Measuring Similarity in Ontologies: A new family of measures

Tahani Alsubait, Bijan Parsia, and Uli Sattler

School of Computer Science, The University of Manchester, United Kingdom
{alsubait,bparsia,sattler}@cs.man.ac.uk

1 Introduction

Similarity measurement is important for numerous applications. Be it classical information retrieval, clustering, ontology matching or various other applications. It is also known that similarity measurement is difficult. This can be easily seen by looking at the several attempts that have been made to develop similarity measures, see for example [2, 4]. The problem is also well-founded in psychology and a number of psychological models of similarity have been already developed, see for example [3]. Rather than adopting a psychological model for similarity as a foundation, we noticed that some existing similarity measures for ontologies are ad-hoc and unprincipled. In addition, there is still a need for similarity measures which are applicable to expressive Description Logics (DLs) (i.e., beyond $\mathcal{EL}$) and which are terminological (i.e., do not require an *ABox*). To address these requirements, we have developed a new family of similarity measures which are founded on the feature-based psychological model [3]. The individual measures vary in their accuracy/computational cost based on which *features* they consider.

To date, there has been no thorough empirical investigation of similarity measures. This has motivated us to carry out two separate empirical studies. First, we compare the new measures along with some existing measures against a gold-standard. Second, we examine the practicality of using the new measures over an independently motivated corpus of ontologies (BioPortal library) which contains over 300 ontologies. We also examine whether cheap measures can be an approximation of some more computationally expensive measures. In addition, we explore what could possibly could wrong when using a cheap similarity measure.

2 A new family of similarity measures

The new measures are based on Jaccard's similarity coefficient which has been proved to be a proper metric (i.e., satisfies the properties: equivalence closure, symmetry and triangle inequality). Jaccard's coefficient, which maps similarity to a value in the range [0,1], is defined as follows (for sets of "features" A',B' of A,B, i.e., subsumers of A and B):

$J(A, B) = \frac{|(A' \cap B')|}{|(A' \cup B')|}$

We aim at similarity measures for general OWL ontologies and thus a naive implementation of this approach would be trivialised because a concept has infinitely many subsumers. To overcome this, we present refinements for the similarity function in which we do not count all subsumers but consider subsumers from a set of (possibly complex) concepts of a concept language $\mathcal{L}$. Let C and D be concepts, let $\mathcal{O}$ be an ontology and let $\mathcal{L}$ be a concept language. We set:

$$\begin{aligned} S(C,\mathcal{O},\mathcal{L}) &= \{D \in \mathcal{L}(\widetilde{\mathcal{O}}) \mid \mathcal{O} \models C \sqsubseteq D\} \\ \mathrm{Com}(C,D,\mathcal{O},\mathcal{L}) &= S(C,\mathcal{O},\mathcal{L}) \cap S(D,\mathcal{O},\mathcal{L}) \\ \mathrm{Union}(C,D,\mathcal{O},\mathcal{L}) &= S(C,\mathcal{O},\mathcal{L}) \cup S(D,\mathcal{O},\mathcal{L}) \\ \mathrm{Sim}(C,D,\mathcal{O},\mathcal{L}) &= \frac{|Com(C,D,\mathcal{O},\mathcal{L})|}{|Union(C,D,\mathcal{O},\mathcal{L})|} \end{aligned}$$

To design a new measure, it remains to specify the set $\mathcal{L}$. For example:

$$\begin{aligned} AtomicSim(C,D) &= Sim(C,D,\mathcal{O},\mathcal{L}_{\text{Atomic}}(\widetilde{\mathcal{O}})), \text{ and } \mathcal{L}_{\text{Atomic}}(\widetilde{\mathcal{O}}) = \widetilde{\mathcal{O}} \cap N_C. \\ SubSim(C,D) &= Sim(C,D,\mathcal{O},\mathcal{L}_{\text{Sub}}(\widetilde{\mathcal{O}})), \text{ and } \mathcal{L}_{\text{Sub}}(\widetilde{\mathcal{O}}) = Sub(\mathcal{O}). \\ GrSim(C,D) &= Sim(C,D,\mathcal{O},\mathcal{L}_{\text{G}}(\widetilde{\mathcal{O}})), \text{ and } \mathcal{L}_{\text{G}}(\widetilde{\mathcal{O}}) = \{E \mid E \in Sub(\mathcal{O}) \\ &\text{or } E = \exists r.F, \text{for some } r \in \widetilde{\mathcal{O}} \cap N_R \text{ and } F \in Sub(\mathcal{O})\}. \end{aligned}$$

