Annals of University of Craiova, Math. Comp. Sci. Ser.
Volume 33, 2006, Pages 67–78
ISSN: 1223-6934

Metrics for Sets of Atoms and Logic Programs

Mircea Preda

Abstract. The information systems that represent entities by using logic programs sometimes need to know if two entities are similar in order to facilitate the knowledge transfer between them. The similarities between such entities must be measured by measuring the distances between the logic programs representing them. The paper proposes a configurable framework to define distances for the class of the logic programs whose answers are or can be converted to sets of ground atoms. The distance between two logic programs is measured regarding to a set of criteria, each criterion being described by a measure function. The qualities of the presented framework are illustrated by comparing it with a well known distance between Herbrand interpretations.

2000 Mathematics Subject Classification. Primary 68T30; Secondary 68Q55.
Key words and phrases. logic programming, predicate logic, software metrics, similarity measures, knowledge reuse.

1. Introduction

Numerous modern information systems need to deal with information that admits logical representations. For example it is common in the case of the Semantic Web systems to represent the knowledge about users by using logic programs. If such Web support systems need to know if two users have similar preferences then they must measure if the two logic programs that represent the users are similar. This paper proposes a configurable framework to define distances for a large class of logic programs. Its configurability allows us to use it in situations that require a careful analysis of the domain's particularities, a usage example for a Web recommendation system being presented in [3]. The similarities between two logic programs are measured regarding a set of criteria, each criterion having attached a measure function that indicates the degree of satisfaction of the criterion by the logic program. These measures are combined in a metric on the set of the logic programs, two combination variants being presented in the paper. The presentation concludes with a comparison between the new proposed framework and a well known distance between Herbrand interpretations for logic programs [2].

2. Distances between logic programs

Two logic programs may be considered similar if they provide in almost all cases similar answers to queries. The *logic equivalence* is a stronger notion, two logic programs are said to be logic equivalent if they provide same answers to queries. Usually, the answers provided by logic programs are sets of atoms or sets of literals. Consequently, a similarity measure on the logic programs space can be defined by defining similarity measures for atoms and sets of atoms. This topic was extensively discussed

Received: July 15, 2005.

in [4, 5, 1]. However, the methods proposed until now cannot be easily adapted to the specificities of a particular domain and they are not efficient for a large number of applications that require a careful study of the domain's characteristics.

In the following paragraphs we will propose an intuitive and configurable framework for measuring similarities between sets of atoms. The similarities between two atoms are given by the similarities between their structures measured according with a set of similarity features. These features will be combined using elements from multi criteria decision theory. Only the ground atoms case will be considered in the paper.

2.1. Mathematical background. Let (A, d) be a metric space where A is a set and d a metric (distance function). A metric space is bounded if the metric is bounded, it exists $m \in \mathbb{R}_+$ such that $d(x, y) \leq m, \forall x, y \in A$.

Definition 2.1. *Let (A, d) be a metric space bounded by a constant m and let $C(A)$ be the family of the all closed subsets of A. The mapping $h : C(A) \times C(A) \to \mathbb{R}$ defined by*

$$h(S,T) = \begin{cases} \max\{\sup_{x\in S}\{\inf_{y\in T} d(x,y)\}, \\ \sup_{x\in T}\{\inf_{y\in S} d(x,y)\}\} \text{ if } S \neq \emptyset, T \neq \emptyset \\ 0 \text{ if } S = T = \emptyset \\ m \text{ if } S = \emptyset \neq T \text{ or } T = \emptyset \neq S \end{cases},$$

is named the Hausdorff *metric induced by d.*

Definition 2.2. *Let (A, d) be a metric space bounded by a constant m and let $C(A)$ be the family of the all closed subsets of A. The function $md : C(A) \times C(A) \to \mathbb{R}$ defined by:*

$$md(S,T) = \begin{cases} \frac{1}{2\max\{|S|,|T|\}}\cdot \\ \left(\sum\limits_{x\in S}\inf\limits_{y\in T} d(x,y) + \sum\limits_{x\in T}\inf\limits_{y\in S} d(x,y)\right) \\ \text{if } S \neq \emptyset, T \neq \emptyset \\ 0 \text{ if } S = T = \emptyset \\ m \text{ if } S = \emptyset \neq T \text{ or } T = \emptyset \neq S \end{cases},$$

is named the minimum distances sum *metric induced by d.*

2.2. A metric for sets of atoms.

Definition 2.3. *Let $\mathcal{F}$ be a finite set of functors (function symbols) and let $\mathcal{P}$ be a finite set of predicate (relation) symbols. A ground term has the form $f(t_1, ..., t_n)$, $n \geq 0$ with f/n an n-arity functor and t_1, ..., t_n ground terms. A ground atom is a construction $p(t_1, ..., t_n)$, $n \geq 0$ with p/n an n-arity predicate symbol and t_1, ..., t_n ground terms. $\mathcal{A}_g$ describes the set of the all ground atoms constructed based on $\mathcal{F}$ and $\mathcal{P}$. $\mathbb{A}_g = 2^{\mathcal{A}_g}$ represents the family of the all subsets of atoms. The function symbols with 0 arity (with 0 arguments) are also named constant symbols or constants.*

Let $\mathbb{V} = \{v_1, v_2, ..., v_n\}$ be a finite nonempty set of criteria used to study the similarities between two sets of atoms. We consider that each criteria $v \in \mathbb{V}$ has attached a quadruple $\mathcal{X}_v = (X_v, 0_v, \oplus, \leq)$ composed from a nonempty set X_v, an internal operation $\oplus : X_v \times X_v \to X_v$, which is associative, commutative and has a neutral element 0_v, and $\leq$ a partial order relation on pe X_v with 0_v the smallest element. Formally, the following properties are satisfied:

i) $x \oplus (y \oplus z) = (x \oplus y) \oplus z, \forall x, y, z \in X_v$;
ii) $x \oplus y = y \oplus x, \forall x, y \in X_v$;

iii) $x \oplus 0_v = x, \forall x \in X_v$. In accordance with the properties i), ii), iii) $(X_v, \oplus)$ is commutative monoid;
iv) $0_v \leq x, \forall x \in X_v$;
v) $\oplus$ maintains the order relation $\leq$. If $x \leq u$ and $y \leq v$ then $x \oplus y \leq u \oplus v$.

