

8-2006

# Semantic Similarity of Spatial Scenes

Konstantinos A. Nedas

Follow this and additional works at: <http://digitalcommons.library.umaine.edu/etd>



Part of the [Databases and Information Systems Commons](#), and the [Geographic Information Sciences Commons](#)

---

## Recommended Citation

Nedas, Konstantinos A., "Semantic Similarity of Spatial Scenes" (2006). *Electronic Theses and Dissertations*. 566.  
<http://digitalcommons.library.umaine.edu/etd/566>

This Open-Access Dissertation is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of DigitalCommons@UMaine.

# **SEMANTIC SIMILARITY OF SPATIAL SCENES**

By

Konstantinos A. Nedas

Dipl. Eng. Aristotle University of Thessaloniki, 2000

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

The University of Maine

August, 2006

Advisory Committee:

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

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

Sudarshan S. Chawathe, Assistant Professor of Computer Science

Kathleen Stewart Hornsby, Assistant Research Professor, National Center for  
Geographic Information and Analysis

Michael F. Worboys, Professor of Spatial Information Science and Engineering

© 2006 Konstantinos A. Nedas

All Rights Reserved

## **LIBRARY RIGHTS STATEMENT**

I am presenting this thesis in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for “fair use” copying of this thesis for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Konstantinos A. Nedas

June 2, 2006



# **SEMANTIC SIMILARITY OF SPATIAL SCENES**

By Konstantinos A. Nedas

Thesis Advisor: Dr. Max J. Egenhofer

An Abstract of the Thesis Presented  
in Partial Fulfillment of the Requirements for the  
Degree of Doctor of Philosophy  
(in Spatial Information Science and Engineering)  
August, 2006

The formalization of similarity in spatial information systems can unleash their functionality and contribute technology not only useful, but also desirable by broad groups of users. As a paradigm for information retrieval, similarity supersedes tedious querying techniques and unveils novel ways for user-system interaction by naturally supporting modalities such as speech and sketching. As a tool within the scope of a broader objective, it can facilitate such diverse tasks as data integration, landmark determination, and prediction making.

This potential motivated the development of several similarity models within the geospatial and computer science communities. Despite the merit of these studies, their cognitive plausibility can be limited due to neglect of well-established psychological principles about properties and behaviors of similarity. Moreover, such approaches are typically guided by experience, intuition, and observation, thereby often relying on more narrow perspectives or restrictive assumptions that produce inflexible and incompatible measures.

This thesis consolidates such fragmentary efforts and integrates them along with novel formalisms into a scalable, comprehensive, and cognitively-sensitive framework for similarity queries in spatial information systems. Three conceptually different similarity queries at the levels of attributes, objects, and scenes are distinguished. An analysis of the relationship between similarity and change provides a unifying basis for the approach and a theoretical foundation for measures satisfying important similarity properties such as asymmetry and context dependence. The classification of attributes into categories with common structural and cognitive characteristics drives the implementation of a small core of generic functions, able to perform any type of attribute value assessment. Appropriate techniques combine such atomic assessments to compute similarities at the object level and to handle more complex inquiries with multiple constraints. These techniques, along with a solid graph-theoretical methodology adapted to the particularities of the geospatial domain, provide the foundation for reasoning about scene similarity queries.

Provisions are made so that all methods comply with major psychological findings about people's perceptions of similarity. An experimental evaluation supplies the main result of this thesis, which separates psychological findings with a major impact on the results from those that can be safely incorporated into the framework through computationally simpler alternatives.

*In memory of my beloved father, Apostolos K. Nedas*

*He was the man I wish to be*

## ACKNOWLEDGMENTS

The completion of this thesis comes only a few months after the loss of my father. He was in many ways the greatest teacher I ever had. No amount of words can express my gratitude to him, or convey what he meant to me. He deserves the greatest thank you for many reasons, but primarily for two things. First, for believing in me, more than anyone ever did and anyone ever will. Second, for his inspiring example of perseverance and courage through difficult times, the only source of strength and determination that kept me going when I felt that there was absolutely nothing and no one to hold on to. I also thank my mother, Haido, for her unlimited love and support, and my younger brother, Emmanuel, for undertaking many pressing responsibilities so that I could finish this task.

