Towards Measuring Similarity in Description Logics

Alex Borgida Thomas J. Walsh Haym Hirsh
Dept.of Computer Science
Rutgers University

Abstract

We review several kinds of previously studied concept similarity measures, and then rephrase them in terms of a simple DL. We discuss the difficulties encountered in trying to generalize these formulations to more complex DLs, and settle on one based on probability/information theory as being the most principled.

1 Motivation and Goals

The idea of measuring *concept similarity* has received considerable attention in several domains, including psychology, cognitive science, and computational linguistics and information retrieval. This work has been applied and extended recently in the field of Information Integration, which often relies on ontologies and hence concepts described in DLs. Most past work has concentrated on the similarity of "atomic" concepts (word senses), rather than composite, defined concepts, which are the stock-in-trade of DLs. This has sometimes been characterized as the difference between considering nouns appearing in a dictionary, e.g. WordNet, and general arbitrary noun phrases.

The modest goal of this paper is to consider the problem of generalizing previous efforts to define similarity for primitive concepts to composite ones. To this end, we review three classes of approaches found in the current literature, rephrase them in terms of a very simple description logic, and then try to generalize this to a complex DL. These extensions invite a host of new questions that we leave open for further research.

To begin with, we make several observations about possible real-valued function(s) $sim(C, D)$ that would measure concept similarity in general. First, any effort to assess simlarity via absolute numeric values seems ill-advised — in small experiments even human judgments have been shown to correlate only in the .91 to .94 range. A less exacting target is to use sim to provide relative orderings over concepts. Some researchers have studied a complementary measure, distance $dist(C, D)$, which can similarly be used for ordering alternatives. To the extent that two functions provide similar orderings (e.g., consistently differing by a constant factor) they are essentially indistinguishable for our purposes. Nonetheless, there are two important general properties that are widely agreed to be desirable for $sim(C,D)$:

1. $sim(C,D)$ is positively correlated with the amount of commonality between C and D.
2. $sim(C,D)$ is negatively correlated with the amount of difference between C and D.

The key question then becomes how to measure "commonality"; in turn, this is related to how information about the concept is captured.[1]

2 Three approaches to concept similarity

Clearly, in order to compare concepts we need to consider what information we have available about them. We follow Rodriguez [14] in categorizing a variety of approaches, and for each we give a few of the best known functions, which we shall follow.

Feature-based Models In this approach, a concept C is characterized by a a set of features $\mathit{ftrs}(C)$.

In his pioneering work on similarity, the psychologist Amos Tversky [16] introduced two families of measures: the "contrast model", where the similarity of C and D is a linear function: $contrast_{tv}(C,D) = \theta f(\mathit{ftrs}(C) \cap \mathit{ftrs}(D)) - \alpha f(\mathit{ftrs}(C) \setminus \mathit{ftrs}(D)) - \beta f(\mathit{ftrs}(D) \setminus \mathit{ftrs}(C))$, where $\setminus$ is set difference, θ, α, and β are non-negative constants, and where $f(.)$ is often taken as the count of features in the set (written $|.|$). Tversky also proposed a normalized "ratio model", where similarity is a fraction involving these sets:

$$sim_{tv}(C,D) =_{def} \frac{f(\mathit{ftrs}(C) \cap \mathit{ftrs}(D))}{f(\mathit{ftrs}(C) \cap \mathit{ftrs}(D)) + \alpha f(\mathit{ftrs}(C) \setminus \mathit{ftrs}(D)) + \beta f(\mathit{ftrs}(D) \setminus \mathit{ftrs}(C))}$$

In cases where asymmetry of similarity is not desired (as in this paper), we can normalize things, and assume $\alpha = \beta = 0.5$. Under the assumption that f is distributive over disjoint sets ($f(V \cup W) = f(V) + f(W)$), sim_{tv} is more commonly written as

$$sim_{tv}(C,D) =_{def} \frac{2 \times f(\mathit{ftrs}(C) \cap \mathit{ftrs}(D))}{f(\mathit{ftrs}(C)) + f(\mathit{ftrs}(D))}$$

[1] Other properties, such $sim(C,D)$ being a metric have been hotly contested, and will not be explicitly considered here.

