ASSESSING SEMANTIC SIMILARITY AMONG SPATIAL ENTITY CLASSES

By

María Andrea Rodríguez

B.S. Universidad de Concepción – Chile, 1987

M.S. University of Maine, 1997

A THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
(in Spatial Information Science and Engineering)

The Graduate School
University of Maine
May, 2000

Advisory Committee:

Max J. Egenhofer, Professor of Spatial Information Science and Engineering, Advisor

Paul C. Bauschatz, Associate Professor of English

M. Kate Beard-Tisdale, Professor of Spatial Information Science and Engineering

Werner Kuhn, Professor of Geoinformatics and Digital Cartography, University of Münster, Germany

David M. Mark, Professor of Geography, State University of New York, Buffalo

Robert D. Rugg, Professor of Urban Studies and Planning, Virginia Commonwealth University, Richmond

Acknowledgments

I would like to thank many people who provided me with support, guidance, and care that help me to conclude this work.

First, I gratefully acknowledge the guidance and support from the members of my advisory committee, Max Egenhofer, Paul Bauschatz, Kate Beard-Tisdale, David Mark, Werner Kuhn, and Robert Rugg. I am particularly grateful to Max Egenhofer for being a friendly, helpful, and critical advisor, who gave me the freedom needed to develop my research.

Second, I would like to thank the interest and encouragement from Dick Berg and Walt Senus of the National Imagery and Mapping Agency and from Mitchell Werther, Kristin Fishburn, and George Nehrebeckyj of Lockheed Martin. I also would like to express my gratitude to Pat Hayes and Gio Wiederhold for their feedback on my work and to George Miller and Nicola Guarino for their prompt responses to my questions.

Third, to all my colleagues and friends in the Department of Spatial Information Science I would like to thank you for sharing the good and bad moments of my study life. I feel fortunate for having being part of a friendly environment that made my Ph.D. program an enjoyable and unforgettable experience.

Fourth, I thank the support and funding from the University of Concepción, Chile, and the initial funding from the Fulbright foundation. Further funding from the National Center of Geographic Information and Analysis, the National Imagery and Mapping Agency, and Lockheed Martin are gratefully acknowledged.

Most important, I thank the continuous support, love, and patience of Christian and Alicia. This long journey would not have been possible without them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Assessing similarity is a judgment process that requires two “things” to be decomposed into elements in which they are the same and elements in which they are different. These types of judgments are typically intuitive, subjective, and part of the everyday life such that they usually display no strict mathematical models (Tversky 1977). In information systems, similarity assessment is part of several processes, such as information retrieval, information integration, and data maintenance. Similarity assessment is particularly important for geographic information systems (GISs), because users of spatial data have diverse backgrounds and no precise definitions underlie the matter of discourse. Satisfactory definitions of geographic phenomena, such as a mountain, the extent of a village, and the boundary of a valley, are difficult to obtain (Fisher and Wood 1998), and spatial properties, such as shape, location, and spatial relations, have varied formalizations. As a result, data stored in a spatial database represent particular views of reality. By using spatial query languages users are able to express an approximation of what they want to retrieve, which is likely an inexact match with any stored data.

This thesis focuses on the semantics of spatial entities and proposes a computational model for assessing semantic similarity among spatial entity classes. Much past research in spatial information science that is concerned with similarity

assessments has focused on the geometric properties of spatial information. Examples of these studies are topological equivalence (Paiva 1998), cardinal direction between extended spatial objects (Egenhofer and Goyal in press), metric details of spatial relations (Egenhofer and Shariff 1998), and content-based image retrieval (Flickner *et al*. 1995). While omitting the geometric properties of spatial objects, this work concentrates on the cognitive properties of the semantic similarity assessment that relate to the spatial domain and leaves for future work the integration of geometric and semantic similarity.

For this thesis, the term entity classes denotes concepts about the real world. These concepts about the real world are cognitive representations that people use to recognize and categorize entities or events in the real world (Dahlgren 1988). In this sense, this work has a top-down approach by starting from the semantics of entities in the real world instead of the semantics of data stored in a database (Sheth 1995). Consequently, this thesis considers studies done by cognitive scientists in the area of knowledge and behavior as well as by computer scientists in the domain of artificial intelligence.

The main motivation for this thesis is the need to enhance geographic information systems with better mechanisms for information retrieval and integration. 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 the research of information retrieval stress the advantages of using domain knowledge and semantic similarity functions to compare words or documents (Ginsberg 1993, Lee *et al.* 1993, Richardson and Smeaton 1995, Voorhees 1998). By introducing the semantic

knowledge of spatial concepts, this thesis creates a similarity model that can obtain flexible and better matching between user-expectated 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.

Models for semantic similarity among entities have usually been addressed from two different perspectives. Psychologists and cognitive scientists have analyzed how people evaluate similarity and have defined models based on features or descriptors of concepts. This approach is in contrast to the work by computer scientists who usually define semantic similarity as the semantic distance among concepts within a hierarchical structure. These two different approaches have advantages and disadvantages that complement each other. This thesis defines a similarity measure that combines distinguishing features with semantic relations to create a model that is not only computationally feasible, but also satisfies cognitive properties of similarity assessment.

1.1 Similarity Assessment in Geographic Information Systems

New trends in science and technology have produced an increasing expectation for more intelligent, efficient, and reliable information systems. People are not expecting to retrieve data, but to find information, that is, data that are meaningful to them. The large amount of data and the need for the integration of autonomous and heterogeneous databases have increased the requirements and made information retrieval and integration essential components of current information systems.

1.1.1 Information Retrieval

In traditional information systems users express what information they need through queries, which can be a set of a Boolean combination of keywords, natural language statements, or user-system dialogs. With the advance of technology, systems deal with more diverse types of digital information (e.g., images, maps, sounds, and characters) causing a growing interest in new forms of user interfaces (Blaser *et al.* in press, Bruns and Egenhofer 1997, Egenhofer 1997). A desirable characteristic of query languages and user interfaces is that users can retrieve and search information without the requirement of knowing the name and structure within which data are stored. This characteristic of query languages is a basic principle for the design of data-manipulation languages of database systems, where the logical access of data is separated from their physical access (Silberschatz *et al.* 1996). Due to the enormous amount of data stored in a database and the fact that names of the data structure may not reflect the nature of the information they contain, it is unrealistic to expect that users could know and directly use those names. Therefore, in order to make progress in this area users should be able to express queries in terms that are familiar to them (Mark and Gould 1991, Richardson and Smeaton 1996).

Once a user has expressed a query, the system performs a matching process between the query and the internal representation. In the past, most approaches to information retrieval have computed similarity between queries and stored data based on a statistical analysis of index terms and have treated terms in isolation from their contexts (Meadow *et al.* 2000). These approaches soon reach their limits since they deal with syntactic but not semantic correspondences, and users may express the same concept in different ways (Lee *et al.* 1993). A falacy of today's methods for querying information is the systems' assumption that the user's query represents precisely what

the user wants. It is common, however, that a system does not find an exact match or that users are interested in data that match the query partially. For example, a user query for a geographic database could be to find cities in the state of New York with at least one university. If the retrieval process is constrained by searching for an exact match, it will ignore towns or colleges in generating the query answer. Semantic similarity assessment goes beyond the determination of an exact matching between queries and stored data, because it provides a range of possible answers depending on conceptually similar terms and gives the users the possibility to choose among them. Thus, a semantic similarity function is a tool for exploratory access to data. It resembles browsing, because users do not know in advance what they are looking for (Schenkelaars and Egenhofer 1997); however, browsing is highly interactive and leaves all the choices to the users.

Geographic information systems manipulate large collections of spatial scenes. Spatial scenes consist of sets of objects represented by their spatial relations —topological relations, distance relations, and direction relations—as well as by other geometric characteristics—shape, size, and density—and attributes specifying the semantics of the spatial object—entity type classification. Initially geographers investigated similarity assessment of point sets for spatial analysis (Unwin 1981). More recent studies have investigated the spatial similarity for content-based image retrieval (Bimbo *et al.* 1994, Bruns and Egenhofer 1996, Faloustsos *et al.* 1994, Papadias *et al.* 1998, Park and Golshani 1997). In those studies, the visual similarity of images usually relies on a judgment in terms of visual descriptions, such as shape, size, texture, and color. For similarity of spatial configurations, on the other hand, the spatial arrangement of objects becomes the subject of comparison. This spatial arrangement is typically expressed by a set of constraints about directions (e.g., north and south), topology (e.g., inside, overlap), and distances (e.g., 5 miles).

Semantic similarity assessment ignores some of the spatial datasets' geometric properties, such as density, dispersion, and pattern derived from representative subsets (Flewelling 1997) and extent and location displayed by magic lenses (Schenkelaars and Egenhofer 1997). The classification of geographic entities, however, is spatial, even when no geometry is involved. Non-geometric concepts, such as building, road, and place, are spatial concepts that are used for describing the semantics of spatial objects. By studying the similarity among spatial concepts that underlie people's spatial descriptions, this research lies in the field of *Naive Geography* (Egenhofer and Mark 1995), a field of study that is concerned with formal models of commonsense worlds.

Computer scientists working on traditional information retrieval have addressed similarity assessment for semantic information (Kim and Kim 1990, Lee *et al.* 1993, Richardson and Smeaton 1995). The main problems faced in those studies are the resolution of ambiguous terms and the multiple ways in which the same concept can be expressed. Recent studies have investigated the use of knowledge bases and semantic similarity functions as a mechanism to compare terms (Jiang and Conrath 1997, Lee *et al.* 1993, Smeaton and Quigley 1996, Voorhees 1998). Many strategies involving a knowledge base and a similarity function aim at solving the problem of information retrieval for a general domain. They have searched for an automatically constructed knowledge base that contains entries for all concepts used in natural language (Jiang and Conrath 1997, Richardson and Smeaton 1995, Richardson *et al.* 1994). Another strategy for a knowledge-based approach to information retrieval has been to work on a specific domain and create a controlled vocabulary (Monarch and Carbonelli 1987, Rada *et al.* 1989). This thesis focuses on the spatial information domain to avoid the pitfalls of trying to obtain a general knowledge base that satisfies and represents the information requirements for each domain. The limited success of C, a ten-year project

of generating a generic common-sense knowledge base (Lenat and Guha 1990), is testimony for the need of alternative approaches.

In order to clarify the use of similarity assessment in a GIS, consider a user who wants to retrieve information from a spatial database about hospitals that are within her district. This query is composed of the semantic components (i.e., the notion of hospital) and the geometric component (i.e., the spatial location defined by the user's district). Based on the user's query, possible scenarios for the retrieval of information are:

- The database contains one or more hospitals within the user's district.
- The database contains hospitals, but they are outside of the user's district. Among the existing hospitals outside of the district, some are closer and some further away from the district.
- The database does not contain hospitals in the user's district, but it contains clinics and health centers.
- The database does not contain hospitals, even nearby the user's district, but it contains clinics or health centers in adjacent districts.

Only the first scenario satisfies an exact matching (spatial and semantic match) of the query and is addressed by today's spatial query languages, such as the spatial SQL (Egenhofer 1994). Although the second scenario does not correspond exactly to what the user requested, it may provide relevant information about hospitals that are close to the user's geographic area of interest. This type of scenario requires spatial similarity methods such as those addressed by sketch-based query languages and image retrieval (Bruns and Egenhofer 1996, Park and Golshani 1997). The third scenario

requires the identification of semantically similar objects that could also be relevant for the user, because it may be sufficient for a user to find a clinic in the district. The last scenario combines spatial and semantic similarity models and represents the ultimate goal for the design of a spatial query language. Among these four scenarios, this thesis contributes primarily to the third one and provides the foundation to solving in the future queries of the fourth type.

As in the case of a single database, environments with multiple and heterogeneous databases require mechanisms for information retrieval that allow the identification of semantically similar objects. For instance, consider the same user who wants to retrieve hospitals within her district. It may happen that the information retrieval is done from different databases and that a user distinguishes between hospitals, clinics, and health centers, whereas a database groups them together into the concept of a health care provider. In such a case, the user would expect that the system indicates that health care providers are semantically similar to the objects she requested.

1.1.2 Information Integration

Information integration is a basic requirement for modern information systems (Sheth 1999) that differs from data integration, because it combines only the selected information that is derived from data sources (Wiederhold and Jannink in press). Some of the main reasons for the growing interest in information integration are the improvement in the interconnection of distributed computing systems (i.e., the Internet) and the need for the reuse and sharing of data. Heterogeneity among data stored in information systems makes the integration of information a challenging area of research. In the spatial domain, in particular, the complexity and diversity of spatial data are major issues for interoperating GISs.

In environments with multiple and autonomous databases three different architectures for locating and accessing information have emerged: (1) global schema integration, (2) federated database systems, and (3) multidatabase languages (Elmagarmid *et al.* 1999). A global schema integration provides a consistent and uniform view of and access to data through a single view of multiple databases (Spaccapietra and Parent 1994). This approach constrains the autonomy of databases and becomes impractical as many databases are interconnected and databases update their local data. A federated database system (FDBS) is a collection of cooperating but autonomous heterogeneous database systems (Sheth and Kashyap 1992, Sheth and Larson 1990) that represent a compromise between total integration and no integration (Bouguettaya *et al.* 1998). The level of integration depends on how tightly or loosely coupled the databases are. A tightly-coupled architecture provides a stable interaction through the definition of a single federated schema controlled by the federation administrators. As with the global schema, whenever there are changes in the export schema of a tightly-coupled architecture, integration needs to be redone. A loosely-coupled architecture is a flexible approach that achieves interoperability by defining multiple views over databases. In this architecture it is the user who has the control of the federation. A shortcoming of the loosely-coupled architecture is the assumption that users know exactly what they are looking for and what each database contains. A multidatabase language represents a more loosely coupled integration than the loosely-coupled FDBS approach, because it does not use a partial or global schema (Litwin 1994). Similarly to the loosely-couple FDBS, however, a multidatabase language lacks the transparency for locating information, because users have to know *a priori* where the data are stored.

Syntax, schema, and semantics are a global definition of different levels of interoperability (Bishr 1997). If any of these levels cannot be solved, interoperability

remains unsolved as well. At the lowest level, syntactic definitions involve classic data structures (e.g., field and object based approaches). Schematic definitions refer to class hierarchies and elements that are used to represent real world entities (e.g., classes, attributes, and relations). Finally, semantic definitions concern the relationship between instances of a class and the real world objects (Meersman 1995).

Since the first studies on interoperability, progress has been made concerning syntactic interoperability (i.e., data types and formats) and structural interoperability (i.e., schematic integration, query languages, and interfaces) (Sheth 1999). As current information systems increasingly confront information and knowledge issues, semantic interoperability becomes the challenge for a new generation of interoperable systems (Egenhofer 1999). The problem of semantic interoperability is the identification of semantically similar objects belonging to different databases and the resolution of their schematic differences (Kashyap and Sheth 1996). Schematic heterogeneity can only exist, and therefore be solved, for semantically similar objects or schema elements (Bishr 1997, Bouguettaya *et al.* 1998). Thus, semantic similarity is introduced as a tool to determine what data can be integrated.

Some methods to solve semantic integration use the semantics underlying the data representation to determine semantic equivalence. For example, attribute equivalence is defined by comparing domain, constraints, and operations (Larson *et al.* 1989). Context and domain definitions are also combined in order to evaluate semantic equivalence (Ouksel and Naiman 1994, Sciore *et al.* 1994, Sheth and Kashyap 1992). Finally, some researchers have suggested comparing data semantics in terms of the behavior that characterizes the data stored in a database (Kuhn 1994). All these semantic similarity methods—attribute-based, context-based, and behavior-based—rely on the way data are modeled in a database.

From a different perspective, some researchers have investigated semantic similarity in databases based on term definitions and their interrelations (Bishr 1997, Bright *et al.* 1994, Collet *et al.* 1991, Fankhauser and Neuhold 1993, Weinstein and Birmingham 1999). The general approach has been to map the local terms in a database onto a shared ontology. An ontology captures the view of the world, supports intensional queries regarding the content of a database, defines semantics independently of data representation, and reflects the relevance of data without accessing them (Goñi *et al.* 1997). Once a common ontology is defined, the interrelationships among terms in the ontology are translated into their semantic similarities. One effort to create this common ontology is to create a knowledge base in terms of a global and domain-independent ontology. An example of this approach is Cyc (Lenat and Guha 1990, Lenat *et al.* 1995), which consists of approximately 40,000 objects. Using Cyc an entity of an information resource is mapped onto concepts of the global ontology by a set of articulation axioms (Collet *et al.* 1991). Another way to deal with ontology-based semantic integration is to work with existing ontologies, which are linked to create an integrated ontology. OBSERVER is a system that enables interoperation across independent pre-existing ontologies based on terminological relationships (i.e., synonyms, hyponyms, and hypernyms) that connect terms in different ontologies (Kashyap and Sheth 1998, Mena *et al.* 1996).

This thesis focuses on the spatial domain and follows an ontological approach to semantic integration. It pursues the definition of a method that finds similar entity classes that could link entities in independent databases to achieve information integration. In this sense, a similarity measure is a tool for loosely-coupled architectures of database integration.

1.2 Motivation

The main motivation of this thesis is the need to enhance geographic information systems at two levels of operation: (1) information retrieval and (2) information integration. For information retrieval this thesis creates a mechanism that allows users to express a query in an intuitive way by using terms of their natural (English) language. These terms should be semantically associated with terms used in the stored data to retrieve the desired information. For information integration, the model of semantic similarity provides the formalization for the identification and computational assessment of semantically similar objects. Furthermore, the semantic similarity model can be used to compare different data models, since it provides indices of how similar the objects embedded in those data models are.

Previous work in the assessment of semantic similarity lacks the following characteristics, which constitute the ground for the investigations of this thesis:

- Context dependence. Although some models consider context in the semantic representation of entities (Kashyap and Sheth 1996), few of them have introduced the context influence on the way the similarity assessment is performed. In this sense, context affects what aspects are more relevant than others in a similarity judgement. These aspects may be the concepts' descriptors (e.g., functions and parts) or cognitive properties (e.g., commonalities vs. differences).

- Asymmetric evaluation for cases of subclass-class and part-whole relations. Most semantic similarity models define symmetric similarity functions. Psychologists, however, have argued that similarity often needs asymmetric measures. Some cases in the spatial domain, such as building vs. museum and building vs. building complex, are examples for the need of an asymmetric evaluation.

- Adequate semantic representation for spatial concepts. Most models based on features or descriptors have an ambiguous explanation of what these features are. These models are usually applied to a broad domain and do not address the particular properties of concepts in the spatial domain. Likewise, models based on semantic relations usually include two types of relation: synonymy (equivalence) and hyponymy (is-a). In the spatial domain the meronymic relation (i.e., part-whole) represents another important semantic relation that needs to be considered in order to provide a more satisfactory representation of the interrelations among spatial concepts.

- Evaluation across multiple and autonomous definitions. Most current models for similarity assessment are based on the use of a shared ontology that semantically interconnects concepts. This approach has limitations in dynamic environments, such as the Internet, where scalability and variability are frequent properties of ontologies.

1.3 Goal and Hypothesis

The goal of this thesis is to create a formal model for the assessment of semantic similarity among spatial entity classes. This model should reflect properties of people's similarity judgments and a solid computational formalism. Major questions that drive the development of this thesis are:

- What are the desirable properties of a similarity model among spatial entity classes?

- What are the main components that semantically distinguish spatial-entity classes?

- What are the advantages and disadvantages of current models for semantic similarity? Can advantages of current models be integrated into a new similarity model?

- How does context affect similarity assessment?

The answers to these questions yield the definition of the Matching-Distance model that combines distinguishing features with semantic distance (Chapters 3 and 4). This model produces asymmetric evaluations and considers contextual information for the determination of the relevant features in the similarity assessment. The hypothesis of this work is therefore that

the Matching-Distance model matches people's judgments of similarity.

This hypothesis is supported by the statistical analysis of a human-subject experiment (Chapter 5).

1.4 Research Approach

This thesis develops a mathematical model to evaluate semantic similarity of spatial entity classes. The model is strongly influenced by studies in cognitive psychology and natural-language processing. This influence is due to the belief that a similarity model that employs elements of people's mental models could produce results that are well accepted and commonly desired. If a system simulates the way people reason and communicate about spatial concepts, the system is most likely to give users their desired answers (Mark 1989). This thesis shares the assumption by Talmy (1983) and Herskovits (1997) that the language we speak reflects our conceptual system; that is, we can treat concepts as linguistic terms and represent their semantics. This work,

however, focuses on spatial entities expressed as nouns, rather than spatial relations expressed as prepositions in natural language.

This thesis considers similarity assessment as a process in which common and different distinguishing features among entity classes are analyzed (Tversky 1977). In addition to distinguishing features, entity classes are defined by their semantic interrelations. We call this set of entity class definitions an ontology. In artificial intelligence, the term ontology has been used in many different ways. Ontology has been defined as a "specification of a conceptualization" (Gruber 1995a) and as a "logical theory which gives an explicit, partial account of a conceptualization" (Guarino and Giaretta 1995). Thus, an ontology is a kind of knowledge base that has an underlying conceptualization. For the purpose of this work, an ontology will be used as a body of knowledge that defines (1) primitive symbols used in the representation of meaning, and (2) a rich system of semantic relations interconnecting those symbols. Unlike the philosophical notion of ontology (Milligan 1992, Smith and Mulligan 1983), this definition relaxes the idea that an ontology describes a unique and task-independent reality. Instead, it allows us to have different ontologies, each of the ontologies having its own perspective for partially describing the same entity classes.

A natural idea for organizing concepts is to use a hierarchical structure derived from the hyponymic (is-a) relation among entity classes. Although linguists and computer scientists have commonly used lexical hierarchies for organizing nominal meanings, cognitive scientists have questioned the inheritance assumption implicit in those hierarchies (Miller 1998). Furthermore, the idea of typicality or prototyping has been suggested to better represent a concept (Lakoff 1987, Rosch 1973). Under the typicality theory, a concept is represented by its focal instances, which are the best examples of the concept.

Despite all arguments against hierarchies, this thesis follows this approach since practical work has shown the usefulness and importance of lexical hierarchies for nominal concepts (Miller 1998) and hierarchical structures for cognitive maps (Hirtle and Jonides 1985). These hierarchies, however, should not only include associations based on shared features, but also associations among concepts regarding the context in which they are used. The idea of using prototyping as part of the conceptual representation is valuable and is also considered. For this thesis, prototyping is assumed to be part of the definition of the typical distinguishing features of a concept. Furthermore, the effect of prototyping over the similarity assessment among concepts has also influenced our model, such that the similarity assessment between a variant and its prototype, or vice versa, results in an asymmetric evaluation (Rosch and Mervis 1975).

The foundation for many semantic distance approaches to similarity assessment (Rada *et al.* 1989, Rips *et al.* 1973) is that distance in a lexical hierarchy can be translated into the response time that associates two concepts (Collins and Quillian 1969). Objections to this assumption soon appeared, though (Smith and Medin 1981). Those studies argued that the time of response of associated concepts is influenced by the typicality of the concepts. Although our model uses the semantic distance among concepts, it embeds this distance into a feature-matching process (Tversky 1977). A feature-based approach to semantic similarity can distinguish entity classes even when they are all grouped under the same superclass, can produce asymmetric evaluations, and can use contextual information that affects the similarity assessment. The semantic distance is basically used to identify the relation between a variant (subclass) and its prototype or more general concept (class). Based on the semantic distance, a feature-matching process can be adjusted by using weights between non-common features that reflect the asymmetric evaluation of a similarity assessment.

The notion of context is also an important issue for the evaluation of semantic similarity (Shoham 1991). The definition of context in this thesis pursues the determination of the relevance of features for the similarity assessment. Context in this thesis is specified by the user's intended operations, because the meaning of a term is strongly affected by how the term is used (Miller and Charles 1991). This work describes an application by the set of tasks and the entity classes in the tasks' domain that characterize this application. The determination of feature relevance can then be obtained by two different approaches: (1) commonality and (2) variability, of distinguishing features in the domain of the application. In addition to the features' relevance, contextual information can partially resolve word-sense ambiguity, since entity classes of the application domain may limit the possible senses of polysemous terms.

Using a common and single ontology constrains the use of the model to individual databases or to multiple, homogenous databases. In multiple and heterogeneous databases different classifications or entities are defined, which leads to diverse conceptual models. Even if databases have the same conceptual models, the issue of *scalability* of the ontology is critical as new information resources enter to form part of the federation of databases (Kashyap and Sheth 1998). This thesis extends the basic model for similarity evaluations within an ontology to create a model that finds the most similar entity classes across ontologies. This model compares names, features, and semantic neighborhoods of entity classes using a matching process. Through the matching process, the model avoids disconnected hierarchical structures and proposes a set of similar entity classes that create anchors for ontology integration.

1.5 Scope of the Thesis

This thesis is concerned with the definition of spatial entities and, therefore, limits its domain of discourse to the set of entities that are part of standard spatial catalogs, such as the Spatial Data Transfer Standard (SDTS) (USGS 1998). These spatial entities are concepts expressed in the English language. Through the concepts of synonymy and polysemy, this thesis permits the distinction of regional differences in the use of language. Polysemy arises when the same word has more than one meaning (different *senses*) and synonymy corresponds to the case when two different words have the same meaning (Miller *et al.* 1990). For example, while the term for a small stream in the South of United States is *creek*, in New England a small stream is called a *brook*. This thesis links the terms *creek* and *brook* by a synonymy relation.

For the purpose of this thesis, we distinguish between similarity of entity classes and similarity of entity instances. While entity classes refer to concepts in the real world, entity instances denote physical objects in the real world. Since this study focuses on entity classes, this thesis does not address the similarity assessment among attribute values of specific instances of a class. For example, when assessing the similarity between a *sports arena* and an *office building*, this study considers what type of structural components (e.g., ceiling, color, floor, external material, and type of architecture), functional descriptors (e.g., to play, train, and work), and attributes (e.g., owner, color, and age) belong to both concepts, while it disregards the similarity assessment among values associated with structural elements, functions, or attributes. For example, this thesis does not address the similarity of colors red and blue.

1.6 Major Results

This thesis develops two models for calculating semantic similarity among entity classes: (1) the Matching-Distance model (MD) for evaluations within a single ontology and (2) the Triple Matching-Distance model (MD3) for evaluations across multiple ontologies. The MD model gives similarity values among entity classes as a function of the combination of the matching process over distinguishing features and the semantic distance of entity classes in a hierarchical structure. The MD3 model extends the MD model such that not only the distinguishing features but also the names and the semantic neighborhoods among entity classes are compared. The models have been implemented in an object-oriented prototype written in C++. This prototype allows users to check the semantic similarity among entity classes based on either a single ontology (user-defined or pre-defined) or across existing ontologies, such as WordNet (Miller *et al.* 1990) and SDTS (USGS 1998).

A human subject experiment supported the hypothesis that the MD model matches people's judgments. This result suggests that at different levels of generalization expressed in terms of is-a relations, semantic-similarity evaluations among entity classes produce asymmetric values. This claim resembles Rosh's (1973) hypothesis that in the similarity assessment a prototype (superclass) is less similar to its variants (classes) than its variants are to the prototype. For entity classes related by part-whole relations, however, asymmetric evaluations vary depending on the number of common distinguishing features among classes. For entity classes that share many of their distinguishing features, such as building and building complex, the similarity assessment tends to give similar asymmetric results as the results found when the classes are related by is-a relations. Since generalization (is-a relation) and aggregation (part-whole relation) are common abstraction mechanisms for handling spatial entities

(Egenhofer and Frank 1992), the MD model is well-suited for detecting semantically similar spatial entities.

The domain of entity classes that are involved in an application (i.e., the context domain) affects the results of similarity evaluations. While this effect is small but significant, the major determinant for a good similarity evaluation is the correct definition of entity classes in terms of distinguishing features. Commonality or variability may be the right approach to the determination of feature relevance depending on the specificity of the application.

The performance of similarity evaluations across ontologies depends on the level of formalization and explicitness of the ontologies. Although the MD3 model detects similar entity classes correctly, it is not clear if the model can detect all entity classes that are indeed similar. While distinguishing features are a basis for detecting similarity within a single ontology, lexicon and semantic neighborhood appeared to be better parameters for cross-ontology evaluations.

1.7 Intended Audience

The intended audience of this thesis is any person interested in information retrieval in general and in similarity assessment for spatial objects in particular. This may include a multidisciplinary group of computer scientists and geographers. This thesis is of particular interest to designers of spatial database systems and spatial query languages, as well as researchers from the fields of geographic information science, artificial intelligence, interoperating information systems, natural language understanding, and cognitive science.

1.8 Organization of Remaining Chapters

The remainder of the thesis is organized into six chapters.

