Similarity and Dissimilarity Measures for Concept Descriptions in Ontological Knowledge

Claudia d'Amato

Tutor: Floriana Esposito
Co-Tutor: Nicola Fanizzi

Dipartimento di Informatica • *Università degli Studi di Bari*
Campus Universitario, Via Orabona 4, 70125 Bari, Italy

Università degli Studi di Milano

Contents

Why the attention to Similarity Measures...

- **Information Retrieval**
- **Information Integration**
  - often relied on ontologies (described by means of DL)
- **Clustering** by means of *partitional* or *agglomerative* algorithms based on a distance

⋮

- **Semantic Web** *Service discovery* (in OWL-S Profile Registry)

...Why the attention to Similarity Measures

- Past works have concentrated on defining similarity of "atomic concepts" (words sense)
- New similarity/dissimilarity measures applicable to composite, defined concepts are necessary
  - defined concepts are the stock-in-trade of DL and hence of ontologies

...Related Work...

- *Path distance measures* [Bright,94]: applied to terms represented in a built hierarchical structure underlying the KB
- *Feature matching measures* [Tversky,77]: consider both common and discriminant features to compute similarity
- *Information Content measures* [Resnik,99]: compute similarity for concepts within a hierarchy, in terms of the amount of information conveyed by their immediate super-concept

...Related Work...

Path distance measures [Bright,94]: MAIN IDEA

- measure the similarity value between single words (and not complex concept definitions)
- concepts (words) are organized in a taxonomy using hypernym/hyponym and synoym links.
- the measure is a weighted count of the links in the path between two terms
  - terms with only a few links separating them are semantically similar
  - terms with many links between them have less similar meanings
  - link counts are weighted because different relationships have different implications for semantic similarity.

...Related Work...

Path distance measures [Bright,94]: WEAKNESS

- the similarity value is subjective due to the taxonomic ad-hoc representation
- the introduction of news term can change similarity values
- the similarity measures cannot be applied directly to the knowledge representation
  - it needs of an intermediate step which is building the term taxonomy structure
- only "linguistic" relations among terms are considered; there are not relations whose semantics models domain

...Related Work...

Feature Matching measures [Tversky,77]

- a feature-based *contrast model* of similarity is proposed
  - common features tend to increase the perceived similarity of two concepts
  - feature differences tend to diminish perceived similarity
  - feature commonalities increase perceived similarity more than feature differences can diminish it
- feature vector is the used representation (not expressive enough)

...Related Work...

Information Content measures [Resnik,99]...

- measure semantic similarity of concepts in an *is-a* taxonomy by the use of notion of *Information Content (IC)*
- similarity of two concepts is given by the information that they share
  - the shared information is represented by a highly specific super-concept that subsumes both concepts
- *similarity value* is given by the *IC of the least common super-concept*
  - *IC for a concept is determined* considering the probability that an instance belongs to the concept

...Related Work...

...Information Content measures [Resnik,99]

- use a criterion similar to those used in *path distance measures*,
- differently from *path distance measures*, the use of probabilities avoids the unreliability of counting edge when changing in the hierarchy occur
- the considered relation among concepts is only *is-a* relation
  - more semantically expressive relations cannot be considered

Motivations

- Ontological knowledge
  - Result of a complex process of knowledge acquisition
  - Plays a key role for interoperability in the Semantic Web perspective
  - Is expressed by standard ontology mark-up languages which are supported by well-founded semantics of Description Logics (DLs)
- Need of services able to build knowledge bases automatically or semi-automatically
  - This can be done by the use of inductive inference services

Objectives...

- Induction of structural knowledge is known is ML (concept formation).
  - This is generally applied on zero-order representations.
- *our Goal* $\longrightarrow$ to make clusters of concepts or individuals asserted in ontological knowledge
- *Problem* $\longrightarrow$ to define a similarity/dissimilarity measure applicable to ontology languages

...Objectives

- Already defined similalrity/dissimilalrity measures cannot be directly applied to ontological knowledge
  - They define similarity value between atomic concepts
  - They are defined for representation less expressive than ontology representation
  - They cannot exploit all the expressiveness of the ontological representation
- Defining new measures that are really semantic is necessary

Why *ALC* Logic

- Knowledge representation by means of Description Logic (ALC)
- Description Logic is the theoretical foundation of OWL language
  - standard de facto for the knowledge representation in the Semantic Web

The Representation Language

- Primitive *concepts* $N_C = \{C, D, \ldots\}$: subsets of a domain
- Primitive *roles* $N_R = \{R, S, \ldots\}$: binary relations on the domain
- *Interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ where
  $\Delta^{\mathcal{I}}$: *domain* of the interpretation and $\cdot^{\mathcal{I}}$: *interpretation function*:

Name	Syntax	Semantics
top concept	$\top$	$\Delta^{\mathcal{I}}$
bottom concept	$\bot$	$\emptyset$
concept	C	$C^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
concept negation	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
concept conjunction	$C_1 \sqcap C_2$	$C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$
concept disjunction	$C_1 \sqcup C_2$	$C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$
existential restriction	$\exists R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}}((x,y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}})\}$
universal restriction	$\forall R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y \in \Delta^{\mathcal{I}}((x,y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}})\}$

Knowledge Base & Subsumption

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

- *T-box* $\mathcal{T}$ is a set of definitions $C \equiv D$, meaning $C^{\mathcal{I}} = D^{\mathcal{I}}$, where C is the concept name and D is a description
- *A-box* $\mathcal{A}$ contains extensional assertions on concepts and roles e.g. $C(a)$ and $R(a, b)$, meaning, resp., that $a^{\mathcal{I}} \in C^{\mathcal{I}}$ and $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$.

