Determining Semantic Similarity Among Entity Classes from Different Ontologies[*]

M. Andrea Rodríguez[124] and Max J. Egenhofer[123]

[1]National Center for Geographic Information and Analysis
[2]Department of Spatial Information Science and Engineering
[3]Department of Computer Science
University of Maine
Orono, ME 04469-5711, USA

[4]Department of Computer Science
Universidad de Concepción
Concepción, Chile

Abstract

Semantic similarity measures play an important role in information retrieval and information integration. Traditional approaches to modeling semantic similarity compute the semantic distance between definitions within a single ontology. This single ontology is either a domain-independent ontology or the result of the integration of existing ontologies. We present an approach to computing semantic similarity that relaxes the requirement of a single ontology and accounts for differences in the levels of explicitness and formalization of the different ontology specifications. A similarity function determines similar entity classes by using a matching process over synonym sets, semantic neighborhoods, and distinguishing features that are classified into parts, functions, and attributes. Experimental results with different ontologies indicate that the model gives good results when ontologies have complete and detailed representations of entity classes. While the combination of word matching and semantic neighborhood matching is adequate for detecting equivalent entity classes, feature matching allows us to discriminate among similar, but not necessarily equivalent, entity classes.

1. Introduction

With the growing access to heterogeneous and independent data repositories, the treatment of differences in the structure and semantics of the data stored in those repositories plays a major role in information systems. Since the first studies on interoperating information systems, progress has been made concerning syntactic (i.e., data types and formats) and structural heterogeneities (i.e., schematic integration, query languages, and interfaces) [1]. As interoperating information systems increasingly confront more complex knowledge management issues, the technology needed to deal successfully with these issues must focus on the semantics underlying the data used by those systems [2].

Recent investigations in information retrieval and data integration have emphasized the use of ontologies and semantic similarity functions as a mechanism for comparing objects that can be retrieved or integrated across heterogeneous repositories [3-7]. In this context, an ontology is a type of knowledge base that describes concepts through definitions that are sufficiently detailed to capture the semantics of a domain. An ontology captures a certain view of the world, supports

[*] This work was partially supported by the National Imagery and Mapping Agency under grant number NMA202-97-1-1023 and the National Science Foundation (NSF) under grant number IIS-9970123. Max Egenhofer's research is further supported by NSF grants IRI-9613646, SBR-9700465, BDI-9723873, and EIA-9876707; the National Institute for Environmental Health Sciences, NIH under grant number 1-R-01-ES09816-01; and by Lockheed Martin Management and Data Systems.

intensional queries regarding the content of a database, and reflects the relevance of data by providing a declarative description of semantic information independent of the data representation [8].

There is great variation among both the level of detail and logic of different ontology representations. For example, a terminological ontology is a collection of categories organized by a partial order that is induced by inclusion. Examples of such ontologies include the WordNet ontology for nouns [9] and the SENSUS ontology for machine translation [10]. A different and more detailed ontology is an axiomatized ontology. An axiomatized ontology is a terminological ontology whose categories are distinguished by axioms and definitions stated in logic or in some language that could be automatically translated into logic [11]. Examples of axiomatized ontologies include the GALEN core model [12], the PSL ontology [13], and Cyc [14].

Our current work is motivated by the need of new tools that can improve the retrieval and integration of information. In this work we focus on ontologies whose specification components include entity classes, semantic relations among these classes, and distinguishing features that describe these classes, while we leave for future work the treatment of more complex axiomatized ontologies. The term entity class refers to concepts that group entities or objects of the real world into classes of entities. Some examples of entity classes are the concepts of *building*, *lake*, and *city*. Although these entity classes may be associated with entities or classes in a database schema, they are usually richer in their semantics because they represent concepts independently of data representation or modeling.

In environments with multiple information systems, independent systems may have their own intended models and, therefore, their own ontologies [15]. In such environments, the general approach to data integration has been to map the local terms of distinct ontologies onto a single shared ontology. Then, the semantic similarity is typically determined as a function of the path distance between terms in the hierarchical structure underlying this single ontology [16-19]. Other methods to assess semantic similarity within a single ontology are feature matching [20] and information content [5, 21]. The feature-matching approach uses common and different characteristics between objects or entities to compute semantic similarity, and information content uses information theory [22] to define a similarity measure in terms of the degree of informativeness of the immediate super-concept that subsumes two concepts being compared.

The use of a single ontology does ensure complete integration across heterogeneous information systems; however, this type of ontology is costly if not impractical, since information systems are forced to commit to this single ontology and compromises are difficult to maintain when new concepts are considered. Using another approach, which considers scalability issues in building an ontology, some studies create a shared ontology by integrating existing ones [23-26]. Studies that pursue ontology integration have to treat overlapping concepts and inconsistencies across ontologies. Like semantic heterogeneity in the database field [27], ontology mismatches occur when two ontologies have different definitions but their terms that denote categories, the components of the category definitions, or the ontological concepts are the same [28].

A strategy for ontology integration is the mapping of local ontologies onto a more generic ontology [15, 26, 29]. For example, ONIONS [26] is a methodology for ontology analysis and integration that has been applied to large medical terminologies. Ontology integration in ONIONS is done by formally representing all concepts and by ontologically integrating these concepts through a set of generic ontologies. The use of semantic interrelations is another approach for ontology integration [25, 30]. For example, OBSERVER is an ontology-based system that is enhanced with relationships for vocabulary heterogeneity resolution [23, 24, 30]. It uses terminological relations (hyponymy and hypernymy) to map the non-translated terms in a user ontology onto terms (which are not synonymous) in a target component ontology. This translation process is recursive and consists of substituting non-translated terms with the intersection of their immediate parents or the union of their immediate children.

Once ontologies have been integrated, similarity measures are applied to compare concepts. A recent work presents different measures for comparing concepts whose formal definitions support inferences of subsumption, and local concepts in differentied ontologies inherit their definitional structures from concepts in a shared ontology [29]. This study assumes that the set intersection of concepts' instances is an indication of these concepts' correspondence. Three main types of measures for comparing concepts descriptions are discussed in this work: (1) filter measures based on a path distance, (2) matching measures based on graph matching that make one-to-one correspondence between elements of concepts' descriptions, and (3) probabilistic measures that give the correspondence in terms of the joint distribution of concepts.

Although there have been previous attempts to compare items from different ontologies, these studies are based on an integrated ontology derived from a manual or semi-automatic process. This work aims at creating a computational model that assesses semantic similar among entity classes from different and independent ontologies without constructing *a priori* a shared ontology. Our approach to modeling similarity is based on a matching process [20] that uses the available information from various ontology specifications (i.e., synonym sets, distinguishing features, and semantic relations of entity classes). Such similarity modeling establishes links among ontologies while keeping each ontology autonomous. It is a weak form of integration, because it does not allow deep processes, that is, it cannot be used for making inferences about the relationship among other entity classes within a given ontology and cannot guarantee computations that require particular components of the entity class representation. It provides, on the other hand, a systematic way to detect which entity classes are most similar to each other and, therefore, which entity classes are the best candidates for establishing an integration across the ontologies. Our measure of similarity could be used as a first step in a strong integration of ontologies with user input as refinements. It is also useful in dynamic environments, such as the World Wide Web (WWW), where it may be impractical to force users to subscribe *a priori* to a shared ontology. In such environment, an agent may request information specified as a concept description, and broker agents that know about the information available in the WWW space will compare and recommend possible candidates to respond the request.

The remainder of this paper is structured as follows: the description of the entity class representations in Section 2 is followed by the presentation of the components of similarity assessment in Section 3. Section 4 explains the matching-based approach to modeling similarity and defines a similarity function for cross-ontology evaluations. An evaluation of the model using different ontologies and a human-subject experiment is presented in Section 5. Conclusions and future work are presented in Section 6.

2. Entity Class Representation

In a previous work, we define three basic components for the representation of entity classes in an ontology: (1) a set of synonym words (synset) that denotes an entity class, (2) a set of semantic interrelations among these entity classes, and (3) a set of distinguishing features that characterize entity classes [31]. The use of a set of words to denote entity classes addresses the issue of polysemy and synonymy in the process of linking words to meaning. Polysemy occurs when the same word denotes more than one meaning, and synonymy occurs when different words denote the same or very similar entity classes [32, 33]. For example, while the word *bank* can denote more than one concept (e.g., a *financial institution*, a *building* of a financial institution, or a *sloping land*), the set of synonyms constituted by *bank, banking company, depository financial institution* identify a unique concept (i.e., a financial institution that accepts deposits and channels the money into lending activities).

Two semantic relations play an important role in the specification of ontologies. Hyponymy, also called the is-a relation [34], is the most common relation used in an ontology. This relation goes from a specific to a more general concept. The is-a relation is transitive and asymmetric and defines a hierarchical structure where terms inherit all the characteristics from their superordinate terms. Meronymy is a partial ordering of concept types by the part-whole relation

[35]. Studying the transitive property of part-whole relations, researchers have argued that part-whole relations are not one type of relations, but a family of relations, and that transitive property holds for some but not all of these part-whole relations [36-38].

