Similarity and Metrics in Case-Based Reasoning

Gavin Finnie,* Zhaohao Sun†
School of Information Technology, Bond University, Qld 4229, Australia

Similarity is a core concept in case-based reasoning (CBR), because case base building, case retrieval, and even case adaptation all use similarity or similarity-based reasoning. However, there is some confusion using similarity, similarity measures, and similarity metrics in CBR, in particular in domain-dependent CBR systems. This article attempts to resolve this confusion by providing a unified framework for similarity, similarity relations, similarity measures, and similarity metrics, and their relationship. This article also extends some of the well-known results in the theory of relations to similarity metrics. It appears that such extension may be of significance in case base building and case retrieval in CBR, as well as in various applied areas in which similarity plays an important role in system behavior. © 2002 Wiley Periodicals, Inc.

1. INTRODUCTION

The concepts of similarity and similarity relations play a fundamental role in many fields of pure and applied science. The notion of a metric or distance, $d(x, y)$, between objects x and y has long been used in many contexts as a measure of similarity or dissimilarity between elements of a set. Thus, there exists a wide variety of techniques for dealing with problems involving similarity, similarity relations, similarity measures, and similarity metrics. For example, fuzzy logic and case-based reasoning (CBR) provide a number of concepts and techniques for dealing with similarity relations, similarity measures, and similarity metrics, many of which are quite effective in dealing with the particular classes of problems that motivated their development.

The present article does not intend to add still another technique to the vast armamentarium that is already available. Its purpose is rather to introduce a unifying point of view based on the available theory and application of fuzzy logic[29] and CBR.[14] This is accomplished by examining the notions of similarity, similarity relations, similarity measures, and similarity metrics or distance functions in Refs. 6, 8, 14, 17, 27, 29, 30, thereby discussing the relationships between these concepts and

*e-mail: gfinnie@bond.edu.au.
†Author to whom correspondence should be addressed; e-mail: zsun@bond.edu.au.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 17, 273–287 (2002)
 • DOI: 10.1002/int.1085

influences on CBR. The main contribution of our approach consists of providing a unified conceptual framework for the study of fuzzy similarity relations and similarity metrics (or measures), thereby facilitating research and development of CBR and fuzzy logic with their applications.

In what follows, our attention will be focused primarily on reviewing and defining some of the basic notions with this conceptual framework, and exploring their elementary implications and the relationships between them. Although our approach might be of significance in areas such as pattern recognition, decision processes, intelligent information retrieval, natural language processing, system modeling, approximation, and multi-agent systems, we will make no attempt in the present article to discuss its applications in these or related problem areas. Because this work was motivated when we attempted to develop algebraic CBR and its application to e-commerce, we will use CBR and e-sales as scenarios, if required.

The article is organized as follows: Section 2 examines similarity relations and metrics, as well as fuzzy similarity relations, concisely. Section 3 reviews similarity and the nearest neighbor algorithm, which plays a major part in CBR. Section 4 assesses similarity based on metrics. Section 5 introduces the possible world of problems and solutions, and examines similarity metrics in order to resolve the confusion among similarity, similarity measures, and similarity metrics in CBR. Section 6 investigates the relationship between similarity metrics and Euclidean metrics, and proposes a couple of definitions for similarity metrics. Section 7 ends this article with a few concluding remarks.

2. FUNDAMENTALS OF SIMILARITY AND METRICS

Similarity is the core concept in CBR, because it is used not only in case retrieval but also in case adaptation and case base building.[24] In what follows, we will examine similarity relations, fuzzy similarity relations, and similarity metrics, which are all necessary for investigating CBR, and in particular, case base building and case retrieval. They are also fundamentally important for the further development of this work.

2.1. Similarity Relations

The concept of a similarity is a natural generalization of similarity between two triangles and two matrices in mathematics.[24] More precisely:

DEFINITION 1. *A binary relation S on a non-empty set X is called a* similarity *relation provided it satisfies*

(R) $\forall x, xSx$
(S) *If* xSy, *then* ySx
(T) *If* xSy, ySz, *then* xSz

The conditions (R), (S), *and* (T) *are the reflexive*, *symmetric*, *and transitive laws. If* xSy, *we say that* x *and* y *are similar.*[19]

Example 1. Matrices B and C in $M_{n,n}$ are *similar* if $C = PBP^{-1}$ for an invertible P, in which case we write $B \sim C$. It is easy to prove that $\sim$ is a similarity relation on $M_{n,n}$.[19]

This example implies that the concept of a similarity relation here is a generalization of the similarity between matrices in $M_{n,n}$.

Example 2. Let f be a function with domain A and codomain B, namely, $f : A \to B$, and define xSy if $f(x) = f(y)$. Then S is a similarity relation on A.

