

A Dissimilarity Measure for Concept Descriptions in Expressive ontology languages

Claudia d'Amato, Nicola Fanizzi, Floriana Esposito

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

KCAP Workshop @ KCAP 2005 ◊ Banff

Contents

- 1 Introduction & Motivation
 - Motivations
 - Objectives
- 2 The Reference Representation Language
 - Knowledge Base & Subsumption
 - Normal Form
- 3 A Dissimilarity Measure for \mathcal{ALC}
 - Overlap Function
 - Dissimilarity Measure
 - Meaning of Dissimilarity Measure
 - Dissimilarity Measure: example
 - Measures Involving Individuals
 - Complexity
- 4 Conclusions and Further Developments
 - Conclusion

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* → to make clusters of concepts or individuals asserted by mean ontological knowledge
- *Problem* → to define a similarity/dissimilarity measure applicable to ontology languages

Why \mathcal{ALC} Logic

- Knowledge representation by mean Description Logic (\mathcal{ALC})
- Description Logic is the theoretical framework of OWL language
 - standard de facto for the knowledge representation in the Semantic Web

The Representation Language

- Primitive *concepts* $N_C = \{C, D, \dots\}$: subsets of a domain
- Primitive *roles* $N_R = \{R, S, \dots\}$: 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	\perp	\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 individuals 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 $MSC_{\mathcal{A}}(a)$, such that $\mathcal{A} \models C(a)$ and $C \sqsubseteq D$, $\forall D$ such that $\mathcal{A} \models D(a)$.

Normal Form

D is in \mathcal{ALC} *normal form* iff $D \equiv \perp$ or $D \equiv \top$ or if
 $D = D_1 \sqcup \dots \sqcup D_n$ ($\forall i = 1, \dots, n, D_i \not\equiv \perp$) 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 $C_1 \sqcap \dots \sqcap C_n$ 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

$\mathcal{L} = \mathcal{ALC}/\equiv$ 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 $\mathcal{L} \equiv$

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

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

Overlap Function / II

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

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

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

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

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:

$$d(C, D) := \left\{ \begin{array}{l} 0 \\ 1 \\ \frac{1}{f(C, D)} \end{array} \right. \left| \begin{array}{l} f(C, D) = \infty \\ f(C, D) = 0 \\ \textit{otherwise} \end{array} \right.$$

where f is the function overlapping

Meaning of Dissimilarity Measure

- If $C \equiv D$ (namely $C \sqsubseteq D$ e $D \sqsubseteq C$) (semantic equivalence) $d(C, D) = 0$, rather d assigns the minimum value
- If $C \sqcap D \equiv \perp$ 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.

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.

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

...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 conjunctive descriptions inside existential restriction 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 simplified.

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

Measures Involving Individuals

Let c and d two individuals in a given A-Box.

We can consider $C^* = MSC^*(c)$ and $D^* = MSC^*(d)$:

$$d(c, d) := d(C^*, D^*) = d(MSC^*(c), MSC^*(d))$$

Analogously:

$$\forall a : d(c, D) := d(MSC^*(c), D)$$

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 $Cmpl(d) = 2 \cdot Cmpl(\sqsupset)$
- 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: $Cmpl(d) = nm \cdot Cmpl(f_{\sqcap}) =$
 $nm \cdot [Cmpl(f_P) + Cmpl(f_V) + Cmpl(f_{\sqsupset})]$

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 $N = |\text{ex}_R(C_i)|$ and $M = |\text{ex}_R(D_j)|$
- 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 and Further Developments

Conclusions...

- Presentation of a semantic dissimilarity measure d suitable for defining dissimilarity value between concepts, individuals and concept and individual
 - d **dissimilarity measure** besides it is definite positive, symmetric, and has minimal value only when the concepts are equal (in the sense of semantic equivalence)
- d can be applied to knowledge bases expressed in OWL and \mathcal{ALC} DL

Conclusions and Further Developments

...Conclusions

- Complexity of d depends from the complexity of the instance check operator and subsumption operator for \mathcal{ALC} DL
- d is defined using the set theory and reasoning operators
 - **It uses a numerical approach but is applied on symbolic representations**

Conclusions and Further Developments

Future Work...

- d is applicable for both the concepts to individual dissimilarity and between individuals one, it is suitable for agglomerative clustering and for divisional clustering too.
- Development of a clustering algorithm that uses the proposed dissimilarity measure in order to apply it for clustering services (described in OWL-S)
- Extension of d for most expressive DL such as \mathcal{ALCN}
- d is defined by using f function which is a recursive function that measure overlapping at every nested level of the concept descriptions. For the future we would introduce a reduction factor in order to minimize the weight of the overlap value of deep nested element of the concept descriptions.

The End

Thank you.
For Attention