

Distance measures between atoms

*Ramon J.
Bruynooghe M.
Van Laer W.*

Report CW 264, May 1998



Katholieke Universiteit Leuven
Department of Computer Science

Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

Distance measures between atoms

*Ramon J.
Bruynooghe M.
Van Laer W.*

Report CW 264, May 1998

Department of Computer Science, K.U.Leuven

Abstract

Many learning systems, e.g. systems based on clustering and instance based learning systems, need a measure for the distance between objects. Adequate measures are available for attribute value learners. In recent years there is a growing interest in first order learners, however existing proposals for distances between non-ground atoms have some drawbacks. In this paper we develop a new measure for the distance between non-ground atoms.

Keywords : Machine learning, distances, first order logic.

1 Introduction

In learning systems based on clustering (e.g. C0.5 [3], KBG [1]) and in instance based learning (e.g. [9, ch.4], RIBL [6]), a measure of the distance between objects is an essential component. Good measures exist for distances between objects in an attribute value representation (see e.g. [9, ch. 4]). Recently there is a growing interest in using more expressive first order representations of objects and in upgrading propositional learning systems into first order learning systems (e.g. TILDE [2], ICL [5] and CLAUDIEN [4]). Some ad-hoc similarity measures exist for distances between first order objects [6], but they do not have all the desirable mathematical properties (e.g. the triangle inequality and the positive definiteness property)¹, as a consequence their use may lead to sub-optimal/inconsistent results.

A first step in defining a good distance between first order objects is to define a distance between first order atoms. Some proposals exist for this, e.g. [10] and [8], but they do not handle variables in a satisfactory way. In this paper we propose a better measure for distances between non-ground atoms and prove that it has all the desirable properties of a *metric*. As in [8], the distance between two atoms is based on their distance to the least general generalisation. However, our distances are pairs (F, V) , where F accounts for the differences between the functors of both atoms and V for the difference due to the variables.

For example, the distance between the atoms $p(x, x, y)$ and $p(u, u, v)$ should be 0 as the two atoms are renamings of each other, and $p(a)$ should be closer to $p(x)$ than to $p(b)$ because $p(x)$ subsumes $p(a)$ while $p(a)$ and $p(b)$ are distinct. None of the existing measures produce a distance which is in agreement with these intuitions.

In section 2, we recall some basic concepts about orders, distances and first order logic. In section 3 we prove that a distance can be derived from a semi-distance which is strictly order preserving and satisfies a so called diamond property. In section 4 we develop such a semi-distance for non-ground atoms. We summarize our achievements and discuss future work in Section 5.

2 Preliminaries

We recall some elementary definitions about **order relations**.

Definition 1 (\mathcal{N}, \leq) (a set \mathcal{N} equipped with a binary relation \leq) is a *partial order* iff \leq is reflexive, anti-symmetric and transitive. It is a *total order* iff it is a partial order and $\forall a, b \in \mathcal{N} : (a \leq b) \vee (b \leq a)$.

Example 1 \mathbb{R}, \leq is a total ordered set.

Now we define a special case of string-ordering (only for strings of equal length)

Definition 2 (Order on n-tuples) Let \mathcal{N}^n be the set of n -tuples of elements of \mathcal{N} . On \mathcal{N}^n we define an order \leq_n based on the order \leq on \mathcal{N} : $\forall (u_1, \dots, u_n)$,

¹positive definite: $d(x, y) \geq 0$ and $(d(x, y) = 0 \text{ iff } x = y)$

$(v_1, \dots, v_n) \in \mathcal{N}^n : (u_1, \dots, u_n) \leq_n (v_1, \dots, v_n) \Leftrightarrow$ if $n > 1$ then $(u_1 < v_1) \vee [(u_1 = v_1) \wedge (u_2, \dots, u_n) \leq_{n-1} (v_2, \dots, v_n)]$ else $u_1 \leq v_1$.

Definition 3 (monotonic) Given two partially ordered sets \mathcal{N}_1, \leq_1 and \mathcal{N}_2, \leq_2 and a function $f : \mathcal{N}_1 \rightarrow \mathcal{N}_2$. We say f is monotonic iff $\forall a, b \in \mathcal{N}_1 : a \leq_1 b \Rightarrow f(a) \leq_2 f(b)$, it is anti-monotonic if $\forall a, b \in \mathcal{N}_1 : a \leq_1 b \Rightarrow f(b) \leq_2 f(a)$. In addition, it is strict if $a <_1 b$ implies $f(a) <_2 f(b)$.

Example 2 The function $f : \mathbb{R} \rightarrow \mathbb{R}$ with $f(x) = 2x$ is strictly monotonic. The function $g : \mathbb{R} \rightarrow \mathbb{R}$ with $g(x) = \lfloor x \rfloor$ ($\lfloor x \rfloor$ is the greatest integer smaller than x) is monotonic but not strictly monotonic. The function $h : \mathbb{R} \rightarrow \mathbb{R}$ with $h(x) = -x$ is strictly anti-monotonic.

Definition 4 (convex) A function $\mathbb{N} \rightarrow \mathbb{R}$ is convex iff $\forall m, n, k, l : n \geq k \wedge m \geq l \Rightarrow v(n+m) - v(n) \geq v(k+l) - v(k)$

Example 3 The function $p(x) : \mathbb{R} \rightarrow \mathbb{R}$ with $p(x) = x^2$ is convex because $\forall m, n, k, l : n \geq k \wedge m \geq l \Rightarrow m(2n+m) \geq l(2k+l) \Rightarrow n^2 + 2mn + m^2 - n^2 \geq k^2 + 2kl + l^2 - k^2 \Rightarrow p(n+m) - p(n) \geq p(k+l) - p(k)$

A **distance function** is intended to quantify the difference between two objects. One distinguishes two kinds.

Definition 5 (semi-distance) A semi-distance d over a set of objects \mathcal{O} is a mapping $\mathcal{O} \times \mathcal{O} \rightarrow (\mathcal{N}, +, \leq)$ with $(\mathcal{N}, +)$ a commutative group with neutral element $0_{\mathcal{N}}$, and \leq a total order on \mathcal{N} such that "+" is order-preserving ($a \leq b \wedge c \leq d \Rightarrow a+c \leq b+d$), such that $\forall a, b, c \in \mathcal{O}$:

1. $d(a, a) = 0_{\mathcal{N}}$ and $d(a, b) \geq 0_{\mathcal{N}}$.
2. $d(a, b) = d(b, a)$ (symmetry)
3. $d(a, c) \leq d(a, b) + d(b, c)$ (triangle inequality)

Definition 6 (distance or metric) A distance or metric d is a mapping $\mathcal{O} \times \mathcal{O} \rightarrow (\mathcal{N}, +, \leq)$ which is a semi-distance and $\forall a, b \in \mathcal{O} : d(a, b) = 0_{\mathcal{N}} \Rightarrow a = b$.

