Concept Dissimilarity with Triangle Inequality

Felix Distel
Institute of Theoretical Computer Science
Faculty of Computer Science
TU Dresden, Germany
felix@tcs.inf.tu-dresden.de

Jamal Atif
Université Paris Sud, LRI
TAO, Orsay, France
jamal.atif@lri.fr

Isabelle Bloch
Institut Mines Télécom
Télécom ParisTech
CNRS LTCI, Paris, France
isabelle.bloch@telecom-paristech.fr

Abstract

Several researchers have developed properties that ensure compatibility of a concept similarity or dissimilarity measure with the formal semantics of Description Logics. While these authors have highlighted the relevance of the triangle inequality, none of their proposed dissimilarity measures satisfy it. In this work we present a theoretical framework for dissimilarity measures with this property. Our approach is based on concept relaxations, operators that perform stepwise generalizations on concepts. We prove that from any relaxation we can derive a dissimilarity measure that satisfies a number or properties that are important when comparing concepts.

1 Introduction

By nature description logics are well equipped for representing precise knowledge in a formal manner. As ontologies and description logics (DL) reach out to a broader audience some limitations become evident. In practice, it often occurs that two concepts have similar meanings, but no precise logical relationship can be established. Similarity measures, or dually dissimilarity measures, are attempts to quantify the differences between concepts. They are crucial in areas such as information retrieval in ontologies, ontology alignment, inductive logic programming and for some tasks in non-monotonic reasoning such as model-based revision or aggregation.

In a DL setting similarity can be defined between individuals, concepts, or even ontologies. In this work we focus exclusively on concept similarity. A large number of concept similarity measures has been developed, most of which are tailored to the specific needs of a particular field, such as biomedicine (Pesquita et al. 2009), or geospatial reasoning (Janowicz and Wilkes 2009). These approaches can be classified according to various criteria, such as the ones given in (Borgida, Walsh, and Hirsh 2005). Initially, the quality of similarity measures has only been measured in terms of empirical evaluations. Increasingly, researchers are starting to look at theoretical properties that ensure compatibility of a similarity measure with the formal semantics of description logics. Works such as (d'Amato, Staab, and Fanizzi 2008) and (Lehmann and Turhan 2012) list amongst others the properties of a metric, in particular the triangle inequality, as well as soundness with respect to equivalence and subsumption.

The triangle inequality has been somewhat controversial and in some applications such as (Janowicz and Wilkes 2009) it is not needed. In other applications such as metric-based conceptual clustering and distance-based optimization methods it is crucial (Fayyad et al. 1996). Unfortunately, even the measures presented in (Lehmann and Turhan 2012) and (d'Amato, Staab, and Fanizzi 2008) with their otherwise good theoretical properties do not satisfy the triangle inequality. Our results aim to provide knowledge engineers from these fields with an adequate measure.

In this work, we give a general framework that can be used to construct concept dissimilarity measures with good theoretical properties, including the triangle inequality. The framework is based on concept relaxations, operators that can be used to successively make concepts more general. A directed distance between two concepts C and D can then be defined as the number of times D needs to be relaxed before it subsumes C. We show that the maximum of the two directed distances yields a good dissimilarity measure. Finally, we demonstrate ways to instantiate the framework.

2 Preliminaries

2.1 Description Logics

Description logics are a family of knowledge representation formalisms (Baader 2003). Every description logic $\mathcal{L}$ provides a set of *concepts* $\mathsf{C}(\mathcal{L})$. Concept descriptions are recursively obtained from a set of *concept names* $\mathcal{N}_C$ and a set of *role names* $\mathcal{N}_R$ using concept constructors such as conjunction $\sqcap$, existential restrictions $\exists$ or the top concept $\top$, among others. The description logic that only allows for these three constructors is called $\mathcal{EL}$. In $\mathcal{EL}$, concepts can be visualized as *$\mathcal{EL}$-Description Trees* where node labels represent concept names and edges represent roles. For example the tree in Figure 1 represents the concept

$$\mathsf{Person} \sqcap \exists \mathsf{c}.\mathsf{Male} \sqcap \exists \mathsf{c}.\exists \mathsf{c}.\mathsf{Female}. \qquad (1)$$

Using a model based semantics one can define a generality relation on concepts. If D is more general than C, in other words D subsumes C, we write $C \sqsubseteq D$. We say that C and D are *equivalent* (denoted by $C \equiv D$) if both $C \sqsubseteq D$ and $D \sqsubseteq C$ hold.