Let us suppose that we can define a function $f(v) : \mathbb{A}_g \to X_v$ that satisfies the following properties:

a) $f(v)(A) \geq 0_v, \forall A \subseteq \mathcal{A}_g$;
b) $f(v)(\emptyset) = 0_v$;
c) If $A_{i,i\geq 0} \subseteq \mathbb{A}_g, \forall i, j : A_i \cap A_j = \emptyset$ then $f(v)(\cup_i A_i) = \oplus_i f(v)(A_i)$.

$f(v), v \in V_i, i \in \overline{1,n}$ are pseudo measures over the universe $\mathcal{A}_g$. They are not measures because the property $f(v)(A) = 0_v \Rightarrow A = \emptyset, \forall A \subseteq \mathcal{A}_g$ is not imposed. The purpose of these functions is to describe a set of atoms regarding to a criterion v (the degree of satisfaction of the criterion v by the set of atoms A).

The pseudo measures $f(v)$ can be relatively easy defined by considering the structure of the atoms. The directed graphs provide a good representation of the elements and relationships involved by structures. Consequently, we will present the graph representation of an atom.

Definition 2.4. *Let Γ_t be the graph attached to a term $t = f(t_1, ..., t_n)$. Γ_t is defined by the following 4 rules:*

(1) *If f is a functor with arity 0 (a constant) then Γ_t contains only one node (*the root*) labeled with f.*
(2) *If f is a functor with arity > 0 then Γ_t is composed from a node (*the root*) labeled with $f(t_1, ..., t_n)$ and conected by directed edges with the roots of the graphs $\Gamma_{t_1}, ..., \Gamma_{t_n}$. These directed edges are labeled with the symbols f_1,..., f_n.*
(3) *Γ_t does not include two nodes with same label. Each node from Γ_t can be uniquely identified by its label.*
(4) *Γ_t cannot include two edges with same source, same destination and same label. A directed edge is identified by a triple (s, t, e) with s - the source node label, t - the destination node label and the edge's label e.*

Definition 2.5. *Let Γ_a be the graph attached to a ground atom $a = p(t_1, ..., t_n)$. Γ_a is defined by the following two rules:*

(1) *If p is a predicate symbol with arity 0 then Γ_a consists in a single node labeled with p.*
(2) *If p is a predicate symbol with arity > 0 then Γ_a consists from a node labeled with $p(t_1, ..., t_n)$ that is connected by directed edges with the root nodes of the graphs $\Gamma_{t_1}, ..., \Gamma_{t_n}$. These edges are labeled with the symbols p_1, ...,p_n.*

Γ_a can be represented as a pair $\Gamma_a = (V_a, E_a)$ where V_a and E_a are the sets of nodes, and, respectively, of edges from Γ_a.

As it is illustrated in the following examples, a criterion can be associated to an element that can be found in the graphs attached to the atoms from $\mathcal{A}_g$, an element that can be a node label or an edge label or a subgraph.

Example 2.1. *Let us consider $\mathbb{V} = \{a, b\}$. $\mathcal{X}_a = \mathcal{X}_b = (\mathcal{L}, \emptyset, \cup, \subseteq)$ unde $\mathcal{L}$ is the family of the all sets of sequences of edge labels. We define $f(a)(A)$ to be the set of the sequences of edge labels from the paths that unite the root of a graph attached to an atom from A with a node labeled with the function symbol a.*

$$f(a)(\{g(f(a), g(a, b))\}) = \{g_2 g_1, g_1 f_1\}.$$

$$f(a)(\{g(f(b),b)\}) = \emptyset.$$

The criterion a compares two sets of atoms from the point of view of the position of the function symbol a inside of their structures. Similarly, $f(b)(A)$ is defined as the set of the sequences attached to the paths that unite the root of a graph corresponding to an atom from A with a node labeled with the function symbol b.

$$f(b)(\{g(f(a),g(a,b))\}) = \{g_2g_2\}.$$
$$f(b)(\{g(f(b),b)\}) = \{g_1f_1,g_2\}.$$

Example 2.2. *If the criteria $v \in \mathbb{V}$ is represented by a node label then we can consider $\mathcal{X}_v = (\mathbb{N}, 0, +, \leq)$ and define*

$$f(v)(A) = |\{a \in A | \Gamma_a = (V_a, E_a), v \in V_a\}|. \tag{1}$$

$f(v)(A)$ represents the number of the occurrences of a node labeled with v in the set of the graphs that are attached to the ground atoms from A. By $|.|$ it is represented the cardinal function for sets. Similarly, if v is an edge label, $\mathcal{X}_v = (\mathbb{N}, 0, +, \leq)$ and $f(v)$ can be defined as follows:

$$f(v)(A) = |\{a \in A | \Gamma_a = (V_a, E_a), v \in E_a\}|, \tag{2}$$

the number of occurrences of the edge label v in the graphs $\Gamma_a, a \in A$.

