Representation of Concept Specialization Distance through Resemblance Relations

Miguel-Ángel Sicilia[1], Elena García[2], Paloma Díaz[1], and Ignacio Aedo[1]

[1] DEI Laboratory, Computer Science Department, Carlos III University,
Avd. Universidad 30 – 28911 Leganés, Madrid (Spain)
{msicilia, pdp}@inf.uc3m.es, aedo@ia.uc3m.es

[2] Computer Science Department, University of Alcalá
Ctra. Barcelona km. 33.600 – 28871, Alcalá de Henares, Madrid (Spain)
elena.garciab@uah.es

Summary. Generalization-specialization (*gen-spec*) relationships between pairs of classifiers can be assigned a grade of strength or relative distance, somewhat representing the level of similarity between the class and its subclass. In many cases, these distances can not be adequately computed from the structural features or properties of the classifiers, since class-subclass discrimination semantics are often not represented explicitly, and distances can only be properly assessed subjectively by humans. In this work, we describe a novel approach that uses resemblance relations to model graded specializations, both from a specific classifier (locally) and also along a subset of the generalization hierarchy (globally). We also show how that approach can be combined with current Web-enabled ontology description languages to carry out adaptive behaviors that involve crawling the *gen-spec* hierarchy.

Key words: Generalization/specialization relationship, subtyping, ontology, resemblance relations, RDF

1 Introduction

The concept of generalization (and its inverse specialization) plays a central role in current approaches to knowledge representation using ontologies, in general-purpose object-oriented modelling notations (like the U*nified Modeling Language*, UML [12]), and also in other related fields like object-oriented databases [11] or programming languages. In addition, the resulting taxonomic relations have been integrated within reasoning in approaches like many-sorted logic [10], order-sorted logic [2] and description logic [6]. Generalization (often called '*is-a*' or generalization/specialization – *gen-spec* –) is a relation between classes (or classifiers, in a more generic sense) that implies a taxonomic relation, and its subsequent inheritance semantics. This notion has been studied in the field of object-oriented programming and design under the name of *type/subtype* relation, and the essential property has been considered to be the *subtype requirement* [8], according to which if $f(x)$ is a property provable about objects x of type T, then $f(y)$ should be true for objects y of type S, where S is an specialization of T (this is essentially the same interpretation underlying *subsumption* [2]). This basic assumption appears in one form or another in all the above mentioned fields. Without breaking that requirement, some approaches allow a specialization to have an empty set of extensions. For example, the DAML+OIL ontology markup language [7] allows the subclass relation between classes to be acyclic (while the RDF [14] language do not), providing a way to assert class equality, but this can be considered an extreme case.

The common understanding of the '*is-a*' relation considers it as '*all-or-nothing*', in the sense that the relation is equally strong between a class and any of its subclasses, and also at every level of the hierarchy. This assumption is in many cases an oversimplification of the psychological account of the real-world relations we are modelling. In other words, some sub-classes can be considered to be closer to a given super-class than others. As a somewhat extreme example, let's suppose we have a hierarchy rooted in the `mammal` class, with sub-classes `domestic-cat`, and `primate`, and `siamese-cat` as a subclass of `domestic-cat`. We can (subjectively) consider that the first specialization level represents a bigger step than the second, and that the distance from the abstract `mammal` category to `primate` is somewhat shorter than its distance to `domestic-cat`[1] (in the sense that the latter is a more specific category, while the former is still rather abstract). Most current *gen-spec* semantics simply neglect this fact, resulting in a subtle problem of *epistemological adequacy* (using the term in the sense given in [9]).

As a second example, let's suppose a Web shopping recommender agent is operating on the UNSPSC[2] product ontology to recommend items to the user according to a set of products the user is believed to like. If a user has a preference on the class `Pianos`, the agent may crawl to its direct superclass `Keyboard-instruments`, and then try to show related products by going down in the hierarchy to `Musical-organs` and `Accordions`. The *distance* at that level can be considered short, in comparison to crawling the generalization level from `Keyboard-instruments` to its superclass `Musical-instruments`, and in turn, this latter level is somewhat at a shorter distance than the one from `Musical-instrument` to the general category represented by the class `Musical Instruments, Recreational Equipment, Supplies and Accessories` (which include disparate products categories like fitness equipment and toys). In consequence, the agent would tend to crawl

[1] Obviously, this example is an oversimplification of the *mammals* taxonomy, but analogous cases are very often found in other contexts.