Chapter 2 reviews previous work on semantic similarity assessment. This review includes the topics of ontology, cognitive properties of similarity, and models to assess semantic similarity. Models for semantic similarity assessment are analyzed in terms of the type of information they require and their main properties as compared with the cognitive property of similarity assessment.

Chapter 3 introduces the MD model for semantic similarity assessment among spatial entity classes. It explains the considerations and the components of the entity class representation. Subsequently, Chapter 3 presents the mathematical model for similarity assessment and its theoretical basis. Finally, an example illustrates the use of the MD model.

Chapter 4 complements the MD model with the contextual information in the similarity assessment. It describes the approach to modeling context and the use of context in similarity assessments. This chapter discusses the effect of context specification with examples of similarity evaluations over the same set of entity classes but under different contexts.

Chapter 5 presents the evaluation of the MD model by a human-subject experiment. This chapter describes the design of a survey given to 72 students, the results of both the subjects' responses and the MD model for the same questions, and the statistical analysis that compares these results.

Chapter 6 extends the MD model to account for semantic similarity assessment across multiple ontologies and defines the MD3 model. This chapter explains

additional components of entity class representations, the mathematical model for cross-ontology evaluations, and a test of the MD3 model with analyses over different combinations of ontologies.

Chapter 7 presents conclusions and further research directions. It discusses the main contributions and limitations of the MD and MD3 models, and addresses needs for future research.

Chapter 2

Modeling Semantic Similarity

Semantic similarity involves an assessment based on what is known about concepts. In information systems, this knowledge is expressed in the ontology that describes the conceptualization of the world the system is trying to represent. This chapter starts by reviewing the concept and use of ontologies. It focuses on the use of ontologies for information retrieval and information integration. Subsequently, this chapter presents properties of similarity assessments described by theories of knowledge and behavior. These properties constitute desirable characteristics of similarity models and are used as parameters for a comparison of current models. The discussion of current models for similarity assessment includes only those models that consider concept definitions and interrelations. Hence, this review excludes semantic similarity definitions among data modeled in databases that have been carried out by computer scientists in the area of heterogeneous and autonomous information systems (Elmagarmid *et al.* 1999).

2.1 Ontology

In a philosophical sense Ontology is the discipline that concerns the definition of a particular system of categories accounting for a certain vision of the world (Milligan 1992). Under this definition, an ontology is independent of a language used to describe it. The artificial intelligence community, in contrast, defines an ontology in regard to a specific vocabulary that describe a certain reality. Gruber (1995b) defines an ontology

as an explicit specification of a *conceptualization*. Distinguishing a *conceptualization* from an *ontology*, Guarino and Giaretta (1995) modified Gruber's definition and described an ontology as "a logical theory designed to account for the intended meaning of a vocabulary; i.e., its *ontological commitment* to a particular *conceptualization* of the world." They suggested that a conceptualization is "an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality." Figure 2.1 clarifies the relationship among conceptualization, language, and ontology (Guarino 1998). A relationship between the philosophical and engineering senses of an ontology exists if a conceptualization is associated with the philosophical sense of an ontology.

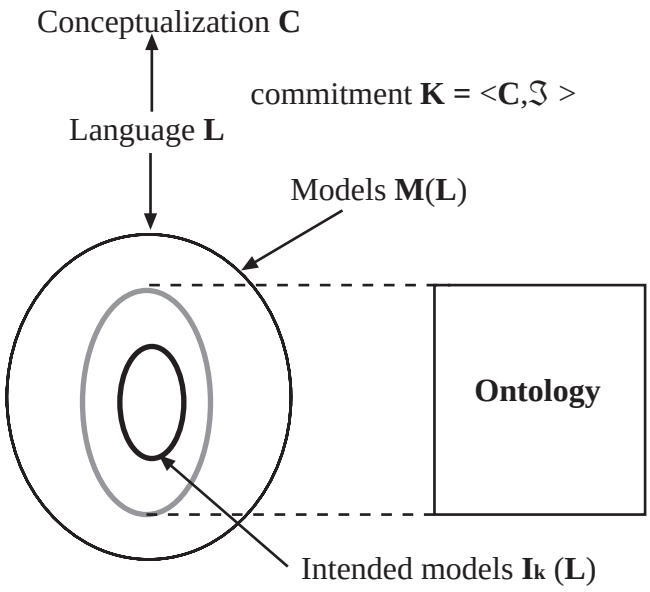

Figure 2.1: Relationship among conceptualization, language, and ontology.

Although an ontology is seen as a kind of knowledge base, an ontology contains state-independent information. It describes facts that are assumed to be always true by a community of users. A knowledge base, in contrast, may also include facts and assertions related to a particular state of affair. In a simple case, an ontology describes a hierarchy of concepts created by a generalization process. A more complex

ontology introduces axioms that relate concepts and constrain their interpretations. A classification of ontologies in terms of their level of explicitness and formalization is the following (modified from the classification done by Gangeni *et al.* (1998)):

- Catalog of normalized terms. A list of normalized terms without inclusion, axioms, and glosses.

- Glossed catalog. A catalog with natural glosses (e.g., dictionary of medicine).

- Taxonomy. A collection of concepts organized by a partial order induced by inclusion, such as WordNet (Miller *et al.* 1990) and SENSUS ontology for machine translation (Knight and Luk 1994).

- Characterized taxonomy. A collection of concepts, relations, and properties that characterize concepts, such as Mikroskomos (Mahesh 1996) and the ontology for the $(KA)^2$ community (Bejamins and Fensel 1998).

- Axiomatized taxonomy. A collection of concepts, semantic relations, properties, and axioms, such as the GALEN core model (Rector *et al.* 1993) and the PSL ontology (Schlenoff *et al.* 1998).

- Context (or ontology) library. A set of axiomatized taxonomies with relations among them, such as Cyc (Lenat *et al.* 1990).

A generic form of an ontology specification is given by a 5-tuple $O = <CD, RD, FD, ID, AD>$, in which CD is a set of class definitions, RD is a set of relation definitions, FD is a set of function definitions, ID is a set of instances definitions, and AD is a set of axioms definitions (Gruber 1992, Visser *et al.* 1998). A definition consists of a *definiendum* (i.e., a term that refers to the concept being defined) and a set of *definiens* (i.e., terms used to define the definiendum).

The interest in ontologies in the computer science community is reflected by the increasing use of ontologies in such diverse areas as knowledge representation (Guarino 1995), knowledge engineering (Gruber 1995b, Uschold *et al.* 1998), language engineering (Lang 1991, Mahesh and Nirenburg 1995, Milligan 1992), information retrieval and extraction (Guarino 1997, Guarino *et al.* 1999, Welty 1998), and information integration (Bergamaschi *et al.* 1998, Mena *et al.* 1998, Wiederhold 1994). The following sections focus on the uses of ontology for information retrieval and information integration that apply to this thesis.

2.1.1 Ontology-Based Information Retrieval

An ontology-based information retrieval, also called knowledge retrieval, uses primitives of an ontology to specify queries and resource descriptions. These primitives are semantically rich so that a better semantic matching between query and data stored can be accomplished. In current ontology-based information systems, semantic matching has meant the agreement on the vocabulary used by different agents. Thus, it implies sharing the same conceptualization, or agreeing to adopt a common conceptualization, which is the intersection of the original conceptualizations (Guarino 1997).

An initiative for ontology-based information (knowledge) retrieval in the World-Wide Web is $(KA)^2$ (Bejamins and Fensel 1998). Using a shared ontology, a web-crawler accesses the web pages and uses the ontology to infer answers. Depending on the level of specification of the ontology, the web-crawler may infer new information that is not explicitly stored on the Web. FindUr (McGuinness 1998) is another initiative in ontology-based information retrieval in the Web. FindUr uses an ontology to perform retrieval by abstracting classes, organizing content, and maintaining a knowledge base that captures the domain knowledge that is needed for

all services on the site. The experience with FindUr shows that an ontology-based information retrieval improves recall (i.e., the proportion of relevant material actually retrieved in the answer to a search request) and precision (i.e., the proportion of retrieved material that is actually relevant). These improvements are observed when the documents' lengths are short, there are few content words per document that are related, documents use an unfamiliar vocabulary, there is variability in the specificity of documents, meta-tagging is inconsistent or irregular, or general documents have higher (relevant) values over specific documents.

Concentrating on online yellow pages on the Web, Guarino *et al.* (1999) discussed the advantages of using a linguistic ontology such as WordNet and a structured representation formalism for information retrieval. They conclude that:

- users can express queries by using the most common English words rather than the data vocabulary;
- recall increases by exploiting the hierarchy to make generic queries and recognizing synonyms; and
- precision increases by a disambiguation mechanism and the ability to navigate the hierarchy to select specific queries. There is a further increment in precision if the system considers the structure of queries and descriptions.

In general, recent research indicates that ontology-based systems are suitable for obtaining effective information (knowledge) retrieval. These studies assume that users subscribe to a common ontology. Moreover, these studies emphasize the shareable nature of ontologies (Gruber 1995a), which may not be the case for all existing ontologies.

2.1.2 Ontology-Based Information Integration

A major application of ontologies is the area of information integration. Ontologies capture the semantics of data sources and are the basis for the link among diverse sources (Wiederhold 1994, Wiederhold and Jannink in press). As current information systems increasingly confront information and knowledge issues, semantic integration becomes the challenge for a new generation of interoperable systems. The problem of semantic integration is the identification of semantically similar objects that belong to different systems and the resolution of their schematic differences (Kashyap and Sheth 1996).

The general approach to semantic integration has been to map the local terms in a database onto a shared ontology. Most of these approaches use the terms' interrelationships to determine semantic similarity (Bishr 1997, Bright *et al.* 1994, Collet *et al.* 1991, Fankhauser and Neuhold 1993). Other approaches are measures based on graph matches and probabilistic measures that predict the probability that an instance of a concept in differentiated ontology will satisfy a request (Weinstein and Birmingham 1999). Efforts that create the shared ontology define a knowledge base in terms of a global and domain-independent ontology, such as Cyc (Lenat and Guha 1990, Lenat *et al.* 1995), LILOG (Lang 1991), and WordNet (Miller 1990). Although a shared ontology ensures complete integration, this type of ontology is costly if not impractical, because information systems are forced to commit to the shared ontology and compromises are difficult to maintain when new concepts are considered.

In environments with multiple and independent information systems, each system may have its own conceptualization and, therefore, its own intended model. Different intended models result in multiple ontologies that describe specific domains, such as an engineering ontology (Borst *et al.* 1997) and a medical ontology

(Zweigenbaum *et al.* 1995). These existing ontologies are well defined and their integration may reduce the cost of building a global ontology from scratch (Bergamaschi *et al.* 1998, Kashyap and Sheth 1998, Mena *et al.* 1996). The ontology integration, however, is a complex task, because concepts can overlap or definitions of concepts may be inconsistent across ontologies (Visser *et al.* 1998). Since there may be several ways to integrate ontologies, the definition of a systematic and consistent methodology for this integration becomes a real challenge.

ONIONS (Gangemi *et al.* 1998) 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. ONIONS's methodology includes the following steps: extraction of relevant set of terms from terminological sources, local definitions of terms, multi-local definitions of terms by triggering theories related to distinctions made in local definitions, and multi-local definitions of terms by triggering theories for the design of top-level categories.

A systematic approach to integrating ontologies is the use of the degree of overlap among ontologies (Wiederhold 1994). This approach considers intersection points and mutual exclusion points between various ontologies based on matching rules (Figure 2.2).

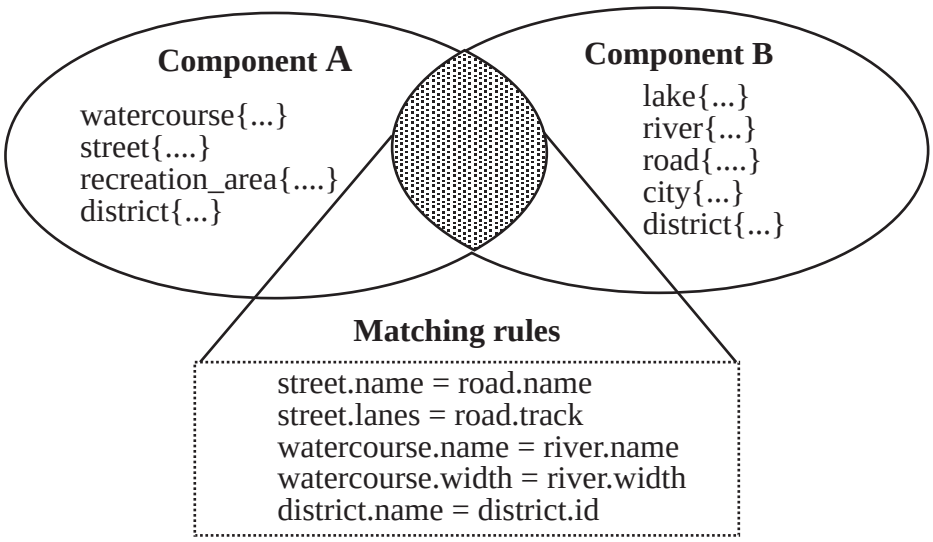

Figure 2.2: Example of ontology integration based on matching rules.

Subconcept-superconcept relationship (Hammer *et al.* 1994) is an approach to ontology integration that defines a concept as a collection of types determined to be similar by a common advisor. The similarity between types is based on heuristics with user inputs as required. The heuristics assess the distinguishing capability of a property of a concept that depends on the inter-concept dissimilarity among concepts and the intra-concept similarity within a concept. Then, a concept hierarchy is generated based on a subconcept-superconcept hierarchy (Figure 2.3).

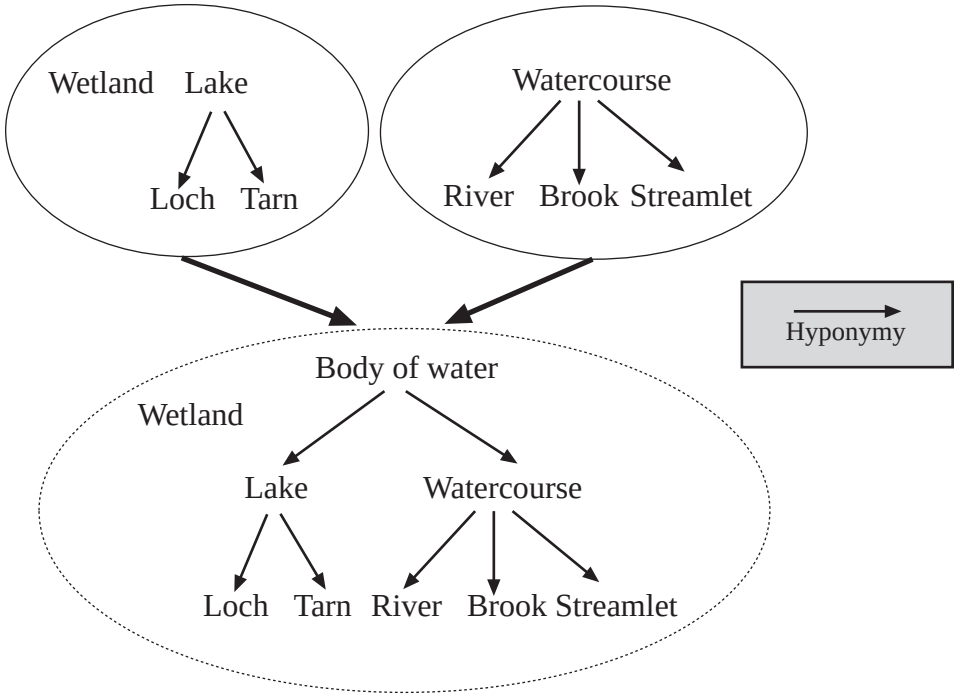

Figure 2.3: Example of ontology integration based on superconcept-subconcept relationship.

The use of semantic interrelations is yet another approach for ontology integration. OBSERVER is an ontology-based system that is enhanced with relationships for vocabulary heterogeneity resolution (Kashyap and Sheth 1998, Mena *et al.* 1996). 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. The loss of information is evaluated for both cases, and the translation with the least loss of information is chosen.

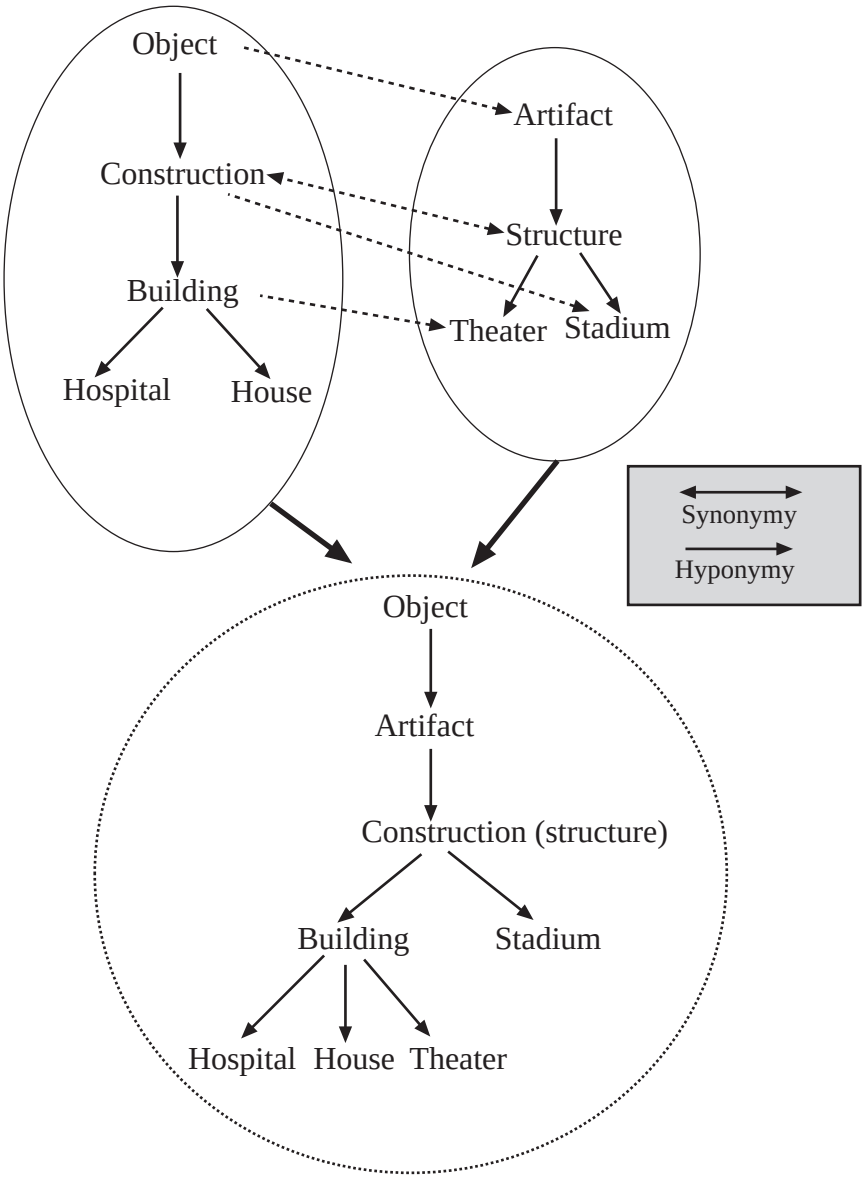

Figure 2.4: OBSERVER's terminological relationships for the integration of two ontologies.

As OBSERVER, Bergamashi *et al.* (1998) used synonymy and hyponymy terminological relations for ontological integration, but they also included a relation of positive association that connects terms generally used in the same context. Their approach is semiautomatic and starts with the extraction of hyponyms and associated terms from the source schema. Synonyms and domain-related knowledge are introduced by a person responsible for the integration. A validation of terminological

relations defined for attributes in the ontology is then followed by the inference of new relations.

2.2 Properties of Similarity Assessment

The study of similarity judgment has been an important area of investigation for psychologists and cognitive scientists. They have pursued the questions of how people classify objects, form concepts, solve problems, and make generalizations. A result of these studies has been the constant debate about the properties of similarity assessment. Many studies focus on the analysis of whether similarity satisfies properties of a metric distance function *d* (Equations 2.1a-c)

$$d(a,b) \geq d(a,a) \quad \text{(i.e., minimality)} \tag{2.1a}$$

$$d(a,b) = d(b,a) \quad \text{(i.e., symmetry)} \tag{2.1b}$$

$$d(a,b) + d(b,c) \geq d(a,c) \quad \text{(i.e., triangle inequality)} \tag{2.1c}$$

Although most studies assume that similarity satisfies minimality, Tversky (1977) argued that the same self-similarity for all objects implied by the minimality property does not hold for some similarity evaluations. Likewise, Krumhansl (1978) stated that the observed measure of similarity between an object and itself may be related to the status of the object within the domain. Thus, the self-similarity measure may not be the same for all objects, and the variation of the self-similarity may be related to the prototyping characteristics of the object within the domain. Krumhansl considered, however, that similarity satisfies the minimality property, because what really matters about minimality is that the self-similarity must be larger than the similarity between two different objects.

Similarity is not always a symmetric relation (Tversky 1977). 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). In the study of semantic categories, Rosch (1973) supported the view that categories are naturally formed and defined in terms of focal points or prototypes. She hypothesized that (1) in sentences such as "a is essentially b," the focal stimuli (i.e., prototypes) appear in the second position, and (2) the perceived distance from the prototype to the variant is greater than the perceived distance from the variant to the prototype. Asymmetry of similarity may result from searching for properties or features that characterize two objects (Krumhansl 1978). The transformation from one feature to another plays a role in similarity measures, because the need for less transformations between two objects results in a higher similarity judgment. Rada *et al.* (1989), however, argued that when similarity is limited to a feature comparison process, it is symmetric. They believe that the asymmetric problem of similarity found by Tversky (1977) is a result of the existence of another asymmetric relation. For example, a metaphor relating two concepts by a "like" relation involves a selective rather than an unconstrained comparison process. In other cases, people use a fuzzy category-membership (Zadeh 1965) rather than an evaluation of similarity.

The validity of the triangle inequality as a foundation for similarity models has been discussed (Tversky 1977). The triangle inequality implies that if *a* is quite similar to *b*, and *b* is quite similar to *c*, then *a* and *c* cannot be very dissimilar from each other (Equation 2.1c). Based on this property if a *sports field* is similar to a *gym* (because of their roles) and a *gym* is similar to a *building* (due to their structural definitions), then the *sports field* must be somehow similar to a *building*, a statement hard to accept. This example also reflects that similarity is not always transitive. Supporters of the triangle inequality property of similarity argue that the triangle inequality property fails due to

the different emphases on features and dimensions that are used to evaluate similarity (Krumhansl 1978, Rada *et al.* 1989). For instance, in the previous example *role* was used to evaluate semantic similarity between the *sports field* and the *gym*, whereas *structural characteristics* were used between the *gym* and the *building*.

Similarity vs. difference, context, and correspondence are also characteristics of similarity assessment discussed in the literature. In general, the often assumed inverse relation between similarity and difference is inaccurate. Naturally, an increase in the measure of the common features increases the similarity and decreases the difference, whereas an increase in the measure of distinction decreases similarity and increases difference. The relative values of these two semantic relations, however, may differ. While subjects may pay more attention to the similar features in the assessment of similarity among objects, they may pay more attention to their distinctive features in the assessment of difference (Krumhansl 1978, Tversky 1977).

Context and the frame of reference determine the relevant features for the evaluation of similarity. Sometimes the relevant frame of reference is explicitly specified (Tversky 1977). For example, how similar are an apple and a pear with respect to taste? Features or dimensions may be given different weights in different stimulus contexts (Krumhansl 1978). A suggestion is that weights are determined by how diagnostic the feature is for a particular set of objects under consideration (Goldstone *et al.* 1997, Tversky 1977). The diagnosticity of a feature refers to the classificatory significance of the feature or the degree of informativeness of a dimension. Tversky described an *extensive effect,* according to which features influence similarity judgment more when they vary within an entire set of stimuli. Likewise, Goldstone suggested that a dimension is highlighted when it presents a variability within a context.

The context effect of range and frequency is associated with the categorical judgments along a single dimension. The range-frequency theory states that a person (1) tends to divide his or her psychological range into a fixed number of subranges of equal size, and (2) employs the alternative categories with equal frequency (Krumhansl 1978). In terms of similarity, the first principle means that if the range of stimuli increases by adding more extreme stimuli, the similarity judgment of stimuli that are common to the original should increase. The second principle states that the similarity value between two objects in a relatively dense region of stimuli should be lower than the similarity value between two objects that differ in an equivalent amount, but occupy a less dense region.

When similarity assessment involves the comparison between scenes, correspondence should be consistent (Goldstone 1994). The similarity between two scenes cannot be determined before the parts of the scenes are placed in correspondence. In spatial scenes, correspondence refers to the spatial distribution of the parts. The degree of importance of a correspondence for the similarity assessment between two scenes depends on the consistency with respect to the emerging pattern of other correspondences between the scenes. Thus, the matching of corresponding features has a greater contribution to the similarity rate than the matching of features that do not correspond.

Another factor found to influence similarity judgment is classification. The diagnostic value of a feature is determined by the prevalence of the classification that is based on it. Thus, similarity has two faces, causal and derivative. It serves as a basis to classify objects, but it is also influenced by the adopted classification (Tversky 1977).

2.3 Models for Semantic Similarity Assessment

A general classification of models for semantic similarity assessment distinguishes models based on features, based on semantic relations, based on information content, and based on contextual information. Feature-based models have been proposed by cognitive psychologists who judge similarity in terms of distinguishing features of concepts or objects, such as properties, role, and rules. Models based on semantic relations, on the other hand, have primarily arisen from the computer science domain. These semantic relations are typically organized in a semantic network where nodes denote concepts and links represent semantic relations. Derived from the use of semantic networks, recent studies relate information content to semantic similarity determination. Finally, an approach to semantic similarity coming from the cognitive-linguistic domain presents a model for similarity assessment that considers the contextual representation of words within sentences.

2.3.1 Feature-Based Models

Using set theory, Tversky (1977) defined a similarity measure as a feature-matching process. It produces a similarity value that is not only the result of common features, but also the result of the differences between two objects. Taking two objects *a* and *b*, the matching process is defined by the two set-theory functions of intersection ($A \cap B$) —the set of features common to both *a* and *b*— and set difference ($A - B$) —the set of features that belong to *a* but not to *b*. Tversky's *contrast model* defines the similarity between two objects $S(a,b)$ (Equation 2.2).

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A), \text{ for } \theta, \alpha, \text{ and } \beta \geq 0 \qquad (2.2)$$

The terms θ, α, and β refer to the weights for common and different features between the two objects. These weights allow the definition of an asymmetric measure

for similarity. The asymmetry property is the result of the relative salience of the stimuli or classificatory significance of the feature. Under the assumption that all objects are of equal salience, similarity between objects is a linear function of the measure of their common features. Another matching function that normalizes the value of similarity is the *ratio model* (Equation 2.3).

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}, \text{ for } \alpha \text{ and } \beta \geq 0 \qquad (2.3)$$

A different strategy for feature-based models is to determine a semantic distance between concepts as their Euclidean distance in a semantic, multidimensional space (Rips *et al.* 1973). This approach describes a similarity measure by a monotonic function of the interpoint distance within a multidimensional space, where the axes in this space describe features of concepts. The distance between two points in the multidimensional space is typically computed by Equation 2.4, where *n* is the number of dimensions and $X_{i,j}$ is the value of object *i* in dimension *j*. The distance model in a semantic space satisfies the usual properties of a distance —minimality, symmetry, and triangle inequality (Equations 2.1a-c).

$$d(a,b) = \left[\sum_{k=1}^{n} \left| X_{a,k} - X_{b,k} \right|^2 \right]^{(1/2)} \qquad (2.4)$$

Krumhansl (1978) also suggested a distance function for similarity assessment that complements the interpoint distance with the spatial density of the space, called the distance-density model. This model assumes that within dense regions of a stimulus range finer discriminations are made than within relatively less dense subregions. The distance-density model defines a distance function $\bar{d}$ (Equation 2.5), where $d(a,b)$ is the normal distance, $\partial(a)$ is the density function, and α and β are relative weights of the density function.

$$\bar{d}(a,b) = d(a,b) + \alpha\partial(a) + \beta\partial(b), \text{ with } \alpha \text{ and } \beta \geq 0 \qquad (2.5)$$

Krumhansl (1978) argued that the distance-density model may be able to account for variations on self-similarity with the condition that the self-similarity is larger than the similarity between any two objects. The asymmetric property of similarity may be reflected in the distance-density model by considering that the density around one point affects the similarity more than the density around the other point in a directional evaluation of similarity.