It is obvious that the similarity relation S in this example has the following property: if x_1 and x_2 are similar in the sense of S, then x_1 and x_2 have the same solution, that is, $f(x_1) = f(x_2)$. This reflects that "similar problems have the same solution" in the e-sale settings, at least in some cases. For example, in a shoe shop, the seller may put many different pairs of shoes together and sell these for the same price, for example, $88.00. In this case, the seller views those mentioned shoes as "similar."

It should be noted that the similarity relation proposed here is identical to the equivalence relation in discrete mathematics.[19] However, we prefer the former rather than the latter in the context of CBR, because similarity relations rather than equivalence relations play an important role in CBR. Thus, this treatment is different from the idea of Zadeh,[29] in that Zadeh considered a similarity relation as a fuzzy one and as a generalization of the concept of an equivalence relation, while we view Zadeh's similarity relations as fuzzy similarity relations (See Section 2.2).

2.2. Fuzzy Similarity Relations

As an extension of similarity relations, fuzzy similarity relations were introduced by Zadeh in 1971[29] and have attracted much attention since then.[6,16,30] Fuzzy similarity relations have also been used in CBR, in particular in case base building[24] and case retrieval.[8,17] For the sake of brevity, we use standard fuzzy set theory notation for operations min and max, although there are many alternative choices for these operations available in fuzzy set theory.[30] S is still used to denote a fuzzy similarity relation if there is not any confusion arising.

DEFINITION 2. *A fuzzy binary relation S on a non-empty set X is a fuzzy similarity relation*[a] *in X if it is reflexive, symmetric, and transitive*;[16,29] *that is*:

$$S(x, x) = 1 \tag{1}$$

$$S(x, y) = S(y, x) \tag{2}$$

$$S \geq S \circ S \tag{3}$$

where $\circ$ is the composition operation of fuzzy binary relations based on min and max operations.

[a]In this article we use the notation $S(p, q)$ for the membership $\mu_S(p, q)$, although the latter is commonly used in the fuzzy set literature.

A more explicit form of Equation 3 is

$$S(x, z) \geq \bigvee_{y} (S(x, y) \wedge S(y, z)) \tag{4}$$

Equation 4 is called max-min transitivity.[29]

The revised form of this definition was given by Ovchinnikov in 1991.[16] Dubois and Prade[6] used the revised form for fuzzy similarity relations directly in 1994. The main difference between the definitions of Zadeh and Ovchinnikov lies in that instead of Equation 4, Ovchinnikov viewed the following model as max-min transitivity:

$$S(x, z) \geq S(x, y) \wedge S(y, z) \tag{5}$$

This is simpler than that used by Zadeh, because if the cardinality of the set is less than or equal to 3, then Equation 4 coincides with Equation 5.

Example 3. Let $X = \{x_1, x_2, x_3\}$. Suppose a binary relation S on X is defined by

$$S = \begin{bmatrix} 1 & a & b \\ a & 1 & a \\ b & a & 1 \end{bmatrix}$$

Then S is a fuzzy similarity relation on X if and only if $a \leq b$.[16]

Zadeh considered Equation 4 as max-min transitivity based on the composition operation of fuzzy relations,[29] while Ovchinnikov, and Dubois and Prade, used Equation 5 for max-min transitivity without any explanation. Unfortunately, they have ignored the influence of the concept of metrics or distance on their definition, which we will discuss at somewhat greater length in Section 6.

Dubois, et al.[8] believe that transitivity in CBR is not always compulsory. However, we emphasize that transitivity is necessary in some cases in CBR, for example, in case base building,[24] because the first step for case base building is to use fuzzy similarity relations to partition the possible world of problems (for detail see Ref. 24). But we also require the separating property $\forall(x, y \in X)$, $S(x, y) = 1$ if and only if $x = y$,[8] because we assume that x is identical to y, if $S(x, y) = 1$.

It should be noted that a fuzzy similarity relation does not satisfy the traditional transitivity law, although it is quasi-transitive (or $\otimes$-transitive[8]), which, unfortunately, is so weak that sequential use will lead to *fuzzy degeneration*. In other words, if fuzzy reasoning is performed for many steps sequentially using the traditional transitive law, the consequence will easily lose validity. For example, there are no exercises of fuzzy reasoning with ten inference steps (even more than one) in many textbooks of fuzzy logic, such as Ref. 30. We also like to illustrate an interesting example as follows. In normal life we can easily say "10,001 is similar to 10,000" without taking membership degree into account. Here, "is similar to" is a fuzzy similarity relation according to our *intuitive expectation,*[29] because it reflects what we think about "similarity." Using this similarity and the transitive law we can at once conclude "10,001 is similar to 9,999," since "10,000 is similar to 9,999." After having performed this similarity-based reasoning 10,000 times, we come to the conclusion that "10,001 is similar to 1." This is a paradox, which leads to a *fuzzy*

degeneration. If we take membership degree into account and replace the min with a product (a t-norm), then from Equation 4, the membership value of compound similarities decreases. In this example, if we assume $\mu(10{,}001, 10{,}000) = 0.99$ then

$$\mu(10{,}001,\ 9{,}999) = t[\mu(10{,}001,\ 10{,}000), \mu(10{,}000,\ 9{,}999)] = 0.99^2$$

and finally we have $\mu(10{,}001, 1) = 0.99^{10{,}000} \approx 0$. Therefore, the degree of similarity between 10,001 and 1 is, in essence, zero, which is the same as our intuitive expectation. However, if we consider t-norm as a min-max function, then $\mu(10{,}001, 1) = 0.99$.[24]