Subsumption

Given two concept descriptions C and D, C *subsumes* D, denoted by $C \sqsupseteq D$, iff for every interpretation $\mathcal{I}$, it holds that $C^{\mathcal{I}} \supseteq D^{\mathcal{I}}$

Examples

An instance of concept definition:
$\text{Father} \equiv \text{Male} \sqcap \exists \text{hasChild}.\text{Person}$
"a father is a male (person) that has some persons as his children"

The following are instances of simple assertions:
$\text{Male}(\text{Leonardo})$, $\text{Male}(\text{Vito})$, $\text{hasChild}(\text{Leonardo}, \text{Vito})$

Supposing $\text{Male} \sqsubseteq \text{Person}$:
$\text{Person}(\text{Leonardo})$, $\text{Person}(\text{Vito})$ and then $\text{Father}(\text{Leonardo})$

Other related concepts: $\text{Parent} \equiv \text{Person} \sqcap \exists \text{hasChild}.\text{Person}$ and
$\text{FatherWithoutSons} \equiv \text{Male} \sqcap \exists \text{hasChild}.\text{Person} \sqcap \forall \text{hasChild}.(\neg \text{Male})$

It is easy to see that the following relationships hold:
$\text{Parent} \sqsupseteq \text{Father}$ and $\text{Father} \sqsupseteq \text{FatherWithoutSons}$.

Other Inference Services

instance checking decide whether an individual is an instance of a concept

retrieval find all invididuals instance of a concept

realization problem finding the concepts which an individual belongs to, especially the most specific one, if any:

most specific concept

Given an A-Box $\mathcal{A}$ and an individual a, the *most specific concept* of a w.r.t. $\mathcal{A}$ is the concept C, denoted $\mathrm{MSC}_{\mathcal{A}}(a)$, such that $\mathcal{A} \models C(a)$ and $C \sqsubseteq D$, $\forall D$ such that $\mathcal{A} \models D(a)$.

Similarity Measure between Concepts: Needs

- Necessity to have a measure really based on Semantics
- Considering [Tversky'77]:
  - common features tend to increase the perceived similarity of two concepts
  - feature differences tend to diminish perceived similarity
  - feature commonalities increase perceived similarity more than feature differences can diminish it
- The proposed similarity measure is:

Similarity Measure between Concepts

Definition [d'Amato'05 @ CILC 2005]: Let $\mathcal{L}$ be the set of all concepts in $\mathcal{ALC}$ and let $\mathcal{A}$ be an A-Box with canonical interpretation $\mathcal{I}$. The *Semantic Similarity Measure* s is a function

$$s : \mathcal{L} \times \mathcal{L} \mapsto [0, 1]$$

defined as follows:

$$s(C, D) = \frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}| + |D^{\mathcal{I}}| - |I^{\mathcal{I}}|} \cdot \max\left(\frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}|}, \frac{|I^{\mathcal{I}}|}{|D^{\mathcal{I}}|}\right)$$

where $I = C \sqcap D$ and $(\cdot)^{\mathcal{I}}$ computes the concept extension wrt the interpretation $\mathcal{I}$.

Similarity Measure: Meaning

- If $C \equiv D$ ($C \sqsubseteq D$ and $D \sqsubseteq C$)then $s(C, D) = 1$, i.e. the maximum value of the similarity is assigned.
- If $C \sqcap D = \bot$ then $s(C, D) = 0$, i.e. the minimum similarity value is assigned because concepts are totally different.
- Otherwise $s(C, D) \in]0, 1[$. The similarity value is proportional to the overlapping amount of the concept extetions reduced by a quantity representing how the two concepts are near to the overlap. This means considering similarity not as an absolute value but as weighted w.r.t. a degree of non-similarity.

Similarity Measure: Example...

Primitive Concepts: $N_C = \{\text{Female}, \text{Male}, \text{Human}\}$.
Primitive Roles:
$N_R = \{\text{HasChild}, \text{HasParent}, \text{HasGrandParent}, \text{HasUncle}\}$.
$\mathcal{T} = \{$ Woman $\equiv$ Human $\sqcap$ Female; Man $\equiv$ Human $\sqcap$ Male
Parent $\equiv$ Human $\sqcap \exists$HasChild.Human
Mother $\equiv$ Woman $\sqcap$ Parent $\exists$HasChild.Human
Father $\equiv$ Man $\sqcap$ Parent
Child $\equiv$ Human $\sqcap \exists$HasParent.Parent
Grandparent $\equiv$ Parent $\sqcap \exists$HasChild.($\exists$ HasChild.Human)
Sibling $\equiv$ Child $\sqcap \exists$HasParent.($\exists$ HasChild ≥ 2)
Niece $\equiv$ Human $\sqcap \exists$HasGrandParent.Parent $\sqcup \exists$HasUncle.Uncle
Cousin $\equiv$ Niece $\sqcap \exists$HasUncle.($\exists$ HasChild.Human)$\}$.

...Similarity Measure: Example...

$\mathcal{A} = \{$Woman(Claudia), Woman(Tiziana), Father(Leonardo), Father(Antonio), Father(AntonioB), Mother(Maria), Mother(Giovanna), Child(Valentina), Sibling(Martina), Sibling(Vito), HasParent(Claudia,Giovanna), HasParent(Leonardo,AntonioB), HasParent(Martina,Maria), HasParent(Giovanna,Antonio), HasParent(Vito,AntonioB), HasParent(Tiziana,Giovanna), HasParent(Tiziana,Leonardo), HasParent(Valentina,Maria), HasParent(Maria,Antonio), HasSibling(Leonardo,Vito), HasSibling(Martina,Valentina), HasSibling(Giovanna,Maria), HasSibling(Vito,Leonardo), HasSibling(Tiziana,Claudia), HasSibling(Valentina,Martina), HasChild(Leonardo,Tiziana), HasChild(Antonio,Giovanna), HasChild(Antonio,Maria), HasChild(Giovanna,Tiziana), HasChild(Giovanna,Claudia), HasChild(AntonioB,Vito), HasChild(AntonioB,Leonardo), HasChild(Maria,Valentina), HasUncle(Martina,Giovanna), HasUncle(Valentina,Giovanna)$\}$

...Similarity Measure: Example

$$s(\text{Grandparent}, \text{Father}) = \frac{|(\text{Grandparent} \sqcap \text{Father})^{\mathcal{I}}|}{|\text{Granparent}^{\mathcal{I}}| + |\text{Father}^{\mathcal{I}}| - |(\text{Grandarent} \sqcap \text{Father})^{\mathcal{I}}|} \cdot$$

$$\cdot\ max\left(\frac{|(\text{Grandparent} \sqcap \text{Father})^{\mathcal{I}}|}{|\text{Grandparent}^{\mathcal{I}}|}, \frac{|(\text{Grandparent} \sqcap \text{Father})^{\mathcal{I}}|}{|\text{Father}^{\mathcal{I}}|}\right) =$$

$$= \frac{2}{2+3-2} \cdot max\left(\frac{2}{2}, \frac{2}{3}\right) = 0.67$$

Similarity Measure between Individuals

Let c and d two individuals in a given A-Box.
We can consider $C^* = \text{MSC}^*(c)$ and $D^* = \text{MSC}^*(d)$:

$$s(c, d) := s(C^*, D^*) = s(\text{MSC}^*(c), \text{MSC}^*(d))$$

Analogously:

$$\forall a: \; s(c, D) := s(\text{MSC}^*(c), D)$$

Discussion...