In a more recent work Goldstone (1994) proposed a new model for similarity assessment of scenes that shares many characteristics with the cognitive process of analogical reasoning. He argues that neither feature-matching nor distance approaches of feature-based models account for the correspondence between scenes. This type of correspondence becomes relevant as propositionally and hierarchically structured scenes are compared. Propositional representations contain relational predicates such as the spatial relations above, below, left, and right. Hierarchical representations involve entities that are embedded into one another, such as *X* is part of *Y* or *X* is a kind of *Y*. Goldstone's model, called SIAM, evaluates similarity as an interactive activation and mapping between features, objects, and role correspondences. The overall similarity between two objects is determined by feature-to-feature matching between the objects, adjusted by the importance of the similarity in terms of the degree of alignment.

A shared disadvantage of feature-based models is that two entities are seen to be similar if they have common features; however, it may be argued that the extent to which a concept possesses or is associated with a feature may be a matter of degree (Krumhansl 1978). Consequently, a specific feature can be more important to the meaning of an entity than another. On the other hand, the consideration of common

features between entity classes seems to be cognitively sensible for the way people assess similarity.

2.3.2 Models Based on Semantic Relations

The semantic distance results in an intuitive and direct way of evaluating similarity in a hierarchical semantic network. This type of hierarchy is a common and efficient way to organize and connect concepts (Collins and Quillian 1969). For a semantic network with only is-a relations, the semantic relatedness and semantic distance are equivalent and one can use the latter as a measure of the former (Rada *et al.* 1989). In this context, conceptual distance is the length of the shortest path between two nodes in the semantic network (Figure 2.5).

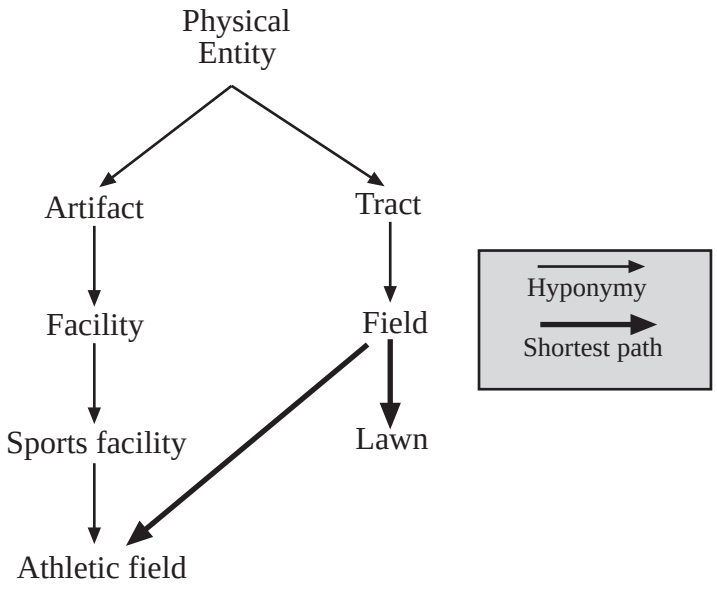

Figure 2.5: Shortest path between the concepts *athletic field* and *lawn.*

Although the semantic distance model has been supported by a number of experiments and has shown to be well suited for specific domains, it has the

disadvantage of being highly sensitive to the predefined hierarchical network. In a realistic scenario, adjacent nodes are not necessarily equidistant. Irregular density results in unexpected conceptual distance measures. The density effect suggests that the greater the density, the closer the distance between the nodes. With respect to the depth of a hierarchy, the distance shrinks as one descends the hierarchy, because the differences between nodes are based on finer details. By contrast, most concepts in the middle to high sections of the hierarchical network, being spatially close to each other, are deemed to be conceptually similar to each other.

In order to account for the underlying architecture of a hierarchical network, the semantic distance model should allow for weighted indexing schema and variable edge weights (Lee *et al.* 1993). To determine weights the structural characteristics of the network, such as the local density, the depth of a node in a hierarchy, the type of link (i.e., type of semantic relation), and the strength of an edge link (i.e., closeness between a child and its parent node), are typically considered.

Some studies have considered weighted distances in a semantic network (Jiang and Conrath 1997, Sussna 1993). Jiang and Conrath (1997) proposed to assign weights to the edges as a function of the link strength (*LS*), the depth of the node (*dp*), the local density (*LD*) of a node, the overall density (*WD*), and the type of link (Equation 2.6). The parameters α ($\alpha \geq 0$) and β ($0 \leq \beta \leq 1$) control the degree to which the node depth and density factors contribute to the edge weighting computation.

$$wt(c,p) = \left(\beta + (1-\beta)\frac{WD}{LD(p)}\right)\left(\frac{dp(p)+1}{dp(p)}\right)^{\alpha} LS(c,p)T(c,p) \qquad (2.6)$$

In Equation 2.7, the strength of the link (*LS*) from a child to its parent is proportional to the conditional probability of encountering an instance of the child concept c_i given an instance of its parent p (Equation 2.7)

$$LS(c_i, p) = -\log\left(P(c_i \mid p) = \frac{P(c_i \cap p)}{P(p)} = \frac{P(c_i)}{P(p)} \right) \tag{2.7}$$

Sussna (1993) defined a similarity measure in terms of the weighted distance in a semantic network that considers the local density, the depth in the hierarchy, and the type of relations. The weighted link between two nodes of a hierarchy is defined by Equation 2.8 and 2.9 where R is a relation, $\overline{R}$ is its inverse; dp is the depth of the deeper of the two nodes; max and min are the maximum and minimum weights possible for a specific relation R; and $n_R(x)$ is the number of relations R leaving from node x.

$$wt(c_1, c_2) = \frac{wt(c_1 R c_2) + wt(c_2 \overline{R} c_1)}{2dp} \tag{2.8}$$

Given

$$wt(xRy) = \max_R - \frac{\max_R - \min_R}{n_R(x)} \tag{2.9}$$

Semantic-distance based models have been widely used in information systems (Bishr 1997, Bright *et al.* 1994, Collet *et al.* 1991, Fankhauser and Neuhold 1993, Guarino *et al.* 1999); however, they present same important disadvantages with respect to cognitive properties of similarity assessment. Semantic-distance models satisfy all metric properties (i.e., minimality, symmetry, and triangle inequality), they are context independent, they are highly sensitive to the semantic structure, they consider only is-a relations among concepts, and they give coarse values of similarity for concepts that have a same superordinate.

2.3.3 Models Based on Information Content

Information-based models use a hierarchical network and information theory to define a measure for semantic similarity (Resnik 1999, Richardson and Smeaton 1996). The basic idea is that the more information two concepts share, the more similar they are.

Conceptual similarity is considered in terms of class similarity. The similarity between two classes is approximated by the information content of the first superclass in the hierarchy that subsumes both classes. The general idea of the information content is that, as the probability of occurrence of a concept in a corpus increases, informativeness decreases, such that the more abstract a concept, the lower its information content. For example, the information content of the abstract concept *entity* is less than the information content of more concrete concepts such as *road* and *house*. The information content of this superordinate is derived from the statistical analysis of word frequency occurrences in a corpus. In mathematical terms, information content is computed by Equation 2.10, where $P(c)$ is the probability of the occurrence of c in a corpus.

$$IC(c) = -\log \frac{1}{P(c)} \tag{2.10}$$

In the case of multiple inheritance (Cardelli 1984), similarity can be determined by the best similarity value among all possible senses to which the classes belong. Equation 2.11 defines the similarity function of the information-based model, where $Sup(c_1,c_2)$ is the set of concepts that subsume both c_1 and c_2, and IC is the information content of a concept or class.

$$S(c_1, c_2) = \max_{c \in Sup(c_1,c_2)} [IC(c)] \tag{2.11}$$

The information-content model requires less information on the detailed structure of the network. The determination of information content can adapt a static knowledge structure to multiple contexts (Resnik 1999). On the other hand, many polysemous words and multi-worded classes have an exaggerated information content value. The information-content model can generate a coarse result for the comparison of classes, because it does not differentiate the similarity values of any pair of classes

in a sub-hierarchy as long as their “smallest common denominator” is the same (Jiang and Conrath 1997).

2.3.4 Context-Based Models

Studying the relation between semantic similarity and contextual similarity, Miller and Charles (1991) discussed a contextual approach to semantic similarity. Contextual representation of a word comprises syntactic, semantic, pragmatic, and stylistic conditions that affect the use of that word. Although they found that the similarity among contextual representations is one of several factors for similarity assessment among words, their work revealed a clear relationship between semantic similarity and contextual similarity, when the words belong to the same syntactic category (i.e., nouns, verbs, adjectives, or adverbs). For such words, the similarity assessment is defined in terms of the degree of substitutability of words in sentences. The more often a word can be substituted by another word in the same context, the more similar the words are. The problem with this similarity measure is that it is difficult to define a systematic way to calculate it.

2.4 Summary

In information systems, ontologies capture the semantics of data sources and are a basis for information retrieval and integration. For this work we confine the definition of an ontology to be a kind of knowledge base that describes a certain reality in terms of a set of entity classes and their interrelations. Models for semantic similarity assessment have usually compared objects or concepts by considering the concepts’ descriptors (features) or the concepts’ interrelations (semantic relations). Important characteristics of similarity assessment are asymmetry, non-transitivity, and context dependence. Besides some feature-based models, models for similarity assessment

have been characterized by being symmetric and context independent. The next chapter introduces a new approach to create a computational model for semantic similarity assessment among entity classes that overcomes limitations of current models to account for symmetric and asymmetric evaluations and part-whole relations. The model is further expanded in Chapter 4 to include contextual information.

Chapter 3

A Computational Model for Semantic Similarity among Entity Classes

A computational model for semantic similarity provides a systematic way to determine quantitative values of semantic similarity. This mapping into the domain of numbers enables an ordering as well as limited inferences about degrees of similarity. This chapter presents a computational model for the determination of semantic similarity among spatial entity classes, called the Matching-Distance model (MD). It assumes a single ontology for the evaluation of similarity between two entity classes, i.e., the same conceptualization underlies the definition of both entity classes. The goal of the computational model is to provide a similarity measure that reflects cognitive properties of similarity judgments, in particular cases of asymmetric evaluations and contextual dependence. It is also expected that the computational model can make use of already available information about entity classes, such as the information found in lexical databases, taxonomies, thesauri, or catalogs. Thus, the model would be not only cognitively plausible, but also computationally achievable.

3.1 Components of the Entity Class Representation

For this work the purpose of the semantic representation of entity classes is to capture sufficient knowledge about entity classes in order to differentiate them. In this thesis no attempt is made to create a knowledge base that allows a person or a machine to completely capture the entity classes' semantics. Thus, this thesis distinguishes two

approaches to the semantic representation of entity classes: differential and constructive (Miller *et al.* 1990). The former approach has more modest requirements since it considers only the relevant knowledge that distinguishes two entity classes.

This work represents entity classes by defining two main components: (1) semantic relations among entity classes and (2) distinguishing features of entity classes (Table 3.1). It organizes entity classes based on their semantic interrelations and describes the set of entity classes and their semantic relations as an ontology.

Components	**Description**
Definiendum	Term or synonym terms that refer to an entity class
Definiens	What is used to define an entity class
Semantic Relations	Relations to other entity classes
Distinguishing Features	Properties of the entity classes

Table 3.1: Components of entity class representations.

Since entity classes are associated with concepts represented in natural language by words, this thesis takes into account two linguistic concepts—synonymy and polysemy—that characterize the mapping between words and meanings (Miller *et al.* 1990). The class-entity representation incorporates synonyms, such as *parking lot* and *parking area*, and different senses of entity classes, such as the case when a *bank* may be an *elevation of the seafloor*, a *sloping margin of a river*, a *financial institution*, or a *building that houses a financial institution*.

3.1.1 Semantic Relations

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). The MD model refers to entity classes by using synonym sets, which are interrelated by hyponymy and meronymy relations. It has been suggested that the two abstraction mechanisms of object-oriented theory (Dittrich 1986) that are associated with hyponomy and meronymy relations (i.e., generalization and aggregation, respectively) are fundamental for adequately modeling spatial configurations (Egenhofer and Frank 1992). The hyponymy relation, usually called is-a relation (Smith and Smith 1977), is the relation most commonly 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.

Mereology, the study of part-whole relations, also plays an important role in an ontology (Guarino 1995). Studies have usually assumed that part-whole relations are transitive such 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. Winston *et al.* (1987) defined six types of part-whole relations based on three main aspects: functional relation, homeomerous property (i.e., whether parts and whole are of the same type) and separable property (Table 3.2).

Relation	**Example**	**Relation Elements**		
		Functional	Homeomerous	Separable
Component - Object	pedal - bicycle	√	-	√
Member - Collection	tree - forest	-	-	√
Portion - Mass	slice - pie	-	√	√
Stuff - Object	steel - bike	-	-	-
Feature - Activity	paying - shopping	√	-	-
Place - Area	oasis - desert	-	√	-

Table 3.2: Types of meronymy relations defined by Winston *et al.* (1987).

In addition to defining types of meronymy relations, Winston *et al.* (1987) discussed similarity between meronymy relations and other semantic relations. They suggested a partial classification of semantic relations (Figure 3.1) and defined a transitive property among these semantic relations. Transitivity among the semantic relations holds if (1) the same type of semantic relation is used for the two premises of the syllogisms, or (2) the conclusion contains the relation that is lower in the hierarchy of inclusion relations. The hierarchy of inclusion relations establishes that spatial inclusion, meronymy inclusion, and class inclusion are the lower, medium, and higher relations, respectively. Iris *et al.* (1988), however, showed contradictions in Winston's transitivity hypothesis. For example, consider that a handle is part of a door (component-object) and the door is part of the house (component-object). By transitivity, the handle would be part of a house, which is debatable because it is spatially part of the house, but not functionally.

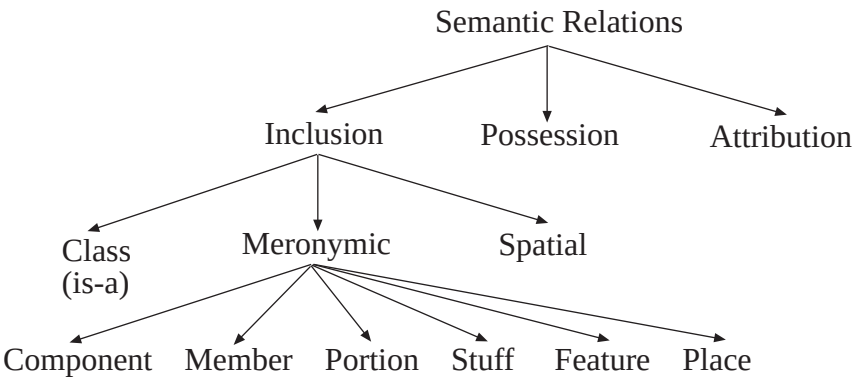

Figure 3.1: Partial classification of semantic relations (Winston *et al.* 1987).

Among all types of part-whole relations, this thesis considers the component-object and stuff-object relations with the properties of asymmetry and (with some reservations) transitivity. When describing the semantic relations among entity classes, the model distinguishes the two relations “part-of” and “whole-of” to be able 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, not all *buildings* are part of a *building complex*.

The MD model organizes the is-a and part-whole relations in an acyclic graph (Figure 3.2). It uses is-a and part-whole relations for hierarchically comparing entity classes such that a factor of asymmetry is determined.

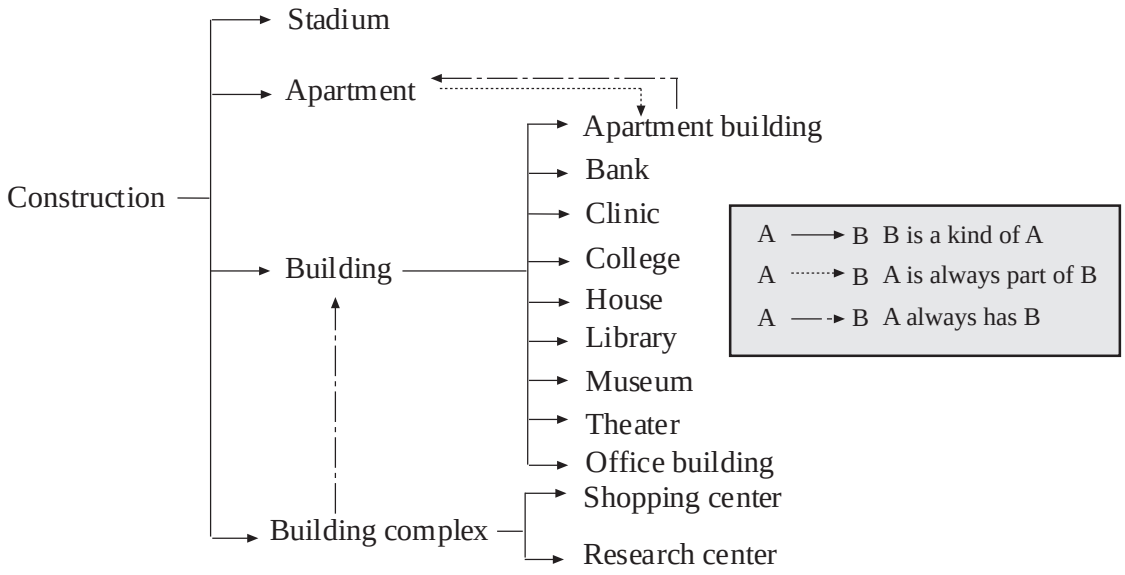

Figure 3.2: Fragment of a hierarchical network with is-a and part-whole relations based on WordNet.

3.1.2 Distinguishing Features

Although the general organization of entity classes is given by their is-a and part-whole interrelations, this information 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 entity classes, and their values describe the properties of individual objects (i.e., instances of an entity class). Attributes can be also seen as relations. By treating attributes separately from relations we distinguish between the organization of entity classes using semantic relations and

the description of entity classes in terms of distinguishing features. This thesis suggests a finer identification of distinguishing features and classifies them into functions, parts, and attributes. This classification attempts to facilitate the implementation of the entity class representation as well as to enable the separate manipulation of each type of distinguishing feature. Considering that entity classes correspond to nouns in linguistic terms, this work matches Miller's (1990) description of nouns. 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 entity classes, more than one term may denote the same feature (i.e., synonymy) or a term may denote more than one feature (i.e., polysemy).

The notion of use-based semantics (Kuhn 1994) leads this thesis to consider functions as one of the distinguishing features of an entity class representation. Function features are intended to represent what is done to or with a class. For example, the function of a *college* is to *educate*. Thus, function features can be related to other terms such as *affordances* (Gibson 1979) and *behavior* (Khoshafian and Abnous 1990). In the spatial domain, parts play an important role for the description of spatial entities. Parts are structural elements of a class, such as *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"). This thesis focuses on mandatory parts that are associated with part-whole relations. While the 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. Finally, attributes 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*.

The representation of entity classes does not contain the values of attributes because these values are associated with specific instances of the entity classes. For example, the representation of the concept *building* specifies an attribute *age*, but it does not store the value for *age*. Consequently, the evaluation of similarity is done at a higher level of abstraction than the similarity assessment among instances of entity classes.

3.2 The Matching-Distance Model

This thesis introduces a computational model that assesses similarity by combining a feature-matching process with a semantic-distance measurement. While this model uses the number of common and different features between two entity classes, it defines the relevance of the different features in terms of the distance among entities in a hierarchical structure. The global similarity function $S(c_1,c_2)$ is a weighted sum of the similarity values for parts, functions, and attributes (Equation 3.1), where ω_p, ω_f, and ω_a are 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 weights all together must add up to 1.

$$S(c_1,c_2) = \omega_p \cdot S_p(c_1,c_2) + \omega_f \cdot S_f(c_1,c_2) + \omega_a \cdot S_a(c_1,c_2) \tag{3.1}$$

For each type of distinguishing features we use a similarity function $S_t(c_1,c_2)$ (Equation 3.2) that 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 ($C_1 - C_2$), defined as the set of all elements that belong to C_1 but not to C_2.

$$S_t(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha(c_1, c_2) \cdot |C_1 - C_2| + (1 - \alpha(c_1, c_2)) \cdot |C_2 - C_1|} \quad (3.2)$$

The function α is determined 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 (l.u.b.) 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 (l.u.b.) between them. For instance, consider the hierarchical structure shown in Figure 3.2. The immediate superclass between *stadium* and *house* is *construction*. In like manner, the immediate superclass between *building* and *museum* is *building*. The distance of each entity class to the l.u.b. is normalized by the total distance between the two classes, such that we obtain values in the range between 0 and 1. Then, the final value of α is defined by a symmetric function (Equation 3.3).

$$\alpha(c_1, c_2) = \begin{cases} \dfrac{d(c_1, l.u.b.)}{d(c_1, c_2)} & d(c_1, l.u.b.) \leq d(c_2, l.u.b.) \\ 1 - \dfrac{d(c_1, l.u.b.)}{d(c_1, c_2)} & d(c_1, l.u.b.) > d(c_2, l.u.b.) \end{cases} \quad (3.3)$$

The determination of α is based on the idea that similarity is not necessarily a symmetric relation (Tversky 1977). For example, “a hospital is similar to a building” is a more generally accepted than “a building is similar to a hospital.” It has been suggested that the perceived distance from the prototype to the variant is greater than the perceived distance from the variant to the prototype, and that the prototype is commonly used as a second argument of the evaluation of similarity (Krumhansl 1978, Rosch and Mervis 1975). Hence, this work assumes that a prototype is generally a superclass for a variant and that the concept used as a reference (i.e., the second argument) should be more relevant in the evaluation.

The similarity function (Equation 3.2) yields values between 0 and 1. The extreme value 1 represents the case when all distinguishing features are 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 everything is different between two entity classes. An interesting case occurs when comparing a class with its superclass or vice versa. Since subclasses inherit features from their superclasses, only subclasses may have non-common features. It can easily be seen that when comparing a class with its superclass (e.g., a *clinic* with a *building*), 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. By considering the direction of the similarity evaluation, a class is more similar to its superclass than the same superclass is to the class.

For the purpose of calculating α, part-whole relations are treated like is-a relations, because they also represent a hierarchy among concepts. For this model, the main difference between is-a relations and part-whole relations depends upon the inheritance property of the former. While subclasses usually inherit all the behavior and properties of their superclasses, the same principle does not apply to composite and compound entities in part-whole relations (Egenhofer and Frank 1992). To determine a class that subsumes two classes under comparison, not only the is-a relation, but also the part-of and whole-of relations are checked. In Figure 3.2, the superclass between *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 for the same two classes, however, would yield the superclass *construction*. Unlike the comparison between class and superclass, evaluations between parts and wholes, or vice versa, follow unpredictable behavior, since parts do not necessarily share features with their wholes.

In the MD model, synonym sets denote entity classes and distinguishing features. A set of synonyms contains more semantic information than a single term. Since the model does not assess similarity of distinguishing features, we expect that a set of synonyms can identify a distinguishing feature with little ambiguity. Words with different semantics or senses (polysemy) are also included. Different senses of an entity class are handled as independent entity classes with a common name. For parts, functions, and attributes, the model first matches the senses of the terms, and then it evaluates the set-intersection or set-difference operation among the set of features. Furthermore, a term in one sense might have a set of synonyms such that the model matches terms or their synonyms that belong to the same sense. For example, the function *play* associated with a sports facility might have different senses in a database, *play* for recreation and *play* for competition. For any entity class that has the function *play* (e.g., *sports arena*, *stadium*, *park*, and *sports field*), the knowledge base also identifies the sense of the word so the model can find the synonyms of *play* for the respective sense.

Since the MD model is based on the comparison of distinguishing features, the lack of distinguishing features in an entity class's definition produces a similarity value with respect to any other entity class in the ontology equal to zero. 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 the MD model, the model's strength is the capability to assess the similarity among concepts located at or below Rosch's (Rosch 1975) basic level of a hierarchical structure, such as the concepts found in spatial catalogs (e.g., Spatial Data Transfer Standard (USGS 1998)). This characteristic of the MD model is 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.

3.3 Using the Matching-Distance Model

To experiment with the MD model, a prototype has been implemented in C++, and an ontology with 257 entity-class definitions has been derived from two readily available resources: the Spatial Data Transfer Standard (SDTS) (USGS 1998) and WordNet (Miller 1995). SDTS was adopted by the American National Standard Institute 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. Included terms in SDTS can be either synonyms or subclasses. For this work, however, we assume all included terms to be subclasses, which increases the ontology without affecting the similarity, because included terms hold the same definitions as their standard term. SDTS narrows the domain of the ontology in the MD model. Thus, SDTS gives the list of the entity classes to be defined, their partial definition of is-a relations, and their attributes.

WordNet is an on-line lexical reference system that was developed by the Cognitive Science Laboratory at Princeton University. WordNet organizes concepts in sets of synonyms (synsets) connected by semantic relations. It contains approximately 118,000 words organized into 90,000 sets of synonyms. These synonym sets are semantically interrelated depending on their syntactic category (Table 3.3). The application of WordNet in an information system is found in areas such as text retrieval (Richardson and Smeaton 1995, Voorhees 1998), word sense disambiguation (Basili *et al.* 1997, Leacock and Chodorow 1998), and conceptual modeling (Burg and Riet 1998). This work extracts synonym sets as well as hyponyny and meronymy relations from WordNet's definitions to complement definitions of entity types in SDTS.

Semantic Relation	Syntactic Category	Example
Synonymy	nouns, verbs, adjectives, adverbs	building - edifice
Antonymy	adjectives, adverbs (nouns, verbs)	bright - dark
Hyponymy	nouns	hospital - building
Meronymy	nouns	apartment - apartment building
Troponomy	verbs	march - walk
Entailment	verbs	buy - pay

Table 3.3: Semantic relations in WordNet (Miller 1995).

To complete the entity class definitions, functions are derived from verbs explicitly used in the natural-language descriptions of entity classes, augmented by common sense. A partial hierarchical structure of the ontology that was created is shown in Figure 3.3. The hierarchical structure includes is-a and strict part-whole relations. It presents a case of polysemy involving the term *bank* (i.e., *bank* as a building and *bank* as a financial institution) and cases where an entity class has more than one superclass, such as the case of a *parking area*, which is a *facility* and a *lot*. Figure 3.4 shows the complete description of the entity class *stadium*, i.e., its distinguishing features, semantic relations, and synonyms.

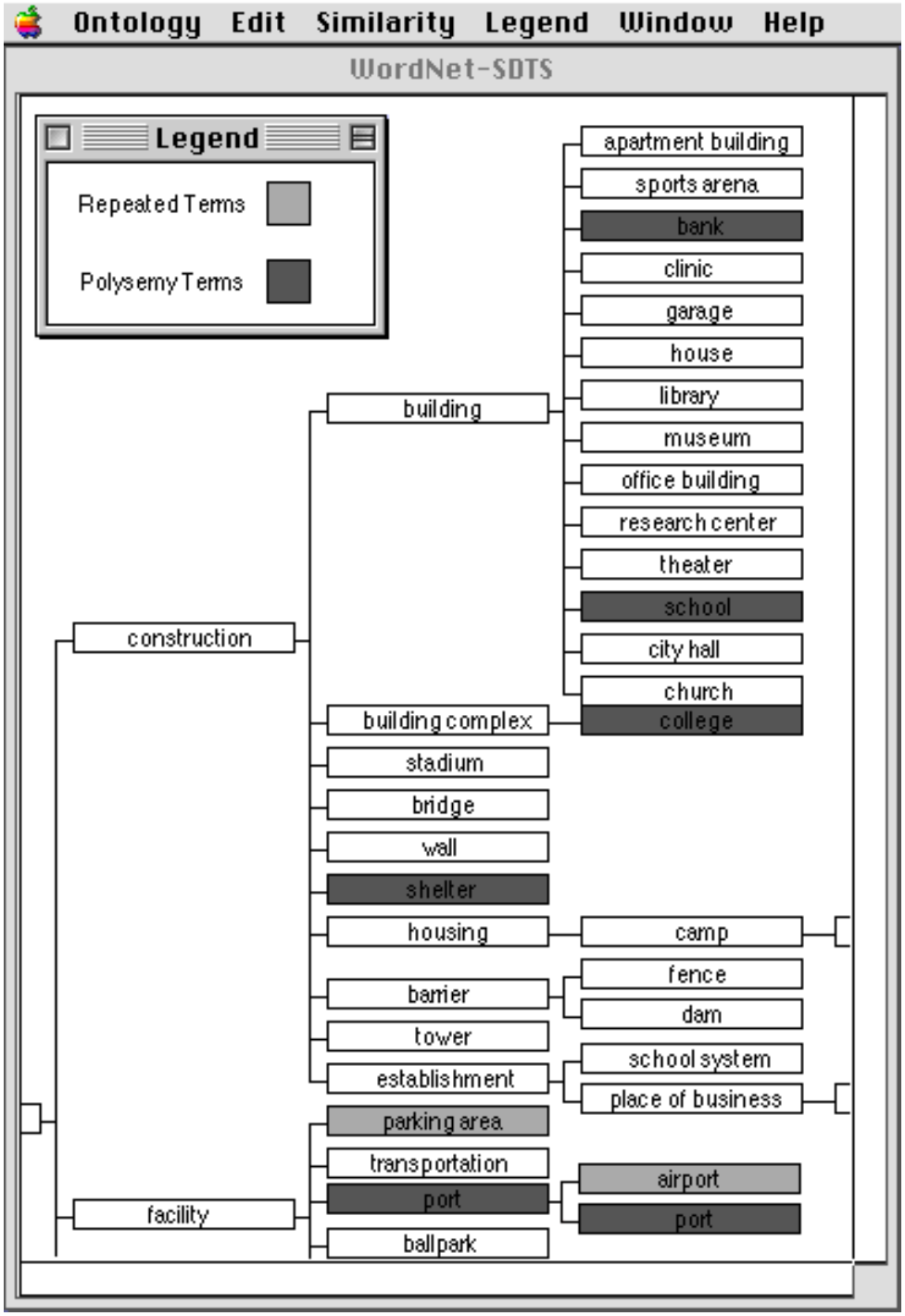

Figure 3.3: Portion of the ontology derived from the combination of SDTS and WordNet.