2.3. Metric and Metric Space

A *metric space* is a non-empty set X in which a *metric* (or distance function) d is defined, with the following properties:[20]

(a) $0 \leq d(x, y) \leq \infty$ for all x and $y \in X$
(b) $d(x, y) = 0$ if and only if $x = y$
(c) $d(x, y) = d(y, x)$ for all x and $y \in X$
(d) $d(x, y) \leq d(x, y) + d(y, z)$ for all x, y, and $z \in X$

As is well known, property (d) is called the *triangle inequality,* which is based on the property of Euclidean geometry.

The popular metric space is Euclidean space R^n, and R^1 is the real line, while R^2 is the plane in Euclidean geometry. Almost all CBR systems are based on Euclidean space R^n.

3. SIMILARITY AND THE NEAREST NEIGHBOR ALGORITHM

In this section we review similarity and the nearest neighbor algorithm, which plays a major role in CBR.

Similarity can be formalized in a relational and in a functional way.[14] The relational approach uses a four-place relation, $R(x, y, u, v)$, meaning "x and y are at least as similar as u and v are." This allows the definition of the nearest neighbor notion:

$$NN(x, z) \Leftrightarrow \forall y R(x, z, x, y) \tag{6}$$

meaning z is a nearest neighbor to x. If the nearest neighbor is unique, then NN is also used as a function symbol. A refinement is when k nearest neighbors are considered for some $k (k \geq 1)$.

In our own view, similarity formalized in a relational way is a binary relation. It is irrelevant how near any two neighbors are. Nearness involves a distance concept and should be assessed by a metric or distance function.[b] We will discuss it again in more detail in the following sections.

[b]Metric is more mathematical flavor than distance function, although they are same in Ref. 20.

The typical nearest neighbor algorithm, which was first implemented in REMIND (Cognitive Systems, 1992),[9,11,28] is shown in Equation 7:

$$\frac{\sum_{i=1}^{n} w_i \times \mathrm{sim}\left(f_i^I, f_i^R\right)}{\sum_{i=1}^{n} w_i} \tag{7}$$

where w_i is the importance weighting of a feature i represented as numerical values between 0 and 1. Nearer neighbors have values closer to 1, while more distant neighbors have values closer to 0. f_i^I and f_i^R are the values for feature i in the input and retrieved cases, respectively, and sim is the similarity function for primitives. It is this similarity function that makes the nearest neighbor algorithm different from a mathematical expectation formula, which is the generalization of arithmetic average value. However, it is regrettable that there appears to be no research telling us how to formalize this similarity function mathematically, based on problem domains.

In the relational approach,[14] similarity is treated as a partial ordering. Such partial orderings can be realized by numerical functions that are called similarity (or dual distance) measures $\mathrm{sim}(x, y)$ or $d(x, y)$, respectively. Both similarity measures sim and distance measures d induce four-place relations R_{sim} and R_d in an obvious way. If $R_{\mathrm{sim}}(x, y, u, v) \Leftrightarrow R_d(x, y, u, v)$ holds, R_{sim} and R_d are called compatible.

It should be noted that from the above discussion we do not yet know the relationship between similarity and distance, although a new concept of similarity measure has been introduced.

In order to reduce arbitrariness, some assumptions are common:

(1) $0 \leq \mathrm{sim}(x, y) \leq 1$
(2) $\mathrm{sim}(x, x) = 1$

The intention of Assumption 1 is normalization and Assumption 2 implies that each object is itself its own nearest neighbor. This is less often the case for the following conditions:

(3) $\mathrm{sim}(x, y) = \mathrm{sim}(y, x)$ (symmetry)
(4) $d(x, z) \leq d(x, y) + d(y, z)$ (*triangle inequality,* in terms of distance measures)

The notion of a distance function $d(x, y)$ is dual; here, simply all orderings are reversed. For an attribute-value representation, a simple distance function is the Hamming distance. If problems are coded as n-dimensional real vectors, classical mathematical metrics like the Euclidean or the Manhattan distance are often used.

From the above discussion we observe that it seems there is no discussion on the relationship between similarity and metric in a mathematical way, although Richter's idea[14] that "from a mathematical viewpoint, the notion of similarity is equivalent to the dual distance concept. However, both notions emphasize different aspects and have given rise to different computational approaches" is correct. Furthermore, we believe that the notion of the nearest neighbor should be directly based on the notion of either "distance" or metrics rather than on the notion of similarity, which will be discussed once again in Section 6.

