(Dis-)Similarity Measures for Description Logics Representation

Claudia d'Amato

Computer Science Department • *University of Bari*

Poznan, 22 June 2011

Contents

Starting Point

- **Problem:** Similarity measures for complex concept descriptions (as those in the ontologies) not deeply investigated **[Borgida et al. 2005]**

Approaches for Computing Similarities

- **Dimension Representation**: feature vectors, strings, sets, trees, clauses...
- **Dimension Computation**: geometric models, feature matching, semantic relations, Information Content, alignment and transformational models, contextual information...
- Distinction: *Propositional* and *Relational* setting
    - analysis of computational models

Propositional Setting: Measures based on Geometric Model

- **Propositional Setting**: Data are represented as n-tuple of fixed length in an n-dimentional space
- **Geometric Model:** objects are seen as *points in an n-dimentional space*.
  - The *similarity* between a pair of objects is considered *inversely related to the distance* between two objects points in the space.
  - Best known distance measures: *Minkowski* measure, *Manhattan* measure, *Euclidean* measure.
- Applied to vectors whose *features* are *all continuous*.

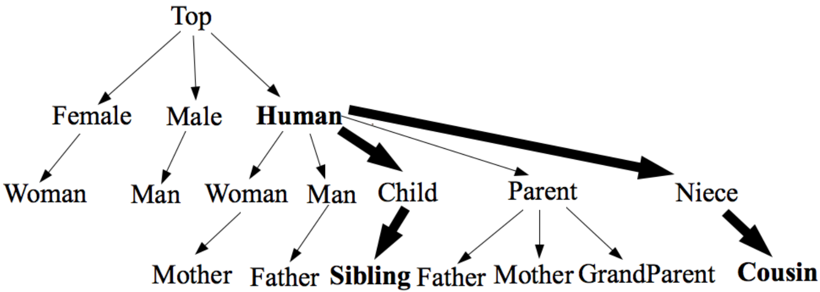
Similarity Measures based on Feature Matching Model

- **Features** can be of **different types**: binary, nominal, ordinal
- *Tversky's Similarity Measure* **[Tversky,77]**: based on the notion of *contrast model*
  - **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
  - it is assumed that *all features have the same importance*
- **Measures in propositional setting are not able to capture expressive relationships among data** that typically characterize most complex languages.

Relational Setting: Measures Based on Semantic Relations

- Also called **Path distance measures** **[Bright,94]**
- Measure the *similarity* value between single words (*elementary concepts*)
- 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* w.r.t. the most specific ancestor
  - terms with 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.

Measures Based on Semantic Relations: Example



Measures Based on Semantic Relations: WEAKNESS

- the similarity value is subjective due to the taxonomic ad-hoc representation
- the introduction of new terms 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

Measures Based on Information Content...

- Measure semantic similarity of concepts in an *is-a* taxonomy by the use of notion of *Information Content (IC)* [Resnik,99]
- Concepts similarity is given by the shared information
  - 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

...Measures Based on Information Content

- 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**

Similarity Measures for Very Low Expressive DLs...

- Measures for complex concept descriptions **[Borgida et al. 2005]**
  - A DL allowing only *concept conjunction* is considered (propositional DL)
- **Feature Matching Approach**:
  - features are represented by atomic concepts
  - An ordinary concept is the conjunction of its features
  - *Set intersection* and *difference* corresponds to the *LCS* and *concept difference*
- **Semantic Network Model** and **IC models**
  - The *most specific ancestor* is given by the *LCS*

...Similarity Measures for Very Low Expressive DLs

OPEN PROBLEMS in considering most expressive DLs:

- What is a *feature* in most expressive DLs?
  - i.e. $(\leq 3R), (\leq 4R)$ and $(\leq 9R)$ are three different features? or $(\leq 3R), (\leq 4R)$ are more similar w.r.t $(\leq 9R)$?
  - How to assess similarity in presence of role restrictions? i.e. $\forall R.(\forall R.A)$ and $\forall R.A$
- *IC-based model*: how to compute the value $p(C)$ for assessing the IC?