- **The presented function is a similarity measure**
  1. $f(a,b) \geq 0 \quad \forall a,b \in E$ *(positive definiteness)*
  2. $f(a,b) = f(b,a) \quad \forall a,b \in E$ *(symmetry)*
  3. $\forall a,b \in E : f(a,b) \leq f(a,a)$

1. It is satisfied by the definition of s
2. $$s(C,D) = \frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}|+|D^{\mathcal{I}}|-|I^{\mathcal{I}}|} \cdot \max\left(\frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}|}, \frac{|I^{\mathcal{I}}|}{|D^{\mathcal{I}}|}\right) = \frac{|I^{\mathcal{I}}|}{|D^{\mathcal{I}}|+|C^{\mathcal{I}}|-|I^{\mathcal{I}}|} \cdot \max\left(\frac{|I^{\mathcal{I}}|}{|D^{\mathcal{I}}|}, \frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}|}\right) = s(D,C)$$
   where I remains the same because of the commutativity of intersection.
3. It is satisfied because s assigns the maximum value when the concepts are equivalent

...Discussion

- **Computational Complexity**
  - Similarity between concepts: $Compl(s) = 3 \cdot Compl(IC)$
  - Similarity individual-concept: $Compl(s) = Compl(MSC^*) + 3 \cdot Compl(IC)$
  - Similarity between individuals: $Compl(s) = 2 \cdot Compl(MSC^*) + 3 \cdot Compl(IC)$

Similarity Measure: Conclusions...

- s is a *Semantic* Similarity measure
  - It uses only *semantic inference* (Instance Checking) for determining similarity values
  - It does not make use of the syntactic structure of the concept descriptions
  - It does not add complexity besides of the complexity of used inference operator

...Similarity Measure: Conclusions

- Experimental evaluations demonstrate that s works satisfying when it is applied between concepts
- s applied to individuals is often zero even in case of similar individuals
  - The MSC^* is so specific that often covers only the considered individual and not similar individuals
- The *new idea* is to measure the similarity (dissimilarity) of the subconcepts that build the MSC^* concepts in order to find their similarity (dissimilarity)

MSC^* : *AnExample*

$$\text{MSC*(Claudia)} = \text{Woman} \sqcap \text{Sibling} \sqcap \exists \text{HasParent}(\text{Mother} \sqcap \text{Sibling} \sqcap \exists \text{HasSibling}(\text{C1}) \sqcap \exists \text{HasParent}(\text{C2}) \sqcap \exists \text{HasChild}(\text{C3}))$$

$$\text{C1} \equiv \text{Mother} \sqcap \text{Sibling} \sqcap \exists \text{HasParent}(\text{Father} \sqcap \text{Parent}) \sqcap \exists \text{HasChild}(\text{Cousin} \sqcap \exists \text{HasSibling}(\text{Cousin} \sqcap \text{Sibling} \sqcap \exists \text{HasSibling}.\top))$$

$$\text{C2} \equiv \text{Father} \sqcap \exists \text{HasChild}(\text{Mother} \sqcap \text{Sibling})$$

$$\text{C3} \equiv \text{Woman} \sqcap \text{Sibling} \sqcap \exists \text{HasSibling}.\top \sqcap \exists \text{HasParent}(\text{C4})$$

$$\text{C4} \equiv \text{Father} \sqcap \text{Sibling} \sqcap \exists \text{HasSibling}(\text{Uncle} \sqcap \text{Sibling} \sqcap \exists \text{HasParent}(\text{Father} \sqcap \text{Grandparent})) \sqcap \exists \text{HasParent}(\text{Father} \sqcap \text{Grandparent} \sqcap \exists \text{HasChild}(\text{Uncle} \sqcap \text{Sibling}))$$

Normal Form

D is in $\mathcal{ALC}$ *normal form* iff $D \equiv \bot$ or $D \equiv \top$ or if $D = D_1 \sqcup \cdots \sqcup D_n$ $(\forall i = 1, \ldots, n,\ D_i \not\equiv \bot)$ with

$$D_i = \prod_{A \in \text{prim}(D_i)} A \sqcap \prod_{R \in N_R} \left[\forall R.\text{val}_R(D_i) \sqcap \prod_{E \in \text{ex}_R(D_i)} \exists R.E \right]$$

where:

- $\text{prim}(C)$ set of all (negated) atoms occurring at C's top-level
- $\text{val}_R(C)$ conjunction $C_1 \sqcap \cdots \sqcap C_n$ in the value restriction on R, if any (o.w. $\text{val}_R(C) = \top$);
- $\text{ex}_R(C)$ set of concepts in the value restriction of the role R

For any R, every sub-description in $\text{ex}_R(D_i)$ and $\text{val}_R(D_i)$ is in normal form.

Overlap Function

Definition [d'Amato'05 @ KCAP 2005 Workshop]:
$\mathcal{L} = \mathcal{ALC}/_{\equiv}$ the set of all concepts in $\mathcal{ALC}$ normal form
$\mathcal{I}$ canonical interpretation of A-Box $\mathcal{A}$

$f : \mathcal{L} \times \mathcal{L} \mapsto R^+$ defined $\forall C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$ in $\mathcal{L}_{\equiv}$

$$f(C,D) := f_{\sqcup}(C,D) = \begin{cases} \infty & C \equiv D \\ 0 & C \sqcap D \equiv \bot \\ \max\limits_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}} f_{\sqcap}(C_i, D_j) & \text{o.w.} \end{cases}$$

$$f_{\sqcap}(C_i, D_j) \quad := \quad f_P(\text{prim}(C_i), \text{prim}(D_j)) + f_{\forall}(C_i, D_j) + f_{\exists}(C_i, D_j)$$

