \sim_{price} y \mathrel{\hat{=}} -abs(x \; div \; 2500 - y \; div \; 2500)$$

This was the approach we used in [3], but it suffers the same weakness that we explain above.

$$x \sim_{price} y \mathrel{\hat{=}} -\,abs(x-y)\ div\ 2500$$

The problem with these (equivalent) formulations is that they still have points where small differences in the degrees of similarity are decisive. For example, similarity degrees of -2499 and -2500 would be mapped to distinct values (0 and -1), whereas similarity degrees of -2500 and -4999 are both mapped to -1.

More reasonable is an ϵ-*equality*, similar to that used for comparing floating point numbers, i.e. we are indifferent between two degrees of similarity, s_1 and s_2, iff $abs(s_1 - s_2) < \epsilon$. For our price metric, ϵ could be -2500. Now, we would be indifferent between similarity degrees of -2499 and -2500, and of -2500 and -4999, but not of -2499 and -4999.

We note that this relation is reflexive and symmetric, but *not* transitive. Instead, it satisfies the generally weaker property of *convexity*. Define a relation R to be P-convex, over any partial order P, iff:

$$\forall x, y, z \in P : x \sqsubseteq_P y \sqsubseteq_P z, x\ R\ z \Rightarrow x\ R\ y, y\ R\ z$$

To generalise then, what we require of an indifference relation on degrees of similarity, $\simeq_P$, is that it satisfy reflexivity, symmetry, and convexity. It need not satisfy transitivity.

Generalised Prioritisation

As argued earlier, an operator is desired which, to some extent, favours its first argument, but not necessarily to the extent of strictly lexicographical preference. Our idea for a more flexible scheme of prioritisation is to use the indifference relation: essentially, given degrees of similarity that are pairs of values, we wish to order those degrees of similarity according to the first value in the pair, unless we are indifferent between the first values, in which case we wish to take both components of the pair into account.

Our definition of generalised prioritisation is:

$$\langle x_1, x_2\rangle \sqsubseteq_{A \gg_{\simeq} B} \langle y_1, y_1\rangle \mathrel{\hat{=}} x_1 \sqsubseteq y_1 \wedge (x_1 \simeq_A y_1 \Rightarrow x_2 \sqsubseteq_B y_2)$$

Suppose, using this definition, we prioritise the price metric over the amenities metric, using ϵ-equality with $\epsilon = 2500$ as our indifference relation. An IRP80,000 house and an IRP85,000 house have a higher degree of similarity than that which exists between an IRP80,000 house and an IRP90,000 house, irrespective of the amenities they offer: the price metric takes precedence and is decisive. An IRP80,000 house and an IRP81,000 house are strictly more similar than an IRP80,000 house and an IRP82,000 house, but we are indifferent between these two degrees of similarity. Hence, for the first pair of houses to be more similar overall to each other than the second pair are to each other, we require the first pair to be more similar on their amenities as well.

To further explore the new definition, we will turn to a more abstract example. It has the advantage of using small, finite sets and is therefore amenable to a diagrammatic treatment.

Consider the partially ordered set **3**, equal to $0 \sqsubset 1 \sqsubset 2 \sqsubset 3$. Remember, this set would be the result type of a metric: very similar objects would be similar to degree 3; dissimilar objects would be similar to degree 0. We will consider prioritising **3** over itself. We will use different indifference relations, and show the different effects that we get.

Note first of all that if we prioritise order A over order B but we put $\simeq_A$ equal to $=_A$ in our definition, after simplification, the definition reduces to our previous definition for strict prioritisation. So, prioritising **3** over **3**, using equality as the indifference relation, gives the ordering on pairs depicted in Figure 1a, in which you can see that the second ordering is used only to resolve strict ties.

Furthermore, if we prioritise order A over order B but we define $x \simeq_A y = \mathbf{True}$, then we instead obtain an ordering identical to that on $A \times B$, the product of the two orders. The ordering obtained by prioritising **3** over **3** using this indifference relation is depicted in Figure 1b.

These are the 'extreme' situations, as we discussed them in the earlier section: strict prioritisation and product. They are special cases of the generalised form of prioritisation, where we

