Comparing Geospatial Entity Classes:

An Asymmetric and Context-Dependent Similarity Measure

M. Andrea Rodríguez
Department of Computer Science, Universidad de Concepción, Casilla 160-C, Correo 3, Concepcion, Chile, andrea@udec.cl

Max J. Egenhofer
National Center for Geographic Information and Analysis, Department of Spatial Information Science and Engineering, and Department of Computer Science, University of Maine, Orono, ME 04469-5711, USA, max@spatial.maine.edu

Abstract

Semantic similarity plays an important role in geographic information systems as it supports the identification of objects that are conceptually close, but not identical. Similarity assessments are particularly important for retrieval of geospatial data in such settings as digital libraries, heterogeneous databases, and the World Wide Web. Although some computational models for semantic similarity assessment exist, these models are typically limited by their inability to handle such important cognitive properties of similarity judgments as their inherent asymmetry and their dependence on context. This paper defines the Matching-Distance Similarity Measure (MDSM) for determining semantic similarity among spatial entity classes, taking into account the distinguishing features of these classes (parts, functions, and attributes) and their semantic interrelations (is-a and part-whole relations). A matching process is combined with a semantic-distance calculation to obtain asymmetric values of similarity that depend on the degree of generalization of entity classes. MDSM's matching process is also driven by contextual considerations, where the context determines the relative importance of distinguishing features. Based on a human-subject experiment, MDSM results correlate well with people's judgments of similarity. When contextual information is used for determining the importance of distinguishing features, this correlation increases; however, the major component of the correlation between MDSM results and people's judgments is due to a detailed definition of entity classes.

1. Introduction

In information systems, similarity assessment is an integral part of such processes as information retrieval, natural-language processing, information integration, and data maintenance. For geographic information systems (GISs), similarity assessment is particularly important because typically users of spatial data have diverse backgrounds and no precise definitions underlie the matter of discourse. A semantic similarity model facilitates the comparison among entities and allows information retrieval and information integration to handle entities that are semantically similar. Traditional methods for information retrieval have been primarily based on query-string matching and statistical analysis. New trends in of information retrieval stress the advantages of using domain knowledge and semantic similarity functions to compare words or documents (Lee *et al.* 1993, Richardson and Smeaton 1995, Voorhees 1998). Within this context, the goal of a similarity model is to obtain flexible and better matches between user-expected and system-retrieved information. In addition, heterogeneous spatial databases could achieve real information integration, because they would be able to identify similar objects that can be exchanged, without compromising semantics. Unlike approaches that integrate different conceptualizations into a common schema for data integration (Mena *et al.* 1996, Bergamaschi *et al.* 1998, Kavouras and Kokla 2002), the use of semantic similarity measures is a strategy that allows us to associate dynamically entities from different conceptualizations while keeping these conceptualizations independently (Rodríguez and Egenhofer 2003).

The Semantic Geospatial Web (Egenhofer 2002, Fonseca and Sheth 2003) is envisioned as a new information-retrieval environment that will facilitate meaningful access to geospatial information. While the World Wide Web currently provides good access to data through a variety of search engines as long as the user knows the keywords that the data

providers used, it falls short as a reliable access mechanism to information when purely syntactic comparisons cannot resolve ambiguities or fail to build connections to related or similar items that a data provider did not foresee. The Semantic Web (Berner-Lees *et al.* 2001) aims to overcome the current limitations by incorporating explicitly-modeled expressions of semantics into the search process. The provision of such explicit semantics may be seen as a much richer metadata model, with the goal to offer machine-readable and machine-executable metadata. The domain of geospatial information is particularly rich in this respect due to the varieties in human spatial languages for expressing and communicating spatial information.

In order to address geospatial semantics, one needs computational methods that go beyond syntax comparisons. In the case of the Semantic Geospatial Web, three types of geospatial semantics are distinguished, each requiring different computational methods (Egenhofer 2002):

- Semantics of geospatial entity classes
- Semantics of spatial predicates
- Semantics of geospatial names

This paper investigates cognitively plausible methods for making comparisons of geospatial entity classes. In combination with geospatial ontologies, (Smith and Mark 2001, Frank 2001, Kuhn 2000), these methods of spatial similarity provide a flexible framework to bridge between the conceptual models of the data providers and the conceptual models of the users (Fonseca *et al.* 2002). The paper focuses on the semantics of spatial entities and introduces a measure for assessing semantic similarity among spatial entity classes. Semantic similarity assessment ignores some of the geometric properties of spatial databases, such as density, dispersion, and pattern derived from representative subsets (Flewelling 1999) and extent and location displayed by magic lenses (Schenkelaars and Egenhofer 1997). The classification of geographic entities, however, is geospatial, even when no geometry is involved. Non-geometric concepts, such as building, road, and place, are geospatial concepts that are used for describing the semantics of geospatial objects.

Much past research in geographic information science that is concerned with similarity assessments has focused on the geometric properties of geospatial information. Examples of these studies are topological equivalence (Egenhofer and Franzosa 1995), cardinal direction between extended spatial objects (Goyal and Egenhofer 2001), metric details of spatial relations (Egenhofer and Shariff 1998), and content-based image retrieval (El-Kwae and Kabuka 1999, Yoshitaka and Ichikawa 1999). Omitting the geometric properties of geospatial objects, this paper concentrates on the cognitive properties of the semantic similarity assessment that relate to the geospatial domain and leaves for future work the integration of geometric and semantic similarity.

The study of similarity models has been an important area of investigation for psychologists and computer scientists. While psychologists have pursued the identification of how people classify objects, form concepts, solve problems, and make generalizations (Rosch and Mervis 1975, Osherson and Smith 1981, Smith and Osherson 1984, Goldstone *et al.* 1997), computer scientists have relied on similarity measures in natural-language processing (Sussna 1993, Resnik 1999), information retrieval (Kim and Kim 1990, Richardson and Smeaton 1995), and information integration (Mena *et al.* 1996, Weinstein and Birmingham 1999). Most similarity models defined by psychologists are based on features or descriptors of concepts (Tversky 1977, Krumhansl 1978, Goldstone 1994). This approach to similarity modeling is in contrast to the work done by computer scientists, who typically use the concept interrelations in a hierarchical structure to define a similarity measure (Rada *et al.* 1989, Lee *et al.* 1993, Smeaton and Quigley 1996, Resnik 1999).

Two key characteristics of similarity models are whether or not they are symmetric and context independent. Although most similarity models based on concepts' interrelations

assume symmetric evaluations, psychologists argue that similarity is not always a symmetric relation (Tversky 1977, Krumhansl 1978). For example, the statement "a hospital is similar to a building" is more generally accepted than "a building is similar to a hospital." In the naive view of the world, distance as well as similarity defined in terms of a conceptual distance are frequently asymmetric (Egenhofer and Mark 1995). Although the similarity evaluation between a class and its superclass may seem odd, it is not unusual to compare objects whose classifications have been assigned by different users and with a different degree of generalization. For example, while one user may classify some objects from a dataset as buildings, another user may go further in her or his classification and distinguish among types of buildings in another dataset (e.g., hospital, theater, house, and so on). Then, a reasonable request is to compare the two datasets.

The literature also indicates that, in addition to such asymmetries, context and frame of reference should be considered when evaluating similarity assessments (Tversky 1977, Goldstone *et al.* 1997, Medin *et al.* 1993). For example, how similar are an apple and a pear with respect to taste? In feature-based models, features may be given different weights in different stimulus contexts (Krumhansl 1978) and these weights could be determined by how diagnostic the feature is for a particular set of objects under consideration (Tversky 1977, Goldstone *et al.* 1997). Models based on semantic distance (Rada *et al.* 1989, Lee *et al.* 1993) ignore any contextual dependence of the similarity assessment and rely on a well-structured hierarchy. Although a recent information-based approach to semantic similarity shows that the content information of concepts could have some implications for the consideration of context (Resnik 1999, Lin 1998), no further study has yet been done in this direction.

This paper extends previous work on semantic similarity (Rodríguez and Egenhofer 1999, Rodríguez *et al.* 1999) by introducing the Matching-Distance Similarity Measure (MDSM) to determine semantic similarity among spatial entity classes. It addresses the asymmetry of similarity judgments as well as the role of context in such judgments. For this work, entity classes correspond to cognitive representations that people use to recognize and categorize objects or events (Dahlgren 1988). The work is strongly influenced by studies in cognitive psychology and natural-language processing. It shares Talmy's (1983) and Herskovits's (1997) assumptions that the language we speak reflects our conceptual system; that is, we can treat concepts as linguistic terms and represent these terms' semantics.