Why New Measures

- **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
  - **There are no measure for assessing similarity between individuals**
- **Defining new measures that are really semantic is necessary**

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 et al. @ 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:

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

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

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 $\geq$ 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

$$
\begin{aligned}
s(\mathsf{Grandparent}, \mathsf{Father}) &= \frac{|(\mathsf{Grandparent} \sqcap \mathsf{Father})^{\mathcal{I}}|}{|\mathsf{Granparent}^{\mathcal{I}}| + |\mathsf{Father}^{\mathcal{I}}| - |(\mathsf{Grandarent} \sqcap \mathsf{Father})^{\mathcal{I}}|} \cdot \\
&\cdot\ max\left(\frac{|(\mathsf{Grandparent} \sqcap \mathsf{Father})^{\mathcal{I}}|}{|\mathsf{Grandparent}^{\mathcal{I}}|}, \frac{|(\mathsf{Grandparent} \sqcap \mathsf{Father})^{\mathcal{I}}|}{|\mathsf{Father}^{\mathcal{I}}|}\right) = \\
&= \frac{2}{2+3-2} \cdot max\left(\frac{2}{2}, \frac{2}{3}\right) = 0.67
\end{aligned}
$$

Similarity Measure between Individuals

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

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

Analogously:

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

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)
  - *Intuition*: **Concepts defined by almost the same sub-concepts will be probably similar**

MSC : An Example*

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

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

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

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

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

$\mathcal{ALC}$ 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

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

where:

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

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

Overlap Function

Definition [d'Amato et al. @ 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 $\mathcal{L}_{\equiv}$

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

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

Overlap Function / II

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

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

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

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

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:

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

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.

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

...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.

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

Dissimilarity Measure: Conclusions

- Experimental evaluations demonstrate that d *works quite well* both for concepts and individuals
- *However*, for complex descriptions (such as MSC^*), deeply nested subconcepts could increase the dissimilarity value
- **New idea:** differentiate the weight of the subconcepts wrt their levels in the descriptions for determining the final dissimilarity value
  - **Solve the problem:** *how differences in concept structure might impact concept (dis-)similarity? i.e. considering the series* $dist(B, B \sqcap A), dist(B, B \sqcap \forall R.A), dist(B, B \sqcap \forall R.\forall R.A)$ *this should become smaller since more deeply nested restrictions ought to represent smaller differences.*" **[Borgida et al. 2005]**

The weighted Dissimilarity Measure

Overlap Function Definition [d'Amato et al. @ 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 $\mathcal{L}_{\equiv}$

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

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

Looking toward Information Content: Motivation

- *The use of Information Content* is presented as *the most effective way for measuring complex concept descriptions* **[Borgida et al. 2005]**
- The necessity of considering concepts in normal form for computing their (dis-)similarity is argued **[Borgida et al. 2005]**
  - *confirmation* of the used approach in the previous measure
- **A dissimilarity measure for complex descriptions grounded on IC has been defined**

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.
  - $pr(C) = |C^{\mathcal{I}}|/|\Delta^{\mathcal{I}}|$
- A function for measuring the *IC variation* between concepts is defined

Function Definition /I

[d'Amato et al. @ 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 $\mathcal{L}_{\equiv}$

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

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

Function Definition / II

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

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

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

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

Other Structural-Based Similarity Measures

- By exploiting a similar approach measures for more expressive DLs have been set up:
  - A Similarity Measure for $\mathcal{ALN}$ **[Fanizzi et. al @ CILC 2006]**
  - A similarity measure for $\mathcal{ALCNR}$ **[Janowicz, 06]**
  - A similarity measure for $\mathcal{ALCHQ}$ **[Janowicz et al., 07]**
- The "trick" consists in assessing an overlap function for each construtor of the considered logic and then aggregate the results of the overlap functions
- **Lesson Learnt:** a new measure has to be defined for each available logic $\Rightarrow$ *The measure does not easily scale to more expressive DLs*

The GCS-based Similarity Measure: Rationale

Two concepts are more similar as much their extensions are similar

- the similarity value is given by the variation of the number of instances in the concept extensions w.r.t. the number of instances in the extension of their common super-concept
  - Common super-concept $\Rightarrow$ the GCS of the concepts **[Baader et al. 2004]**

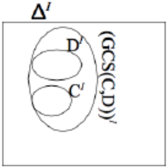


Fig. 1. Concepts C $\equiv$credit-card-payment, D $\equiv$debit-card-payment are similar as the extension of their GCS$\equiv$card-payment does not include many other instances besides of those of their extensions.