[2] See `http://www.unspsc.org`

shorter levels more than larger ones, since the recommendation proximity of the products decrease as distance from the superclass increases.

A third example is the class Bulldozer in the *Cyc Transportation Ontology*[3], which is a subclass of the classes named RoadWork-Vehicle and TransportationDevice-Vehicle. Intuitively, it's clear that the first superclass is closer to the class than the latter, and thus, an agent would decide to take the first path – before the second – in a shallow reasoning process.

The just presented notion of 'graded' specializations has been somewhat addressed *at the instance level* in fuzzy conceptual modelling, by constraining the membership grades of a *fuzzy subclass* [5], with a threshold that represent the minimum distance from the superclass. But it would be more convenient to separate it clearly from that notion of fuzzy subclass and deal with it *at the class-level*, to enrich conceptual definitions with semantics that can be used with no regard to the instance level.

In this work, we describe the semantics of an approach to generalization in ontologies that allows the definition of graded specializations at the class (or term) level, measuring *how much* a classifier specializes a more general concept (that is, a *distance* notion). In Section 2, the problem is described, and resemblance relations are proposed as a measure for closeness between classes in classifier hierarchies. Section 3 describes how these fuzzy resemblance measures can be integrated in a modern ontology definition language and sketches a simple scenario that takes advantage from them. Finally, conclusions and future work are provided in Section 4.

2 Resemblance as a Metric for Specialization Distance

We'll denote a generalization relationship (or link) between two classifiers in the universal set C of classifiers as (informally) defined in (1).

$$a \succ^d b \quad a, b \in C \tag{1}$$

$$d = \{\phi_i(a, y) \mid a \succ^d y \ \wedge \ y \in C\} \tag{2}$$

The discriminator d determines the taxonomic criterion that justifies the relationship, and can be represented in the most general case by a set of predicates (2) – one for each direct specialization – that determines the specific properties of the instances of each subclass. Each of the predicates ϕi characterize one of the subclasses discriminated, and these characterizations can be expressed (for example, discriminating by ranges, 'a viola is a *alto* string-instrument', while 'a violin is a *treble* string instrument', and so on with the

[3] A modified version of the Cyc's taxonomy of transportation devices has been used as a case study for the approach described in this paper. It's available at http://opencyc.sourceforge.net/daml/cyc-transportation.daml

cello - *tenor* - and the bass). Note that on a single specific (super-)class, an arbitrary number of discriminators $d_1, d_2, \ldots, d_n$ can be defined, corresponding to different specialization criteria.

Discriminators are considered to be predicates on the structure of the involved classifiers, but in practice, they are often denoted simply as a set of constants or labels. That is, an specific discriminator d_j is not described as a set of predicates, but as the simple enumeration of the subclasses that participate in the discrimination (3). We'll use this definition from here on, for simplicity's sake.

$$d_j = \{c_1, c_2, \ldots c_k\},\ c_i \in C \tag{3}$$

This approach correspond to the way in which they are represented in the UML by the discriminator *meta-attribute* in *meta-class Generalization* [12].

For example, the *Cyc Transportation Ontology* includes the concepts of `RoadVehicle-Electric`, `RoadVehicle-ICE` (*internal-combustion-engine*), `Bus-RoadVehicle`, `Automobile` and `Motorcycle`, as subclasses of the class `RoadVehicle`. At least two discriminators originate the taxonomy at that level. The first and second subclasses can are clearly discriminated by *motor-type* (4), while the third, fourth and fifth are chiefly distinguished by the room they provide for passengers (5) (according to the ontology documentation), so that we have two discriminators that originate from `RoadVehicle`).

$$d_{motor-type} = \{\mathtt{RoadVehicle-Electric}, \mathtt{RoadVehicle-ICE}\} \tag{4}$$

$$d_{room} = \{\mathtt{Bus-RoadVehicle}, \mathtt{Automobile}, \mathtt{Motorcycle}\} \tag{5}$$

Given a classifier, its specialization links to its direct subclasses are divided in disjoint sets (partitions), according to their discriminators. Each partition represents an orthogonal dimension of specialization, and as such, should be handled separately. P denotes the set of (local) partitions of a model or ontology defined on a set C of classifiers (6). For example, $p_{(motor-type,\mathtt{RoadVehicle})} = \{\mathtt{RoadVehicle-Electric}, \mathtt{RoadVehicle-ICE}\}$.