Properties that distinguish entity classes from the same superclass are called distinguishing features or differentiae [11]. Although the general organization of entity classes is given by their semantic interrelations, this information alone may be insufficient to distinguish one class from another. For example, a *hospital* and an *apartment building* have a common superclass *building*; however, this information falls short when trying to differentiate a *hospital* from an *apartment building*, since the is-a relation does not indicate the important difference in terms of the entity classes' functionality (i.e., a *hospital* is a building where medical care is given and an *apartment building* is a group of apartments that serves as living quarters).

Usually, attributes describe different types of distinguishing features of a class. They provide the opportunity to capture details about classes, and their values describe the properties of individual objects (i.e., instances of a class). We suggest a finer identification of distinguishing features than the typical single classification of features into attributes, and we classify them into functions, parts, and attributes. Functions are intended to represent what is done to or with instances of a class. For example, the function of a *college* is to educate. Thus, function features can be related to other terms such as affordances [39] and behavior [40]. Parts are structural elements of a class, such as the roof and floor of a building. While the part-whole relations work at the level of entity class representations, part features can have items that are not always defined as entity classes in an ontology. For example, although roof and floor are part features of a *building*, they may not be necessarily defined as entity classes in the model and, therefore, they are connected to a *building* through a part-of relation. Treating part functions different from part-of relations is need and not an option; otherwise, we will produce an endless process of defining entity class. Finally, attributes correspond to additional characteristics of an entity class that are not considered by either the set of parts or functions. For example, some of the attributes of a building are age, user type, owner type, and architectural properties.

We identify different types of distinguishing features (i.e., parts, functions, and attributes) to enable the separate manipulation of them. This distinction, however, has the drawback of articulating new types of mismatches associated with the classification of features. While an ontology may group all features under attributes, as does the Spatial Data Transfer Standard [41], another ontology may classify them into attributes and parts, such as WordNet [9]. Although new types of mismatches may occur, we propose the classification of distinguishing features in order to support a comparison between corresponding characteristics of entity classes. Another benefit of considering different types of distinguishing features is that weights could be assigned to these types of distinguishing features to reflect how important they are in particular contexts [42].

Our representation of entity classes can be clearly associated with the definition of classes in the object-oriented paradigm. Is-a and part-whole relations are extracted from basic paradigms of the object-oriented theory (inheritance and composition, respectively), while the distinguishing features of our entity class representation, with the exception of parts, could be associated with attributes or methods of classes in object orientation. A formal syntax of an entity class definition using BNF notation is presented in Table 1 together with an example of the definition of the entity class *stadium* derived from a combination of WordNet and SDTS. In this specification, primitives of our language are pointers and words.

BNF Notation	**Example: *Stadium***
<entity_class>::= **entity_class {**	**entity_class {**
name: {<syn_set>}	**name:** {stadium,bowl,arena}
description: <description>	**description:** large often unroofed structure in which athletic events are held
is_a: <is-a>	**is_a:** {construction*}
part_of: <part_of>	**part_of:** {}
whole_of: <whole_of>	**whole_of:** {athletic_field*}
parts: <parts>	**parts:** {{athletic_field,sports_field,playing_field}, {dressing_room},{foundation}, {midfield},{spectator_stands,stands}, {ticket_office, box_office,ticket_booth}}
functions: <functions>	**functions:** {{play,compete},{play,practise}, {recreate,play}}
attributes: <attributes>**}**	**attributes:** {{architectural_property}, {covered/uncovered}, {name}, {lighted/unlighted}**,**{owner_type}, {sports_type},{user_type}}**}**
<is_a>::= {}|{<pts_entity_classes>}	
<part_of>::= {}|{< pts_entity_classes >}	
<whole_of>::= {}{< pts _entity_classes >}	
<parts>::= {}|{<syn_sets>}	
<functions>::= {}|{<syn_sets>}	
<attributes>::= {}|{<syn_sets>}	
<syn_sets>::= {<syn_set>}|<syn_sets>,{syn_set}	
<syn_set>::= <word>|<syn_set>,<word>	
<description>::= <word>|<description> <word>	
<pt_to_entity_classes>::= <pointer>|<pt_to_entity_classes>, <pointer>	

Table 1: Entity_class definition in BNF notation and an example with the definition of *stadium*. (*x** denotes a pointer to the entity class *x*)

3. Comparing Entity Class Representations

Different levels of explicitness and formalization of the ontologies influence the way entity classes can be compared. Similarity evaluations across ontologies can only be achieved if their representations of entity classes share some components. A natural way to exploit the full expressiveness of the entity class representations for a similarity evaluation is to compare each component in those representations. Thus, two different ontologies that have at least one common specification component can still be compared.

Our approach to comparing entity classes across ontologies uses three independent similarity assessments with respect to synonym sets that denote entity classes, distinguishing features of entity classes, and semantic relations among entity classes. Synonym sets are themselves groupings of semantically equivalent or very similar words [9]. Thus, our model considers synonyms as the same entity class and that the similarity between an entity class and itself is always maximum. The goal of comparing synonym sets in a cross-ontology evaluation is to exploit the general agreement in the use of words and detects equivalent words that likely refer to the same entity class. Thus, similar synonym sets can only be used to detect equivalent or synonym entity classes across ontologies. As such it provides a very basic level of similarity assessment, which is an inconclusive form of similarity assessment, since words can be quiet different, while the entity classes can still be semantically similar. An example is *clinic* and *hospital*, which have only a few characters in common and, therefore, their string similarity is very low. Their semantic similarity, however, is fairly high. Inversely, words in synonym sets can be the same, whereas the corresponding entity classes are semantically unrelated.

Incorporating semantics into the similarity measure, we can use distinguishing features as another indicator of how similar entity classes are. Unlike synonym-set similarity with a binary resolution of similarity (same or different words), a feature similarity handles grades of similarity, since semantically similar entity classes with quite different names are likely to have some common features. For example, knowing that *stadium* and *sports_arena* are places where people can play a sport makes these two concepts similar. Diverse feature-based models for semantic similarity [20, 43-45] that have pointed the need for considering context dependence of the relative importance of distinguishing features and asymmetric characteristics of similarity assessments.

Our approach treats semantic relations themselves as the subject of comparison. Since the types of semantic relations are known (e.g., is-a or part-whole relations), the interesting aspect of comparing semantic relations is whether target entity classes (i.e., entity classes that are the subject of comparison) are related to the same set of entity classes. If target entity classes are related to the same set of entity classes, they may be semantically similar. For example, *hospital* and *house* are related to the same superclass *building*, and they are semantically similar. Thus, comparing semantic relations becomes a comparison between the semantic neighborhoods of entity classes.

The semantic neighborhood (N) of an entity class a^o is the set of entity classes c_i^o whose distance $d()$ to the entity class a^o is less than or equal to a non negative integer r, called the radius of the semantic neighborhood (Equation 1).

$$N(a^o, r) = \left\{c_i^o\right\} \text{ such that } \forall i \ d(a^o, c_i^o) \leq r \qquad (1)$$

The distance between two entity classes in the ontology is measured along the shortest path, which is formed by the smallest number of undirected arcs that connect the entity classes. These arcs represent subclass-superclass or part-whole relations, and so the shortest path can represent two sorts of hierarchical relationships. Since distance is a metric function that satisfies the property of minimality (i.e., the self-distance is equal to zero), the semantic neighborhood of an entity class also contains this entity class. For example, the immediate semantic neighborhood (i.e., semantic neighborhood of radius 1) of *stadium* in a portion of the WordNet ontology includes the *stadium*, its superclass *structure* and, its parts *athletic field* and *sports arena* (Figure 1).

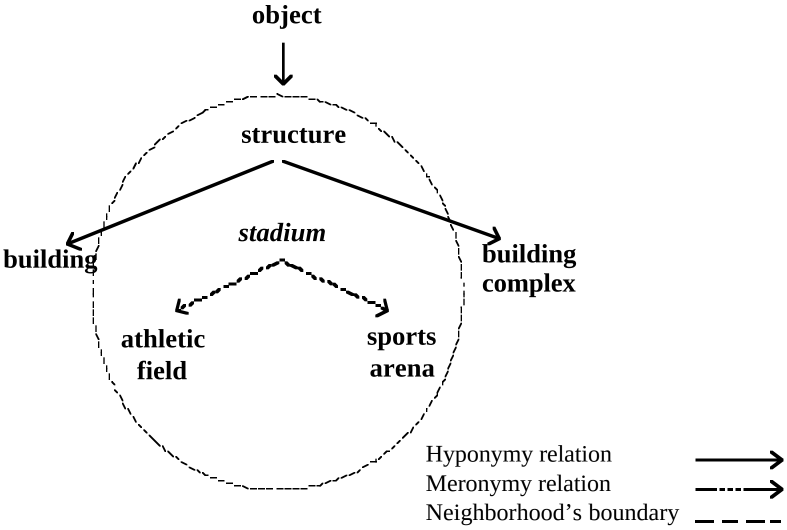


Figure 1: Example of the immediate semantic neighborhood of stadium in a portion of the WordNet Ontology.