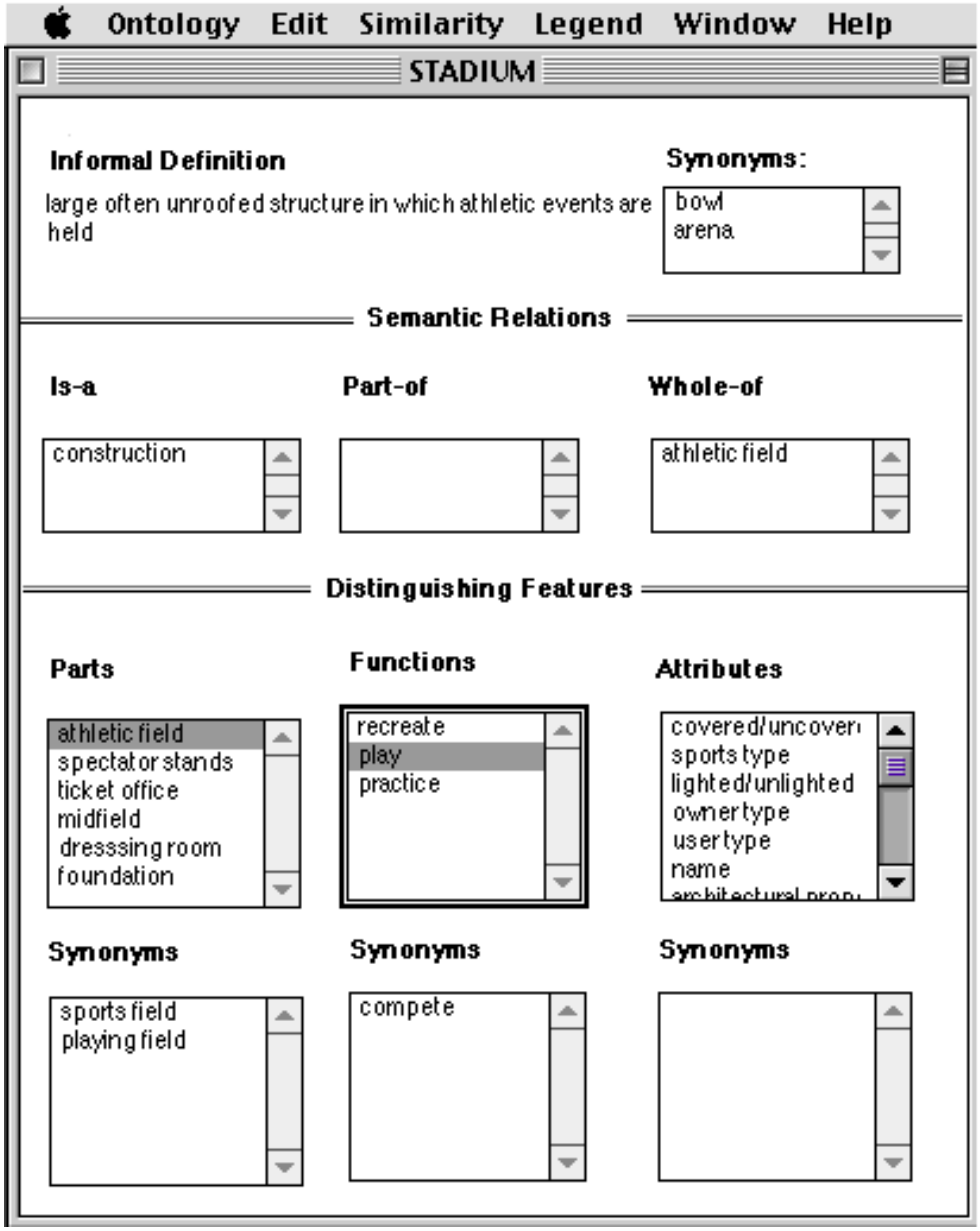

Figure 3.4: Definition of a *stadium*.

By default the MD model assigns the same weight to each type of distinguishing feature. Figure 3.5 shows an example of a similarity evaluation with default settings between a *stadium* and the rest of the entity classes in the ontology.

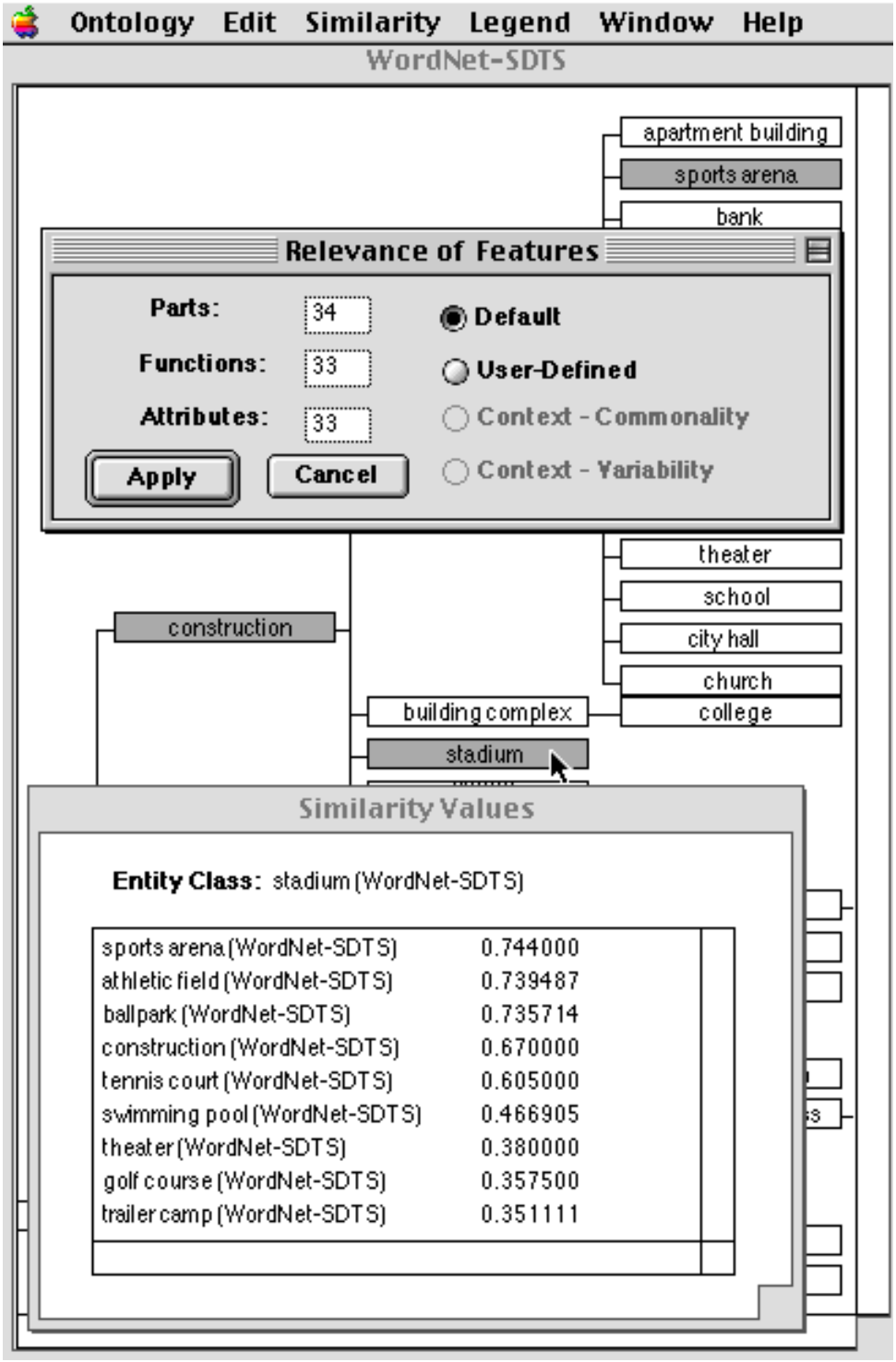

Figure 3.5: Results of the similarity between *stadium* and a portion of the WordNet-SDTS ontology.

Two characteristics of the MD model are (1) the asymmetric evaluation of entity classes located at different levels in the hierarchical structure and (2) the use of weights for the relative importance of distinguishing features. Table 3.4 shows some results that demonstrate the asymmetric evaluation of the MD model. For example, *sports arena* and *theater* are subclasses of *building* at the same level in the hierarchical structure, so the similarity evaluation between them is symmetric. The evaluations between *sports arena* and *building* or between *theater* and *building*, however, are asymmetric. For all evaluations that go from a class to a superclass (i.e., from *theater* to *building*) the similarity value is greater than the similarity value from the superclass to the class (i.e., from *building* to *theater*). In the case of part-whole relations, Table 3.4 shows that the similarity value from the whole to its part (e.g., from *stadium* to *athletic field*) is greater than the value from the part to the whole (e.g., from *athletic field* to *stadium*). Bear in mind, however, that this situation may not always occur, since there is not a general relationship between the distinguishing features of entity classes related by a part-whole relation.

	Athletic field	Ballpark	Building	Road	Sports arena	Stadium	Theater
Athletic field	1.00	0.83	0.17	0.12	0.49	0.70	0.16
Ballpark	0.84	1.00	0.16	0.10	0.49	0.74	0.14
Building	0.18	0.16	1.00	0.10	0.48	0.30	0.44
Road	0.17	0.18	0.10	1.00	0.10	0.14	0.10
Sports arena	0.52	0.50	0.67	0.10	1.00	0.78	0.58
Stadium	0.74	0.74	0.30	0.12	0.74	1.00	0.38
Theater	0.19	0.17	0.67	0.10	0.58	0.42	1.00

Table 3.4: Example of similarity values for a subset of the WordNet-SDTS ontology.

The sensitivity of the model to the distinguishing features' weights is shown by performing a set of evaluations that considers only one type of distinguishing feature (i.e., only parts, functions, or attributes) and the combination of these types (i.e., parts-attributes, parts-functions, functions-attributes, and parts-functions-attributes). Table 3.5 shows the evaluation with different sets of weights between a *stadium* and a portion of the ontology.

Weights			Athletic field	Ballpark	Building	Road	Sports arena	Theater
ω_p	ω_f	ω_a						
100%	0%	0%	0.33(4)	0.50(2)	0.22(5)	0.00(6)	0.60(1)	0.50(2)
0%	100%	0%	1.00(1)	1.00(1)	0.00(4)	0.00(4)	1.00(1)	0.00(4)
0%	0%	100%	0.90(1)	0.71(2)	0.67(3)	0.36(6)	0.64(4)	0.64(4)
50%	50%	0%	0.67(3)	0.75(2)	0.11(5)	0.00(6)	0.80(1)	0.25(4)
0%	50%	50%	0.62(1)	0.61(3)	0.44(5)	0.18(6)	0.62(1)	0.57(4)
50%	0%	50%	0.95(1)	0.86(2)	0.33(4)	0.18(6)	0.82(3)	0.32(5)
33%	33%	33%	0.74(1)	0.74(1)	0.30(5)	0.12(6)	0.74(1)	0.38(4)

Table 3.5: Similarity evaluations with different distinguishing features' weights between a *stadium* and a portion of the ontology. (Numbers in parentheses denote the rank in each horizontal combination.)

Table 3.5 indicates that important variations may occur, either in absolute values or ranks, as a result of different weights for distinguishing features. When an ontology has been designed for a specific application, distinguishing features in the entity class definitions are already selected as important for the application. Thus, we could have a good approximation of the similarity assessment by assuming that distinguishing features are equally important. When an application-independent ontology is used, in contrast, distinguishing features may be more or less important for some particular application.

3.4 Summary

The basic characteristic of the MD model is the combination of two different approaches to similarity assessment: (1) a feature-matching process and (2) a semantic-distance determination. This model for semantic similarity has a strong basis in linguistics. It introduces synonyms and different meanings (senses) in the use of terms. The model also provides a first approach to handle part-whole relations in the evaluation of semantic similarity. It defines a semantic-similarity function that is asymmetric for classes that belong to different levels of generalization in a hierarchical structure. This model organizes information about distinguishing features of an entity class into parts, functions, and attributes such that different relevance weights can be assigned to them. The next chapter discusses context as the determinaning factor of weight definitions and proposes two approaches —commonality and variability— to obtain weights for distinguishing features.

Chapter 4

Integrating Context into the Similarity Model

Context is an important aspect for such diverse areas as natural language processing (NLP), knowledge-based problem solving, database systems, and information retrieval. Despite this recognition, the meaning of context in information systems is usually left to the user's interpretation and its role may vary among different domains (Akman and Surav 1996). For NLP, context has a sense-disambiguation function (Leech 1981) so that otherwise ambiguous statements become meaningful and precise. Studies in NLP analyze the meaning of words within either a topical context or the local context of a corpus (Leacock and Chodorow 1998). Knowledge representation involves statements and axioms that hold in certain contexts; therefore, context determines the truth or falsity of a statement as well as its meaning (McCarthy 1987). For knowledge-based problem solving, context is usually defined as the situations or circumstances that surround a reasoning process (Aïmeur and Frasson 1995, Dojat and Pachet 1995, Turner 1998). Recent studies on data semantics and interoperability have stressed the importance of context to describe data content. In this domain, context is the knowledge needed to reason about another system (Ouksel and Naiman 1994), the intentional description of database objects (Kashyap and Sheth 1996), and the extent of validity of an ontology (Wiederhold and Jannink in press). For information retrieval, context provides a framework for well-defined queries and, therefore, improves the

matching process between a user's query and the data stored in a database (Hearst 1994).

Following the idea of Naive Physics (Hayes 1990) and Naive Geography (Egenhofer and Mark 1995), it is possible to derive common sense definitions of entity classes such that entity classes are described by their essential properties. Using these common sense definitions, we could expect to obtain a good approximation of the similarity assessment among entity classes by considering the essential properties as equally important. Psychologists and cognitive scientists, however, have pointed out that some features may be more important than others depending on context (Krumhansl 1978, Tversky 1977), since the classificatory significance of features varies with the set of entity classes under consideration.

This chapter presents an integration of contextual information into the MD model. The first section describes the thesis approach to modeling context through a user's intended operation. Subsequently, two approaches to determining relevant features are presented and explained with examples.

4.1 Modeling Context

Similar to the analysis of word meaning within statements (Leacock *et al.* 1993), similarity assessment is analyzed within a domain of discourse. In experimental studies of how people assess similarity, the domain of discourse is the set of entities that the subject observes and compares. Using information systems, however, it is unlikely that users could know the set of entities against which their queries will be compared. This work defines the domain of discourse (application domain as the set of entity classes that are subjects of the user's interest. Since a domain of discourse may change among applications, the similarity assessment changes as well.

This work derives the domain of discourse from the user's intended operations. The user's intended 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 operations, since they can change the operations' meaning depending on the kinds of noun arguments with which they co-occur (Fellbaum 1990). For example, different senses of the verb *play* are *play a role*, *play the flute*, and *play a game*. Hence, verbs together with their noun arguments describe the underlying goal for the use of the similarity assessment.

Contextual information (C) is specified as a set of tuples over operations (op_i) associated with their respective noun arguments (e_j) (Equation 4.1). The nouns correspond to entity classes in the MD model, while the operations refer to verbs that are associated as methods to these classes.

$$C = \left\langle \left(op_i, \{e_1, \mathsf{K}\ , e_i\}\right), \mathsf{K}\ , \left(op_j, \{e_k, \mathsf{K}\ , e_l\}\right) \right\rangle \tag{4.1}$$

In the specification of context an entity-class argument may be empty; if no, further explanation is needed to describe the intended operation. Since the context specification uses operations and entity classes, the knowledge base used by the entity-class representation of the MD model can be extended to represent the components of the context specification. For example, if a user wants to analyze some on-line datasets with the purpose of purchasing a cottage, she would describe her intention by C = <(*purchase*, {*cottage*})>. By using the hierarchical structure of the knowledge base, an operation's argument can be expressed at different levels of generalization. For example, a user may be looking for *sports facilities* and in such a case, she can specify C = <(*search*, {*sports facility*})> or C = <(*search*, {*athletic field*, *bowl park*, *tennis court*, *sports arena*, *stadium*})>. Another user's intention can be described by using

operations without arguments, such as $C = <(play, \{\})>$. In this case, the operation *play* corresponds to a common function that characterizes the entity classes the user is looking for.

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 because 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 children of the entity classes found in (1) and (2).

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 meaning (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. Unfortunately, this approach may not distinguish polysemous terms that are semantically similar and belong to the same domain of discourse.

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 to play a sport and may use a stadium as the prototypical entity to search in a database. In an information retrieval process, stadium will be compared with other entities 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

The variability approach relates the relevance of a feature to the degree of the feature's informativeness, such that if a feature is shared by all entity classes of the domain, its relevance decreases. For example, consider a small domain with buildings that differ in their structural characteristics, but have a common function (e.g., they all serve as sport facilities). 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, function, and attributes (ω_p ,ω_f, and ω_a of Equation 3.1) by analyzing the variability of distinguishing features within the application domain. In this sense, the type of distinguishing features that presents greater variability is more important in the similarity assessment than the type of features that do not contribute significantly to distinguishing entity classes. The variability of a type of feature *t* (P_t^v) is based on the

converse of the frequency with which each distinguishing feature of this type characterizes an entity class in the domain (Equation 4.2). 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.

$$P_t^v = 1 - \sum_{i=1}^{l} \frac{o_i}{n} \tag{4.2}$$

The final weights ω_p, ω_f, and ω_a (Equation 3.1) are functions of the variability of a type of feature with respect to the variability of the other two types of features (Equation 4.3a-c).

$$\omega_p = \frac{P_p^v}{(P_p^v + P_f^v + P_a^v)} \tag{4.3a}$$

$$\omega_f = \frac{P_f^v}{(P_p^v + P_f^v + P_a^v)} \tag{4.3b}$$

$$\omega_a = \frac{P_a^v}{(P_p^v + P_f^v + P_a^v)} \tag{4.3c}$$

When the application domain has maximum variability, that is, no feature is shared by entity classes or only one entity class is part of the application domain, the relevance for parts, functions, and attributes are equally assigned. Similar results occur without variability. In such a case, equal weights are assigned to the different types of distinguishing features.

4.2.2 Commonality

The commonality 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 to play a sport, a greater weight for this type of distinguishing 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 of Equation 3.1) by analyzing the frequency with which each distinguishing feature type characterizes an entity class in an application domain, that is, the converse of the measure given by the variability approach (Equation 4.4). High frequency is translated into a high relevance. In P_t^c, o_i is the number of occurrence of a feature in the entity class definitions, n is the number of entity classes, and l is the number of features in a domain of discourse.

$$P_t^c = \sum_{i=1}^{l} \frac{o_i}{n} = 1 - P_t^v \tag{4.4}$$

As in the variability approach, the final weights ω_p, ω_f, and ω_a in Equation 3.1 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 (Equation 4.5a-c).

$$\omega_p = \frac{P_p^c}{(P_p^c + P_f^c + P_a^c)} \tag{4.5a}$$

$$\omega_f = \frac{P_f^c}{(P_p^c + P_f^c + P_a^c)} \tag{4.5b}$$

$$\omega_a = \frac{P_a^c}{(P_p^c + P_f^c + P_a^c)} \tag{4.5c}$$

A special case occurs with maximum variability; that is, when each distinguishing feature characterizes only one entity class. In such a case, P_p^v, P_f^v, and P_a^v are 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, independently of the assignment of weights.

4.3 Using Contextual Information with the Matching-Distance Model

To illustrate the integration of context into the MD model, this section presents different specifications of context with their corresponding results of the MD model. These examples of context specification use the ontology derived from SDTS (USGS 1998) and WordNet (Miller 1995) described in Section 3.3.

The evaluations take a set of entity classes and apply a number of similarity assessments that use different context specifications. The scenarios are the following:

- 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 specifying that all entity classes in the domain have the function *play* (Figure 4.1), that is, an intentional specification of context, or by listing all the entity classes in the ontology that satisfied this condition (Figure 4.2), that is, an extensional

specification of context. What matters is to obtain an application domain with all the entity classes that are in fact of interest for 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 4.3. In this case, the application domain corresponds to 3% of the entire ontology.

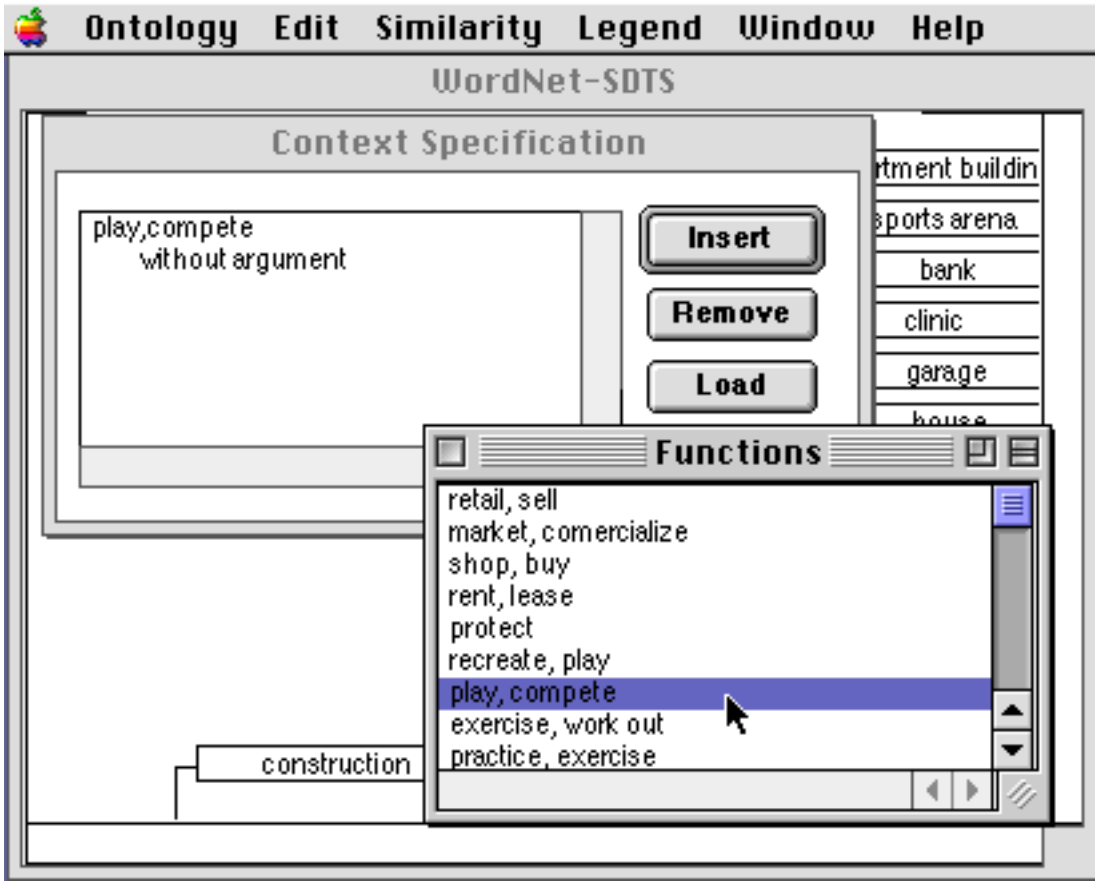

Figure 4.1: Intentional specification of context for a user who searches for a place to play a sport.

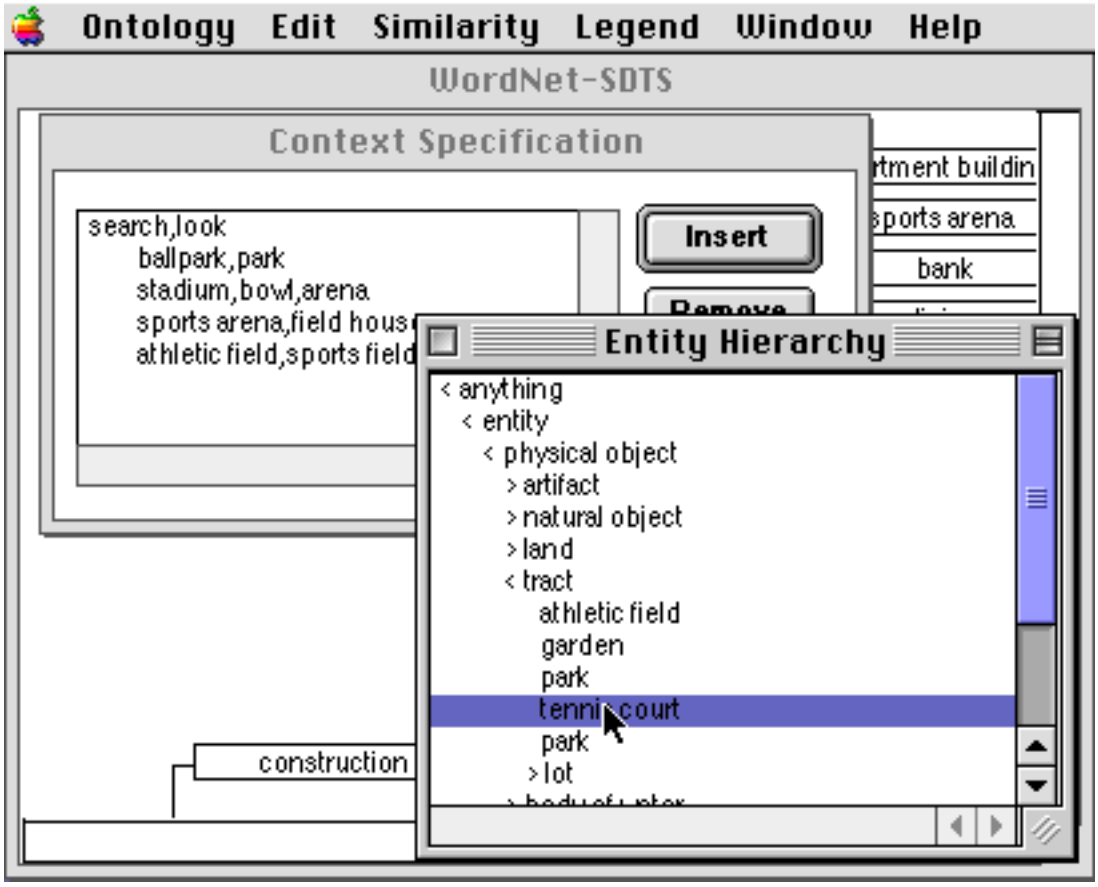

Figure 4.2: Extensional specification of context for a user who searches for a place to play a sport.

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). Figures 4.4 and Figure 4.5 present partial application domains for both context specifications. The application domain in the case of Context-2 represents 30% of the ontology and in the case of Context-3 7% of the ontology.

