Dimensions of ontological similarity

Pawel Szmeja*, Maria Ganzha*‡, Marcin Paprzycki* and Wieslaw Pawłowski†

*Institute of System Research, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland
{name.surname}@ibspan.waw.pl

†Institute of Informatics, University of Gdańsk, ul. Wita Stwosza 57, 80-952 Gdańsk, Poland
w.pawlowski@inf.ug.edu.pl

‡Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland

Abstract—Thus far, many methods for calculating semantic similarity have been proposed. Most of them produce a single value that is to “represent all aspects of similarity.” In opposition to this approach we argue that semantic entities can be similar in different ways, each with a separate score (and meaning / interpretation). We propose conceptualization of semantic similarity through separate dimensions that denote / capture different “aspects of similarity.”

I. INTRODUCTION

OVER the years, multiple methods of calculating semantic similarity have been proposed. For ontologies, similarity of pairs of entities is based either on simple terms or complex concepts. This note concerns measurement of semantic similarity between entities in description logic (DL). Let us stress that we are not interested in ontology matching, understood as finding similarities between ontologies (e.g. as in [27]). Instead, we consider a single ontology and answers to questions like: what is the similarity between two (or more) entities within it? or, for a given entity, which entity (form a given set) is most similar to it?

II. DESCRIPTION LOGIC

Let us start from a few comments about description logic (DL). In DL, knowledge is stored in knowledge bases (KB) that contain ontologies. KBs are partitioned into TBox, ABox and RBox. TBox contains declarations and descriptions of concepts (classes). Each concept description is constructed from concept names, role names, constants and a set of constructors. A hierarchy of concepts is called a taxonomy. ABox contains assertions about individuals built from concept names, role names and constants. An individual is an instance of one, or more classes, and can have roles (properties) assigned to it. RBox describes relationships between roles. In what follows, concept names are denoted by A, B, ..., Z, individual names by a, b, ..., o, and role names by p, r, ..., z.

In DL (see, [9]), there are two definitions relevant to calculation of similarity – the least common subsumer (LCS; [10]) and the most specific concept (MSC; [11]). The LCS of entities X and Y is the most specific entity that is an ancestor to both X and Y. In a taxonomy, it is a concept that shares the most types with the compared concepts. The MSC of an individual is a concept whose description is built from assertions about the individual in a way that includes every such assertion. In other words, it is a class that is built specifically to contain that individual.

III. SIMILARITY MODEL

Let us now discuss selected models of similarity, used in practice (for a more comprehensive list, see [5]). In Tversky’s model [12] concepts are represented by sets of features. Similarity of two sets of features X and Y is given by

\[ S_{TV}(X,Y) = \frac{\alpha f(X \cap Y) + \beta f(X - Y) + \gamma f(Y - X)}{f(X) + f(Y)} \]

where \( f \) is a monotonically increasing function while \( \alpha, \beta \) and \( \gamma \) are positive coefficients. For different choices of coefficients \( S_{TV}(X,Y) \) produces different formulas; e.g. for \( \alpha = \beta = \gamma = 1 \) and \( f = | \cdot | \) the model becomes the Jaccard index \( J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \).

In the Information Content (IC) model, proposed in [13], similarity is related to the “amount of information” an entity provides. The IC of an entity \( e \) is computed from its probability \( p(e) \): \( IC(e) = -\log(p(e)) \). When applied to a textual entity in a corpus, \( p(e) \) is the probability that this entity appears in a given document from the corpus. In a taxonomy, probability of an entity is inversely proportional to the number of entities it subsumes. It is calculated from the IC of the most informative common ancestor (MICA) – a common subsumer that has the maximum IC: \( S_{Res}(X,Y) = IC(MICA(X,Y)) \).

Edge-based models assume that the edge distance in an ontology graph is meaningful for similarity. The simplest approach considers similarity to be equal to the length of the shortest path between a pair of concepts \( S_{Res}(X,Y) = \min(paths(X,Y)) \) (6), where \( paths(X,Y) \) is the set of path lengths in an IS-A graph (see, Section IV). More sophisticated methods, such as, e.g. \( Sim_{Wu}(X,Y) = \frac{2 \cdot depth(LCS(X,Y))}{\text{depth}(X) + \text{depth}(Y)} \) (7), involve normalization and take into account depth of the compared entities, depth of their LCS, or length of path between the root, LCS and entities. Finally, [8] utilizes multiple relations (not just IS-A) in a graph (multigraph).