Our work synthesizes and expands in two significant ways the work done by psychologists and computer scientists. First, we supplement the feature-based approach by systematically treating the asymmetry of the similarity judgments. We obtain asymmetric values for similarity assessments of spatial entity classes based on their degree of generalization within a hierarchical structure. Second, we extend the feature-based approach by explicitly incorporating context into the similarity assessment, where context leads to the determination of entity classes in the domain of an application and, consequently, determines the relative importance of distinguishing features in the similarity judgments. Such a new measure of semantic similarity matches with people's judgments of similarity, and it is useful for comparing concepts that are organized by their semantic interrelations and described by their distinguishing features.

The remainder of this paper is organized as follows. Section 2 describes the components of spatial entity class representations. Section 3 describes MDSM, followed by the specification of contextual information in Section 4. An example in Section 5 illustrates the use of MDSM under different contexts. Section 6 evaluates MDSM with a human-subject experiment and compares the results with previous similarity models. Conclusions and future work are addressed in Section 7.

2. Spatial Entity Class Representation

We organize spatial entity classes based on their semantic interrelations and describe the set of spatial entity classes and their semantic relations as an *ontology* (Gruber 1995, Guarino and Giaretta 1995). For this work, we consider an ontology to be a type of knowledge base that

describes concepts through definitions that are sufficiently detailed to capture the semantics of a domain. In this sense, an ontology captures a view of the world, supports intentional queries regarding the content of a database, defines semantics independently of data representation, and reflects the relevance of data without needing to access them (Goñi *et al.* 1998).

Semantic relations are a typical way to describe knowledge about concepts. In natural-language communication, for instance, synonymy, antonymy, hyponymy, meronymy, and entailment are examples of semantic relations used to define terms (Miller 1995). We refer to entity classes by words or sets of synonyms, which are interrelated by hyponymy and meronymy relations. The hyponymy relation, usually called is-a relation (Smith and Smith 1977), is the most common relation used in an ontology. The is-a relation is transitive and asymmetric and defines a hierarchical structure where terms inherit all the characteristics from their superordinate terms. Mereology, the study of part-whole relations, also plays an important role in the definition of an ontology (Guarino 1995). Studies have usually assumed that part-whole relations are transitive, so that if *a* is part of *b* and *b* is part of *c*, then *a* is part of *c* as well. Linguists, however, have expressed concerns about this assumption (Cruse 1979, Iris *et al.* 1988). Explanations of the transitive problem rely on the idea that part-whole relations are not one type of relation, but a family of relations. Among all types of part-whole relations, this work considers the component-object and stuff-object relations (Winston *et al.* 1987) with the properties of asymmetry and, for some cases, transitivity. When describing the semantic relations among entity classes, MDSM distinguishes the two relations "part-of" and "whole-of" to account for cases when the converseness of part-whole and whole-part relations does not hold. For example, we can say that a *building complex* has *buildings* (i.e., *building complex* is the whole for a set of *buildings*); however, we cannot claim that all *buildings* are part of a *building complex*.

Although the general organization of entity classes is given by their is-a and part-whole interrelations, this information may not be sufficient 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' functionalities (i.e., a *hospital* is a building where medical care is given and an *apartment building* is a group of apartments that serve as living quarters).

Typically, *attributes* describe different types of *distinguishing features* of a class. Attributes capture details about entity classes, while their values describe the properties of individual objects (i.e., instances of an entity class). We suggest a finer identification of distinguishing features and classify them into functions, parts, and attributes. Function features are intended to represent what is done to or with a class. For example, the function of a *college* is to *educate*. These function features can be related to other terms, such as *affordances* (Gibson 1979) and *behavior* of the object-orientation paradigm (Khoshafian and Abnous 1990). Parts are structural elements of a class, such as the *roof* and *floor* of a *building*. It is possible to make a further distinction between "things" that a class may have ("optional") or must have ("mandatory"). In this work, we focus exclusively on mandatory parts, which are associated with part-whole relations. While part-whole relations work at the level of entity class representations and force us to define all the entity classes involved, part features can have items that are not always defined as entity classes in this model. For example, although *roof* and *floor* are part features of a building, they may not be necessarily defined as entity classes in the model. Finally, some attributes can correspond to additional characteristics of a 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*. This classification of distinguishing features into parts, functions, and attributes attempts to facilitate the implementation of the entity class representation, as well as enable the separate manipulation of each type of distinguishing feature. Even more, context can affect the importance of each feature, depending on the *role* that an object plays in a particular context (Fonseca *et al.* 2002).

Considering that spatial entity classes correspond to nouns in linguistic terms, this work matches Miller's (1990) description of nouns, since entity classes are semantically interrelated by hyponomy and meronomy relations. Likewise, the representation of spatial entity classes that underlies MDMS relates to the *qualia structure* of the Generative Lexicon Theory (GLT) (Pustejovsky 1995). The *qualia structure* of a lexical meaning in GLT is composed of *CONST* (i.e., the role that expresses the relationship between an object and its parts), FORMAL (i.e., the role that distinguishes an object within a large domain), TELIC (i.e., the role that indicates the object's function or goal), and AGENTIVE (i.e., the origin or "bringing about" of an object). Although the representation of spatial entity classes used in MDSM does not explicitly included the AGENTIVE role of the *qualia structure*, it does so as it semantically relates entity classes by the *is-a* relation; for example, if entities are classified into *artifacts* and *natural kinds*.

Using a lexical categorization, parts are given by nouns, functions by verbs, and attributes by nouns whose associated values are given by adjectives or other nouns. As with spatial entity classes, synonym sets identify distinguishing features, because these sets carry more semantic information than a single term. This work does not address the semantic matching among distinguishing features. We expect that a set of synonyms can identify a distinguishing feature with little ambiguity such that a matching among synonym sets can identify equivalent distinguishing features.

The representation of entity classes used in MDSM can be clearly associated with the definition of classes in object-oriented theory (Khoshafian and Abnous 1990). Is-a and part-whole relations are extracted from basic paradigms of object-oriented theory (i.e., inheritance and composition, respectively), while the distinguishing features of the MSDM entity class representation, with the exception of parts, correspond to attributes or methods of classes in object orientation. The definition of spatial entity classes uses the two inclusion relations that have been considered more relevant in a semantic specification. The hierarchy of inclusion relations establishes that spatial inclusion, meronymy inclusion, and class inclusion are the lower, medium, and higher relations, respectively (Winston *et al*. 1987). The definition of entity classes excluded the spatial components, because these properties—topology, size, shape, orientation, distance, and direction—are mostly associated with instantiations rather than definitions of entity classes.

Table 1 presents a formal syntax of an entity class definition using BNF notation, with an example of the definition of the entity class *stadium* that is derived from a combination of WordNet and SDTS. In this specification, primitives of the MDSM language are *pointers* and *words*.

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

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. The Matching-Distance Measure for Semantic Similarity

We define a new computational model that assesses similarity by combining the comparison of distinguishing features classified by types. The global similarity function $S(c_1,c_2)$ is a weighted sum of the similarity values for parts, functions, and attributes (Equation 1), where ω_p, ω_f, and ω_a are the weights of the similarity values for parts, functions, and attributes, respectively. These weights define the relative importance of parts, functions, and attributes that may vary among different contexts. The sum of the weights must equal 1.

$$S(c_1,c_2) = \omega_p \cdot S_p(c_1,c_2) + \omega_f \cdot S_f(c_1,c_2) + \omega_a \cdot S_a(c_1,c_2) \tag{1}$$

For each type of distinguishing feature we use a similarity function $S_t(c_1,c_2)$ (Equation 2), which is based on the *ratio model* of a feature-matching process (Tversky 1977). In $S_t(c_1,c_2)$, c_1 and c_2 are two entity classes, t symbolizes the type of features, and C_1 and C_2 are the respective sets of features of type t for c_1 and c_2. The matching process determines the cardinality ($|\ |$) of the set intersection ($C_1 \cap C_2$) and the set difference ($C_1 \setminus C_2$).

$$S_t(c_1,c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha(c_1,c_2)\cdot|C_1 \setminus C_2| + (1-\alpha(c_1,c_2))\cdot|C_2 \setminus C_1|} \tag{2}$$

Like Tversky's model (1977), MDSM uses the number of common and different features between two entity classes; however, it differs from Tversky's model in that it defines the relevance of the different features in terms of the distance among entity classes in a hierarchical structure. Thus, we take further the general formulation of the ratio model by completely defining a function α that determines the relative importance of different features between entity classes. This function α is defined in terms of the distance between the entity classes (c_1 and c_2) and the immediate superclass that subsumes both classes. The immediate

common superclass corresponds to the least upper bound (*lub*) between two entity classes in partially ordered sets (Birkhoff 1967). When one of the concepts is the superclass of the other, the former is also considered the immediate superclass (*lub*) between them. For instance, consider the hierarchical structure shown in Figure 1.