Example 4 In the n -dimensional Euclidian space E_n , well known metrics are the euclidian distance ($d_e : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$) and the manhattan distance ($d_m : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$). Let $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n) \in E_n$.

$$d_e(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}$$

$$d_m(x, y) = |x_1 - y_1| + \dots + |x_n - y_n|$$

Definition 7 (order preserving) A (semi-)distance $d : \mathcal{O} \times \mathcal{O} \rightarrow \mathcal{N}$ on a partial ordered set \mathcal{O}, \leq is order preserving iff $\forall a, b, c \in \mathcal{O} : a \leq b \leq c \Rightarrow d(a, b) \leq d(a, c) \wedge d(b, c) \leq d(a, c)$.

Definition 8 (strictly order preserving) A semi-distance d is strictly order preserving iff it is order preserving and $\forall a, b \in \mathcal{O} : a < b \Rightarrow d(a, b) > 0_{\mathcal{N}}$.

We also recall some terminology from **logic**. The set of terms \mathcal{T} is built from the set of variables \mathcal{V} and the set of functors \mathcal{F} : a variable is a term and, with f/n a functor and t_1, \dots, t_n terms, $f(t_1, \dots, t_n)$ is a term. An *atom* is of the form $p(t_1, \dots, t_n)$, with p/n a predicate and t_1, \dots, t_n terms. We denote the set of all atoms with \mathcal{A} .

We will use the notion of *position* as defined in [7]. *Positions* are sequences of positive integers (e.g. [2,3,2]), elements of N_+^* . We use the symbols u, v, w, \dots to denote positions. λ denotes the empty position, and \cdot the concatenation operation on positions. The relation \sqsubseteq in N_+^* defined by $u \sqsubseteq v \Leftrightarrow \exists w, v = u \cdot w$ is the prefix order. With t a term or atom, the set of positions of t , $\mathcal{O}cc(t)$ and the subterm of t at position u , t/u are defined as follows:

- If t is a variable or a constant, then $\mathcal{O}cc(t) = \{\lambda\}$ and $t/\lambda = t$.
- if $t = f(t_1, \dots, t_n)$, then $\mathcal{O}cc(t) = \{\lambda\} \cup \{i \cdot u \mid 1 \leq i \leq n \wedge u \in \mathcal{O}cc(t_i)\}$, $t/\lambda = t$ and $t/(i \cdot u) = t_i/u$.

The subset of positions selecting a subterm which is a variable is denoted $\mathcal{O}cc_V$, those selecting subterms which are non-variable terms are denoted $\mathcal{O}cc_F$.

Let $S, G \in \mathcal{A}$ be logical objects. G is *more general* than S ($G \theta$ -*subsumes* S) ($G \succeq S$) iff there is a substitution θ such that $G\theta = S$. It is a preorder which can be used to define an equivalence relation (*equivalent modulo renaming*). The induced partial order over the quotient set is also denoted \succeq . The Least General Generalisation of two objects, $\text{lgg}(A, B)$ is equal to G iff $G \succeq A$ and $G \succeq B$ and $\forall L \in \mathcal{A} : L \succeq A \wedge L \succeq B \Rightarrow L \succeq G$. The Least Specific Specialisation of two objects, $\text{lss}(A, B)$ is equal to S iff $A \succeq S$ and $B \succeq S$ and $\forall L \in \mathcal{A} : L \preceq A \wedge L \preceq B \Rightarrow L \preceq S$. We extend \mathcal{A} with a special atom \top (treated as a variable) which is more general than all others, so that $\text{lgg}(A, B)$ is always defined.

$\text{Var}(t)$ is a predicate which is true iff t is a variable, $\text{Vars}(t)$ is a function which returns the set of variables occurring in t . $\text{mgu}(t_1, t_2)$ denotes the most general unifier of t_1 and t_2 . Components of substitutions are represented as $x \rightarrow t$.

3 A distance based on a semi-distance

As argued in the introduction, our interest is in a distance between equivalence classes (modulo renaming) of atoms. Following [8] we define a distance based on the notion of Least General Generalisation. In this section we prove a result which holds for any set of logical objects ordered by the more general relation. It shows that a strictly order preserving semi-distance which satisfies a so called *diamond inequality* can be the basis for a distance.

Definition 9 (d_s) *Given a mapping size that maps elements of \mathcal{O} to the n -dimensional space \mathbb{R}^n . We then define $d_s(A, B) = |\text{size}(A) - \text{size}(B)| = \max(\text{size}(A) - \text{size}(B), \text{size}(B) - \text{size}(A))$.*

Example 5 *Let $\text{size}(a) = (1, 2, 3)$ and $\text{size}(b) = (1, 4, 1)$. We have $(1, 2, 3) <_3 (1, 4, 1)$. $d_s(a, b) = (1, 4, 1) - (1, 2, 3) = (0, 2, -2)$.*

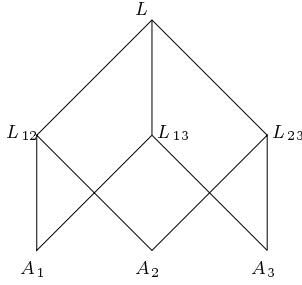


Figure 1: Triangle inequality

Lemma 1 d_s is a semi-distance. If size is strictly monotonic or strictly anti-monotonic, then d_s is strictly order preserving.

Definition 10 (Diamond inequality) Given a partial order (\mathcal{O}, \leq) such that $\text{lgg}(A, B)$ always exists. A semi-distance $d_s : \mathcal{O} \times \mathcal{O} \rightarrow \mathcal{N}$, satisfies the diamond inequality iff the existence of $\text{lss}(A, B)$ implies that $d_s(A, \text{lgg}(A, B)) + d_s(\text{lgg}(A, B), B) \leq d_s(A, \text{lss}(A, B)) + d_s(\text{lss}(A, B), B)$.

Definition 11 (d_l) Given a semi-distance d_s on logical objects. We define $d_l(A, B) = d_s(A, \text{lgg}(A, B)) + d_s(\text{lgg}(A, B), B)$.

Theorem 1 Let d_s be a strictly order preserving semi-distance. If d_s satisfies the diamond inequality then the d_l based on this d_s is a distance.

PROOF

We must prove that d_l satisfies the four properties of a distance as given in definitions 5 and 6.