Semantic-network based models In this approach, background information is provided in the form of a semantic network involving concepts and at least is-a edges. Sometimes more complex relationships are considered (as in WordNet).

Similarity measures in this setting usually involve measuring path lengths in the network. In particular, one of the earliest and best known is the proposal by Rada et al [12], which locates the *most specific is-a ancestor* node E=*msa*(C,D) of C and D, and then defines their similarity as the length of the path from C to E plus the length of the path from D to E.[2] More recent proposals also take into account the depth of *msa*(C,D) (in order to apply principle 2), the density of edges at nodes, and possibly edge weights.

Information-content based models In this case we have available both a semantic network, and also information $pr(C)$ about the probability of an individual being described by a specific concept/word C. (Such a probability is usually estimated from some task-specific corpus.) Resnick [13] focused on E=*msa*(C,D) as the representative of the similarity of C and D, but suggested that $pr(E)$ is a better basic measure than the depth of the concept in the Is-A hierarchy, since it is not affected by later changes to the hierarchy. Resnick went on to show that rather than using the probability *pr*(E), one obtains results that correlate better with human judgments by using as a similarity measure *information content*

$$sim_{res}(C,D) = IC(E) =_{def} \ -\log pr(E)$$

Jiang and Conrath [6] proposed a more refined measure of distance compensating for factors like concept depth and density, but a simplified version of this formula dealing only with IC extends Resnick's measure by also satisfying Property 2:

$$dist_{jc}(C,D) =_{def} \ IC(C) + IC(D) - 2 \times IC(msa(C,D))$$

Lin [10] derived mathematically from 6 basic axioms (some about the proposed properties of similarity, some about the form of the function) the related formula

$$sim_{lin}(C,D) =_{def} \ \frac{2 \times IC(msa(C,D))}{IC(C) + IC(D)}$$

In general, when specific similarity measures are proposed, they are experimentally compared with human judgments of similarity and other, previously proposed measures. The recent popularity of IC-based similarity measures is indicated by papers such as [4], which finds $dist_{jc}$ empirically best on a spelling correction task, and [11], which uses sim_{res} and the gene ontology GO.

[2] Both here and below, we only consider the simple case where the semantic net is a tree, so that E is unique.

3 Specifying similarity for descriptions

We now propose to take a very simple DL, $\mathcal{A}$, involving only conjunction, and show how each of the above measures can be applied to it. The important point is that one can then consider what it would take to extend each approach when the DL was generalized from $\mathcal{A}$ to something more complex.

The DL $\mathcal{A}$ allows conjunction of concepts, which can be atomic or defined in an acyclic T-box using expressions of the form $D := C_1 \sqcap \ldots$. We will use A and B for atomic concepts, and C and D for possibly more complex expressions or named defined concepts.

This language admits a simple structural subsumption algorithm, where the normal form *nf*(C) of a concept C is the *set* of atoms appearing in its definition, and $C \sqsubseteq D$ is decided by testing whether $\mathit{nf}(C) \subseteq \mathit{nf}(D)$.

Reformulation in $\mathcal{A}$ For the feature model, we will view features as atomic concepts, and then an ordinary concept is just the conjunction of its features. A simple, but important, observation is that set intersection and difference of the atom sets corresponds, at least in this simple case, to computing the *least common subsumer*[5, 8] and *concept difference* [18, 3] in $\mathcal{A}$. As a result, we can translate into DL notation Tversky's measures as follows:

$$\mathit{contrast}_{tv}(C,D) =_{def} \; f(\mathit{lcs}(C,D)) - 0.5 f(\mathit{diff}(C,D)) - 0.5 f(\mathit{diff}(D,C))$$

$$\mathit{sim}_{tv}(C,D) =_{def} \; \frac{2 \times f(\mathit{lcs}(C,D))}{2 \times f(\mathit{lcs}(C,D)) + f(\mathit{diff}(C,D)) + f(\mathit{diff}(D,C))}$$

Recall that normally in this case f is taken as the *count* of (possibly weighted) features, in this case atomic concepts.