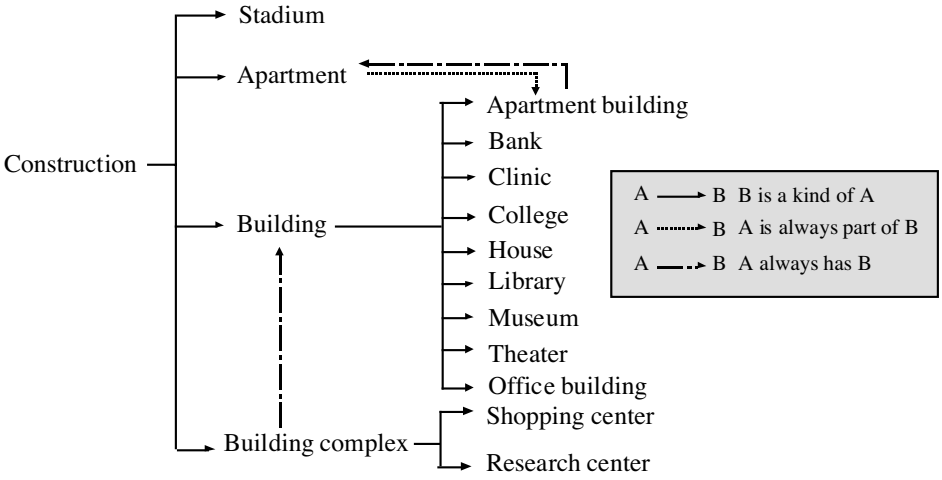


Figure 1: Portion of an Ontology with Is-a and Part-Whole Relations.

The immediate superclass for *stadium* and *house* is *construction*. In like manner, the immediate superclass for *building* and *museum* is *building*. The distance of each entity class to the *lub* is normalized by the total distance between the two classes, such that we obtain values in the range between 0 and 1. In this calculation, distance is given by the number of arcs along the shortest path between entity classes**.** Then, the final value of α is defined by a symmetric function (Equation 3).

$$\alpha(c_1,c_2) = \begin{cases} \dfrac{d(c_1,lub)}{d(c_1,c_2)} & d(c_1,lub) \le d(c_2,lub) \\ 1 - \dfrac{d(c_1,lub)}{d(c_1,c_2)} & d(c_1,lub) > d(c_2,lub) \end{cases} \quad \text{, where} \quad d(c_1,c_2) = d(c_1,lub) + d(c_2,lub) \qquad (3)$$

The parameter αweighs the importance of difference with respect to common distinguishing features of entity classes, considering that difference should be less important than commonality in a similarity assessment (Tversky 1977, Krumhansl 1978). The values of α are between 0 and 0.5, where a value of 0 represents the case when differences of one entity with respect to the other entity are the only important differences for the similarity evaluation, whereas a value of 0.5 represents the case when differences of both entity classes and equally important. The determination of α is based on the observation that similarity is not necessarily a symmetric relation. Cognitive psychologists have given different explanations for the asymmetric evaluations of similarity. Asymmetry can be explained by the relative size and salience of distinctive features sets (Tverski 1977), by potential stimulus biases, such as density and prototypicality (Krumhansl 1978, Holman 1979), by a natural reference point or landmark for members of a category (Rosch 1975), and by the direction of maximum informativity (Bowdle and Gentner 1997). Common to all these explanations is the different role that the *target*[1] and *base* positions play in a similarity evaluation. The most salient term, the item with larger bias, the prototypical term, and the term that provides information to understand the target are always in the base position. Our work assumes that a prototype used

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

as a base of a similarity evaluation is generally a superclass for a variant and that the perceived similarity from the variant to the prototype is greater than the perceived similarity from the prototype to the variant. Thus, the common, as opposed to the different, features between a subclass with respect to its superclass have a larger contribution to the similarity evaluation than the common features in an inverse direction. Using this assumption, we provide a systematic approach to determine the asymmetry of a similarity evaluation that is based on the distance from the *target* and *base* terms to the immediate superclass that subsumes both terms.

The similarity function (Equation 2) yields values between 0 and 1. The extreme value 1 represents the case when everything is common between two entity classes, or when the non-common features do not affect the similarity value (i.e., the coefficient of the non-common feature is zero). The value 0, on the other hand, occurs when nothing is common between two entity classes.

To illustrate the behavior of the model, consider the hierarchical semantic structure in Figure 1 and the definitions of the entity classes *building*, *building_complex, sports arena*, and *theater* given by the set of distinguishing features in Table 2. A set of similarity assessment evaluations between pairs of these entity classes is shown in Table 3. An interesting case occurs when comparing a class with its superclass or vice versa (e.g., *theater* versus *building* or vice versa). Since subclasses inherit features from their superclasses, the set difference between a subclass and its superclass may be greater than zero, whereas the set difference between the superclass and the subclass is zero. It can be easily seen that for this pair of entity classes the weight associated with the non-common features of the first argument α is 0 and the weight for the non-common features of the second argument $(1-\alpha)$ is 1, because the immediate superclass between a class and its superclass is this superclass itself. By considering the direction of the similarity evaluation, a class is seen to be more similar to its superclass than that same superclass is to the class. For entity classes at the same level of generalization (e.g., *theater* versus *sports arena* or vice versa), on the other hand, the same weight is assigned to the different features of entity classes and, therefore, a symmetric evaluation of similarity that depends on the number of common and different distinguishing features is performed.

Entity Class	Parts	Functions	Attributes
Building	Foundation Roof Wall		Architectural properties External material construction Height location Name Owner type Structure type User type
Building complex	Building Foundation Roof Wall		Architectural properties Area External material construction location Name Number of buildings Owner type Structure type User type
Theater	Dressing room Entrance hall Foundation Orchestra Roof Spectator stands Stage Ticket office Wall	Perform Present Recreate	Architectural properties External material construction Height location Name Owner type Structure type User type
Sport arena	Court Dressing room Foundation Roof Spectator stands Wall	Play Practice Recreate	Architectural properties External material construction Height location Name Owner type Structure type User type

Table 2: Example of Distinguishing Features for *Building*, *Building Complex*, *Theater*, and *Sport Arena*.

a versus b	α	$S_p(a,b)$	$S_f(a,b)$	$S_a(a,b)$
a = building	0.0	0.75	0.0	0.78
b = building complex				
a = building complex	0.0	1.0	0.0	0.85
b = building				
a = theater	0.0	1.0	0.0	1.0
b = building				
a = building	0.0	0.33	0.0	1.0
b = theater				
a = sport arena	0.5	0.53	0.33	1.0
b = theater				
a = theater	0.5	0.53	0.33	1.0
b = sport arena				

Table 3: Examples of Similarity among Parts, Functions, and Attributes.

In the calculation of α we use as basis not only the hierarchical nature of is-a relations, but also the hierarchical structure of part-whole relations. To determine a class that subsumes two classes under comparison, the is-a relation as well as the part-of and whole-of relations are checked. In Figure 1, the immediate superclass for *building* and *building complex* is *building complex*, since the closest path between the two classes is given by the link *building complex has always building*(*s*). Considering only is-a relations, however, would yield the superclass *construction*. Although we consider is-a and part-whole relations in the same way to determine the immediate superclass, we cannot infer that the similarity between a part to its whole will be always greater than the similarity between a whole to its part, since part-whole relations do not always have the strict inheritance properties that is-a relations do.

MDSM is based on the comparison of distinguishing features rather than on the shortest path in a hierarchical structure. It only uses this shortest path for determining the relative importance of the differences between distinguishing features. The lack of distinguishing features in an entity class's definition, however, produces a similarity value of zero with respect to any other entity class in the ontology. This is a common situation for entity classes that are general concepts located at the top level of the hierarchical structure, such as *entity* and *natural entity*. Although this can be seen as a drawback of MDSM, the MDSM's strength is its capability to assess the similarity among concepts located at or below Rosch's (1975) basic level of a hierarchical structure, such as *building*, *office building*, road, and *lake*. In MDSM, different concepts should have at least one different feature that distinguishes them. So, even if a class inherits all features from its superclass, we expect to have distinguishing features in the class representation that differ from the features in its superclass's representation; otherwise, they are considered equivalent. These characteristics of MDSM are in contrast to previous models based on semantic distance (Rada *et al*. 1989). While semantic distance can determine similarity among general concepts of a hierarchical structure, it usually assigns the same similarity value to any pair of entity classes that have a common superclass.