- $d_l(A, A) = d_s(A, A) + d_s(A, A) = 0_{\mathcal{N}}$ and $d_l(A, B) = d_s(A, \text{lgg}(A, B)) + d_s(\text{lgg}(A, B), B) \geq 0_{\mathcal{N}}$.
- if $d_l(A, B) = 0$ then $d_s(A, \text{lgg}(A, B)) = d_s(\text{lgg}(A, B), B) = 0$. Since d_s is strictly order preserving, we must have $\text{lgg}(A, B) = A = B$.
- $d_l(A, B) = d_s(A, \text{lgg}(A, B)) + d_s(\text{lgg}(A, B), B) = d_s(B, \text{lgg}(B, A)) + d_s(\text{lgg}(B, A), A) = d_l(B, A)$.
- The triangle inequality remains to be proven:

$$d_l(A_1, A_3) \leq d_l(A_1, A_2) + d_l(A_2, A_3) \quad (1)$$

Let $L_{12} = \text{lgg}(A_1, A_2)$, $L_{13} = \text{lgg}(A_1, A_3)$, $L_{23} = \text{lgg}(A_2, A_3)$, and $L = \text{lgg}(A_1, A_2, A_3)$ (see figure 1).

We start with applying the diamond inequality on L_{12} and L_{23} :

$$d_s(L_{12}, L) + d_s(L, L_{23}) \leq d_s(L_{12}, \text{lss}(L_{12}, L_{23})) + d_s(\text{lss}(L_{12}, L_{23}), L_{23}) \quad (2)$$

We have $L_{23} \succeq \text{lss}(L_{12}, L_{23}) = \text{lss}(\text{lgg}(A_1, A_2), \text{lgg}(A_2, A_3)) \succeq A_2$ and similarly, $L_{12} \succeq \text{lss}(L_{12}, L_{23}) \succeq A_2$. As d_s is strictly order preserving, it

follows that $d_s(L_{23}, \text{lss}(L_{12}, L_{23})) \leq d_s(A_2, L_{23})$ and $d_s(L_{12}, \text{lss}(L_{12}, L_{23})) \leq d_s(L_{12}, A_2)$. This allows to reduce the righthandside of (2), yielding:

$$d_s(L_{12}, L) + d_s(L, L_{23}) \leq d_s(L_{12}, A_2) + d_s(A_2, L_{23})$$

Adding $d_s(A_1, L_{12}) + d_s(A_3, L_{23})$ to both sides gives

$$\begin{aligned} d_s(L, L_{12}) + d_s(L, L_{23}) + d_s(A_1, L_{12}) + d_s(A_3, L_{23}) \\ \leq \\ d_s(A_1, L_{12}) + d_s(L_{12}, A_2) + d_s(A_3, L_{23}) + d_s(L_{23}, A_2) \end{aligned} \quad (3)$$

d_s is a semi-distance, so $d_s(L, A_1) \leq d_s(L, L_{12}) + d_s(L_{12}, A_1)$ and $d_s(L, A_3) \leq d_s(L, L_{23}) + d_s(L_{23}, A_3)$

This allows to simplify the lefthandside of equation 3:

$$d_s(L, A_1) + d_s(L, A_3) \leq d_s(A_1, L_{12}) + d_s(L_{12}, A_2) + d_s(A_3, L_{23}) + d_s(L_{23}, A_2)$$

We have that $L \succeq L_{13} \succeq A_1$ and $L \succeq L_{13} \succeq A_3$. As d_s is strictly order preserving, it holds that $d_s(L, A_1) \geq d_s(L_{13}, A_1)$ and $d_s(L, A_3) \geq d_s(L_{13}, A_3)$.

This allows to further reduce the lefthandside:

$$d_s(L_{13}, A_1) + d_s(L_{13}, A_3) \leq d_s(A_1, L_{12}) + d_s(L_{12}, A_2) + d_s(A_3, L_{23}) + d_s(L_{23}, A_2)$$

Applying three times the definition of d_l , the equation reduces to equation 1 which was to be proven. \square .

4 Distance between atoms

According to theorem 1, if we find a strictly order preserving semi-distance between atoms which satisfies the diamond equality, then we can derive a distance between atoms. We will make use of Lemma 1 and base our semi-distance on a strictly anti-monotonic function from the set of atoms \mathcal{A} to the 2-dimensional space \mathbb{R}^2 . The first component considers the functors and to allow the flexibility of giving different importance to components in different positions, we associate a set of positive weights $w_{f,0}, w_{f,1}, \dots, w_{f,n}$ with each functor f/n and a set of positive weights $w_{p,0}, \dots, w_{p,n}$ with each predicate p/n . The definition is parametrised by these weights.

Definition 12 (F-component) $F : \mathcal{A} \cup \mathcal{T} \rightarrow \mathbb{R}$ is defined as:

$$\begin{aligned} F(t) = & \text{if } t = p(t_1, \dots, t_n) \text{ then } w_{p,0} + \sum_{i=1}^n w_{p,i} F(t_i) \\ & \text{else if } t = f(t_1, \dots, t_n) \text{ then } w_{f,0} + \sum_{i=1}^n w_{f,i} F(t_i) \\ & \text{else (a variable) } 0. \end{aligned}$$

Example 6 We use weights 1 in all examples. $F(f(g(a, x), h(x), y)) = w_{f,0} + w_{f,1} \cdot F(g(a, x)) + w_{f,2} \cdot F(h(x)) + w_{f,3} \cdot F(y) = 1 + w_{g,0} + w_{g,1} \cdot F(a) + w_{g,2} \cdot F(x) + w_{h,0} + w_{h,1} \cdot F(x) + 0 = 2 + w_{a,0} + 0 + 1 + 0 = 4$

Lemma 2 F is strictly monotonic.

Note that we can define a $d_{s,g}(A, B) = |f(A) - f(B)|$ and a $d_g(A, B) = d_s(A, \text{lgg}(A, B)) + d(\text{lgg}(A, B), B)$ on ground atoms. If we choose all weights as follows: $w_{p,0} = w_{f,0} = 1/2$, $w_{f/n,i} = w_{p/n,i} = 1/2n$, we can prove that d_g is equal to the distance on ground atoms defined in [10].

The second component takes into account the variables and their multiplicity. It is parametrised by a function $v : \mathbb{N} \rightarrow \mathbb{R}$. The definition makes use of a function $\text{frq}(O, x) =$ the number of positions of variable x in the object O .

Definition 13 (V-component) $V : \mathcal{A} \times (\mathcal{A} \cup \mathcal{T}) \rightarrow \mathbb{R}$ is defined as:

$$V(A, t) = \begin{cases} \text{if } t = p(t_1, \dots, t_n) \text{ then } \sum_{i=1}^n V(A, t_i) \\ \text{else if } t = f(t_1, \dots, t_n) \text{ then } \sum_{i=1}^n V(A, t_i) \\ \text{else (a variable) } v(\text{frq}(A, t)). \end{cases}$$

Lemma 3 $V(A, A) = \sum_{x \in \text{vars}(A)} \text{frq}(A, x) \cdot v(\text{frq}(A, x))$ with $\text{vars}(A)$ the set of all variables occurring in A .