In the case of a semantic network model, we will perform a well-known encoding: whenever node F in the network has is-a parent nodes $G_1, \ldots, G_n$, introduce atomic concept F^*, and now *define* concept F as $F := F^* \sqcap G_1 \sqcap \ldots \sqcap G_n$. In the resulting T-box, the defined concepts have the same subsumption hierarchy as the original corresponding nodes in the semantic network; moreover, if in the original network there was a path $U_1, U_2, \ldots U_n = \top$ to the root of the is-a hierarchy then the normal form of U_1 in the new DL is $\mathit{nf}(U_1) = U_1^* \sqcap U_2^* \sqcap \ldots \sqcap U_{n-1}^*$. In other words, if the network was a tree, then $\mid \mathit{nf}(C) \mid$ is the length of the path from U_1 to the root. Since the paths from C and D to the root first intersect at *msa*(C,D), which once again is the same as *lcs*(C,D), we get

$$\mathit{dist}_{rada}(C,D) =_{def} \; \mid C \mid \; + \; \mid D \mid \; - \; 2 \times \mid \mathit{lcs}(C,D)) \mid$$

where $\mid X \mid$ measures the cardinality of the normal form of X.

For the IC models, notice again the parallel between *msa*(C,D) in a semantic net and *lcs*(C,D) in a DL, so translating IC measures to $\mathcal{A}$ yields:

$$\mathit{dist}_{jc}(C,D) =_{def} \; IC(C) + IC(D) - 2 \times IC(\mathit{lcs}(C,D))$$

$$sim_{lin}(C, D) =_{def} \frac{2 \times IC(lcs(C, D))}{IC(C) + IC(D)}$$

Generalizing beyond $\mathcal{A}$ Let us consider what obstacles we face if we want to extend the above three kinds of approaches to a more complex DL. On the positive side, notions such as *lcs* are available in arbitrary DLs; therefore we can focus on applying the various measures beyond simple atomic names.

In feature based models, the key issues are what counts as a feature, and what are valid decompositions into features. In a propositional DL, one might take minimal disjunctive normal form, and count literals; but it is much less clear what to do with terms constructed using roles. For example, if (**atleast** 3 R), (**atleast** 4 R) and (**atleast** 9 R) each count as a single feature, how would one account for the fact that the first and second would be judged to be more similar than the second and the third? And what should one do about nested role restrictions: $\forall R.\forall R.A$ vs. $\forall R.A$? Clearly, more information is required concerning the salience of roles, and how to combine such measures in the cases of enumenration and nesting to produce a legitimate measure of feature set "size" .

Similarly, in network-based measures, such as $dist_{rada}$, a key problem is that of assigning a useful size for the various concepts in the description. To elucidate the difficulties, we note that in the one paper [17] we have found which tried to address similarity in a DL involving atoms, conjunctions, and existential restrictions ($\mathcal{AE}$), the concepts were converted to a graph and at least three metrics evaluating the size of the lcs were explored, including the sum of the length of the role paths, the number of roles from the root, and the number of roles bearing value restrictions; none of these performed well. This arbitrariness in choosing a size measure for complex concepts appears to be a substantial obstacle to this approach. Once again, we require some mechanism for measureing size beyond what is available in the pure structural form.

In such situations, information content appears to provide a much more sensible basis for the size measure. Essentially, we wish to allow the IC of complex concepts to drive our notions of component salience and allow the laws of probability in a given domain to govern the combination of these weights into a proper measure of "size". Most interestingly, this results in exactly the IC-based measures of similarity in certain simple cases. In order to use an IC-based approach, we need however to be able to estimate the probability pr(C) of an object being an instance of an arbitrary concepts C.