4. SIMILARITY AND METRICS

In many applications such as CBR, we use metrics to measure the similarity between two objects, such as cases in CBR. Then it would be reasonable to assume that:

$$S_1(x, y) = 1 - |x - y| \tag{8}$$

and say that x and y are similar with respect to S_1 if $|x - y| < \varepsilon$, where $|\ |$ is the Euclidean distance function, and ε is a small number (in relation to $|x - y|$). But then, S_1 is not transitive from a mathematical viewpoint, which is inconsistent with our intuitive expectation that the similarity relation is transitive. However, the fuzzy similarity relation S_2 in the following example is transitive in some sense.[29]

Example 4. A fuzzy similarity relation possessing transitivity. Suppose that:

$$S_2(x, y) = e^{-\beta|x-y|}, \quad x, y \in X \tag{9}$$

where β is any positive number. In the definition the max-product transitivity is employed. Under this condition, S_2 satisfies Equation 5, and therefore it is a fuzzy similarity relation.

In practice, we sometimes prefer to use S_1 rather than S_2 to measure the similarity between two objects. Thus, two questions arise as follows:

(1) What is the difference between S_1 and S_2?
(2) Can we resolve the inconsistency between transitivity and our intuitive expectation?[29]

Let's try to answer them. Assume that $\beta = 1$ for convenience. As is well known, the Taylor series expansion of the function $e^{-|x-y|}$ is[20]:

$$\sum_{n=0}^{\infty} \frac{(-|x - y|)^n}{n!} = 1 - |x - y| + \frac{|x - y|^2}{2!} + \cdots \tag{10}$$

Then

$$S_2 - S_1 = \frac{|x - y|^2}{2!} + \cdots = O(|x - y|^2) \tag{11}$$

Thus, if we choose ε as small as possible, that is, p and q are quite similar, almost equal, then the difference between S_1 and S_2 can be insignificant. In other words, the inconsistency between transitivity and our intuitive expectation results from the insignificant difference in Equation 11; that is, $O(|x - y|^2)$. Therefore, we can refer to S_1 as a fuzzy similarity relation and perform similarity-based reasoning with transitivity using S_1 only if we limit the number of transitive reasoning steps. This result also can be easily extended as follows:

Let

$$S_3(x, y) = 1 - d(x, y) \tag{12}$$

and say that x and y are similar with respect to S_3 if $d(x, y) < \varepsilon$, where d is a metric and ε is a small number.

5. SIMILARITY METRICS IN CBR

In this section we discuss the case base of a CBR system in a broader domain; that is, the possible world of problems and the possible world of solutions. Then we discuss similarity metrics in CBR based on integration of similarity relations and metrics mentioned. We argue that it is similarity metrics rather than similarity measures that should be used to assess the similarity between problems or solutions in CBR.

5.1. Possible World of Problems and Solutions

After the failure of GPS (general problem solver) in early AI to capture general purpose reasoning or intelligence, intelligent systems have only served to solve certain types of problems in a special field or in a narrow *domain*.[24] Any CBR system can thus only give the answers to problems in a *possible world*,[c] which corresponds to a scenario in the real world. Based on this idea, the possible world of problems, W_p, and the possible world of solutions, W_s, are the whole world of an agent (See Ref. 15) to use CBR to do everything that he can. If an agent considers a CBR system as a function or transformation, h, from W_p to W_s, it is meaningless to discuss the image of $h(x)$ if $x \notin W_p$. Therefore, the agent can only know and play in the world $W_p \times W_s$. For example, in a CBR e-sale system, the possible world of problems, W_p, might consist of:

- Properties of goods
- Normalized queries of customers
- Knowledge of customer behavior
- General knowledge of business (similar to K in Ref. 17), and so on.

And the possible world of solutions, W_s, consists of:

- Price of goods
- Customized answers to the queries of customers
- General strategies for attracting customers to buy the goods, and so on.

It should be noted that if we denote the case base in the CBR system as $C = (P, Q)$, where P is the subset of problem descriptions and Q is the subset of solution descriptions, then it is obvious that P and Q are subsets of W_p and W_s, respectively.[24] In fact, cases in context are usually denoted as $n + m$-tuples of completely, incompletely, or fuzzily described attribute values, with this set of attributes being divided into two non-empty disjoint subsets: the subset of problem description attributes (n-tuples) and the subset of solution or outcome attributes (m-tuples), denoted by P and Q, respectively. A case, c, can be denoted as an ordered pair (p, q), where $p \in P$ and $q \in Q$. The case base $C = (P, Q)$ is the set of known cases.[8,24]

[c]This term is affected by the terminology in modal logic and AI.