Notation 1 In what follows we use $V(A)$ as a shorthand for $V(A, A)$.

Definition 14 (Size for atoms) $\text{size} : \mathcal{A} \rightarrow \mathbb{R}^2$: $\text{size}(A) = (F(A), V(A))$.

Example 7 Let $A = f(g(a, x), h(x), y)$. $\text{size}(A) = (4, v(1) + 2v(2))$ because $F(A) = 4$ (see example 6) and $V(A) = V(A, g(a, x)) + V(A, h(x)) + v(A, y) = V(A, a) + V(A, x) + V(A, x) + v(A, y) = v(\text{frq}(A, x)) + v(\text{frq}(A, x)) + v(\text{frq}(A, y)) = v(2) + v(2) + v(1) = 2v(2) + v(1)$. We can also apply lemma 3 to calculate $V(A)$: $V(A) = \text{frq}(A, x) \cdot v(\text{frq}(A, x)) + \text{frq}(A, y) \cdot v(\text{frq}(A, y)) = v(1) + 2v(2)$.

Lemma 4 If $v(n)$ is strictly monotonic for $n \geq 1$ then size is strictly anti-monotonic.

PROOF

Consider atoms A_1 and A_2 such that $A_1 \succ A_2$. Let $\theta = \text{mgu}(A_1, A_2)$.

- If θ has a component $x \rightarrow f(t_1, \dots, t_n)$ then there is at least one position u such that $A_1/u = x$ and $A_2/u = f(t_1, \dots, t_n)$. For such positions, $F(A_1/u) < F(A_2/u)$. As a consequence, $F(A_1) < F(A_2)$ and $\text{size}(A_1) < \text{size}(A_2)$.
- If θ has no such component, then $F(A_1) = F(A_2)$. However, then A_1 has at least two variables, say x and y such that $x\theta = y\theta$ as A_1 is strictly more general than A_2 . So there is at least one positions u (e.g. the position of x in A_1) such that $\text{frq}(A_1, A_1/u) < \text{frq}(A_2, A_2/u)$ and thus also $V(A_1, A_1/u) = v(\text{frq}(A_1, A_1/u)) < V(A_2, A_2/u) = v(\text{frq}(A_2, A_2/u))$ because v is strictly monotonic. As a consequence, $V(A_1) < V(A_2)$ and $\text{size}(A_1) < \text{size}(A_2)$. \square

As a consequence of Lemma 1 and lemma 4 we have also:

Corollary 1 If $v(n)$ is strictly monotonic for $n \geq 1$ then $d_s(A, B) = |\text{size}(A) - \text{size}(B)|$ is a strictly order preserving semi-distance.

Next we analyse which property the function v must have to ensure that d_s also satisfies the diamond inequality. First we prove some lemmas. In these lemma's, we assume A and B unifiable atoms and adopt the following notations: $G = \text{lgs}(A, B)$, $S = \text{lss}(A, B)$.

Lemma 5 d_s satisfies the diamond inequality iff $\text{size}(A) + \text{size}(B) \leq \text{size}(G) + \text{size}(S)$.

PROOF

Using the definition of d_s , the diamond inequality reduces to:

$$|size(A) - size(G)| + |size(B) - size(G)| \leq |size(A) - size(S)| + |size(B) - size(S)|$$

As $A \preceq G$ we have $size(A) \geq size(G)$ and similarly $size(B) \geq size(G)$, $size(S) \geq size(A)$, and $size(S) \geq size(B)$.

This allows to rewrite the inequality as:

$$size(A) - size(G) + size(B) - size(G) \leq size(S) - size(A) + size(S) - size(B)$$

or

$$size(A) + size(B) \leq size(G) + size(S)$$

□

Notation 2 With Q an atom and u a position: $U_Q(u) = \{v | v \in \mathcal{O}cc(Q) \text{ and } Q/v = Q/u\}$, i.e. the set of all positions which select the same subterm as u .

Lemma 6 Let A and B be unifiable atoms without common variables. Then $F(A) + F(B) \leq F(G) + F(S)$. Moreover, if $F(A) + F(B) = F(G) + F(S)$ then $\forall u : (u \in \mathcal{O}cc_V(G) \Rightarrow (F(S/u) = F(A/u) \vee F(S/u) = F(B/u)))$.

PROOF

Differences in value between $F(A)$, $F(B)$, $F(G)$, and $F(S)$ are due to the different values of F at positions $v \in \mathcal{O}cc_V(G)$. Now choose such a position $v \in \mathcal{O}cc_V(G)$.

Let $U = U_S(v) \cap \mathcal{O}cc_V(G)$. As for all $u \in \mathcal{O}cc_V(G)$, A/u and B/u are unifiable, and their generalisation is a variable, at least one of them is a variable. Therefore, we can assume $U = \{u_1, \dots, u_k, u_{k+1}, \dots, u_l, u_{l+1}, \dots, u_n\}$ such that $0 \leq k \leq l \leq n + 1$ and $\{u_1, \dots, u_k\} \subseteq \mathcal{O}cc_F(A) \cap \mathcal{O}cc_V(B)$, $\{u_{k+1}, \dots, u_l\} \subseteq \mathcal{O}cc_V(B) \cap \mathcal{O}cc_V(A)$ and $\{u_{l+1}, \dots, u_n\} \subseteq \mathcal{O}cc_V(A) \cap \mathcal{O}cc_F(B)$. We have $\sum_{i=1}^n (F(G/u_i) + F(S/u_i)) = \sum_{i=1}^n F(S/u_i)$ as G/u_i is a variable.

Also $\sum_{i=1}^n (F(A/u_i) + F(B/u_i)) = \sum_{i=1}^k F(A/u_i) + \sum_{i=l+1}^n F(B/u_i)$. Because $A/u_i \succeq S/u_i$, we have $F(A/u_i) \leq F(S/u_i)$ and similarly $F(B/u_i) \leq F(S/u_i)$. As this holds for all v and corresponding U , we have $F(A) + F(B) \leq F(G) + F(S)$. Moreover, if the equality holds, then for $1 \leq i \leq k : F(A/u_i) = F(S/u_i) \wedge F(B/u_i) = 0$, $k + 1 \leq i \leq l : F(A/u_i) = F(B/u_i) = F(S/u_i) = 0$, $l + 1 \leq i \leq n : F(A/u_i) = 0 \wedge F(B/u_i) = F(S/u_i)$ i.e. $\forall i : F(A/u_i) = F(S/u_i) \vee F(B/u_i) = F(S/u_i)$

all choices of v and u_i , either $F(A/u_i) = F(B/u_i) = F(S/u_i) = 0$ ($k = 0$ and $l = n + 1$) or $F(A/u_i) = F(S/u_i) \vee F(B/u_i) = F(S/u_i)$. □

Lemma 7 Let A and B be unifiable atoms without common variables. Given: $\mathcal{O}cc_V(G) = \mathcal{O}cc_V(S)$. If v is convex, then $V(A) + V(B) \leq V(G) + V(S)$.

