# Approximate Measures of Semantic Dissimilarity under Uncertainty

Nicola Fanizzi, Claudia d'Amato and Floriana Esposito

*Lacam Lab – Department of Computer Science*
*University of Bari – Italy*

**Uncertainty Reasoning for the Semantic Web**
Workshop ◇ ISWC, 12 November 2007

# Contents

# Introduction & Motivations

- In the context of Reasoning in the SW, it is growing the interest in alternative *inductive procedures* (i.e. case-based reasoning, retrieval, conceptual clustering, ontology matching...)
  - Many of them are based on the notion of *similarity*
- Most of the measures able to assess similarity in DL representation focus on *similarity between atomic concepts*
  - Inductive learning methods often need for a notion of **similarity among individuals**
- **A new family of dissimilarity measures for semantically annotated resources has been devised**

# Knowledge Base Representation

**Assumption:** resources, concepts and relationships are defined in terms of a representation that can be mapped to some DL language (with the standard model-theoretic semantics)

$\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$

- *T-box* $\mathcal{T}$ is a set of definitions $C \equiv D$
- *A-box* $\mathcal{A}$ contains extensional assertions on concepts and roles e.g. $C(a)$ and $R(a, b)$
- The set of the individuals (resources) occurring in $\mathcal{A}$ will be denoted $\mathrm{Ind}(\mathcal{A})$

*Instance checking* and *retrieval* inference services will be used

# Semantic Distance Measure: Main Idea

- **IDEA**: *on a semantic level, similar individuals should behave similarly w.r.t. the same concepts*
- Following HDD **[Sebag 1997]**: individuals can be compared on the grounds of their behavior w.r.t. a given set of hypotheses $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, that is a collection of (primitive or defined) concept descriptions
  - $F$ stands as a group of *discriminating features* expressed in the considered language
- As such, the new measure *totally depends on semantic* aspects of the individuals in the KB

# The Projection Function

## Projection Function

Given a concept $F_i \in \mathsf{F}$, the related *projection function* $\pi_i : \mathsf{Ind}(\mathcal{A}) \mapsto \{0, 1/2, 1\}$ is defined, $\forall a \in \mathsf{Ind}(\mathcal{A})$

$$\pi_i(a) := \begin{cases} 1 & \mathcal{K} \models F_i(a) \\ 0 & \mathcal{K} \models \neg F_i(a) \\ 1/2 & otherwise \end{cases}$$

- Case: $\pi_i(a) = 1/2 \Rightarrow$ the reasoner cannot give the truth value for a certain membership query
  - This is due to the *OWA* normally made in this context
- Hence, as in the classic probabilistic models, *uncertainty* is coped with by considering a *uniform distribution* over the possible cases.

# Semantic Dinstance Measure: Definition

**[Fanizzi et al. @ DL 2007]** Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a KB. Given sets of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_k\}$, a *family of semi-distance functions* $d_p^{\mathsf{F}} : \mathrm{Ind}(\mathcal{A}) \times \mathrm{Ind}(\mathcal{A}) \mapsto \mathbb{R}$, inspired to Minkoswi's distance, is defined as follows:

$$\forall a, b \in \mathrm{Ind}(\mathcal{A}) \quad d_p^{\mathsf{F}}(a, b) := \frac{1}{k} \sqrt[p]{\sum_{i=1}^{k} \delta_i(a, b)^p}$$

where $p > 0$ ¡and $\delta_i : \mathrm{Ind}(\mathcal{A}) \times \mathrm{Ind}(\mathcal{A}) \mapsto [0, 1]$ is the **discernibility function**: $\forall (a, b) \in \mathrm{Ind}(\mathcal{A}) \times \mathrm{Ind}(\mathcal{A})$

$$\delta_i(a, b) = |\pi_i(a) - \pi_i(b)|$$

that compare two individuals $(a, b)$ w.r.t. a feature concept $F_i \in \mathsf{F}$

# Distance Measure: Example

$\mathcal{T} = \{$ Female $\equiv \neg$Male, Parent $\equiv \forall$child.Being $\sqcap \exists$child.Being,
Father $\equiv$ Male $\sqcap$ Parent,
FatherWithoutSons $\equiv$ Father $\sqcap \forall$child.Female$\}$

$\mathcal{A} = \{$ Being(ZEUS), Being(APOLLO), Being(HERCULES), Being(HERA),
Male(ZEUS), Male(APOLLO), Male(HERCULES),
Parent(ZEUS), Parent(APOLLO), $\neg$Father(HERA),
God(ZEUS), God(APOLLO), God(HERA), $\neg$God(HERCULES),
hasChild(ZEUS, APOLLO), hasChild(HERA, APOLLO),
hasChild(ZEUS, HERCULES), $\}$

Suppose $\mathsf{F} = \{F_1, F_2, F_3, F_4\} = \{$Male, God, Parent, FatherWithoutSons$\}$.