5.2. Similarity Metrics

Similarity is frequently used in mathematics, from similarity between two triangles to similarity between two matrices. Similarity is also frequently used in CBR from case base building to case retrieval, and so on. Generally speaking, similarity in mathematics is considered as a relation, while similarity in CBR is considered both a relation[8] and a measure,[3] as well as a metric.[2] This confusion between similarity relation, similarity measure,[3,22,27] and similarity metric[2,27] is so popular that these three concepts are de facto the same in CBR. However, there is still no general definition for the concept of a similarity metric in CBR. At least we have not found such a concept, although many CBR publications are involved in it.[13,14,22,27] No one seems to have any idea about how to differentiate these concepts or what the relationship is between them, although Richter[14] seems to have been aware of the difference between similarity measure and distance or metric (See Section 3). However, he has not investigated them in a mathematical way. In what follows, we will fill this gap.

Roughly speaking, measures assess the size of any subset in a mathematical system, for example, a Borel field,[5] while metrics evaluate the distance between any two elements in a mathematical system, for example, a Banach space.[20] In Euclidean space R^2, a measure can be considered as the generalization of the notion of the area, while a metric can be viewed as the distance between two points. In Euclidean space R^1, we can consider a measure as the generalization of the notion of "*length*" *of any interval* [if the interval I has endpoints x and y, where $x \leq y$, then the length of I is $l(I) = y - x$)[20]] and as a function from a certain subset of the power set of R^1 to $[0, \infty)$, while a metric d is only a function $R^1 \times R^1 \to R^1$ with some conditions. Thus, measures and metrics are two different concepts. We should use a similarity metric rather than a similarity measure to investigate the similarity involved in CBR.

DEFINITION 3. *A relation, denoted by S_m, on non-empty X, is a similarity metric if it satisfies*:

(1) *S_m is a similarity relation in X*
(2) *$1 - S_m$ is a metric on X; that is, it is a function from $X \times X$ to* $[0, 1]$, *provided that*:
- *For any $x, y \in X$, $S_m(x, y) = 1$ if and only if $x = y$*
- *For all $x, y \in X$, $S_m(x, y) = S_m(y, x)$*
- *For all $x, y, z \in X$,*

$$S_m(x, z) \geq S_m(x, y) \wedge S_m(y, z) \tag{13}$$

where $\wedge$ is min operator. Equation 13 *in this definition is called the* similarity inequality. *It should be noted that the similarity metric here, S_m, can not directly satisfy the* triangle inequality (*See Section* 6). *Equation* 13 *is motivated by the concept of fuzzy similarity relations given in Ref.* 16, *which is the same as Equation* 5 *in this article.*

In comparison with the definition of fuzzy similarity relations given in Refs. 8 and 16 and in Section 2.2, we emphasize that the similarity metric here is first a

traditional similarity relation, and also just a metric, maybe to some extent, because we believe that the similarity between two elements is the necessary condition to further discuss how similar they are, which coincides with our intuitive expectation. In practice, our first concern is whether x and y are similar, then we ask how similar they are. In fact, in some cases, such as case base building,[24] we only care for similarity relations rather than for similarity metrics. However, metrics, and in particular similarity metrics, play an important role in case retrieval in CBR.[25] Therefore, the integration of similarity relations and metrics into similarity metrics is of practical significance.

6. RELATIONSHIPS BETWEEN SIMILARITY METRICS AND EUCLIDEAN METRICS

Usually we use similarity metrics to evaluate the similarity between two cases in CBR.[25] The question arises, what is the relationship between the proposed similarity metric and the Euclidean metric? In what follows, we discuss this at somewhat greater length.

Let

$$d(x, y) = 1 - S_m(x, y), \quad \forall x, y \in X \tag{14}$$

Then we have:

- For any $x \in X$, $d(x, x) = 1 - S_m(x, x) = 0$
- For any $x, y \in X$, $d(x, y) = 1 - S_m(x, y) = 1 - S_m(y, x) = d(y, x)$
- For any $x, y, z \in X$, using Equation 13, we have

$$\begin{aligned} d(x, z) &= 1 - S_m(x, z) \leq 1 - (S_m(x, y) \wedge S_m(y, z)) \\ &= (1 - S_m(x, y)) \vee (1 - S_m(y, z)) \\ &= d(x, y) \vee d(y, z) \leq d((x, y) + d(y, z)) \end{aligned}$$

That is,

$$d(x, z) \leq d(x, y) + d(y, z) \tag{15}$$

The *triangle inequality* is valid. Thus we conclude with the following:

PROPOSITION. *$1 - S_m$ is a metric or distance function.*

This proposition demonstrates that the similarity inequality implies the triangle inequality indirectly. On the other hand, the triangle inequality is the generalization of the R^2 property that the "sum of the lengths of any two sides of a triangle is greater than the length of the remaining side,"[5] demonstrated in Figure 1. However, if we define

$$S_m(x, y) = 1 - d(x, y), \quad \forall x, y \in X \tag{16}$$

then we find that the longest edge, $d(x, z)$, in the sense of distance function d,

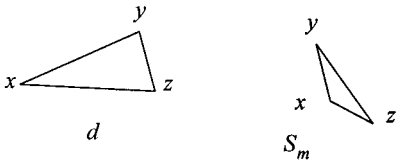

Figure 1. Triangle inequality and similarity inequality.