PROOF

Set $A_G = A$, $S_B = S$ and observe that the following properties hold:

$$\mathcal{O}cc_V(A) = \mathcal{O}cc_V(S) = \mathcal{O}cc_V(S_B) = \mathcal{O}cc_V(A_G). \quad (0)$$

$$V(A) + V(S_B) \leq V(A_G) + V(S). \quad (1)$$

$S \preceq A \preceq A_G \preceq G$, $S \preceq S_B \preceq B \preceq G$, $A_G = \text{lgg}(A, S_B)$, and $S = \text{lss}(A, S_B)$. (2)

$\forall u \in \mathcal{O}ccV(G)$: $U_{A_G}(u) \subseteq U_{S_B}(u)$ (3)

If S_B is a renaming of B then A_G is a renaming of G and the lemma trivially holds. Otherwise, consider the following piece of code:

```

while  $S_B \prec B$  do
   $S_{B,o} = S_B$ ;  $A_{G,o} = A_G$ ;
  Select positions  $u_0, v_0 \in \mathcal{O}ccV(G)$  such that
     $S_{B,o}/u_0 = S_B/v_0$  and  $B/u_0 \neq B/v_0$ ;
  Let  $x, y$  be fresh variables;
   $S_B = S_{B,o}$  except at positions  $u \in U_{S_{B,o}}(u_0)$  where
     $S_B/u = (\text{if } B/u = B/v_0 \text{ then } x \text{ else } y)$ ;
   $A_G = \text{lgg}(A, S_B)$ 

```

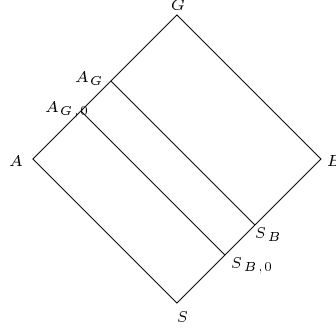


Figure 2: diagram for lemma 7

Observe that u_0 and v_0 always exist as $S_B \prec B$. We show that the while loop preserves the properties (0) - (3). This is fairly straightforward for (0), (2), and (3). Concerning (1), the differences in the V -value of S_B and $S_{B,o}$, A_G and $A_{G,o}$ are due to the V -values at the positions of $U_{S_{B,o}}(u_0)$. Let $\text{frq}(S_{B,o}, S_{B,o}/u_0) = f$, $\text{frq}(S_B, x) = f_x$, and $\text{frq}(S_B, y) = f_y$ with $f = f_x + f_y$. Because of (3), there exists sets U_1, \dots, U_k which are a partition of $U_{S_{B,o}}(u_0)$ and such that for each i : $\exists u_i \in U_{S_{B,o}}(u_0) : U_i = U_{A_{G,o}}(u_i)$. Let $\text{frq}(A_G, A_G/u_i) = f_i$ ($\sum_{i=1}^k f_i = f$). In A_G , the lgg of A and S_B , each set of positions $U_{A_{G,o}}(u_i)$ is split in two sets, a set with size $f_{x,i} (\geq 0)$ where the term is a fresh variable x_i and a set with size $f_{y,i} (\geq 0)$ where the term is a fresh variable y_i ($f_{x,i} + f_{y,i} = f_i$). (1) holds at the beginning of the loop, i.e. $V(A) + V(S_{B,o}) \leq V(A_{G,o}) + V(S)$. To prove that (1) is preserved, it suffices to show that $V(S_B) - V(S_{B,o}) \leq V(A_G) - V(A_{G,o})$ or $\sum_{u \in U_{S_{B,o}}(u_0)} V(S_B, S_B/u) - \sum_{u \in U_{S_{B,o}}(u_0)} V(S_{B,o}, S_{B,o}/u) \leq \sum_{u \in U_{S_{B,o}}(u_0)} V(A_G, A_G/u) - \sum_{u \in U_{S_{B,o}}(u_0)} V(A_{G,o}, A_{G,o}/u)$

This can be written as:

$$\begin{aligned}
& f_x v(f_x) + f_y v(f_y) - f v(f) \leq \sum_{i=1}^k [f_{x,i} v(f_{x,i}) + f_{y,i} v(f_{y,i})] - \sum_{i=1}^k f_i v(f_i) \text{ or} \\
& \sum_{i=1}^k f_{x,i} v(\sum_{i=1}^k f_{x,i}) + \sum_{i=1}^k f_{y,i} v(\sum_{i=1}^k f_{y,i}) - (\sum_{i=1}^k f_{x,i} + \sum_{i=1}^k f_{y,i}) v(\sum_{i=1}^k f_{x,i} + \sum_{i=1}^k f_{y,i}) \\
& \leq \sum_{i=1}^k [f_{x,i} v(f_{x,i}) + f_{y,i} v(f_{y,i})] - \sum_{i=1}^k (f_{x,i} + f_{y,i}) v(f_{x,i} + f_{y,i}) \text{ or} \\
& \sum_{i=1}^k f_{x,i} [v(\sum_{i=1}^k f_{x,i} + \sum_{i=1}^k f_{y,i}) - v(\sum_{i=1}^k f_{x,i}) - v(f_{x,i} + f_{y,i}) + v(f_{x,i})] + \\
& \sum_{i=1}^k f_{y,i} [v(\sum_{i=1}^k f_{x,i} + \sum_{i=1}^k f_{y,i}) - v(\sum_{i=1}^k f_{y,i}) - v(f_{x,i} + f_{y,i}) + v(f_{y,i})] \geq 0
\end{aligned}$$

which is true when

$v(\sum_{i=1}^k f_{x,i} + \sum_{i=1}^k f_{y,i}) - v(\sum_{i=1}^k f_{x,i}) \geq v(f_{x,i} + f_{y,i}) - v(f_{x,i})$ and $v(\sum_{i=1}^k f_{x,i} + \sum_{i=1}^k f_{y,i}) - v(\sum_{i=1}^k f_{y,i}) \geq v(f_{x,i} + f_{y,i}) - v(f_{y,i})$. Both inequalities follow from the fact that v is convex. As a term can only have a finite number of

generalisations, at some point, $S_B = B$ and, from (2), $A_G = G$. This completes the proof. \square

Notation 3 $t[x_1, \dots, x_l]$ represents a term t with $\text{Vars}(t) = \{x_1, \dots, x_l\}$.