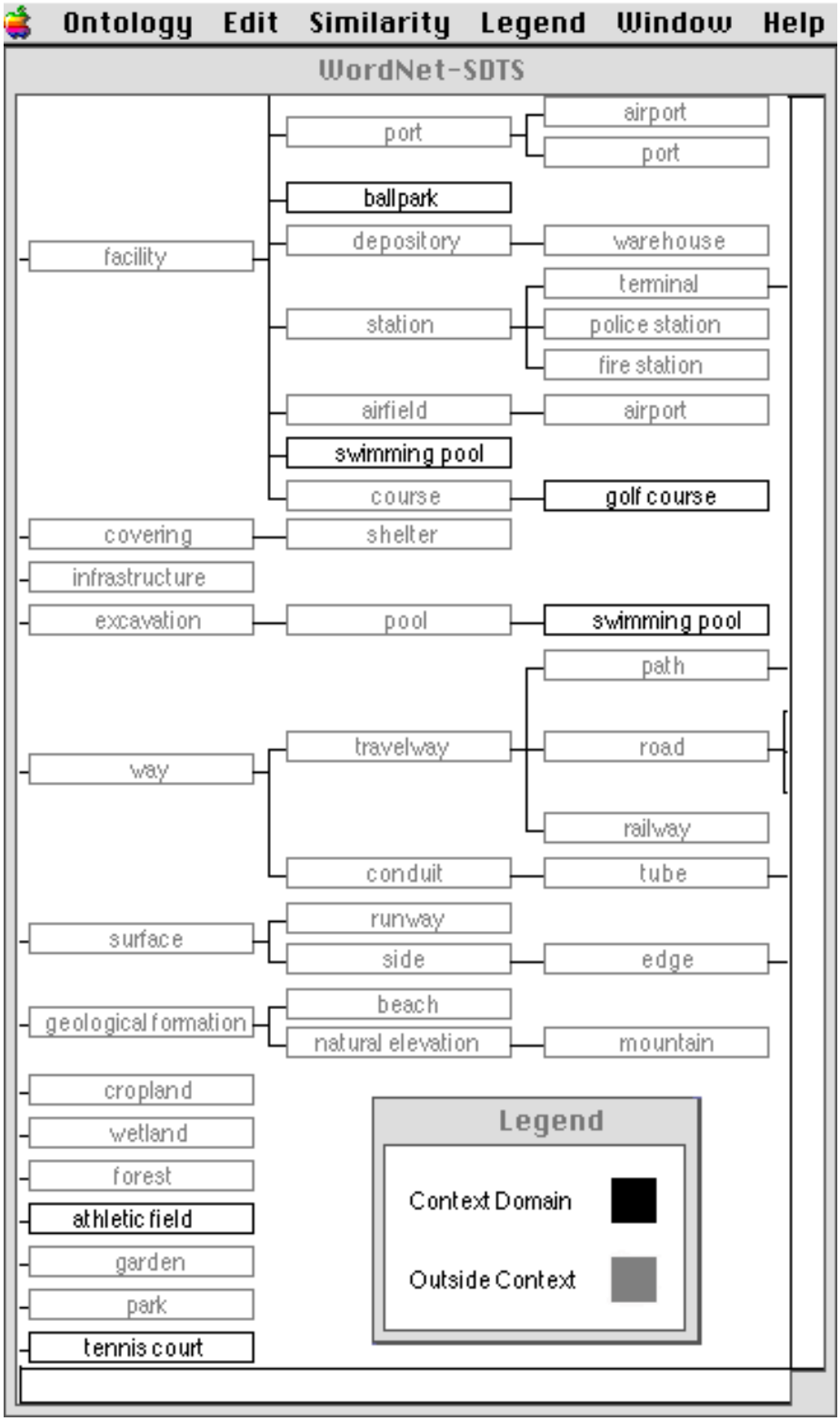

Figure 4.3: Application domain for a user who searches for a place to play a sport.

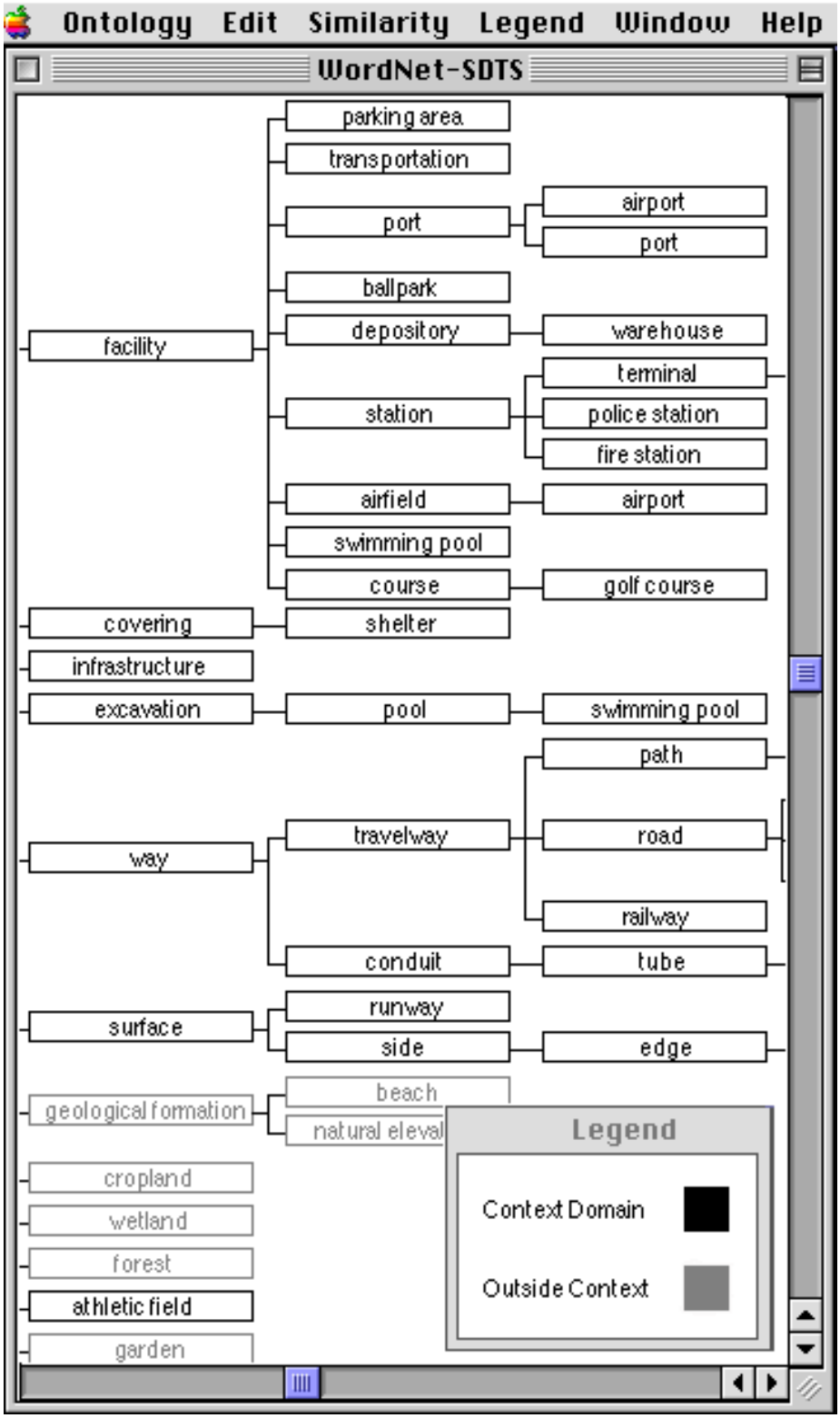

Figure 4.4: Application domain for a user who compares downtowns.

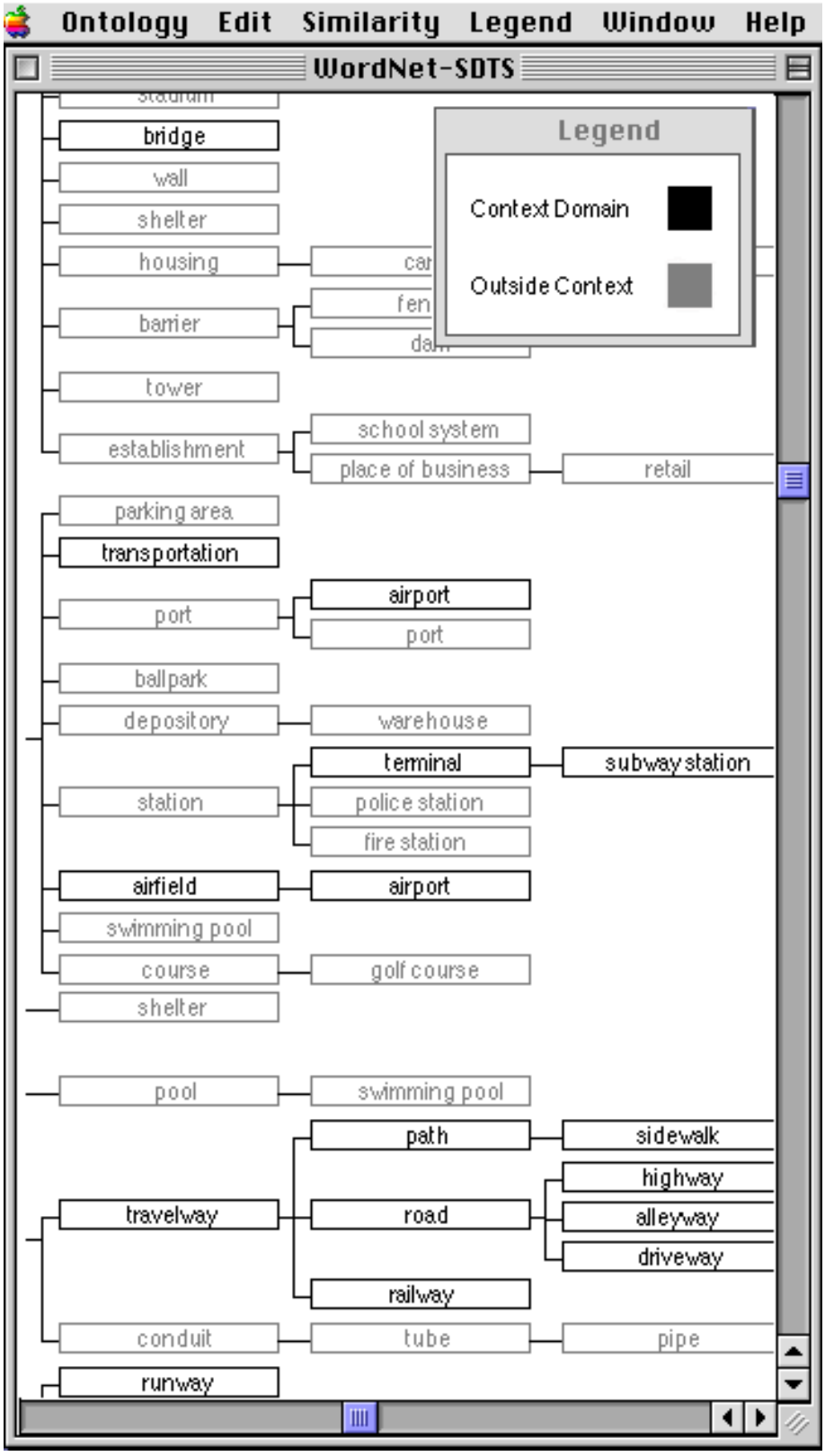

Figure 4.5: Application domain for a user who assesses a transportation system.

Table 4.1 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. An obvious observation is that a high weight in the commonality approach yields a low weight in the variability approach.

	Commonality			**Variability**		
Context	ω_p	ω_f	ω_f	ω_p	ω_f	ω_a
1	9%	62%	29%	46%	19%	35%
2	10%	13%	77%	36%	35%	29%
3	4%	29%	67%	45%	35%	20%

Table 4.1: Weights (%) for different specifications of context based on the commonality and variability approaches.

Table 4.2 presents results of the similarity evaluation 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 likelihood that increases the similarity values. Although absolute values are likely to vary with different approaches to weight determination, relative values in terms of ranks could remain invariable. When determining ranks, if ties occur, each tied rank is assigned the mean of the rank positions for which it is tied (Daniel 1978). For example, if the three most similar entity classes have the same value, the rank assigned to these entity classes is 2.

Table 4.2 shows similarity in terms of absolute values between 0 and 1 and Figure 4.6 presents the results in ranks. These results indicate that similarity values vary not only in terms of absolute values, but also in terms of ranks. Figure 4.6 suggests that changes occur depending on context specification as well as in terms of

approaches to weight determination. In terms of weight determination, the commonality approach produces more variation in the ranks than the variability approach. Overall, drastic changes are rare, and it is still possible to distinguish the group of most similar entity classes. In the next chapter, a human-subject experiment is used to evaluate the sensibility of the MD model with respect to people's judgments under different contexts.

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 4.2: 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 commonality and symbol *v* denotes variability.)

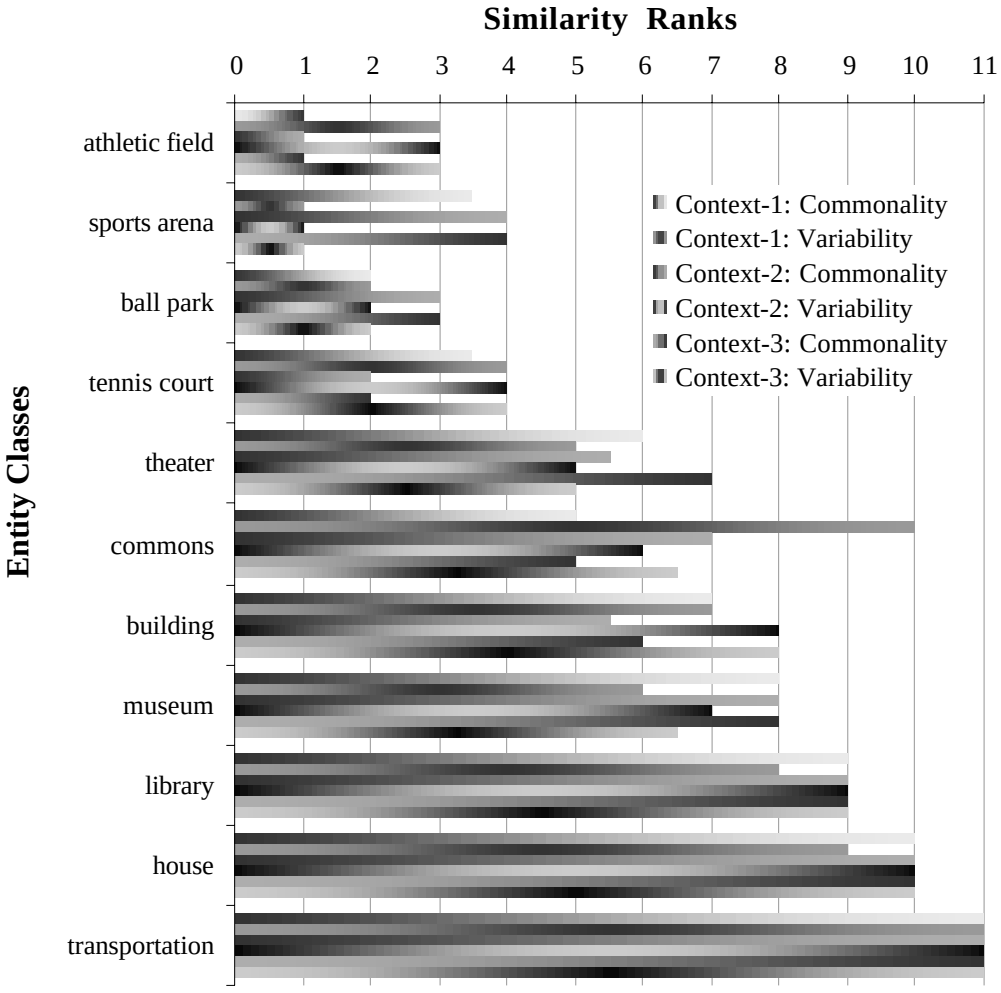

Figure 4.6: Results in ranks between a *stadium* and a portion of the WordNet-SDTS ontology for different context specifications and different approaches to weight determination.

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, evaluations that use the specification of Context-1 (a narrow application domain), but with slightly different ontologies, are performed. 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 4.3 shows the changes of weights for parts, functions, and attributes based on the commonality and the variability approaches and using subsets of the default application domain.

		Commonality			**Variability**		
Case	**Application Domain**	ω_p	ω_f	ω_f	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 4.3: 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 ontologies, that is, commonality highlights functions whereas variability highlights parts. Although changes occur depending on the set of entity classes in the ontology, the model 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.

4.4 Summary

The feature-distance model has been complemented with contextual information. Context is defined as the set of tasks and corresponding entity classes to which the tasks apply. The set of entity classes that belong to the application domain reduces the problem of word-sense ambiguity, since only these entity classes are considered in the similarity assessment. The variability or commonality of the entity class features that belong to the application domain determines the weights for the similarity of parts, functions, and attributes. The next section describes a human-subject experiment that tests whether the results given by the MD model are compatible with people's judgments.

Chapter 5

Assessment of the Matching-Distance Model

A model for similarity assessment is useful when it gives results that match people's judgments. Such cognitive evaluation of a computational model cannot be accomplished without comparing the model's results with people's judgments of similarity. The cognitive plausibility of the MD model is analyzed with a human-subject experiment whose design addresses the context dependence of similarity evaluations. The following sections describe the experiment and present subjects' responses. Subsequently, an evaluation of the MD model is based on the statistical analysis of these results.

5.1 Experimental Design

The experiment consisted of five questions with sets of entity classes that subjects were asked to rank according to their judgments of similarity (See Appendix). Four of the five questions (Questions 1–4) involve entity classes of a constructed kind, such as a building and a road. The last question addresses the similarity assessment among large geographic entities, such as a lake, a desert, and a forest. In this sense, the experiment attempts to capture any divergence in the similarity assessment of objects of a different kinds—natural vs. constructed.

The first three questions ask users to judge the same set of entity classes, but using different contextual information. Question 1 represents the default case of similarity assessment with no explicit contextual information. Questions 2 and 3 specify context defined as desired operations (i.e., “play a sport” and “compare constructions,” respectively). Question 4 uses a set of transportation-type entities, which becomes the contextual information of this question. As in the first question, the last question assumes the default case of a similarity assessment (i.e., no explicit contextual information).

In the MD model contextual information that is described as a natural-language statement is mapped onto a context specification. Table 5.1 shows the questions and the derived context specifications in the MD model. This mapping is manual, and future work should explore the automatic extraction of contextual information from natural-language statements. Although we assume that no context is given in Questions 1 and 5, contexts could be extracted in terms of the entity classes *place* and *entity*, respectively. The terms *place* and *entity*, however, are used in a generic way such that no particular application domain is implied.

Question	Natural-Language Statement	MD's Specification
1	How similar is a *stadium* (an *athletic field*) to the following places?	C = <>
2	How similar is a *stadium* (an *athletic field*) to the following places if you want to **play a sport**?	C = <(play*,{})>
3	How similar is a *stadium* (an *athletic field*) to the following places if you **compare *constructions***?	C=<(compare, {*construction*})>
4	How similar is a *travelway* (*path*) to these other ***transportation-type*** **entities**?	C=<(compare, {*transportation***})>
5	How similar is a *lake* to these other entities?	C = <>

Table 5.1: Contextual information as a natural-language statement and a formal specification in the MD model. (Symbol * denotes the sense of playing a sport and symbol ** denotes the sense of a transportation system.)

We can characterize questions by comparing the set of entity classes that are actually compared and the application domain that is derived from the context specification in the MD model. This comparison may yield some interesting conclusions, since the sets of entity classes that are actually compared in each question have also been described as contextual information that may influence the similarity evaluations (Krumhansl 1978, Tversky 1977). 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.

In order to keep the experiment short and check asymmetric evaluations, 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 ask for entity classes that are similar to a *stadium*, while Questions 1-3 in Survey B ask 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 asks 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 MD model. This ontology was derived from the combination of WordNet (Miller *et al.* 1990) and SDTS (USGS 1998) (Section 3.3). Since the goal of the experiment is to evaluate the similarity model rather than the entity class definitions, subjects were asked to judge similarity based on the set of definitions that were provided to them during the experiment and used by the MD model.

Seventy-two students (forty-three female and twenty-nine male) of an undergraduate English class at the University of Maine 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 is their mother tongue and their ages range from 18 to 36 years old. Subjects were paid for participating in the experiment ($2.00) and answered the questions at the same time and in less than twenty minutes.

5.2 Subjects' Responses

To avoid ambiguities, incomplete answers were eliminated. Thirty-three completed questionnaires were considered for Questions 1, 2, 4, and 5 in each survey. For Question 3 there were 32 completed answers. 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 (Table 5.2).

Type of rank	**Sports arena**	**Ball park**	**Athletic field**	**Tennis court**	**Theater**	**Library**	**Museum**	**Building**	**Commons**	**Transportation**	**House**
Original rank	1	1	2	2	3	3	3	4	4	4	5
Normalized rank	1.5	1.5	3.5	3.5	6	6	6	9	9	9	11

Table 5.2: An example of the normalization of subjects' responses.

The normalized answers were averaged and compared against the MD model. We found no significant evidence for differences based on gender, so the following results consider the total of responses for each survey. Figures 5.1 and 5.2 show the answers to Questions 1-3 of Survey A and Survey B, respectively.

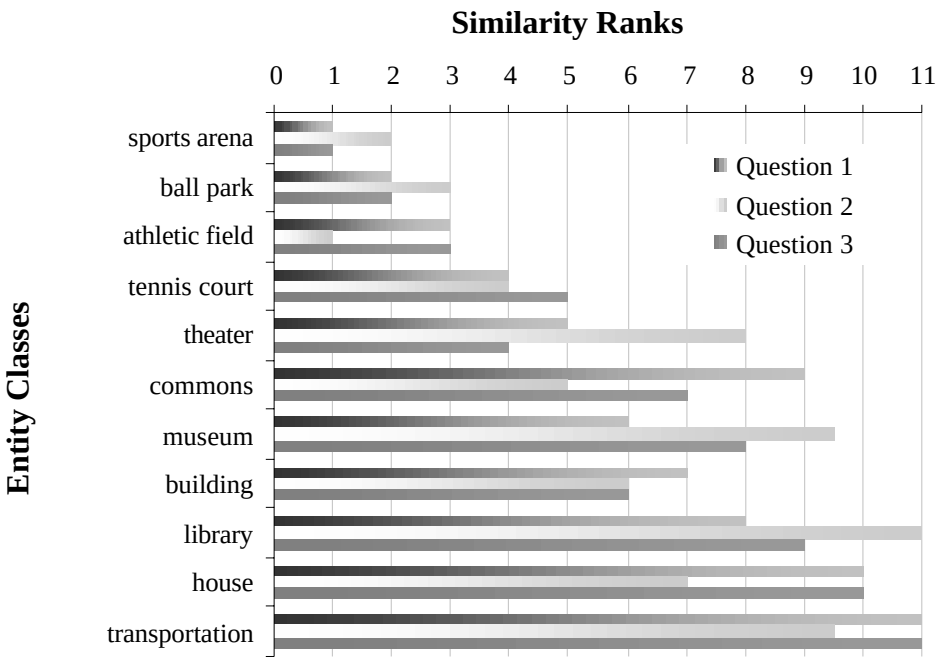

Figure 5.1: Subjects' responses to Questions 1, 2, and 3 of Survey A.

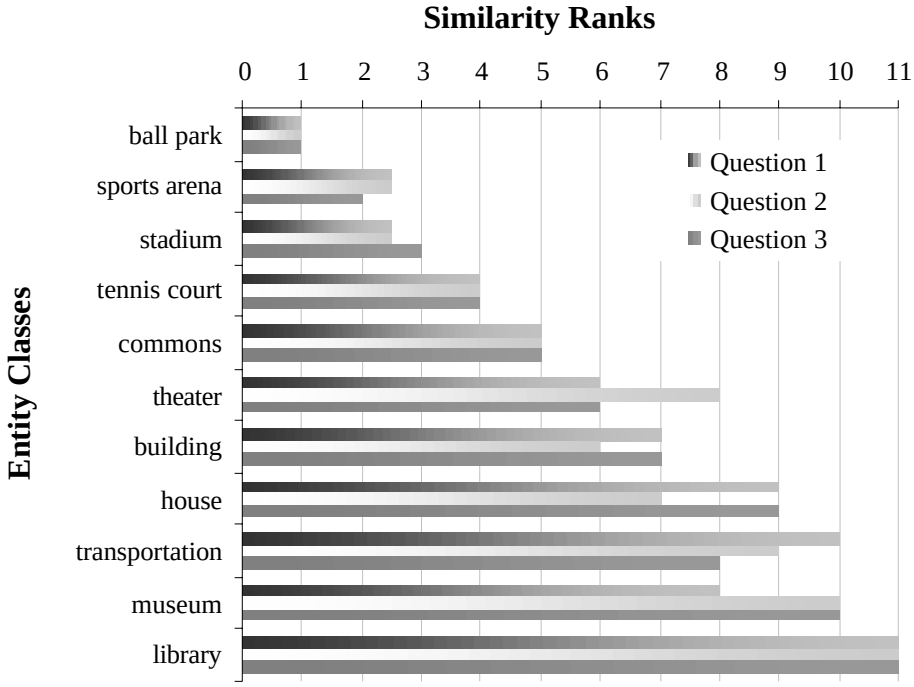

Figure 5.2: Subjects' responses to Questions 1, 2, and 3 of Survey B.

For the ranking of entity classes there is less variation across the answers to Questions 1-3 in Survey B than there is across the answers to Questions 1-3 in Survey A. While the three most similar entity classes in Survey A are the same independently of context, large variations exist for ranks higher than rank 4. In Survey B, in contrast, variations of ranks are confined between rank 5 and rank 9.

Table 5.3 presents the subjects' answers to question 4 in Survey A and Survey B. In both surveys *road* was the most similar entity class to either a *travelway* or a *path.* This was an unanticipated result, since *path* was explicitly defined as a subclass of *travelway* that is used for the "passage of persons or animals on land," whereas a road is also a subclass of *travelway,* but used for the "passage of vehicles on land."

Figure 5.3 corresponds to the subjects' responses to Question 5 in Survey A and Survey B. As expected, answers in Survey A and Survey B were very similar. Small variations are due to swapping of ranks 3-4 and 5-6.

Rank	Survey A: Question 4	Survey B: Question 4
1	road	road
2	highway	travelway
3	path	bridge
4	railway	highway
5	bridge	railway
6	transportation	transportation
7	subway station	subway station
8	airport *	terminal
9	terminal *	port
10	port	airport

Table 5.3: Answers to Question 4 in Survey A and Survey B. (Symbol * denotes tied ranks.)

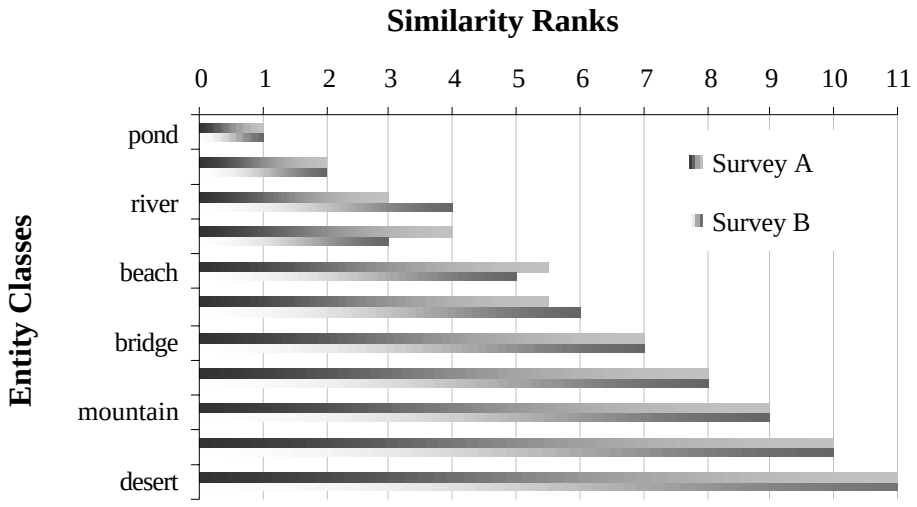

Figure 5.3: Subjects' responses to Question 5 in Survey A and Survey B.

5.3 Analysis

Four research hypotheses were formulated and statistically tested. While the first two hypotheses compare answers among subjects, the last two hypotheses compare subject's answers with respect to the MD model.

Hypothesis 1: Answers are associated.

The hypothesis is tested with Kendall's coefficient of concordance *W* for multiple rankings (Daniel 1978). The coefficient *W* leads to a non-parametric test; that is, it is a valid test under very general assumptions. The value of *W* is a measure of association whose extreme values 0 and 1 mean no association and perfect association, respectively. For situations with ties, the test statistic is given in Equation 5.1, where *m* is the number of sets of rankings, *n* is the number of objects that are ranked, R_j is the sum of the ranks assigned to the *j*th object, and t^3 is the is the number of observations in any set of rankings tied for a given rank.

$$W = \frac{12\sum_{j=1}^{n} R_j^2 - 3m^2 n(n+1)^2}{m^2 n(n^2-1) - m\sum(t^3 - t)} \tag{5.1}$$

For large samples (n > 30), a chi-square value is computed (Equation 5.2) and compared for significance with tabulated values of chi-square with *n*-1 degrees of freedom.

$$X^2 = m(n-1)W \tag{5.2}$$

This test uses the normalized subjects' responses such that each entity class in a question has 33 or 32 different ranks. The test statistic *W* and its corresponding X^2

value for each question in Survey A and Survey B are shown in Table 5.4. Based on the corresponding degrees of freedom, the research hypothesis is accepted with a probability of Type I equal to 0.005; that is, a probability of 0.005 that we accept the hypothesis when it is false. The large number of answers makes the test statistically significant; however, the values of W for Questions 3 and 4 are small (under 0.5), which means less agreement in these answers.

