SEMilarity: Towards a Model-Driven Approach to Similarity

Rudi Araújo and H. Sofia Pinto

INESC-ID
Rua Alves Redol 9, Apartado 13069, 1000-029 Lisboa, Portugal
{rnna,sofia}@algos.inesc-id.pt

Abstract. Enabling the Semantic Web requires solving the semantic heterogeneity problem, for which ontology matching methods have been proposed. These methods rely on similarity measures that are mainly focused on terminological, structural and extensional properties of the ontologies. Semantics rarely play a direct role on the ontology matching process, albeit some algorithms have been proposed. On the other hand, many ontology engineers choose representation languages that have an underlying formal logic, providing well-defined model-theoretic semantics. Since semantics are a key advantage of ontologies, we believe that semantics-based similarity measures are crucial. In this paper, we present a novel approach to semantic similarity.

Key words: ontology matching, semantic similarity

1 Introduction

Given the semi-anarchic organisation of the current World Wide Web, it is unrealistic to expect that the Semantic Web, its envisioned evolution, will not suffer from semantic heterogeneity, which can, if it is not properly tackled, hinder its acceptance and consequently its growth and, in the worst case, preclude its development. Ontology matching and alignment is an area that deals with this problem by establishing relations (usually equivalence and subsumption relations) between elements in different ontologies. According to [1], ontology alignment techniques can be categorised in two major groups: *local*, which focuses on similarities of individual elements and/or their relations to other elements, and *global*, dealing with the whole ontology or parts of it. The local alignment techniques are further classified as *terminological*, *structural*, *extensional* or *semantics*. Terminological methods are twofold: many rely in string-matching techniques, such as sub-string matching, Jaccard Distance, Edit Distance, etc; others use external linguistic resources, such as dictionaries or thesauri. Structural techniques rely on the structure of the elements, and their relations to other elements, recurring, for example, to graph matching techniques. Extensional techniques focus on the extensions (instances) of concepts to assess their likelihood. Finally, semantics-based alignment approaches are aware and make use of the semantics underlying

the representation language, which enables them to resort to deduction services, such as subsumption and consistency checking.

Similarity measures are used to assess the likelihood of elements of ontologies or the ontologies themselves. In this paper we present our preliminary work on defining an ontology similarity measure that is purely based in the semantics of concepts. We should note at this point that we are committing to the notion of semantics as defined by a formal logic system, and not as its pragmatical meaning as approached in [2]. For the target representation language, we chose a Description Logics formalism for mainly three reasons: it is the backbone of the current most prominent ontology representation language for the Semantic Web – the OWL language –, it is the most active family of languages in the community and it provides well-defined model-theoretic semantics. Our target representation language is $\mathcal{ALC}$ without roles. Although this is a rather inexpressive logic, we stress that this work is only preliminary and that we plan to extend it towards more expressive languages. Note that this logic is equivalent to propositional logic, which, inexpressive as it is, can still find application in the real world, since it allows to describe taxonomies (web directories are examples of this). In the following, we assume that a TBox is a set of subsumption and equivalence axioms, that relate atomic and complex concepts. The concepts we are considering are $\bot$, $\top$, A, $C \sqcap D$, $C \sqcup D$, $\neg C$, where A is a concept name, and C and D are concepts. Their semantics are defined as usually [3]. The term *ontology* is often used to refer to a number of different artifacts that may include, for example, a glossary of terms, the conceptual and coded model and the documentation. For simplicity, in this paper we will restrict the notion of *ontology* to an $\mathcal{ALC}$ TBox without roles. In the following, it is assumed that the set of concepts $\mathcal{C}$ contained in an ontology is finite.

The paper is organised as follows: section 2 presents the theoretical underpinning of the work presented here, followed by a toy example demonstrating how it works in practice. The implementation of the algorithm is the subject of section 3. We conducted an experiment with average-sized ontologies, using the proposed similarity measure, described in section 4. Section 5 comprises an evaluation and discussion of the proposed measure. We summarise related work in section 6 and finish the paper with conclusions and future directions in section 7.

