Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Metrics for classifying heterogeneous objects

M. Bezem, K. Blok, M. Keijzer

Probability, Networks and Algorithms (PNA)

PNA-R9804 May 31, 1998

Report PNA-R9804
ISSN 1386-3711

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Metrics for Classifying Heterogeneous Objects

Marc Bezem
CWI
Utrecht University
P.O. Box 80126, 3508 TC Utrecht, The Netherlands
bezem@phil.uu.nl

Karnik Blok
Syllogic B.V.
P.O. Box 26, 3990 DA Houten, The Netherlands
karnik@syllogic.nl

Maarten Keijzer
Cap Gemini B.V., Advanced Technology Services
P.O. Box 2575, 3500 GN Utrecht, The Netherlands
MKeijzer@inetgate.capgemini.nl

ABSTRACT
We propose and discuss distance measures to compare objects that have heterogeneous sets of characteristics, such as encountered in, for example, medical diagnosis and information retrieval. We treat both boolean-valued and (scaled) real-valued characteristics. Weighting of characteristics is accomodated. The paper is a modified and extended version of [1].

1991 Mathematics Subject Classification: 54E35, 54E50
1991 Computing Reviews Classification System: H.3.3
Keywords and Phrases: Metric, Classification, Information Retrieval

1. Introduction

Consider two bitstrings $\vec{s}_1$ and $\vec{s}_2$ of equal length, say n. The *Hamming distance* $d(\vec{s}_1, \vec{s}_2)$ between $\vec{s}_1$ and $\vec{s}_2$ is by definition the number of positions in which $\vec{s}_1$ and $\vec{s}_2$ differ. It is well-known (and, moreover, easily verified) that $d(\vec{s}_1, \vec{s}_2)$ satisfies the following general conditions axiomatising a real-valued metric on a set S, in the special case here the set $\{0,1\}^n$ of bitstrings of length n:

(i) $\forall x, y \in S \;\; (d(x,y) = 0 \iff x = y)$

(ii) $\forall x, y \in S \;\; d(x,y) = d(y,x)$ (symmetry)

(iii) $\forall x, y, z \in S \;\; d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality)

The notion of Hamming distance suggests a scheme for classification in the following straightforward way. Given bitstrings $\vec{s}_1, \ldots, \vec{s}_k$, the set $\{0,1\}^n$ can be partitioned into subsets $N_1, \ldots, N_k$ (so-called neighbourhoods), in such a way that each N_i consists of all bitstrings whose Hamming distance to $\vec{s}_i$ is minimal, as compared with their distance to $\vec{s}_1, \ldots, \vec{s}_k$. (The fact that there may be more than one bitstring among $\vec{s}_1, \ldots, \vec{s}_k$ at minimal Hamming distance, is considered irrelevant for this paper. In such a case, any choice is good enough for our purposes.)

As running example of a classification problem we take medical diagnosis, 'the recognition of a disease from its symptoms' [13]. In this context, the clinical picture of a patient should be compared with a number of known syndromes. A syndrome is 'a group of disease symptoms commonly found in association with one another' [13]. The diagnosis is the syndrome that matches best with the clinical picture of the patient.

Bitstrings and Hamming distance provide a rudimentary formalism for diagnosis. Assume we have n boolean-valued symptoms, numbered 1 through n. Every syndrome as well as any clinical picture corresponds to a bitstring encoding the presence/absence of each of the n symptoms. Given a finite set $\mathcal{S}$ of syndromes, every clinical picture can be classified in terms of the nearest syndrome in the sense of the Hamming distance of the corresponding bitstrings. It will be clear that this rudimentary formalism has considerable drawbacks for diagnosis. For example, the simple fact that some symptoms are more important than others is not taken into account. Moreover, some symptoms may not be discrete and should be described by continuous quantities. Even worse, in many cases not all symptoms will be known, or some symptoms may not be applicable at all. It is the purpose of this paper to remedy these shortcomings by proposing and exploring several metrics for objects having boolean- and real-valued characteristics.

In abstract terms, syndromes as well as clinical pictures of patients are expressions in the same representation formalism. Therefore, whenever we talk in the sequel about such expressions as syndroms, it is tacitly understood that we also talk about clinical pictures.

Before giving several metrics, let us discuss the adequacy of the three conditions (i)–(iii) above for the purpose of classification. First observe that, in the above situation, the partitioning of $\{0,1\}^n$ into subsets $N_1, \ldots, N_k$ can be done on the basis of any arbitrary real-valued mapping d. Since (i) implies that the distance of a syndrome to itself is zero, and (ii) that the distance of one syndrome to some other is equal to the distance of the other syndrome to the first, mappings that do not satisfy (i) and (ii) can hardly be taken into consideration. However, (i) is in fact stronger than stated above, for it expresses that every object x is the one and only object at distance 0 of x. This is a bit too restrictive, as we can well imagine two patients having exactly the same clinical picture (or two documents having exactly the same profile, see Section 6). This situation can formally be dealt with by identifying objects that are at distance 0 of each other, which seems to be an allowable abstraction in many applications. Thus doing, we maintain condition (i) at the expense of a slight abstraction on the nature of the objects. Condition (iii), the triangle inequality, requires some more elaborate justification. We shall argue that (iii) is not only important for geometry, but also for classification. In particular, the triangle inequality provides ground for notions like locality and approximation.

Let us illustrate this with an example. Consider a set of syndromes $\mathcal{S}$ and a syndrome $S \in \mathcal{S}$ having distance at least l to any other syndrome. Assume the clinical picture of a patient has distance $\frac{1}{10}l$ to S. By the triangle inequality, the distance to any other diagnosis S' is at least $\frac{9}{10}l$, so 9 times the distance to S. This might provide sufficient evidence in favour of S and against any other diagnosis. Without the triangle inequality one could not exclude so easily the existence of other syndromes than S at a short distance. Of course one has always the possibility to compute all distances between a clinical picture and any syndrome, but this is very inefficient and actually against the idea of approximation.