There exit arguments against the use of path distance in similarity assessments [7, 21], which have been addressed by considering weighted indexing schema and variable edge weights [5, 46]. Although we use path distance to identity the semantic neighborhood of entity classes within their own ontologies, we do not define the similarity measure between neighborhoods based on this path distance. Path distance determines the neighborhoods, and the similarity of entity classes depends on the similarity of the entity classes in their neighborhoods.

In order to integrate the information obtained from the similarity assessments of synonym sets, distinguishing features, and semantic neighborhoods, we propose a similarity function that is defined by the weighted sum of the similarity of each specification component (Equation 2). The functions S_w, S_u, and S_n are the similarity between synonym sets, features, and semantic neighborhoods between entity classes *a* of ontology *p* and *b* of ontology q, and ω_l, ω_u, and ω_n are the respective weights of the similarity of each specification component.

$$S(a^p,b^q) = \omega_w \cdot S_w(a^p,b^q) + \omega_u \cdot S_u(a^p,b^q) + \omega_n \cdot S_n(a^p,b^q) \quad \text{for } \omega_l, \omega_u, \text{and } \omega_n \geq 0 \qquad (2)$$

Weights assigned to S_w, S_u, and S_n depend on the characteristics of the ontologies. Only common specification components can be used in a similarity assessment and their respective weights add up to 1.0. Similarity of synonym sets can always be a factor of the similarity assessment, but when polysemous terms occur within an ontology, this similarity is less likely an indication of semantic similarity among entity classes. For example, one ontology may include different meanings of the word *bank* (e.g., a financial institution, a sloping of land, and a building), whereas another ontology may contain only one meaning of *bank* (e.g., a financial institution). Measuring only similarity of synonym sets, we would assign maximum similarity between each of the meanings of *bank* in the first ontology and the single meaning of *bank* in the second ontology, which is clearly incorrect. Similarity of synonym sets complemented with feature and semantic-neighborhood similarity, on the other hand, can highlight the similarity between corresponding senses of the term bank. Through experiments, Section 6 attempts to analyze the best setting of weights.

4. A Matching Approach to Similarity Assessment

Using set theory, Tversky [20] defined a similarity measure in terms of a matching process. This measure produces a similarity value that is not only the result of the common, but also the result of the different characteristics between objects, which is in agreement to an information-theoretic definition of similarity [47]. Unlike traditional models based on semantic distance [48], the matching model is not forced to satisfy metric properties (i.e., minimality, symmetry, and triangle inequality). Thus, for example, the similarity between an *office building* and a *building* can be greater than the similarity between a *building* and an *office building* (i.e., an asymmetric evaluation). Although an *athletic field* is similar to a *stadium* (because both are sports facilities) and a *stadium* is similar to a *theater* (because both are constructions where people go to attend events), an *athletic field* and a *theater* are not necessarily similar (i.e., a non-transitive evaluation).

A similarity measure based on the normalization of Tversky's model and the set-theory functions of intersection ($A \cap B$) and difference (A / B) is given in Equation 3, where a and b are entity classes; A and B corresponds to description sets of a and b (i.e., synonym sets, set of distinguishing features, of set of entity classes in the semantic neighborhood); | | is the cardinality of a set; and α is a function that defines the relative importance of the non-common characteristics.

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a,b)\,|A/B| + (1-\alpha(a,b))\,|B/A|}, \text{ for } 0 \leq \alpha \leq 1 \qquad (3)$$

The relative importance of the non-common characteristics (shown in the second and third terms of the denominator on the right hand side of Equation 3) allows the asymmetric evaluation of semantic similarity. Incorporating such an asymmetric measure is important because if we want to make similarity evaluations sensible to people judgments, we have to consider cognitive properties of similarity. In this sense, studies have shown that the perceived similarity from a class to its superclass is greater than the perceived similarity from the superclass to the class, and that the superclass is commonly used as *base*[1] of the similarity evaluation [44, 49]. There have given different explanations for the asymmetric evaluations of similarity. Asymmetry can be explained by

[1] The first term of a comparison is referred to as the target and the second term as the base.

the relative size and salience of distinctive features sets [20], by potential stimulus biases, such as density and prototypicality [44, 50], by a natural reference point or landmark for members of a category [49], and by the direction of maximum informativiness [51]. Common to all these explanations is the different role that the *target* and *base* positions play in a similarity evaluation. The most salient term, the item with larger bias, the prototypical term, or the term that provides information to understand the *target* are always in the *base* position.

Our work considers that a prototype used as a *base* of a similarity evaluation is a more general concept in a hierarchical structure and that the perceived similarity from a class to its superclass (i.e., a more general concept) is greater than the perceived similarity from this superclass to the class. Thus, the common, as opposed to the different, component definitions between a class with respect to its superclass have a larger contribution to the similarity evaluation than the common components in an inverse direction. A natural approach to comparing the degree of generalization between entity classes is to determine the distance from these entity classes to the immediate superclass that subsumes them, that is, their least upper bound in a partially ordered set [52]. In a cross-ontology evaluation, however, there is no such common superclass between entity classes. An approximation to obtain the level of generalization of entity classes is to consider that the two independent ontologies are connected by making each of their roots a direct descendant of an imaginary and more general entity class *anything* (Figure 2).

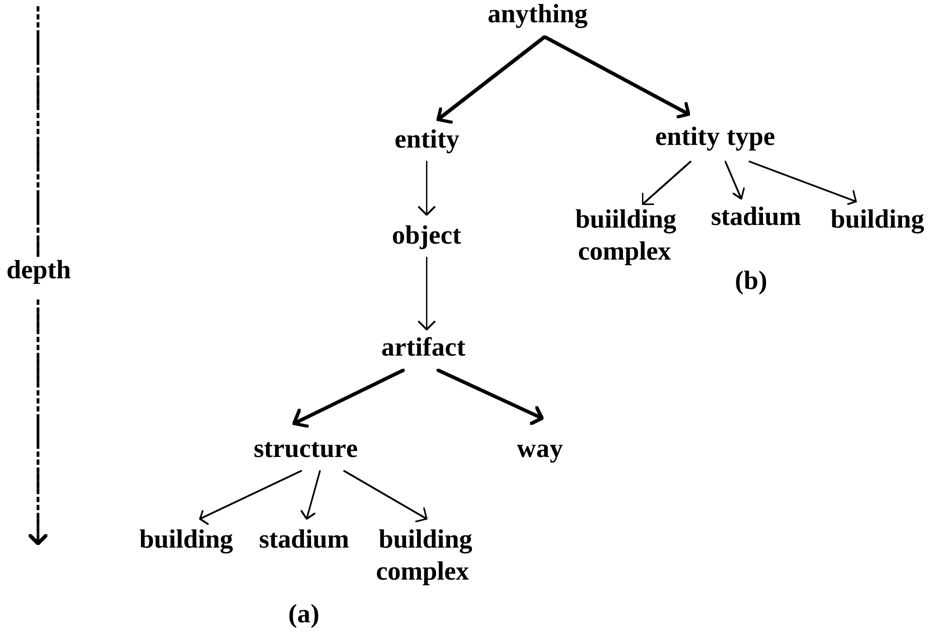


Figure 2: Connecting independent ontologies: (a) partial WorNet ontology and (b) partial SDTS ontology. (*Anything** denotes an imaginary root)

Using this connected ontology, the function α of the matching model can be expressed in terms of the *depth* of the entity classes (Equation 4). The function *depth*() corresponds to the shortest path from the entity class to the imaginary root. This *depth* reflects the degree of granularity upon which the ontology was designed. For example, consider the ontologies in Figure 2. While WordNet's hierarchy has multiple levels, SDTS defines a large number of concepts that are unrelated, which yields a shallow hierarchy. When *building* in WordNet ($building^w$) is compared to *building* in SDTS ($building^s$), $depth(building^w)$ is 5 whereas $depth(building^s)$ is 2, such that α (*building*^{w}, *building*^{s}) is 0.28. With this definition of α, evaluations from deep to shallow ontologies usually result in greater values of similarity than evaluations from shallow to deep ontologies.

$$\alpha(a^p,b^q)=\begin{cases}\dfrac{depth(a^p)}{depth(a^p)+depth(b^q)} & depth(a^p)\le depth(b^q)\\[2ex] 1-\dfrac{depth(a^p)}{depth(a^p)+depth(b^q)} & depth(a^p)>depth(b^q)\end{cases} \tag{4}$$

Values of α are greater than 0 and less than or equal to 0.5, which corresponds to the case when entity classes have the same depth in their respective hierarchies. The non-common characteristics between entity classes are considered less important than the common characteristics (α and 1–α are less than 1), because we follow the finding that subjects pay more attention to the similar than to the different characteristics in a similarity assessment [20, 44].