Let $v \in \mathbb{V}$. The binary relation $=_v \subseteq \mathbb{A}_g \times \mathbb{A}_g$ defined by $A =_v B$ if and only if $f(v)(A) = f(v)(B)$ is an equivalence relation (reflexive, symmetric and transitive). Let $d : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+^{n_d}$, for an arbitrary n_d, $n_d \in \mathbb{N}^*$. d is an $=_v$-distance defined on $\mathbb{A}_g$ if and only if the following 3 properties are satisfied:

a) $d(A,B) \geq 0^{n_d}$, $d(A,B) = 0^{n_d} \Leftrightarrow A =_v B, \forall A, B \subseteq \mathcal{A}_g$,
b) $d(A,B) = d(B,A), \forall A, B \subseteq \mathcal{A}_g$,
c) $d(A,B) \leq d(A,C) + d(C,B), \forall A, B, C \subseteq \mathcal{A}_g$.

d is $=_{\mathbb{V}}$-distance if and only if d is $=_v$ - distance $\forall v \in V_i, \forall i \in \overline{1,n}$. If d is $=_{\mathbb{V}}$ - distance then the equivalence $d(A,B) = 0^{n_d} \Leftrightarrow A =_v B, \forall v \in V_i, \forall i \in \overline{1,n}$ is true. We denote by D_v the family of the all functions $d : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+^{n_d}, n_d \in \mathbb{N}^*$ that are $=_v$ - distances. $D_{\mathbb{V}} = \bigcap_{v \in \mathbb{V}} D_v$ represents the family of the all $=_{\mathbb{V}}$ - distances.

Definition 2.6. *We name distance synthesis operator an application $\Theta : D_{v_1} \times ... \times D_{v_n} \to D_{\mathbb{V}}$. If d_{v_i} are $=_{v_i}$ - distances, $\forall i \in \overline{1,n}$ and Θ is a distance synthesis operator then $\Theta(d_{v_1}, ..., d_{v_n})$ is an $=_{\mathbb{V}}$ - distance defined on $\mathbb{A}_g$.*

The following propositions present examples of distance synthesis operators.

Proposition 2.1. *Let $\mathbb{V}$ be a set of criteria partitioned in m families V_1, V_2,..., V_m of related criteria, $V_i \cap V_j = \emptyset, i, j \in \overline{1,m}, i \neq j$ and $\bigcup_{i=1}^m V_i = \mathbb{V}$. The applications $d_a^p : D_{v_1} \times ... \times D_{v_n} \to D_{\mathbb{V}}$ defined by*

$$d_a^p(d_{v_1}, ..., d_{v_n}) : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+^s$$
$$d_a^p(d_{v_1}, ..., d_{v_n})(A,B) = (d_1^p(d_{v_1}, ..., d_{v_n})(A,B), ..., d_m^p(d_{v_1}, ..., d_{v_n})(A,B)) \in \mathbb{R}_+^s \tag{3}$$

where

$$d_i^p(d_{v_1}, ..., d_{v_n})(A,B) = (\textstyle\sum_{v \in V_i} d_v(A,B)^p)^{1/p}, \quad \forall i \in \overline{1,m}, p \in \mathbb{N}^* \tag{4}$$

and

$$d_i^\infty(d_{v_1}, ..., d_{v_n})(A,B) = \max_{v \in V_i}\{d_v(A,B)\}, \forall i \in \overline{1,m}, p = \infty \tag{5}$$

are distance synthesis operators $\forall p \in \mathbb{N}^* \cup \{\infty\}$. The definition domain of these operators is represented by the all sets of distances $(d_{v_1}, ..., d_{v_n})$ with the property that

$\forall d_{v'}, d_{v''} \in V_i, i \in \overline{1,m}\ n_{d_{v'}} = n_{d_{v''}} = n_i$ *(every two distances from same partition subset have same codomain). In this case* $s = n_1 + ... + n_m$.

Proof: During the demonstration the following two notations will be used:

$$d_a^p(d_{v_1}, ..., d_{v_n})(A,B) \stackrel{not}{=} d_a^p(A,B) \text{ and}$$
$$d_i^p(d_{v_1}, ..., d_{v_n})(A,B) \stackrel{not}{=} d_i^p(A,B), \forall i \in \overline{1,m}.$$

The order relation $\leq \subseteq \mathbb{R}_+^n \times \mathbb{R}_+^n$ is defined by $(x_1, ..., x_n) \leq (y_1, ..., y_n)$ if and only if $\exists j \in \overline{1,n}$ such that $x_i = y_i, \forall i \in \overline{1,j}$ and $x_{j+1} < y_{j+1}$.

i) $d_v(A,B) \geq 0^{n_{d_v}}, \forall A, B \subseteq \mathcal{A}_g, \forall v \in \mathbb{V}$. Consequently, $d_i^p(A,B) \geq 0^{n_i}, \forall A, B \subseteq \mathcal{A}_g, \forall i \in \overline{1,m}, \forall p \in \mathbb{N}^* \cup \{\infty\}$. Results that $d_a^p(A,B) \geq (0, ..., 0), \forall A, B \subseteq \mathcal{A}_g, \forall p \in \mathbb{N}^* \cup \{\infty\}$. Moreover, $d_a^p(A,B) = (0, ..., 0)$ involve $d_i^p(A,B) = 0^{n_i}, \forall i \in \overline{1,m}$. Consequently, $d_v(A,B) = 0^{n_{d_v}}, \forall v \in \mathbb{V}$ or, equivalently, $A =_v B, \forall v \in \mathbb{V}$;

ii) Let $A, B, C \subseteq \mathcal{A}_g$ be subsets of ground atoms, $p \in \mathbb{N}^*$ şi $i \in \overline{1,n}$. Then,

$$\begin{aligned} d_i^p(A,B) &= (\sum_{v \in V_i} d_v(A,B)^p)^{\frac{1}{p}} \\ &\leq (\sum_{v \in V_i} (d_v(A,C) + d_v(C,B))^p)^{\frac{1}{p}} \\ &\leq (\sum_{v \in V_i} d_v(A,C)^p)^{\frac{1}{p}} + (\sum_{v \in V_i} d_v(C,B)^p)^{\frac{1}{p}} \\ &= d_i^p(A,C) + d_i^p(C,B). \end{aligned}$$

In same manner

$$\begin{aligned} d_i^\infty(A,B) &= \max_{v \in V_i}\{d_v(A,B)\} \\ &\leq \max_{v \in V_i}\{d_v(A,C) + d_v(C,B)\} \\ &\leq \max_{v \in V_i}\{d_v(A,C)\} + \max_{v \in V_i}\{d_v(C,B)\} \\ &= d_i^\infty(A,C) + d_i^\infty(C,B). \end{aligned}$$