In the next section we generalize the Hamming distance, and we prove the preferred generalization to be a metric in Section 3. The next section discusses further generalization to the real-valued case, with proofs in Section 5. The last sections are devoted to applications.

2. Generalizing Hamming distance

First we consider the situation in which not all symptoms are known, or not all symptoms do apply. Assume we have a set S of symptoms. As we do not wish to fix the number of symptoms in advance, we allow S to be countably infinite. We may just as well number the symptoms, which will be done by simply taking $S = \{0, 1, 2, \ldots\}$, the set of natural numbers. Again we assume that all symptoms are boolean-valued. It seems reasonable to represent syndromes and clinical pictures as finite, single-valued, sets of pairs consisting of a symptom and a boolean. Let $\mathcal{S}$ be the set of all finite, single-valued,

sets of such pairs. Whenever convenient, we shall use a string notation for elements of $\mathcal{S}$, for example, the string uuftutt denotes $\{(2,\mathtt{f}),(3,\mathtt{t}),(5,\mathtt{t}),(6,\mathtt{t})\}$. Here u is to be interpreted as 'undefined', as opposed to the defined values t for 'true' and f for 'false'. Strings can be made of equal length by postfixing them with u's.

The first attempt to generalize the Hamming distance to this new situation is by restricting two given strings s_1, s_2 to the positions in which they are both defined, and then computing their Hamming distance as bitstrings. This attempt fails since, for example, the unequal strings fu and ut would have distance 0, thus violating condition (i) of the definition of a metric. Similarly, fu and uf would have distance 0, whereas uf and ut would have distance 1. This violates condition (iii) as well.

The second attempt is to take the new strings as three-valued strings and count the number of positions in which two strings differ, exactly as in the two-valued case. This attempt succeeds mathematically in the sense that indeed a metric is obtained, but this metric is unsatisfactory from the point of view of classification. For example, ut and ft are at the same distance from tt, and this is undesirable in a case in which the first symptom in the clinical picture ut does not apply to the patient in question. Another drawback is that the distance between uuuuf and uuuut is exactly the same as the distance between ttttf and ttttt, whereas in the latter case 4 out of 5 defined values (symptoms) match. It is not so obvious how to norm the metric in such a way that a larger number of matching *defined* values reduces the distance.

For the third attempt we need the notion of *symmetric difference* $X \triangle Y$ of two sets X and Y. By definition, $X \triangle Y = (X-Y)\cup(Y-X)$, equivalently characterized by $X \triangle Y = (X \cup Y) - (X \cap Y)$. For any finite set Z, let $|Z|$ denote the number of its elements. Define $d(X,Y) = |X \triangle Y|$ for all $X, Y \in \mathcal{S}$. Now d is a metric according to [8, Chapter 10]. Moreover, $d(\mathtt{ut},\mathtt{tt}) = 1$ and $d(\mathtt{ft},\mathtt{tt}) = 2$, so d is not as bad as the second attempt. However, we still have that $d(\mathtt{uuuut},\mathtt{uuuuf}) = d(\mathtt{ttttt},\mathtt{ttttf}) = 2$. The idea is now to norm this metric.

The fourth attempt is $d(X,Y) = |X \triangle Y| \div (|X| + |Y|)$, a normed version of the previous attempt. Unfortunately this is not a metric, as the triangle inequality fails: $d(\mathtt{ut},\mathtt{tu}) = 1$ and $d(\mathtt{ut},\mathtt{tt}) = d(\mathtt{tt},\mathtt{tu}) = \frac{1}{3}$. This failure is remarkable, since d coincides with the (normed) Hamming distance h for two-valued strings. More precisely, for bitstrings $\vec{s}_1$, $\vec{s}_2$ of length n, with corresponding single-valued sets of pairs X_1, X_2, we have $d(X_1,X_2) = h(\vec{s}_1,\vec{s}_2) \div n$.[1]

As fifth and final attempt we propose the metric[2] with $d(\varnothing,\varnothing) = 0$ and in the non-empty case

$$d(X,Y) = |X \triangle Y| \div |X \cup Y|$$

In the next section we prove that d is indeed a metric and show how to accomodate weighting of symptoms. Note that $d(\mathtt{ut},\mathtt{tt}) = \frac{1}{2}$, $d(\mathtt{ft},\mathtt{tt}) = \frac{2}{3}$, $d(\mathtt{uuuut},\mathtt{uuuuf}) = 1$ and $d(\mathtt{ttttt},\mathtt{ttttf}) = \frac{1}{3}$, all in accordance with the desiderata above. For further examination of the adequacy of d, consider a syndrome ut and clinical pictures uf, tu and fu. Then $d(\mathtt{ut},\mathtt{uf}) = d(\mathtt{ut},\mathtt{tu}) = d(\mathtt{ut},\mathtt{fu}) = 1$. On the basis of this example it could be argued that d is inadequate, by putting forward that the plain conflict of one symptom in $d(\mathtt{ut},\mathtt{uf})$ should outweigh the incomparability of symptoms in $d(\mathtt{ut},\mathtt{tu})$ and $d(\mathtt{ut},\mathtt{fu})$, respectively. However, we give three arguments in favour of the adequacy of the proposed metric. First, in none of the three clinical pictures there is any positive evidence for the syndrome ut, and so all three are justifiably at maximal distance 1 of ut. Second, not only the clinical pictures, but also the syndrome could be incompletely described by two symptoms, which makes the conflicting symptom less important. Third, if one symptom is really more important than some others, then this could be expressed by attaching more weight to the more important symptoms.