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

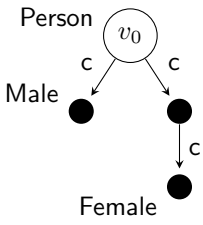

Figure 1: $\mathcal{EL}$-Description Tree for (1)

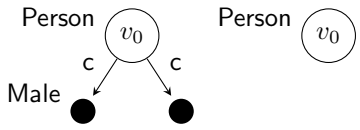

Figure 2: Consecutive applications of ρ_{depth} to (1)

In description logics, axioms are typically stored in ontologies, which can be divided into TBoxes and ABoxes. We define our framework in the absence of background ontologies.

2.2 Similarity and Dissimilarity on Concepts

When similarity measures were first investigated within the DL community, researchers mainly focused on adaptations of existing measures from other fields (cf. (Borgida, Walsh, and Hirsh 2005) for a survey). The quality of these measures was mainly examined in an empirical way, showing that they perform well in a given setting, but providing little transferable insight. It was only in (d'Amato, Staab, and Fanizzi 2008) that qualitative criteria were developed, based on the ones given by (Bock and Diday 2000). The following definition is slightly adapted to dissimilarity between concepts.

Definition 1 (Dissimilarity (Bock and Diday 2000))**.** *Let $\mathcal{L}$ be a DL language. A function $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called a* dissimilarity measure *if it satisfies the following properties for all $C, D \in \mathsf{C}(\mathcal{L})$.*

- positiveness: $d(C, D) \geq 0$
- reflexivity: $d(C, C) = 0$*, and*
- symmetry: $d(C, D) = d(D, C)$.

These properties can be expected to hold for any dissimilarity measure. In a description logics context it should also be compatible with the semantics of the logic. To ensure this, (d'Amato, Staab, and Fanizzi 2008) and more recently (Lehmann and Turhan 2012) have introduced an extended set of properties. These properties are originally stated for similarity measures, here we present their equivalents for dissimilarity measures.

Definition 2. *A dissimilarity measure $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called*

- equivalence closed *if $d(C, D) = 0 \implies C \equiv D$,*
- equivalence sound *if $D \equiv E \implies d(C, D) = d(C, E)$,*
- subsumption preserving *if $C \sqsubseteq D \sqsubseteq E \implies d(C, D) \leq d(C, E)$*
- reverse subsumption preserving *if $C \sqsubseteq D \sqsubseteq E \implies d(D, E) \leq d(C, E)$*
- structurally dependent *if for all sequences $(C_n)_n$ of atoms with $C_i \not\sqsubseteq C_j$ for all $i, j \in \mathbb{N}$, $i \neq j$ the concepts*

$$D_n = \prod_{i \leq n} C_i \sqcap D, E_n = \prod_{i \leq n} C_i \sqcap E$$

satisfy $\lim_{n\to\infty} d(D_n, E_n) = 0$ for all $C, D, E \in \mathsf{C}(\mathcal{L})$.
- *We say that d fulfills the* triangle inequality *if $d(C, E) \leq d(C, D) + d(D, E)$ for all $C, D, E \in \mathsf{C}(\mathcal{L})$.*

A dissimilarity d is a *metric* if it satisfies the triangle inequality and is additionally strict, i.e. $d(x, y) = 0$ implies $x = y$.

A desirable feature of a good dissimilarity measure is that concepts with more common features should be less dissimilar than concepts with few common features. Structural dependence is a formalization of this idea. Another attempt has been introduced in (d'Amato, Staab, and Fanizzi 2008), where it is formalized in terms of common subsumers.

Definition 3 ((Strict) Monotonicity)**.** *A dissimilarity measure $d\colon \mathsf{C}(\mathcal{L}) \times \mathsf{C}(\mathcal{L}) \to \mathbb{R}$ is called* (strictly) monotone *if for all $C, D, E \in \mathsf{C}(\mathcal{L})$ that satisfy*

- *every common subsumer of C and E also subsumes D,*
- *there is a common subsumer of C and D that does not subsume E,*

it holds that $d(C, D) \leq d(C, E)$, respectively $d(C, D) < d(C, E)$.

3 General Framework

We provide a general framework for defining dissimilarity measures. All dissimilarity measures obtained within this framework have all properties from Section 2.2, except monotonicity and structural dependence. The framework is based on *concept relaxation operators*, operators that allow a stepwise generalization of concepts.