IV. SEMANTIC SIMILARITY DIMENSIONS

This note draws on the idea of knowledge dimensions introduced in [2], where ontological knowledge was divided into subsets (dimensions). This idea was applied (in [1]) to combine “dimensional scores” into a (single) similarity score.

The idea of semantic similarity dimensions is based on a conjecture that different “aspects of similarity” can be...
“grouped together.” Those groups represent different dimensions of semantic similarity. A simple example of this approach is a comparison of two physical objects with respect to shape and color. Both features have independent similarity scores (objects can have similar color and different shape). It can be shown that, in DL, different dimensions use sets of data that are “separable.” Let us follow this idea with further details.

External similarity dimensions

External similarity dimensions involve information from outside of the knowledge base or the ontology, that contains compared entities. External methods use a small (likely atomic) part of an entity description, that serves as an identifier, to find information about it in external sources.

Here, a pair of natural language labels can be subject to the lexical methods. They can utilize dictionaries and lexical ontologies (such as WordNet) to assess similarity of entities (see [1]). A simple lexical method could, for instance, extract concept labels and use WordNet’s synsets of these labels as features, in a feature-based method. Lexical dimension is most useful when entities have uniquely identifying labels (e.g. “dolphin” and “porpoise”) and gives poor results for non-unique labels (e.g. human names).

Co-occurrence methods (known for questionable results [3]) are also external. Here, similarity is calculated assuming that entities that appear together in high number of contexts (e.g. words in many text corpora) are “similar.” For instance, web search that measures the number of web pages that contain two (or more) terms is a co-occurrence method.

The biggest weakness of the external similarity dimensions is that they operate on small terms rather than full descriptions. Usefulness of methods in external dimension relies on existence of a good unique IDs for the considered entities.

Internal dimensions

Similarity in the hierarchical sort dimension (also called hierarchical or taxonomical) is based on the underlying taxonomy. Here, in the sort dimension, similarity grows with each IS-A relation that points to the same concept for both entities and decreases with each relation that holds true only for one. Many edge-counting methods (e.g. [19]), some IC methods (e.g. [14], [26]) and feature methods (e.g. [4]) can be used in this dimension. The sort dimension includes information about subsumption (⊆), inclusion (∋) and equivalence (≡). Similarity measures that work on subsumptions usually take into account subsumers of measured classes, rather than children. Some IC methods make use of number of children (subsumed classes) to calculate “probability” of a node.

The descriptive dimension encapsulates properties that an entity “has” as opposed to what it “is.” It compares attributes, or properties of objects. The more disparate attributes between entities, the less similarity and vice-versa. More formally, in descriptive dimension one can treat “descriptive” expressions as sets of features, either for a TBox or an ABox. However, existing methods usually do not distinguish between sort and descriptive data, so no purely descriptive methods have been found.

Another way to create a dimension is to isolate a part of the descriptive knowledge into its own dimension. Any subdivision of the descriptive dimension generally means loss of universality, i.e. one cannot apply the new dimension to every ontology. An example is the compositional dimension that takes into account roles that denote “having parts,” “having ingredients,” etc. Compositional dimension has a very clear interpretation and humans often look at composition of a physical object. In practice this dimension is applicable only if the knowledge base contains specific role that fits this dimension (hasPart, isPartOf, etc.). Both SSN [20] and WordNet [16] have such roles.

Instead of the role (hasPart) one might choose a set of roles to represent a dimension. For instance, in a sensor ontology [20], a physical dimension would describe size, mass and shape of a sensor. In this case, other features such as location or interface type would be irrelevant.

A dimension that is not a subdivision of a descriptive dimension, is the membership dimension. It can be used to measure similarity (only) between concepts by comparing sets of individuals that are of a specific type. The membership dimension is implicitly used in [15] to build feature sets (and to calculate similarity as in the Tversky’s feature method).