With notations as in the previous attempt one can easily see that for bitstrings of length n the metric d satisfies $d(X_1,X_2) = 2 \cdot h(\vec{s}_1,\vec{s}_2) \div (n + h(\vec{s}_1,\vec{s}_2))$. This equality shows that the Hamming

[1] In [9, page 40], $|X \triangle Y| \div (|X| + |Y|)$ is falsely claimed to be a metric, explicitly referring to the triangle inequality. The counterexample above can be rephrased in the setting of [9] as $d(\{a\},\{b\}) = 1$ and $d(\{a\},\{a,b\}) = d(\{a,b\},\{b\}) = \frac{1}{3}$.

[2] Priority goes to [7], where this metric first appears with an application to the comparison of biotopes. The metric was indepently rediscovered in [5] and was proved to satisfy the triangle inequality in [1] with a simpler argument than in [7].

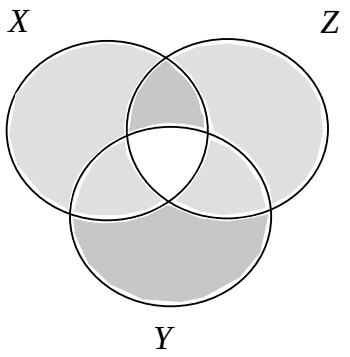

Figure 1: Shaded areas $X \triangle Y$, $Y \triangle Z$, with doubly shaded intersection.

distance is generalized in a quite unexpected way.

3. A METRIC FOR FINITE SETS

For any two finite sets that are not both empty, define as above

$$d(X,Y) = |X \triangle Y| \div |X \cup Y|$$

and complete the definition of d by putting $d(\varnothing,\varnothing) = 0$. In this section we prove that d is a metric. By definition, $d(X,Y)$ is a rational number between 0 and 1. We have $d(X,Y) = 1$ if and only if X and Y are disjoint, in particular so if X or Y is empty but not both. Obviously, d is symmetric (since $\cup$ and $\triangle$ are so), and we have $d(X,Y) = 0$ if and only if $X = Y$. In order to prove that d is a metric, it remains to show that d satisfies the triangle inequality $d(X,Z) \leq d(X,Y) + d(Y,Z)$ for all finite sets X, Y, Z. Below we shall use $X \uplus Y$ for the union of *disjoint* sets X and Y. Note that this so-called disjoint union is symmetric and associative. We write $X \uplus Y \uplus Z$ for the union of the sets X, Y, Z when these sets are pairwise disjoint. For the proof of the triangle inequality we need the following lemma.

LEMMA 1. For all sets X, Y, Z we have $(X - Y) \cup (Y - X) \cup (Z - Y) \cup (Y - Z) = ((X \cup Z) - (X \cap Z)) \uplus (Y - (X \cup Z)) \uplus ((X \cap Z) - Y)$

PROOF. Both sides are equal to the set $S = (X \cup Y \cup Z) - (X \cap Y \cap Z)$. The left hand side is equal to S since this set consist of those elements that occur in at least one and at most two of the sets X, Y, Z. The right hand side is equal to S on the basis of the following calculation: $(X \cup Y \cup Z) - (X \cap Y \cap Z) = ((X \cup Z) - (X \cap Y \cap Z)) \uplus ((Y - (X \cup Z)) - (X \cap Y \cap Z)) = ((X \cup Z) - (X \cap Y \cap Z)) \uplus (Y - (X \cup Z)) = ((X \cup Z) - (X \cap Z)) \uplus ((X \cap Z) - Y) \uplus (Y - (X \cup Z))$. Perhaps most convincing is Figure 1 above. □

If one or more of the sets X, Y, Z is empty, then the triangle inequality trivially holds. Now assume that X, Y, Z are finite non-empty sets. Using the elementary fact $|A \uplus B| = |A| + |B|$ (additivity), and its immediate consequences $|A \cup B| \leq |A| + |B|$ and $A \subseteq B \Rightarrow |A| \leq |B|$ as well as the lemma above, we can calculate

$$\begin{aligned} d(X,Y) + d(Y,Z) &= \frac{|(X-Y) \cup (Y-X)|}{|X \cup Y|} + \frac{|(Y-Z) \cup (Z-Y)|}{|Y \cup Z|} \\ &\geq \frac{|(X-Y) \cup (Y-X) \cup (Y-Z) \cup (Z-Y)|}{|X \cup Y \cup Z|} \end{aligned}$$

$$
\begin{aligned}
(1) \qquad & = \frac{|((X \cup Z) - (X \cap Z)) \uplus (Y - (X \cup Z)) \uplus ((X \cap Z) - Y)|}{|X \cup Y \cup Z|} \\
& \geq \frac{|((X \cup Z) - (X \cap Z)) \uplus (Y - (X \cup Z))|}{|X \cup Y \cup Z|} \\
& = \frac{|((X \cup Z) - (X \cap Z)) \uplus (Y - (X \cup Z))|}{|(X \cup Z) \uplus (Y - (X \cup Z))|} \\
(2) \qquad & \geq \frac{|(X \cup Z) - (X \cap Z)|}{|X \cup Z|} \\
& = d(X, Z)
\end{aligned}
$$

The numbered formulas above should be explained a bit more: (1) uses the lemma; (2) uses that $\frac{x+y}{z+y}$ is monotone in $y \geq 0$ for $z \geq x \geq 0$. This concludes the proof of the triangle inequality.