	Questions in Survey A					**Questions in Survey B**				
	1	2	3	4	5	1	2	3	4	5
W	0.70	0.76	0.37	0.45	0.64	0.69	0.64	0.33	0.45	0.70
X^2	230	252	120	134	210	226	210	107	135	231

Table 5.4: Test statistic W and the corresponding X^2 value for each question of Survey A and Survey B.

The standard deviations of the normalized rankings of each question in Survey A and Survey B (Figures 5.4 and 5.5) are an indication of whether people's judgments are more associated with some particular ranks. For instance, we expected that people would agree on the first and last ranks, but would have discrepancies in the similarity evaluations of the middle ranks. Figures 5.4 and 5.5, however, show that the agreement across ranks does not have a clear pattern. There is only a slight tendency to have more agreement among the first four ranks. As the coefficient of concordance W indicates, Questions 3 and 4 in both surveys have the largest standard deviations.

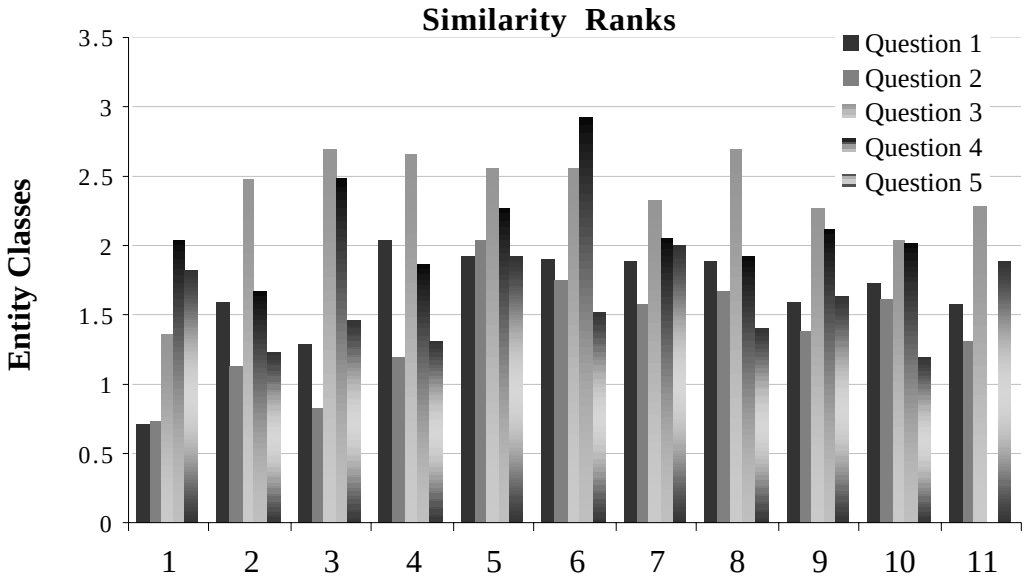

Figure 5.4: Standard deviations of questions in Survey A.

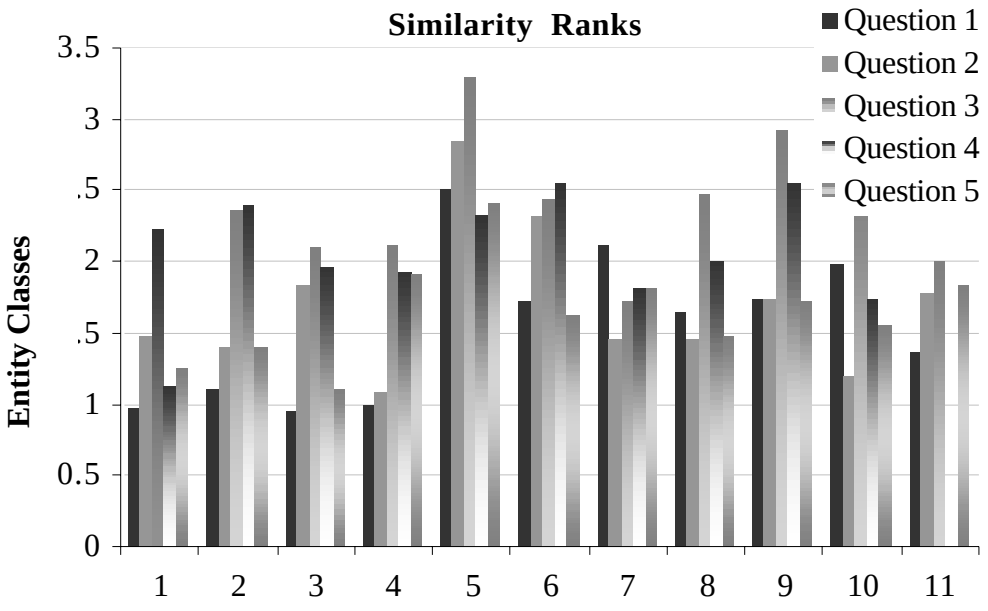

Figure 5.5: Standard deviations of questions in Survey B.

Hypothesis 2: People's judgments of similarity are context dependent.

This research hypothesis assumes that if people's judgments of similarity depend on context, their answers should vary across context. To test the hypothesis we compare the subjects' answers to questions with the same set of entity classes and different contextual information; that is, we compare Questions 1-3 of each survey. Since the same subject answers the three questions, we assume that his or her answers should be in perfect agreement to reject the hypothesis of context dependence. The normalized responses were averaged and this average was compared for each ranked entity class across different contexts using Kendall's coefficient of concordance *W*. The value of *W* for Survey A is 0.88 and for Survey B is 0.96. These values of *W* are high and suggest that the ranks are associated with a probability greater than 0.99 in both surveys.

To make sure that the similarity among contexts could not affect the result of the test statistic, we compare only the two questions with explicitly different context (i.e., Questions 2 and 3). Question 1, without explicit contextual information, could be a default context that is implicit in the contextual information of other questions. For comparing two sets of rankings, the test statistic is the Spearman rank correlation coefficient (Gibbons 1976). Equation 5.3 gives the expression of the Spearman coefficient under the presence of ties, where *n* is the number of objects that are ranked, *D* is the difference between paired ranks, and *u* and *v* are the numbers of observations in each set of rankings that are tied for a given rank. The test statistic is also a measure of association. As such, r_s should be equal to +1 when there is a perfect direct relationship between rankings.

$$r_s = \frac{n(n^2-1) - 6\sum D_i^2 - 6(u' + v')}{\sqrt{n(n^2-1) - 12u'}\sqrt{n(n^2-1) - 12u'}} \qquad \text{with}$$

$$u' = \frac{\left(\sum u^3 - \sum u\right)}{12} \quad \text{and} \quad v' = \frac{\left(\sum v^3 - \sum v\right)}{12} \tag{5.3}$$

The values of r_s in Survey A and Survey B are 0.80 and 0.95, respectively. Like the values of *W*, values of r_s are high, which suggests that ranks are associated with a probability of accepting the association when it is false equal to 0.01. Such a high level of agreement suggests that people would give more or less the same evaluation under different contexts, which is against the hypothesis.

Hypothesis 3: People's judgments of similarity and results of the MD model with default weights are correlated.

This test compares people's judgments of similarity with the results of the model, assuming that distinguishing features are equally important and people's judgments are independent of context. The comparison is based on the average of the normalized responses that were given to entity classes in each question. The average of the normalized responses is used since, by the first research hypothesis, we found that responses were associated. This hypothesis is tested with the Spearman rank correlation coefficient (Equation 5.3). The test statistic r_s for each question is shown in Table 5.5. All values of r_s are over 0.75, which supports the hypothesis that people's judgment and the MD model are associated with a probability that accepting the hypothesis when it is fail 0.01.

Question in Survey A					**Question in Survey B**				
1	2	3	4	5	1	2	3	4	5
0.96	0.83	0.95	0.90	0.78	0.92	0.87	0.90	0.88	0.86

Table 5.5: Spearman rank correlation coefficients between subjects' responses and the MD model with default weights.

Hypothesis 4: The correlation between people's judgments and the MD model improves when context is considered.

This test uses the same approach of the evaluation of hypothesis 3, but considering only questions with explicit contextual information (i.e., Questions 2-4). The goal is to compare the correlation between the subjects' responses and the MD model when weights of distinguishing features are calculated based on contextual information. The correlation of the results is calculated with the Spearman rank correlation coefficient (Equation 5.3). Table 5.6 shows that all values of r_s are high, which represents a high association between the subjects' answers and the computational model.

		Default	**Commonality**	**Variability**
Question 2	A	0.83	**0.85**	0.68
	B	0.87	0.87	**0.88**
Question 3	A	0.95	0.87	**0.96**
	B	0.90	0.87	**0.91**
Question 4	A	0.90	0.78	**0.96**
	B	0.88	0.84	**0.91**

Table 5.6: Spearman rank correlation coefficients between subjects' responses and the MD model with different approaches for weight determination.

The commonality and variability approaches have opposite effects on the results. While one of the approaches increases the correlation, the other one decreases it. The default setting and the weights derived from the variability approach in Question 3 have no significant difference. The correlation between people and the MD

model for Question 2 in Survey B has no significant difference across contexts. The greatest improvement of the correlation is found in Question 4 when the variability approach is used for the determination of the weights of distinguishing features.

Another observation from Table 5.6 is that a wrong strategy for weight determination may decrease more strongly the correlation than a right strategy may increase it. For instance, in Question 4 of Survey A the commonality approach decreases by 13% the correlation between the subjects' answers and the default evaluation of the MD model. The variability approach for the same question, in contrast, increases the correlation by 7%.

In summary, the results support the thesis hypothesis stated in Chapter 1 that the model matches people's judgments. The model can better represent people's judgments when context and the right approach for weight determination are considered. In general, the experiment has suggested that the variability approach produces a better correlation between the computational model and the subjects' answers when context specifies a particular type of entity classes. The commonality approach, in contrast, works well for context specifications based on particular functions of the entity classes.

5.4 Discussion

Previous work on semantic similarity has also applied human-subject experiments to determine the effectiveness of computational models. The experiments found a correlation of 0.60 using a semantic distance approach, 0.79 using an information content approach, and 0.83 using a combined distance approach (Jiang and Conrath 1997, Resnik 1999). Our human-subject experiment is not comparable to these experiments, because it has focused on a narrow domain, spatial entity classes, and has

considered contextual information for the similarity assessment. In addition, while most previous experiments evaluate similarity among quite different concepts (e.g., car, brother, coast, and journey), our experiment uses entity classes that are semantically related (e.g., *ballpark*, *stadium*, and *athletic field*) to study the performance of the model at a detailed scale.

The results of the human-subject experiment support the use of the MD model for semantic similarity among entity classes. Correlation between the model and the subjects' answers was 0.78 in the worst case and 0.96 in the best case. An important observation is that although subjects' responses are associated, the degree of concordance among subjects' answers is unsatisfactory (0.33 – 0.76) when compared to previous experiments on semantic similarity (e.g., 0.90 in Resnik's (1999) experiment). This low degree of concordance may be due to the large number of entity classes that were evaluated with respect to the same target and the use of entity classes that are semantically related.

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; however, the major determinant for the high correlation between the MD model and subjects' answers seems to be the correct identification of distinguishing features of entity classes. For example, an important difference between the model and subjects' answers was the least similar entity class 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 (Mark *et al.* 1999, Rosch 1973, Rosch and Mervis 1975).

5.5 Summary

This chapter has evaluated the performance of the MD model with a set of similarity evaluations under different contexts. The experiment confirmed that the MD model gives a good approximation of human subjects' similarity assessment among spatial entity classes. A small improvement of the correlation was found when contextual information was used to determine weights of distinguishing features. The experiment suggests that the major factor for the high correlation of the computational model with people's judgments is the correct characterization of entity classes through distinguishing features. Next chapter further expands the basic MD model to account for definitions that come from different ontologies.

Chapter 6

A Computational Model for Semantic Similarity Across Ontologies

With the increasing interest in providing seamless access to distributed information, information integration has become more relevant. At the semantic level, information systems that have different conceptualizations also differ in the intended models of these conceptualizations, that is, in their underlying ontologies (Guarino 1998). Current approaches to dealing with definitions that come from different ontologies make the original ontologies subscribe to a shared ontology (Bishr 1997, Bright *et al.* 1994, Collet *et al.* 1991, Fankhauser and Neuhold 1993, Weinstein and Birmingham 1999) or create a global ontology from the integration of the existing ones (Bergamaschi *et al.* 1998, Kashyap and Sheth 1998, Mena *et al.* 1996). Both approaches have limitations when updates in the original ontologies occur, since changes may invalidate the relationships between the existing ontologies and the global or integrated ontology. They also require off-line user intervention for choosing the terms to integrate or for formalizing the shared ontology; therefore, alternative methods are needed to allow information access across ontologies.

This chapter introduces the Triple Matching-Distance model (MD3) that connects existing ontologies through a concept of similarity. The MD3 model extends the MD model to evaluate semantic similarity across independent ontologies. The similarity model aims at finding the most similar entity classes across ontologies by

using the common specification components of the entity class representations. Such a similarity relation establishes anchors between 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 concepts in the ontology and cannot insure computations that require particular components of the entity class representation. It provides, on the other hand, a systematic way to detect what terms are the most similar and, therefore, what terms are the best candidates for establishing an integration across the ontologies. This form of integration is particularly useful in dynamic environments, such as the Internet, where it would be unrealistic to force users to subscribe to a single, global ontology.

The following section discusses ontology mismatches that affect similarity evaluations across ontologies. This discussion is followed by the strategy that is used to extend the MD model and the presentation of the computational formalization of the MD3 model. Subsequently, the performance of the MD3 model is tested with three different ontologies: WordNet (Miller *et al.* 1990), the Spatial Data Transfer Standard (USGS 1998), and a combination of WordNet and SDTS (Section 3.3).

6.1 Ontology Mismatches

Handling multiple ontologies at the same time requires solving discrepancies in the definition of entity classes. To illustrate different scenarios when comparing two ontologies, assume two entity classes E_1 and E_2 belonging to two independent ontologies. The possible scenarios are:

- E_1 and E_2 are the same entity class that is represented in the same way,
- E_1 and E_2 are the same entity class that is represented in different ways,

- E_1 and E_2 are different entity classes that are represented in the same way, and
- E_1 and E_2 are different entity classes that are represented in different ways.

Among these scenarios, the second and third scenarios represent ontology mismatches. Visser *et al.* (1998) gave a comprehensive description and classification of ontology mismatches in terms of the two processes that are involved in the creation of an ontology: (1) conceptualizing a domain and (2) explicating the conceptualization (Table 6.1). This classification of ontology mismatches resembles the studies of semantic heterogeneity in the database field (Ceri and Widom 1993, Kim and Seo 1991). Indeed, ontology mismatches are present in any form of conceptualization, such as databases and knowledge bases.

This research compares representations of entity classes and defines a similarity model in terms of the commonality among these representations. Since the semantics of an entity class may be represented in more than one way, equivalent or similar entity classes whose definiens are different are not detected, which establish a *definiens mismatch*. For example, consider an ontology in which entity classes are represented in a hierarchical structure of is-a relations and a second ontology in which entity classes are represented by attributes. Although both ontologies may include similar entity classes, their representations are different and, therefore, comparing representations is insufficient for identifying any similarity between them.

	Mismatches	Description
Conceptualization mismatches	Class	Class and subclass distinction
	Categorization	Same class and different subclasses
	Aggregation-level	Classes at different levels of abstraction
	Relation	Relation distinction
	Structure	Same classes with different relations
	Attribute assignment	Same attribute, but different classes
	Attribute type	Same attribute with different instantiations
Explication mismatches	Concept-Term	Same definiens for different terms and concepts
	Concept-Definiens	Same term for different concepts and definiens
	Concept	Same term and definiens for different concepts
	Term-Definiens	Same concept with different terms and definiens
	Term	Same concept and definiens with different terms
	Definies	Same concept and terms with different definiens

Table 6.1: Types of ontology mismatches. (Term refers to concepts' names and definiens refers to the elements that define concepts.)

6.2 Extending the Matching-Distance Model

The MD model applies over an ontology in which entity classes are semantically interrelated in a hierarchical structure. This hierarchical structure corresponds to a partially ordered set (Birkhoff 1967) in which any two entity classes can be linked by a common superclass (i.e., least upper bound). In a cross-ontology evaluation, however, there is no such common superclass between entity classes, which constrains the MD

model to environments with a single ontology. Since a common superclass in the MD model is used for determining the relative level of abstraction of entity classes, it is possible to obtain an approximation of this level of abstraction by considering that two independent ontologies are connected by an imaginary and more general entity class *anything* (Figure 6.1).

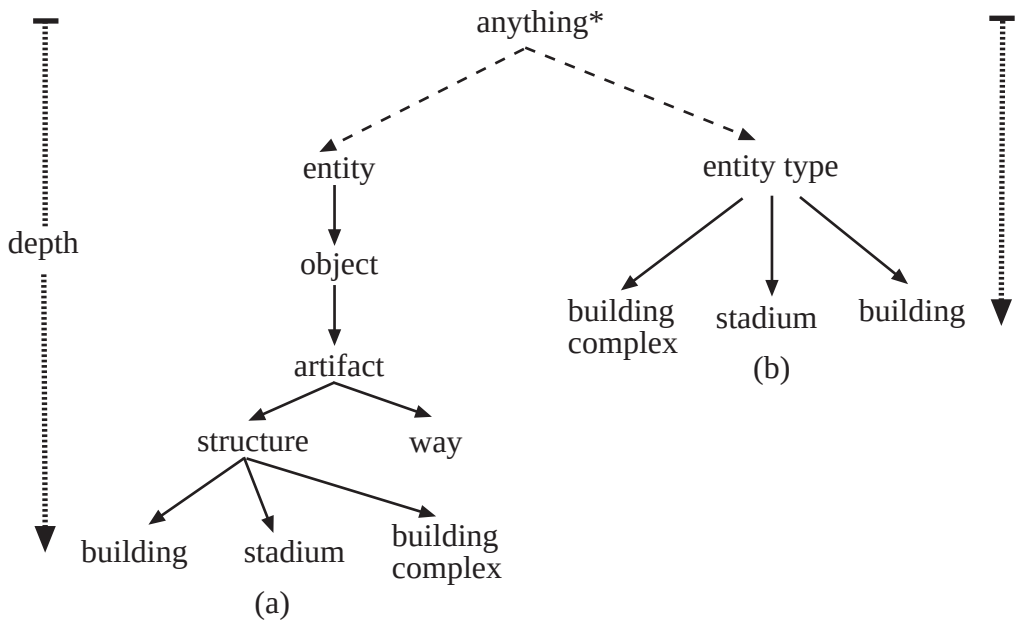

Figure 6.1: Connecting independent ontologies: (a) partial WordNet ontology and (b) partial SDTS ontology. (Anything* denotes an imaginary root.)

Using this connected ontology, α of the MD model (Equation 3.2) could be expressed as a function of the *depth* of the entity classes (Equation 6.1). The function *depth*() corresponds to the shortest distance from the entity class to the imaginary root. Equation 6.1 is equivalent to Equation 3.3 of the MD model, since the imaginary root is the only common superclass between entity classes of independent ontologies. Equation 6.1, however, results in values greater than zero and less than or equal to 0.5. This range of α excludes the extreme values (i.e., 0 and 1) that characterize α and

$(1-\alpha)$ of the MD model when an entity class is subsumed by another one, which is an impossible situation for cross-ontology evaluations. For example, consider the ontologies in Figure 6.1. 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.

$$\alpha(a^p, b^q) = \begin{cases} \dfrac{depth(a^p)}{depth(a^p) + depth(b^q)} & depth(a^p) \le depth(b^q) \\ 1 - \dfrac{depth(a^p)}{depth(a^p) + depth(b^q)} & depth(a^p) > depth(b^q) \end{cases} \quad (6.1)$$

Different levels of explicitness and formalization of the ontologies influence the way entity classes can be compared. This type of discrepancy in the entity class representation becomes more important when comparing entity classes from different ontologies. Indeed, similarity evaluations across ontology can only be achieved if the representation of entity classes in those ontologies share some common components. Since the MD model gives similarity values in terms of common and different distinguishing features (i.e., parts, functions, and attributes), it is unable to assess similarity in existing ontologies whose definitions exclude distinguishing features, such as the SENSUS taxonomy (Knight and Luk 1994) and the UMLS Metathesaurus (NLM 1997). An approach to overcome this limitation of the MD model is to consider all components of the entity class representation such that the chance of having common elements upon which similarity could be determined is increased. Section 3 discussed the main components of the entity class representation, which are summarized in Table 6.2.

Components	Description
Definiendum	Lexicon or synonym set that refers to an entity class
Definiens	What is used to define an entity class
Semantic Relations	Relations to other entity classes
Hyponymy	Is-a relation (Smith and Smith 1977)
Meronymy	Component-object and stuff-object relations (Winston *et al.* 1987)
Distinguishing Features	Property of the entity classes
Parts	Structural elements
Functions	What is done to or with instances of a class
Attributes	General characteristics of a class

Table 6.2: Components of the entity class representations.

Besides features, lexicons and semantic relations are components of entity class representations that can be compared. Comparing entity class lexicons is an inconclusive form of similarity assessment, since lexicons can be different, but the entity classes can still be semantically similar. An example is *building* and *hospital*, which have only a few characters in common and, therefore, their string matching is very low. Their semantic similarity, however, is fairly high. Inversely, entity class lexicons can be the same whereas the entity classes are semantically unrelated (polysemous terms). In a cross-ontology evaluation, comparing entity class lexicons exploits the general agreement in the use of terms and detects equivalent terms that likely refer to the same entity class. Thus, similar entity class lexicons can be used to

detect equivalent or synonym entity classes across ontologies. As such it makes a syntactic comparison and provides a very basic level of similarity assessment.

Unlike approaches that use semantic relations to determining semantic distances in a hierarchical structure (Rada *et al.* 1989), our approach treats the semantic relations themselves as the subject of comparison. Since the types of semantic relations are predefined, 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 of an entity class in a semantic network is the set of entity classes whose *distance* to the entity class is less than or equal to a specified value, a value called the *radius* of the semantic neighborhood. The *distance* between two entity classes in the semantic network is measured as 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. 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. Equation 6.2 gives a formal definition of the semantic neighborhood (N), where a^o and c_i^o are entity classes in an ontology o, r is the specified radius, and $d()$ is the distance between the two entity classes.

$$N(a^o, r) = \left\{c_i^o\right\} \text{ such that } \forall i \; d(a^o, c_i^o) \leq r \tag{6.2}$$

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. 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 (Miller *et al.* 1990) includes the *stadium*, its superclass *structure* and, its parts *athletic field* and *sports arena* (Figure 6.2).

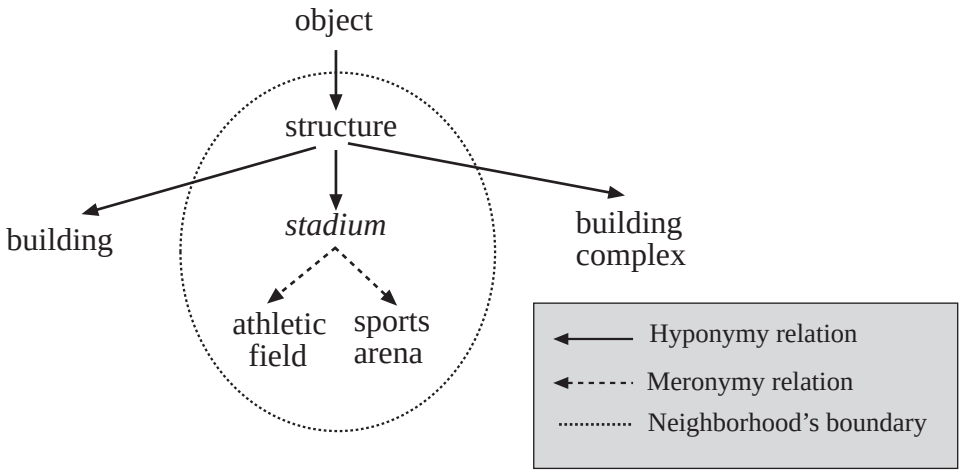

Figure 6.2: Example of the immediate semantic neighborhood of *stadium* in a portion of the WordNet ontology.

Logically, there is an inverse relation between the similarity of semantic neighborhoods and the determination of a semantic distance (Rada *et al.* 1989). As the semantic distance increases between entity classes, the semantic neighborhood becomes less similar. Unlike semantic distance, however, semantic neighborhood does not require a connecting path between entity classes.

Comparing all components of the entity class representation raises the issue of dependence among these components (Figure 6.3). This type of dependence implies that comparing semantic neighborhood is a recursive process, since it involves the similarity assessment of entity classes in the neighborhood. Since part features may also be entity classes, comparing parts may also involve a recursive process. This work, however, considers part features in the same way as the other types of features whose representation is given by the term or the synonym terms that refer to them.

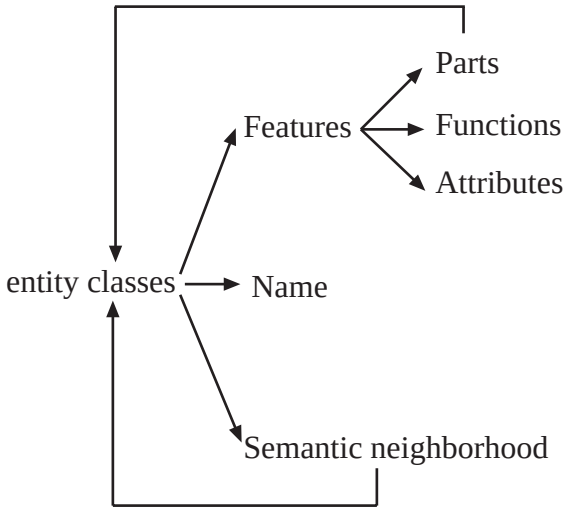

Figure 6.3: Dependence among components of the entity class representation.

The following section describes an approach for the similarity assessment among components of the entity class representation in independent ontologies, which is called the Tripe Matching-Distance model.

6.3 The Triple Matching-Distance Model

As in the case of a single ontology, the MD3 model compares components of entity class representations in terms of a matching process (Tversky 1977). A matching process is exempt from having the ontologies' hierarchies interconnected, which is a characteristic of models based on semantic distance. A matching process can result in asymmetric values and account for context dependencies (Section 2.2).

For cross-ontology evaluations the matching process is applied in successive steps to different specification components: (1) lexicon matching, (2) feature matching, and (3) semantic-neighborhood matching. The global similarity is then a weighted sum of the similarity of each component (Equation 6.3). The parameters S_l, S_u, and S_n are the similarity among names, features, and semantic neighborhoods, respectively, and their weights ω_l, ω_u, and ω_n add up to 1.0. A threshold over the global similarity can be used to avoid irrelevant calculations.

$$S(a^p, b^q) = \omega_l \cdot S_l(a^p, b^q) + \omega_u \cdot S_u(a^p, b^q) + \omega_n \cdot S_n(a^p, b^q) \quad \text{for } \omega_l, \omega_u, \text{ and } \omega_n \geq 0 \quad (6.3)$$

Weights assigned to S_l, 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 should add up to 1.0. Lexicon similarity can always be a factor of the similarity assessment, but when polysemous terms occur within an ontology, lexicon similarity is a less likely indication of semantic similarity among entity classes. For example, one ontology may include different meanings of the term *bank* (e.g., a financial institution, a sloping of land, and a pile), whereas another ontology may contain only one meaning of *bank* (e.g., a financial institution). Using only lexicon similarity, 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. Lexicon similarity complemented with feature and semantic-neighborhood similarity, on the other hand, can highlight the similarity between corresponding senses of the term *bank*.