Consequently, $d_a^p(A,B) \leq d_a^p(A,C) + d_a^p(C,B), \forall A, B, C \subseteq \mathcal{A}_g$ and $\forall p \in \mathbb{N}^* \cup \{\infty\}$. □

Example 2.3. *(An extension of the example 2.1)*

Let $\mathcal{L}$ *be the set of the sequences of edge labels. The function* $l_p : \mathcal{L} \times \mathcal{L} \to \mathbb{N}$ *maps to each pair of sequences the length of their common prefix. Let* $l_1, l_2 \in \mathcal{L}$, $l_1 = a_1^1, ..., a_{m_1}^1$, $l_2 = a_1^2, ..., a_{m_2}^2$ *be two sequences of labels. If* $l_p(l_1, l_2) = j$ *then* $a_i^1 = a_i^2, \forall i \in \overline{1,j}$ *şi* $a_{j+1}^1 \neq a_{j+1}^2$. *For example,* $l_p(g_2g_2, g_2) = 1$. l_p *satisfies the following two properties:*

i) $l_p(l_1, l_2) = l_p(l_2, l_1), \forall l_1, l_2 \in \mathcal{L}$;

ii) $l_p(l_1, l_2) \geq \min\{l_p(l_1, l_3), l_p(l_2, l_3)\}, \forall l_1, l_2, l_3 \in \mathcal{L}$.

Proposition 2.2. *The function* $d_l : \mathcal{L} \times \mathcal{L} \to [0,1]$ *defined by:*

$$d_l(l_1, l_2) = \frac{\max\{m_1, m_2\} - l_p(l_1, l_2)}{\max\{m_1, m_2\}}, \tag{6}$$

where $\max\{m_1, m_2\}$ *represents the maximum of the lengths of the sequences* l_1 *and* l_2, *is a distance.*

Proof:

(1) $d_l(l_1, l_2) \geq 0, \forall l_1, l_2 \in \mathcal{L}$. $d_l(l_1, l_2) = 0$ *involves* $\max\{m_1, m_2\} = l_p(l_1, l_2)$ *from where* $l_1 = l_2$;

(2) $d_l(l_1, l_2) = d_l(l_2, l_1), \forall l_1, l_2 \in \mathcal{L}$;

(3) $d_l(l_1,l_2) \leq d_l(l_1,l_3)+d_l(l_3,l_2), \forall l_1,l_2,l_3 \in \mathcal{L}$. *The following notations are used:* $M_{ij} = \max\{m_i,m_j\}, l_{ij} = l_p(l_i,l_j), \forall i,j \in \overline{1,3}$. *In these conditions, we must prove that:*

$$\frac{M_{12}-l_{12}}{M_{12}} \leq \frac{M_{13}-l_{13}}{M_{13}} + \frac{M_{32}-l_{32}}{M_{32}} \Leftrightarrow$$
$$0 \leq M_{12}M_{13}M_{32} + M_{13}M_{32}l_{12} -$$
$$M_{12}M_{32}l_{13} - M_{12}M_{13}l_{32}.$$

The last inequality is true according with the properties of the function l_p. □

Let d_l^h *be the Hausdorff metric induced by* d_l. *For the set of criteria* $\mathbb{V} = \{\{a,b\}\}$ *from the example 2.1 we define*

$$d(a) : \mathbb{A}_g \times \mathbb{A}_g \to [0,1], d_a(A,B) = d_l^h(f(a)(A), f(a)(B))$$

and

$$d(b) : \mathbb{A}_g \times \mathbb{A}_g \to [0,1], d_b(A,B) = d_l^h(f(b)(A), f(b)(B)).$$

In these settings

$$\begin{aligned} d_a^2(\{g(f(a),g(a,b))\},\{g(f(b),b)\}) = \\ = \quad & (d(a)(\{g(f(a),g(a,b))\},\{g(f(b),b)\})^2 + \\ & +d(b)(\{g(f(a),g(a,b))\},\{g(f(b),b)\})^2)^{\frac{1}{2}} \\ = \quad & \sqrt{d_l^h(\{g_2g_1,g_1f_1\},\emptyset) + d_l^h(\{g_2g_2\},\{g_1f_1,g_2\})} \\ = \quad & \sqrt{2} \approx 1.4142. \end{aligned}$$

An obvious problem of the Hausdorff metric is its sensitivity to the extreme points of the sets. In fact, these points establish the distance between two sets.

If the functions $d(a)$ *and* $d(b)$ *are defined using the minimum distances sum metric* d_l^{md} *induced by* d_l*:*

$$d(a) : \mathbb{A}_g \times \mathbb{A}_g \to [0,1], d_a(A,B) = d_l^{md}(f(a)(A), f(a)(B))$$

and

$$d(b) : \mathbb{A}_g \times \mathbb{A}_g \to [0,1], d_b(A,B) = d_l^{md}(f(b)(A), f(b)(B)),$$

then

$$\begin{aligned} d_a^2(\{g(f(a),g(a,b))\},\{g(f(b),b)\}) = \\ = \quad & (d(a)(\{g(f(a),g(a,b))\},\{g(f(b),b)\})^2 + \\ & +d(b)(\{g(f(a),g(a,b))\},\{g(f(b),b)\})^2)^{\frac{1}{2}} \\ = \quad & \sqrt{d_l^{md}(\{g_2g_1,g_1f_1\},\emptyset) + d_l^{md}(\{g_2g_2\},\{g_1f_1,g_2\})} \\ = \quad & \sqrt{1^2 + (\tfrac{1}{4}(\tfrac{1}{2}+1+\tfrac{1}{2}))^2} = \sqrt{\tfrac{5}{4}} \approx 1.1180. \end{aligned}$$

Because $d_a^2(S,T) \in [0,\sqrt{2}], \forall S,T \in \mathbb{A}_g$, *we can scale the results in the range [0,1] and obtain* $d_a^2(\{g(f(a),g(a,b))\},\{g(f(b),b)\}) \approx 0.7805$.