becomes the shortest edge, $S_m(x, z)$, in the sense of S_m and R^2 (See Figure 1). This characteristic leads us to consider

$$S_m(x, z) \leq S_m(x, y) \wedge S_m(y, z) \tag{17}$$

as an important feature in the similarity metric, when $d(x, z) \geq d(x, y) \vee d(y, z)$, demonstrated in Figure 1. In other cases, that is, $d(x, z) \leq d(x, y) \vee d(y, z)$, Equation 17 is not valid, for example, when the distance of $d(x, z)$ is the shortest among them. However, in such cases, the similarity inequality in Equations 5 or 13 is satisfied. These results have not previously been discussed from the viewpoint of fuzzy similarity relations.[16,29] Thus, it is necessary to examine the relationship between $d(x, z)$, $d(x, y)$, and $d(y, z)$, and $S_m(x, z)$, $S_m(x, y)$, and $S_m(y, z)$ in a unified way. Taking the commutativity of $\vee$ and $\wedge$ into account, it is sufficient for us to consider the following cases:

- $d(x, z) \geq d(x, y) \vee d(y, z)$
- $d(x, y) \leq d(x, z) \leq d(y, z)$
- $d(x, z) \leq d(x, y) \wedge d(y, z)$

In what follows, we consider each of these cases in more detail. We leave the consideration of their relationship with the identity relation "=" to the readers.

Case 1. $d(x, z) \geq d(x, y) \vee d(y, z)$. In this case, we have:

$$S_m(x, z) \leq S_m(x, y) \wedge S_m(y, z) \tag{18}$$

which leads to:

$$S_m(x, z) \leq S_m(x, y) \vee S_m(y, z). \tag{19}$$

Case 2. $d(x, y) \leq d(x, z) \leq d(y, z)$. In this case, we have:

$$S_m(x, z) \leq S_m(x, y) \vee S_m(y, z) \tag{20}$$

but $S_m(x, z) \leq S_m(x, y) \wedge S_m(y, z)$ is not valid; however, we have:

$$S_m(x, z) \geq S_m(x, y) \wedge S_m(y, z) \tag{21}$$

as mentioned earlier, Equation 21 was, in essence, used to define the transitivity in the fuzzy similarity relations by Ovchinnikov.[16]

Case 3. $d(x, z) \leq d(x, y) \wedge d(y, z)$. In this case, both

$$S_m(x, z) \leq S_m(x, y) \wedge S_m(y, z) \quad \text{and} \quad S_m(x, z) \leq S_m(x, y) \vee S_m(y, z) \tag{22}$$

are not valid. However, we have:

$$S_m(x, z) \geq S_m(x, y) \vee S_m(y, z) \tag{23}$$

which leads to:

$$S_m(x, z) \geq S_m(x, y) \wedge S_m(y, z) \tag{24}$$

These results imply that the definition of fuzzy similarity relations in Ref. 16 is irrelevant to the triangle inequality. Because the definition of fuzzy similarity relations in Ref. 16 is a simpler form of the definition of a similarity relation given by Zadeh,[29] the latter is also irrelevant to the triangle inequality. In fact, we can assert that the definition of fuzzy similarity relations in Refs. 16 and 29 only reflects one of the above results, that is, Equation 21. It should be noted that Jacas and Valverde[10] extended Equation 19 to define a S-triangular inequality and then a S-pseudometric.

We have already come to two points. The first is that we hope to define the similarity metric as not only a similarity relation from a traditional viewpoint, but also as a metric or distance function, to some extent. The second is that we have found that Equations 18 and 24 are valid under different conditions if we take Euclidean space into account. These two results suggest that we introduce the following definition:

DEFINITION 4. *A relation, denoted by* S_m, *on non-empty* X, *is a similarity metric if it satisfies*:

(1) S_m *is a similarity relation in* X
(2) $1 - S_m$ *is a metric on* X; *that is, it is a function from* $X \times X$ *to* [0, 1], *provided that*:
- *For any* $x, y \in X$, $S_m(x, y) = 1$ *if and only if* $x = y$
- *For all* $x, y \in X$, $S_m(x, y) = S_m(y, x)$
- *For all* $x, y, z \in X$, *either*

$$S_m(x, z) \leq S_m(x, y) \wedge S_m(y, z) \quad \text{or} \quad S_m(x, z) \geq S_m(x, y) \vee S_m(y, z) \tag{25}$$

where $\vee$ *and* $\wedge$ *are max and min operators. Equation* 25 *in this definition is also called the* similarity inequality.