On the academic side, my special thanks go to my advisor, and mentor, Max Egenhofer. I thank him for his positive support and encouragement throughout this effort, for his trust and patience, and for his guidance and advice that were available, whenever I needed them. He gave me plenty of room to pursue my own interests, but always made sure I did not go astray. I will never forget the sparking discussions and how motivated for research I felt after our meetings. Many thanks go also to the other members of my committee: Kathleen Hornsby for her valuable scientific input and her constant encouragement, Mike Worboys for always answering my requests on short notice and for his keen observations, Kate Beard for her fruitful suggestions and for always being eager to help, Sudarshan Chawathe for his insightful comments, and Stephan Winter, the external examiner from the University of Melbourne, Australia, for the thorough review he provided on a very tight schedule. I would also like to express my appreciation to two of my professors back home: Apostolos Arvanitis, who directed my attention to UMaine, and Konstantinos Katsambalos, who strongly encouraged me to take the next step in academia. Karen Kidder and Blane Shaw deserve my gratitude for helping me solve all my bureaucratic problems and making sure that I did not have much to worry about.

My clique of friends in Orono might have not been one of maximum cardinality, but it certainly was one of maximum weight. Special thanks to my closest buddy and roommate David for all the fun moments, for extending my cooking skills beyond French fries, and mostly, for the unforgettable and fiery Friday night discussions about anything, from politics and religion to the discovery of the lost intersection, always, of course, accompanied with an overwhelming dose of gin and tonic. A great word of thanks goes to Cecilia for being an exceptional and inspiring friend and for still giving me trouble finding an issue that we could possibly disagree upon, to Brenna for her ever-cheerful presence and for lending her voice to my music, and to Sotiris for having an answer to all my questions and helping me cope with life in Orono. I am also greatly indebted to my wonderful friend Maria, for making sure everything went smooth during my sudden and prolonged absence and for being so compassionate and caring throughout some of my darkest times. Thank you to all my other friends, Giorgos, Marcus, Arda, Sibel, Metin, Tanya, Dominik, Harris, Katie, Chris, and also to Peggy for the delicious Easter dinners and to Tony for the endless supply of new sounds. It was exciting to be part of such a multicultural and international company of people that not only made my staying here enjoyable, but also helped me escape my narrow view and understanding of this world.

From my friends in Greece I would like to thank especially Kostas and Yiannis for their cordial and exemplary friendship from my early childhood until today and for always staying in touch, but also Harris, Petros, Michalis, Vassiliki, Emilia and Panayotis. My aunt Eleftheria, my uncle Fr. Demetrios, and my cousins Tassos, Katerina, and Manolis deserve my gratitude for providing for me a home away from home. I will never forget their generous hospitality, love, and help during my years in the USA. A heartfelt thank you goes to Fani, Maria, Magda, Christos, Dinos, and Fr. Anastasios for staying close to me and my family through the struggles and hardships of the last months. Last but not least, a special thank you to Dina for her love, support, and patience all those years.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS .....	iv
LIST OF TABLES .....	xii
LIST OF FIGURES .....	xiii
CHAPTER 1 INTRODUCTION .....	1
1.1 Terminology.....	1
1.2 Information Retrieval in Geographic Information Systems .....	3
1.3 A Framework for Similarity-Enhanced Retrieval in Spatial Information Systems .....	5
1.3.1 Concept .....	6
1.3.2 Motivation.....	8
1.3.3 Goal.....	11
1.3.4 Hypothesis.....	11
1.3.5 Research Questions.....	16
1.4 Approach.....	17
1.5 Scope.....	20
1.6 Intended Audience .....	22
1.7 Organization of the Thesis .....	22
CHAPTER 2 SEMANTIC SIMILARITY IN INFORMATION SYSTEMS.....	25
2.1 Ontologies .....	25
2.1.1 Defining an Ontology .....	26
2.1.2 Common Misconceptions about Ontologies.....	28
2.1.3 Ontology Types.....	28