V. Example of multi-dimensional similarity

Let us now consider an example of dimensional similarity scores in a mock-up biological ontology (see Fig. 1), which is an extract of a phylogenetic ontology, with added roles. We compare three concepts – short-beaked common dolphin [21], silvertip shark [22] and lesser electric ray [24] denoted D, S and R respectively. Data originates from [21], [22], [23], [24], [25]. This example is meant to demonstrate usage and indicate usefulness of similarity dimensions. The taxonomic methods use data in the hierarchy of concepts. In this example there are 20 phylogenetic concepts (classes, including ⊤). According to Resnik’s method [13] $S_{Res}(D, S) = IC(MICA(Y, X))$, $IC(e) = - log(p(e))$ similarity scores are as follows: $S_{Res}(D, S) = IC(CHORDATA) = - log(\frac{18}{20}) \approx 0.105$, $S_{Res}(S, R) = IC(SELACHIMORPHA) = - log(\frac{9}{20}) \approx 0.799$, $S_{Res}(D, R) = IC(CHORDATA) \approx 0.105$. Note that this example contains only a fraction of phylogenetic classes and in a full ontology Resnik’s method would give a different! score. Calculation of Jaccard index $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, where $A_f$ is a set of features of $A$ and assuming that each type of a concept (including ⊤) is a feature, gives the following results: $J(D, S) = \frac{4}{16} \approx 0.188$, $J(S, R) = \frac{6}{11} \approx 0.429$, $J(D, R) = \frac{3}{16} \approx 0.188$.

In the descriptive dimension we use role restrictions. We can, again, use the Jaccard index, using roles as features. Hence, the similarity dimensions are, in fact, independent of the method. Here, D and S have 6 roles each, while R has 4 roles.
Let us also consider the physical dimension represented by roles for mass, length and coloration. To better represent differences between numerical values, a simple ratio method is used for the data values of the same role. This similarity is equal to the smaller value divided by the larger one 

\[ \text{Sim}(k', l') = \frac{\min(k', l')}{\max(k', l')} \]

where \( k' \) and \( l' \) are values of role restrictions or assertions, about the same role \( r \). For instance, similarity of average weight between \( D \) and \( S \) is 

\[ \text{Sim}(D, S) = \frac{0.2}{0.907} \approx 0.225 \]

Total similarity, is this dimension, is calculated by taking arithmetic average over similarity of each relevant role. The scores are:

\[ \text{Sim}_{ph}(D, S) = \frac{0.105 + 0.799 + 0.079}{3} = 0.363 \]
\[ \text{Sim}_{ph}(S, R) = \frac{0.075 + 0.0 + 0.0}{3} = 0.025 \]

Obtained similarity scores are summarized in Table I. Observe that each method produces different similarity scores, even in the same dimension; e.g. Resnik’s method gives different results than Jaccard method. This is, for instance, because of the assumption of Resnik that distance to the root in an ontology (a level) is significant. This also means that the results produced by the Resnik’s method depend on “size of the taxonomy.”

The levels of example concepts do not correspond with levels of phylogenetic classification, e.g. the dolphin does not have a subclass or superorder, so technically its order (CETACEA) is on the same ontological level as subclass of the shark (ELASMOBRANCHII), even though intuitively (and in accordance with biological research) an order should be more informative than a subclass.

Explanation of those results lays in the fact that the descriptive features were not used when constructing phylogeny. Features such as diet, type of reproduction, coloration, period of gestation, and others vary in the same genus, so species are not classified based on those characteristics. Purely taxonomical methods (such as Resnik’s) do not take such features into account at all. Consequently, in this case, the descriptive results are independent of the taxonomy.

Note also that the physical dimension score does not coincide with the descriptive score, even though the former is, theoretically, a subdimension of the latter. This difference stems from the difference in methods used in each dimension. The physical method takes into account degree of difference between values corresponding to the same role, while the descriptive one does not. Note that there is no good and universal method that would compare complex descriptions in expressive DLs in an in-depth manner.