However, it is not easy for us to verify if the similarity inequality in Equation 25 is valid from a pragmatic viewpoint. Therefore we introduce the following definition:

DEFINITION 5. *A relation, denoted by* S_m, *on non-empty* X, *is a similarity metric if it satisfies*:

(1) S_m *is a similarity relation in* X
(2) $1 - S_m$ *is a metric on* X; *that is, it is a function from* $X \times X$ *to* [0, 1], *provided that*:
- *For any* $x, y \in X$, $S_m(x, y) = 1$ *if and only if* $x = y$
- *For all* $x, y \in X$, $S_m(x, y) = S_m(y, x)$
- *For all* $x, y, z \in X$,

$$S_m(x, z) \geq S_m(x, y) + S_m(y, z) - 1 \tag{26}$$

Equation 26 *in this definition is called the* similarity inequality. *It is worth noting that Equation* 26 *is a variant of the following equation*:

$$1 - S_m(x, z) \leq (1 - S_m(x, y)) + (1 - S_m(y, z)) \tag{27}$$

Therefore it satisfies the triangle inequality.

The previous discussion is, in essence, based upon the *triangle inequality* in Euclidean space R^2. In fact, there is another R^2 property that is parallel to the *triangle inequality;* that is, "the difference of the lengths of any two sides of a triangle is less than the length of the remaining side;" more formally:

$$d(x, z) \geq d(x, y) - d(y, z) \tag{28}$$

If we assume

$$S_m(x, y) = 1 - d(x, y), \quad \forall x, y \in X \tag{29}$$

then, based on Equation 28 we come to the conclusion:

$$S_m(x, z) \leq S_m(x, y) - S_m(y, z) + 1$$

Therefore, we introduce another definition of a similarity metric:

DEFINITION 6. *A relation, denoted by* S_m, *on non-empty* X, *is a similarity metric if it satisfies*:

(1) S_m *is a similarity relation in* X
(2) $1 - S_m$ *is a metric on* X, *that is, it is a function from* $X \times X$ *to* [0, 1], *provided that*:
- *For any* $x, y \in X$, $S_m(x, y) = 1$ *if and only if* $x = y$
- *For all* $x, y \in X$, $S_m(x, y) = S_m(y, x)$
- *For all* $x, y, z \in X$,

$$S_m(x, z) \leq S_m(x, y) - S_m(y, z) + 1 \tag{30}$$

Equation 30 *in this definition is also called the* similarity inequality.

We have examined similarity measures, similarity metrics, and distance functions in a novel way, and built an important relationship between similarity metrics and Euclidean metrics, or distance functions. The core idea behind our work is the integration between similarity and metrics. As is known, metrics have played a vital role in mathematics and engineering, in particular in functional analysis and engineering computation, while similarity plays a similar role in many fields in computer science, such as CBR, IR, and pattern recognition. However, there is no theoretical insight into similarity metrics. Almost all similarity metrics and measures for "neighborhood" are domain dependent. There is also a misunderstanding about similarity measures and similarity metrics in CBR. Our results basically fill this gap. We also believe that the similarity metrics introduced in the last two definitions can be used easily in CBR.

Finally, in contrast to the nearest neighbor algorithm mentioned previously in Section 3, we introduce a most similar problem model (MSPM) for case retrieval in CBR as follows.

Let S_m be a similarity metric on the possible world of problems, W_p, and p_0 be a current problem (a normalized enquiry) and similar to $p \in W_p$ in the sense of S_m as a similarity relation. Then the most similar problem p_{most} is the problem that satisfies:

$$\max\{S_m(p_0, q), \forall q \in [p]\} \tag{31}$$

where $[p]$ is the similarity class with the representative p.

7. CONCLUDING REMARKS

In this article we examined similarity relations, similarity measures, similarity metrics, and distance functions in a unified way, and built an important relationship between similarity metrics and Euclidean metrics. The core idea behind our work is the integration of similarity relations and metrics into similarity metrics, based on investigation of similarity relations and Euclidean metrics used in CBR. We also proposed a model for the most similar problem, MSPM, which is based on a similarity metric both as a similarity relation and as a metric to some extent. Because similarity measures and metrics are frequently used to assess the similarity between two objects in a confused way, we suggest that it is better to use similarity metrics rather than similarity measures in CBR. We also advise that similarity metrics introduced in the last two definitions can easily be used in CBR. The preceding analysis extended some of the well-known results in the theory of relations to similarity metrics. It appears that such extension may be of use in case base building and case retrieval in CBR, as well as in various applied areas in which similarity plays an important role in system behavior.

Acknowledgment

This work was supported by an Australian Research Council Small Grant and OPRS of Australian government.