Using this matching model, we then define similarity functions for each of the components of the entity class representation (i.e., we define the elements of the set intersection of Equation 3), which we have called word matching, feature matching, and semantic-neighborhood matching.

4.1 Word Matching

Word matching (S_w) checks the number of common and different words in the synonym sets that denote entity classes. For the ontologies in Figure 2, the word matching between building of WordNet (buildingw) and building complex of SDTS (building_complexs) is 0.58 for α = 0.28 (Equation 5). Likewise, word matching between *stadium*^{w} and *stadium*^{s} results in 1.0, independent of the value for α.

$$\begin{aligned} S_l(building^w, building\_complex^s) &= \frac{|\{building\}|}{|\{building\}|+0.28\,|\{\}|+0.72\,|\{complex\}|} \\ &= \frac{1}{1.72}=0.58 \end{aligned} \tag{5}$$

In cases when more than one word exit in the respective synonym sets of entity classes, word matching finds the most similar terms between synonym sets. For example, if *edifice* is used as a synonym for *building* in the WordNet ontology (Figure 2), then the word matching between *edifice*^{w} and *building_complex*^{s} is 0.58, which is the highest value of word matching between $S_l(edifice^w, building\_complex^s)$ and $S_l(building^w, building\_complex^s)$.

4.2 Feature Matching

Feature matching (S_u) applies a matching process over corresponding types of distinguishing features such that *A* and *B* of Equation 3 are the sets of features of entity classes *a* and *b*, respectively. When both ontologies classify features into parts, functions, and attributes, the feature matching is given by Equation 6, where S_p, S_f, and S_a are the similarity measures of parts, functions, and attributes, respectively, and ω_p, ω_f, and ω_a are their corresponding weights. This Equation 6, represents a refinement in the level of detail of feature similarity (S_w in Equation 2), since it establishes a composition of feature matching by subtype of features. By default, the types of distinguishing features that are present in the specifications of ontologies are considered equally important (i.e., $\omega_p=\omega_f=\omega_a=0.33$). In a previous work, we discussed how contextual information can be used to determine weights of distinguishing features as a function of the degree of *informativeness* or *diagnosticity* of the features within the domain of an application [42]. When no classification of distinguishing features is given, a global feature-matching process is performed, that is, all distinguishing features are considered of the same type.

$$S_u(a^p,b^q) = \omega_p \cdot S_p(a^p,b^q) + \omega_f \cdot S_f(a^p,b^q) + \omega_a \cdot S_a(a^p,b^q)$$
$$\text{for } \omega_p,\ \omega_f, \text{ and } \omega_a \geq 0 \text{ and } \omega_p + \omega_f + \omega_a = 1.0 \qquad (6)$$

In this work we have made a lexicographic, rather than semantic, representation of distinguishing features. Thus, a distinguishing feature is represented by a synonym set, and the feature matching process applies a string-matching operation over the words in these synonym sets that refer to the features. String matching over distinguishing features is a strict string matching in the sense that distinguishing features match only if they are represented by the same word or by synonym sets that intersect. This process ignores similarity between compound terms, such as between lane and number of lanes. A major virtue of such strict string matching is a fast comparison of feature names for large ontologies where the percentage of partial string matching among feature names is limited.

To see in detail how we assess the similarity of distinguishing features, we present an extended example. Consider the definitions of *stadium* in WordNet (*stadium*^{w}) and our *ad-hoc* ontology WS (*stadium*^{ws}). While WS identifies parts, functions, and attributes of entity classes, WordNet has only parts and, therefore, feature matching is confined to the comparison among parts of entity classes (Table 2).

Stadium (WS)	**Stadium (WordNet)**
entity_class {	**entity_class {**
name: {stadium,bowl,arena}	**name:** {stadium,bowl,arena}
description: large often unroofed structure in which athletic events are held	**description:** large often unroofed structure in which athletic events are held
is_a: {*construction**}	**is_a:** {*construction**}
part_of: {}	**part_of:** {}
whole_of: {*athletic_field**}	**whole_of:** {*athletic_field**, *sports_arena**}
parts: {{athletic_field,sports_field,playing_field}, {dressing_room},{foundation}, {midfield},{spectator_stands,stands}, {ticket_office, box_office,ticket_booth}}	**parts:** {{athletic_field,sports_field,playing_field}, {foundation},{midfield},{plate}, {sports_arena,field_house},{stands}, {structural_elements}, {standing_room},{tiered_seats}}
functions: {{play,compete},{play,practise}, {recreate,play}}	**functions:** { }
attributes: {{architectural_property}, {covered/uncovered}, {name}, {lighted/unlighted},{owner_type}, {sports_type},{user_type}}**}**	**attributes:** {}**}**

Table 2: Entity_class definition of *stadium* in WS and WordNet. (*x** denotes a pointer to the entity class *x*)

Distinguishing features in both ontologies are denoted by synonym sets. We say that two distinguishing features are equivalent if the intersection of their synonym sets is not empty. Thus, between *stadium*^{w} and *stadium*^{ws} there are four no-empty synonym sets (i.e., four common features) (Equations 7).

$$X = stadium^{ws}.\text{parts} \cap stadium^{w}.\text{parts} = \{\{athletic\_field, playing\_field, field\}, \{foundation\}, \{midfield\}, \{stands\}\} \qquad (7)$$

The set difference between features of *stadium*^{w} and *stadium*^{ws}, or vice versa, is defined by the set of features that belong to *stadium*^{w} and not to *stadium*^{ws} . Thus, there are five parts in *stadium*^{w} that are not in *stadium*^{ws} and, inversely, there are two parts in *stadium*^{ws} that are not *stadium*^{w} (Equations 8a-b).

$$Y = \textit{stadium}^{w}.\text{parts} - \textit{stadium}^{ws}.\text{parts} = \{\{\textit{plate}\},\{\textit{sports\_area},\textit{field\_hourse}\},\{\textit{standing\_room}\},\{\textit{structural\_elements}\},\{\textit{tiered\_seats}\}\} \quad (8a)$$

$$Z = \textit{stadium}^{ws}.\text{parts} - \textit{stadium}^{w}.\text{parts} = \{\{\textit{dressing\_room}\},\{\textit{ticket\_office},\textit{box\_office},\textit{ticket\_booth}\}\} \quad (8b)$$

The similarity measure between distinguishing features of *stadium*^{w} and *stadium*^{ws} is then determined by Equation 9 for α equal to 0.45. This equation is equivalent to Equation 3 when *A* and *B* are replaced by the set of parts of *stadium*^{w} and *stadium*^{ws}, respectively.

$$\begin{aligned} S_u(stadium^{w}, stadium^{ws}) &= S_p(stadium^{w}, stadium^{ws}) \\ &= \frac{|X|}{|X| + 0.45|Y| + 0.55|Z|} = \frac{4}{4 + 0.45*5 + 0.55*2} = 0.54 \end{aligned} \quad (9)$$

4.3 Semantic-Neighborhood Matching

Semantic-neighborhood matching (S_n) compares entity classes in semantic neighborhoods based on synonym_set or feature matching. Semantic-neighborhood matching (S_n) with radius *r* between entity classes a^p and b^q of ontologies *p* and *q*, respectively, is a function of the cardinality (| |) of the semantic neighborhoods (*N*) and the approximate cardinality of the set intersection ($\cap_n$) between these semantic neighborhoods (Equation 10).

$$S_n(a^p,b^q,r) = \frac{a^p \cap_n b^q}{a^p \cap_n b^q + \alpha(a^p,b^q)\cdot\delta(a^p,a^p \cap_n b^q,r) + (1-\alpha(a^p,b^q))\cdot\delta(b^q,a^p \cap_n b^q,r)} \quad \text{with}$$

$$\delta(a^p,a^p \cap_n b^q,r) = \begin{cases} |N(a^p,r)| - |a^p \cap_n b^q| & \text{if } |N(a^p,r)| > |a^p \cap_n b^q| \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

The intersection over semantic neighborhoods is approximated by the similarity of entity classes across neighborhoods (Equation 11), where *S*() is the semantic similarity of entity classes; a_i^p and b_j^q are entity classes in the semantic neighborhood of a^p and b^q, respectively; and *n* and *m* are the numbers of entity classes in the corresponding semantic neighborhoods.

$$|a^p \cap_n b^q| = \left[\sum_{i\le n} \max_{j\le m} S(a_i^p,b_j^q)\right] - \varphi S(a^p,b^q) \text{, where}$$

$$\varphi = \begin{cases} 1 & \text{if } S(a^p,b^q) = \max_{j\le m} S(a^p,b_j^q) \\ 0 & \text{otherwise} \end{cases} \quad \text{and}$$

$$S(a_i^p,b_j^q) = \omega_l^{'} S_l(a_i^p,b_j^q) + \omega_u^{'} S_u(a_i^p,b_j^q) \text{ with } 0 < \omega_l^{'} + \omega_u^{'} \le 1 \quad (11)$$

Since *S*() in Equation 11 is an asymmetric function, the approximate cardinality of the set-intersection is also asymmetric. The approximate set intersection matches entity classes with maximum similarity. This matching excludes the similarity between the two entity classes that are actually being compared, which would be a redundant evaluation. It allows multiple entity classes in a semantic neighborhood to match the same entity class in a second semantic neighborhood. Thus, the approximate cardinality set intersection may reach a value greater than the actual cardinality of the set of entity classes in the second semantic neighborhood. In such a case, the model considers the maximum between the approximate cardinality of the set intersection and the cardinality of the semantic neighborhood. No matching between entity classes of the same role (i.e., superclass-superclass or subclass-subclass) is enforced, because this type of correspondence emphasizes similarity among classes with the same superclass while ignoring similarity between classes and their superclasses.