2.1.4 WordNet.....	30
2.1.5 Problems of Ontologies.....	31
2.1.6 The Role of Ontologies in Information Systems .....	32
2.2 Modeling Similarity .....	33
2.2.1 Properties of Similarity .....	33
2.2.1.1 Minimality.....	34
2.2.1.2 Symmetry .....	34
2.2.1.3 Transitivity .....	36
2.2.1.4 The Relationship of Similarity to Difference, Dissimilarity, and Distance .....	36
2.2.1.5 Similarity and Context .....	39
2.2.1.6 Similarity in Classification .....	41
2.2.2 Models for Similarity Assessment .....	42
2.2.2.1 Geometric Models.....	42
2.2.2.2 Featural Models .....	44
2.2.2.3 Transformational Models.....	47
2.2.2.4 Models Based on Semantic Networks .....	47
2.2.2.5 Integrated Approaches—The Matching Distance Model .....	50
2.2.2.6 Alignment Models and Configuration Similarity .....	52
2.3 Mathematics for Similarity .....	56
2.3.1 Fuzzy Set Theory .....	57
2.3.2 Graph Theory .....	59
2.4 Summary .....	63

CHAPTER 3 SEMANTIC SIMILARITY AMONG ATOMIC ATTRIBUTE VALUES	64
3.1 Similarity versus Change .....	65
3.2 Similarity Functions.....	72
3.2.1 Specification and Properties .....	72
3.2.2 Mathematical Formalization .....	74
3.2.3 Thresholds and Normalization.....	76
3.3 Classifications of Attributes.....	77
3.4 Similarity Assessment for Ratio Values .....	79
3.4.1 Similarity for Ratio Quantities.....	82
3.4.2 Similarity for Ratio Magnitudes .....	85
3.5 Similarity Assessment for Interval Values .....	86
3.6 Similarity Assessment for Ordinal Values.....	90
3.7 Similarity Assessment for Nominal Values.....	93
3.7.1 Similarity Assessment for Nominal Classifiers .....	94
3.7.2 Similarity Assessment for Boolean Attributes.....	98
3.7.3 Similarity Assessment for Nominal Identifiers.....	99
3.8 Similarity Assessment for Cyclic Values .....	100
3.9 Attribute Considerations beyond the Five Levels of Measurement .....	103
3.10 Null Values in Similarity Assessments.....	105
3.11 Summary .....	110
CHAPTER 4 SEMANTIC SIMILARITY AMONG OBJECTS.....	112
4.1 Queries Expressed through Relational Operators.....	112
4.2 Queries Expressed through Logical Operators .....	114
4.2.1. Queries with <i>AND</i> on Different Attributes .....	114



4.2.1.1 Locally-Better Conjunctive Matching .....	115
4.2.1.2 Globally-Better Conjunctive Matching.....	116
4.2.1.3 Other Approaches to Conjunctive Matching .....	119
4.2.2. Queries with <i>AND</i> on the Same Attribute.....	120
4.2.2.1 Conjunctive Queries on Multi-valued Attributes.....	121
4.2.2.2 Conjunctive Queries on Composite Attributes .....	126
4.2.3 Queries with OR on the Same Attribute .....	128
4.2.4 Queries with OR on Different Attributes.....	129
4.2.5 Queries with NOT .....	130
4.3 Attribute Weights.....	131
4.4 Summary .....	133
CHAPTER 5 SEMANTIC SIMILARITY AMONG SPATIAL SCENES .....	135
5.1 Spatial Scene Queries .....	136
5.1.1 Types of Spatial Scene Queries .....	136
5.1.2 Components of a Spatial Scene Query.....	138
5.1.3 Formulating Spatial Scene Queries.....	139
5.1.4 Representing a Spatial Scene as a Graph.....	140
5.2 Types of Solutions for a Scene Query .....	141
5.3 Types of Retrieval.....	144
5.4 Relaxation .....	147
5.4.1 Relaxation for Spatial Objects .....	147
5.4.2 Relaxation for Spatial Relations .....	148
5.5 Query Execution .....	152
5.5.1 Query Preprocessing .....	153