References

1. Allen BP. Case-based reasoning: Business applications. Commun ACM 1994;37(3):40–42.
2. Bento C, Costa E. A similarity metric for retrieval of cases imperfectly described and explained. In: Proc 1st European Workshop on CBR EWCBR'93. Kaiserslautern, Germany; November 1993. Berlin: Springer-Verlag; 1994. pp 8–13.
3. Bonissone PP, Ayub S. Similarity measures for case-based reasoning systems. In: Bouchon_Meunier et al, editors. Advanced Methods in AI: 4th Int Conf on Information Processing and Management of Uncertainty in KBS (IPMU'92). Palma de Mallorca, Spain, LNAI 682 July 1992. Berlin: Springer-Verlag; 1992. pp 483–487.
4. Burkhard H. Extending some concepts of CBR—Foundations of case retrieval nets. In: Lenz M, Bartsch-Spörl B, Burkhard HD, Wess S, editors. Case-based reasoning technology, from foundations to applications. Berlin: Springer; 1998 pp 17–50.
5. DePree JD, Swartz CW. Introduction to real analysis. New York: John Wiley & Sons; 1988.
6. Dubois D, Prade H. Similarity-based approximate reasoning. In: Zurada JM, Marks RJ, II, Robinson XCJ, editors. Computational intelligence: Imitating life. New York: IEEE Press; 1994. pp 69–80.

7. Dubois D, Esteva F, Garcia P, Godo L, López de Mántaras R, Prade H. Fuzzy set-based models in case-based reasoning. In: 2nd Int Conf on Case-Based Reasoning (ICCBR'97). Providence, Rhode Island; July 25–27, 1997, LNAI 1266, 1997. pp 599–610.
8. Dubois D, Esteva F, Garcia P, Godo L, López de Mántaras R, Prade H. Case-based reasoning: A fuzzy approach. In: Ralescu AL, Shanahan JG, editors. Fuzzy logic in artificial intelligence, IJCAI'97 Workshop. Berlin: Springer-Verlag; 1999. pp 79–90.
9. Finnie G, Wittig G. Intelligent support for Internet marketing with case based reasoning. In: Proc 2nd Annual CollECTeR Conf on Electronic Commerce. Sydney; September 1998. pp 6–14.
10. Jacas J, Valverde L. On fuzzy relations, metrics and cluster analisys, http://dmi.uib.es/people/valverde/gran1/GRAN1.html, 1996.
11. Kolodner J. Case-based reasoning. San Mateo: Morgan-Kaufmann; 1993.
12. Leake D. Case-based reasoning: Experiences, lessons & future direction. Menlo Park, CA: AAAI Press/MIT Press, 1996.
13. Leake DB, Plaza E, editors. Case-based reasoning research and development. In: Proc 2nd Int Conf on Case-Based Reasoning (ICCBR'97). Providence, USA; July 1997. Springer; 1997.
14. Lenz M, Bartsch-Spörl B, Burkhard HD, Wess S, editors. Case-based reasoning technology, from foundations to applications. Berlin: Springer; 1998.
15. Nilsson NJ. Artificial intelligence, a new synthesis. San Francisco, CA: Morgen Kaufmann; 1998.
16. Ovchinnikov S. Similarity relations, fuzzy partitions, and fuzzy orderings. Fuzzy Sets Syst 1991;40:107–126.
17. Plaza E, Esteva F, Garcia P, Godo L, López de Mántaras R. A logical approach to case-based reasoning using fuzzy similarity relations, Inf Sci 1996;106:105–122. http://www.iiia.csic.es
18. Plaza E, Arcos JL, Martín F. Cooperative case-based reasoning. In: Distributed artificial intelligence meets machine learning, LNAI 1221. Berlin: Springer-Verlag; 1997. pp 180–201.
19. Ross KA, Wright CRB. Discrete mathematics. Englewood Cliffs, NJ: Prentice Hall; 1988.
20. Rudin W. Real complex analysis. New York: McGraw-Hill; 1987.
21. Smith I, Faltings B, editors. Advances in case-based reasoning, LNAI 1168. Berlin: Springer-Verlag; 1996.
22. Smyth B, Cunningham P, editors. Advances in case-based reasoning, LNAI 1488. Berlin: Springer-Verlag; 1998.
23. Sun Z, Finnie G. ES = MAS? In: Shi Z, Faltings B, Musen M, editors. Proc Conf on Intell Inform Processing (IIP'2000). Beijing; August 21–25, 2000. pp 541–548.
24. Sun Z, Finnie G, Weber K. Case base building with similarity relations, to appear 2002.
25. Sun Z, Finnie G. Rule-based models for case retrieval. In: Baba N, Jain LC, Howlett RJ, editors. Knowledge-based Intelligent Information Engineering Systems & Allied Technologies (KES'01). Frontiers in Artificial Intelligence and Application. Amsterdam: IOS Press; 2001. Vol 69, pp 1511–1515.
26. Voss A. Towards a methodology for case adaptation. In: Wahlster W, editor. Proc 12th European Conf on Artificial Intelligence (ECAI'96); August 1996, Chichester: John Wiley & Sons; 1996. pp 147–151.
27. Waston I, editor. Progress in case-based reasoning. Berlin: Springer; 1995.
28. Waston I. An introduction to case-based reasoning. In: Waston I D, editor. Progress in case-based reasoning. Berlin: Springer; 1995.
29. Zadeh LA. Similarity relations and fuzzy orderings. Inf Sci 1971;3:177–200.
30. Zimmermann HJ. Fuzzy set theory and its applications. Boston/Dordrecht/London: Kluwer; 1991.