4. Integrating Context into the Similarity Model

Using common-sense definitions, one could expect to obtain a good approximation of the similarity assessment among entity classes by considering the essential properties of distinguishing features as equally important. Some features, however, may be more important than others, depending on context (Tversky 1977, Krumhansl 1978), since the classificatory significance of features varies with the set of entity classes under consideration. Similar to the analysis of word meaning within statements (Leacock *et al*. 1993), we analyze the similarity assessment within an *application domain*. This work defines the application domain as the set of entity classes that are subjects of the user's interest. Since an application domain may change among applications, the value of similarity assessment can change as well.

4.1 Determining the Domain of an Application

This work derives the domain of discourse from the user's intended actions or operations. These operations may be abstract, high-level intentions (e.g., "analyze" or "compare") or detailed plans (e.g., "purchase a house"). From a linguistic point of view, the user's intended operations are associated with verbs that denote actions. Verbs alone, however, may not be enough to completely describe these operations, since they can change the operations' meaning depending on the kinds of noun arguments with which they co-occur (Fellbaum 1990).

Contextual information (C) is specified as a set of tuples over operations (op_i) associated with their respective noun arguments (e_j) (Equation 4). The nouns correspond to entity classes in MDSM, while the operations refer to verbs that are associated with methods of these classes.

$$C = \left\langle \left(op_1, \{e_1, ..., e_m\}\right), ..., \left(op_n, \{e_1, ..., e_l\}\right) \right\rangle \quad (4)$$

Since the context specification uses operations and entity classes, the ontology used by MDSM can be extended to contain all components of the context specification. The context specification defines the domain of the application based on the operations that characterize the entity classes and the semantic relations among entity classes. These semantic relations provide a flexible way to describe context since the specification of one entity class can be used to obtain other entity classes that are semantically related. Following a top-down approach in the hierarchical structure of interrelated entity classes, the domain of the application is given by:

- entity classes whose functions correspond to the intended user's operations,
- entity classes that are parameters of the operations in the context specification, and
- entity classes derived from a recursive search of parts and subclasses of the entity classes found in the previous steps.

For example, if a user is looking for *sports facilities*, she can specify $C = <(search, \{sports\ facility\})>$. Using the hierarchical structure, the system will associate the domain of application with the entity class *sports facility* and its subclasses or parts. Another contextual specification is $C = <(search, \{athletic\ field, ball\ park, tennis\ court, sports\ arena, stadium\})>$. This specification is a more extensive description of the user's interests since it contains explicitly the entity classes that are part of the application domain. A contextual specification based on only operations is $C = <(play, \{\})>$. In this case, the operation *play* corresponds to a common function that characterizes the entity classes the user is looking for.

Like the topical context of word-sense disambiguation (Gale *et al.* 1992), the domain of the application helps to select among senses of a term with multiple meanings (i.e., polysemous terms). Since the domain of the application is usually a subset of the entire knowledge base, the contextual specification decreases the number of entity classes that possess the same name and are part of the application domain. Unfortunately, this approach may not distinguish polysemous terms that are semantically similar and belong to the same application domain.

4.2 Determining Feature Relevance

Tversky (1977) and later Goldstone *et al.* (1997) pointed out that the relevance of a feature is associated with how *diagnostic* the feature is for a particular set under consideration. The diagnosticity of features refers to the classificatory significance of features, which is highly sensitive to the particular entity classes under consideration. The previous section presented a method to derive the entity classes of interest for an application (i.e., application domain). This application domain may or may not be the set of entity classes that are compared in the similarity assessment. For example, a user may be looking for places where to play a sport and so chooses a stadium as the prototypical entity to search in a database. In an information retrieval process, stadium will be compared with any entity in the database, where these entities may be either inside or outside the application domain. Based on the characteristics of the application domain and the database, two different approaches to determining features' relevance are *variability* and *commonality*.

4.2.1 Variability

This approach relates the relevance of a feature to the degree of the feature's informativeness, such that a feature's relevance decreases if it is shared by all entity classes of the domain. For example, consider a small domain with buildings that have a common function (e.g., they all serve as sports facilities), but differ in their structural characteristics. Based on this approach, the buildings' structural characteristics are more relevant for the similarity assessment than the buildings' functional characteristics.

This approach defines weighted values for the similarity among parts, functions, and attributes (ω_p ,ω_f, and ω_a in Equation 1) by analyzing the variability of distinguishing features within the application domain. In this sense, the type of distinguishing features that presents greater variability among definitions of entity classes is considered more important in the similarity assessment than the type of features that does not contribute significantly to distinguish these entity classes. The variability of a type of feature t (P_t^v) is based on the inverse of the frequency with which each distinguishing feature of this type characterizes an entity class in the domain (Equation 5). In P_t^v, o_i is the number of occurrences of a feature in the entity class representations, n is the number of entity classes, and l is the number of features in the application domain.

$$P_t^v = 1 - \sum_{i=1}^{l} \frac{o_i}{n \cdot l} \tag{5}$$

The final weights ω_p, ω_f, and ω_a (Equation 1) are functions of the variability of a type of feature with respect to the variability of the other two types of features (Equations 6a-c).

$$\omega_p = \frac{P_p^v}{(P_p^v + P_f^v + P_a^v)} \tag{6a}$$

$$\omega_f = \frac{P_f^v}{(P_p^v + P_f^v + P_a^v)} \tag{6b}$$

$$\omega_a = \frac{P_a^v}{(P_p^v + P_f^v + P_a^v)} \tag{6c}$$

As an example, consider the set of entity class definitions in Table 2. Assuming that the application domain is formed only by these entity classes, the number of entity classes in the application domain is 4 (n), the number of different parts, functions, and attributes in this domain (l) are 12, 5, and 10, respectively, and P_a^v, P_f^v, and P_a^v are 0.54, 0.67, and 0.15, respectively. Finally, weights ω_p, ω_f, and ω_a of Equations 6a-c are 0.40, 0.49, and 0.11, respectively.

When the application domain has maximum variability (i.e., when no feature is shared by entity classes or only one entity class is part of the application domain) the weights for parts, functions, and attributes turn out to be equal. Similar results occur when the application domain has no variability.

4.2.2 Commonality

This approach associates the relevance of distinguishing features with the feature's contribution to the characterization of the application domain. When users specify an application domain, they are implicitly classifying entity classes that are of interest to the application. These entity classes share some features that make them subjects of interest. For example, when the user's intention is to find a place where to play a sport, a greater weight for this feature in the similarity assessment results in higher similarity values among those entity classes where people can *play a sport*.

This approach defines weighted values for the similarity among parts, functions, and attributes (ω_p, ω_f, and ω_a in Equation 1) by analyzing the frequency with which each distinguishing feature of this type characterizes an entity class in an application domain, that

is, the inverse of the measure given by the variability approach (Equation 7). High frequency translates into high relevance. In P_t^c, o_I is the number of occurrences of a feature in the entity class definitions, n is the number of entity classes, and l is the number of features in an application domain.

$$P_t^c = \sum_{i=1}^{l} \frac{o_i}{n \cdot l} = 1 - P_t^v \tag{7}$$

As in the variability approach, the final weights (ω_p, ω_f, and ω_a in Equation 1) are functions of the frequency of occurrence of a type of feature with respect to the frequency of occurrence of the other two types of features (Equations 8a-c).

$$\omega_p = \frac{P_p^c}{(P_p^c + P_f^c + P_a^c)} \tag{8a}$$

$$\omega_f = \frac{P_f^c}{(P_p^c + P_f^c + P_a^c)} \tag{8b}$$

$$\omega_a = \frac{P_a^c}{(P_p^c + P_f^c + P_a^c)} \tag{8c}$$

Consider again the example given by the set of entity class definitions in Table 2. The parameters n and l of Equation 5 are also applicable to Equation 7, P_p^c, P_f^c, and P_a^c result in 0.46, 0.33, and 0.85, respectively, and weights ω_p, ω_f, and ω_a in Equations 8a-c are 0.28, 0.20, and 0.52, respectively.

A special case is when the maximum variability occurs, that is, each distinguishing feature characterizes only one entity class. In such a case, the commonality among parts, functions, and attributes is zero and the model assigns equal importance to parts,, functions, and attributes. The same weights are also obtained when either an application domain has only one entity class or entity classes share all features. When there are no common features among the entity classes, the similarity values are zero, regardless of the assignment of weights. Likewise, when features are shared by all entity classes, the similarity values are 1.0, independent of the assignment of weights.

The use of frequency for the determination of weights in the similarity model resembles the weight determination in models for information retrieval (Baeza-Yates and Ribeiro-Neto 1999). Unlike information retrieval models that calculate weights independently of a domain application by considering always the whole set of documents in a data collection, this work makes the determination of weights context-dependent, since the domain of concepts from where the frequency of distinguishing features is determined changes depending on the user's intention.

