A Similarity Measure for the $\mathcal{ALN}$ Description Logic

Nicola Fanizzi, Claudia d'Amato

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

CILC 2006 ◇ Bari

Contents

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 by mean ontological knowledge
- *Problem* $\longrightarrow$ to define a similarity/dissimilarity measure applicable to ontology languages

Why $\mathcal{ALN}$ Logic

- Knowledge representation by mean Description Logic ($\mathcal{ALN}$)
- Description Logic is the counterpart framework 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	$\bot$	$\emptyset$
primitive concept	A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
primitive negation	$\neg A$	$\Delta^{\mathcal{I}} \setminus A^{\mathcal{I}}$
concept conjunction	$C_1 \sqcap C_2$	$C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$
universal restriction	$\forall R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y \in \Delta^{\mathcal{I}}((x, y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}})\}$
at-most restriction	$\leq n.R$	$\{x \in \Delta^{\mathcal{I}} \mid \lvert\{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R^{\mathcal{I}}\}\rvert \leq n\}$
at-least restriction	$\geq n.R$	$\{x \in \Delta^{\mathcal{I}} \mid \lvert\{y \in \Delta^{\mathcal{I}} \mid (x, y) \in R^{\mathcal{I}}\}\rvert \geq n\}$

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

Instances of concept definitions:
$\mathsf{Single} \equiv \mathsf{Person} \sqcap \leq 0.\mathsf{isMarriedTo}$
$\mathsf{Polygamist} \equiv \mathsf{Person} \sqcap \forall \mathsf{isMarriedTo}.\mathsf{Person} \sqcap \geq 2.\mathsf{isMarriedTo}$
$\mathsf{Bigamist} \equiv \mathsf{Person} \sqcap \forall \mathsf{isMarriedTo}.\mathsf{Person} \sqcap = 2.\mathsf{isMarriedTo}$
$\mathsf{MalePolygamist} \equiv \mathsf{Male} \sqcap \mathsf{Person} \sqcap \forall \mathsf{isMarriedTo}.\mathsf{Person} \sqcap \geq 2.\mathsf{isMarriedTo}$

The following are instances of simple assertions:
$\mathsf{Male}(\mathsf{Bob})$, $\mathsf{Person}(\mathsf{Mary})$, $\mathsf{Single}(\mathsf{Jhon})$, $\mathsf{isMarriedTo}(\mathsf{Bob}, \mathsf{Mary})$

It is easy to see that the following relationship holds:
$\mathsf{Poligamist} \sqsupseteq \mathsf{MalePolygamist}$.

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 $\mathcal{A} \models D(a)$.

Normal Form

C is in $\mathcal{ALN}$ *normal form* iff $C \equiv \bot$ or $C \equiv \top$ or if

$$C = \bigsqcap_{P \in \mathrm{prim}(C)} P \sqcap \bigsqcap_{R \in N_R} (\forall R.C_R \sqcap \geq n.R \sqcap \leq m.R)$$

where:
$C_R = \mathrm{val}_R(C)$, $n = \mathrm{min}_R(C)$ and $m = \mathrm{max}_R(C)$

- $\mathrm{prim}(C)$ 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. $\mathrm{val}_R(C) = \top$);
- $\mathrm{min}_R(C)$ $= \max\{n \in N \mid C \sqsubseteq (\geq n.R)\}$ (always finite number);
- $\mathrm{max}_R(C)$ $= \min\{n \in N \mid C \sqsubseteq (\leq n.R)\}$ (if unlimited $\mathrm{max}_R(C) = \infty$)