Observe that the essential properties of $|\,.\,|$ we have used are positivity, $|A| > 0$ for non-empty A and $|\varnothing| = 0$, as well as additivity, $|A \uplus B| = |A| + |B|$. This allows us to accomodate weighting in a very satisfying way. Assume every element x has weight $w(x) > 0$. Define $|\,.\,|_w$ by $|\{x_1, \ldots, x_n\}|_w = w(x_1) + \cdots + w(x_n)$. Then $|\,.\,|_w$ is positive and additive and everything above goes through with $|\,.\,|_w$ instead of $|\,.\,|$.

We finish this section with an easy lemma showing that sets differing more than a factor 2 in size are at distance at least $\frac{1}{2}$. This lemma implies that from a certain point adding more and more symptoms to a clinical picture doesn't necessarily result in a better diagnosis.

LEMMA 2. For all finite sets X and Y, if $d(X,Y) < \frac{1}{2}$, then $|X| \leq 2 \cdot |Y|$.

PROOF. If X and Y are both empty, then the lemma trivially holds. Otherwise, observe $X \cup Y = (X - Y) \uplus (X \cap Y) \uplus (Y - X)$ and put $m = |X - Y|$, $n = |Y - X|$, $k = |X \cap Y|$. Then $d(X,Y) = \frac{m+n}{m+k+n} < \frac{1}{2}$ is equivalent to $m + n < k$, which easily implies $m + k < 2 \cdot (n + k)$, so $|X| \leq 2 \cdot |Y|$. □

4. FROM DISCRETE TO CONTINUOUS

Up to now we have considered objects (syndromes, clinical pictures) that are defined in terms of boolean-valued symptoms. The expressive power of boolean values is limited. Many phenomena are most naturally expressed in terms of continuous quantities. We just mention, in the medical domain: blood pressure, pulse, temperature. Of course one has always the possibility of discretization (such as: blood pressure above or below 130), but this is awkward or even unsound. (Soundness is at stake when the discretization involves more than one boolean, as the symptoms are in our approach essentially independent.) One would like to improve on this by allowing also real values. Thus we are led to reconsider our objects along the lines set out in the following paragraph.

Recall that the set of symptoms is $S = \{0, 1, 2, \ldots\}$. Assuming that symptoms are real-valued, syndromes and clinical pictures can be represented as finite, single-valued, sets of pairs consisting of a symptom and a real value. Such objects can also be viewed as partial functions $X, Y, Z, \ldots$ from S to the real numbers, with the extra restriction that domains are finite. We will write $X(n)$ for the value that X takes at n, that is, the unique value (if any) such that $(n, X(n)) \in X$. The domain of X is by definition the set of symptoms where X is defined, denoted by $\underline{X}$. Other useful notations are $\overline{X}$ for the *total* (= everywhere defined) function extending X with default values on all arguments where X is undefined, and $\|X\|$ for the *norm* of X, being the sum $\sum_{n \in \underline{X}} X(n)$ of all values for arguments in the domain of X.

The notations introduced above are already sufficient to consider possible generalizations of some of the attempts in Section 2. Define

$$f(X,Y) = \textstyle\sum_{n \in \underline{X} \cap \underline{Y}} |X(n) - Y(n)|.$$

Here $|x|$ stands for the absolute value of the real number x, not to be confused with the number $|X|$ of elements of the set X. The function f is a natural extension of the first attempt from Section 2 to the continuous case: only symptoms where X and Y are both defined are taken into account as the summation takes place over $n \in \underline{X} \cap \underline{Y}$. For the same reason as mentioned in Section 2, f is not a metric. One is tempted to consider

$$g(X,Y) = \sum_n |\overline{X}(n) - \overline{Y}(n)|,$$

where the summation is assumed to take place over *all* natural numbers n. The function g is not a metric either, as undefined cannot be distinguished from the default value. However, g is a metric for total functions having finite norm, and as such a natural extension of Hamming distance. When partial functions having the same extension are identified, then g is also a metric for partial functions.

The function g cannot be seen as a generalization of the second attempt in Section 2, since the treatment of `u` as a third value different from `t`, `f` is essentially different from taking default values for `u`. Much better in this respect is

$$h_1(X,Y) = |\underline{X} \triangle \underline{Y}| + \sum_{n \in \underline{X} \cap \underline{Y}} |X(n) - Y(n)|.$$

The first summand counts the number of positions where only one of X, Y is defined and the other is undefined, in the second the differences in value in arguments where X and Y are both defined are summed. The function h_1 is not a metric since $h_1(\{(0,1)\}, \{(0,100)\}) = 99$, whereas $h_1(\varnothing, \{(0,1)\}) = h_1(\varnothing, \{(0,100)\}) = 1$, which violates the triangle inequality. However, h_1 is a metric for partial $[0,1]$-valued (that is, with function values between 0 and 1, inclusive) functions with finite norm. The latter result will follow as a corollary to the proof in the next section, see footnote[3].

The third attempt in Section 2 yields two suggestions for metrics, namely $|\underline{X} \triangle \underline{Y}|$ and $|X \triangle Y|$. The former is obviously not a metric since only the domains and not the differences in values of X and Y are taken into account. The latter yields a metric, but in addition to the shortcomings described already in Section 2 we have that $|\{(0,0)\} \triangle \{(0,x)\}| = 2$ for any $x \neq 0$, which is unacceptable in cases in which the value x really matters. A more realistic generalization of $|X \triangle Y|$ is

$$h_2(X,Y) = |\underline{X} \triangle \underline{Y}| + 2 \cdot \sum_{n \in \underline{X} \cap \underline{Y}} |X(n) - Y(n)|.$$

Note the factor 2 before the $\sum$-symbol, which is the only difference between h_2 and h_1. For boolean-valued X, Y, putting $\mathtt{t} = 1, \mathtt{f} = 0$, we have $|X \triangle Y| = h_2(X,Y)$ (see Lemma 3 below). Surprisingly, h_2 is a metric for partial $[0,1]$-valued functions with finite norm, again by a corollary to the proof in the next section, see footnote[3].