2 Theory

Given the set of possible ontologies in the language we are considering, $\mathcal{O}$, our aim is to define a similarity measure $\sigma : \mathcal{O} \times \mathcal{O} \rightarrow [0, 1]$, which is purely based on semantics. This similarity measure is required to take the highest value for equivalent ontologies, i.e. given three ontologies $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$, if $\mathcal{T}_2 \equiv \mathcal{T}_3$:[1]

1. $\sigma(\mathcal{T}_1, \mathcal{T}_2) = \sigma(\mathcal{T}_1, \mathcal{T}_3)$.
2. $\sigma(\mathcal{T}_2, \mathcal{T}_3) = 1$.

[1] Assume that $\mathcal{T}_1 \equiv \mathcal{T}_2$ is equivalent to $\mathcal{T}_1 \models \mathcal{T}_2$ and $\mathcal{T}_2 \models \mathcal{T}_1$.

Our approach starts by considering a simple version of the aimed similarity function, defined as follows:

$$\sigma'(\mathcal{T}_1, \mathcal{T}_2) = \begin{cases} 1 & \text{if } \mathcal{T}_1 \equiv \mathcal{T}_2 \\ 0.5 & \text{else if } \mathcal{T}_1 \models \mathcal{T}_2 \text{ or } \mathcal{T}_2 \models \mathcal{T}_1 \\ 0 & \text{otherwise .} \end{cases} \quad (1)$$

This similarity function is not sufficiently discriminative, which is due to the fact that the definition of the entailment operator requires *every* model of $\mathcal{T}_1$ to be a model of $\mathcal{T}_2$ so that $\mathcal{T}_1 \models \mathcal{T}_2$. What we wish to achieve is a similarity operator that is a function of the *quantity* of models of $\mathcal{T}_1$ and $\mathcal{T}_2$. However, the amount of models of an ontology is usually infinite. Our approach is to consider a kind of Herbrand interpretation to get around this problem, but instead of redefining the whole logic system, as it is done with Herbrand logic to deal with Herbrand models, we choose to define syntactic elements (concepts), rather than semantic ones (interpretations). Consider the following definitions.

Definition 1 (Characteristic Concept). *Let $\mathcal{C}$ be a set of DL concept names. A* characteristic concept *wrt $\mathcal{C}$ is a concept conjunction of the form $C_1 \sqcap \ldots \sqcap C_n$, where C_i is either A or $\neg\mathsf{A}$, with $\mathsf{A} \in \mathcal{C}$, $n = |\mathcal{C}|$, and for every $i \neq j$, $C_i \neq C_j$ and $C_i \neq \neg C_j$. $\zeta(\mathcal{C})$ is the set of all possible characteristic concepts wrt $\mathcal{C}$.*

Definition 2 (Characteristic Disjunction and Axiom). *Let $\mathcal{C}$ be a set of DL concept names and $S \subseteq \zeta(\mathcal{C})$. The* characteristic disjunction *of S, $U(S)$, is the concept $\bigsqcup_{C \in S} C$. The* characteristic axiom *of S, $\theta(U(S))$, is the axiom $\top \sqsubseteq U(S)$.*

Definition 3 (Characteristic Acceptance Set). *Let $\mathcal{T}$ be an ontology containing the set of DL concept names $\mathcal{C}$. The* characteristic acceptance set *of $\mathcal{T}$, written $Z(\mathcal{T})$, is such that $Z(\mathcal{T}) \subseteq \zeta(\mathcal{C})$ and $\mathcal{T} \equiv \theta(Z(\mathcal{T}))$.*

In other words, a characteristic concept wrt a set of concept names is one of the most specific concepts that is possible to build from them. The characteristic disjunction is the concept disjunction of all the characteristic concepts. Finally, the characteristic acceptance set of an ontology is the set of all characteristic concepts consistent in that ontology. We should note that the characteristic disjunction can also be interpreted as a formula in the disjunctive normal form (DNF). Although normal forms are usually very large, we show how to circumvent this in section 3. Note that the elements of the characteristic acceptance set can be thought of as Herbrand models (with an arbitrary constant). Since the similarity measure we present is heavily based on the characteristic acceptance set, the following lemmas must hold.