where $\widetilde{\mathcal{O}}$ is the signature of $\mathcal{O}$, N_C is the set of concept names and $Sub(\mathcal{O})$ is the set of concept expressions in $\mathcal{O}$. The rationale of $SubSim(\cdot)$ is that it provides similarity measurements that are sensitive to the modeller's focus. To capture more possible subsumers, one can use $GrSim(\cdot)$ for which the grammar can be extended easily.

3 Approximations of similarity measures

Some measures might be practically inefficient due to the large number of candidate subsumers. For this reason, it would be nice if we can examine whether a "cheap" measure can be a good approximation for a more expensive one.

Definition 1 *Given two similarity functions $Sim(\cdot)$, $Sim'(\cdot)$, we say that:*

- $Sim'(\cdot)$ preserves the order of $Sim(\cdot)$ *if* $\forall A_1, B_1, A_2, B_2 \in \widetilde{\mathcal{O}}$: $Sim(A_1,B_1) \leq Sim(A_2,B_2) \implies Sim'(A_1,B_1) \leq Sim'(A_2,B_2)$.
- $Sim'(\cdot)$ approximates $Sim(\cdot)$ from above *if* $\forall A, B \in \widetilde{\mathcal{O}}$: $Sim(A,B) \leq Sim'(A,B)$.
- $Sim'(\cdot)$ approximates $Sim(\cdot)$ from below *if* $\forall A, B \in \widetilde{\mathcal{O}}$: $Sim(A,B) \geq Sim'(A,B)$.

Consider $AtomicSim(\cdot)$ and $SubSim(\cdot)$. The first thing to notice is that the set of candidate subsumers for the first measure is actually a subset of the set of candidate subsumers for the second measure ($\widetilde{\mathcal{O}} \cap N_C \subseteq Sub(\mathcal{O})$). However, we need to notice also that the number of entailed subsumers in the two cases need not to be proportionally related. Hence, the above examples of similarity measures are, theoretically, non-approximations of each other.

4 Empirical evaluation

We carry out a comparison between the three measures $GrSim(\cdot)$, $SubSim(\cdot)$ and $AtomicSim(\cdot)$ against human similarity judgments. We also include two existing similarity measures in this comparison (Rada [2] and Wu & Palmer [4]). We also study in detail the behaviour of our new family of measures in practice. $GrSim(\cdot)$ is considered as the expensive and most precise measure in this study.

To study the relation between the different measures *in practice*, we examine the following properties: order-preservation, approximation from above/below and correlation (using Pearson's coefficient).

4.1 Experimental set-up

Part 1: Comparison against a gold-standard The similarity of 19 SNOMED-CT concept pairs was calculated using the three methods along with Rada [2] and Wu & Palmer [4] measures. We compare these similarities to human judgements taken from the Pedersen et al.[1] test set.

Part 2: Cheap vs. expensive measures A snapshot of BioPortal from November 2012 was used as a corpus. It contains a total of 293 ontologies. We excluded 86 ontologies which have only atomic subsumptions as for such ontologies the behaviour of the considered measures will be identical, i.e., we already know that $AtomicSim(\cdot)$ is good and cheap. Due to the large number of classes and difficulty of spotting interesting patterns by eye, we calculated the pairwise similarity for a sample of concepts from the corpus. The size of the sample is 1,843 concepts with 99% confidence level. To ensure that the sample encompasses concepts with different characteristics, we picked 14 concepts from each ontology. The selection was not purely random. Instead, we picked 2 random concepts and for each random concept we picked some neighbour concepts.