The employed criteria have a strong influence on the degree of similarity between two sets of atoms. For example, let us consider $\mathbb{V} = \{ab\}$ *where* $f(ab) : \mathbb{A}_g \to \mathcal{L}$, $f(ab)(A) = f(a)(A) \cup f(b)(A)$. *The criterion "ab" compares two sets of atoms regarding how the function symbols* f *and* g *are composed for constructing the atoms.*

$$f(ab)(\{g(f(a),g(a,b))\}) = \{g_1f_1, g_2g_1, g_2g_2\} \text{ and}$$
$$f(ab)(\{g(f(b),b)\}) = \{g_1f_1, g_2\}.$$

The distance

$$\begin{aligned} d_a^2(\{g(f(a),g(a,b))\},\{g(f(b),b)\}) = \\ = \quad & \sqrt{d_l^{md}(\{g_1f_1,g_2g_1,g_2g_2\},\{g_1f_1,g_2\})^2} \\ = \quad & \tfrac{1}{6}(0+\tfrac{1}{2}+\tfrac{1}{2}+0+\tfrac{1}{2}) = \tfrac{1}{4} = 0.25. \end{aligned}$$

It can be observed that, regarding to this criterion, the atoms $g(f(a), g(a, b))$ *and* $g(f(b), b)$ *are very similar.*

Remark 2.1. *If the set* $\mathbb{V}$ *satisfies the property:* $\forall A, B \subseteq \mathcal{A}_g : A =_v B, \forall v \in V_i, \forall i \in \overline{1,n} \Rightarrow A = B$ *then the functions* d_a^p *are distances (metrics),*$\forall p \in \mathbb{N}^* \cup \{\infty\}$.

Proposition 2.3. *The distances* $d_a^p, p \in \mathbb{N}^* \cup \{\infty\}$ *are similar with the following meaning: for every* $p, q \in \mathbb{N}^* \cup \{\infty\}$, $\exists a, b \in \mathbb{R}$ *such that* $a \cdot d_a^q(A, B) \leq d_a^p(A, B) \leq b \cdot d_a^q(A, B), \forall A, B \subseteq \mathcal{A}_g$.

Example 2.4. *If* $k = \max_{i \in \overline{1,n}}\{|V_i|\}$, *then* $d_a^\infty(A, B) \leq d_a^2(A, B) \leq \sqrt{k} \cdot d_a^\infty(A, B)$ *and* $d_a^\infty(A, B) \leq d_a^1(A, B) \leq k \cdot d_a^\infty(A, B), \forall A, B \subseteq \mathcal{A}_g$.

In the following paragraphs, we will present several alternate ways to obtain distances defined on the set of the all subsets of atoms starting from pseudo measures that are attached to the criteria from $\mathbb{V}$.

Proposition 2.4. *Let* $v \in \mathbb{V}$ *and* $f(v) : \mathbb{A}_g \to X_v$ *a pseudo measure. The application* $d_v : \mathbb{A}_g \times \mathbb{A}_g \to X_v$, $d_v(A, B) = f(v)((A \setminus B) \cup (B \setminus A))$ *satisfies the three properties that define an* $=_v$ *- distance.*

Proof:

(1) $d_v(A, B) \geq 0_v, \forall A, B \in \mathbb{A}_g$ because 0_v is the smallest element from X_v. Moreover, $d_v(A, B) = 0_v$ involves $f(v)(A \setminus B) \oplus f(v)(B \setminus A) = 0_v$ or, equivalently, $f(v)(A \setminus B) = f(v)(B \setminus A) = 0_v$. Consequently, $f(v)(A) = f(v)(B)$.

(2) $d_v(A, B) = d_v(B, A), \forall A, B \in \mathbb{A}_g$.

(3) Let $A, B, C \in \mathbb{A}_g$. The inequality $d_v(A, B) \leq d_v(A, C) \oplus d_v(C, B)$ can be restated as $f(v)(A \setminus B) \oplus f(v)(B \setminus A) \leq f(v)(A \setminus C) \oplus f(v)(C \setminus A) \oplus f(v)(C \setminus B) \oplus f(v)(B \setminus C)$. This last inequality is true because the internal composition operation $\oplus$ maintains the inequality relation $\leq$.

□

Proposition 2.5. *Let* $v \in \mathbb{V}$ *be a criteria,* $\mathcal{X}_v = (\mathbb{N}, 0, +, \leq)$ *and* $f(v) : \mathbb{A}_g \to \mathbb{N}$ *a pseudo measure. The functions* $d(v) : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+$ *defined by*

$$d(v)(A, B) = \frac{|f(v)(A) - f(v)(B)|}{\max\{f(v)(A), f(v)(B)\}}, \tag{7}$$

$$d(v)(A, B) = \frac{|f(v)(A) - f(v)(B)|}{f(v)(A \cup B)}, \tag{8}$$

$$d(v)(A, B) = f(v)((A \setminus B) \cup (B \setminus A)), \tag{9}$$

$$d(v)(A, B) = \frac{f(v)((A \setminus B) \cup (B \setminus A))}{f(v)(A \cup B)}. \tag{10}$$

are $=_v$ *- distances on* $\mathbb{A}_g$.

Proposition 2.6. *Let* $\mathbb{V}$ *be a set of criteria and* $\alpha_1, \alpha_2, ..., \alpha_n \in \mathbb{R}_+$ *a set of real coefficients. The applications* $d_a^p : D_{v_1} \times ... \times D_{v_n} \to D_{\mathbb{V}}$ *defined by*

$$d_a^p(d_{v_1}, ..., d_{v_n}) : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+^s$$

$$d_a^p(d_{v_1}, ..., d_{v_n})(A, B) = \left(\sum_{i=1}^{n} \alpha_i^p d_{v_i}(A, B)^p\right)^{1/p}, p \in \mathbb{N}^* \tag{11}$$

and $d_a^p(d_{v_1}, ..., d_{v_n})(A, B) = \max_{i \in \overline{1,n}}\{\alpha_i d_{v_i}(A, B)\}, p = \infty$ *are partially defined distance synthesis operators* $\forall p \in \mathbb{N}^* \cup \{\infty\}$. *The definition domain of these operators is represented by the all sets of distances* $(d_{v_1}, ..., d_{v_n})$ *with the property that* $n_{d_{v_1}} = n_{d_{v_2}} = ... = n_{d_{v_n}} = s$ *(all distances have same codomain).*

Remark 2.2. *The coefficient α_i attached to a criterion $v_i, i \in \overline{1,n}$ can be used to specify the relative degree of importance of the criterion. The relative importance coefficients can be determined using methods from multi criteria decision theory like Analitic Hierarchy Process (AHP)* [6].