Lemma 1. *Let $\mathcal{T}$ be a consistent ontology containing the DL concept names $\mathcal{C}$. $Z(\mathcal{T})$ exists and is unique.*

Lemma 2. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be consistent ontologies containing the concepts $\mathcal{C}$.*

i. *$\mathcal{T}_1 \models \mathcal{T}_2$ iff $Z(\mathcal{T}_1) \subseteq Z(\mathcal{T}_2)$;*

ii. $\mathcal{T}_1 \equiv \mathcal{T}_2$ *iff* $Z(\mathcal{T}_1) = Z(\mathcal{T}_2)$.

The proofs of these lemmas can be found in [4]. The acceptance set depends on the number of models of the ontology, since it is the set of most specific consistent concepts in the ontology (i.e., for which there are at least one model). Given these definitions, we are now able to expand the definition of our similarity measure.

Definition 4 (Semantic Similarity). *Let* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *be consistent ontologies containing the concepts* $\mathcal{C}$*. Let* $Z_1 = Z(\mathcal{T}_1)$ *and* $Z_2 = Z(\mathcal{T}_2)$*. The* semantic similarity measure $\sigma : \mathcal{O} \times \mathcal{O} \to [0, 1]$ *is defined as follows:*

$$\sigma(\mathcal{T}_1, \mathcal{T}_2) = 1 - (|Z_1 - Z_2| + |Z_2 - Z_1|)/2^{|\mathcal{C}|} \ . \tag{2}$$

Intuitively, equation 2 measures the accordance of characteristic concepts between both ontologies.

Theorem 1. *Let* $\mathcal{T}_1$*,* $\mathcal{T}_2$ *and* $\mathcal{T}_3$ *be consistent ontologies containing the DL concept names* $\mathcal{C}$*. If* $\mathcal{T}_2 \equiv \mathcal{T}_3$ *then* $\sigma(\mathcal{T}_1, \mathcal{T}_2) = \sigma(\mathcal{T}_1, \mathcal{T}_3)$ *(i.e.,* σ *is purely based on semantics).*

Proof. The result follows immediately from Lemma 2.

Example 1. Consider the following ontologies:

$\mathcal{T}_1$	$\mathcal{T}_2$
$\neg\mathsf{Male} \sqsubseteq \mathsf{Female}$	$\mathsf{Person} \sqsubseteq \mathsf{Male} \sqcup \mathsf{Female}$
$\mathsf{Man} \doteq \mathsf{Person} \sqcap \mathsf{Male}$	$\mathsf{Man} \doteq \mathsf{Person} \sqcap \mathsf{Male}$
$\mathsf{Woman} \doteq \mathsf{Person} \sqcap \mathsf{Female}$	$\mathsf{Female} \doteq \neg\mathsf{Male}$
$\mathsf{MaleCat} \doteq \mathsf{Cat} \sqcap \mathsf{Male}$	$\mathsf{Woman} \doteq \mathsf{Person} \sqcap \neg\mathsf{Man}$
	$\mathsf{MaleCat} \sqsubseteq \mathsf{Cat}$

Although similar, these two ontologies display subtle differences. In particular, Male or Female are necessary in $\mathcal{T}_1$ (each individual has to be either one or the other, *or both*), but in $\mathcal{T}_2$ the definition is stricter: each individual is exclusively one or the other. Furthermore, in $\mathcal{T}_2$ we define Woman as a Person and not a Man. In both ontologies, the concept of Man is defined as the intersection of Person and Male, but in $\mathcal{T}_1$, some members of Man can also be Female. However, in $\mathcal{T}_2$ it is forbidden for a Male to be Female, so it restricts the concept of Man to individuals who are Male and, consequently, not Female. Finally, MaleCat's definition in $\mathcal{T}_2$ is incomplete wrt $\mathcal{T}_1$.

Table 1 shows the $Z(\mathcal{T}_1)$ and $Z(\mathcal{T}_2)$ sets. Each row is a concept name and each column is a characteristic concept, in such a way that if $+$(resp. $-$) is in the intersection of a concept name C and a characteristic concept D, then C appears in D as a positive (resp. negative) literal.