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

A Similarity Measure for $\mathcal{ALN}$: Definition / I

$\mathcal{L} = \mathcal{ALN}/_{\equiv}$ the set of all concepts in $\mathcal{ALN}$ normal form
$\mathcal{I}$ canonical interpretation of $\mathcal{A}$ A-Box $s : \mathcal{L} \times \mathcal{L} \mapsto [0,1]$ defined $\forall C, D \in \mathcal{L}$:

$$
\begin{aligned}
s(C,D) \quad := \quad & \lambda[s_P(\mathrm{prim}(C), \mathrm{prim}(D)) \; + \\
+ \quad & \frac{1}{|N_R|} \sum_{R \in N_R} s(\mathrm{val}_R(C), \mathrm{val}_R(D)) \; + \frac{1}{|N_R|} \cdot \\
\cdot \quad & \sum_{R \in N_R} s_N((\min_R(C), \max_R(C)), (\min_R(D), \max_R(D)))]
\end{aligned}
$$

where $\lambda \in]0,1]$ (let $\lambda = 1/3$),

A Similarity Measure for $\mathcal{ALN}$: Defintion / II

$$s_P(\mathrm{prim}(C), \mathrm{prim}(D)) := \frac{|\bigcap_{P_C \in \mathrm{prim}(C)} P_C^{\mathcal{I}} \cap \bigcap_{Q_D \in \mathrm{prim}(D)} Q_D^{\mathcal{I}}|}{|\bigcap_{P_C \in \mathrm{prim}(C)} P_C^{\mathcal{I}} \cup \bigcap_{Q_D \in \mathrm{prim}(D)} Q_D^{\mathcal{I}}|}$$

$$s_N((m_C, M_C), (m_D, M_D)) := \frac{\min(M_C, M_D) - \max(m_C, m_D) + 1}{\max(M_C, M_D) - \min(m_C, m_D) + 1}$$

$$s_N((m_C, M_C), (m_D, M_D)) := 0 \text{ if } \min(M_C, M_D) > \max(m_C, m_D)$$

Similarity Measure: example...

Let $\mathcal{A}$ be the considered ABox

Person(Meg), ¬Male(Meg), hasChild(Meg,Bob), hasChild(Meg,Pat),
Person(Bob), Male(Bob), hasChild(Bob,Ann),
Person(Pat), Male(Pat), hasChild(Pat,Gwen),
Person(Gwen), ¬Male(Gwen),
Person(Ann), ¬Male(Ann), hasChild(Ann,Sue), marriedTo(Ann,Tom),
Person(Sue), ¬Male(Sue),
Person(Tom), Male(Tom)

and let C and D be two descriptions in $\mathcal{ALN}$ normal form:

$$C \equiv \text{Person} \sqcap \forall\text{marriedTo.Person} \sqcap \leq 1.\text{hasChild}$$
$$D \equiv \text{Male} \sqcap \forall\text{marriedTo.}(\text{Person} \sqcap \neg\text{Male}) \sqcap \leq 2.\text{hasChild}$$

...Similarity Measure: example...

In order to compute $s(C, D)$ let us consider:

- Let be $\lambda := \frac{1}{3}$
- $N_R = \{\text{hasChild}, \text{marriedTo}\} \rightarrow |N_R| = 2$

$$\begin{aligned} s(C, D) \quad &:= \quad \frac{1}{3} \left[s_P(\text{prim}(C), \text{prim}(D)) + \frac{1}{2} \sum_{R \in N_R} s(\text{val}_R(C), \text{val}_R(D)) + \right. \\ &+ \quad \left. \frac{1}{2} \sum_{R \in N_R} s_N((\text{min}_R(C), \text{max}_R(C)), (\text{min}_R(D), \text{max}_R(D))) \right] \end{aligned}$$

...Similarity Measure: example...

In order to compute s_P let us note that:

- prim(C) = Person
- prim(D) = Male

$s_P(\{\text{Person}\}, \{\text{Male}\}) =$

$$= \frac{|\{\text{Meg,Bob,Pat,Gwen,Ann,Sue,Tom}\} \cap \{\text{Bob,Pat,Tom}\}|}{|\{\text{Meg,Bob,Pat,Gwen,Ann,Sue,Tom}\} \cup \{\text{Bob,Pat,Tom}\}|} =$$

$$= \frac{|\{\text{Bob,Pat,Tom}\}|}{|\{\text{Meg,Bob,Pat,Gwen,Ann,Sue,Tom}\}|} = 3/7$$

...Similarity Measure: example...