Example 2.5. *Let us consider the vocabulary $\mathcal{P} = \{p/2\}$, $\mathcal{F} = \{a/0, b/0, f/1, g/2\}$ and the atoms $p(f(f(a)), g(a, f(b)))$, $p(f(f(b)), g(b, f(a)))$, $p(a, g(a, f(b)))$. The degree of similarity of these atoms will be evaluated regarding to three criteria $\mathbb{V} = \{str, a, b\}$, where str describes the structure of a set of atoms and the two criteria a and b describe the positions of the associated constant symbols. By applying the pseudo measures attached to the three criteria at the previous atoms we obtain the following sets of sequences of labels:*

$$
\begin{array}{l}
f(str)(\{p(f(f(a)), g(a, f(b)))\}) = \{p_1f_1f_1, p_2g_1, p_2g_2f_1\},\\
f(str)(\{p(f(f(b)), g(b, f(a)))\}) = \{p_1f_1f_1, p_2g_1, p_2g_2f_1\},\\
f(str)(\{p(a, g(a, f(b)))\}) = \{p_1, p_2g_1, p_2g_2f_1\},\\
f(a)(\{p(f(f(a)), g(a, f(b)))\}) = \{p_1f_1f_1, p_2g_1\},\\
f(a)(\{p(f(f(b)), g(b, f(a)))\}) = \{p_2g_2f_1\},\\
f(a)(\{p(a, g(a, f(b)))\}) = \{p_1, p_2g_1\},\\
f(b)(\{p(f(f(a)), g(a, f(b)))\}) = \{p_2g_2f_1\},\\
f(b)(\{p(f(f(b)), g(b, f(a)))\}) = \{p_1f_1f_1, p_2g_1\},\\
f(b)(\{p(a, g(a, f(b)))\}) = \{p_2g_2f_1\}.
\end{array}
$$

The distance functions attached to the criteria are constructed using the distance d_l between sequences of labels and its extension at sets of sequences of labels using the minimum distances sum:

$$
\begin{array}{l}
d_{str} : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+\\
d_{str}(A, B) = d_l^{md}(f(str)(A), f(str)(B)),\\
d_a : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+\\
d_a(A, B) = d_l^{md}(f(a)(A), f(a)(B)),\\
d_b : \mathbb{A}_g \times \mathbb{A}_g \to \mathbb{R}_+\\
d_b(A, B) = d_l^{md}(f(b)(A), f(b)(B)).
\end{array}
$$

Restricting at the three atoms that we want to be compared:

$$
\begin{array}{l}
d_{str}(\{p(f(f(a)), g(a, f(b)))\}, \{p(f(f(b)), g(b, f(a)))\})\\
= \frac{1}{6}(0 + 0 + 0 + 0 + 0 + 0) = 0,\\
d_{str}(\{p(f(f(a)), g(a, f(b)))\}, \{p(a, g(a, f(b)))\})\\
= \frac{1}{6}\left(\frac{2}{3} + 0 + 0 + \frac{2}{3} + 0 + 0\right) = \frac{2}{9} = 0.2222,\\
d_a(\{p(f(f(a)), g(a, f(b)))\}, \{p(f(f(b)), g(b, f(a)))\})\\
= \frac{1}{4}\left(1 + \frac{2}{3} + \frac{2}{3}\right) = \frac{7}{12} = 0.5833\\
d_a(\{p(f(f(a)), g(a, f(b)))\}, \{p(a, g(a, f(b)))\})\\
= \frac{1}{4}\left(\frac{2}{3} + 0 + \frac{2}{3} + 0\right) = \frac{1}{3} = 0.3333,\\
d_b(\{p(f(f(a)), g(a, f(b)))\}, \{p(f(f(b)), g(b, f(a)))\})\\
= \frac{1}{4}\left(\frac{2}{3} + 1 + \frac{2}{3}\right) = \frac{7}{12} = 0.5833,\\
d_b(\{p(f(f(a)), g(a, f(b)))\}, \{p(a, g(a, f(b)))\})\\
= \frac{1}{2}(0 + 0) = 0.
\end{array}
$$

Let us suppose that, intuitively, the structural similarity criterion is more important than the constant symbols positions criteria. The relative importance coefficients will be computed using AHP [6].

The matrix

	S.S.	*C.P.S.*
Structural similarity (S.S.)	1/1	5/1
Constants' positions similarity (C.P.S.)	1/5	1/1

presents the relative importance of the criteria structural similarity *and* constants positions similarity*, where the value 5 has the meaning* strongly preferred*. The importance coefficients attached to these two criteria are obtained by computing the eigenvector of the relative importance matrix, which, in our case, is* $[0.8333 \;\; 0.1667]^T$.