As can be seen from the table, $|Z_1 - Z_2| = |Z_2 - Z_1| = 4$. Equation 2 yields:

$$\sigma(\mathcal{T}_1, \mathcal{T}_2) = 1 - (|Z_1 - Z_2| + |Z_2 - Z_1|)/2^{|\mathcal{C}|} = 1 - (4 + 4)/128 = 93.75\% \ .$$

Table 1. The characteristic acceptance sets for $\mathcal{T}_1$ and $\mathcal{T}_2$.

Concept	$Z_1 - Z_2$	$Z_1 \cap Z_2$	$Z_2 - Z_1$
Man	+ − − +	+ + − − − − − −	− − + −
Male	+ + + +	+ + − − − − + +	− − + +
Person	+ − − +	+ + + + − − − −	− + + −
Cat	+ + − −	+ − − + − + − +	+ + + +
Female	+ + + +	− − + + + + − −	+ + − −
Woman	+ − − +	− − + + − − − −	− + − −
MaleCat	+ + − −	+ − − − − − − +	+ + − −

3 Implementation

A naive implementation of this theory could potentially be very inefficient, since $Z(\mathcal{T})$ grows exponentially in proportion to $|\mathcal{C}|$. However, we only need the size of a sub-set of $Z(\mathcal{T})$. Given that the characteristic disjunction of a sub-set of $Z(\mathcal{T})$ is equivalent to a DNF formula, we can use #SAT, which computes the size of the set.

Given two ontologies $\mathcal{T}_1$ and $\mathcal{T}_2$, our implementation starts by computing the concepts C_1 and C_2 such that $\mathcal{T}_1 \equiv \top \sqsubseteq C_1$ and $\mathcal{T}_2 \equiv \top \sqsubseteq C_2$. The purpose is to count the characteristic concepts that are subsumed by $C_1 \sqcap C_2$ (i.e., that are both in $Z(\mathcal{T}_1)$ and $Z(\mathcal{T}_2)$) and the ones that are subsumed by $\neg C_1 \sqcap \neg C_2$ (i.e., that are neither in $Z(\mathcal{T}_1)$ nor $Z(\mathcal{T}_2)$). To achieve this, we represent the concept $C_1 \sqcap C_2 \sqcup \neg C_1 \sqcap \neg C_2$ as a CNF formula and feed it to a #SAT solver. To transform the concept into CNF we use the Definitional CNF Transformation algorithm (CNF with naming). Note that the following holds:

$$\sigma(\mathcal{T}_1, \mathcal{T}_2) = \mathrm{mc}(\mathrm{cnf}(C_1 \sqcap C_2 \sqcup \neg C_1 \sqcap \neg C_2))/2^{|\mathcal{C}|} \ , \tag{3}$$

where mc is the model count and cnf is the CNF representation of the formula. To perform model counting we use the RELSAT tool [5]. We should note that the computation is not performed exactly as defined in equation 3. We observed that RELSAT performed considerably faster using the following equivalent equation:

$$\sigma(\mathcal{T}_1, \mathcal{T}_2) = \Big((2^{|\mathcal{C}|} - \mathrm{mc}(\mathrm{cnf}(C_1))) - \mathrm{mc}(\mathrm{cnf}(C_2)) + 2 \times \mathrm{mc}(\mathrm{cnf}(C_1 \sqcap C_2)) \Big) / 2^{|\mathcal{C}|} \ .$$

4 Experiment

In this experiment, we were aiming at evaluating our similarity measure against an intuition of similarity. This measure is only applicable to ontologies sharing the same concept names, but the lack of such ontologies thwarts the direct employment of the measure. It is thus necessary to map a set of ontologies in the same domain. Then, the set of concepts involved in the mapping are cropped, so that the ontologies that are to be compared contain the same set of concept names.