Note that h_1 and h_2 are only metrics for partial $[0,1]$-valued functions with finite norm. We have not yet considered the case in which function values are not restricted to $[0,1]$. In that case there is an incompatibility between the bounded, discrete character of defined versus undefined and the unbounded, continuous character of the function value if it exists. More explicitly, if $X(0)$ is 99 and Y is not defined in 0, then we do not know what to do with the value 99. If, on the other hand, $Y(0) = 100$, then the value 99 becomes of crucial importance for the distance between X and Y. This discussion can equally well be viewed as a plea for $[0,1]$-valued functions, as the difference $|X(0) - Y(0)|$ should be scaled to the maximal range of the continuous quantity behind symptom number 0. For truly real-valued functions we have found no better metric than (essentially) applying the metric for functions defined in [7] to the extensions $\overline{X}$ and $\overline{Y}$. See d_a, d_p at the end of the next section. These are only metrics for partial functions when functions X having the same extension $\overline{X}$ are identified. For a number of applications this identification is unrealistic, for example in the medical domain there is no sensible default value for characteristics like age, pregnancy, etcetera.

Like in Section 2, the idea is to norm the metric h_2. The fourth attempt tells us how *not* to do this. The fifth attempt will be generalized in the next section.

5. A METRIC FOR PARTIAL $[0,1]$-VALUED FUNCTIONS WITH FINITE DOMAINS

For all real-valued functions X, Y with finite norm, define

$$\Delta(X,Y) = \sum_{n \in \underline{X} \cap \underline{Y}} |X(n) - Y(n)|.$$

In fact, Δ is the function f rejected above as metric. We prepare the generalization of $d(X,Y) = |X \triangle Y| \div |X \cup Y|$ by the following lemma.

LEMMA 3 For boolean-valued X, Y as above, putting $\mathtt{t} = 1, \mathtt{f} = 0$, we have $|X \triangle Y| = |\underline{X} \triangle \underline{Y}| + 2 \cdot \Delta(X,Y)$ and $|X \cup Y| = |\underline{X} \cup \underline{Y}| + \Delta(X,Y)$.

PROOF. We have $(n, \cdot) \in X \triangle Y$ if and only if either $n \in \underline{X} \triangle \underline{Y}$, or $n \in \underline{X} \cap \underline{Y}$ and $X(n) \neq Y(n)$. In the latter case, the set $X \triangle Y$ contains the two different pairs $(n, X(n))$, $(n, Y(n))$ and $|X(n) - Y(n)| = 1$, which explains the factor 2. The second identity is even simpler. □

For partial $[0,1]$-valued functions X, Y with finite domains, X and Y not both empty, define

$$d(X,Y) = \frac{|\underline{X} \triangle \underline{Y}| + 2 \cdot \Delta(X,Y)}{|\underline{X} \cup \underline{Y}| + \Delta(X,Y)},$$

putting $d(\varnothing, \varnothing) = 0$. By Lemma 3 above, d is an extension of the metric d from Section 3, obviously satisfying conditions (i) and (ii) from Section 1. In order to prove that d is a metric for $[0,1]$-valued partial functions with finite domains, it remains to verify condition (iii), the triangle inequality.

Recall that $|A|$ denotes the number of elements of the finite set A, and $|x|$ the absolute value of the number x. In order to prepare the proof of the triangle inequality, we treat the counters of the fractions involved first, and then the numerators. Using $|A| + |B| = |A \cup B| + |A \cap B|$ for sets A, B, we have

$$|\underline{X} \triangle \underline{Y}| + |\underline{Y} \triangle \underline{Z}| = |(\underline{X} \triangle \underline{Y}) \cup (\underline{Y} \triangle \underline{Z})| + |(\underline{X} \triangle \underline{Y}) \cap (\underline{Y} \triangle \underline{Z})|$$

which equals

$$|(\underline{X} \triangle \underline{Z}) \uplus (\underline{Y} - (\underline{X} \cup \underline{Z})) \uplus ((\underline{X} \cap \underline{Z}) - \underline{Y})| + |(\underline{Y} - (\underline{X} \cup \underline{Z})) \uplus ((\underline{X} \cap \underline{Z}) - \underline{Y})|$$

by Lemma 1 and an easy variation thereof, see Figure 1.

The sum $\Delta(X,Y) + \Delta(Y,Z)$ can be estimated as follows.

$$\begin{aligned} \Delta(X,Y) + \Delta(Y,Z) &\geq \sum_{n \in \underline{X} \cap \underline{Y} \cap \underline{Z}} (|X(n) - Y(n)| + |Y(n) - Z(n)|) \\ &\geq \sum_{n \in \underline{X} \cap \underline{Y} \cap \underline{Z}} |X(n) - Z(n)| \\ &= \sum_{n \in \underline{X} \cap \underline{Z}} |X(n) - Z(n)| - \sum_{n \in (\underline{X} \cap \underline{Z}) - \underline{Y}} |X(n) - Z(n)| \\ &\geq \Delta(X,Z) - |(\underline{X} \cap \underline{Z}) - \underline{Y}| \qquad (3) \end{aligned}$$

In the last step we use that the functions are $[0,1]$-valued. Note that $(\underline{X} \cap \underline{Z}) - \underline{Y}$ occurs two times in the previous paragraph.