For example, consider WordNet and SDTS and the evaluation between $stadium^w$ and $stadium^s$ (Figure 2). In a first instance, we consider a radius of 1 and compare how many entity classes in the immediate neighborhood (i.e., immediate superclasses, subclasses, parts, and wholes) are common between $stadium^w$ and $stadium^s$ (Equations 12a-b). Semantic-neighborhood matching takes each entity class in $N(stadium^w,1)$ and finds the corresponding most similar entity class in $N(stadium^s,1)$. Based on word and feature matching, the only similar entity classes in the neighborhoods $N(stadium^w,1)$ and $N(stadium^s,1)$ are $stadium^w$ and $stadium^s$ entity classes themselves, which are the original entity classes that are compared. In this case, φ is zero and, therefore the semantic-neighborhood matching is also equal to zero.

$$N(stadium^w,1) = \{stadium^w, structure^w, athletic\_field^w, sports\_arena^w\} \tag{12a}$$

$$N(stadium^s,1) = \{stadium^s, entity\_type^s\} \tag{12b}$$

Analogous to the notion of *shallow* and *deep* equality in object orientation [53, 54], semantic-neighborhood matching defines shallow and deep matching depending on the radius of the semantic neighborhood. Shallow matching corresponds to an evaluation that is based on the similarity of the immediate neighborhood of entity classes (i.e., radius is 1). For semantic neighborhoods with radius greater than 1, deep matching is the evaluation that is based on the similarity of the end nodes (i.e., leaves) of the semantic neighborhood. These nodes are the entity classes located at the end of the path in the network of semantic relations that connect the entity classes in the semantic neighborhood. A similar notion of shallow and deep could be applied to the feature matching among parts if we had used a semantic evaluation there instead of a lexicographic evaluation.

5. Cross-Ontology Evaluations

There are few studies that have addressed the quality of results of similarity assessments. In cases of evaluations within a single ontology, these studies analyze the correlation between the computational similarity and answers of a human-subject testing [5, 21, 55]. For cross-ontology evaluations, however, no work has attempted to correlate computational similarity with people's judgments. In the context of cross-ontology evaluations, quality of evaluations has been addressed on the basis of an intensional or extensional analysis of query expansion to multiple ontologies. OBSERVER [30] uses intensional as well extensional analysis to define lower and upper bounds of query expansion based on a manually defined subsumption relation. In an effort to creating an environment to study algorithms that compute description compatibility, Weinstein and Birmingham [29] used an automatic generation of ontologies and compared different measures for determining semantic compatibility, which they define as the probability that an instance of a recommended answer satisfies a request. Unfortunately, it is unclear the generality of their results due to the unrealistic scenarios taking from the automatic generation of ontologies.

We designed new experiments that differ significantly from the previous experiments. First, we have a model for similarity evaluations across independent ontologies that are not linked to a top level shared ontology. Second, the model creates automatically, as opposed to manually, associations across ontologies. Third, our experiments use already available ontologies (i.e., WordNert and SDTS) that differ in their ontology specification as well as level of specificity of their intended purposes (i.e., general versus specific domain). Forth, we use a human-subject testing that defines sensible results of the computational model for evaluations among semantically related entity classes. Finally, we use an intensional approach (i.e., comparing only entity class definitions rather than instances of classes) in our experiments.

Our application work is focused on the spatial domain so, our experiments employ subsets of the two readily available resources, WordNet (334 definitions) [9] and SDTS (498 definitions) [41] that deal with spatial concepts. WordNet is a widely used terminological ontology [4, 56-58] that organizes concepts in sets of synonyms (synsets) connected by semantic relations. It contains approximately 118,000 words organized into 90,000 sets of synonyms, which are semantically interrelated depending on their syntactic category. SDTS was created to provide a common classification and definitions of spatial features used in processes of spatial data transfer. It contains a set of entity types (approximately 200 standard terms and 1300 "included" terms) and their corresponding attributes. We selected all the standard terms of SDTS plus included terms that match terms in the WordNet ontology. From WordNet, we selected all entity classes whose names match terms in SDTS. Although the selection of concepts based on word matching already establishes a degree of similarity between concepts, our experiments will show that word matching is useful, but insufficient, to identify corresponding entity classes across ontologies.

Finally, we create a new ontology WS (257 definitions) from the combination of WordNet and SDTS (WS) to exploit a more complete definition of entity classes (i.e., semantic relations as well as distinguishing features). This new ontology has less entity classes than the union of WordNet and SDTS, since we group some of the intermediate entity classes in the hierarchical structure derived from WordNet that have just one subclass. To this new WS, SDTS brings the list of entity classes to be defined, their partial definition via is-a relations, and their attributes. WordNet complements these definitions with synonym, part-whole, and is-a relations. In addition, functions in the WS definitions were derived from verbs explicitly used in the natural-language description of entity classes, augmented by common sense. Since the ontologies used in these experiments vary in terms of domain (i.e., general vs. specific) and specificity (semantic relations vs. distinguishing features) (Table 3), the potential conclusions of these experiments can provide a good indication of the behavior of the similarity model when used with such different kinds of ontologies.

Characteristcs	SDTS	WordNet	WS
Words			
Synonymy		√	√
Polysemy	√	√	√
Relations			
Is-a	√	√	√
Part-of		√	√
Whole-of		√	√
Features			
Parts		√	√
Functions			√
Attributes	√		√

Table 3: Characteristics of the specification components of SDTS, WordNet, and WS.

Two types of experiments were performed that correspond to two different goals: (1) search for equivalent or most similar entity classes across ontologies and (2) rank similarity between an entity class in one ontology and a set of entity classes in a second ontology. The first type of similarity evaluation is useful for ontology integration, since most similar entity classes are the best candidate for such integration. The second experiment analyzes how well the model performs for finding similar, and not necessarily the most similar, entity classes across ontologies. This type of calculation is useful for information retrieval, because it provides a range of possible answers depending on conceptually similar terms and gives the users the possibility to choose among them. For example, consider the case of a user who is looking for a stadium in a certain location. A system can search in one or multiple resources and find that there is not only a *stadium*, but also other kinds of sports facilities, such as an *athletic field* or a *tennis court*. To do so, the system should be able to calculate semantic similarity and give the user a set of ranked answers. To run these experiments, a prototype of the similarity model was implemented in C++.

5.1 Experiment 1: Equivalent or Most Similar Entity Classes

The experiment was done by using different combinations of ontologies in cross-ontology evaluations (Table 4). These combinations correspond to diverse grades of similarity among entity classes and components of the entity class representations. They include identical ontologies (1-2), ontology and sub ontology (3), overlapping ontologies (4), and different ontologies (5).

Case	Ontology-Ontology	Description
1	WordNet-WordNet	Same ontology with is-a and part-whole relations
2	SDTS-SDTS	Same ontology with is-a relations and attributes
3	WordNet-WordNet*	Subset with same specification components
4	WordNet*-WS	Overlapping semantic relations and attributes
5	WordNet*-SDTS*	Different ontologies and specification components

Table 4: Cases of cross-ontology evaluations. Symbol * denotes small subsets of the initial ontology.