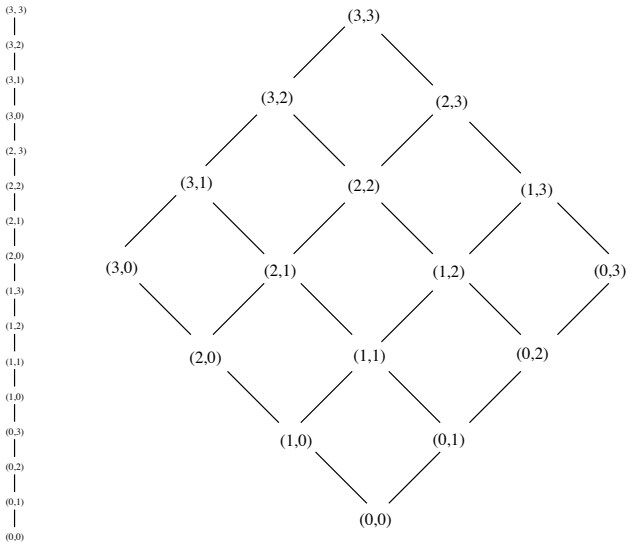

Figure 1: (a) Strict prioritisation (b) Product

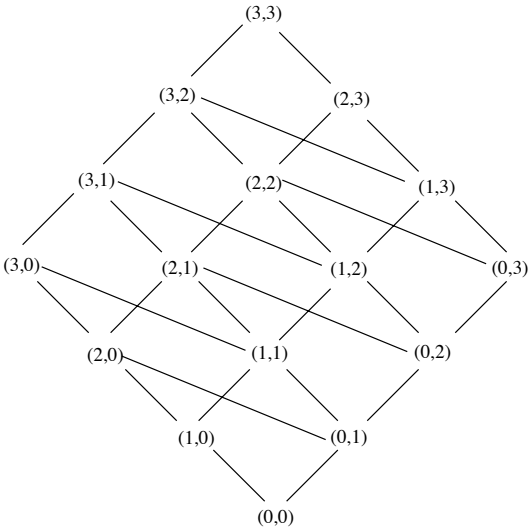

Figure 2: Prioritisation using $\simeq_{\mathbf{3}}$

use indifference relations that are minimally and maximally indifferent, respectively.

But, we can use other indifference relations to get orders which are (at least in some sense) intermediate between strict prioritisation and product.

Consider a non-trivial indifference relation on **3**, an ϵ-equality with $\epsilon = 2$:

$$s_1 \simeq_{\mathbf{3}} s_2 \hat{=} abs(x - y) < 2$$

That is, we now have, e.g., that $1 \simeq_{\mathbf{3}} 2$, and $2 \simeq_{\mathbf{3}} 3$ (but $1 \not\simeq_{\mathbf{3}} 3$). If we prioritise **3** over **3** but use $\simeq_{\mathbf{3}}$ as the indifference relation, we get the new partial order shown in Figure 2. Certain pairs which were ordered in strict prioritisation (Figure 1a) are now incomparable and thus, in this sense, 'equally good', in the generalised prioritisation. Certain pairs that were unordered in product (Figure 1b) are now ordered and thus reflect the dominance of the first ordering.

This figure also illustrates an interesting technical point. If we consider pairs $\langle 2,0\rangle$ and $\langle 1,1\rangle$ in Figure 2, we can see these are bounded above by $\langle 3,0\rangle$ and by $\langle 2,1\rangle$. As these two upper bounds are themselves incomparable ($\langle 3,0\rangle \not\sqsubseteq \langle 2,1\rangle$, and $\langle 2,1\rangle \not\sqsubseteq \langle 3,0\rangle$), then $\langle 2,0\rangle$ and $\langle 1,1\rangle$ have no *least* upper bound. Hence, this partial order is not a lattice. This is a key part of our motivation for choosing partial orders as the basis of our framework.[4]

References

[1] Bridge, D.G.: Defining and Combining Symmetric and Asymmetric Similarity Measures, in B.Smyth & P.Cunningham (eds.), *Advances in Case-Based Reasoning (Procs. of 4th European Workshop on Case-Based Reasoning)*, LNAI-1488, pp.52-63, Springer, 1998

[2] Matuschek,D. & Jantke,K.P.: Axiomatic Characterization of Structural Similarity for Case-Based Reasoning, in D.D.Dankel (ed.), *Procs. of Florida AI Research Symposium*, Florida AI Research Society, pp.432-436, 1997

[3] Osborne, H. & Bridge, D.G.: We're All Going on a Summer Holiday: An Exercise in Non-Cardinal Case Base Retrieval, in G.Grahne (ed.), *Frontiers in Artificial Intelligence and Applications (Procs. of Sixth Scandinavian Conference on Artificial Intelligence)*, pp.209-219, IOS Press, 1997

[4] Plaza, E.: Cases as terms: A feature term approach to the structured representation of cases, in M.Veloso & A.Aamodt (eds.), *Case-Based Reasoning Research and Development (Procs. of First International Conference on Case-Based Reasoning)*, LNAI-1010, pp.265-276, Springer, 1995

[5] Wilson, D.R. & Martinez, T.R.: Improved Heterogeneous Distance Functions, *Journal of Artificial Intelligence Research*, vol.6, pp.1-34, 1997

[4] Contra our earlier work [3, 1].