Combining the results of the last two paragraphs, we have the following estimation[3] for the sum of the counters in $d(X,Y) + d(Y,Z)$:

$$|\underline{X} \triangle \underline{Y}| + 2 \cdot \Delta(X,Y) + |\underline{Y} \triangle \underline{Z}| + 2 \cdot \Delta(Y,Z) \geq |\underline{X} \triangle \underline{Z}| + 2 \cdot \Delta(X,Z) = cnt_1.$$

Apart from cnt_1 we need another lower bound for the sum of the counters in $d(X,Y) + d(Y,Z)$, namely

$$|\underline{X} \triangle \underline{Z}| + 2 \cdot (|\underline{Y} - (\underline{X} \cup \underline{Z})| + \Delta(X,Y) + \Delta(Y,Z)) = cnt_2.$$

[3] This inequality holds with and without the factors 2 and implies the triangle inequality for h_1, h_2 in the previous section.

The numerators in $d(X,Y)+d(Y,Z)$ have the following upper bound

$$\begin{aligned} num &= |\underline{X} \cup \underline{Y} \cup \underline{Z}| + \Delta(X,Y) + \Delta(Y,Z) \\ &= |\underline{X} \cup \underline{Z}| + |\underline{Y} - (\underline{X} \cup \underline{Z})| + \Delta(X,Y) + \Delta(Y,Z) \end{aligned}$$

In order to verify the triangle inequality for d we distinquish two cases. In the first case we have $\Delta(X,Z) \geq |\underline{Y} - (\underline{X} \cup \underline{Z})| + \Delta(X,Y) + \Delta(Y,Z)$. Then

$$d(X,Y) + d(Y,Z) \geq \frac{cnt_1}{num} \geq \frac{|\underline{X} \triangle \underline{Z}| + 2 \cdot \Delta(X,Z)}{|\underline{X} \cup \underline{Z}| + \Delta(X,Z)} = d(X,Z)$$

by using $num \leq |\underline{X} \cup \underline{Z}| + \Delta(X,Z)$. In the (complementary) second case we have $\Delta(X,Z) < |\underline{Y} - (\underline{X} \cup \underline{Z})| + \Delta(X,Y) + \Delta(Y,Z)$. Then we have

$$d(X,Y) + d(Y,Z) \geq \frac{cnt_2}{num} \geq \frac{|\underline{X} \triangle \underline{Z}| + 2 \cdot \Delta(X,Z))}{|\underline{X} \cup \underline{Z}| + \Delta(X,Z)} = d(X,Z)$$

by using that $\frac{x+2y}{z+y}$ is monotone in $y \geq 0$ for $z \geq x \geq 0$. This completes the proof of the triangle inequality.

The introduction of weights is slightly more complicated than in the case of the metric for finite sets. Assume every natural number n has weight $w(n) > 0$. Define $|\,.\,|_w$ by $|\{n_1, \ldots, n_k\}|_w = w(n_1) + \cdots + w(n_k)$. Then $|\,.\,|_w$ is positive and additive, and the only critical step in the proof above is the last inequality (3) in the estimation of $\Delta(X,Y) + \Delta(Y,Z)$, which relies on

$$\textstyle\sum_{n \in (\underline{X} \cap \underline{Z}) - \underline{Y}} |X(n) - Z(n)| \leq |(\underline{X} \cap \underline{Z}) - \underline{Y}|$$

This suggests how d should be redefined as d_w in the weighted case:

$$d_w(X,Y) = \frac{|\underline{X} \triangle \underline{Y}|_w + \sum_{n \in \underline{X} \cap \underline{Y}} w(n) \cdot |X(n) - Y(n)|}{|\underline{X} \cup \underline{Y}|_w}$$

Then we have

$$\textstyle\sum_{n \in (\underline{X} \cap \underline{Z}) - \underline{Y}} w(n) \cdot |X(n) - Z(n)| \leq |(\underline{X} \cap \underline{Z}) - \underline{Y}|_w$$

and one can easily verify that the whole proof above for d goes through for d_w.

In addition to the introduction of weights, several (minor) variations of d are possible. We mention two of them, $d_{1,1}$ and $d_{1,0}$. The proofs of the triangle inequality for $d_{1,1}$ and $d_{1,0}$ are very similar to the proof above. For the sake of completeness we also provide two metrics which follow directly from [7]. Metric d_a is a metric for total functions taking arbitrary real values, Metric d_p is the restriction of d_a to non-negative total functions. The latter two metrics should be used with care. They are defined in terms of total extensions. Only finitely many symptoms can have default value different from 0, as the norm of every extension must be finite. Partial functions having the same extension are identified.

$$d_{1,1}(X,Y) = \frac{|\underline{X} \triangle \underline{Y}| + \Delta(X,Y)}{|\underline{X} \cup \underline{Y}| + \Delta(X,Y)},$$

$$d_{1,0}(X,Y) = \frac{|\underline{X} \triangle \underline{Y}| + \Delta(X,Y)}{|\underline{X} \cup \underline{Y}|},$$

$$d_a(X,Y) = \frac{\sum_n |\overline{X}(n) - \overline{Y}(n)|}{\sum_n \max\left(|\overline{X}(n)|, |\overline{Y}(n)|, |\overline{X}(n) - \overline{Y}(n)|\right)}$$

$$d_p(X,Y) = \frac{\sum_n |\overline{X}(n) - \overline{Y}(n)|}{\sum_n \max\left(\overline{X}(n), \overline{Y}(n)\right)}$$

6. Application to information retrieval

In the previous sections several metrics have been proposed which are able to deal with heterogeneous objects. In this section we discuss applying these distance measures in the field of information retrieval, more precisely to the purpose of *document retrieval*. The goal is to retrieve documents 'on the same subject' from a large document base.