Analogously to standard evaluation measures in information retrieval based on the relevance of data retrieved [59], we adapt the concepts of *recall* and *precision* to evaluate the results of the model. For this work, *recall* corresponds to the proportion of similar entity classes that are detected by the model (Equation 13a), while precision is the proportion of entity classes detected by the model that are actually similar (Equation 13b), where *A* is the set of similar entity classes; *B* is the set of similar entity classes calculated by the model; and | | is the counting measure.

$$recall = \frac{|A \cap B|}{|A|} \tag{13a}$$

$$precision = \frac{|A \cap B|}{|B|} \tag{13b}$$

A critical issue for calculating recall and precision is to know what entity classes are in fact similar, which corresponds to the idea of knowing the relevance of data in information retrieval. This determination is simplified by the fact that we want to detect synonyms or equivalent entity classes. For example, building in WordNet ($building^w$) is similar to building ($building^s$) and building_complex ($building\_complex^s$) in SDTS; however, only $building^w$-$building^s$ is considered, because this pair has the highest similarity.

In the first two evaluations (i.e., WordNet-WordNet and SDTS-SDTS), each entity class in the first ontology has its corresponding entity class in the second ontology, since we compare the ontologies with themselves and we expect to obtain the highest value of recall and precision (i.e., an upper bound for cases with equivalent components of entity class specification). When the definitions in the first ontology are a superset of the definitions in the second ontology (i.e., WordNet-WordNet*), the model should find the corresponding entity classes of the sub-ontology in the super-ontology. Case 4, WordNet*-WS*, represents the combination of ontologies where the specification components in the first ontology are a subset of the specification components in the second ontology. In this case, WordNet has parts and semantic relations, whereas WS has parts, functions, and attributes as well as semantic relations. From the manual integration of WordNet and SDTS into WS we specified which entity classes in WordNet correspond to what entity classes in WS. A more complex situation occurs when specification components have major differences (i.e., WordNet*-SDTS*). To simplify this task, we consider a particular application that deal with spatial entity classes present on a university campus map. Thus, forty-eight entity classes in SDTS where selected, and a manual process found twenty-two corresponding entity classes between WordNet and SDTS.

The experiment compare all entity classes across ontologies using different weights for synonym-set, feature, and semantic-neighborhood matching. We show in this paper only those results that represent lower and upper bounds in terms of recall and precision for each of the combination of ontologies. Table 5 shows results using a threshold of 75%, that is, entity classes with lower similarity than 75% were disregarded. Using a lower threshold increases recall, but decreases precision.

Case	**Weights (%)**			**Recall**	**Precision**
	ω_w	ω_u	ω_n	(%)	(%)
WordNet-WordNet	50	0	50	100	97
WordNet-WordNet	0	100	0	48	10
SDTS-SDTS	50	0	50	100	100
SDTS-SDTS	0	0	100	100	1
WordNet-WordNet*	50	0	50	99	98
WordNet-WordNet*	0	50	50	28	14
WordNet*-WS	100	0	0	100	78
WordNet*-WS	50	0	50	55	98
WordNet*-WS	0	50	50	0	0
WordNet*-SDTS*	100	0	0	100	42
WordNet*-SDTS*	50	0	50	50	92
WordNet*-SDTS*	0	100	0	0	0

Table 5: Recall and precision of evaluations with threshold equal to 75%. Symbol * denotes small subsets of the entire ontology.

We obtained obvious results for cases of comparing ontologies with themselves. Recall based on word matching is 100%, since corresponding entity classes have the same names. Precision, however, is not necessarily 100% for cases with identical ontologies due to the presence of polysemous terms. A more general observation indicates that feature matching alone is insufficient for detecting the most similar entity classes across ontologies. Many entity classes share common features or have a common superclass from which they inherit common features. This situation is particularly true for the SDTS ontology, which has a low value for precision. SDTS has distinguishing features in its entity classes' definitions, but the intrinsically nature of the general top level entity classes without features and the lack of precision of features to distinguish entity classes produce bad results.

Recall and precision decrease drastically for combinations of weights that ignore word matching. The combination of word and semantic-neighborhood matching obtains, in most cases, the best evaluations of recall and precision. Complementing word matching with feature matching tends to increase precision, but decreases recall. As was expected, the worst results are associated with evaluations over different ontologies (i.e., WordNet*-WS and WordNet*-SDTS*). In these cases, precision is still over 85%, but recall is considerably lower (50%-55%). For different ontologies, introducing feature matching had a negative effect in the performance of the model.

This experiment has shown that the results of the similarity model are highly sensitive to the components of the entity class representations. As ontologies share more components in their entity class specifications, the model produces more accurate results. Thus, in an environment with multiple ontologies, a similarity function should emphasize those components of an entity class representation that are likely shared by all ontologies. In an ideal scenario where ontology specifications are complete (i.e., entity class representation contains semantic relations and distinguishing features) and detailed (i.e., features differentiate entity classes), the similarity model is a good estimator for similarity. In a realistic scenario with different ontologies, however, the test found that synonym sets and semantic neighborhood are more stable specification components than the set of features associated with entity classes. Thus, semantic organization of entity classes is more similar across ontologies than the distinguishing features used to describe those entity classes.

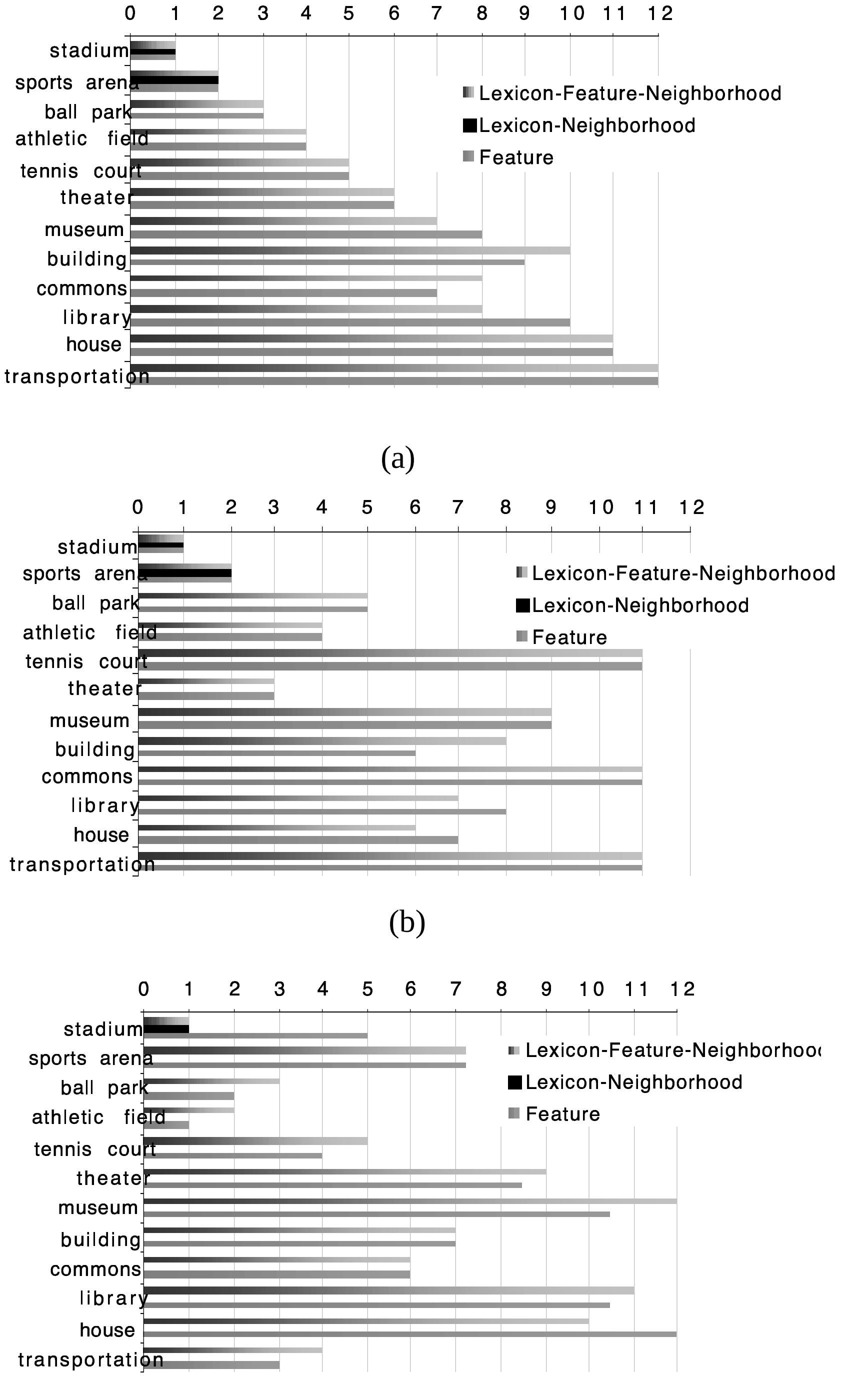
5.2 Experiment 2: Rank of Similarity

This experiment consists in cross-ontology evaluations that are transformed into a rank of similarity. The evaluations compare an entity class in an ontology (e.g., *stadium*) with a reduced set of entity classes defined in a different ontology (e.g., *stadium*, *athletic field*, *ballpark*, *tennis court*, *commons*, *building*, *theater*, *museum*, *library*, *transportation system*, *house*, *sport arena*). We selected evaluations between SDTS-WS and WordNet-WS, because they represent different levels of detail in the entity class representations and because WS has our proposed representation of entity classes that allows the exploration of each of the components of the evaluation across ontologies (i.e., word matching, feature matching, and semantic-neighborhood matching). We also considered the evaluation between WS-WS, since it corresponds to the best scenario with equivalent definitions and can indicate how well the similarity model works for evaluations within a single ontology. We chose the entity class *stadium* as the target of our evaluations, since this entity classes was found equivalent across ontologies so, similar entity classes to *stadium* in one ontology should be similar to *stadium* in a second ontology.