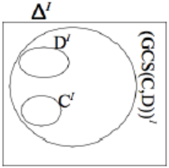


Fig. 2. Concepts C $\equiv$car-transfer, D $\equiv$ debit-card-payment are different as the extension of their GCS$\equiv$service includes many other instances besides of those of the extension of C and D.

The GCS-based Similarity Measure: Defintion

Definition: [d'Amato et al. @ SMR2 WS at ISWC 2007]

Let $\mathcal{T}$ be an $\mathcal{ALC}$ TBox. For all C and D $\mathcal{ALE}(\mathcal{T})$-concept descriptions, the function $s : \mathcal{ALE}(\mathcal{T}) \times \mathcal{ALE}(\mathcal{T}) \rightarrow [0,1]$ is a *Semantic Similarity Measure* defined as follow:

$$s(C,D) = \frac{\min(|C^I|,|D^I|)}{|(GCS(C,D))^I|} \cdot (1 - \frac{|(GCS(C,D))^I|}{|\Delta^I|} \cdot (1 - \frac{\min(|C^I|,|D^I|)}{|(GCS(C,D))^I|}))$$

where $(\cdot)^I$ computes the concept extension w.r.t. the interpretation I (canonical interpretation).

Semi-Distance Measure: Motivations

- Most of the presented measures are grounded on concept structures $\Rightarrow$ hardly scalable w.r.t. most expressive DLs
- **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

Semantic Semi-Dinstance Measure: Definition

[Fanizzi et al. @ DL 2007] Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a KB and let $\mathsf{Ind}(\mathcal{A})$ be the set of the individuals in $\mathcal{A}$. Given sets of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$ in $\mathcal{T}$, a *family of semi-distance functions* $d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto \mathbb{R}$ is defined as follows:

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \quad d_p^{\mathsf{F}}(a, b) := \frac{1}{m} \left[\sum_{i=1}^{m} \mid \pi_i(a) - \pi_i(b) \mid^p \right]^{1/p}$$

where $p > 0$ and $\forall i \in \{1, \ldots, m\}$ the *projection function* π_i is defined by:

$$\forall a \in \mathsf{Ind}(\mathcal{A}) \quad \pi_i(a) = \begin{cases} 1 & F_i(a) \in \mathcal{A} \quad (\mathcal{K} \models F_i(a)) \\ 0 & \neg F_i(a) \in \mathcal{A} \quad (\mathcal{K} \models \neg F_i(a)) \\ \frac{1}{2} & \textit{otherwise} \end{cases}$$

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

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

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

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

Semi-Distance Measure: Discussion

- The measure is a semi-distance
    - $d_p(a, b) \geq 0$ and $d_p(a, b) = 0$ if $a = b$
    - $d_p(a, b) = d_p(b, a)$
    - $d_p(a, c) \leq d_p(a, b) + d_p(b, c)$
- *it does not guaranties* that if $d_p^{\mathsf{F}}(a, b) = 0 \Rightarrow a = b$

Defining the Weights

- To take into account the **discriminating power of each feature** **[d'Amato et al. @ ESWC'08]**
  1. the **weights reflect the amount of information conveyed by each feature** (quantity estimated by the entropy of the features)

     $$H(F_i) = P^i_{-1}\log(1/P^i_{-1}) + P^i_0\log(1/P^i_0) + P^i_{+1}\log(1/P^i_{+1})$$

     where $P^i_v = (\mathrm{check}(a \in F_i) = v)/\mathsf{Ind}(\mathcal{A})$ and $v = \{-1, 0, +1\}$
     then, the weights are set as: $w_i := H(F_i)/\sum_j H(F_j)$, for $i = 1, \ldots, m$.
  2. **estimate of the feature variance**

     $$\widehat{\mathrm{var}}(F_i) = \frac{1}{2 \cdot |\mathsf{Ind}(\mathcal{A})|^2} \sum_{a \in \mathsf{Ind}(\mathcal{A})} \sum_{b \in \mathsf{Ind}(\mathcal{A})} [\pi_i(a) - \pi_i(b)]^2$$

     which induces the choice of weights: $w_i = 1/(2 \cdot \widehat{\mathrm{var}}(F_i))$, for $i = 1, \ldots, m$.