5. Using the Matching-Distance Similarity Measure

To illustrate how contextual information is specified and used to derive the domain of an application, we use an example with an ontology derived from the combination of definitions in WordNet (Miller 1990) and in the Spatial Data Transfer Standard (SDTS) (USGS 1998). This ontology focuses on the spatial domain and contains 257 definitions of entity classes that are associated with the entities defined in SDTS, complemented with synonyms, part-whole relations, and is-a relations obtained from WordNet. Functions of entity classes were

determined by analyzing the natural-language definitions and extracting the verbs in those definitions, augmented by common sense. For example, Table 4 gives an example of the definition of a *stadium* that was derived from the combination of WordNet and SDTS. Attributes of *stadium* were derived from the attributes of *stadium* and semantically related entities in SDTS. These semantically related entities were determined by using the semantic relations defined in WordNet. For example, some attributes were added to the definition of *stadium*, because they were inherited as attributes coming from a superclass (e.g., attributes of construction). The distinguishing feature *function* was derived as an action (verb) associated with sports and entertainment events.

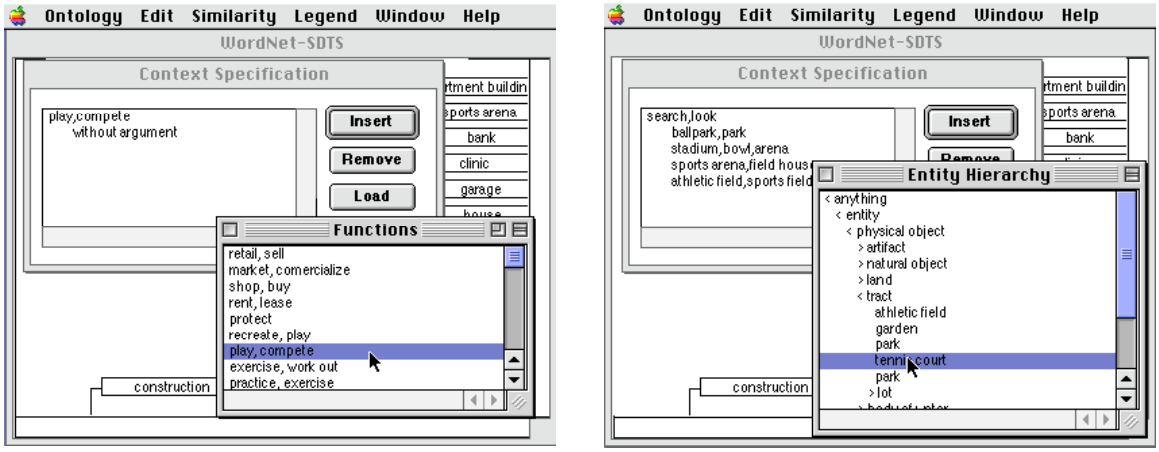
Stadium (WordNet + SDTS)	**Stadium (WordNet)**
entity_class { **name:** {stadium,ball,arena} **description:** large often unroofed structure in which athletic events are held **is_a:** {*construction**} **part_of:** {} **whole_of:** {*athletic_field**} **parts:** {{athletic_field,sports_field,playing_field},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

Table 4: Entity_class definition of *stadium* in WS and WordNet. (*x** denotes a pointer to the entity class *x*)

Based on this ontology, we consider three different scenarios of contextual information:

- Context-1: The user's intention is to play a sport.
- Context-2: The user's intention is to compare downtowns.
- Context-3: The user's intention is to assess a transportation system.

The first scenario (Context-1) represents a domain of entity classes where a person can play a sport. The contextual information for this scenario could be expressed by an intentional specification of context (i.e., by specifying that all entity classes in the domain have the function *play*) (Figure 2a), or by an extensional specification of context (i.e., by listing all the entity classes in the ontology that are of the user' interest) (Figure 2b).



(a) (b)

Figure 2: Context Specification for a User who Searches for a Place to Play a Sport: (a) Intentional Specification of Context and (b) Extensional Specification of Context.

What matters is to obtain an application domain with all the entity classes that are in fact of interest to the user. The latter context specification is more tedious, and in some cases, impractical. It may be, on the other hand, a more accurate specification of the user's interest than an intentional context specification. A portion of the application domain derived from the intentional context specification is shown in Figure 3. In this case, the application domain corresponds to 3% of the entire ontology.

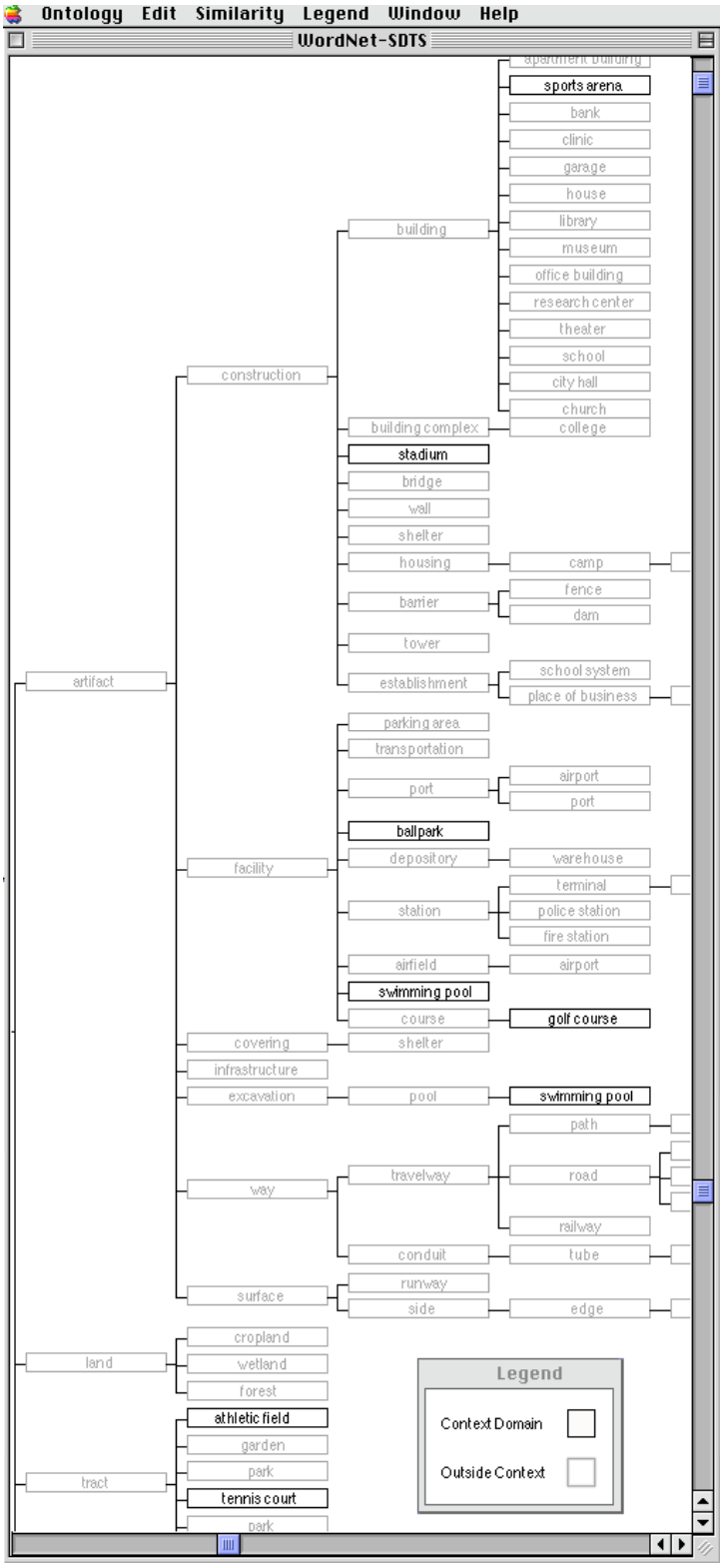


Figure 3: Application Domain for a User who Searches for a Place to Play a Sport.

In the same way that Context-1 was specified, Context-2 and Context-3 were defined in an intentional manner. The specification is done with a general operation (i.e., *compare* and *assess* for Context-2 and Context-3, respectively) and a general entity class whose subclasses or parts are included in the application domain (i.e., *downtown* and *transportation system* for Context-2 and Context-3, respectively). The application domain in the case of Context-2 represents 30% of the ontology and in the case of Context-3, it represents 7% of the ontology.