Overlap Function / II

$$f_P(\text{prim}(C_i), \text{prim}(D_j)) := \frac{|(\text{prim}(C_i))^{\mathcal{I}} \cup (\text{prim}(D_j))^{\mathcal{I}}|}{|((\text{prim}(C_i))^{\mathcal{I}} \cup (\text{prim}(D_j))^{\mathcal{I}}) \setminus ((\text{prim}(C_i))^{\mathcal{I}} \cap (\text{prim}(D_j))^{\mathcal{I}})|}$$

$$f_P(\text{prim}(C_i), \text{prim}(D_j)) := \infty \text{ if } (\text{prim}(C_i))^{\mathcal{I}} = (\text{prim}(D_j))^{\mathcal{I}}$$

$$f_{\forall}(C_i, D_j) := \sum_{R \in N_R} f_{\sqcup}(\text{val}_R(C_i), \text{val}_R(D_j))$$

$$f_{\exists}(C_i, D_j) := \sum_{R \in N_R} \sum_{k=1}^{N} \max_{p=1,\ldots,M} f_{\sqcup}(C_i^k, D_j^p)$$

where $C_i^k \in \text{ex}_R(C_i)$ and $D_j^p \in \text{ex}_R(D_j)$ and wlog.
$N = |\text{ex}_R(C_i)| \geq |\text{ex}_R(D_j)| = M$, otherwise exchange N with M

Dissimilarity Measure

The *dissimilarity measure* d is a function $d : \mathcal{L} \times \mathcal{L} \mapsto [0, 1]$ such that, for all $C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$ concept descriptions in $\mathcal{ALC}$ normal form:

$$d(C, D) := \begin{cases} 0 & f(C, D) = \infty \\ 1 & f(C, D) = 0 \\ \frac{1}{f(C,D)} & \textit{otherwise} \end{cases}$$

where f is the function overlapping

Dissimilarity Measure: example...

$C \equiv A_2 \sqcap \exists R.B_1 \sqcap \forall T.(\forall Q.(A_4 \sqcap B_5)) \sqcup A_1$
$D \equiv A_1 \sqcap B_2 \sqcap \exists R.A_3 \sqcap \exists R.B_2 \sqcap \forall S.B_3 \sqcap \forall T.(B_6 \sqcap B_4) \sqcup B_2$
where A_i and B_j are all primitive concepts.

$C_1 := A_2 \sqcap \exists R.B_1 \sqcap \forall T.(\forall Q.(A_4 \sqcap B_5))$
$D_1 := A_1 \sqcap B_2 \sqcap \exists R.A_3 \sqcap \exists R.B_2 \sqcap \forall S.B_3 \sqcap \forall T.(B_6 \sqcap B_4)$

$$f(C,D) := f_{\sqcup}(C,D) = \max\{\ f_{\sqcap}(C_1,D_1), f_{\sqcap}(C_1,B_2), f_{\sqcap}(A_1,D_1), f_{\sqcap}(A_1,B_2)\ \}$$

...Dissimilarity Measure: example...

For brevity, we consider the computation of $f_{\sqcap}(C_1, D_1)$.

$f_{\sqcap}(C_1, D_1) = f_P(\text{prim}(C_1), \text{prim}(D_1)) + f_{\forall}(C_1, D_1) + f_{\exists}(C_1, D_1)$
Suppose that $(A_2)^{\mathcal{I}} \neq (A_1 \sqcap B_2)^{\mathcal{I}}$. Then:

$$\begin{aligned} f_P(C_1, D_1) &= f_P(\text{prim}(C_1), \text{prim}(D_1)) \\ &= f_P(A_2, A_1 \sqcap B_2) \\ &= \frac{|I|}{|I \setminus ((A_2)^{\mathcal{I}} \cap (A_1 \sqcap B_2)^{\mathcal{I}})|} \end{aligned}$$

where $I := (A_2)^{\mathcal{I}} \cup (A_1 \sqcap B_2)^{\mathcal{I}}$

...Dissimilarity Measure: example...

In order to calculate $f_\forall$ it is important to note that

- There are two different role at the same level T and S
- So the summation over the different roles is made by two terms.

$$\begin{aligned} f_\forall(C_1, D_1) &= \sum_{R \in N_R} f_\sqcup(\mathrm{val}_R(C_1), \mathrm{val}_R(D_1)) = \\ &= f_\sqcup(\mathrm{val}_T(C_1), \mathrm{val}_T(D_1)) + \\ &+ f_\sqcup(\mathrm{val}_S(C_1), \mathrm{val}_S(D_1)) = \\ &= f_\sqcup(\forall Q.(A_4 \sqcap B_5), B_6 \sqcap B_4) + f_\sqcup(\top, B_3) \end{aligned}$$

...Dissimilarity Measure: example

In order to calculate $f_{\exists}$ it is important to note that

- There is only a single one role R so the first summation of its definition collapses in a single element
- N and M (numbers of existential concept descriptions w.r.t the same role (R)) are $N = 2$ and $M = 1$
  - So we have to find the max value of a single element, that can be semplifyed.

$$\begin{aligned} f_{\exists}(C_1, D_1) &= \sum_{k=1}^{2} f_{\sqcup}(\text{ex}_R(C_1), \text{ex}_R(D_1^k)) = \\ &= f_{\sqcup}(B_1, A_3) + f_{\sqcup}(B_1, B_2) \end{aligned}$$

Discussion...

- If C $\equiv$ D (namely C $\sqsubseteq$ D e D $\sqsubseteq$ C) (semantic equivalence) $d(C, D) = 0$, rather d assigns the minimun value
- If $C \sqcap D \equiv \bot$ then $d(C, D) = 1$, rather d assigns the maximum value because concepts involved are totally different
- Otherwise $d(C, D) \in]0, 1[$ rather dissimilarity is inversely proportional to the quantity of concept overlap, measured considering the entire definitions and their subconcepts.