Analyzing the criteria related on the constants positions we obtain the relative importance matrix

	"a"	*"b"*
Position "a"	1/1	3/1
Position "b"	1/3	1/1

with the eigenvector $[0.75 \;\; 0.25]^T$*, where the value 3 has the meaning* moderate importance*. Following these computations, the resulting importance coefficients are:*

$$\begin{aligned} \alpha_{str} &= 0.8333, \\ \alpha_a &= 0.1667 \cdot 0.75 = 0.125, \\ \alpha_b &= 0.1667 \cdot 0.25 = 0.0417. \end{aligned}$$

Replacing these coefficients in the formula 11, we obtain the distances:

$$\begin{aligned} &d_a^2(d_{str}, d_a, d_b) \\ &(\{p(f(f(a)), g(a, f(b)))\}, \{p(f(f(b)), g(b, f(a)))\}) = \\ &= \sqrt{0.8333^2 \cdot 0 + 0.125^2 \cdot 0.5833^2 + 0.0417^2 \cdot 0.5833^2} \\ &= 0.0762, \\ &d_a^2(d_{str}, d_a, d_b) \\ &(\{p(f(f(a)), g(a, f(b)))\}, \{p(a, g(a, f(b)))\}) = \\ &= \sqrt{0.8333^2 \cdot 0.2222^2 + 0.125^2 \cdot 0.3333^2 + 0.0417^2 \cdot 0} \\ &= 0.1895. \end{aligned}$$

Due to our preference for the structural similarity criterion, the atom $p(f(f(a)), g(a, f(b)))$ *is considered closer to the atom* $p(f(f(b)), g(b, f(a)))$ *than to the atom* $p(a, g(a, f(b)))$.

Let us reverse now our preferences and consider that the positions of the constants are more important than the structural similarity criterion. The relative importance matrix is

	S.S.	*C.P.S.*
Structural similarity (S.S.)	1/1	1/5
Constants' positions similarity (C.P.S.)	5/1	1/1

with the eigenvector $[0.1667 \;\; 0.8333]^T$*. The relative importance coefficients are now:*

$$\begin{aligned} \alpha_{str} &= 0.1667, \\ \alpha_a &= 0.8333 \cdot 0.75 = 0.6249, \\ \alpha_b &= 0.8333 \cdot 0.25 = 0.2084. \end{aligned}$$

The new values for the distances are:

$$\begin{array}{l} d_a^2(d_{str}, d_a, d_b) \\ (\{p(f(f(a)), g(a, f(b)))\}, \{p(f(f(b)), g(b, f(a)))\}) = \\ = \sqrt{0.1667^2 \cdot 0 + 0.6249^2 \cdot 0.5833^2 + 0.2084^2 \cdot 0.5833^2} \\ = 0.3841, \\ d_a^2(d_{str}, d_a, d_b) \\ (\{p(f(f(a)), g(a, f(b)))\}, \{p(a, g(a, f(b)))\}) = \\ = \sqrt{0.1667^2 \cdot 0.2222^2 + 0.6249^2 \cdot 0.3333^2 + 0.2084^2 \cdot 0} \\ = 0.2112. \end{array}$$

The new distances show that the atom $p(f(f(a)), g(a, f(b)))$ is considered closer to the atom $p(a, g(a, f(b)))$ than to the atom $p(f(f(b)), g(b, f(a)))$.

2.3. Comparative study. In this section we will compare the performances of the proposed similarities measuring method with a metric with same objectives presented in [2] and described in the following definition.

Definition 2.7. *Let us define $d_c : \mathcal{A}_g \times \mathcal{A}_g \to \mathbb{R}$ by:*
(1) *$d_c(a, a) = 0, \forall a \in \mathcal{A}_g$;*
(2) *If $p \neq q$ then $d_c(p(s_1, ..., s_n), q(t_1, ..., t_m)) = 1$;*
(3) *$d_c(p(s_1, ..., s_n), p(t_1, ..., t_n)) = \frac{1}{2n} \sum_{i=1}^{n} d_c(s_i, t_i)$.*
The distance between two subsets of $\mathcal{A}_g$ is defined by the Hausdorff metric induced by d_c.

Example 2.6. *$d_c(g(f(a), g(a, b)), g(f(b), b)) = \frac{3}{8} \approx 0.375$. In the example 2.3, the degree of similarity between the same two atoms was 0.25 and 0.7805, function of the employed similarity criteria. These values show that the sensitivity of the new proposed method can be adjusted by changing the similarity criteria that are used.*

Example 2.7. *Let us consider the vocabulary $\mathcal{P} = \{p/2\}$ şi $\mathcal{F} = \{a/0, b/0, c/0, d/0\}$ and the atoms $e = p(a, b)$, $e' = p(a, c)$ and $e'' = p(c, d)$. We have $d_c(e, e') = \frac{1}{4}$ şi $d_c(e, e'') = \frac{1}{2}$.*

For the our similarities measuring method we use the family of criteria $\mathbb{V} = \{\{a, b, c, d\}\}$. The pseudo measures $f(a)$, $f(b)$, $f(c)$, $f(d)$ will be defined as in the example 2.1:

$$\begin{array}{llllll} f(a)(\{e\}) & = & \{p_1\} & f(b)(\{e\}) & = & \{p_2\} \\ f(a)(\{e'\}) & = & \{p_1\} & f(b)(\{e'\}) & = & \emptyset \\ f(a)(\{e''\}) & = & \emptyset & f(b)(\{e''\}) & = & \emptyset \end{array}$$

$$\begin{array}{llllll} f(c)(\{e\}) & = & \emptyset & f(d)(\{e\}) & = & \emptyset \\ f(c)(\{e'\}) & = & \{p_2\} & f(d)(\{e'\}) & = & \emptyset \\ f(c)(\{e''\}) & = & \{p_1\} & f(d)(\{e''\}) & = & \{p_2\} \end{array}$$

d_a^2 can be defined using minimum distances sum metric or the Hausdorff metric:

$$d_a^2(\{e\}, \{e'\}) = \sqrt{0^2 + 1^2 + 1^2 + 0^2} = \sqrt{2},$$

$$d_a^2(\{e\}, \{e''\}) = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2.$$

The both similarities measuring methods are able to conclude that the atom e is more closer to e' than to e''.

Example 2.8. *Let us consider the vocabulary from the previous example and the atoms* $e = p(a, a)$, $e' = p(b, b)$ *and* $e'' = p(c, d)$. *The metric* d_c *considers the pairs* (e, e') *and* (e, e'') *as having the same degree of similarity. The distances are* $d_c(e, e') = \frac{1}{4}(1 + 1) = \frac{1}{2} = d_c(e, e'')$.