Since we wanted to evaluate the quality of the results derived from the computational model, we use a human-subject testing. We decided to design a new experiment rather than using previous experiments of similarity assessment within a single ontology [5, 21, 55], because these previous studies compare quite different entity classes (e.g., *car, automobile*, *food*, *birth*, *brother*, *noon*, so on) without focusing on distinguishing more related entity classes (e.g., *stadium*, *athletic field*, *sports arena*, *park*, so on). We asked subjects to rank the similarity among the set of entity classes based on the definitions in the WS ontology that were given at the beginning of the experiment. Thirty-seven students (twenty female and seventeen male) of an undergraduate English class at the University of Maine participated in the experiment. For all subjects, U.S. English is their mother tongue and their ages range from 18 to 36 years old. Subjects were paid for participating in the experiment and answered the questions at the same time and in less than 10 minutes without pressure.

The subjects' answers varied in the number of ranks used to classify entity classes. Most of them, however, assigned to each entity class a different rank. To compare subjects' answers, tied ranks were normalized by the mean of the ranks for which they tie, assuming a number of ranks equal to the number of entity classes compared [60]. The normalized answers were averaged, then ranked and normalized, if needed, to obtain the final ranks, which are compared against the similarity model. We found no significant evidence for differences based on gender, so the result is given as the total of responses. Subjects found that the most similar entity classes to a *stadium* in decreasing order were *sports arena*, *ball park*, *athletic field*, *tennis court*, *theater*, *museum*, *building*, *commons*, *library*, *house*, and *transportation*.

The model evaluations used three types of weight settings: the default (i.e., ω_w: 0.33; ω_u: 0.33; ω_n: 0.33), the best combination of weights that was found in the previous experiment (i.e., ω_w: 0.5; ω_u: 0; ω_n: 0.5), and feature-based similarity evaluation (i.e., ω_w: 0; ω_u: 100; ω_n: 0). Figures 3a-c present the model's results with the three different settings and the combinations SDTS-WS, WordNet-WS, and WS-WS, respectively. In these graphs, the ordering of entity classes in the axis corresponds to the subjects' responses in decreasing order.



(a)

(b)

(c)

Figure 3: Evaluations between a *stadium* and a set of entity classes: (a) SDTS-WS; (b) WordNet-WS; and (c) WS-WS.

The correlation between the model's results and the subjects' responses was estimated by the Spearman rank correlation coefficient [61], since this coefficient allows the statistical test based on ranked answers. The test statistic is also a measure of association such that it is equal to +1 when there is a perfect direct relationship between rankings. Table 6 gives the correlation coefficient under the presence of ties for each combination of ontologies and each weight setting.

Ontologies	**Synset-Feature-Neighborhood**	**Synset-Neighborhood**	**Feature**
	ω_w: 33.3, ω_u: 33.3, ω_n: 33.3	ω_w: 50, ω_n: 50	ω_u: 100
SDTS-WS	0.48	-0.34	0.37
WordNet-WS	0.68	-0.34	0.71
WS-WS	0.96	-0.34	0.97

Table 6: Correlation coefficient for similarity ranks in cross-ontology evaluations.

Like the former experiment, this experiment has shown that the performance of the model depends on how compatible are the ontology specifications. As expected the best results are between WS and WS whereas the worst results are between SDTS and WS. The best combination of weights detected in the former experiment gave the worst values of correlation for each of the ontology combinations. This bad correlation is due to the fact that the model with this combination of weights detects the most similar entity class and nothing else.

Since the comparison between the model and the subjects' responses is only possible in terms of the entity classes that subjects were asked to rank, ranking of the model's results is done over the similarity values obtained for this set of entity classes. Therefore, an entity class that was ranked second within the small set of entity classes could be ranked fifth with respect to the whole ontology. This situation could mislead the interpretation of the results based on the measures of recall and precision; however, the most important conclusion of this experiment is that feature matching is important for detecting similar entity classes within an ontology or the similarity of semantically related entity classes across ontologies. The assignment of the weights to the similarity of the specification components cannot only depend on the ontology characteristics, but also on the goal of the similarity assessment (i.e., ontology integration vs. information retrieval).

6. Conclusions and Future Work

We have presented a model for semantic similarity across different ontologies. The similarity model provides a systematic way to detect similar entity classes across ontologies based on the matching process of each of the specification components in the entity class representations (i.e., synonym sets, distinguishing features, and semantic neighborhoods). The similarity model is useful as a first step in an ontology integration, since it may detect most similar entity classes across ontologies. These similar entity classes could be then analyzed with user input to derive semantic relations, such as is-a relation or synonym relations, to create an integrated of ontology.

Experiments using the similarity model with different ontologies indicated that different components of entity class representations have different effects on the similarity evaluations. Synonym sets and semantic neighborhoods are good components to use for detecting equivalent or most similar entity classes across ontologies. Distinguishing features are suitable for detecting entity classes that are somewhat similar, that is, entity classes that are not synonyms and that are located far apart in the hierarchical structure (e.g., *stadium* and *athletic field* in the WordNet ontology).

This work has concentrated on entity classes and has compared distinguishing features in terms of a strict string matching between synonym sets that refer to those features. The semantic similarity among features, however, has been left for future work. For example, parts are also entity classes that could be semantically compared in a recursive process. Verbs could be related by the semantic relation *entailment* [33] (e.g., buy and pay) or could be formally specified such that they could be semantically compared. Likewise, the specification of attributes in terms of their domains (i.e., the set of possible values) could lead to exhaustive similarity evaluations among entity classes.

7. References

[1] Sheth, A., *Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics*, in M. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman (eds.), *Interoperating Geographic Information Systems*. 1999. Kluwer Academic Publishers: Norwell, MA. p. 5-30.

[2] Sheth, A. and V. Kashyap. *So Far (Schematically) Yet So Near (Semantically)*. in D. Hsiao, E. Neuhold, and R. Sacks-Davis (eds.), *IFIP WG2.6 Database Semantics Conference on Interoperable Database Systems*. 1992. North-Holland: Lorne, Victoria, Australia.

[3] Guarino, N., C. Masolo, and G. Verete, OntoSeek: Content-Based Access to the Web. *IEEE Intelligent Systems*, 1999. **14**(3): p. 70-80.

[4] Voorhees, E., *Using WordNet for Text Retrieval*, in C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database* 1998, The MIT Press: Cambridge, MA. p. 285-303.

[5] Jiang, J. and D. Conrath. *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. *International Conference on Computational Linguistics (ROCLING X)*. 1997. Taiwan.

[6] Smeaton, A. and I. Quigley. *Experiment on Using Semantic Distance Between Words in Image Caption Retrieval*. in *19th International Conference on Research and Development in Information Retrieval SIGIR'96*. 1996. Zurich, Switzerland.

[7] Lee, J., M. Kim, and Y. Lee, Information Retrieval Based on Conceptual Distance in IS-A Hierarchies. *Journal of Documentation*, 1993. **49**(2): p. 188-207.

[8] Goñi, A., E. Mena, and A. Illarramendi. *Querying Heterogeneous and Distributed Data Repositories Using Ontologies*. in P.-J. Charrel and H. Jaakkola (eds.), *Information Modelling and Knowledge Base IX*. 1998: IOS Press.

[9] Miller, G., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 1990. **3**(4): p. 235-244.

[10] Knight, K. and S. Luk. *Building a Large-Scale Knowledge Base for Machine Translation*. in *National Conference on Artificial Intelligence AAAI-94*. 1994. Seattle, WA.

[11] Sowa, J., *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. 2000, Pacific Grove, CA: Brook/Cole, a division of Thomson Learning.

[12] Rector, A., W. Nowlan, and A. Glowinski. *Goals for Concept Representation in the GALEN Project*. in *17th Annual Symposium on Computer Applications in Medical Care SCAMC'93*. 1993. Washington.

[13] Schlenoff, C., A. Knutilla, and S. Ray. *A Robust Ontology for Manufacturing Systems Integration*. in *2nd International Conference on Engineering Design and Automation*. 1998. Mai, HI.