Table 5 displays the sets of weights for parts, functions, and attributes that result from the definition of the three scenarios and using the variability and commonality approaches.

	Commonality			**Variability**		
Context	ω_p	ω_f	ω_a	ω_p	ω_f	ω_a
(play,{sport})	9%	62%	29%	46%	19%	35%
(compare,{downtown})	10%	13%	77%	36%	35%	29%
(assess,{transportation_system})	4%	29%	67%	45%	35%	20%

Table 5: Weights (%) for Different Specifications of Context Based on the Commonality and Variability Approaches.

Table 6 presents results of the similarity evaluations between a *stadium* and a portion of the entire ontology based on the commonality and variability approaches. While variability highlights differences that decrease the similarity values, commonality emphasizes similarities that increase the similarity values. Similarity values among entity classes vary not only in terms of absolute values, but also in terms of ranks. These results suggest that changes occur depending on context specification as well as in terms of approaches to determining weights. In terms of weight determination, the commonality approach produces more variation in the ranks than the variability approach. Overall, drastic changes in terms of ranks are rare, and it is still possible to distinguish the group of most similar entity classes.

Entity	Context-1		Context-2		Context-3	
	c	*v*	*c*	*v*	*c*	*v*
Sports arena	0.86	0.69	0.68	0.75	0.74	0.75
Athletic field	0.91	0.66	0.85	0.73	0.90	0.68
Theater	0.23	0.45	0.54	0.36	0.45	0.35
Ball park	0.88	0.67	0.73	0.74	0.79	0.72
Commons	0.43	0.29	0.52	0.32	0.53	0.27
Museum	0.20	0.36	0.50	0.29	0.42	0.27
Tennis court	0.86	0.48	0.77	0.59	0.84	0.52
Transportation	0.10	0.07	0.15	0.06	0.13	0.04
Library	0.19	0.31	0.48	0.25	0.41	0.22
Building	0.21	0.34	0.54	0.27	0.46	0.23
House	0.18	0.30	0.46	0.24	0.39	0.21

Table 6: Example of Similarity Values between a *Stadium* and a Portion of the WordNet-SDTS Ontology for Three Different Scenarios of Contextual Information. (Symbol *c* denotes the commonality approach and symbol *v* denotes the variability approach)

A characteristic of the commonality and variability approaches is their sensitivity to the set of entity classes defined in the ontology. This sensitivity becomes more important for a narrow application domain, where the omission of one entity class may affect the determination of common and different features of the application domain. To check this sensitivity, we performed similarity evaluations that involve the same narrow application domain (i.e., Context-1), but using slightly different ontologies. The first case contains the default ontology that contains seven entity classes in the application domain: *sports arena*, *stadium*, *athletic field*, *swimming pool*, *golf course*, *ballpark*, and *tennis court*. Subsequent cases eliminate one by one entity classes of the ontology to reduce the application domain (i.e., *sports arena*, *golf course, tennis court*, and *athletic field* are eliminated). Table 7 shows the changes of weights for parts, functions, and attributes based on the commonality and the variability approaches, using subsets of the default application domain.

Case	Application Domain	Commonality			Variability		
		ω_p	ω_f	ω_a	ω_p	ω_f	ω_a
1	Default	9%	62%	29%	46%	19%	35%
2	(1) – *sports arena*	8%	57%	35%	48%	20%	32%
3	(2) – *golf course*	12%	52%	36%	47%	22%	31%
4	(3) – *tennis court*	15%	52%	33%	44%	23%	33%
5	(4) – *athletic field*	19%	56%	25%	37%	27%	36%

Table 7: Weights Based on the Same Context Specification and Different Ontologies.

The main trend in the weights of distinguishing features for Context-1 remains stable across different application domains (i.e., commonality highlights functions whereas variability highlights parts). Although changes may occur depending on the set of entity classes in the ontology, MDSM is robust enough to capture the main property of the application domain and allows a systematic way to determine the features' relevance for similarity assessment.

The next section describes a human-subject experiment that tests whether the results given by MDSM are compatible with people's judgments.

6. Assessing the Matching-Distance Similarity Measure

Previous studies on similarity assessment have studied the correlation between computational similarity models and people's judgments of similarity. The models have used the WordNet's taxonomy (Miller *et al*. 1990) and the Brown Corpus of American English (Francis and Kucera 1982) for the determination of word frequency. For the human-subject experiment, these studies have used the experiment by Miller and Charles (1991), which gave to 38 undergraduate subjects 30 pairs of nouns that were chosen to cover high, intermediate, and low levels of similarity as determined by a previous study (Rubenstein and Goodenough 1965). These studies found a correlation of 0.60 using a semantic-distance approach, 0.79 using an information-content approach, and 0.83 using an extended-distance approach (Jiang and Conrath 1997, Resnik 1999).

In assessing MDSM, we designed a new experiment, because our goal was to evaluate similarity under different contexts and because previous studies have not analyzed similarity within a specific domain (i.e., entity classes of a specific domain that are semantically related). The experiment consisted of five questions with sets of entity classes that subjects were asked to rank according to their judgments of similarity with a specified base entity class. Four of the five questions involved entity classes of a constructed kind, such as a building and a road. The last question addressed the similarity assessment among large geographic entities, such as a *lake*, a *desert*, and a *forest*. In this sense, the experiment attempted to capture any divergence in the similarity assessment of objects of different kinds—natural vs. constructed. We explored this divergence between natural and constructed objects, because we had previously detected differences in how people describe and relate objects from different scale and nature (Rodríguez and Egenhofer 2000). We expect, for example, that characterizing objects in terms of what a user can do with them may be easier for constructed objects than for natural objects.

The first three questions asked users to judge the same set of entity classes, but using different contextual information. Question 1 represented the default case of similarity assessment with no explicit contextual information. Questions 2 and 3 specified context defined as desired operations (i.e., "play a sport" and "compare constructions," respectively). Question 4 used a set of transportation-type entities, which became the contextual information for this question. As in the first question, the last question assumed the default case of a similarity assessment (i.e., no explicit contextual information). Questions also differed in the relation between the set of entity classes that were actually compared and the application domain that was derived from the context specification in MDSM. This relationship may yield some interesting conclusions, since the sets of entity classes that were actually compared in each question were described as contextual information that may influence the similarity evaluations (Tversky 1977, Krumhansl 1978). For instance, Question 2 contains *ballpark* (i.e., an entity class in the application domain) and *library* (i.e., an entity class outside of the application domain). Among all entity classes evaluated, Question 2 includes 50% of entity classes that are outside of the application domain, Question 3 has 45% of entity classes that are outside of the application domain, and Question 4 contains only entity classes in the application domain.

Two questionnaires were prepared (Survey A and Survey B) with the same set of entity classes, but with different targets for the similarity evaluations. These different targets are related by either an is-a relation or a part-whole relation. For example, Questions 1-3 in

Survey A asked for entity classes that are similar to a *stadium*, while Questions 1-3 in Survey B asked for entity classes that are similar to an *athletic field*, which is part of a *stadium*. Likewise, Question 4 in Survey A asks for entity classes that are similar to a *travelway*, whereas Question 4 in Survey B asked for entity classes similar to a *path*, which is a subclass of *travelway*. Each entity class used in the experiment has its corresponding definition in the ontology of the similarity measure. As in the example of the previous section, we used the ontology derived from the combination of WordNet and SDTS, since it contains all desired components of the entity class representation. Since the goal of the experiment was to evaluate the similarity measure rather than the entity class definitions per se, subjects were asked to judge similarity based on the set of definitions that were provided to them during the experiment and used by MDSM.

Seventy-two students (forty-three female and twenty-nine male) of an undergraduate English class participated in the experiment. A group of thirty-seven students (twenty female and seventeen male) answered Survey A and a group of thirty-five students (twenty-three female and twelve male) answered Survey B. For all subjects U.S. English was their mother tongue, and their ages ranged from 18 to 36 years. Each subject was paid for participating in the experiment and answered the questions at the same time and in less than twenty minutes.

We analyzed the results by Kendall's coefficient of concordance for studying the association among subjects' responses and by the Spearman rank correlation coefficient for studying the correlation between the model and subjects' responses (Gibbons 1976, Daniel 1978). We normalized tied ranks by using the mean of the ranks for which they tie, assuming a number of ranks equal to the number of entity class comparisons. As the best estimator of the true similarity rank of entity classes, we consider the average rank assigned to an entity class by subjects. Table 8 summarizes these results for each question in each survey. For the similarity evaluations with MDSM we use the default setting, in which distinguishing features are equally important, and we use the settings obtained from the commonality and variability approaches to weigh determination.