Definition 4 (Relaxation)**.** *A* (concept) relaxation *is an operator $\rho\colon \mathsf{C}(\mathcal{L}) \to \mathsf{C}(\mathcal{L})$ that satisfies the following three properties for all $C, D \in \mathcal{L}$.*

1. *ρ is* non-decreasing*, i.e. $C \sqsubseteq D$ implies $\rho(C) \sqsubseteq \rho(D)$,*
2. *ρ is* extensive*, i.e. $C \sqsubseteq \rho(C)$, and*
3. *ρ is* exhaustive*, i.e. $\exists k \in \mathbb{N}_0\colon \top \sqsubseteq \rho^k(C)$, where ρ^k denotes ρ applied k times, and ρ^0 is the identity.*

Examples for relaxation operators that can be used to instantiate the framework are presented in Section 4.

A dissimilarity measure that is equivalence sound and closed should have the value $d(C, D) = 0$ if and only if $C \equiv D$, i.e. iff $C \sqsubseteq D$ and $D \sqsubseteq C$. Like (Lehmann and Turhan 2012) and (Suntisrivaraporn 2013) we first introduce directed measures d^d_ρ that capture how "far" D is from being a subsumer of C, and vice versa. If both $C \sqsubseteq D$ and $D \sqsubseteq C$ hold, then both directed measures will be 0. The directed measure $d^d_\rho(C, D)$ counts how often we need to successively relax D to reach a subsumer of C. If we think of concepts in terms of sets of individuals, then the intuition behind successive relaxations can be visualized as in Figure 3.

Definition 5 (Directed measure)**.** *Let ρ be a relaxation on $\mathsf{C}(\mathcal{L})$. For $C, D \in \mathsf{C}(\mathcal{L})$ the* directed measure *$d^d_\rho(C, D)$ is defined as*

$$d^d_\rho(C, D) = \min\{k \in \mathbb{N}_0 \mid C \sqsubseteq \rho^k(D)\},$$

where ρ^k denotes ρ applied k times, and ρ^0 is the identity.

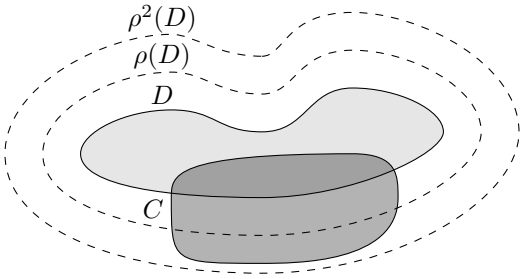

Figure 3: D needs to be relaxed twice before it subsumes C, i.e. $d^d_\rho(C, D) = 2$

The directed measure is always finite because ρ is exhaustive. We can then define the *relaxation dissimilarity* based on a relaxation operator simply as the maximum of the two directed measures.

Definition 6 (Relaxation Dissimilarity)**.** *Let $\rho\colon \mathcal{L} \to \mathcal{L}$ be a relaxation on $\mathsf{C}(\mathcal{L})$. For two concepts C and D the* relaxation dissimilarity $d_\rho(C, D)$ *is defined as*

$$d_\rho(C, D) = \max\{d^d_\rho(C, D), d^d_\rho(D, C)\}.$$

Theorem 1. *For every relaxation ρ the operator d_ρ is a dissimilarity measure, that is equivalence sound, equivalence closed, subsumption preserving and reverse subsumption preserving, and satisfies the triangle inequality.*

Proof. Positiveness, reflexivity and symmetry follow trivially from Definitions 5 and 6, and therefore d_ρ is a dissimilarity measure.

We have the following chain of equivalences: $C \equiv D$, iff $C \sqsubseteq D$ and $D \sqsubseteq C$, iff $C \sqsubseteq \rho^0(D)$ and $D \sqsubseteq \rho^0(C)$, iff $d^d_\rho(C, D) = d^d_\rho(D, C) = 0$, iff $d_\rho(C, D) = 0$. Thus d_ρ is both equivalence sound and equivalence closed.

To prove the triangle inequality, let C, D, E be concept descriptions and let $d_\rho(C, D) = d_1$, $d_\rho(D, E) = d_2$. Then in particular, $d^d_\rho(C, D) \leq d_1$ and thus $C \sqsubseteq \rho^{d_1}(D)$ by extensivity. Similarly, we obtain $D \sqsubseteq \rho^{d_2}(E)$. Since relaxations are non-decreasing we obtain from the latter

$$\rho^{d_1}(D) \sqsubseteq \rho^{d_1+d_2}(E)$$

and therefore $C \sqsubseteq \rho^{d_1+d_2}(E)$, i.e. $d^d_\rho(C, E) \leq d_1 + d_2$. Analogously, it can be shown that $d^d_\rho(E, C) \leq d_1 + d_2$ and thus $d_\rho(C, E) \leq d_1 + d_2 = d_\rho(C, D) + d_\rho(D, E)$.