[14] Lenat, D. and R. Guha, *Building Large Knowledge Based Systems: Representation and Inference in the CYC Project*. 1990, Reading, MA: Addison-Wesley Publishing Company.

[15] Guarino, N. *Formal Ontology in Information Systems*. in N. Guarino (ed.), *Formal Ontology in Information Systems*. 1998. Trento, Italy: IOS Press.

[16] Bishr, Y., *Semantic Aspects of Interoperable GIS*, Ph.D. Thesis, Wageningen Agricultural University and ITC, The Netherlands, 1997.

[17] Bright, M., A. Hurson, and S. Pakzad, Automated Resolution of Semantic Heterogeneity in Multidatabases. *ACM Transactions on Database Systems*, 1994. **19**(2): p. 212-253.

[18] Fankhauser, P. and E. Neuhold. *Knowledge Based Integration of Heterogeneous Databases*. in H. Hsiao, E. Neuhold, and R. Sacks-Davis (eds.), *Database Semantics*

Conference on Interoperable Database Systems IFIP WG2.6. 1992. Victoria, Australia: Elsevier Science Publishers, North-Holland.

[19] Collet, C., M. Huhns, and W. Shen, Resource Integration Using a Large Knowledge Base in Carnot. *Computer*, 1991. **24**(12): p. 55-62.

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

[21] Resnik, O., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research*, 1999. **11**: p. 95-130.

[22] Ross, S., *A First Course in Probability*. 1976, New York: Macmillan.

[23] Mena, E., V. Kashyap, and A. Sheth. *OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies*. in *International Conference on Cooperative Information Systems (CoopIS'96)*. 1996. Brussels, Belgium: IEEE Computer Society Press.

[24] Kashyap, V. and A. Sheth, *Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context, and Ontologies*, in M. Papazoglou and G. Schlageter (eds.), *Cooperative Information Systems: Tends and Directions* 1998, Academic Press: London, UK. p. 139-178.

[25] Bergamaschi, B., S. Castano, S.D.C.d. Vermercati, S. Montanari, and M. Vicini. *An Intelligent Approach to Information Integration*. in N. Guarino (ed.), *First International Conference on Formal Ontology in Information Systems*. 1998. Trento, Italy: IOS Press.

[26] Gangemi, A., D. Pisanelli, and G. Steve. *Ontology Integration: Experiences with Medical Terminologies*. in N. Guarino (ed.), *Formal Ontology in Information Systems*. 1998. Trento, Italy: IOS Press.

[27] Kim, W. and J. Seo, Classifying Schematic and Data Heterogeneity in Multidatabase Systems. *IEEE Computer*, 1991. **24**: p. 12-18.

[28] Visser, P., D. Jones, T. Bench-Capon, and M. Shave, *Assessing Heterogeneity by Classifying Ontology Mismatches*, in N. Guarino (ed.), *Formal Ontology in Information Systems* 1998, IOS Press: Amsterdam, the Netherlands. p. 148-162.

[29] Weinstein, P. and P. Birmingham. *Comparing Concepts in Differentiated Ontologies*. in *12th Workshop on Knowledge Acquisition, Modeling, and Management*. 1999. Banff, Canada.

[30] Mena, E., A. Illarramendi, V. Kashyap, and A. Sheth, OBSERVER: An Apprioach for Query Processing in Global Information Systems Based on Interoperation across Pre-existing Ontologies. *Distributed and Parallel Databases*, 2000. **8**(2): p. 223-271.

[31] Rodríguez, A., M. Egenhofer, and R. Rugg, *Assessing Semantic Similarity Among Geospatial Entity Class Definitions*, in A. Vckovski, K. Brassel, and H.-J. Schek (eds.), *Interoperating Geographic Information Systems INTEROP99, Zurich, Switzerland*. Lecture Notes in Computer Science 1580, 1999, Springer-Verlag, Berlin: Berlin. p. 189-202.

[32] Miller, G., *Nouns in WordNet*, in C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database* 1998, The MIT Press: Cambridge, MA. p. 23-46.

[33] Fellbaum, C., *A Semantic Network of English Verbs*, in C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database* 1998, The MIT Press: Cambridge, MA. p. 69-104.

[34] Smith, J. and D. Smith, Database Abstractions: Aggregation and Generalization. *ACM Transactions of Database Systems*, 1977. **2**(2): p. 105-133.

[35] Guarino, N., Formal Ontology, Conceptual Analysis, and Knowledge Representation. *International Journal of Human and Computer Studies*, 1995. **43**(5/6): p. 625-640.

[36] Winston, M., R. Chaffin, and D. Herramann, A Taxonomy of Part-Whole Relations. *Cognitive Science*, 1987. **11**: p. 417-444.

[37] Cruse, D., On The Transitivity of the Part-Whole Relation. *Linguistics*, 1979. **15**: p. 29-38.

[38] Iris, M., B. Litowitz, and M. Evens, *Problem of the Part-Whole Relation*, in M. Evens (ed.), *Relational Models of the Lexicon: Representing Knowledge in Semantic Network* 1988, Cambridge University Press: Cambridge, MA. p. 261-288.

[39] Gibson, J., *The Ecological Approach to Visual Perception*. 1979, Boston, MA: Houghton Mifflin.

[40] Khoshafian, S. and R. Abnous, *Object Orientation: Concepts, Languages, Databases, and User Interfaces*. 1990, New York: John Wiley & Sons.

[41] USGS, *View of the Spatial Data Transfer Standard (SDTS) Document*, http://mcmcweb.er.usgs.gov/sdts/standard.html, 1998.

[42] Rodríguez, A. and M. Egenhofer, *Putting Similarity Assessment into Context: Matching Functions with the User's Intended Operations*, in P. Bouquet, L. Sefarini, O. Brezillon, and F. Castellano (eds.), *Modeling and Using Context CONTEXT99, Trento, Italy*. Lecture Notes in Computer Science 1688, 1999, Springer-Verlag: Berlin. p. 310-323.

[43] Rips, L., J. Shoben, and E. Smith, Semantic Distance and the Verification of Semantic Relations. *Journal of Verbal Learning and Verbal Behavior*, 1973. **12**: p. 1-20.

[44] Krumhansl, C., Concerning the Applicability of Geometric Models to Similarity Data: The Interrelationship Between Similarity and Spatial Density. *Psychological Review*, 1978. **85**(5): p. 445-4 63.

[45] Goldstone, R., Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1994. **20**(1): p. 3-28.

[46] Sussna, M. *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network*. in *Second International Conference on Information Knowledge Management, CIKM'93*. 1993.

[47] Lin, D. *An Information-Theoretic Definition of Similarity* (eds.), *International Conference on Machine Learning ICML'98*. 1998. Madison, WI.

[48] Rada, R., H. Mili, E. Bicknell, and M. Blettner, Development and Application of a Metric on Semantic Nets. *IEEE Transactions on System, Man, and Cybernetics*, 1989. **19**(1): p. 17-30.

[49] Rosch, E., Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology*, 1975. **104**: p. 192-233.

[50] Holman, E., Monotonic Models for Asymmetric Proximities. *Journal of Mathematical Psychology*, 1979. **20**: p. 1-15.

[51] Bowdle, F. and D. Gentner, Informativity and Asymmetry in Comparisons. *Cognitive Psychology*, 1997. **34**: p. 244-286.

[52] Birkhoff, G., *Lattice Theory*. 1967, Providence, RI: American Mathematical Society.

[53] Khoshafian, S. and G. Copeland. *Object Identity*. in *OOPSLA*. 1986. Portland, OR.

[54] Zdonik, S. and D. Maier, *Fundamentals of Object-Oriented Databases*, in S. Zdonik and D. Maier (eds.), *Readings in Object-Oriented Systems* 1990, Morgan Kaufmann: San Mateo, CA. p. 1-32.

[55] Miller, G. and W. Charles, Contextual Correlates of Semantic Similarity. Language and Cognitive Processes, 1991. **6**(1): p. 1-28.

[56] Leacock, C. and M. Chodorow, *Combining Local Context and WordNet Similarity for Word Sense Identification*, in C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database* 1998, The MIT Press: Cambridge, MA. p. 265-283.

[57] Burg, J. and R.v.d. Riet, *COLOR-X: Using Knowledge from WordNet for Conceptual Modeling*, in C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database* 1998, The MIT Press: Cambridge, MA.

[58] Richardson, R. and A. Smeaton, *Using WordNet in a Knowledge-Based Approach to Information Retrieval*. 1995, Dublin City University, School of Computer Applications: Dublin, Ireland.

[59] Korfhage, R., *Information Storage and Retrieval*. 1997, New York: John Wiley & Sons.

[60] Daniel, W., *Applied Nonparametric Statistics*. 1978, Boston, MA.: Houghton Mifflin.

[61] Gibbons, J., *Nonparametric Methods for Quantitative Analysis*. 1976, Columbus, OH: American Sciences Press.