Let us compute the distances (with $p = 1$):

$d_1^{\mathsf{F}}(\text{HERCULES}, \text{ZEUS}) =$
$(|1-1| + |0-1| + |1/2-1| + |1/2-0|)\,/4 = 1/2$

$d_1^{\mathsf{F}}(\text{HERA}, \text{HERCULES}) =$
$(|0-1| + |1-0| + |1-1/2| + |0-1/2|)\,/4 = 3/4$

# Distance Measure: Discussion

- The measure is a semi-distance (i.e. it does not guaranties that if $d_p^{\mathsf{F}}(a, b) = 0 \Rightarrow a = b$)
- *More similar* the considered *individuals are*, more similar the project function values are $\Rightarrow d_p^{\mathsf{F}} \simeq 0$
- *More different* the considered *individuals are*, more different the projection values are $\Rightarrow$ the value of $d_p^{\mathsf{F}}$ will increase
- The measure complexity mainly depends from the complexity of the *Instance Checking* operator for the chosen DL
  - $\mathit{Compl}(d_p^{\mathsf{F}}) = |\mathsf{F}| \cdot 2 \cdot \mathit{Compl}(\mathsf{IChk})$
- **Optimal discriminating feature set could be learned**

# Measure Optimization: Feature Selection

- **Assumption:** F represents a sufficient number of (possibly redundant) features able to really discriminate individuals.
  - The choice of the features – *feature selection* – may be crucial
- **Proposal of optimization algorithms**that are able to find/build optimal discriminating concept committees **[Fanizzi et al. @ DL 2007 and @ ICSC 2007]**
  - **Idea:** Optimization of a *fitness function* that is based on the *discernibility factor of the committee*, namely
  - Given $\mathrm{Ind}(\mathcal{A})$ (or just a hold-out sample) $HS \subseteq \mathrm{Ind}(\mathcal{A})$ *find* the subset F that maximize the following function:

$$\textsc{discernibility}(\mathsf{F}, HS) := \sum_{(a,b)\in HS^2} \sum_{i=1}^{k} \delta_i(a, b)$$

  - The results obtained with KSs drawn from ontology libraries show that (a selection) of the (primitive and defined) concepts is often sufficient to induce satisfactory dissimilarity measures

# Dissimilarity under Uncertainty: Motivation

- The defined measure deals with uncertainty in a uniform way
  - the degree of discernibility of two individuals is null when they have the same behavior w.r.t. the same feature, even in the presence of total uncertainty of class-membership for both
  - When uncertainty regards only one projection, then they are considered partially (possibly) similar
  - **GOAL:** makes this uncertainty more explicit
- **New Proposal:** The dissimilarity between two individuals is assessed as a *combination* of **degree of evidence** that they differ w.r.t. a feature set
  - The measure is again based on the *degree of belief of discernibility* of individuals w.r.t. the features
    - the notion of *probability masses* of the basic events (class-membership) is exploited

# Computing the Probability Masses

Given the feature set $\mathsf{F} = \{F_1, F_2, \ldots, F_k\}$, the *probability mass* of the basic events "class-membership", $\forall a \in Ind(\mathcal{A})$ and $i \in \{1, 2, \ldots, k\}$ *in case of uncertainty* is given by:

$$m_i(\mathcal{K} \models F_i(a)) \approx |\text{retrieval}(F_i, \mathcal{K})|/|\text{Ind}(\mathcal{A})|$$
$$m_i(\mathcal{K} \models \neg F_i(a)) \approx |\text{retrieval}(\neg F_i, \mathcal{K})|/|\text{Ind}(\mathcal{A})|$$
$$m_i(\mathcal{K} \models F_i(a) \vee \mathcal{K} \models \neg F_i(a)) \approx 1 - m_i(\mathcal{K} \models F_i(a)) - m_i(\mathcal{K} \models \neg F_i(a))$$

**Rationale:** the larger the (estimated) extension the more likely is for individuals to belong to the concept.

**In case of a certain answer received from the reasoner, the probability mass amounts to 0 or 1.**

# The Discernibility Function

The discernibility function (w.r.t. a concept) measures the amount of evidence that two input individuals are separated by that concept