	Question		**Association**	**Correlation**		
	Target	Context		D	C	V
Survey A	Stadium	Null Context	0.70	0.96	-	-
	Stadium	Play a Sport	0.76	0.83	0.85	0.68
	Stadium	Compare constructions	0.37	0.95	0.87	0.96
	Path	Compare transportation systems	0.45	0.90	0.78	0.96
	Lake	Null context	0.64	0.78	-	-
Survey B	Athletic field	Null context	0.69	0.92	-	-
	Athletic field	Play a Sport	0.64	0.87	0.87	0.88
	Athletic field	Compare constructions	0.33	0.90	0.87	0.91
	Travelway	Compare transportation systems	0.45	0.88	0.84	0.91
	Lake	Null context	0.70	0.86	-	-

Table 8: Association Among Subjects' Responses and Correlation between MDSM and Subjects' Responses. (Symbol D denotes default setting, C denotes settings obtained from the commonality approach, and V denotes settings obtained from the variability approach)

The results of the human-subject experiment support the use of MDSM for semantic similarity among entity classes. We found a correlation between 0.78 and 0.96 with the default setting. Taking one of the surveys and evaluations that do not consider contextual information, Table 9 shows the Spearman rank correlation coefficient between subjects' responses and the results in ranks of MDSM, a basic semantic-distance model (Rada *et al*. 1989), and an information-content model (Resnik 1999). An important observation is that although subjects' responses are associated, the degree of concordance among subjects' answers is lower than the degree of concordance of previous experiments on semantic similarity (e.g., 0.90 in Resnik's (1999) experiment). This low degree of concordance may be due to the larger number of entity classes that were evaluated with respect to the same target and due to the use of entity classes that are semantically related.

Question Target	Semantic distance	Information content	Matching distance
Stadium	-0.37	-0.37	0.96
Path	0.78	0.87	0.90
Lake	0.82	0.80	0.78

Table 9: Spearman Rank Correlation Coefficient Using a Semantic-Distance Model, an Information-Content Model, and the Matching-Distance Model.

The experiment shows a small improvement in performance (6% in the best case) when weights of distinguishing features were determined based on contextual information. This improvement is still relevant, since the results are nearing the observed upper bound (i.e., 1.0); however, the major determinant for the high correlation between MDSM and subjects' answers seems to be the correct identification of distinguishing features of entity classes. For example, an important difference between the model and the subjects' answers was the entity class least similar to a *lake*. While the model assigns a *bridge* as the least similar entity class, subjects selected a *desert* as the least similar entity class to a *lake*. This suggests that not only the existence of a prototypical feature, but also the negation of this feature may affect considerably the similarity assessment. In this example, a characteristic of a *desert* is the lack of water, whereas water is the common feature of all entity classes that are similar to a *lake*.

In Question 4 subjects identified a *road* as the most similar entity class to a *path* and *travelway*. This result suggests that although definitions that were given to subjects indicate that *travelway* is a more general concept than *path* and *road*, subjects considered *road* as the prototypical entity for the class transportation. This type of result could lead to a further study that considers the classification of entities in terms of prototypical characteristics rather than necessary and sufficient conditions (Rosch 1975, Mark *et al*. 1999).

The commonality and variability approaches have opposite effects on the results. While one of the approaches increases the correlation, the other one decreases it. Based on the characteristics of the context specification, the commonality approach worked better for context specifications that used specific features, such as playing a sport, whereas the variability approach is appropriate for context specification defined in terms of type of entity classes, such as a construction or transportation type.

7. Conclusions and Future Work

We described the Matching-Distance Similarity Measure (MDSM), a new semantic similarity measure among entity classes. MDSM compares spatial entity classes that are defined in an ontology. Definitions of spatial entity classes are composed of semantic relations (i.e., is-a relations and part/whole relations) and distinguishing features (i.e., attributes, functions, and parts). MDSM combines a feature-based with a distance-based model of similarity. The feature-based model evaluates similarity in terms of distinguishing features that are common or different between definitions, whereas the distance-based model compares entity classes in

terms of their distances in the semantic structure that is defined by the semantic relations. In MDSM, the concept of semantic distance is used as an indicator of the level of abstraction of entity classes that affects the relevance of different features between entity classes.

The main characteristics of MDSM are (1) the handling of asymmetric evaluations for entity classes by taking into account levels of abstraction in the hierarchical structure; (2) the use of is-a and part-whole relations in the entity class representation; (3) the treatment of synonymy and polysemy of entity class names; (4) the weighted contribution to the similarity assessment of distinguishing features; and (5) the systematic use of contextual information for weight determination. An assessment of the similarity measure by using a human-subject experiment showed that the model correlates with people's judgments of similarity.

There are several interesting and useful studies that remain for future work. An important and practical issue in using MDSM is the construction of the ontology. The ontology used in this work was created semi-automatically, by complementing manually, what WordNet and SDTS provide in their concepts' definitions. Functions, however, were added manually. It would be interesting to explore systematic ways for constructing this ontology. Another aspect for future study is an extension of the context specification that considers additional features of entity classes. For example, a user may want to search for sports facilities that have spectator stands. In these cases, although context is still determined by an intended operation, parts and attributes may also describe the desired domain of entity classes. Another interesting area is concerned with similarity reasoning, since such reasoning may allow inferences about the similarity relations among entity classes by using a subset of known similarity relations. For reasoning about similarity, we envision two lines of investigations that are worthwhile to follow. From a cognitive point of view, research could address properties of the composition of semantic relations. In particular, the research question is whether there is any situation or context in which inferences and composition of semantic relations (i.e., is-a and similarity relations) could be solved. From a mathematical point of view, it is interesting to compose measures that result in ranges of possible values of similarity. In this sense, a potential approach is the study of Boolean combinations of graded sets (Fagin 1999) using fuzzy logic (Zadeh 1965). A graded set could be associated with the set of entity classes that have a value of similarity (i.e., grade) with respect to a target.

This study motivates a future area of investigation that is concerned with the semantic comparison of distinguishing features. 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* (Fellbaum 1998) (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.

Finally, we are currently working on the use of MDSM for similarity evaluations of spatial scenes. When scenes are evaluated, an important issue is the structural alignment of objects within these scenes (Goldstone 1994, Markman and Gentner 1993, Gentner and Markman 1997). This structural alignment could be determined based on spatial relations (Bruns and Egenhofer 1996) or based on the role of objects within the scenes.

Acknowledgments

Thanks to Bob Rugg for his contribution to the preliminary work of MDSM, published by Rodríguez *et al.* (1999). Andrea Rodríguez's research was funded by CONICYT, Chile, under FONDECYT project 1010897. Max Egenhofer's research was supported by the National Science Foundation under grant numbers IIS-9613646, IIS-9970123, and EPS-9983432; by the National Imagery and Mapping Agency under grant numbers NMA202-97-1-1023, NMA201-00-1-2009, NMA201-01-1-2003, and NMA401-02-1-2009, the National Institute for Environmental Health Sciences, NIH under grant number 1-R-01-ES09816-01; and by Lockheed Martin Management Data Systems.

References

Baeza-Yates, R., and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York, NY: ACM Press/Addison Wesley.

Bergamaschi, B., Castano, S., De Capitani di Vermercati, Montanari, S., and Vicini, M. (1998). An Intelligent Approach to Information Integration. *In* Guarino, N. (ed.), *International Conference on Formal Ontology in Information Systems*, Tentro, Italy, 253-268.

Berner-Lee, J., Handler, J. and Lassila, O. (2001). The Semantic Web. *Scientific American* 184(5): 34-43.

Birkhoff, G. (1967). *Lattice Theory*. Providence, RI: American Mathematical Society.

Bowdle, F. and Gentner, D. (1997). Informativity and Asymmetry in Comparisons. *Cognitive Psychology* 34: 244-286.

Bruns, T. and Egenhofer, M. (1996). Similarity of Spatial Scenes. in Kraak, M. and Molenaar, M. (eds.), *Seventh International Symposium of Spatial Data Handling*, Delft, The Netherlands, pp. 4A.31-42.

Cruse, D. (1979). On The Transitivity of the Part-Whole Relation. *Linguistics*, 1529-38.

Dahlgren, K. (1988). *Naive Semantics for Natural Language Understanding*. Norwell, MA: Kluwer Academic Publishers.

Daniel, W. (1978). *Applied Nonparametric Statistics*. Boston, MA.: Houghton Mifflin.

Egenhofer, M. (2002). Toward the Semantic Geospatial Web. *In* A. Voisard and S.-C. Chen (eds.), *ACM-GIS*, McLean, VI (in press).

Egenhofer, M. and Shariff, A. (1998). Metric Detail for Natural-Language Spatial Relations. *ACM Transactions on Information Systems* 16(4): 295-321.