...Discussion

The presented function d is a dissimilarity measure

1. $f(a,b) \geq 0 \quad \forall a,b \in E$ *(positive definiteness)*
2. $f(a,b) = f(b,a) \quad \forall a,b \in E$ *(symmetry)*
3. $\forall a,b \in E : a \neq b : f(a,a) < f(a,b)$

1. It is satisfied for the definition of d
2. It is satisfied by the commutativity of the sum and maximum operators.
3. It is satisfied because d assigns the minimum value only when the concepts are equivalent

Measure Involving Individuals

Let c and d two individuals in a given A-Box.
We can consider $C^* = \text{MSC}^*(c)$ and $D^* = \text{MSC}^*(d)$:

$$d(c, d) := d(C^*, D^*) = d(\text{MSC}^*(c), \text{MSC}^*(d))$$

Analogously:

$$\forall a: \; d(c, D) := d(\text{MSC}^*(c), D)$$

Dissimilarity Measure: Conclusions

- Experimental evaluations demonstrate that *d* *works satisfying* both for concepts and individuals
- *However*, for complex concept descriptions (such as MSC^*), deeply nested subconcepts could increase the dissimilarity value
- The *new idea* is to differentiate the weight of the subconcepts wrt their levels in the concept descriptions in order to determine the final dissimilarity value

The weighted Dissimilarity Measure

Overlap Function Definition [d'Amato '05 @ SWAP 2005]:

$\mathcal{L} = \mathcal{ALC}/_{\equiv}$ the set of all concepts in $\mathcal{ALC}$ normal form

$\mathcal{I}$ canonical interpretation of A-Box $\mathcal{A}$

$f : \mathcal{L} \times \mathcal{L} \mapsto R^+$ defined $\forall C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$ in $\mathcal{L}_{\equiv}$

$$f(C,D) := f_{\sqcup}(C,D) = \begin{cases} |\Delta| & C \equiv D \\ 0 & C \sqcap D \equiv \bot \\ 1 + \lambda \cdot \max_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}} f_{\sqcap}(C_i, D_j) & \text{o.w.} \end{cases}$$

$$f_{\sqcap}(C_i, D_j) \quad := \quad f_P(\text{prim}(C_i), \text{prim}(D_j)) + f_{\forall}(C_i, D_j) + f_{\exists}(C_i, D_j)$$

Looking toward Information Content: Motivation

- In [Borgida '05 @ DL 2005] the same necessity of generalize previous efforts to define similarity for primitive concepts to composite ones in presented
- The three classical approaches are applied to a poorly expressive DL, where only conjunction is allowed
- Open problems in defining similarity measures for most complex DL are illustrated
- *The use of Information Content* is presented as *the most effective way for measuring complex concept descriptions*

Information Content: Defintion

- A measure of concept (dis)similarity can be derived from the notion of *Information Content* (IC)
- IC depends on the probability of an individual to belong to a certain concept
  - $IC(C) = -\log pr(C)$
- In order to approximate the probability for a concept C, it is possible to recur to its extension wrt the considered ABox.
  - $pr(C) = |C^{\mathcal{I}}|/|\Delta^{\mathcal{I}}|$

Function Definition /I

[d'Amato '05 @ SAC 2006] $\mathcal{L} = \mathcal{ALC}/_{\equiv}$ the set of all concepts in $\mathcal{ALC}$ normal form

$\mathcal{I}$ canonical interpretation of A-Box $\mathcal{A}$

$f : \mathcal{L} \times \mathcal{L} \mapsto R^+$ defined $\forall C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$ in $\mathcal{L}_{\equiv}$

$$f(C, D) := f_{\sqcup}(C, D) = \begin{cases} 0 & C \equiv D \\ \infty & C \sqcap D \equiv \bot \\ \max_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}} f_{\sqcap}(C_i, D_j) & \text{o.w.} \end{cases}$$

$$f_{\sqcap}(C_i, D_j) \quad := \quad f_P(\text{prim}(C_i), \text{prim}(D_j)) + f_{\forall}(C_i, D_j) + f_{\exists}(C_i, D_j)$$

Function Definition / II

$$f_P(\text{prim}(C_i), \text{prim}(D_j)) := \begin{cases} \infty & \text{if } \text{prim}(C_i) \sqcap \text{prim}(D_j) \equiv \bot \\ \frac{IC(\text{prim}(C_i) \sqcap \text{prim}(D_j)) + 1}{IC(LCS(\text{prim}(C_i), \text{prim}(D_j))) + 1} & \text{o.w.} \end{cases}$$

$$f_{\forall}(C_i, D_j) := \sum_{R \in N_R} f_{\sqcup}(\text{val}_R(C_i), \text{val}_R(D_j))$$

$$f_{\exists}(C_i, D_j) := \sum_{R \in N_R} \sum_{k=1}^{N} \max_{p=1,\ldots,M} f_{\sqcup}(C_i^k, D_j^p)$$

where $C_i^k \in \text{ex}_R(C_i)$ and $D_j^p \in \text{ex}_R(D_j)$ and wlog.
$N = |\text{ex}_R(C_i)| \geq |\text{ex}_R(D_j)| = M$, otherwise exchange N with M

Dissimilarity Measure: Definition

The *dissimilarity measure* d is a function $d : \mathcal{L} \times \mathcal{L} \mapsto [0, 1]$ such that, for all $C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$ concept descriptions in $\mathcal{ALC}$ normal form:

$$d(C, D) := \begin{cases} 0 & f(C, D) = 0 \\ 1 & f(C, D) = \infty \\ 1 - \frac{1}{f(C,D)} & otherwise \end{cases}$$

where f is the function defined previously

Discussion

- $d(C, D) = 0$ iff IC=0 iff C $\equiv$ D (semantic equivalence) rather d assigns the minimun value
- $d(C, D) = 1$ iff IC $\rightarrow \infty$ iff $C \sqcap D \equiv \bot$, rather d assigns the maximum value because concepts involved are totally different
- Otherwise $d(C, D) \in]0, 1[$ rather d tends to 0 if IC tends to 0; d tends to 1 if IC tends to infinity

Measures Involving Individuals

Let c and d two individuals in a given A-Box.
We can consider $C^* = \text{MSC}^*(c)$ and $D^* = \text{MSC}^*(d)$:

$$d(c, d) := d(C^*, D^*) = d(\text{MSC}^*(c), \text{MSC}^*(d))$$

Analogously:

$$\forall a: \; d(c, D) := d(\text{MSC}^*(c), D)$$

Dissimilarity Measures Complexity

Let $C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$ be in normal form:

- *C and D are semantically equivalent* $Compl(d) = 2 \cdot Cmpl(\sqsupseteq)$
- *C and D are disjoint yet not semantically equivalent* same complexity of the previous case
- *C and D are not semantically equivalent nor disjoint.* computing $f_{\sqcap}$ for $n \cdot m$ times: $Cmpl(d) = nm \cdot Cmpl(f_{\sqcap}) = nm \cdot [Cmpl(f_P) + Cmpl(f_{\forall}) + Cmpl(f_{\exists})]$

Complexity / II

- The dominant operation for f_P is instance checking (IC): $C(f_P) = 2 \cdot C(IC)$.
- The computation of $f_\forall$ and $f_\exists$ apply recursively the definition of $f_\sqcup$ on less complex descriptions.
  A maximum of $|N_R|$ calls of $f_\sqcup$ are needed for computing $f_\forall$, while the calls of $f_\sqcup$ needed for $f_\exists$ are $|N_R| \cdot N \cdot M$, where $N = |\mathrm{ex}_R(C_i)|$ and $M = |\mathrm{ex}_R(D_j)|$
- Summing up $Cmpl(d) = nm \cdot [(2 \cdot Cmpl(IC)) + (|N_R| \cdot Cmpl(f_\sqcup)) + (|N_R| \cdot M \cdot N \cdot Cmpl(f_\sqcup))]$

The computation of d depends on IC: P-space $\mathcal{ALC}$
Nevertheless, in practical applications: exploit the statistics that are maintained by the DBMSs query optimizers

Conclusions...

- The presented function are *Dissimilarity Measures*
  - They are definite positive, symmetric, and has minimal value only when the concepts are equal (in the sense of semantic equivalence)
- The presented Dissimilarity Measures are *semantic* and they are able to involve individuals, concepts and individual and concept
- **Dissimilarity Measures** can be applied to knowledge bases expressed in OWL and *ALC* DL
  - They can be applied to any DL which has *IC*, *LCS* (and *MSC*/*MSC**) operators

...Conclusions

- The *Complexity* of Dissimilarity Measures depends from the complexity of the instance checking operator for the chosen DL
- Dissimilarity Measures are defined using the set theory and reasoning operators
  - **They use a numerical approach but are applied on symbolic representations**

Motivations

- New defined similarity and/or dissimilarity measures need to be validated
- Validation w.r.t the human judgment is too subjective because different humans can express different similarity degree of the same object (concept)
- An *automatic validation* is more reliable and less subjective
- Realization of a classification algorithm
  - settled to validate the proposed measures
  - aiming to make the *populating A-Box* task less time consuming, *adding new information (not derivable)*

K-NN: Peculiarities

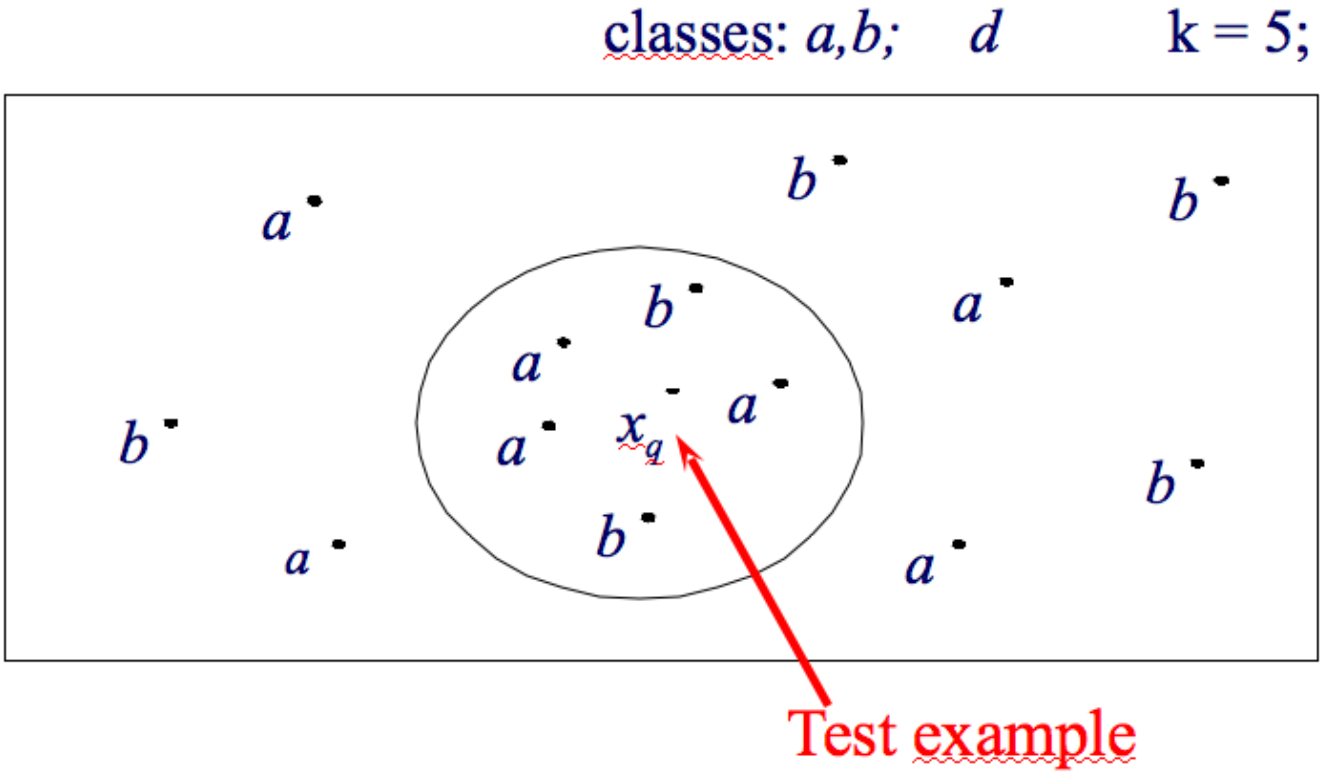
- Lazy Learning Algorithm
  - Learning phase consists in memorizing training example
- Classification results are given by analogy w.r.t. K selected training examples that are most similar to the examples to classify
- Intermediate information and classification results are discarded after the classification of a test example

Classical K-NN algorithm...