## Discernibility Function

Given $F_i \in \mathsf{F}$, the *discernibility function* $\delta_i : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0,1]$ is defined, $\forall (a,b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A})$, as follows:

$$\delta_i(a,b) := \begin{cases} 0 & \text{if } \mathcal{K} \models F_i(a) \wedge \mathcal{K} \models F_i(b) \\ 1 & \text{if } \mathcal{K} \models F_i(a) \wedge \mathcal{K} \models \neg F_i(b) \text{ or viceversa} \\ m_i(\mathcal{K} \models \neg F_i(b)) & \text{else if } \mathcal{K} \models F_i(a) \\ m_i(\mathcal{K} \models F_i(b)) & \text{else if } \mathcal{K} \models \neg F_i(a) \\ \delta_i(b,a) & \text{else if } \mathcal{K} \models F_i(b) \vee \mathcal{K} \models \neg F_i(b) \\ 2 \cdot m_i(\mathcal{K} \models F_i(a)) \cdot m_i(\mathcal{K} \models \neg F_i(b)) & \text{otherwise} \end{cases}$$

# Discernibility Function: Interpretation

- The extreme values $\{0, 1\}$ are returned when the answers from the instance-checking service are certain for both individuals.
- If $a$ is an instance of $F_i$ (resp., its complement) $\Rightarrow$ the discernibility depends on the belief of class-membership to the complement concept of $b$.
- If there is uncertainty for $a$ but not for $b$, the function is computed swapping the roles of the two individuals.
- In case of uncertainty for both individuals, the discernibility is computed as the chance that they may belong one to $F_i$ and one to its complement

# Dissimilarity Measure under Uncertainty: Definition

Following the *mixing combination rule*, the degree of belief can be combined for assessing a dissimilarity measure between individuals:

**Dissimilarity Measure under Uncertainty**

Given an ABox $\mathcal{A}$, a dissimilarity measure $d^{\mathsf{F}}_{avg} : \mathrm{Ind}(\mathcal{A}) \times \mathrm{Ind}(\mathcal{A}) \mapsto [0, 1]$, $\forall (a, b) \in \mathrm{Ind}(\mathcal{A}) \times \mathrm{Ind}(\mathcal{A})$, is defined as follows:

$$d^{\mathsf{F}}_{avg}(a, b) := \sum_{i=1}^{k} w_i \delta_i(a, b)$$

Where the weights can be defined uniform as:

- $w_i = 1/k$ or
- $w_i = \frac{u_i}{u}$ where $u_i = \frac{1}{|\mathrm{Ind}(A) \setminus \mathrm{retrieval}(F_k, \mathcal{K})|}$ and $u = \sum_{i=1}^{k} u_i$

# Concept Dissimilarity under Uncertainty

The measures can be extended to the case of concepts, by recurring to the notion of *medoids*.

- The *medoid* of a group of individuals $G = \{a_1, a_2, \ldots, a_n\}$ is the individual that has the highest similarity w.r.t. the others i.e. $\mathrm{medoid}(G) = \mathrm{argmin}_{a \in G} \sum_{j=1}^{n} d(a, a_j)$

## Concept Dissimilarity

Given $C_1$ and $C_2$ concepts, let $R_i = \{a \in \mathsf{Ind}(\mathcal{A}) \mid \mathcal{K} \models C_i(a)\}$ be groups of individuals for $i = 1, 2$ and $m_i = \mathrm{medoid}(R_i)$ their resp. medoids w.r.t. a given measure $d_p^{\mathsf{F}}$. Then the function for concepts can be defined as:

$$d_p^{\mathsf{F}}(C_1, C_2) := d_p^{\mathsf{F}}(m_1, m_2)$$

Similarly, the distance of an individual $a$ to a concept $C$ can be defined:

$$d_p^{\mathsf{F}}(a, C) := d_p^{\mathsf{F}}(a, m)$$

# Conclusions

- The definition of dissimilarity measures over the spaces of individuals in a KB have been proposed
  - The measures are totally semantic (not language-dependent)
  - The measures are parameterized on a committes of concepts
- Optimal committees can be found maximizing a discernibility function, by the use of randomized search methods
- Dissimilarity measures able to cope with cases of uncertainty have been defined
  - based on a simple evidence combination method

# Future Works

Embedding the presented measures in distance-based methods to apply to KBs for:

- setting up logic approaches to ontology matching
- supporting a process of *(semi-)automatic classification* of new data (also as a first step towards ontology evolution)
- ranking the answers provided by a matchmaking algorithm on the ground of the similarity between the query concept and the retrieved individuals

# The End

That's all!