Measure Optimization: Feature Selection

- **Implicit assumption:** F represents a sufficient number of (possibly redundant) features that are really able to discriminate different individuals
- The choice of the concepts to be included in F could be crucial for the correct behavior of the measure
  - a "good" feature committee may discern individuals better
  - a smaller committee yields more efficiency when computing the distance
  - **Proposed optimization algorithms grounded on stochastic search that are able to find/build optimal discriminating concept committees [Fanizzi et al. @ IJSWIS'08]**
- *Experimentally obtained good results by using the very set of both primitive and defined concepts in the ontology*

Optimal Discriminating Feature Set

- **Proposal of optimization algorithms** that are able to find/build optimal discriminating concept committees **[Fanizzi et al. @ IJSWIS'08]**
  - **Idea:** Optimization of a *fitness function* that is based on the *discernibility factor of the committee*, namely
  - Given $\mathsf{Ind}(\mathcal{A})$ (or just a hold-out sample) $HS \subseteq \mathsf{Ind}(\mathcal{A})$ *find* the subset F that maximize the following function:

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

Characterizing a "Semantic Similarity Measure"

[d'Amato et al. @ EKAW 2008]

- *Expected behaviors* of a *semantic similarity measure* applied to *ontological knowledge*
- Current Similarity measures fail (some of) the expected behaviors
- *Formalization of criteria* that a measure has to *satisfy* for correctly coping with ontological representation

Motivating Example

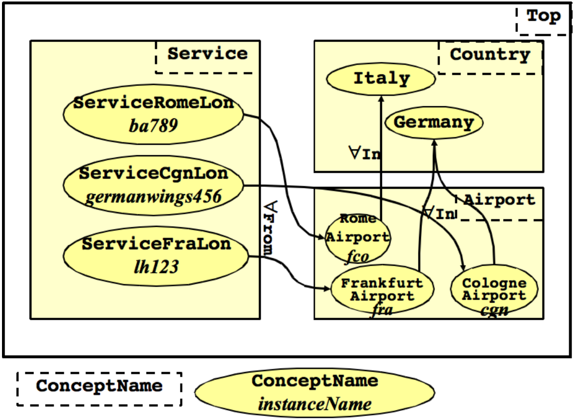
$\mathcal{T} = \{\text{Service} \sqsubseteq \text{Top};\ \text{Airport} \sqsubseteq \text{Top} \sqcap \neg\text{Service}; \text{Town} \sqsubseteq \text{Top} \sqcap \neg\text{Service} \sqcap \neg\text{Airport};$
$\text{Country} \sqsubseteq \text{Top} \sqcap \neg\text{Service} \sqcap \neg\text{Town} \sqcap \neg\text{Airport};\ \text{Germany} \sqsubseteq \text{Country};$
$\text{Italy} \sqsubseteq \text{Country} \sqcap \neg\text{Germany};\ \text{UK} \sqsubseteq \text{Country} \sqcap \neg\text{Germany} \sqcap \neg\text{Italy};$
$\text{CologneAirport} \sqsubseteq \text{Airport} \sqcap \forall\text{In.Germany};\ \text{RomeAirport} \sqsubseteq \text{Airport} \sqcap \forall\text{In.Italy};$
$\text{FrankfurtAirport} \sqsubseteq \text{Airport} \sqcap \forall\text{In.Germany} \sqcap \neg\text{CologneAirport};$
$\text{LondonAirport} \sqsubseteq \text{Airport} \sqcap \forall\text{In.UK}\ \}$

$\mathcal{A} = \{\text{FrankfurtAirport(fra); CologneAirport(cgn); RomeAirport(fco); LondonAirport(lhr)}\}$

$$\begin{aligned}
\text{ServiceFraLon} &= \text{Service} \sqcap \exists\text{From.FrankfurtAirport} \sqcap \forall\text{From.FrankfurtAirport} \sqcap \\
&\quad \sqcap \exists\text{To.LondonAirport} \sqcap \forall\text{To.LondonAirport} \\
\text{ServiceCgnLon} &= \text{Service} \sqcap \exists\text{From.CologneAirport} \sqcap \forall\text{From.CologneAirport} \sqcap \\
&\quad \sqcap \exists\text{To.LondonAirport} \sqcap \forall\text{To.LondonAirport} \\
\text{ServiceRomeLon} &= \text{Service} \sqcap \exists\text{From.RomeAirport} \sqcap \forall\text{From.RomeAirport} \sqcap \\
&\quad \sqcap \exists\text{To.LondonAirport} \sqcap \forall\text{To.LondonAirport}
\end{aligned}$$