To compute s for value restrictions, it is important to note that

- $N_R = \{\mathsf{hasChild}, \mathsf{marriedTo}\}$
- $val_{marriedTo}(C) = \mathsf{Person}$ and $val_{hasChild}(C) = \top$
- $val_{marriedTo}(D) = \mathsf{Person} \sqcap \neg\mathsf{Male}$ and $val_{hasChild}(D) = \top$

$$s(\mathsf{Person}, \mathsf{Person} \sqcap \neg\mathsf{Male}) + s(\top, \top) =$$

$$= \frac{1}{3} \cdot (s_P(\mathsf{Person}, \mathsf{Person} \sqcap \neg\mathsf{Male}) + \frac{1}{2} \cdot (1 + 1) + \frac{1}{2} \cdot (1 + 1)) +$$

$$+ \frac{1}{3} \cdot (1 + 1 + 1) = \frac{1}{3} \cdot (\frac{4}{7} + 1 + 1) + 1 = \frac{13}{7}$$

...Similarity Measure: example...

To compute s for number restrictions it is important to note that

- $N_R = \{\text{hasChild}, \text{marriedTo}\}$
- $min_{marriedTo}(C) = 0; \quad max_{marriedTo}(C) = |\Delta| + 1 = 7 + 1 = 8$
  $min_{hasChild}(C) = 0; \quad max_{hasChild}(C) = 1$
- $min_{marriedTo}(D) = 0; \quad max_{marriedTo}(D) = |\Delta| + 1 = 7 + 1 = 8$
  $min_{hasChild}(D) = 0; \quad max_{hasChild}(D) = 2$
- $\min(M_C, M_D) > \max(m_C, m_D)$

$$s_N((m_{\text{hasChild}}(C), M_{\text{hasChild}}(C)), (m_{\text{hasChild}}(D), M_{\text{hasChild}}(D))) +$$

$$+ s_N((m_{\text{marriedTo}}(C), M_{\text{marriedTo}}(C)), (m_{\text{marriedTo}}(D), M_{\text{marriedTo}}(D))) =$$

$$= \frac{min(M_{\text{hasChild}}(C), M_{\text{hasChild}}(D)) - max(m_{\text{hasChild}}(C), m_{\text{hasChild}}(D)) + 1}{max(M_{\text{hasChild}}(C), M_{\text{hasChild}}(D) - min(m_{\text{hasChild}}(C), m_{\text{hasChild}}(D)) + 1)} + 1 =$$

$$= \frac{min(1,2) - max(0,0) + 1}{max(1,2) - min(0,0) + 1)} + 1 = \frac{2}{3} + 1 = \frac{5}{3}$$

Measure Involving 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:

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

Discussion

- The similarity value is mainly determined as the amount of overlapping sets of individuals that are extension of the concepts involved, considering also their sub-concepts
  - the influence of sub-concepts in determining similarity value decreases w.r.t. their nesting level
- The similarity measure is defined recursively
  - its complexity mainly depends on the complexity of the *Instance checking* operator
    - limited to primitive concepts
    - it can be pre-compiled

Conclusions

- The presented function s is a *Similalrity Measure*
  - it is definite positive, symmetric, and has maximal value only when the concepts are equivalent
- The presented Similarity Measure is based on the A-Box *semantics* and it is applicable also to couples of individuals, or a concepts and an individual
- s is defined using the set theory and reasoning operators
  - **It uses a numerical approach but is applied on symbolic representations**

Further Developments

- *Testing* the Similarity Measure using some *classification* and *clustering algorithms*
- (Ongoing) Extension of the measure for more 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

The End

Thanks

For Your Attention