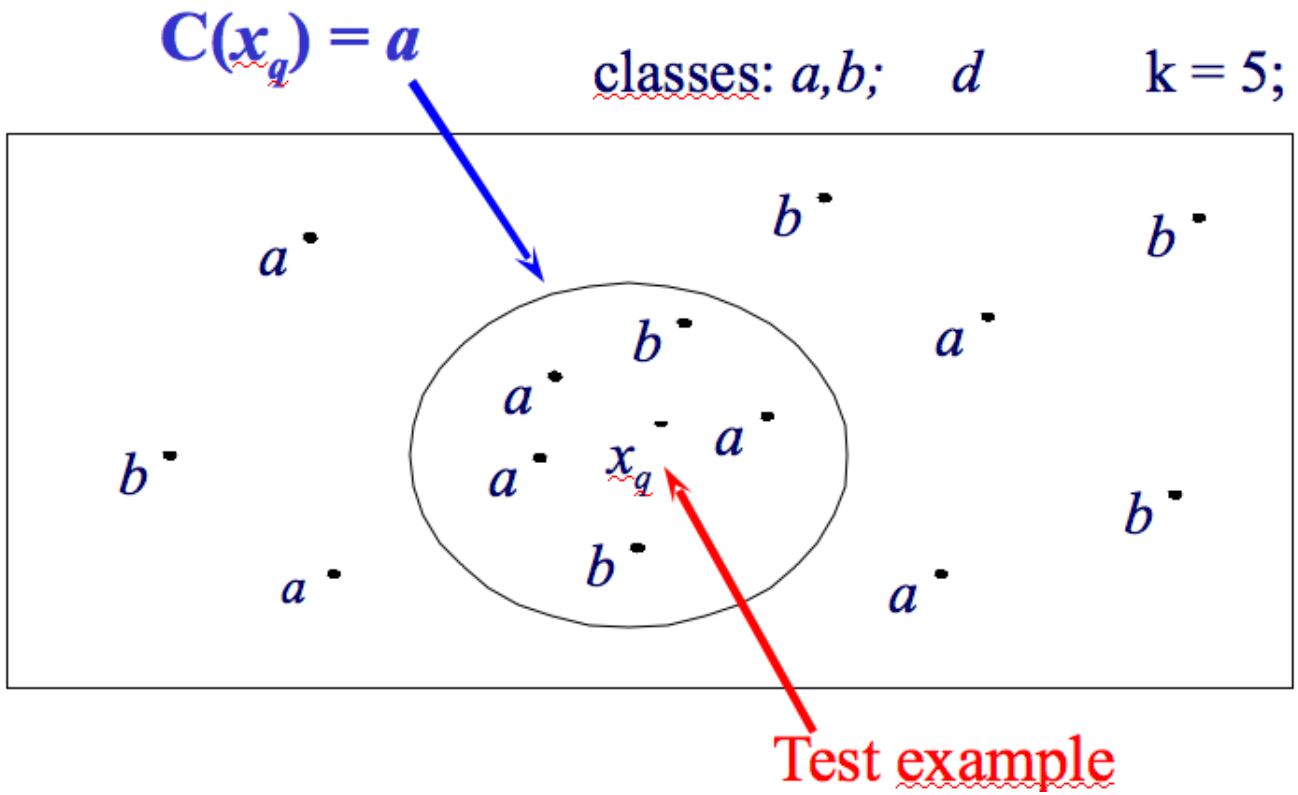
- **Training Phase:** All training examples are memorized jointly with the classes to which they belong to
- **Testing Phase:**
  - Given a test example x_q and a dissimilarity measure d, the k training elements less dissimilar from x_q are determined
  - $C(x_q) = argmax_{v \in V} \sum_{i=1}^{k} \delta(v, C(x_i))$

    where V is the set of known classes; $\delta(a, b) = 1$ if $a = b$; $\delta(a, b) = 0$ if $a \neq b$

...Classical K-NN algorithm...



...Classical K-NN algorithm...



...Classical K-NN algorithm

- Generally applied to feature vector representation
- In classification phase it is assumed that each training and test example belong to a single class, so classes are considered to be disjoint
- An implicit *Closed World Assumption* is made

Difficulties in applying K-NN to Ontological Knowledge

To apply K-NN for classifying individual asserted in an ontological knowledge base

1. It has to find a way for applying K-NN to a most complex and expressive knowledge representation
2. It is not possible to assume disjointness of classes. Individuals in an ontology can belong to more than one class (concept).
3. The classification process has to cope with the *Open World Assumption* charactering Semantic Web area

Choices for applying K-NN to Ontological Knowledge

1. To have similarity and dissimilarity measures applicable to ontological knowledge allows applying K-NN to this kind of knowledge representation
2. A new classification procedure is adopted, decomposing the multi-class classification problem into smaller binary classification problems (one per target concept).
   - For each individual to classify w.r.t each class (concept), classification returns $\{-1,+1\}$
3. A third value 0 representing unknown information is added in the classification results $\{-1,0,+1\}$
4. Hence a majority voting criterion is applied

Realized K-NN Algorithm...

- **Main Idea:** similar individuals, by analogy, should likely belong to similar concepts
  - for every ontology, all individuals are classified to be instances of one or more concepts of the considered ontology
- For each individual in the ontology MSC is computed
- MSC list represents the set of training examples

...Realized K-NN Algorithm

- Each example is classified applying the k-NN method for DLs, adopting the leave-one-out cross validation procedure.

$$\hat{h}_j(x_q) := \operatorname*{argmax}_{v \in V} \sum_{i=1}^{k} \delta(v, h_j(x_i)) \qquad \forall j \in \{1, \ldots, s\} \quad (1)$$

where

$$h_j(x) = \begin{cases} +1 & C_j(x) \in \mathcal{A} \\ 0 & C_j(x) \notin \mathcal{A} \\ -1 & \neg C_j(x) \in \mathcal{A} \end{cases}$$

Experimentation Setting...

- **FSM** ontology (Protege Library): describes finite state machines. It is made up of:
  - 20 concepts (both primitives and defined), some of them are declared to be disjoint
  - 10 object properties, 7 datatype properties
  - 37 individuals. About half are instance of only a single class and are not involved in any property; other half is involved in properties.

...Experimentation Setting...