ServiceFraLon(*lh456*); ServiceCgnLon(*germanwings123*); ServiceRomeLon(*ba789*)

Sketch of the KB



Expected Behavior: *Soundness*

- which service (at the concept level) brings us to London?
- ServiceFraLon ⇒ if Frankfurt airport is not usable
  - ServiceCgnLon *should be favored over* ServiceRomeLon, since it is known from the KB that FrankfurtAirport and CologneAirport are both Airports in Germany
- To do this, *a similarity measure needs to appreciate the underlying ontology semantics*. We call this **expected behavior** of a similarity measure **soundness**

Expected Behavior: *Equivalence Soundness*

Let us assume that the following definition:

$$\begin{aligned}\mathsf{ServiceItLon} = {} & \mathsf{Service} \sqcap \exists \mathsf{From.RomeAirport} \sqcap \forall \mathsf{From.RomeAirport} \sqcap \\ & \sqcap \forall \mathsf{From.ItalianAirport} \sqcap \exists \mathsf{To.LondonAirport} \sqcap \forall \mathsf{To.London}\end{aligned}$$

is semantically equivalent to ServiceRomeLon

we should have
$sim(\mathsf{ServiceItLon}, \mathsf{ServiceCgnLon}) =$
$sim(\mathsf{ServiceRomeLon}, \mathsf{ServiceCgnLon})$

We call this *expected behavior* **equivalence soundness**

Expected Behavior: *disjointness compatibility*

Similarity between disjoint concepts needs not always to be zero

- *Ex.* : Let us suppose ServiceCgnLon $\equiv \neg$ServiceFraLon
- Analyzing ServiceCgnLon and ServiceFraLon, they are not totally different:
  - both perform a flight from a German airport to London
- *Consequently, it should be:*
  $sim(\text{ServiceCgnLon}, \text{ServiceFraLon}) >$
  $sim(\text{ServiceCgnLon}, \text{Service})$ where the only known thing is that ServiceCgnLon *is a* Service

We call the *ability of a similarity measure to recognize similarities between disjoint concepts* **disjointness compatibility**

Extensional-based Similarity Measures

- Basically inspired by the *Jaccard similarity measure* and the Tversky's *contrast model*
- *Similarity measures for DL concept descriptions* assign a value that is mainly proportional to the overlap of the concept extensions *[d'Amato et al.@ CILC'05]*
- **This approach fails the soundness criterion** (it is not able to fully convey the underlying ontology semantics)
  - $sim(\text{ServiceFraLon}, \text{ServiceCgnLon}) = 0$ *since* they do not share any instance.
- **This approach fails the disjointness compatibility criterion**
  - the measures cannot recognize similarities between disjoint concepts

Intentional-based Similarity Measures 1/3

Intentional-based similarity measures exploit the structure of the concept definitions for assessing their similarity

- The *similarity* of two concepts C and D (in a is-taxonomy) is given by the *length of the shortest path connecting* C and D: $sim(C, D) = length(C, E) + length(D, E)$ where E is the *msa* of C and D *[Rada et al.'89]*
  - **This measure violates the soundness criterion**
  - **Ex** : Given ServiceFraLon, ServiceCgnLon and ServiceRomeLon and their *msa* that is Service we have:
    - $sim(\text{ServiceFraLon}, \text{ServiceCgnLon}) = sim(\text{ServiceFraLon}, \text{ServiceRomeLon})$
    - *but*, from the KB, ServiceFraLon and ServiceCgnLon are more semantically similar than ServiceFraLon and ServiceRomeLon