Lemma 8 Let A, B , and $G = \text{lgg}(A, B)$ be atoms which differ only at a finite set of positions U in such a way that for $i \in \{1, \dots, k\}, j \in \{1, \dots, n_i\}$: (1) $A/u_{i,j} = t[x_1, \dots, x_l]\theta_i$ where θ_i are variable to variable substitutions, (2) $B/u_{i,j} =$ if $i = 1$ then z_2 else z_i , (3) $G/u_{i,j} = y_i$, (4) for $i \neq j$: $t[x_1, \dots, x_l]\theta_i \neq t[x_1, \dots, x_l]\theta_j$, and that for $i, j \in \{1, \dots, k\}, p, q \in \{1, \dots, l\}, p \neq q$: $x_p\theta_i \neq x_q\theta_j^2$. Let $S = \text{Amgu}(t[x_1, \dots, x_l]\theta_1, t[x_1, \dots, x_l]\theta_2)$. So, for $i \in \{1, \dots, k\}, j \in \{1, \dots, n_i\}$: $S/u_{i,j} =$ if $i = 1$ then $t[x_1, \dots, x_l]\tau_2$ else $t[x_1, \dots, x_l]\tau_i$ with τ_2, \dots, τ_k also variable to variable substitutions. If v satisfies

$$v(n+m) - v(n) \geq v(k+l) - v(k)$$

for natural numbers n, m, k , and l such that $n \geq k$ and $m \geq l$ (i.e. v is convex), then $V(B) - V(S) \leq V(G) - V(A)$.

PROOF

As differences in V -value between A, B, G , and S are due to difference in the positions U , it suffices to prove $\sum_{u \in U} V(B, B/u) - \sum_{u \in U} V(S, S/u) \leq \sum_{u \in U} V(G, G/u) - \sum_{u \in U} V(A, A/u)$ or $\sum_{u \in U} V(S, S/u) - \sum_{u \in U} V(A, A/u) \geq \sum_{u \in U} V(B, B/u) - \sum_{u \in U} V(G, G/u)$. (1)

The positions relevant for the difference in V -value between B and G are: $u_{1,1}, \dots, u_{1,n_1}, u_{2,1}, u_{2,n_2}$. We have: $\sum_{u \in U} V(B, B/u) - \sum_{u \in U} V(G, G/u) = (n_1 + n_2)v(n_1 + n_2) - n_1v(n_1) - n_2v(n_2)$. (2)

The difference in V value between A and S is due to the variables x_i such that $x_i\theta_1 \neq x_i\theta_2$ (There is at least one such variable). Let $I = \{i | x_i\theta_1 \neq x_i\theta_2\}$. Unifying $x_i\theta_1$ and $x_i\theta_2$ does not affect the frequency of variables $x_p\theta_q$ with $p \neq i$. Let $\text{frq}(t[x_1, \dots, x_l], x_i) = f_i (\geq 1)$ and let $p_{i,1} (\geq 0)$ be the number of subterms of A in the positions $u_{3,1}, \dots, u_{k,n_k}$ containing $x_i\theta_1$, and $p_{i,2}$ the similar value for $x_i\theta_2$. Then $x_i\theta_1$ occurs $f_i(n_1 + p_{i,1})$ times in A , $x_i\theta_2$ occurs $f_i(n_2 + p_{i,2})$ times in A , and $x_i\tau_2$ occurs $f_i(n_1 + p_{i,1} + n_2 + p_{i,2})$ times in S .

Thus $\sum_{u \in U} V(S, S/u) - \sum_{u \in U} V(A, A/u) = \sum_{i \in I} [f_i(n_1 + p_{i,1} + n_2 + p_{i,2})v(f_i(n_1 + p_{i,1} + n_2 + p_{i,2})) - f_i(n_1 + p_{i,1})v(f_i(n_1 + p_{i,1})) - f_i(n_2 + p_{i,2})v(f_i(n_2 + p_{i,2}))]$. (3)

Substituting (2) and (3) in (1) gives: $\sum_{i \in I} [f_i(n_1 + p_{i,1} + n_2 + p_{i,2})v(f_i(n_1 + p_{i,1} + n_2 + p_{i,2})) - f_i(n_1 + p_{i,1})v(f_i(n_1 + p_{i,1})) - f_i(n_2 + p_{i,2})v(f_i(n_2 + p_{i,2}))] \geq (n_1 + n_2)v(n_1 + n_2) - n_1v(n_1) - n_2v(n_2)$.

This is true if:

$\sum_{i \in I} [f_i(n_1 + p_{i,1})v(f_i(n_1 + p_{i,1} + n_2 + p_{i,2})) - f_i(n_1 + p_{i,1})v(f_i(n_1 + p_{i,1}))] \geq n_1v(n_1 + n_2) - n_1v(n_1)$ and

$\sum_{i \in I} [f_i(n_2 + p_{i,2})v(f_i(n_1 + p_{i,1} + n_2 + p_{i,2})) - f_i(n_2 + p_{i,2})v(f_i(n_2 + p_{i,2}))] \geq n_2v(n_1 + n_2) - n_2v(n_2)$.

As I is nonempty, $f_i \geq 1$, and $p_{i,1}, p_{i,2} \geq 0$, both inequalities are true under the given condition for v . \square

Theorem 2 Let A and B be unifiable atoms without common variables. If v is convex, then d_s satisfies the diamond inequality.

²E.g. the atoms A_G, G, G_1 , and A_{G_1} in Figure 3.

PROOF

According to Lemma 5 we have to prove that $size(A) + size(B) \leq size(G) + size(S)$. (1)

It follows from Lemma 6 that $F(A) + F(B) \leq F(G) + F(S)$. If $F(A) + F(B) < F(G) + F(S)$ then $size(A) + size(B) \leq size(G) + size(S)$ trivially follows. Otherwise, $F(A) + F(B) = F(G) + F(S)$. In this case, we have to prove $V(A) + V(B) \leq V(G) + V(S)$ (2)

From lemma 6 then also follows that $u \in \mathcal{O}_{cc_V}(G) \Rightarrow (F(S/u) = F(A/u) \vee F(S/u) = F(B/u))$.

Now let $p = \#(\mathcal{O}_{cc_V}(G) \setminus \mathcal{O}_{cc_V}(S))$. We prove the theorem by induction on p . If $p = 0$, $\mathcal{O}_{cc_V}(G) \subset \mathcal{O}_{cc_V}(S)$ from which $\mathcal{O}_{cc_V}(G) = \mathcal{O}_{cc_V}(S)$, so we can apply lemma 7 and obtain (2).

Now we prove (2) for some value of $p > 0$, assuming that the theorem is proved for all values smaller than p .