As dataset we used three ontologies in the cooking domain. The first one is called OntoChef$_{GS}$ (O_{GS}) and can be seen as a gold standard, as its development was carried out more zealously and by a bigger team than the others [6]. The other two, OntoChef$_1$ and OntoChef$_2$ (O_1 and O_2), were developed by students at an undergraduate course on Knowledge Representation. The selection of these ontologies was based on their correctness and thoroughness. Many contained axioms such as $\mathsf{Preparation} \sqsubseteq \mathsf{Recipe}$, using subsumption incorrectly and were ruled out. The ontologies were required to define at least: recipes, measurements, (kitchen) tools and ingredients/food, so we ruled out the ones that were not sufficiently thorough on (or completely neglected) these topics. Figure 1 shows a section of each ontology. O_1 has 49 concept names, while O_2 has 167 and O_{GS} has 571.

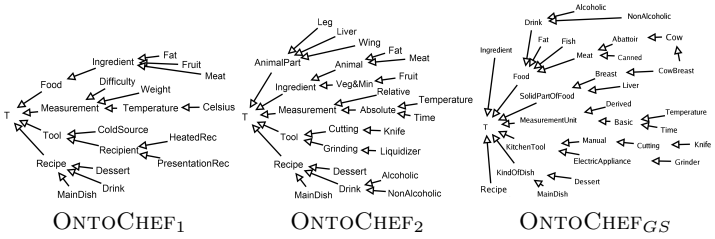

Fig. 1. A relevant part of the ontologies.

Despite obvious dissimilarities, there is an overlap of concepts in the ontologies. For example, both O_1 and O_2 characterise dishes as Recipes, while in O_{GS} these are subsumed by KindOfDish, which intuitively makes more sense. Also, in O_{GS}, Salad is subsumed by Starters, but in O_1 and O_2 they are at the same level as the other kinds of dishes, which shows that even the gold standard can be (and usually is) less than perfect, since salads are not necessarily starters.

The results of applying the similarity measure are as follows:

$$\sigma(O_1, O_2) = 98.1134\%,\ \sigma(O_1, O_{GS}) = 94.8180\%,\ \sigma(O_2, O_{GS}) = 94.0139\%$$

5 Evaluation and Discussion

Although it is not clear from figure 1 that these ontologies are as similar as assessed by the proposed measure, their cropped sections display many similarities. Thus, we can say that the measure is on a par with our intuition of similarity. We can also observe that when a *set* of values is available, comparing the different values is a reasonable way to establish which ontologies are more or less similar to an ontology. In the previous section we observed that the two ontologies built by the undergraduate students were more similar, which is an intuitive outcome. This is mostly due to the fact that each kind of dish is considered as a Recipe in O_1 and O_2, and also that the kinds of Ingredient in these ontologies are considered as Food in O_{GS}. Furthermore, the similarity between O_1 and O_{GS} is slightly higher than the similarity between O_2 and O_{GS}. This

happens mainly because Cup, TeaSpoon and SoupSpoon are represented in O_1 as volume measurement units and in O_2 they are tools.

A possible use-case we envision for our similarity measure is the automatic assessment of a learnt ontology against a gold standard, assuming the learnt ontology has the same concepts as the gold standard, or a sub-set of them. Also, our measure could be used in an ontology merging system that would search an ontology library for similar ontologies and propose extending the source ontology with axioms and concepts from the most similar ontologies.

In [7], the authors present a set of reasonable criteria for assessing the quality of a similarity measure. It can be shown that our measure respects the *proportional error effect* and the *usage of interval* criteria. It is also worth mentioning that the measure is, indeed, a similarity measure as it is usually defined (e.g. [8]).

Although our implementation is based on #SAT, which is NP-Hard, the use of heuristics boost the efficiency of the #SAT solvers, and can deliver results for ontologies containing more than 500 concepts, in less than 10 seconds. We consider this to be acceptable.

6 Related Work

Some alignment algorithms and tools have been developed, many of which are described in [1]. In this survey it is mentioned that only 4 out of the 21 systems analysed rely directly on semantic properties of the ontologies: S-Match [9], Buster [10], Chimarae [11] and KILT [12]. There are also approaches to similarity in DL formalisms, such as [13]. In this work, Hu *et al.* present a method for calculating distances between concepts based on their signatures. A concept signature is the set of elements that a concept is dependent of, which is determined using tableaux-like reasoning rules. Their approach starts by computing the signatures of concepts and counting the times each element (atomic concept and role) appears in the signature and fine-tuning it using information retrieval techniques. They define the distance between ontologies by aggregating the distances between their different components. Herein lies an advantage of their work: they define similarity on many levels; our work focuses on ontologies as wholes. An advantage of our work is that our measure is bounded between 0 and 1, as opposed to their work which can yield any (possibly negative) number, and thus cannot be strictly considered as a similarity measure since it does not hold the positive definiteness condition. Other work in DL similarity can be found in [8, 14, 15].