$$P = \{p_{(d,a)} \mid a \in C\} \quad where\ p_{(d,a)} = \{c \mid c \in C\ \wedge\ a \succ^{d} c\} \tag{6}$$

Mechanisms can be devised for which an estimation of resemblance measures can be automatically obtained [13]. For example, the number of varying structural features from class to subclass could be used as a metric [1]. But all these mechanisms are inherently flawed in practical settings since common conceptual models have incomplete contextual information items that often are not encoded in the model itself (i.e. hierarchy modelers abstract many details that would be needed for the measure of specialization distance, since many specializations seems obvious for them). Therefore, for practical reasons, it would ultimately be required that a human (the modeler or even the users of the model) gives an assessment of specialization measures. But humans find it difficult to give such measures in a global way, since distance is a relative concept.

We have designed a number of small experiments (using the above mentioned ontologies) to gather some evidence about how people tend to assess the relative distance between classes and subclasses. Results have leaded us to sketch an approach for the task, which involves human assessments at two levels:

- At a *micro-level*, in which the distance between a class and its subclasses for a specific discriminator (i.e. for a specific partition $p \in P$) is assessed.
- At a *sub-tree level*, in which distances (obtained at the micro-level) inside a hierarchy tree including descendants of a given classifier are somehow '*harmonized*' (we do not cover this level in detail in this paper).

Due to the subjective nature of such assessments, a large population would be required to come up with a statistically reliable measure, but the acquisition process is outside of the scope of the present work. In order to represent assessments, we have used *resemblance relations* to model specialization distance (note that resemblance or similarity relations can be used also as general semantic relationships that are not related with the subtype requirement, but they're not considered here). A resemblance relation R on a crisp domain D is a binary fuzzy relation (7).

$$R : D \times D \rightarrow [0 \ldots 1] \tag{7}$$

which satisfies reflexive (8) and symmetric (9) properties.

$$R(x, x) = 1 \ \ \forall x \in D \tag{8}$$

$$R(x, y) = R(y, x) \ \ \forall x, y \in D \tag{9}$$

Given this definition, a separate partial resemblance relation R can be obtained (from micro-level assessments) locally for each partition of subclasses, so that we operate on a set of relations (10) in the form (11).

$$\Pi_D = \bigcup_{x \in P} R_x \tag{10}$$

$$R_x : p_{(d,c)} \cup \{c\} \times p_{(d,c)} \cup \{c\} \rightarrow [0 \ldots 1] \tag{11}$$

Relations are labelled partial since they only contain class-subclass relationships, that is, relations are really defined in the form $R_x : \{c\} \times p_{(d,c)} \rightarrow [0 \ldots 1]$, i.e. from a specified super-class to all its subclasses that are discriminated by an specific d (although this can easily be extended to siblings: a complete resemblance relation could be derived for each local partition by the properties of the resemblance relation). This enables a form of stepwise simple reasoning in which concepts at hierarchy level i can be substituted with the closest concept in the $i \pm 1$ level traversing *gen-spec* relations through the different discriminators.

The partial resemblance relations are obtained by directly taking the resemblance grades obtained experimentally from *micro-level* assessments with a sample population knowledgeable in the domain (which requires the conversion of human relative distance assessments – like '*subclass B is rather closer from superclass A than subclass C*' – to numerical values in the unit interval[4]). Once the local resemblance structured, subtree-level assessments can be used to adjust resemblance values entire subtrees of the generalization hierarchy, using only a discriminator for each superclass. This approach only works for limited-size subtrees, since it's necessary to set comparison between all the elements in a set where at least one generalization link is included for each of the partitions.