Let u_1 be such a position. Without loss of generality, we can assume $B/u_1 = y$ and A/u_1 is a non-variable term. Let $U_B(B/u_1) = \{u_{1,1}, \dots, u_{1,n_1}, \dots, u_{k,1}, \dots, u_{k,n_k}\}$ such that for each i , $\{u_{i,1}, \dots, u_{i,n_i}\} = U_G(G/u_{i,1})$ and $G/u_{i,1} = z_i$ (see Figure 3). Because S is the lss of A and B , G their lgg, and because of (2) and (3), there exists a term $t[x_1, \dots, x_l]$ with x_1, \dots, x_l fresh variables and variable to variable substitutions σ and σ_i such that $\forall u \in U_B(B/u_1) : S/u = t[x_1, \dots, x_l]\sigma$ and $\forall u \in U_G(G/u_{i,1}) : A/u = t[x_1, \dots, x_l]\sigma_i$. Now define S_B as $B\{y \leftarrow t[x_1, \dots, x_l]\}$. Because the x_i are fresh variables, their only positions in S_B are inside the subterms at positions in $U_B(B/u_1)$ and differences between the V -value of B and S_B are due to differences at the positions in $U_B(B/u_1)$.

Define A_G as $lgg(A, S_B)$. For $j \in \{1, \dots, n_i\}$, it is possible to write $A_G/u_{i,j}$ as $t[x_1, \dots, x_l]\theta_i$ where the θ_i are also variable to variable substitutions. Moreover, the only difference between G and A_G are at the positions in $U_B(B/u_1)$ and the variables x_i/θ_i cannot occur in A_G outside the subterms A_G/u_i (as the variables x_i do not occur outside the subterms at the positions in $U_B(B/u_1)$ of S_B). So also here, the differences between V -value of G and A_G are due to differences at the positions in $U_B(B/u_1)$.

The number of positions u where $Var(A_G/u) \wedge not(Var(B/u))$ hold is, compared to G and B , reduced by $\sum_{i=1}^k n_i \geq 1$ so we can apply the induction hypotheses (4): $V(A) + V(S_B) \leq V(A_G) + V(S)$, so, to prove $V(A) + V(B) \leq V(G) + V(S)$, it suffices to prove: $V(B) - V(S_B) \leq V(G) - V(A_G)$. (5)

If k , the number of different variables in the positions $u \in U_B(B/u_1)$ of G is 1 then G is a renaming of B and A_G is a renaming of S_B , thus $V(B) - V(S_B) = V(G) - V(A_G)$ and (7) trivially holds. This is the base case for an induction step. Assuming (5) holds up to value $k - 1$, we prove it holds for $k \geq 2$. We define $A_{G_1} = A_G mgu(t[x_1, \dots, x_l]\theta_1, t[x_1, \dots, x_l]\theta_2)$. For $i \in \{2, k\}$, $j \in \{1, n_i\}$ we have $A_{G_1}/u_{i,j} = t[x_1, \dots, x_l]\tau_i$ and for $i = 1$, $j \in \{1, n_1\}$ we have $A_{G_1}/u_{i,j} = t[x_1, \dots, x_l]\tau_2$ for τ_2, \dots, τ_k variable to variable substitutions. We define $G_1 = lgg(A_{G_1}, B)$. For the positions in $U_B(B/u_1)$, G_1 has only $k-1$ different variables (see Figure 3) and, from the induction hypotheses on G_1, A_{G_1}, B , and S_B , it follows that $V(B) - V(S_B) \leq V(G_1) - V(A_{G_1})$. (6)

Lemma 8 is applicable on the atoms G, A_G, G_1 , and A_{G_1} , so $V(G_1) - V(A_{G_1}) \leq V(G) - V(A_G)$. Combining this with (6) one obtains (5). □

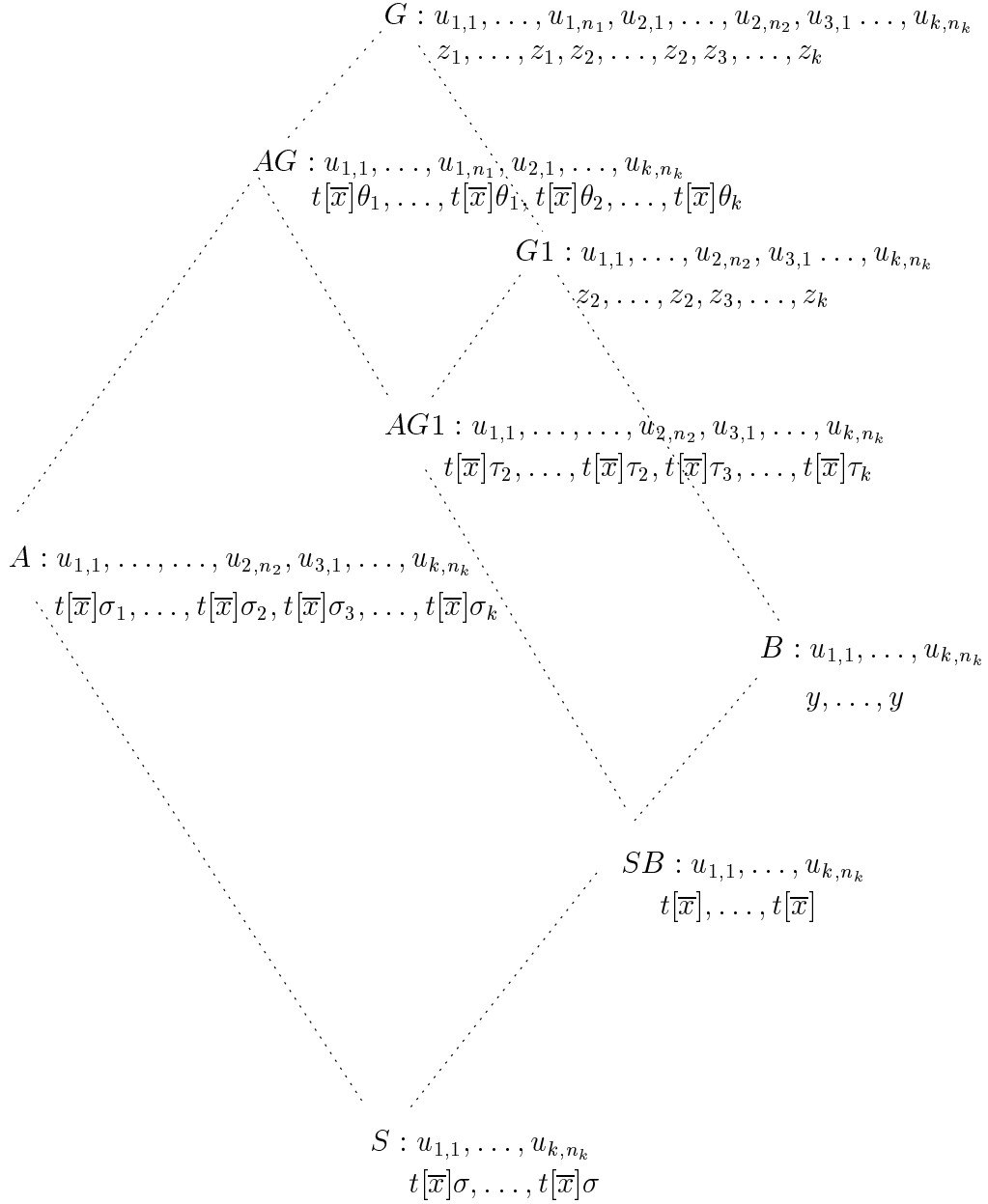


Figure 3: positions $u_{1,1}, \dots, u_{k,n_k}$ where the atoms differ and the subterms at these positions