6.3.1 Lexicon Matching

In the MD3 model, lexicon matching checks the number of common and different words in the names of entity classes. For example, consider the ontologies of Figure 6.1. The lexicon matching between *building* of WordNet ($building^w$) and *building*

complex of SDTS ($building\_complex^s$) is 0.58 for α = 0.28 (Equation 6.4). Likewise, lexicon matching between $stadium^w$ *and* $stadium^s$ results in 1.0, independently of the value α.

$$S_l(building^w, building\_complex^s) = \frac{|\{building\}|}{|\{building\}| + 0.28|\{\}| + 0.72|\{complex\}|} = \frac{1}{1.72} = 0.58 \quad (6.4)$$

In cases where synonym sets refer to entity classes, the lexicon matching finds the most similar terms between synonym sets. For example, given the synonym sets (*Sys*) in Equation 6.5 that refer to the entity classes $building^w$ and $building\_complex^s$ of Figure 6.1, respectively, the result of the lexicon matching is 0.58. This value of lexicon matching results from comparing $building^w$ and $building\_complex^s$, since the lexicon matching between $edifice^w$ and $building\_complex^s$ is zero.

$$Sys(building^w) = \{building, edifice\}$$
$$Sys(building\_complex^s) = \{building\_complex\} \quad (6.5)$$

Giving the result of lexicon matching as a function of only the most similar terms between synonym sets, we consider it unlikely to have the same number of synonyms in different ontologies. Using a stricter approach, the model could also apply a matching process over the synonym sets such that the number of common and different synonyms in the sets would affect the similarity value. This approach may result in low values of lexicon matching, since the mere missing of one of the synonyms in the set reduces the value of lexicon matching considerably. For example, consider the synonym sets of the entity classes $airport^w$ and $airport^s$ (Equation 6.6). Using the most similar terms between synonym sets, lexicon matching between $airport^w$ and $airport^s$ is equal to 1.0, whereas using the common and different synonyms in the sets yields to a lexicon matching equal to 0.5, for α equal to 0.5.

$$SS_{aiport^w} = \{airport, airdrome, aerodrome\}$$
$$SS_{airport^s} = \{airport\} \quad (6.6)$$

6.3.2 Feature Matching

Feature matching is equivalent to the MD model for independent ontologies connected by an imaginary root (Equations 3.1 and 3.2). If both ontologies classify features into parts, attributes, and functions, a weighted sum of the corresponding similarity of each type of features yields the global feature similarity (Equation 3.1). By default, the types of distinguishing features that are present in the ontologies' specifications are considered equally important. When no classification of distinguishing features is given, a global feature-matching process is performed.

Existing ontologies may have schematic conflicts that are the product of different feature classifications. For example, while WordNet's definitions identify parts of entity classes, SDTS denotes the features of entity classes as attributes. In such a case, a comparison of features by type would find no common features in cross-ontology evaluations. In order to avoid this type of schematic conflict, different types of distinguishing features can also be compared. For example, a feature *lane* can be a *part* or an *attribute* in the entity class representation of a *road*.

This work makes a syntactic, rather than semantic, representation of distinguishing features. Thus, a distinguishing feature is represented by a lexicon or a synonym set and the feature matching process applies a string matching over the lexicons or synonym sets that refer to these features. String matching over distinguishing features is a strict string matching such that distinguishing features match only if they are represented by the same lexicon or by synonym sets that intersect. This process ignores similarity between compound terms, such as between *lane* and *number of lanes*; however, strict string matching is a fast comparison of

feature lexicons for large ontologies where the percentage of partial string matching among feature lexicons is limited. For example, consider the distinguishing features of *stadium* in WordNet ($stadium^{w}$) and in an *ad-hoc* ontology called 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 (Equations 6.7a-b).

$$Parts(stadium^{w}) = \begin{Bmatrix} foundation^{w}, midfield^{w}, playing\_field^{w}, plate^{w}, \\ sports\_arena^{w}, stands^{w}, standing\_room^{w}, \\ strutural\_elements^{w}, tiered\_seats^{w} \end{Bmatrix} \quad (6.7a)$$

$$Parts(stadium^{ws}) = \begin{Bmatrix} dressing\_room^{ws}, foundation^{ws}, midfield^{ws}, \\ playing\_field^{ws}, spectator\_stands^{ws}, ticket\_office^{ws} \end{Bmatrix} \quad (6.7b)$$

Both WordNet and WS represent distinguishing features by synonym sets (Equations 6.8a-o), where *Sys* denotes the representation of a distinguishing feature as synonym sets.

$$Sys(foundation^{w}) = \{foundation\} \quad (6.8a)$$

$$Sys(midfield^{w}) = \{midfield\} \quad (6.8b)$$

$$Sys(playing\_field^{w}) = \{playing\_field, athletic\_field, field\} \quad (6.8c)$$

$$Sys(plate^{w}) = \{plate\} \quad (6.8d)$$

$$Sys(sports\_arena^{w}) = \{sports\_arena, field\_house\} \quad (6.8e)$$

$$Sys(stands^{w}) = \{stands\} \quad (6.8f)$$

$$Sys(standing\_room^{w}) = \{standing\_room\} \quad (6.8g)$$

$$Sys(tiered\_seats^{w}) = \{tiered\_seats\} \quad (6.8h)$$

$$Sys(structural\_elements^{w}) = \{structural\_elements\} \quad (6.8i)$$

$$Sys(dressing\_room^{w}) = \{dressing\_room\} \quad (6.8j)$$

$$Sys(foundation^{w}) = \{foundation\} \quad (6.8k)$$

$$Sys(midfield^{ws}) = \{midfield\} \quad (6.8l)$$

$$Sys(spectator\_stands^{ws}) = \{spectator\_stands, stands\} \quad (6.8m)$$

$$Sys(ticket\_office^{ws}) = \{ticket\_office, box\_office, ticket\_booth\} \quad (6.8n)$$

$$Sys(playing\_field^{ws}) = \{playing\_field, athletic\_field, sports\_field\} \quad (6.8o)$$

For representations based on synonym sets, we say that two distinguishing features are the same if the intersection of their synonym sets is different than the empty set (Equation 6.9).

$$F = G \quad iff \quad Sys(F) \cap Sys(G) \neq \{\} \quad (6.9)$$

The set of common distinguishing features between $stadium^{w}$ and $stadium^{ws}$ defines their set intersection (Equation 6.10a-e).

$$X = Parts(stadium^{w}) \cap Parts(stadium^{ws}) = \begin{Bmatrix} foundation^{w}, midfield^{w}, \\ playing\_field^{w}, stands^{w} \end{Bmatrix} \quad \text{with} \quad (6.10a)$$

$$foundation^{w} = foundation^{ws} \quad (6.10b)$$

$$midfield^{w} = midfield^{ws} \quad (6.10c)$$

$$playing\_field^{w} = playing\_field^{ws} \quad (6.10d)$$

$$stands^{w} = spectator\_stands^{ws} \quad (6.10e)$$

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 $stadium^{ws}$ (Equations 6.11a-b).

$$Y = Parts(stadium^{w}) - Parts(stadium^{ws}) = \begin{Bmatrix} plate^{w}, sports\_arena^{w}, \\ standing\_room^{w}, \\ structural\_elements^{w}, \\ tiered\_seats^{w} \end{Bmatrix} \quad (6.11a)$$

$$Z = Parts(stadium^{ws}) - Parts(stadium^{w}) = \{dressing\_room^{ws}, ticket\_office^{ws}\} \quad (11b)$$

The similarty measure between distinguishing features of $stadium^{w}$ and $stadium^{ws}$ is then determined by Equation 6.12 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.

$$\begin{aligned} S_u(stadium^{w}, stadium^{ws}) &= S_p(stadium^{w}, stadium^{ws}) \\ &= \frac{|X|}{|X| + 0.45|Y| + 0.55 \cdot |Z|} \\ &= \frac{4}{4 + 0.45 * 5 + 0.55 * 2} = 0.54 \end{aligned} \quad (6.12)$$

Since string matching for feature names is a weak form of similarity assessment, a further analysis should consider the semantics of distinguishing features. For example, these studies could exploit the use of set intersection between features based on a shared domain (i.e., attribute domain) or semantic interrelationships (e.g., entailment of functions).

6.3.3 Semantic-Neighborhood Matching

Semantic-neighborhood matching (S_n) is a recursive process, because comparing entity classes in the semantic neighborhoods is also a similarity evaluation. This recursion stops when the specified radius is reached, at which point entity classes can be compared based on lexicon 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

function of the cardinality (| |) of the semantic neighborhoods (*N*) and the approximate intersection ($\cap_n$) between these semantic neighborhoods (Equation 6.13).

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

with

$$\delta(a^p, a^p \cap_n b^q, r) = \begin{cases} |N(a^p, r)| - a^p \cap_n b^q & \text{if } |N(a^p, r)| > a^p \cap_n b^q \\ 0 & \text{otherwise} \end{cases} \qquad (6.13)$$

The intersection over semantic neighborhoods is approximated by the similarity of entity classes across neighborhoods (Equation 6.14), 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 number of entity classes in the corresponding semantic neighborhoods.

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

$$\varphi = \begin{cases} 1 & \text{if } S(a^p, b^q) = \max_{j \le m} S(a^p, b_j^q) \\ 0 & \text{otherwise} \end{cases} \qquad (6.14)$$

Since *S*() in Equation 6.14 is an asymmetric function, the approximate set-intersection $\cap_n$ is also asymmetric. The approximate set intersection over semantic neighborhoods matches corresponding 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 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 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 6.15a-b). Semantic-neighborhood matching takes each entity class in $N(stadium^w)$ and finds the corresponding most similar entity class in $N(stadium^s)$. Based on lexicon and feature matching, the only similar entity classes in the neighborhoods are $stadium^w$ and $stadium^s$, which are the original entity classes that are compared; therefore, the semantic-neighborhood matching is equal to zero.

$$N(stadium^w, 1) = \{stadium^w, structure^w, athletic\_field^w, sports\_arena^w\} \quad (6.15a)$$

$$N(stadium^s, 1) = \{stadium^s, entity\_type^s\} \quad (6.15b)$$

Analogously to the notion of *shallow* and *deep* equality in object orientation (Khoshafian and Copeland 1986, Zdonik and Maier 1990), 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 in lieu of a syntactic evaluation.

6.4 Cross-Ontology Evaluations

The tests of the MD3 model address two main questions:

- How does the MD3 model perform with ontologies that differ in their specification components?
- How does the MD3 model perform compared to the MD model?

These tests employ the combination of WordNet and SDTS (WS) described in Section 3.3 (257 definitions) and subsets of the original definitions in WordNet (334 definitions) and SDTS (498 definitions) (Table 6.3). Since the ontologies used in these tests vary in terms of domain (i.e., general vs. specific) and specificity (semantic relations vs. distinguishing features), the potential conclusions of these tests can provide a good indication of the performance of the MD3 model under different scenarios.

The derived ontology from SDTS includes all entity types of SDTS as well as the included terms for *boundary*, *building*, *building complex*, *control point*, *road*, *tower*, *utility*, and *watercourse*. Included terms in SDTS can be either synonyms or subclasses of the corresponding entity type; however, the subsequent evaluations consider all included terms as subclasses. This assumption has the effect of increasing the size of the ontology without altering the similarity evaluations, since all included terms have the same definitions as their respective entity types. In order to create a hierarchy, SDTS's entity classes are interconnected through a superclass *anything*, which contains no particular information.

The ontology derived from WordNet is a set of definitions whose terms match with the names in the VMAP Level 0 specification (NIMA 1999). This ontology also includes the intermediate entity classes that create the hierarchical structure. Like SDTS, WordNet does not have a common superclass for all definitions (multiple hierarchies), therefore, a common superclass *anything* is created.

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

Table 6.3: Characteristics of the specification components of SDTS, WordNet, and WS.

6.4.1 Test 1: Evaluations Using Ontologies with Different Specification Components

The performance of the model is studied by using different combinations of ontologies in cross-ontology similarity evaluations (Table 6.4). These combinations correspond to diverse grade 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 (5-6).

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
6	SDTS*-WordNet*	Different ontologies and specification components (inverse evaluation)

Table 6.4: Cases of cross-ontology evaluations. (Symbol * denotes small subsets of the entire ontology.)

Analogously to evaluations for information retrieval (Korfhage 1997), we use 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 6.12a), while precision is the proportion of entity classes detected by the model that are actually similar (Equation 6.12b). In Equations 612a-b, A is the set of similar entity classes, B is the similar entity classes obtained by the model, and $|\ |$ is the counting measure.

$$recall = \frac{|A \cap B|}{|A|} \tag{6.12a}$$

$$precision = \frac{|A \cap B|}{|B|} \tag{6.12b}$$

A critical issue for calculating recall and precision is to know what entity classes are similar. To simplify this determination, we take only the most similar entity classes across ontologies, that is, 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 cases of the test (i.e., WordNet-WordNet and SDTS-SDTS), each entity class in the first ontology has its corresponding entity class in the second ontology. The most similar entity class of an entity in the first hierarchy should be the entity class with the same name in the second ontology. When the definitions in the first ontology are a super set 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 (Section 3.3) we derive what 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* and SDTS-WordNet*). To simplify this task, the test considers a small portion of the two original ontologies (i.e., 240 WordNet definitions and 48 SDTS definitions). Using these subsets of the ontologies, a manual identification of corresponding entity classes found 22 from the total of 48 entity classes in SDTS whose definitions are also included in WordNet.

The test performs multiple evaluations with different combinations of weights for name, feature, and semantic neighborhood matching. They start with single-matching evaluations over each of the specification components. Table 6.5 shows results for the single-matching evaluations in terms of names, features, and semantic neighborhoods. This table presents the best results obtained from the single matching of semantic neighborhood, that is, when similarity among entity classes in the semantic neighborhoods is determined by the matching of entity class names.

Table 6.5 demonstrates that single matching over entity class names tends to have better measures of recall and precision than single matching over features. Obviously, for identical ontologies recall of the lexicon matching is 100%, since corresponding entity classes have the same names. Precision, however, is not necessarily 100% for cases with identical ontologies (i.e., Cases 1 and 2) due to the presence of polysemous terms. A general observation indicates that entity classes in all ontologies overlap; that is, corresponding entity classes have the same name, but not all entity classes with the same name are in fact semantically similar. Overlapping of entity class names is more likely in situations where an ontology handles synonym sets, such as WordNet, since the chance increases for using one of the terms in the synonym set to refer to an entity class.

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.

Case	Weights (%)			Recall	Precision
	Lexicon	Feature	Neighborhood	(%)	(%)
WordNet-WordNet	100	0	0	100	74
WordNet-WordNet	0	100	0	48	10
WordNet-WordNet	0	0	100	100	97
SDTS-SDTS	100	0	0	100	87
SDTS-SDTS	0	100	0	100	2
SDTS-SDTS	0	0	100	100	1
WordNet-WordNet*	100	0	0	100	74
WordNet-WordNet*	0	100	0	62	10
WordNet-WordNet*	0	0	100	100	94
WordNet*-WS	100	0	0	100	78
WordNet*-WS	0	100	0	17	37
WordNet*-WS	0	0	100	29	12
WordNet*-SDTS*	100	0	0	100	42
WordNet*-SDTS*	0	100	0	0	0
WordNet*-SDTS*	0	0	100	27	2
SDTS*-WordNet*	100	0	0	100	38
SDTS*-WordNet*	0	100	0	0	0
SDTS*-WordNet*	0	0	100	32	3

Table 6.5: Recall and precision of single-matching evaluations and threshold equal to 75%. (Symbol * denotes small subsets of the entire ontology.)

Semantic-neighborhood matching is very sensitive to the hierarchical structure underlying the ontology. In general, neighborhood matching produces unsatisfactory results unless the ontologies are similar and they have detailed identification of hyponymy (is-a) relation, such as in WordNet. When features are shared by many entity classes and ontologies have a shallow semantic hierarchy (e.g., SDTS-SDTS), semantic-neighborhood matching is imprecise.

Single-matching evaluations are followed by double-matching evaluations that combine two specification components: name with feature, name with semantic neighborhood, and feature with semantic neighborhood (Table 6.6). Recall and precision in double-matching evaluations are reduced drastically for combinations that ignore lexicon matching. The combination of name and semantic neighborhood matching obtains the best evaluations of recall and precision. As it was expected, the worst results are associated with evaluations over different ontologies (i.e., WordNet*-WS, WordNet*-SDTS*, and SDTS*-WordNet*). In these cases, precision is still over or equal to 75%, but recall is considerably lower (41%-55%). When differences between ontologies increase, the model loses its effectiveness. Differences in the specification components between WordNet and WS (i.e., Case 4) are less than the differences between WordNet and SDTS (i.e., Case 5) or between SDTS and WordNet (Case 6). Hence, precision (over 90%) and recall (over 50%) for WordNet and WS are better than the measures obtained from the cases with different ontologies (i.e., WordNet*-SDTS* and SDTS*-WodNet*).

Case	Weights (%)			Recall	Precision
	Lexicon	Feature	Neighborhood	(%)	(%)
WordNet-WordNet	50	50	0	46	97
WordNet-WordNet	0	50	50	46	14
WordNet-WordNet	50	0	50	100	97
SDTS-SDTS	50	50	0	100	100
SDTS-SDTS	0	50	50	100	2
SDTS-SDTS	50	0	50	100	100
WordNet-WordNet*	50	50	0	59	97
WordNet-WordNet*	0	50	50	28	14
WordNet-WordNet*	50	0	50	99	98
WordNet*-WS	50	50	0	17	100
WordNet*-WS	0	50	50	0	0
WordNet*-WS	50	0	50	55	95
WordNet*-SDTS*	50	50	0	0	0
WordNet*-SDTS*	0	50	50	0	0
WordNet*-SDTS*	50	0	50	50	92
SDTS*-WordNet*	50	50	0	0	0
SDTS*-WordNet*	0	50	50	0	0
SDTS*-WordNet*	50	0	50	41	75

Table 6.6: Recall and precision of double-matching evaluations and threshold equal to 75%. (Symbol * denotes small subsets of the entire ontology.)

Finally, the triple-matching evaluations combine the names, features, and semantic neighborhoods and assign the same importance to each matching process (Table 6.7). They result in lower values of recall and precision than the values obtained by the best of the double-matching processes (i.e., name with semantic neighborhood matching). Tripe-matching evaluations keep a high value for precision, but a low value for recall, with the exception of the SDTS-SDTS combination. A reason is that many of the entity classes at the top levels of the hierarchical structures (i.e., general concepts) do not have features in their descriptions such that the consideration of features in the similarity assessment decreases instead of increases the chances of finding similar entity classes. Although SDTS has features in all its entity-class definitions, many features are shared by entity classes and, therefore, there is no significant distinction among these entity classes. A lower threshold increases the chances of finding similar entity classes, but also increases the chances of selecting dissimilar entity classes. For triple-matching evaluations with a threshold of 50%, the model does not have better statistics than the results of double-matching evaluations with 75%.

An important observation is the asymmetric result of the model. The model gives slightly better results when the direction of the similarity evaluation goes from SDTS to WordNet, that is, from a shallow to a deep ontology. A general conclusion, however, is impossible, since both ontologies differ strongly in their specifications.

Case	Weights (%)			Recall	Precision
	Lexicon	Feature	Neighborhood	(%)	(%)
WordNet-WordNet	34	33	33	44	97
SDTS-SDTS	34	33	33	100	100
WordNet-WordNet*	34	33	33	57	97
WordNet*-WS*	34	33	33	1	100
WordNet*-SDTS*	33	33	33	0	0
SDTS*-WordNet*	34	33	33	0	0

Table 6.7: Recall and precision of tripe-matching evaluations and threshold equal to 75%. (Symbol * denotes small subsets of the entire ontology.)

This test has shown that the results of the MD3 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 MD3 model is a good estimator for similarity. In a realistic scenario with different ontologies, however, the test found that semantic neighborhood and name are more stable specification components than the set of features associated with entity classes. Moreover, features can be shared by many entity classes within an ontology such that the determination of the most similar

entity class becomes more difficult. High recall with low precision is obtained when only lexicon matching is considered. High precision, however, is obtained as lexicon matching is combined with semantic-neighborhood matching. With this combination and with ontologies that have different specification components, the model has better precision than recall; that is, the model detects less of the total of similar entity classes, but the ones it detects are indeed similar (over 75% for precision in the worse case). Although feature matching proved to be a less adequate method for detecting the most similar entity classes across ontologies, this method may still be suitable for determining the similarity of entity classes within an ontology or the similarity of semantically related entity classes across ontologies.

6.4.2 Test 2: MD3 Model vs. MD Model

Taking an ontology that has semantically similar entity classes, an interesting question is to find similar entity classes across the same ontology and to check whether the model detects the same similar entity classes that were identified with the MD model in a single ontology. Since both MD and MD3 models use a matching process over features, differences between the results of these models may indicate how adequate name and semantic neighborhood matching are for a similarity assessment with a unique ontology.

To carry out this analysis, the combined ontology of WordNet and SDTS (WS) of Section 3 was used, because it gives good results for the MD model with respect to the human-subject experiment (Chapter 4). The test evaluates similarity across the same ontology (WS-WS) with a threshold equal to zero to detect all similar entity classes. The first column in Table 6.8 shows the results of the MD model between *stadium* and the rest of entity classes in the same ontology. The second and third

columns present the results of the MD3 model between *stadium* and the entity classes in the second, but identical, ontology.

MD	**MD3** (33,33,33)	**MD3** (50,0,50)
Sports arena (0.74)	Stadium (1.0)	Stadium (0.50)
Athletic field (0.74)	Sports arena (0.57)	Sports arena (0.34)
Ballpark (0.74)	Athletic field (0.46)	
Construction (0.67)	Ballpark (0.32)	
Tennis court (0.61)	Tennis court (0.26)	

Table 6.8: Most similar entity classes to a *stadium* using the MD model and the MD3 model.

The MD3 model was applied with two sets of weights for name, feature, and semantic neighborhood matching. The first evaluation considers the default case of all three types of matching that are considered equally important (i.e., name: 33; feature: 33; and semantic neighborhood: 33), whereas the second evaluation considers the weights of the best double-matching evaluations found in the previous section (i.e., name: 50; feature: 0; semantic neighborhood: 50). Since the first evaluation considers feature matching, the results of the evaluation come close to the results obtained from the MD model. Using the MD3 model, however, the similarity values decrease. The second evaluation gives a subset of the value obtained from the MD model, since semantic neighborhood is unable to detect similarity when entity classes are far apart in the hierarchical structure (e.g., *stadium* and *athletic field*).

In conclusion, the relationship between the results obtained from the MD and MD3 models varies depending on the components of the entity class representations. In

cases when the MD model is a good estimator of semantic similarity (i.e., when features characterize entity classes) the MD3 model gives a set that is equal to or smaller than the set of similar entity classes that are found by the MD model. The MD3 model, however, is useful for cases when features are not well specified or semantic relations are the main components of the entity class representations.

6.5 Summary

This chapter introduced a model to evaluate semantic similarity across autonomous ontologies. The model, called MD3, uses a matching process over name, feature, and semantic neighborhood. Experiments using SDTS, WordNet, and the combination of SDTS and WordNet suggest that the lexicon and semantic neighborhood matchings are a good approach to detecting the most similar entity classes across ontologies. Feature matching, on the other hand, is most useful in detecting similar entity classes within an ontology.

Chapter 7

Conclusions and Future Research Directions

This thesis created and investigated computational models that assess semantic similarity among spatial entity classes. The thesis took a top-down approach for similarity assessment by concentrating on entity classes that represent concepts in the real world rather than data stored in a database. The study explored the cognitive aspects of a similarity assessment and the computational formalization of semantic similarity measures. Such similarity measures contribute to the design of systems that compare and process information on a semantic basis and, therefore, bring information systems close to users' expectations in terms of information and knowledge management.

7.1 Summary of the Thesis

This thesis defined a novel approach for semantic similarity assessment among spatial entity classes. This approach is based on a matching process that, combined with a semantic distance, produces an asymmetric similarity measure of entity classes. In this thesis, an ontology was defined as the set of entity class representations that are composed of distinguishing features and semantic relations. As a first implementation of this approach, the Matching-Distance (MD) model applies to similarity evaluations within a single ontology. The main characteristics of the MD model are:

- asymmetric evaluations of semantic similarity for entity classes that represent different levels of abstraction in the hierarchical structure,
- use of is-a and part-whole relations in the entity class representation,
- treatment of synonymy and polysemy of entity class names,
- weighted contribution of the similarity assessment among distinguishing features, and
- a systematic approach to weight determination in terms of contextual information.

This work also extended the MD model and defined the Triple Matching-Distance (MD3) model for similarity evaluations across autonomous ontologies. The MD3 model assumes that ontologies may differ in the level of formalization and explicitness of their definitions and; therefore, it evaluates similarity depending on the common components of the entity class representations. Thus, three similarity measures are defined: lexicon matching, feature matching, and semantic-neighborhood matching.

A prototype of the MD and MD3 models was created in C++ and used with diverse ontologies, such as WordNet (Miller *et al.* 1990), SDTS (USGS 1998), and an *ad hoc* ontology created from the combination of both WordNet and SDTS. The thesis tested the cognitive plausibility of the MD model with a human-subject experiment and the performance of the MD3 model with evaluations that combine WordNet and SDTS.

7.2 Major Results

The major contribution of this thesis is the definition of the MD model that considers cognitive properties of similarity assessment (i.e., asymmetry and context dependence)

and that matches people's judgments. The model identifies three types of distinguishing features—parts, functions, and attributes—that characterize spatial entity classes and that allow the independent determination of the features' contributions to similarity assessment. The model calculates distinguishing features' weights according to the specification of contextual information by the user's intended operations. This type of contextual information defines an application domain, which can also partially solve word-sense ambiguity.

The use of synonym sets to refer to entity classes proved to be a practical approach for treating different ways to express the same concepts and polysemous terms. As polysemous terms are allowed, ontological hierarchies become simpler, because each concept tends to have a unique superordinate concept. Even more, a distinction among polysemous terms focuses the similarity evaluation on the distinguishing features of entity classes associated with a particular sense rather than comparing entity classes whose distinguishing features relate to more than one sense.

Although contextual information affects similarity evaluations, the major factor for the MD model's performance is the correct representation of entity classes in terms of distinguishing features and semantic relations. Experiments with existing ontologies demonstrated that accurate definitions of distinguishing features are possible at or below Rosch's (1975) basic level of a hierarchical structure. At the top level, however, more abstract concepts, such as *entity* and *organization*, lack the characterization that make the MD model adequate for similarity evaluations.

Overall, experiments suggested that the selection of the approach for weight determination should consider the type of context specification. While the commonality approach seems to work well for specific applications where users seek

particular features of entity classes, the variability approach produces good results for cases when users seek a type of entity classes.

A disadvantage of the MD model is the lack of existing ontologies that use functions in their entity class representations. Functions, behavior, or affordances were suggested to be determinant for the meaning of objects; however, their formalization in existing ontologies is still missing. Likewise, the MD model is constrained to applications with all entity classes semantically interrelated (i.e., a single ontology) and entity classes characterized by distinguishing features.

Another important contribution of this thesis is the definition of the MD3 model for evaluation across ontologies. The MD3 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., names, distinguishing features, and semantic neighborhoods). The MD3 model is useful as a first step in an ontology integration, since it detects the 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, and used as bass to create a single, integrated ontology.

Examples that used the MD3 model with different ontologies indicated that components of entity class representations have varied effects on the similarity evaluations. While names and semantic neighborhoods are good elements for detecting equivalent or most similar entity classes across ontologies, distinguishing features are suitable for detecting entity classes that are just similar, that is, entity classes that are not synonyms and are located far apart in the hierarchical structure (e.g., *stadium* and *athletic field* in the WordNet ontology).

7.3 Future Work

Several new research questions have resulted from this thesis. They involve extensions of both the MD and MD3 models, comparison of the MD and MD3 models with existing models, data modeling, context specification, ontology integration, reasoning about similarity, and similarity assessment among spatial scenes.

7.3.1 Extensions of the MD and MD3 Models

Inheritance is a powerful feature of a semantic network with is-a relation. This feature, however, might cause problems in situations where subclasses represent exceptions and they do not inherit all properties of their superclasses. A typical example is the case of a "penguin" that is linked to a "bird." Since a typical feature of a bird is "to fly," a penguin would inherit this feature as well. This is obviously a mistake. Thus, the ontology of entity classes could be improved by defining a strategy for exception handling (Durkin 1994) or by considering the classification of entities in terms of prototypical characteristics rather than necessary and sufficient conditions (Mark *et al.* 1999, Rosch 1973, Rosch and Mervis 1975).

The thesis has concentrated on entity classes and has compared distinguishing features in terms of a basic 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* (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.

The extension of the models by considering instances of entity classes is another area for future research. Instances have attributes with associated values. As a first approach, values could be compared with a syntactic approach (i.e., string matching), but a serious effort should compare values depending on the type (e.g., numerical values, range values, and nominal values) and domain.

This study has suggested an entity class representation with components consisting of semantic relations and distinguishing features. Although these components seem to be adequate for a large number of entity class definitions, they may be insufficient to capture the semantics of some entity classes. For example, a historical building is a building whose age is greater than a specific value. This type of semantics is well represented by axioms, which are not incorporated in the MD and MD3 models. If axioms are included into the entity class representations, the model must compare them and infer a similarity value among them.