To show subsumption preservation let $C \sqsubseteq D \sqsubseteq E$ with $d_\rho(C, E) = d$. Then in particular, $E \sqsubseteq \rho^d(C)$ and thus also $D \sqsubseteq \rho^d(C)$. On the other hand, $C \sqsubseteq \rho^0(D) \sqsubseteq \rho^d(D)$ by extensivity, which yields $d_\rho(C, D) \leq k = d_\rho(C, E)$, which proves subsumption preservation. □

A comparison of these properties with those of some existing measures is provided in Table 1.

4 Instantiations

Theorem 1 shows that our framework produces dissimilarity measures with good theoretical properties. The framework can be instantiated with any relaxation operator. The behaviour of the resulting measures can vary greatly depending on the relaxation. This is demonstrated by the following examples. A trivial relaxation in any description logic is the operator $\rho_\top$ that maps every concept to $\top$. It results in a very coarse dissimilarity measure $\rho_\top$ that is 0 iff the concepts are equivalent and 1 otherwise.

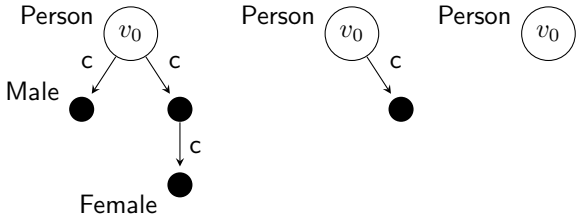

Figure 4: Consecutive application of ρ_{leaves} to (1)

Relaxations from tree operations For the lightweight logic $\mathcal{EL}$ (Baader, Küsters, and Molitor 1999) have proven a close connection between $\mathcal{EL}$ concepts and description trees. Due to this connection any operator that maps description trees to strict subtrees gives rise to a relaxation. One possibility is the operator ρ_{depth} that reduces the role depth of each concept by 1, simply by pruning the description tree (cf. Figure 2). The corresponding dissimilarity ρ_{depth} measures the first depth-level where the concepts differ in the description tree. It gives higher weight to features at a smaller depth. For example, if we compare the concepts $\mathsf{F} := \mathsf{Male} \sqcap \exists\mathsf{hasChild}.\top$ and $\mathsf{HoJ} := \mathsf{Male} \sqcap \exists\mathsf{marriedTo}.(\mathsf{Female} \sqcap \mathsf{Judge})$ to the concept $\exists\mathsf{hasChild}.\top$ the value will be 2 in both cases, since the change occurs at the lowest level, in the concept name Male. This is counterintuitive, since F and $\exists\mathsf{hasChild}.\top$ share more common features than HoJ and $\exists\mathsf{hasChild}.\top$. A slightly better behaviour can be achieved by the relaxation ρ_{leaves} that removes all leaves from a description tree (Figure 4).

Relaxations from distances between models As logics become more expressive, it becomes harder to directly define a relaxation on the concepts. Since the models remain simple labeled graphs, even for complex descriptions, one solution is to identify concepts with the set of their models. Similar to related work from Section 5 one might start with a simple distance between models, e.g. an edit distance, and generalize it to a Hausdorff distance between sets of models. However, since the model space is infinite the Hausdorff distance can often not be computed directly.

A workaround is to use the distance on the model space to define dilations, as used in mathematical morphology (Serra 1982), on sets of models. For some distances, such as a simple tree edit distance, the dilated sets themselves correspond to DL concepts. The operator that maps a concept to the concept corresponding to its dilated set of models can be shown to be a relaxation. All these instantiations will be further studied in our future work.

Table 1: Properties of some (dis-)similarity measures

Measure	Equivalence Sound	Monotone	Equivalence Closed	Subs. Preserving	Rev. Subs. Preserving	Structurally Dependent	Triangle Inequality
(Lehmann and Turhan 2012)	✓	–	✓	✓	✓	✓	–
(d'Amato, Staab, and Fanizzi 2008)	✓	✓	–	✓	✓	–	–
relaxation dissimilarity	✓	–	✓	✓	✓	–	✓

5 Existing Metrics for Other Logics

Outside of description logics several works have proposed metrics between logical objects. Works such as (Nienhuys-Cheng 1998; Ramon and Bruynooghe 1998) exploit the fact that is relatively easy to define a metric on ground expressions in first order logic. They extend these ground distances to sets of atoms, or Herbrand interpretations using constructions such as Hausdorff distances or Manhattan distances.