Theorem 3 *With the proposed mapping $size(A) = (F(A), V(A))$ and $v(n)$ strictly monotonic and convex, d_l is a distance on the set of all atoms.*

PROOF

If $v(n)$ satisfies the given properties, we can apply theorem 2, so $size$ satisfies the diamond inequality. Applying corollary 1, we conclude that d_s is a strictly order preserving semi-distance. Using theorem 1, it follows that d_l is a distance. \square

Example 8 *The function $v(n) = n$ is strictly monotonic and convex.*

5 Conclusion

We have developed a new distance function for atoms. As in [8], the distance between two atoms A and B is based on the difference with their lgg, however, the distance consists of a pair. The first component of the pair is based on the differences between the functors in both terms. It is an extension of the notion of distance used in [10]: it is also defined for non-ground atoms and it introduces weights. The second component is based on the differences in occurrences of variables, it allows to differentiate distances in cases where the first component cannot.

Example 9 $d_l(p(x), q(x)) = d_s(p(x), \top) + d_s(\top, q(x)) = ((1, 1) - (0, 1)) + (1, 1) - (0, 1) = (2, 0)$.
 $d_l(p(x), q(a)) = d_s(p(x), \top) + d_s(\top, q(a)) = ((1, 1) - (0, 1)) + (2, 0) - (0, 1) = (3, -1)$.
 $d_l(p(a, x, x), p(b, y, y)) = d_s(p(a, x, x), p(u, v, v)) + d_s(p(u, v, v), p(b, y, y)) = ((2, 4) - (1, 5)) + ((2, 4) - (1, 5)) = (2, -2)$.
 $d_l(p(a, x, x), p(b, y, z)) = d_s(p(a, x, x), p(u, v, w)) + d_s(p(u, v, w), p(b, y, z)) = ((2, 4) - (1, 3)) + ((2, 2) - (1, 3)) = (2, 0)$.
 $d_l(p(a, x, x), p(a, y, z)) = d_s(p(a, x, x), p(a, v, w)) + d_s(p(a, v, w), p(a, y, z)) = ((2, 4) - (2, 2)) + ((2, 2) - (2, 2)) = (0, 2)$.

The example shows that $p(x)$ is closer to $q(x)$ than to $q(a)$ and that $p(a, x, x)$ is more similar to $p(a, y, z)$ than to $p(b, y, y)$ while $p(b, y, z)$ is still further away. The example shows also that the difference between the arguments a and b counts as much as the difference between the predicate symbols p and q . A better calibration can be obtained by using different weights, e.g. using normalised weights (with sum 1) and giving 0.5 to the predicate/functor while dividing the other 0.5 over the different argument positions.

In future work we will experimentally compare this new distance with those of [8] and [10] in distances for sets of atoms which are parameterised with a distance between atoms. We have described such distances in [11] and [12].

Acknowledgements

We thank Luc De Raedt for the many interesting discussions. We thank also Shan Hwei Nienhuys-Cheng for reading the draft and the useful comments.

Wim Van Laer and Maurice Bruynooghe are supported by the Fund of Scientific Research, Flanders. This work is supported by the European community Esprit project no. 20237, Inductive Logic Programming 2.

References

- [1] G. Bisson. Conceptual clustering in a first order logic representation. In *Proceedings of the 10th European Conference on Artificial Intelligence*, pages 458–462. John Wiley & Sons, 1992.
- [2] H. Blockeel and L. De Raedt. Experiments with top-down induction of logical decision trees. Technical Report CW 247, Dept. of Computer Science, K.U.Leuven, January 1997. Also in Periodic Progress Report ESPRIT Project ILP2, January 1997. <http://www.cs.kuleuven.ac.be/publicaties/rapporten/CW1997.html>.
- [3] L. De Raedt and H. Blockeel. Using logical decision trees for clustering. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, volume 1297 of *Lecture Notes in Artificial Intelligence*, pages 133–141. Springer-Verlag, 1997.
- [4] L. De Raedt and L. Dehaspe. Clausal discovery. *Machine Learning*, 26:99–146, 1997.
- [5] L. De Raedt and W. Van Laer. Inductive constraint logic. In Klaus P. Jantke, Takeshi Shinohara, and Thomas Zeugmann, editors, *Proceedings of the 6th International Workshop on Algorithmic Learning Theory*, volume 997 of *Lecture Notes in Artificial Intelligence*, pages 80–94. Springer-Verlag, 1995.
- [6] W. Emde and D. Wettschereck. Relational instance based learning. In *Proceedings of the 1995 Workshop of the GI Special Interest Group on Machine Learning*, 1995.
- [7] G. Huet. Confluent reductions: Abstract properties and applications to term rewriting systems. *Journal of the Association for Computing Machinery*, 27(4):797–821, 1980.
- [8] A. Hutchinson. Metrics on terms and clauses. In *Proceedings of the 9th European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence, pages 138–145. Springer-Verlag, 1997.
- [9] P. Langley. *Elements of Machine Learning*. Morgan Kaufmann, 1996.
- [10] Shan-Hwei Nienhuys-Cheng. Distance between herbrand interpretations: A measure for approximations to a target concept. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, Lecture Notes in Artificial Intelligence. Springer-Verlag, 1997.
- [11] J. Ramon and M. Bruynooghe. A framework for defining distances between first-order logic objects. In *Proceedings of the 8th International Conference*

on Inductive Logic Programming, Lecture Notes in Artificial Intelligence, pages 271–280. Springer-Verlag, 1998.

- [12] J. Ramon and M. Bruynooghe. A framework for defining distances between first-order logic objects. Technical Report CW 263, Department of Computer Science, Katholieke Universiteit Leuven, 1998. <http://www.cs.kuleuven.ac.be/publicaties/rapporten/-CW1998.html>.