4.2 Results

How good is the expensive measure? Not surprisingly, $GrSim$ and $SubSim$ had the highest correlation values with experts' similarity (Pearson's correlation coefficient $r = 0.87, p < 0.001$). Secondly comes $AtomicSim$ with $r = 0.86$. Finally comes Wu & Palmer then Rada with $r = 0.81$ and $r = 0.64$ respectively. Figure 1 shows the similarity curves for the 6 measures used in this comparison. The new measures along with Wu & Palmer measure preserve the order of human similarity more often than Rada measure. They mostly underestimated similarity whereas the Rada measure was mostly overestimating human similarity.

Cost of the expensive measure The average time per ontology taken to calculate grammar-based similarities was 2.3 minutes (standard deviation $\sigma = 10.6$ minutes, median $m = 0.9$ seconds) and the maximum time was 93 minutes for the Neglected Tropical Disease Ontology which is a $\mathcal{SRIQ}$ ontology with 1237 logical axioms, 252 concepts and 99 object properties. For this ontology, the cost of $AtomicSim(\cdot)$ was only 15.545 sec and 15.549 sec for $SubSim(\cdot)$. 9 out of 196 ontologies took over 1 hour to be processed. One thing to note about these ontologies is the high number of logical axioms and object properties. Clearly, $GrSim(\cdot)$ is far more costly than the other two measures. This is why we want to know how good/bad a cheaper measure can be.

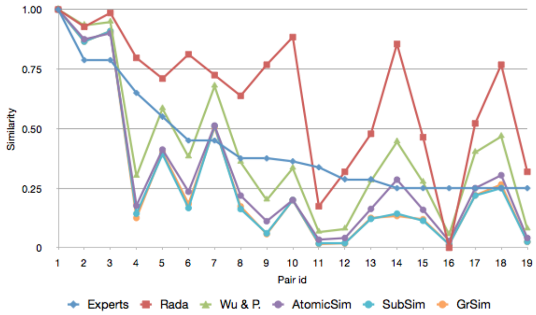

Fig. 1: 6 Curves of similarity for 19 SNOMED clinical terms

How good is a cheap measure? Although we have excluded all ontologies with only atomic subsumptions from the study, in 12% of the ontologies the three measures were perfectly correlated ($r = 1, p < 0.001$). These perfect correlations indicate that, in some cases, the benefit of using an expensive measure is totally neglectable.

$AtomicSim(\cdot)$ and $SubSim(\cdot)$ did not preserve the order of $GrSim(\cdot)$ in 80% and 73% of the ontologies respectively. Also, they were not approximations from above nor from below in 72% and 64% of the ontologies respectively.

Take a look at the African Traditional Medicine ontology in Figure 2. $SubSim(\cdot)$ is 100% order-preserving while $AtomicSim(\cdot)$ is only 99% order-preserving.

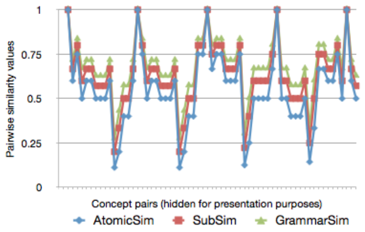

Fig. 2: African Traditional Medicine

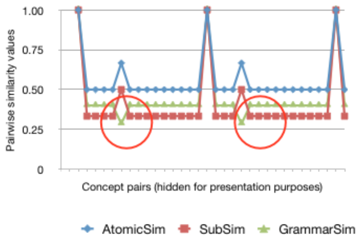

Fig. 3: Platynereis Stage

Note also the Platynereis Stage Ontology in Figure 3 in which both $AtomicSim(\cdot)$ and $SubSim(\cdot)$ are 75% order-preserving. However, $AtomicSim(\cdot)$ was 100% approximating from above while $SubSim(\cdot)$ was 85% approximating from below.

References

1. T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 30(3):288–299, 2007.
2. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. In *IEEE Transaction on Systems, Man, and Cybernetics*, volume 19, page 1730, 1989.
3. A. Tversky. Features of similarity. *Psycological Review by the American Psycological Association, Inc.*, 84(4), July 1977.
4. Z. Wu and MS. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, page 133138, 1994.