A first and important remark is that the full text of a document is the wrong level of abstraction for (semantic) comparison, and this would moreover be computationally unfeasible because of the large number of (possibly long) documents stored in nowadays information disclosure systems. It follows that we need to represent documents by so-called *document profiles*.

Document profiles

A document profile is an abstract representation of a textual document. For our analysis of document content we abstract from words to *terms*. Different words such as 'walking', 'walked' and 'walks' refer to the same semantic term 'walk'. Each document can now be represented by a document profile which contains the relevant terms of the original document, annotated with their so-called *weights*. The weight of a term is a real number between 0, expressing that the term is not important at all, and 1, expressing that the term is very important. Weights can be calculated on the basis of well known techniques for information disclosure systems. See for example [14]. These techniques and their underlying linguistic and statistical theory are beyond the scope of this paper.

There are (at least) three possible ways of representing documents by profiles. The first is called the *inverted file representation*. Let k be the length of the list of all relevant terms in the entire document base. Now represent an arbitrary documents as a k-vector of weights. Representations will thus be sparse k-vectors, as most of the weights can be expected to be 0. Nevertheless, as k can be very large, this is not considered a computationally feasible representation.

The second way of representing documents tries to remedy the computational drawback of the first by restricting the list of relevant terms to those with maximal discriminating potential. This means the restriction to those terms that allow one to clearly distinguish the documents in which they occur (with a certain weight) from other documents. This makes the representation computationally feasible, but at the expense of an inevitable loss of information, and we will call this the *incomplete static representation*. Another disadvantage is that the choice of relevant terms can be very critical, and should be reconsidered when documents are added to the document base. A change in the list of relevant terms requires all document profiles to be recalculated.

The third way is called the *complete dynamic representation*. For this representation we focus on each document separately. First the weight of each term of the document is determined, which is to be understood as the importance of the term *for the document*. Instead of a vector, the representation of a document consists of a finite, single-valued set of pairs consisting of a term and its corresponding weight in the represented document. This is computationally feasible, there is no serious loss of information with respect to the document, and it is unproblematic to add documents to the document base. The only problem is that, unlike the previous two representations, there is no straightforward way to define the distance between two such document profiles. Exactly here the metrics proposed in this paper come in.

Application of the metric d in the discrete case

Point of departure is the complete dynamic representation of each document. One straightforward approach is to select of every document the, say, 20 heaviest terms and consider this set as the document profile. Then the metric d from Section 3 can be used to measure distances between document profiles.

In order to get some insight in the behavior of the metric d, we have applied this distance measure to the Cranfield collection [10]. This physics based collection of about 1400 documents is housed at the University of Glasgow. The Cranfield collection can be used for testing and benchmarking since it is supplied together with a file containing relevancy judgments. For a number of standard queries,

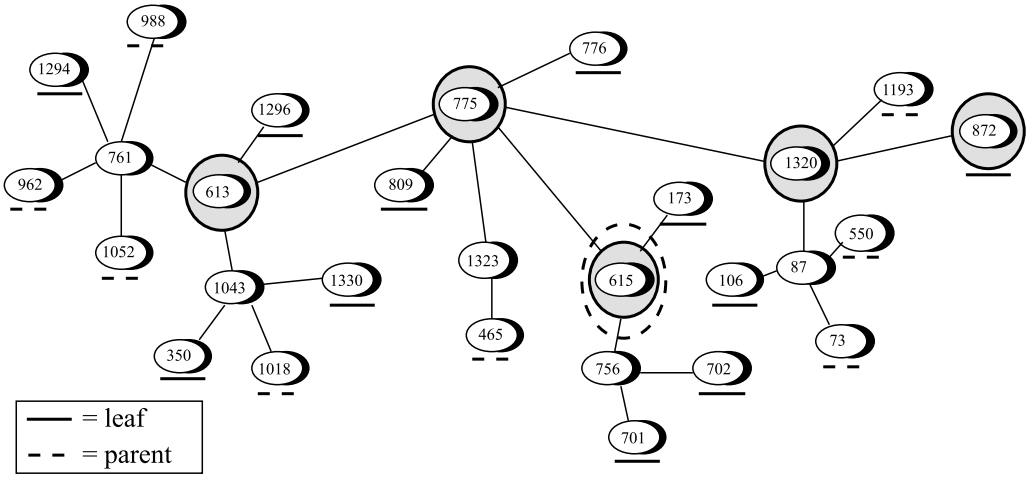

Figure 2: Detail of the acyclic graph representing the cluster structure.

the documents are graded with respect to a relevancy scale ranging from 1 to 5:

1. References which are a complete answer to the query.
2. References of a high degree of relevance, the lack of which either would have made the research impracticable or would have resulted in a considerable amount of extra work.
3. References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.
4. References of minimum interest, for example, those that have been included from an historical viewpoint.
5. References of no interest.

Our experiment consists of establishing, on the basis of the distance measure d, the cluster structure of of the Cranfield document collection. The hope is of course that 'nearby in the cluster structure' coincides with 'high relevancy scores'. Apart from the distance measure, the clustering algorithm is an important parameter of the experiment. A detailed desciption of the clustering algorithm, as well as the quantitative evaluation of the results are beyond the scope of this paper. Therefore we only present the results as an encouraging example, and do not wish to draw scientific conclusions from them. More information on the example can be found in [2]. We also stipulate here the general difficulty of the proper scientific evaluation of information disclosure systems, a priority of future research.