This work has produced a prototype of the MD and MD3 models that uses an object-oriented representation in C++. Future work may consider an implementation of the semantic similarity model that uses a formalism for expressing structured and sharable knowledge, such as description logic or terminological logic (Brachman and Schmolze 1985). Description logic gives a logical basis for frames, semantic networks, and object-oriented representations as well as for semantic models. It can automatically classify definitions with subsumption inferences.

7.3.2 The MD and MD3 Models vs. Existing Models

The study highlighted main differences between the MD model and existing models that are based on semantic distance and information content. Thus, we have claimed that the MD model is a good estimator of the semantic similarity among entity classes

located at the medium and low levels of abstraction in the hierarchical structure. At these levels, the model uses features to distinguish entity classes that belong to the same sub-hierarchy and have the same common superclass. A further study, however, should examine whether or not the performance of the MD model under the same set of evaluations is better than the performance of existing models. Such as study could lead to the conclusion that the different approaches provide complementary answers and that no single model, but multiple approaches to semantic similarity should be considered depending on the semantic organization of entity classes.

The comparison of the MD3 model with current models of similarity across ontologies is another area for further research. An interesting methodology for comparing these models is to calculate the correlation between the models' results for the entity classes that are similar to a user's request and the probability that instances of these entity classes will satisfy the request under a logical interpretation (Weinstein and Birmingham 1999). A major difficulty for comparing different models is that while most of the current models require an integrated or shared ontology, the MD3 model uses unconnected ontologies. Although current models of similarity assessment could have a slightly better performance than the MD3 model, the MD3 model would still be advantageous, because it does not require a pre-processing for creating the integrated or shared ontology.

7.3.3 Ontology vs. Database Schema

The motivation of this thesis is the enhancement of geographic information systems for information retrieval and integration. In order to make use of the similarity models in information systems, entity classes (i.e., concepts in the real world) should be linked to entity classes modeled in those systems.

Ontologies and database schemas are related, but not equivalent. Ontologies have explicit representations of the entity classes' semantics, whereas database schemas usually use implicit semantics and describe entity classes in terms of attributes and operations required by a specific application. Thus, semantic similarity evaluations are better obtained by comparing definitions in an ontology rather than in a database schema. Entity classes in a database could be associated with their corresponding ontological definitions through a semantic directory. The creation and maintenance of these directories are areas for further research as well as solving schematic conflicts that are product of different levels of abstraction in the entity class representations. For example, a conflict occurs when a database schema handles an attribute *type* to distinguish among subclasses, which are explicitly represented in the ontology as entity classes.

7.3.4 Context Specification

Context was specified as the user's intended operations. This type of specification may be extended by considering additional features of entity classes. For example, a user may want to search for sports facilities that have spectator stands. Although context is still determined by an intended operation, parts and attributes may also describe the desired domain of entity classes.

An interesting area of research is the inferences derived from the combination of contexts. Context could be seen as abstract objects and used as any other object (McCarthy 1987). Then a relation between contexts is *specializes* (*c1*, *c2*), which means that "context *c2* involves no more assumption than context *c1* and every proposition meaningful in *c1* is translated into one meaningful in *c2*" (McCarthy 1993). This type of relation is particularly useful for defining *lifting rules* that relate the

propositions and terms in subcontexts to possibly more general propositions and terms in their outer contexts.

7.3.5 Ontology Integration

The determination of similar entity classes across ontologies could be used as a first step for ontology integration. A relation of similarity may overlap with is-a, synonymy, and part-whole relations: for instance, a *hospital* is a *building* and a *hospital* is semantically similar to a *building*. A further analysis of the commonality among similar entity classes could contribute to identify whether these entity classes are also related by synonymy, is-a, or part-whole relations. In this way, a stronger type of ontology integration could be achieved.

7.3.6 Reasoning about Similarity

Reasoning about similarity involves a process in which inferences about the similarity relations among entity classes are determined by using a subset of known similarity relations. These types of inferences are very useful, since they may reduce the process of comparing entity classes; moreover, they may be indispensable for comparing entity classes when no complete information exists about them. Similarity assessment, however, is a subjective judgment that follows no strict logical properties, such as transitivity, symmetry, and reflexivity, defined mathematically. As such, it is very difficult to compose measures of semantic similarity to derive new similarity values.

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.

7.3.7 Similarity Among Spatial Scenes

Geographic information systems deal with geographic scenes, which are described by spatial and non-spatial properties. A next study should consider the similarity evaluation among spatial scenes. This similarity evaluation could be based on the combination of the semantic similarity model with similarity models for geometric characteristics, such as those related to topological relations and cardinal directions (Bruns and Egenhofer 1996, Paiva 1998, Papadias *et al.* 1998). This type of similarity assessment requires a strategy to handle scenes with different numbers of elements and the analysis of correspondences among these elements.

Semantic similarity assessment is obtaining much attention by information scientists, because it has an important effect on many areas of information management. We expect that research in this area will contribute to the design of the next generation of information systems that respond adequately to real user needs. The development of technology that is not only useful, but also desired by broad groups of users, is still an open field for research

References

E. Aïmeur and C. Frasson (1995). Eliciting the Learning Context in Co-Operative Tutoring Systems. in: P. Brezillon and S. Abu-Hakima (eds.), *Workshop on Modelling Context in Knowledge Representation and Reasoning*, Paris, France, pp. 1-11, Institute Blaise Pascal.

V. Akman and M. Surav (1996). Steps Toward Formalizing Context. *AI Magazine* 17(3): 55-72.

R. Basili, M. DellaRocca, and M. Pazienza (1997). Contextual Word Sense Tuning and Disambiguation. *Applied Artificial Intelligence* 11(3): 235-262.

V. Bejamins and D. Fensel (1998). The Ontology Engineering Initiative $(KA)^2$. in: N. Guarino (ed.), *Formal Ontology in Information Systems*, Trento, Italy, pp. 287-301, IOS Press, Amsterdam, The Netherlands.

B. Bergamaschi, S. Castano, S. De Capitani di Vermercati, S. Montanari, and M. Vicini (1998). An Intelligent Approach to Information Integration. in: N. Guarino (ed.), *First International Conference on Formal Ontology in Information Systems*, Trento, Italy, pp. 253-268, IOS Press, Amsterdam, The Netherlands.

A. del Bimbo, E. Vicario, and D. Zingoni (1994). A Spatial Logic for Symbolic Description of Image Content. *Journal of Visual Language and Computing* 5: 267-286.

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

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

A. Blaser, M. Sester, and M. Egenhofer (in press). Visualization in an Early Stage of the Problem Solving Process in GIS. *Computers and Geosciences*.

W. Borst, J. Akkermans, and J. Top (1997). Engineering Ontologies. *International Journal of Human-Computer Studies. Special Issue on Using Explicit Ontologies in KBS Development* 46: 365-406.

A. Bouguettaya, B. Benatallah, and A. Elmagarmid (1998). *Interconnecting Heterogeneous Information Systems*. Kluwer Academic Press, Norwell, MA.

R. Brachman and J. Schmolze (1985). An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science* 9: 179-216.

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

T. Bruns and M. Egenhofer (1996). Similarity of Spatial Scenes. in: M. Kraak and M. Molenaar (eds.), *Seventh International Symposium on Spatial Data Handling (SDH '96)*, Delft, The Netherlands, pp. 4A.31-42.

T. Bruns and M. Egenhofer (1997). User Interfaces for Map Algebra. *Journal of the Urban and Regional Information Associations* 9(1): 44-54.

J. Burg and R. van de Riet (1998). COLOR-X: Using Knowledge fromWordNet for Conceptual Modeling. in: C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

L. Cardelli (1984). A Semantic of Multiple Inheritance. in: G. Kahn, D. McQueen, and G. Plotkin (eds.), *Semantics of Data Type*. Lecture Notes in Computer Science 173, pp. 51-67, Springer-Verlag, New York.

S. Ceri and J. Widom (1993). Managing Semantic Heterogeneity with Production Rules and Persistent Queues. in: *19th Very Large Database Conference VLDB,* Dublin, Ireland, pp. 108-119.

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

A. Collins and M. Quillian (1969). Retrieval Time From Semantic Memory. *Journal of Verbal Learning and Verbal Behavior* 8: 240-247.

D. Cruse (1979). On The Transitivity of the Part-Whole Relation. *Linguistics* 15: 29-38.

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

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

K. Dittrich (1986). Object-Oriented Data Base Systems: The Notation and The Issues. in: K. Dittrich, U. Dayal, and A. Buchmann (eds.), *International Workshop in Object-Oriented Database Systems*, Pacific Grove, CA, Washington, D.C., pp. 2-4, IEEE Computer Society Press.

M. Dojat and F. Pachet (1995). Three Compatible Mechanisms for Representing Medical Context Implicitly. in: P. Brezillon and S. Abu-Hakima (eds.), *Workshop on Modelling Context in Knowledge Representation and Reasoning*, Paris, France, pp. 69-77, Institute Blaise Pascal.

J. Durkin (1994). *Expert Systems: Design and Development*. Prentice-Hall, Englewoood Cliffs, NJ.

M. Egenhofer (1994). Spatial SQL: A Query and Presentation Language. *IEEE Transactions on Knowledge and Data Engineering* 6(1): 86-95.

M. Egenhofer (1997). Query Processing in Spatial-Query-By-Sketch. *Journal of Visual Languages and Computing* 8(4): 403-424.

M. Egenhofer (1999). Introduction, Theory and Concepts. in: M. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman (eds.), *Interoperating Geographic Information Systems*. pp. 1-4, Kluwer Academic Publishers, Norwell, MA.

M. Egenhofer and A. Frank (1992). Object-Oriented Modeling for GIS. *Journal of Urban and Regional Information System Association* 4(2): 3-19.

M. Egenhofer and R. Goyal (in press). Cardinal Directions Between Extended Spatial Objects. *IEEE Transactions on Knowledge and Data Engineering*.

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

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

A. Elmagarmid, M. Rusinkiewicz, and A. Sheth (1999). *Management of Heterogeneous and Autonomous Database Systems*. Morgan Kaufmann, San Mateo, CA.

R. Fagin (1999). Combining Fuzzy Information from Multiple Systems. *Journal of Computer and Systems Sciences* 58: 83-99.

C. Faloustsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petrovic, and W. Equitz (1994). Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems* 3: 231-262.

P. Fankhauser and E. Neuhold (1993). 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*, Victoria, Australia, pp. 155-175, Elsevier Science Publishers, North-Holland.

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

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

P. Fisher and J. Wood (1998). What Is a Mountain? *Geography* 83(3): 247-256.

D. Flewelling (1997). *Comparing Subsets From Digital Spatial Archives: Point Set Similarity*. Ph.D. Thesis, University of Maine, Orono, ME.

M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petrovik, D. Steele, and P. Yanker (1995). Query by Image and Video Content: The QBIC System. *IEEE Computer* 28(9): 23-32.

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

A. Gangemi, D. Pisanelli, and G. Steve (1998). Ontology Integration: Experiences with Medical Terminologies. in: N. Guarino (ed.), *Formal Ontology in Information Systems,* Trento, Italy, pp. 163-178, IOS Press, Amsterdam, The Netherlands.

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

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

A. Ginsberg (1993). A Unified Approach to Automatic Indexing and Information Retrieval. *IEEE Expert* 8(5): 46-56.

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

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

A. Goñi, E. Mena, and A. Illarramendi (1997). Querying Heterogeneous Data Repositories Using Ontologies. in: P.-J. Charrel and H. Jaakkola (eds.), *Conference on Information Modelling and Knowledge Bases IX*, pp. 19-34, IOS Press, Amsterdam, The Netherlands.

T. Gruber (1992). *Ontolingua: A Mechanism to Support Portable Ontologies*. Knwoledge Systems Laboratory, Stanford University, Stanford, CA, Technical Report KSL 91-66.

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

T. Gruber (1995b). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 6: 199-220.

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

N. Guarino (1997). Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. in: M. Pazienza (ed.), *Information Extraction: A Multidisciplinary Approach to an Engineering Information Technology*, Francasi, Italy, pp. 139-170, Springer Verlag.

N. Guarino (1998). Formal Ontology in Information Systems. in: N. Guarino (ed.), *Formal Ontology in Information Systems*, Trento, Italy, pp. 3-15, IOS Press, Amsterdam, The Netherlands.

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

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

J. Hammer, D. McLeod, and A. Soli (1994). An Intelligent System for Identifying and Integrating Non-Local Objects in Federated Database Systems. in: *27th International Conference on System Sciences*, Honolulu, HI, pp. 389-407, Computer Society of IEEE.

P. Hayes (1990). Naive Physics I: Ontology for Liquids. in: D. Weld and J. de Kleer (eds.), *Reading in Qualitative Reasoning about Physical Systems*. pp. 484-502, Morgan Kaufmann, San Mateo, CA.

M. Hearst (1994). *Context and Structure in Automated Full-Text Information Access*. Ph.D. Thesis, Computer Science Division, University of California at Berkeley, Berkeley, CA.

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

S. Hirtle and J. Jonides (1985). Evidence of Hierarchies in Cognitive Maps. *Memory and Cognition* 13(3): 208-217.

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

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

V. Kashyap and A. Sheth (1996). Schematic and Semantic Similarities between Database Objects: A Context-based Approach. *The Very Large Database Journal* 5(4): 276-304.

V. Kashyap and A. Sheth (1998). 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*. pp. 139-178, Academic Press, London, UK.

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

S. Khoshafian and G. Copeland (1986). Object Identity. in: *OOPSLA*, Portland, OR, pp. 406-416.

W. Kim and J. Seo (1991). Classifying Schematic and Data Hetergeneity in Multidatabase Systems. *IEEE Computer* 24: 12-18.

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

K. Knight and S. Luk (1994). Building a Large-Scale Knowledge Base for Machine Translation. in: *National Conference on Artifical Intelligence AAAI-94*, Seattle, WA, pp. 773-778.

R. Korfhage (1997). *Information Storage and Retrieval.* John Wiley and Sons, New York.

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

W. Kuhn (1994). Defining Semantics for Spatial Data Transfers. in: T. Waugh and R. Healey (eds.), *Sixth International Symposium on Spatial Data Handling*, Edinburgh, Scotland, pp. 973-987, International Geographical Union.

G. Lakoff (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press, Chicago, IL.

E. Lang (1991). The LILOG Ontology From a Linguistic Point of View. in: O. Herzo and C. Rollinger (eds.), *Text Understanding in LILOG.* pp. 464-481, Springer-Verlag, Berlin, Germany.

J. Larson, S. Navathe, and S. Elmasri (1989). A Theory of Attribute Equivalence in Database with Application to Schema Integration. *IEEE Transactions on Software Engineering* 15(4): 449-463.

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

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

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

G. Leech (1981). *Semantics: The Study of Meaning*. Penguin, Harmondsworth, U.K.

D. Lenat and R. Guha (1990). *Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley Publishing Company, Reading, MA.

D. Lenat, R. Guha, K. Puttman, D. Pratt, and M. Shepherd (1990). Cyc: Towards Programs with Common Sense. *Communications of the ACM* 33(8): 30-49.

D. Lenat, G. Miller, and T. Yokoi (1995). Cyc, WordNet, and EDR: Critiques and Responses. *Communications of the ACM* 38(11): 45-48.

W. Litwin (1994). *Multidatabse Systems*. Prentice Hall, Englewood Cliffs, NJ.

K. Mahesh (1996). *Ontology Development for Machine Translation: Ideology and Methodlogy*. Computing Research Laboratory, New Mexico State University, Las Cruces, NM, Technical Report MCCS 96-292.

K. Mahesh and S. Nirenburg (1995). A Situated Ontology for Practical NLP. in: *Workshop on Basic Ontological Issues in Knowledge Sharing. International Joint Conference on Artifical Intelligence (IJCAI-95),* Montreal, Canada.

D. Mark (1989). *Cognitive and Linguistic Aspects of Geographic Space*. National Center for Geographic Information and Analysis, Technical Report 1.

D. Mark and M. Gould (1991). Interaction with Geographic Information: A Commentary. *Photogrammetric Engineering & Remote Sensing* 57(11): 1427-1430.

D. Mark, B. Smith, and B. Tversky (1999). Ontology and Geographic Objects: An Empirical Study of Cognitive Category. in: C. Freksa and D. Mark (eds.), *Spatial Information Theory-Cognitive and Computational Foundations of Geographic Information Science, COSIT '99,* Stade, Germany. Lecture Notes in Computer Science V. 1661, pp. 283-298, Springer-Verlag, Berlin, Germany.

J. McCarthy (1987). Generality in Artificial Intelligence. *Communications of the ACM* 30(12): 1030-1035.

J. McCarthy (1993). Notes on Formalizing Context. in: *13ht International Joint Conference on Artificial Intelligence,* Menlo Park, CA, pp. 555-560.

D. McGuinness (1998). Ontological Issues for Knowledge-Enhanced Search. in: N. Guarino (ed.), *Formal Ontology in Information System.* pp. 302-316, IOS Press, Amsterdam, The Netherlands.

C. Meadow, B. Boyce, and D. Kraft (2000). *Text Information Retrieval Systems*. Academic Press, San Diego, CA.

R. Meersman (1995). An Essay on the Role and Evolution of Data(base) Semantics. in: R. Meersman and L. Mark (eds.), *Proceedings of the 6th IFIP Working Conference on Data Semantics*, London, UK, pp. 1-7, Chapman Hall.

E. Mena, V. Kashyap, A. Illarramendi, and A. Sheth (1998). Domain Specific Ontologies for Semantic Information Brokering on the Global Information Infrastructure. in: N. Guarino (ed.), *Formal Ontology Information Systems*, Trento, Italy, pp. 269-286, IOS Press, Amsterdam, The Netherlands.

E. Mena, V. Kashyap, and A. Sheth (1996). 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)*, Brussel, Belgium, pp. 14-25, IEEE Computer Society Press.

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

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

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

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

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

K. Milligan (1992). *Language, Truth and Ontology*. Kluwer, Dordrecht, The Netherlands.

I. Monarch and J. Carbonelli (1987). CoalSORT: A Knowledge-Base Interface. *IEEE Expert* Spring 1997: 39-53.

NIMA (1999). Geospatial Standards and Specifications, 10/14/99, http://www.nima.mil/publications/specs/.

NLM (1997). *UMLS Knowledge Sources*. National Library of Medicine, Bathesda, Maryland.

A. Ouksel and C. Naiman (1994). Coordinating Context Building in Heterogeneous Information Systems. *Journal of Intelligent Information Systems* 3(1): 151-183.

J. Paiva (1998). *Topological Equivalence and Similarity in Multiple Representation Geographic Database*. Ph.D. Thesis, University of Maine, Orono, Maine.

D. Papadias, D. Arkoumanis, and N. Karacapilidis (1998). On the Retrieval of Similar Configurations. in: T. Poiker and N. Chrisman (eds.), *8th International Symposium on Spatial Data Handling*, Vancouver, pp. 510-521, International Geographical Union.

Y. Park and F. Golshani (1997). ImageRoadMap: A New Content-Based Image Retrieval System. in: A. Haneurlaim and A. Tjoa (eds.), *Database and Expert Systems Applications DEXA'97*, Toulouse, France, pp. 225-239, Springer Verlag.

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

A. Rector, W. Nowlan, and A. Glowinski (1993). Goals for Concept Representation in the GALEN Project. in: *17th Annual Symposium on Computer Applications in Madical Care SCAMC'93*, Washington, pp. 414-418,

O. Resnik (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* 11: 95-130.

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

R. Richardson and A. Smeaton (1996). *An Information Retrieval Approach to Locating Information in Large Scale Federated Database Systems*. Dublin City University, School of Computer Applications, Dublin, Ireland, Technical Report CA-0296.

R. Richardson, A. Smeaton, and J. Murphy (1994). *Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words.* Dublin City University, School of Computer Applications, Dublin, Ireland, Technical Report CA-1294.

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

E. Rosch (1973). On the Internal Structure of Perceptual and Semantic Categories. in: T. Moore (ed.), *Cognitive Development and the Acquisition of Language*. Academic Press, New York.

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

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

V. Schenkelaars and M. Egenhofer (1997). Exploratory Access to Geographic Libraries. in: *ACSM/ASPRS Auto-Carto 13,* Seattle, WAASPRS and ACSM.

C. Schlenoff, A. Knutilla, and S. Ray (1998). A Robust Ontolgy for Manufacturing Systems Integration. in: *2nd International Conference on Engineering Design and Autonation*, Mai, Hawaii.

E. Sciore, M. Siegel, and A. Rosenthal (1994). Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information System. *ACM Transactions on Database System* 19(2): 254-290.

A. Sheth (1995). Data Semantics: What, Where, and How? in: R. Meersman and L. Mark (eds.), *Database Application Semantics*. pp. 601-610, Chapman and Hall.

A. Sheth (1999). 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*. pp. 5-30, Kluwer Academic Publishers, Norwell, MA.

A. Sheth and V. Kashyap (1992). 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*, North-Holland, pp. 283-312, Lorne, Victoria, Australia.

A. Sheth and J. Larson (1990). Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys* 22(3): 183-236.

Y. Shoham (1991). Varieties of Context. in: V. Lifschitz (ed.), *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*. pp. 393-407, Academic Press, San Diego, CA.

A. Silberschatz, H. Korth, and S. Sudarshan (1996). *Database System Concepts*. McGraw-Hill, Burr Ridge, IL.

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

B. Smith and K. Mulligan (1983). Framework for Formal Ontology. *Topoi* 2: 73-85.

E. Smith and D. Medin (1981). *Categories and Concepts*. Harvard University Press, Cambridge, MA.

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

S. Spaccapietra and C. Parent (1994). View Integration: A Step Forward in Solving Structural Conflicts. *IEEE Transactions on Knowledge and Data Engineering* 6(2): 258-270.

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

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

R. Turner (1998). Context-Mediated Behavior for Intelligent Agents. *International Journal of Human-Computer Studies* 48: 307-330.

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

D. Unwin (1981). *Introductory Spatial Analysis*. Methuen, New York.

M. Uschold, M. Healy, K. Williamson, P. Clark, and S. Woods (1998). Ontology Reuse and Application. in: N. Guarino (ed.), *Formal Ontology in Information Systems*, Trento, Italy, pp. 163-178, IOS Press, Amsterdam, The Netherlands.

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

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

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

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

C. Welty (1998). The Ontological Nature of Subject Taxonomies. in: N. Guarino (ed.), *Formal Ontology in Information Systems*. pp. 317-327, IOS Press, Amsterdam, the Netherlands.

G. Wiederhold (1994). Interoperation, Mediator, and Ontologies. in: *International Symposium on Fifth Generation Computer Systems FGCS95: Workshop on Heterogeneous Cooperative Knowledge-Bases*, Tokyo, Japan, pp. 33-48.

G. Wiederhold and J. Jannink (in press). Composing Diverse Ontologies. in: *8th Working Conference on Database Semantics (DS-8)*, Rotorua, New ZealandIFIP/Kluwer/Chapman & Hall.

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

L. Zadeh (1965). Fuzzy Sets. *Information and Control* 8: 338-353.

Appendix

Survey

This is not a test that has right or wrong answers. We are studying how people judge similar things and how their views change under different contexts. This survey has 5 parts. Each part describes a situation with a set of places. The list of places and their definitions are also given. Then, you will be asked to rank places according to your judgment of similarity. Start with 1 for the most similar place and assign the same rank for places that you consider equally similar.

The whole test should take less than 20 minutes.

Your completion of this task is voluntary, and you may skip any or all parts you choose to. Your responses will remain anonymous, please do not write your name anywhere on this form.

Please read the description at the top of each page and use the definitions of places that are given. Fill in your evaluation, and then turn the page. Please do not go back.

General Information

Age: ___ years Gender: ____ Female ____Male

Place of birth: _____ Place of residence: _____

Native (first) language spoken: _____

Definitions

1. **Stadium**: large often unroofed structure in which athletic events are held.
2. **Sports arena**: building where games, contest, and other exertions are performed.
3. **Athletic field**: open area where sports events, exercise, or games occur.
4. **Theater**: building for the presentation of plays, motion pictures, or other dramatic performances or spectacles.
5. **Museum**: a depository for collecting and displaying objects having scientific or historical or artistic value.
6. **Ball park**: a facility in which ball games are played (especially baseball games).
7. **Tennis court**: a specially marked area within which tennis is played.
8. **Transportation system**: the roads and equipment necessary for the movement of passengers or goods.
9. **Library**: a facility built to contain books and other materials for reading and study.
10. **Commons** a piece of open land for recreational use in an urban area.

11. **Building**: permanent walled and roofed construction.

12. **House**: building in which something is sheltered or located.

13. **Path**: open way for the passage of persons or animals on land.

14. **Road** open way (generally public) for the passage of vehicles on land.

15. **Port**: landing place provided with terminal and transfer facility for loading and discharging cargo or passengers.

16. **Bridge**: structure erected over a depression or obstacle to carry traffic or some facility such as a pipeline.

17. **Railway**: permanent way having one or more rails which provides a track for cars.

18. **Airport** facility, either on land or water, where aircraft can take off and land.

19. **Terminal** where transport vehicles load or unload passengers or goods.

20. **Subway station**: terminal where subways load and unload passengers.

21. **Highway** major road for any form of motor transport.

22. **Travelway** open way for the passage of vehicles, persons, or animals on land.

23. **Lake**: body of (usually fresh) water surrounded by land.

24. **Forest**: land that is covered with trees and shrubs.

25. **City** large and densely populated urban area

26. **Desert** arid region with little or no vegetation.

27. **River**: large natural stream of water (larger than a creek).

28. **Beach** area of sand sloping down to the water of a sea or lake.

29. **Lagoon**: body of water cut off from a larger body by a reef of sand or coral.

30. **Island**: land mass (smaller than a continent) that is surrounded by water.

31. **Wetland**: low area where the land is saturated with water.

32. **Pond**: small lake.

33. **Mountain**: land mass that projects well above its surroundings; higher than a hill.

Part A

How similar is a stadium to the following places (1: the most similar)?

1. [ ] A sports arena
2. [ ] An athletic field
3. [ ] A theater
4. [ ] A museum
5. [ ] A ball park
6. [ ] A tennis court
7. [ ] A transportation system
8. [ ] A library
9. [ ] A commons
10. [ ] A building
11. [ ] A house

Part B

How similar is a stadium to the following places if you are searching for a place to play a sport (1: the most similar)?

1. [ ] A building
2. [ ] A ball park
3. [ ] A theater
4. [ ] A museum
5. [ ] A house
6. [ ] A sports arena
7. [ ] A transportation system
8. [ ] A commons
9. [ ] An athletic field
10. [ ] A library
11. [ ] A tennis court

Part C

How similar is a stadium to the following places if you are comparing constructions (1: the most similar)?

1. [ ] A commons
2. [ ] A transportation system
3. [ ] A tennis court
4. [ ] A building
5. [ ] A library
6. [ ] A sports arena
7. [ ] A ball park
8. [ ] A museum
9. [ ] A house
10. [ ] An athletic field
11. [ ] A theater

Part D

How similar is a travelway to each of these other transportation-type entities (1: the most similar)?

1. [ ] A road
2. [ ] A port
3. [ ] A bridge
4. [ ] A railway
5. [ ] A transportation system
6. [ ] An airport
7. [ ] A terminal
8. [ ] A subway station
9. [ ] A highway
10. [ ] A path

Part E

How similar is a lake to these other entities (1: the most similar)?

1. [ ] A forest

2. [ ] A city

3. [ ] A desert

4. [ ] A river

5. [ ] A beach

6. [ ] A lagoon

7. [ ] An island

8. [ ] A wetland

9. [ ] A bridge

10. [ ] A pond

11. [ ] A mountain

Thanks for your cooperation.

Biography

M. Andrea Rodríguez was born in Concepción, Chile, on February 23, 1965. She graduated with high distinction from the Universidad de Concepción in 1987 with a degree in Computer Science and in 1989 with the title of Information Engineer. Since her graduation, Andrea has worked as a teaching assistant at the Department of Computer Science and as an information engineer at the Center for Research and Education of Environmental Science, Centro Eula-Chile, Universidad de Concepción. In 1995 she was awarded a Fulbright scholarship by the US government and a leave of absence from the Universidad de Concepción to pursue a Master's degree in Spatial Information Science and Engineering at the University of Maine. In 1997 she obtained her Master of Science degree with a thesis entitled “Image Scemata-Based Inferences: the Container-Surface Algebra for Solid Objects.” After her Master's degree she began her doctoral program and served as a Graduate Research Assistant in the Department of Spatial Information Science and Engineering.

Andrea Rodríguez is a candidate for the Doctor of Philosophy degree in Spatial Information Science and Engineering from the University of Maine in May, 2000.