- **Surface-Water-Model** (Protege Library) describes water quality models. It is made up of:
  - 19 concepts (both primitives and defined), there not specification about disjointness
  - 9 object properties, 115 datatype properties.
  - 115 individuals. All are instances of a single class, only few of them are involve in object properties

...Experimentation Setting

- **Family** (handcrafted ontology) describes family relationship
  - 14 concepts (both primitives and defined), some of them are declared to be disjoint
  - 5 object properties.
  - 39 individuals. Major of them are instances of more that one concept and are involved in more than one object property

Measures for Evaluating Experiments

- **Predictive Accuracy:** measures the number of correctly classified individuals w.r.t. overall number of individuals.
- **Omission Error Rate:** measures the amount of unlabelled individuals $C(x_q) = 0$ with respect to a certain concept C_j while they are instances of C_j in the KB.
- **Commission Error Rate:** measures the amount of individuals labelled as instances of the negation of the target concept C_j, while they belong to C_j or vice-versa.
- **Induction Rate:** measures the amount of individuals that were found to belong to a concept or its negation, while this information is not derivable from the KB.

Experimentation Evaluation

Average results of the trials using KCAP measure

Ontologies	Predictive Accuracy	Omission Error	Induction Rate	Commission Error
FSM	100	0	31	0
S.-W.-M.	100	0	0	0
FAMILY	44.25	55.75	14	0

Average results of the trials employing SAC measure

Ontologies	Predictive Accuracy	Omission Error	Induction Rate	Commission Error
FSM	100	0	31	0
S.-W.-M.	100	0	0	0
FAMILY	49.07	50.93	16.85	0

Experimentation: Discussion...

- for every ontology, the *commission error is null*; the classifier never mades critical mistakes
- **SURFACE-WATER-MODEL Ontology**: the classifier always assigns individuals to the correct concepts; it is never capable to induce new knowledge
  - Because individuals are all instances of a single concept and are involved in a few roles, so MSCs are very similar and so the amount of information they convey is very low

...Experimentation: Discussion...

FSM Ontology

- The classifier always assigns individuals to the correct concepts
  - Because most of individuals are instances of a single concept
- Induction rate is not null so knew knowledge is induced
  - Due mainly to the presence of some concepts that are declared to be mutually disjoint, secondary because some individuals are involved in relations

...Experimentation: Discussion

FAMILY Ontology

- Predictive Accuracy is not so high and Omission Error not null
  - Because instances are more irregularly spread over the classes, so computed MSCs are often very different provoking sometimes incorrect classifications (weakness on K-NN algorithm)
- No Commission Error (but only omission error)
- The *Classifier* is able of *induce new knowledge* that is *not derivable*

Comparing Family Ontology Results...

FAMILY *ontology – KCAP measure.*

	Predictive Accuracy	Omission Error	Induction Rate	Commission Error
Female	64	36	7.69	0
Woman	64	36	7.69	0
Mother	0	100	5.12	0
Male	12.5	87.5	23	0
Man	12.5	87.5	23	0
Father	0	100	23	0
Human	100	0	2.56	0
Child	100	0	25.64	0
Sibling	62.5	37.5	41	0
Parent	29	71	2.56	0
Grandparent	75	25	0	0
Grandchild	100	0	36	0
Cousin	0	100	0	0
UncleAunt	0	100	14	0
average	44.25	55.75	14	0

FAMILY *ontology – SAC measure.*

	Predictive Accuracy	Omission Error	Induction Rate	Commission Error
Female	75	25	30.76	0
Woman	75	25	35.89	0
Mother	0	100	30.76	0
Male	83.36	16.64	30.76	0
Man	83.36	16.64	33.33	0
Father	14.28	85.72	30.76	0
Human	100	0	2.56	0
Child	80.95	19.05	12.82	0
Sibling	0	100	0	0
Parent	37.5	62.5	12.82	0
Grandparent	50	50	5.12	0
Grandchild	37.5	62.5	12.82	0
Cousin	50	50	0	0
UncleAunt	0	100	0	0
average	49.07	50.93	16.85	0

...Comparing Family Ontology Results...

- *SAC measure improves the classification* of most of concepts (classes) w.r.t. KCAP measure
  - Father (+14.28), Man (+70.86), Parent (+8.5), Female (+11), Male (+70, 86), Woman (+11), Cousin (+50)
- The predictive accuracy of only a few classes decreases w.r.t. KCAP measure
  - Child (−19.05), Sibling (−100), Grandchild (−62.5), GrandParent (−25)
- *The average predictive accuracy of SAC measure is not so high* w.r.t. those of KCAP measure because the decreasing of the predictive accuracy is quite high for some classes (e.g. Child)

...Comparing Family Ontology Results...

- SAC measure *increases results in classifying concepts* that have *poorer predictive accuracy w.r.t. KCAP measure* (e.g. see the results for the concepts Male, Man *and Cousin) and vice-versa.*
- *SAC measure classifies poorly concepts* that have *less information* in the ontology
- *SAC measure is less able*, w.r.t. KCAP measure, *to classify concepts* correctly, when they have *few information* (instance and object properties involved);
- *When concepts have enough information, SAC measure classifies notably better* than KCAP measure.

...Comparing Family Ontology Results

- The two measures give the same predictive accuracy for the concepts: Human (100), Uncle (0) and Mother (0).
  - because all individuals in the ontology are instance of Human, while there is scarce information about Mother and Uncle.
- *SAC measure* generates a *higher induction rate* (+2.85) w.r.t. KCAP measure
- Summarizing *SAC measure slightly increases the overall performance* w.r.t. KCAP measure
- Considering the complementarity of the results of the two measures, seems to be interesting the **definition of a new dissimilarity measure that combines**, in some way, **the two** tested **measures**

Future Work

- *Test* Similarity and Dissimilarity Measures using some *clustering algorithms*
- Extention of Similarity and Dissimilarity Measures for most expressive DL such as *ALCN*
- Definition of new Similarity/Dissimilarity Measures for DLs representations, using *Kernel functions* that are a means to express a notion of similarity in some unknown feature space. Thus it could be possible exploiting the efficiency of kernel methods (e.g. SVMs) in a relational setting
- Application of Similarity and Dissimilarity Measures for the *matchmaking and/or composition of services* (described in OWL-S)

The End

That's all!

Claudia d'Amato
claudia.damato@di.uniba.it
Nicola Fanizzi
fanizzi@di.uniba.it