The family of criteria $\mathbb{V} = \{\{a, b, c, d\}, \{n\}\}$ *is used, where* $\mathcal{X}_n = (2^{\mathbb{N}}, \emptyset, \cup, \subseteq)$ *and* $f(n) : \mathbb{A}_g \to 2^{\mathbb{N}}$, $f(n)(A)$ *is the greatest set with the property that* $\forall m \in f(n)(A)$ *exists a graph attached to an atom from* A *that contains exactly* m *nodes.*

$$f(n)(\{e\}) = f(n)(\{e'\}) = \{2\} \qquad f(n)(\{e''\}) = \{3\}.$$

The pseudo measures $f(a)$, $f(b)$, $f(c)$, $f(d)$ *are defined using the model from the example 2.1.*

$$\begin{array}{lcllcl} f(a)(\{e\}) & = & \{p_1, p_2\} & f(b)(\{e\}) & = & \emptyset \\ f(a)(\{e'\}) & = & \emptyset & f(b)(\{e'\}) & = & \{p_1, p_2\} \\ f(a)(\{e''\}) & = & \emptyset & f(b)(\{e''\}) & = & \emptyset \end{array}$$

$$\begin{array}{lcllcl} f(c)(\{e\}) & = & \emptyset & f(d)(\{e\}) & = & \emptyset \\ f(c)(\{e'\}) & = & \emptyset & f(d)(\{e'\}) & = & \emptyset \\ f(c)(\{e''\}) & = & \{p_1\} & f(d)(\{e''\}) & = & \{p_2\} \end{array}$$

The distances $d(a)$, $d(b)$, $d(c)$, $d(d)$, $d(n)$ *can be defined using the Hausdorff metric or the minimum distances sum metric.*

$$d_a^2(\{e\}, \{e'\}) = (\sqrt{1^2 + 1^2 + 0^2 + 0^2}, \sqrt{0}) = (\sqrt{2}, 0),$$
$$d_a^2(\{e\}, \{e''\}) = (\sqrt{1^2 + 0^2 + 1^2 + 1^2}, \sqrt{1}) = (\sqrt{3}, 1).$$

Contrastingly with d_c, *the similarity measure* d_a^2 *considers that the atom* e *is closer to* e' *than to* e''. *Intuitively, this corresponds with the human reasoning model.*

Example 2.9. *Let us consider the vocabulary* $\mathcal{P} = \{p/2, q/2\}$, $\mathcal{F} = \{a/0, b/0, c/0, d/0\}$ *and the atoms* $e = p(a, b)$, $e' = q(a, c)$ *and* $e'' = q(c, d)$. $d_c(e, e') = d_c(e, e'') = 1$ *and the atoms* e' *and* e'' *are at the same distance from* e *regarding the metric* d_c.

$\mathbb{V} = \{\{a, b, c, d\}\}$ *is a set of criteria and the pseudo measures* $f(a)$, $f(b)$, $f(c)$, $f(d)$ *are defined by the formula 1 and the distances* $d(a)$, $d(b)$, $d(c)$, $d(d)$ *are defined by 7.*

$$\begin{array}{lcllcl} f(a)(\{e\}) & = & 1 & f(b)(\{e\}) & = & 1 \\ f(a)(\{e'\}) & = & 1 & f(b)(\{e'\}) & = & 0 \\ f(a)(\{e''\}) & = & 0 & f(b)(\{e''\}) & = & 0 \end{array}$$

$$\begin{array}{lcllcl} f(c)(\{e\}) & = & 0 & f(d)(\{e\}) & = & 0 \\ f(c)(\{e'\}) & = & 1 & f(d)(\{e'\}) & = & 0 \\ f(c)(\{e''\}) & = & 1 & f(d)(\{e''\}) & = & 1 \end{array}$$

$$d_a^2(\{e\}, \{e'\}) = \sqrt{0^2 + 1^2 + 1^2 + 0^2} = \sqrt{2} \text{ and}$$
$$d_a^2(\{e\}, \{e''\}) = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2.$$

The new metric considers that e *is closer to* e' *than to* e''. *This conclusion is natural,* e *and* e' *include same constant symbol* a *and do not use the symbol* d.

3. Conclusion

The information systems that use entities described by logic programs need to know when two entities have numerous characteristics in common in order to perform the knowledge transfer from one entity to a second one similar to the first. A framework for constructing distances between logic programs was defined, the distances being

defined regarding several user specified criteria. Each criterion was described by a pseudo measure and the pseudo measures were combined to obtain a distance. Several examples show that the framework can provide intuitive results.

References

[1] T. Eiter and H. Mannila, *Distance measures for point sets and their computation*, Acta Informatica, 34(2), pages 109-133, 1997

[2] S-H. Nienhuys-Cheng, *Distance Between Herbrand Interpretations: A Measure for Approximations to a Target Concept*, ILP97, Lecture Notes in Artificial Intelligence, Springer Verlag, pages 213-226, 1997

[3] M. Preda and D. Popescu, *Personalized Web Recommendations: Supporting Epistemic Information about End-Users*, Proc. of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Compiegne, France, IEEE Computer Press, ISBN: 0-7695-2415-X, pages 692-695, 2005

[4] J. Ramon and M. Bruynooghe, *A framework for defining distances between first logic objects*, Technical Report, Katholieke Universiteit Leuven, Belgium, 1998

[5] J. Ramon, W. Van Laer and M. Bruynooghe, *Distance measures between atoms*, CompulogNet Area Meeting on Computational Logic and Machine Learing, University of Manchester, UK, pages 35–41, 1998

[6] T.L. Saaty and L.G. Vargas,*Models, Methods, Concepts and Applications of the Analytic Hierarchy Process*, Kluwer Academic, 2000

(Mircea Preda) Department of Computer Science, University of Craiova,
Al.I. Cuza Street, No. 13, Craiova RO-200585, Romania
E-mail address: mircea_cezar_preda@yahoo.com