Originally, with what we now see as atomic concepts, probabilities were estimated from large text corpora. For more complex DLs, we could obtain estimates from databases whose semantics are modeled by concepts and roles of the DL we are considering. For example, if we could define a view over the database for every primitive role and concept, then [2] shows how to translate any complex DL concept C into an SQL query Q_C, which returns tuples for all its instances. The count of such tuples, relative to the total number of possible individuals, yields the desired statistic. Moreover, as suggested to us by Alon Halevy, query optimizers maintain statistics about database contents which allow them to estimate query size, including number of tuples, without actually running the query. So one could obtain appropriate probability estimates quite quickly.

More accurate information about probabilities would have to come from an ontology which actually provides this kinds of information. Bacchus [1] proposed using a probability distribution over the domain of interpretation Δ as semantics for probability statements concerning FOL formulas with free variables. P-Classic [7] instantiated this scheme more congenially for DLs, using Bayes nets and provides algorithms for computing pr(C) for arbitrary descriptions for the subset of Classic it considers, and claims that extensions are easily found for more complex DLs. Therefore, the empirical success of IC-based similarity measures on simple concepts, and the existence of a theory for P-Classic indicates that among the choices considered here the best solution to our original problem – to find the similarity of complex descriptions – relies on using IC-based similarity measures backed up by a P-Classic like ontology, which provides information about the probabilities of objects satisfying properties described by concepts.

4 Applying simplifying estimates

Unfortunately, developing full P-Classic ontologies, such as those illustrated in [7], is likely to be a difficult task. More realistically, existing ontologies, such as WordNet will probably be annotated with rough probabilistic estimates, and drastic assumptions of independence will be made in order to guide appropriate measures of IC.

For example, suppose we assume that every individual belongs to atomic concept A with probability p, and that membership in atomic concepts is independent. Then if $C := A_1 \sqcap \ldots \sqcap A_n$, it is easy to see that $pr(C) = p^{\mid C \mid}$, and hence $IC(C) = \mid C \mid \times (-log\, p)$, where the second factor can be treated as a constant. As a result, it is easy to prove the following

Theorem 4.1 *For the logic $\mathcal{A}$, and with the above uniform independence assumption, $sim_{lin}(C, D) = sim_{tv}(C, D)$. Also, the measures $dist_{jc}$, $dist_{rada}$, and $contrast_{tv}$ with $\theta = 0$ and $\alpha = \beta = \frac{1}{2}$, $f(.) = \mid . \mid$ are within a constant factor of each other.*

In the presence of roles, we can again radically simplify P-Classic, and consider a semantics where the probability of role R having exactly k fillers is $q^k(1 - q)$, for some $0 \leq q \leq 1$. (Conceptually, this corresponds to tossing a coin with probability of heads q in order to decide if the role has more fillers; these probabilities add to 1, as needed.) And we assume that for every individual x, the properties of role fillers are independent of each other and of x.

Consider now a description such as $\forall R.D$. According to the semantics, an individual will have exactly k role fillers of type D with probability $q^k(1-q) \times pr(D)^k$. Therefore $pr(\forall R.D)$ is the sum of this geometric series, reducing to $pr(\forall R.D) = (1-q)/(1-q \times pr(D))$. This formula can be used recursively to compute the probability, hence IC, of nested ALL-descriptions once we can estimate a value for q, or even for $(1-q)/(1-q \times pr(A_0))$, for some concept A_0, where $pr(A_0)$ has been estimated. A significant open problem in our research agenda is how such information can be obtained from available data.

Our model of similarity also can be used to approach, though not necessarily decide, questions such as how differences in concept structure might impact concept similarity. For example, consider the series dist(B,B⊓A), dist(B, B⊓$\forall R.A$), dist(B,B⊓$\forall R.\forall R.A$),..., which one might argue should become smaller since more deeply nested restrictions ought to represent smaller differences. Using $dist_{jc}$, the only thing varying is the series IC(A), IC($\forall R.A$), IC($\forall R.\forall R.A$), ... Interestingly, it can be shown that this series decreases *or* increases asymptotically depending on whether $q(1+p)$ is > 1 or < 1, so that there is in fact no single, probability independent answer.