3 Extending Semantic Markup Languages for Specialization Distance

We have specified an extension to Daml+Oil to encode resemblance relations within Rdf files. The concept of discriminator was added to the language, along with a way to encode local resemblance relations. The following example Rdf fragment sketches the essentials of this extension, which uses a higher-order statement [14] called `withDiscriminator` and `withResemblance` about `subClassOf` statements.

```
<daml:Class rdf:ID="A">
  <rdfs:subClassOf rdf:resource="#B">
       <ext:withDiscriminator rdf:ID="d1"/>
       <ext:withResemblance grade="0.8"/>
  </rdfs:subClassOf>
</daml:Class>

<daml:Class rdf:ID="K">
  <rdfs:subClassOf rdf:resource="#B">
       <ext:withDiscriminator rdf:ID="#d1"/>
       <ext:withResemblance grade="0.7"/>
  </rdfs:subClassOf>
  <rdfs:subClassOf rdf:resource="#D">
       <ext:withDiscriminator rdf:ID="#d2"/>
       <ext:withResemblance grade="very low"/>
  </rdfs:subClassOf>
</daml:Class>
```

[4] We have not already mathematically formalized such a procedure, but it'd possible to do so.

```
<ext:HarmonizedHierarchy rdf:ID="h1">
<ext:fromClass rdf:ID="#B">
<ext:discriminants>
  <rdf:Bag>
    <rdf:li resource="#d1"/>
    <rdf:li resource="#d3"/>
  </rdf:Bag>
</ext:discriminants>
<ext:HarmonizedHierarchy>
```

The example shows that either numeric values or labels in ordered label sets (that could be defined as *XML Schema* dataypes [15]) could be used for the resemblance values. Note also that harmonized hierarchies are explicitly encoded by specifying the root class, and the set of discriminators that have been assessed.

As a case study, a simple personalized recommender agent has been developed, that uses resemblance relations in the filtering process inside a Web application about products. More specifically, the agent takes content items from a database of existing ones for user U that has previously demonstrated interest in a set of classifiers C_U. The items are annotated by terms in the agent's internal ontology. The agent searches for items that match C_U (and that have not been previously visited by U), taking into account the extended annotations by using resemblance to direct subclass or superclasses as a partial match in the absence of full matches. Figure 1 shows an example of the prototype, where the left frame shows changing links about UNSPSC categories, ordered by relevance for the specific user (this is commonly referred to as *adaptive sorting* in adaptive hypermedia research [3]), and taking into account resemblance in filtering related items.

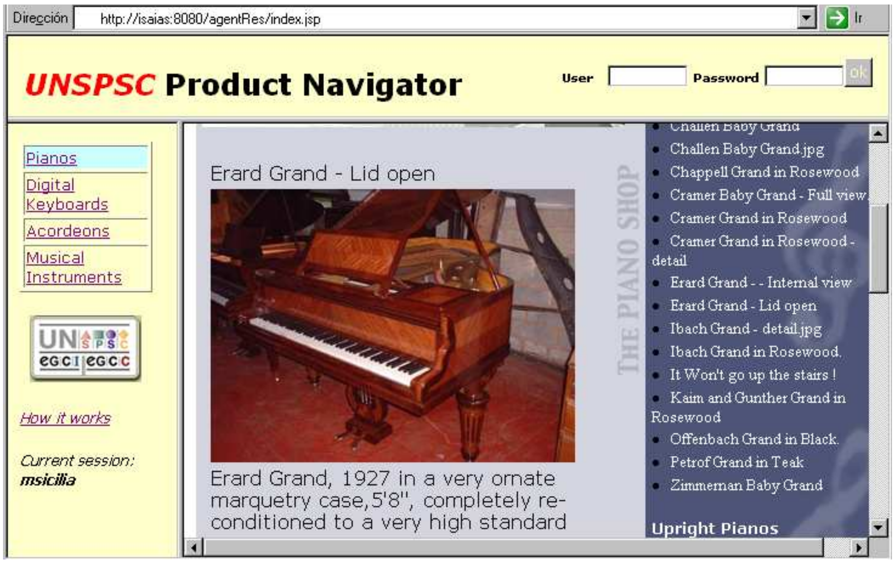
For example, if $piano \in C_U$ then the agent has two options to generate recommendations from the *gen-spec* hierarchy:

- Going 'up' to the more abstract term `Keyboard-instruments`
- Or going 'down' to specializations of `piano` (like spinet, console or grand pianos).

The agent would choose first the shorter distance (i.e. the larger resemblance). In the example, it would choose going 'down' (and sort the `piano` subclasses by descending resemblance), since the way up represents a bigger step, due to the highly abstract nature of the term `Keyboard-instruments`.

This behavior may prevent the agent to jump to excessively abstract categories (e.g. reaching the awkward `Musical Instruments, - Recreational Equipment, Supplies and Accessories` class mentioned above).