In some cases it is straightforward to define a distance between two terms if one is a generalization of the other. To obtain a distance between two arbitrary terms one can simply use the sum of the distances to their least general common generalization. In a general form (Birkhoff 1993) has presented this idea as the classical distance in graded lattices. It is used to define a distance between first order literals by (Hutchinson 1997), who then generalizes it to a distance between clauses using the Hausdorff metric. This idea can also be extended to cases where there is no unique minimally general generalization (De Raedt and Ramon 2009).

6 Discussion

In this work, we have presented a framework for dissimilarity measures whith good theoretical properties (cf. Table 1). Our measures satisfy at the same time the properties of a metric, in particular the triangle inequality, and they are compatible with the semantics of description logics, in particular they are equivalence sound. Some hints for instantiations of the proposed framework have been provided, and will be the focus of future work.

The similarity measures that we have presented here are defined for concepts without TBoxes. If the background ontology is an acyclic TBox, they can trivially be adapted by comparing only unfolded concepts. In principle, it is possible to generalize relaxations with respect to general TBoxes, but it is left for future work how to instantiate them.

Acknowledgments. This work was initiated during the stay of Felix Distel at Telecom ParisTech in summer 2012, supported by a grant from this institution. Felix Distel has also been supported by the Collaborative Research Center 912 'Highly Adaptive Energy-Efficient Computing'. The authors would like to thank Anni-Yasmin Turhan for fruitful discussions.

References

Baader, F.; Küsters, R.; and Molitor, R. 1999. Computing least common subsumers in description logics with existential restrictions. In *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 96–101. Morgan-Kaufmann.

Baader, F. 2003. Description Logic terminology. In Baader, F.; Calvanese, D.; McGuinness, D.; Nardi, D.; and Patel-Schneider, P. F., eds., *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press. 485–495.

Birkhoff, G. 1993. *Lattice theory*, volume 25 of *Colloquium publications*. Providence, Rhode Island: American Mathematical Society, 3rd edition.

Bock, H. H., and Diday, E. 2000. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Berlin: Springer.

Borgida, A.; Walsh, T. J.; and Hirsh, H. 2005. Towards measuring similarity in description logics. In *Proc. of the 2005 Int. Workshop on Description Logics (DL)*.

d'Amato, C.; Staab, S.; and Fanizzi, N. 2008. On the influence of description logics ontologies on conceptual similarity. In *Knowledge Engineering: Practice and Patterns*. Springer. 48–63.

De Raedt, L., and Ramon, J. 2009. Deriving distance metrics from generality relations. *Pattern Recognition Letters* 30(3):187–191.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. 1996. Advances in knowledge discovery and data mining.

Hutchinson, A. 1997. Metrics on terms and clauses. In *Proc. of the European Conf. on Machine Learning (ECML)*, 138–145. Springer.

Janowicz, K., and Wilkes, M. 2009. SIM-DLA: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity. In *The Semantic Web: Research and Applications*. Springer. 353–367.

Lehmann, K., and Turhan, A.-Y. 2012. A framework for semantic-based similarity measures for $\mathcal{ELH}$-concepts. In *Proc. of the 13th European Conf. on Logics in Artificial Intelligence (ECAI)*, LNAI, 307–319. Springer Verlag.

Nienhuys-Cheng, S.-H. 1998. Distances and limits on herbrand interpretations. In *Inductive Logic Programming*. Springer. 250–260.

Pesquita, C.; Faria, D.; Falcao, A. O.; Lord, P.; and Couto, F. M. 2009. Semantic similarity in biomedical ontologies. *PLoS computational biology* 5(7):e1000443.

Ramon, J., and Bruynooghe, M. 1998. A framework for defining distances between first-order logic objects. In *Inductive Logic Programming*. Springer. 271–280.

Serra, J. 1982. *Image analysis and mathematical morphology*. London.: Academic Press.

Suntisrivaraporn, B. 2013. A similarity measure for the description logic $\mathcal{EL}$ with unfoldable terminologies. In *Proc. of the 5th Int. Conf. on Intelligent Networking and Collaborative Systems (INCoS)*, 408–413. IEEE.