7 Conclusions and Future Work

In this paper we present a semantic similarity measure for a sub-set of the $\mathcal{ALC}$ Description Logic. We show some properties of the measure, how it can be applied to average-sized ontologies and that the results yielded roughly correspond to our intuitive notion of similarity.

In the future, we would like to tackle the efficiency and expressiveness problems. A formal analysis of the algorithm should be done in order to provide a deeper understanding of the limitations of the current implementation. We would like to add the possibility of having a weighing factor in the form of a probability distribution over concepts. Finally, we should apply it to a use-case, namely in the automatic assessment of learnt ontologies against a gold standard.

References

1. Ehrig, M., Euzenat, J.: State of the art on ontology alignment. Knowledge Web Deliverable 2.2.3, University of Karlsruhe (2004)
2. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M., Milios, E.: Information retrieval by semantic similarity. Int'l Journal on Semantic Web & Information Systems **2**(3) (2006) 55–73
3. Nardi, D., Brachman, R.J.: An introduction to description logics. In Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: Description Logic Handbook, Cambridge University Press (2003) 1–40
4. Araújo, R., Pinto, H.S.: SEMilarity theory. Technical report, INESC-ID (2007) See `http://algos.inesc-id.pt/~rnna/semilarity-theory.pdf`.
5. Jr., R.J.B., Pehoushek, J.D.: Counting models using connected components. In: AAAI/IAAI. (2000) 157–162
6. Ribeiro, R., Batista, F., Pardal, J.P., Mamede, N.J., Pinto, H.S.: Cooking an ontology. In Euzenat, J., Domingue, J., eds.: AIMSA. Volume 4183 of LNCS., Springer (2006) 213–221
7. Dellschaft, K., Staab, S.: On how to perform a gold standard based evaluation of ontology learning. In Cruz, I.F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L., eds.: Proc. of ISWC 2006. Volume 4273 of LNCS., Springer (2006) 228–241
8. d'Amato, C., Fanizzi, N., Esposito, F.: A dissimilarity measure for ALC concept descriptions. In Haddad, H., ed.: Proc. of the 2006 ACM Symposium on Applied Computing (SAC), ACM (2006) 1695–1699
9. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: S-match: an algorithm and an implementation of semantic matching. Technical report, University of Trento (2004)
10. Vögele, T., Hübner, S., Schuster, G.: BUSTER - an information broker for the semantic web. KI **17**(3) (2003) 31
11. McGuinness, D., Fikes, R., Rice, J., Wilder, S.: An environment for merging and testing large ontologies. In: Proc. of KR 2000, Morgan Kaufmann (2000) 483–493
12. d'Aquin, Mathieu, Bouthier, C., Brachais, S., Lieber, J., Napoli, A.: Knowledge editing and maintenance tools for a semantic portal in oncology. Int'l Journal of Human-Computer Studies **62**(5) (2005) 619–638
13. Hu, B., Kalfoglou, Y., Alani, H., Dupplaw, D., Lewis, P.H., Shadbolt, N.: Semantic metrics. In Staab, S., Svátek, V., eds.: Proc. of EKAW 2006. Volume 4248 of LNCS., Springer (2006) 166–181
14. Borgida, A., Walsh, T., Hirsh, H.: Towards measuring similarity in description logics. In Horrocks, I., Sattler, U., Wolter, F., eds.: Proc. of DL2005. Volume 147 of CEUR Workshop Proceedings., CEUR-WS.org (2005)
15. Janowicz, K.: Sim-DL: Towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval. In Meersman, R., Tari, Z., Herrero, P., eds.: On the Move to Meaningful Internet Systems. Proc., Part II. Volume 4278 of LNCS., Springer (2006) 1681–1692