Egenhofer, M. and Franzosa, R. (1995). On the Equivalence of Topological Relations. *International Journal of Geographical Information Systems* 9 (2): 133-152.

Egenhofer, M. and Mark, D. (1995). Naive Geography. *In* Frank, A. and Kuhn, W. (eds.), *Spatial Information Theory: A Theoretical Basis for Geographic Information Systems, International Conference COSIT'95*, Semmering, Austria, Springer-Verlag: 1-14.

Egenhofer, M. (2002). Toward the Semantic Geospatial Web. *In* Voisard, A. and Chen, S.-C. (eds.), *ACM-GIS*, McLean, VI, ACM press: 1-4.

El-Kwae and Kabuka, M. (1999). A Robust Framework for Content-Based Retrieval by Spatial Similarity in Image Databases. *ACM Transactions on Information Systems* 17(2): 174-198.

Fagin, R. (1999). Combining Fuzzy Information from Multiple Systems. *Journal of Computer and Systems Sciences*, 5883-99.

Fellbaum, C. (1990). English Verbs as a Semantic Net. *International Journal of Lexicography*, 3(4), 270-301.

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

Flewelling, D. (1999). Measuring Similarities of Spatial Datasets. *URISA Journal,* 11(1): 45-52.

Fonseca, F., Egenhofer, M., Agouris, O. and Câmara, G. (2002). Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS* 6(3): 121-151.

Fonseca, F., Martin, J and Rodríguez, A. (2002). From Geo to Eco-Ontologies. *Geographic Information Science 2002*. *Lecture Notes in Computer Science* Vol. 2478, Berlin: Springer: 93-107.

Fonseca, F. and Sheth, A. (2003). *The GeoSpatial Semantic Web*, Research Priorities Series, University Consortium for Geographic Information Science (http://www.personal.psu.edu/faculty/f/u/fuf1/Fonseca-Sheth.pdf).

Francis, W. and Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston, MA: Houghton Miffin.

Frank, A. (2001). Tiers of Ontology and Consistency Constraints in Geographical Information Systems. *International Journal of Geographical Information Science* 15(7): 667-678.

Gale, W., Church, K. and Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and Humanities*, 26(5/6), 415-450.

Gentner, D. and Markman, A. (1997). Structure Mapping in Analogy and Similarity. *American Psychologist* 52(1): 45-56.

Gibbons, J. (1976). *Nonparametric Methods for Quantitative Analysis*. Columbus, OH: American Sciences Press.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

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

Goldstone, R., Medin, D. and Halberstadt, J. (1997). Similarity in Context. *Memory and Cognition*, 25(2), 237-255.

Goñi, A., Mena, E. and Illarramendi, A. (1998). Querying Heterogeneous and Distributed Data Repositories Using Ontologies. *In* Charrel, P.-J. and Jaakkola, H. (eds.), *Information Modelling and Knowledge Base IX*, IOS Press: 19-34.

Goyal, R. and Egenhofer, M. (2001). Similarity of Direction Relations. In C. Jensen, M. Schneider, B. Seeger, V. Tsotras (eds.), *Seventh International Symposium on Spatial and Temporal Databases*, Los Angeles, CA, Lecture Notes in Computer Science, Vol. 2121, Springer: 36-55.

Gruber, T. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human and Computer Studies*, 43(5/6), 907-928.

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

Guarino, N. and Giaretta, P. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. *In* Mars, N. (ed.), *Toward Very Large Knowledge Base: Knowledge Building and Knowledge Sharing*, Amsterdam, The Netherlands, IOS Press: 25-32.

Herskovits, A. (1997). Language, Spatial Cognition, and Vision. *In* Stock, O. (ed.), *Temporal and Spatial Reasoning*. Dordrecht, The Netherlands: Kluwer Academic Press**:** 155-202.

Holman, E. (1979). Monotonic Models for Asymmetric Proximities. *Journal of Mathematical Psychology* 20:1-15.

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

Jiang, J. and Conrath, D. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In Proceedings of the International Conference on Computational Linguistics (ROCLING X)*, Taiwan: 19-35.

Kavouras, M. and Kokla, M. (2002). A Method for the Formalization and Integration of Geographic Categorizations. *International Journal of Geographic Information Science*, 16(5): 439-453.

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

Kuhn, W. (2000). How to Produce Ontologies. *Geographical Domain and Geographic Information Systems —EuroConference on Ontology and Epistemology for Spatial Data Standards*, La-Londe-Les Maures, France.

Kim, Y. and Kim, J. (1990). A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph. *Journal of Documentation*, 46(2), 113-136.

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

Leacock, C., Towell, G. and Voorhees, E. (1993). Corpus-based Statistical Sense Resolution. *In Proceedings of the ARPA Human Language Technology Workshop*, San Francisco, CA, Morgan Kaufmann: 260-265.

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

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *In Proceedings of the Fifteenth International Conference on Machine Learning*, San Francisco, CA, Morgan Kaufmann: 296-304.

Mark, D., Freksa, C., Hirtle, S., Lloyd, R. and Tversky, B. (1999). Cognitive Models of Geographic Space. *International Journal of Geographical Information Science*, 13(8), 747-774.

Markman, A. and Gentner, D. (1993). Splitting the Differences: A Structural Alignment View of Similarity. *Journal of Memory and Language* 32: 517-535.

Medin, D., Goldstone, R. and Gentner, D. (1993). Respects for Similarity. *Psychological Review* 100(2): 254-278.

Mena, E., Kashyap, V. and Sheth, A. (1996). OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies. *In Proceedings of the International Conference on Cooperative Information Systems (CoopIS'96)*, Brussels, Belgium, IEEE Computer Society Press: 14-25.

Miller, G. (1990). Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography*, 3(4), 245-264.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39-41.

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

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

Osherson, D. and Smith, E. (1981). On the Adequacy of Prototype Theory as a Theory of Concepts. *Cognition*, 9(1), 35-58.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

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

Resnik, O. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research*, 1195-130.

Richardson, R. and Smeaton, A. (1995). Using WordNet in a Knowledge-Based Approach to Information Retrieval, Technical, CA-0395, Dublin City University, School of Computer Applications.

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

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

Rodríguez, A., and Egenhofer, M. (2000). Comparison of Inferences about Containers and Surfaces in Small-Scale and Large-Scale Spaces. *Journal of Visual Language and Computing*, 6(6), 639-662.

Rodríguez, A., and Egenhofer, M. (2003). Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2), 442-456.

Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology*, 104, 192-233.

Rosch, E. and Mervis, C. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7, 573-603.

Rubenstein, H. and Goodenough, j. (1965). Contextual Correlates of Synonymy. *CACM*, 8(10), 627-633.

Schenkelaars, V. and Egenhofer, M. (1997). Exploratory Access to Geographic Libraries. *ACSM/ASPRS Autocarto 13*. Seattle, VA.

SDTS (1992). Spatial Data Transfer Standard (SDTS). 173, F. I. P. S. P., U.S. Department of Commerce: U.S. Government Printing Office. FIPS 173.

Smeaton, A. and Quigley, I. (1996). Experiment on Using Semantic Distance Between Words in Image Caption Retrieval. *In Proceedings of the 19th International Conference on Research and Development in Information Retrieval SIGIR'96*, Zurich, Switzerland: 174-180.

Smith, E. and Osherson, D. (1984). Conceptual Combination with Prototype Concepts. *Cognitive Science*, 8, 337-361.

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

Sussna, M. (1993). Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. *In Proceedings of the Second International Conference on Information Knowledge Management, CIKM'93*: 67-74.

Talmy, L. (1983). How Language Structures Space. *In* Pick, H. and Acredolo, L. (eds.), *Spatial Orientation*. New York: Plenum Press**:** 225-282.

Tversky, A. (1977). Features of Similarity. *Psychological Review*, 84(4), 327-352.

Voorhes, E. (1998). Using WordNet for Text Retrieval. *In* Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press: 285-303.

USGS (1998). View of the Spatial Data Transfer Standard (SDTS) Document. 6/12/98, http://mcmcweb.er.usgs.gov/sdts/standard.html

Weinstein, P. and Birmingham, P. (1999). Comparing Concepts in Differentiated Ontologies. *In Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling, and Management*, Banff, Canada:

Winston, M., Chaffin, R. and Herramann, D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11417-444.

Yoshitaka, A. and Ichikawa, T. (1999). A survey on Content-Based Retrieval for Multimedia Databases. *IEEE Transactions on Knowledge and Data Engineering* 11(1): 81-93.

Zadeh, L. (1965). Fuzzy Sets. *Information and Control*, 8338-353.