A standard query (about progress on unsteady dynamics, but this is not essential for the description here) leads to the structure of the Cranfield collection as shown in part in Figure 2. This structure can be described as an acyclic graph with document 615 as entry point, with end points underlined, where dashed lines indicate that the structure continues but has been suppressed in the figure. According to [10], the relevancy scores of documents on the query in question are 613-1, 615-1, 775-2, 872-3, 1320-3. These documents appear within a reasonable distance from the entry point 615, which is indeed one of the two best matching documents. Admittedly, the result would be better with 613 closer to 615, and with some irrelevant documents such as 173 at a larger distance. We nevertheless consider the result as encouraging.

Application of continuous distance measures
In this subsection we briefly touch upon the use of continuous distance measures in the clustering algorithm. As an example, consider document profiles p_1, p_2, p_3 of three documents: $p_1 = \{(t_1, 1.0), (t_2, 0.1)\}$, $p_2 = \{(t_1, 1.0), (t_3, 0.4)\}$ and $p_3 = \{(t_1, 0.3), (t_4, 0.1)\}$. For reasons of simplicity, these profiles only involve $t_1, \ldots, t_4$. For three metrics we have calculated the distances in the triangle of profiles in Figure 3 below. The main advantage of a continuous distance measure over the

distance between:	p_1, p_2	p_1, p_3	p_2, p_3
d, discrete	0.67	0.67	0.67
$d_{1,0}$, continuous	0.67	0.90	0.90
d_p, continuous	0.33	0.75	0.80

Figure 3: Distances in the triangle of profiles p_1, p_2, p_3.

discrete one is a higher resolution. Recall that the discrete distance measure does not take the weights into account. Therefore, as becomes apparent from the first row of the above table, the discrete distance measure is not able to discriminate between the distances of any two of the documents. (The fact that t_1 has the same weight in p_1 and p_2, and quite a different weight in p_3 is not taken into account.) This is counter intuitive, one would expect the distance between p_1 and p_2 to be smaller than the other two distances. This is indeed the case for the continuous distance measures. The continuous distance measures take into account the weights of terms occurring in both document profiles, and see therefore more variety in the distances. We expect generally better results with a continuous than with a discrete distance measure. Comparing $d_{1,0}$ and d_p we observe that the latter also takes into account the term weights of the symmetric difference. It is not straightforward to semantically motivate the need for such a refinement. On one hand one could argue that term weights in the symmetric difference do not contribute to the similarity of two documents. On the other hand, we could observe that using all available information results in a more discriminating distance measure. We need to keep in mind that in case of document profiles, the symmetric difference set usually will be much bigger then the intersection. Therefore the application of $d_{1,0}$ probably results in high rated distances.

7. Other applications

The field of genetic algorithms [4] traditionally uses bitstrings to represent genotypes, therefore the Hamming distance is a natural metric to measure distance between these genotypes. Roughly speaking, a genetic algorithm generates repeatedly new individuals, whose bitstrings can be compared with the other individuals of the population. For example, a new individual could replace an old one that is at minimal Hamming distance, thus preserving the genetic diversity of the population.

Genetic *programming* [6] differs from genetic *algorithms* [4] in one important aspect: the 'individuals' are parse trees of computer programs instead of bitstrings. Because of this different structure (involving variable length), the Hamming distance does not apply directly to genetic programming. In [5] first steps are taken to use the generalized Hamming distance introduced in the previous sections to measure distance between parse trees and populations of parse trees.

With the generalized Hamming distance from Section 3 applied to finite sets of subtrees, it is possible to monitor the exploration rate (defined as the amount of change in genetic material between subsequent generations). Moreover, the likelyhood that two parse trees have a common ancestry can be estimated and several strategies can be developed to maintain the structural diversity of the population.

Acknowledgements

We thank Thierry Coquand for drawing our attention to [7] and Ameen Abu-Hanna for valuable criticism on earlier versions of this paper.

References

1. M.A. Bezem and M. Keijzer. Generalizing Hamming distance to finite sets. In D. Patterson e.a., editors, *Proceedings PACES/SPICIS'97*, pp. 148–153, Nanyang Technological University, Singapore, Februari 1997.
2. K. Blok. *Applying Machine Learning Algorithms for Document Retrieval.* Master Thesis, Utrecht University, 1998.
3. S. Cost and S. Salzberg. *A weighted nearest neighbour algorithm for learning with symbolic features.* Machine Learning 10(1): 57–78, 1993.
4. J. Holland. *Adaptation in natural and artificial systems.* MIT Press, 1976.
5. M. Keijzer. Efficiently representing computer programs in genetic programming. Chapter 13 of *Advances in Genetic Programming*, eds. P.J. Angeline and K.E. Kinnear, Complex Adaptive Systems Series 2, MIT Press, 1996.
6. J. Koza. *Genetic Programming: on the programming of computers by means of natural selection.* MIT Press, 1991.
7. E. Marczewski and H. Steinhaus. *On a certain distance of sets and the corresponding distance of functions.* Colloquium Mathematicum VI: 319–327, 1958.
8. J.C. Oxtoby. *Measure and category.* Graduate texts in mathematics, 2, 1971.
9. C.J. van Rijsbergen, *Information Retrieval (Second Edition).* Butterworths, 1979.
10. M. Sanderson. *Cranfield collection 1400.* Department of Computing Science, University of Glasgow. Available at:
`http://local.dcs.gla.ac.uk/idom/ir_resources/test_collections/`
11. C. Stanfill and D. Waltz. *Towards memory-based reasoning.* Communications of the ACM 29(12): 1213–1228, 1986.
12. W.A. Sutherland. *Introduction to metric and topological spaces.* Oxford University Press, 1976.
13. N.N. Webster. *The new lexicon of the english language.* Lexicon Publications Inc., New York, 1990.
14. I.H. Witten and A. Moffat and T.C. Bell. *Managing Gigabytes, Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, New York, 1994.