The “final answer” to the question of “how similar are two concepts?,” for the dolphin and shark, according to Resnik’s method is 0.105. According to the method of dimensional similarity \( D \) and \( S \) have taxonomical similarity of 0.188, descriptive similarity of 0.2 and physical similarity of 0.932.

Overall the results suggest that short beaked dolphins and silvertip sharks look similar (high physical similarity), but are evolutionary different (low or average taxonomical score). Note that this statement is possible because separate dimensions of similarity have been independently evaluated.

One dimension that was not used (for sake of brevity) is the compositional dimension. It would comprise of physical “components” of the animals with additional details, for instance, fins (e.g. small pointed dorsal fin), details of bone structure (e.g. serrated teeth), specific organs and functions. Note that this dimension has a specific interpretation in the context of phylogeny.

## VI. COMBINING DIFFERENT DIMENSIONS

Separate similarity scores can be combined into a single number. In [1] authors used a weighted sum of 5 dimensional similarity scores, and weights trained against human-based similarity scores. The suggested advantage of this approach is that the final score combines “all available knowledge.” The main disadvantage is that there is no reason to believe that weights calculated for one ontology will “work” for a different one. There is also a question: what is the actual meaning of these weights?

Let us now propose that representing similarity scores in a similarity vector (containing all dimensions as separate values) would avoid aforementioned problems. One would also posses precise information as to what similarity dimension each score represents. For instance, as can be seen above, the lexical dimension answers “how similar are the names,” co-occurrence answers “how often do they appear together,” taxonomical dimension deals with “classification or types of entities.”

![Phylogeny ontology example](image-url)
In the example from section V the taxonomical score has an interpretation of phylogenetic similarity (i.e. evolutionary ancestry). Understanding of what is phylogeny and how it is constructed further improves the understanding of this dimension. In SSN [20], the taxonomical dimension has a different interpretation and describes the kind of sensor, process or model (depending on module). Descriptive dimension answers general questions about roles, properties, characteristics, etc. and, similarly to taxonomical dimension might answer a more specific question depending on the context.

VII. CONCLUDING REMARKS

The notion of (semantic) similarity is, by nature, vague and ambiguous. Many semantic similarity measuring methods have been proposed and work well for ontology-specific or domain-specific applications. Their results, however, do not generalize across domains (or ontologies). Similarity dimensions address this problem and attempt to rectify the ambiguity of similarity scores. A combined, single score suggests how similar two entities are, but does not answer the question: in what way are the entities actually similar? Multi-dimensional similarity scores may provide such answers by treating each dimension separately. Thus, it is possible to capture the fact that being similar, or not, the entities actually similar? Multi-dimensional similarity scores specify not only how similar entities are, but also why.

Furthermore, as is apparent from the phylogeny example (section V), in practice, known methods utilize only a part of available knowledge. For instance Resnik’s method is purely taxonomical, but this fact is apparent only after analysis of the method. Hence, merely by explicitly labeling this method as taxonomical one gains valuable information. Someone not familiar with details of Resnik’s method would not know why similarity does not change, even if one adds a lot of roles into the KB. Labeling it as taxonomical informs us that it is insensitive to roles.

It should be stressed that the proposed approach is flexible as: (i) it still allows attempts at domain/ontology specific methods to combine separate scores into a single one (as in, [1], [2]), (ii) it is possible to restrict similarity dimensions that are actually considered in a given domain (e.g. only taxonomic and compositional dimensions), based on the “nature of the application.”

Note that accurate dimensional characterization of complex methods, such as IntellIGO [18] (a domain specific method for the Gene Ontology [17] similarity), may turn out very difficult. Note that this is not a judgment about correctness or usefulness of such methods. As a matter of fact, these application specific methods may work best in capturing the “nature of the specific problem” (similarly to training weights in [1]).

This, in turn, may lead to the following general conclusion. There are two ways of dealing with semantic similarity. First, the general approach (e.g. the one proposed in this note), based on application of similarity dimensions, with separate scores to understand how similar entities are and in what way. Second, development of domain / ontology specific methods that capture only the nature of the problem of interest.

REFERENCES

silvertipshark.html