Finally, note that, as in P-Classic, it is not valid to compute $pr($ (**atleast** 3 R) ⊓(**atmost** 5 R)) as the product of pr((**atleast** 3 R)) and pr((**atmost** 5 R)) since the cases they cover are not exclusive. Therefore the concepts need to be considered in a normal form where all information about role R is in a single component, which is in some sense the structural normal form required for computing similarity.

5 Summary

There are literally dozens of proposals for similarity measures, and we could have tried to carry out the same program for each as we did here. For example [9] examines over 10 different non-linear combinations of properties such as depth and path length in WordNet, as well as Information Content, and after 5,000 hours of computation produces a formula that correlates well with the results of human experiments involving 56 words [15]. In trying to find a direction for measuring similarity of complex descriptions in DLs, we considered a few of the best known previous measures and tried to apply them to DLs. This is just a beginning. To some extent, our success stems from transforming the original problem into one of finding estimates for probabilities, and thereby intimately tying the similarity of descriptions not just to their structure but also their distribution in the real world. Numerous other problems have been left open, including dealing with non-tree IsA hierarchies, as well as experimental evaluation of similarity functions applied to noun phrases (as examples of composite concepts).

References

[1] F. Bacchus, “Lp, a logic for representing and reasoning with statistical knowledge”, *Computational Intelligence 6*: 209-231 (1990)

[2] A. Borgida & R. Brachman, “Loading Data into Description Reasoners”, *Proc. SIGMOD 1993*: 217-226

[3] S. Brandt, R. Kusters, A.-Y. Turhan, “Approximation and Difference in Description Logics”, *KR 2002*: 203-214

[4] A. Budanitsky & G. Hirst, ”Semantic distance in WordNet ”, *Workshop on Wordnet, NAACL*, 2001.

[5] W. Cohen, A. Borgida, H. Hirsh: “Computing Least Common Subsumers in Description Logics”, *AAAI 1992*: 754-760

[6] J. Jiang & D. Conrath, “Semantic Similarity based on corpus statistics, and lexical taxonomy”, *Proc. ICCL*, Taiwan, 1997.

[7] D. Koller, A. Levy, & A. Pfeffer, “P-Classic: a tractable probabilistic description logic”, *Proc. AAAI 1997.*

[8] R. Kusters & R. Molitor, “Computing Least Common Subsumers in ALEN”, *IJCAI 2001*: 219-224

[9] Y. Li, Z. Bandar, D. McLean, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”, *IEEE Trans. Knowl. Data Eng. 15(4)*: 871-882 (2003)

[10] D. Lin, “An Information theoretic definition of similarity”, *Proc. ICML 1998* .

[11] P. W. Lord, R. D. Stevens, A. Brass & C. A. Goble, “Investigating semantic similarity measures across the Gene Ontology”, *Bioinformatics 19(10)*: 12751283 (2003)

[12] R. Rada, H. Milli, E. Bicknell, M. Blettner, ”Development and Application of a metric on Semantic Nets”, *IEEE Trans. on Systems, Man, and Cybernetics*, 19(1): 17-30 (1989)

[13] P. Resnik, ”Using Information Content to Evaluate Semantic Similarity”, *Proc. IJCAI 1995*: 448-453

[14] A. Rodriguez, “Assessing semantic similarity between spatial entity classes”, Ph.D. Thesis, University of Maine, 1997.

[15] G. Miller & W.G. Charles, ”Contextual correlates of semantic similarity”, *Language and Cognitive Processes*, 6, 1-28, 1991.

[16] A. Tversky, “Features of Similarity”, *Psychological Review 84(4)*: 327-352, 1977.

[17] P. Weinstein & P. Birmingham, “Comparing Concepts in Differentiated Ontologies”, *Proc. KAW’99*, Banff, Canada.

[18] G. Teege. “Making the Difference: A Subtraction Operation for Description Logics”, *Proc. KR 1994*: 540-550