Other resemblance-filtering schemes can be implemented and studied, and existing ontologies can be easily annotated to support this approach, but we have found that ontology-level overall relations are difficult to acquire from users.



Dirección http://isaias:8080/agentRes/index.jsp

UNSPSC Product Navigator

Erard Grand - Lid open

Erard Grand, 1927 in a very ornate marquetry case,5'8", completely re-conditioned to a very high standard

Current session: msicilia

Fig. 1. Overall layout of the Web prototype that uses resemblance relations to navigate product categories.

4 Conclusions and Future Work

Current approaches to generalization in conceptual modelling and ontology engineering lack the notion of distance between classes and subclasses. This notion, in many cases, cannot be computed from the structural characteristics of the ontology (or model) due, for example, to incompleteness or ill-defined hierarchies.

We have described the notion of distance from a class to its subclasses, and how it can be represented through partial local resemblance (fuzzy) relations, and added to current Rdf-based ontology description languages. This distance notion can be assessed and applied easily to filtering processes, but further empirical user testing is needed to come up with a measure of its impact from the user-interaction perspective. Our approach can be considered as complementary to fuzzy subtyping schemes [4], in which uncertainty and/or partial truth about types of objects is considered.

Further research should address additional properties of *gen-spec* hierarchies, for example, the relationship between discriminators at various levels of the hierarchy, or the implications of the notion of distance in '*sibling*' classes inside the ontology.

References

1. AlGhamdi J, Elish M, Ahmed M (2002) A tool for measuring inheritance coupling in object-oriented systems. Information Sciences, 140(3-4):217–227
2. Beierle C (1995) Type inferencing for polymorphic order-sorted logic programs. In: Sterling L (editor) Proceedings of the Twelfth International Conference on Logic Programming. Springer-Verlag, Lecture Notes in Computer Science 2401:765–780
3. Brusilovsky P (2001) Adaptive hypermedia. User Modeling and User Adapted Interaction 11(1–2):87–110
4. Cao T H, Creasy P N (2000) Fuzzy types: a framework for handling uncertainty about types of objects. International Journal of Approximate Reasoning, Elsevier Science, 25(3):217–253
5. Chen G (1998) Fuzzy logic in data modeling: semantics, constraints, and database design. Kluwer Academic Publishers
6. Guarino N, Welty C (2000) Ontological analysis of taxonomic relationships. In: Laender A, Storey V (editors) Proceedings of the 19th International Conference on Conceptual Modeling. Springer-Verlag, Lecture Notes in Computer Science 1920: 210–224
7. Horrocks I (2002) DAML+OIL: a reason-able Web ontology language. In: Jensen, C S et al. (eds.) Proceedings of the 8th International Conference on Extending Database Technology. Springer-Verlag Lecture Notes in Computer Science 2287: 2–13
8. Liskov B, Wing J M (1994) A behavioral notion of subtyping. ACM Transactions on Programming Languages and Systems 16(6):1811–1841
9. McCarthy J, Hayes P (1969) Some philosophical problems from the standpoint of artificial intelligence. Machine Intelligence 4:463–502
10. Meinke K, Tucker J V (1993) Many-sorted logic and its applications. Wiley, Chichester
11. Norrie M C, Reimer U, Lippuner P, Rys M, Schek H J (1994) Frames, objects and relations: three semantic levels for knowledge-based systems. In: Baader F et al. (eds.) Proceedings of the Workshop on Reasoning about Structured Objects: Knowledge Bases meets Databases, CEUR Workshop Proceedings
12. Object Management Group (OMG) (2001) The unified modeling language specification, version 1.4, available at http://www.uml.org
13. Spanoudakis G, Constantopoulos P (1994) Similarity for analogical software reuse: a computational model. In: Proceedings of the 11th European Conference on Artificial Intelligence(ECAI '94),John Wiley and Sons: 18–22
14. World Wide Web Consortium (W3C) (1999) Resource Description Framework (RDF) Model and Syntax. W3C Recommendation, 22 February 1999, available at http://www.w3.org/TR/1999/REC-rdf-syntax-19990222
15. World Wide Web Consortium (W3C) (2001) XML Schema Part 2: Datatypes. W3C Recommendation, 2 May 2001, available at http://www.w3.org/TR/xmlschema-2/