Intentional-Based Similarity Measures 2/3

- Other similarity measures compute concept similarity by comparing the syntactic DL concept descriptions. *[d'Amato et al. @ SAC'06, Janowicz'06, Janowicz et al. '07]*
- The *similarity* value *is computed by comparing the building blocks of the concept descriptions* (primitive concepts, universal and existential value restrictions...)
- **These measures fail the equivalence soundness criterion**
  - **EX** : given the concept $\mathsf{Parent} \equiv \mathsf{Human} \sqcap \exists \mathsf{hasChild.Human}$ and the following equivalent descriptions
    $\mathsf{Parent} \sqcap \mathsf{Man}$
    $\mathsf{Human} \sqcap \exists \mathsf{hasChild.Human} \sqcap \mathsf{Man}$
    the similarity value of each of them w.r.t. a third concept i.e. $\mathsf{Parent} \sqcap \mathsf{Man} \sqcap \exists \mathsf{hasChild.(Human} \sqcap \neg \mathsf{Man)}$ is different *because they are written in different ways*

Intentional-Based Similarity Measures 3/3

- Another approach consists in *measuring concept dissimilarities as vector distances* in high dimensional spaces *[Hu et al.'06]*
  - Concepts C and D are unfolded, so that only primitive concept and role names appear
  - each concept is represented as a feature vector where each feature is a primitive concept or role and its value is the number of occurrences in the unfolded concept description
- **This measure fails the soundness criterion**
  - given ServiceFraLon and ServiceCgnLon, the unfolding does not take advantage of the fact that CologneAirport and FrankfurtAirport are German airports *since inclusion axioms are only used*

Behaviors of Similarity Measures

Table: Intentional and extensional based similarity measures and their behavior w.r.t. semantic criteria. "$\surd$" stands for criterion satisfied; "X" stands for criterion not satisfied.

	MEASURE	*Soundness*	*Equiv. soundness*	*Disj. Incompatibility*
EXT.	d'Amato et al.'05 CILC	X	$\surd$	X
	d'Amato et al.'06	$\surd$	$\surd$	X
INT.-BASED	Rada et al.'89	X	$\surd$	$\surd$
	Maedche et al.'02	X	$\surd$	$\surd$
	d'Amato et al.'05 KCAP	$\surd$	X	X
	Janowicz et al.'06-'07	$\surd$	X	$\surd$
	Hu et al.'06	X	$\surd$	$\surd$

Equivalence Soundness Criterion: Formalization

Equivalence Soundness Criterion

Let $(\mathcal{C}, d)$ a metric space where $\mathcal{C}$ is the set of DL concept descriptions expressible in the given language. A dissimilarity measure $d : \mathcal{C} \times \mathcal{C} \to [0, 1]$ obeys the criterion of equivalence soundness iff: $\forall C, D, E \in \mathcal{C} : D \equiv E \Rightarrow d(C, D) = d(C, E)$.

- **It can be proved that**
  If the triangle inequality holds for a given dissimilarity measure d then it satisfies the equivalence soundness criterion

Monotonicity Criterion: Formalization

Monotonicity Criterion

Let $(\mathcal{C}, d)$ a metric space, $\mathcal{C}$ set of DL concept descriptions. A dissimilarity measure $d : \mathcal{C} \times \mathcal{C} \to [0, 1]$ obeys the monotonicity criterion iff given the concepts $C, D, E, L, U \in \mathcal{C}$ s.t:

1. $C \sqsubseteq L, D \sqsubseteq L, C \sqsubseteq U, D \sqsubseteq U$,
2. $E \sqsubseteq U$, and $E \not\sqsubseteq L$
3. $\nexists H \in \mathcal{C}$ s.t. $C \sqsubseteq H \wedge E \sqsubseteq H \wedge D \not\sqsubseteq H$

imply that $d(C, D) \leq d(C, E)$.

- This criterion asserts that, if given the concepts C, D, E, the concept generalizing C and D is more specific (w.r.t. the subsumption relationship) than the one generalizing C and E, than $d(C, D) \leq d(C, E)$

Strict Monotonicity Criterion: Formalization

Given $(\mathcal{C}, d)$ metric space, $\mathcal{C}$ set of DL concept descriptions. A dissimilarity measure $d : \mathcal{C} \times \mathcal{C} \to [0, 1]$ obeys the soundness and disjointness compatibility expected behaviors iff $\forall C, D, E, L, U \in \mathcal{C}$ s.t:

1. $C \sqsubset L, D \sqsubset L, C \sqsubset U, D \sqsubset U$,
2. $E \sqsubset U$, and $E \not\sqsubset L$
3. $\nexists H \in \mathcal{C}$ s.t. $C \sqsubset H \land E \sqsubset H \land D \not\sqsubset H$

imply that $d(C, D) < d(C, E)$

- Given ServiceCgnLon, ServiceFraLon, ServiceRomeLon $\Rightarrow$ dis(ServiceCgnLon, ServiceFraLon) $<$ dis(ServiceCgnLon, ServiceRomeLon) *is valid although* ServiceCgnLon and ServiceFraLon *do not have common instances*
  - *Strict Monotonicity* allows that also empty extension intersections have a value lower than the maximum

Open Issue

(Strict) Monotonicy Criteria pose an open issue: "**how to compute a concept generalization that is able to take into account both the concept definitions and the TBox?**"

1. *LCS of the considered concepts*. However:
   - *for DLs allowing for concept disjunction*, it is given by the disjunction of the considered concepts ⇒ 1) it does not take into account the TBox of reference; 2) it does not add further information besides of that given by the considered concepts.
   - *if less expressive DLs* (i.e. those do not allow for concept disjunction) *are considered*, it is computed in a structural way
2. *A possible generalization able to satisfy our requirements is the Good Common Subsumer (GCS)*. However:
   - it is defined only for $\mathcal{ALE}(\mathcal{T})$ concept descriptions. *If most expressive DLs are considered the problem remains still open*

The GCS-based Similarity Measure: Rationale

Lesson Learnt: A semantic similarity measure should be defined in a way that is neither structural nor extensional

Two concepts are more similar as much their extensions are similar

- the similarity value is given by the variation of the number of instances in the concept extensions w.r.t. the number of instances in the extension of their common super-concept
  - Common super-concept $\Rightarrow$ the GCS of the concepts

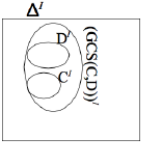


Fig. 1. Concepts C ≡credit-card-payment, D ≡debit-card-payment are similar as the extension of their GCS≡card-payment does not include many other instances besides of those of their extensions.

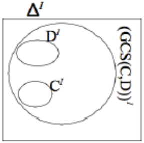


Fig. 2. Concepts C ≡car-transfer, D ≡ debit-card-payment are different as the extension of their GCS≡service includes many other instances besides of those of the extension of C and D.

The GCS-based Similarity Measure: Discussion

The **GCS-based similarity** is a *semantic similarity measure*, namely it **satisfies the semantic criteria**

- given C, D, E s.t. $D \equiv E \Rightarrow^{Def} GCS(C, D) \equiv GCS(C, E) \Rightarrow$ the *equivalence soundness criterion is satisfied*
- Given the Tbox $\mathcal{T} = \{$Human$\sqsubseteq$ Top; Female$\sqsubseteq$ Top; Male$\sqsubseteq$ Top; Table$\sqsubseteq$ Top; Woman$\equiv$ Human $\sqcap$ Female; Man$\equiv$ Human $\sqcap$ Male;$\}$ and the concepts Woman and Man (disjoint in the KB) $\Rightarrow$ $s(\text{Woman}, \text{Man}) \neq 0 \Rightarrow$ the *disjointness compatibility criterion is satisfied*
- By considering the GCS as concept generalization $\Rightarrow$ The *monotonicity criterion is* straightforwardly *satisfied*; indeed
  - $s(\text{ServiceFraLon}, \text{ServiceCgnLon}) > s(\text{ServiceCgnLon}, \text{Service})$
- The GCS-based similarity measure can be used for assessing individual similarity by first computing the *MSCs*

Conclusions

- A set of semantic (dis-)similarity measures for DLs has been presented
  - Able to assess (dis-)similarity between complex concepts, individuals and concept/individual
- The attended behaviors of a similarity measure for ontological knowledge have been analyzed
  - The notions of *(equivalence) soundness* and *disjointness compatibility* have been introduced
- Most of the current measures do not fully satisfy these attended behaviors
- Defined a set of criteria (*equivalence soundness*, *(strict) monotonicity*) that a measure needs to fulfill to be compliant with the attended behaviors
- A new semantic similarity measure satisfying the "semantic" criteria have been introduced

The End

That's all!

Claudia d'Amato

claudia.damato@di.uniba.it