5.5.2 Creating the Association Graph and Extracting the Maximal Cliques .....	153
5.5.3 Post-Processing of Results .....	154
5.5.3.1 Component Similarity .....	154
5.5.3.2 Scene Completeness.....	156
5.5.3.3 Scene Similarity .....	158
5.5.3.4 Filtering and Presentation .....	159
5.6 An Example of Processing a Spatial Scene Query .....	160
5.7 Summary .....	165
CHAPTER 6 MODEL EVALUATION .....	167
6.1 Measures of Incompatibility .....	167
6.2 Experimental Design.....	170
6.3 Experiments at the Object Level.....	177
6.3.1 Setup .....	177
6.3.2 Results and Discussion .....	181
6.3.2.1 Results of Experiment $E_1$ and Interpretation .....	181
6.3.2.2 Results of Experiments $E_{2A}$ and $E_{2B}$ and Interpretation.....	189
6.3.2.3 Results of Experiment $E_3$ and Interpretation .....	201
6.4 Experiments at the Scene Level.....	207
6.4.1 Setup .....	207
6.4.2 Results of Experiment $E_4$ and Interpretation .....	210
6.5 Summary .....	224
CHAPTER 7 CONCLUSIONS .....	226
7.1 Summary of the Thesis .....	226
7.2 Major Results .....	231

7.3 Future Research .....	236
7.3.1 Similarity Models for Detailed Spatial Relations .....	236
7.3.2 Automated Weight Calibration and Constraint Significance.....	238
7.3.3 Efficient Execution of Similarity Queries.....	240
7.3.4 Extension to Heterogeneous Database Systems .....	242
7.3.5 Formalizing Similarity in Ontologies for the Semantic Web .....	243
7.3.6 Discovering Additional Applications of Similarity .....	245
7.3.7 Evolution of Similarity Models .....	247
REFERENCES .....	248
BIOGRAPHY OF THE AUTHOR.....	286

## LIST OF TABLES

Table 3.1:	A possible similarity matrix for the attribute <i>Topological_Relation</i> (Egenhofer and Al-Taha 1992). ....	73
Table 3.2:	Similarity measures obtained from WordNet with network models versus those obtained from a spatial ontology with the Matching Distance model.....	98
Table 3.3:	Alternative approaches to similarity assessment for the various cases of nominal attributes. A + represents a feasible approach for a case, whereas a ++ represents the recommended approach for that case. A – means that the approach does not apply.....	100
Table 3.4:	Relation <i>accommodations</i> with attributes that include null values.....	109
Table 3.5:	The different attribute types, their corresponding chapter sections, and the recommended methods for performing similarity assessments between their values. ....	111
Table 4.1:	Similar results to a logical <i>or</i> -query involving two attributes. ....	130
Table 4.2:	The different types of constraint connectives, their corresponding chapter sections, and the recommended methods for similarity assessments with each type. ....	134

## LIST OF FIGURES

Figure 1.1:	The three conceptual levels of a geographic information system: scenes, objects, and attributes. ....	6
Figure 1.2:	A simple configuration of spatial objects and the attributes used to capture (a) the topological properties of spatial relations and (b) the metric refinements that apply to the topological properties (modified from Egenhofer (1997)). ....	14
Figure 2.1:	Types of ontology according to their level of generality (Guarino 1998). ....	29
Figure 2.2:	Similarity versus distance (dissimilarity) as expressed by (a) a linear, (b) an exponential, and (c) a Gaussian function. ....	38
Figure 2.3:	Shortest path and is-a relationships in a hierarchical network structure.....	48
Figure 2.4:	Various types of graphs: (a) complete graph, (b) digraph, (c) pseudograph, (d) ARG, (e) bipartite graph, and (f) complete bipartite graph. ....	60
Figure 2.5:	Matchings in bipartite graphs: (a) a simple matching and (b) a maximum-weight matching. ....	61
Figure 2.6:	Demonstration of the concepts of component, maximum clique, maximum-weight clique, maximal clique, and clique. ....	62
Figure 3.1:	Assessing similarity based on absolute differences. ....	67
Figure 3.2:	Forms of change: (a) generation and destruction, (b) expansion and contraction, (c) alteration with no intermediate states, and (d) alteration with intermediate states. ....	71
Figure 3.3:	Graph representation of a similarity matrix. ....	75

Figure 3.4:	An ostensibly ordinal scale is ratio if the origin indicates absence of the property and the intervals between consecutive values represent equal amounts of change to the degree that the property is fulfilled. ....	81
Figure 3.5:	Similarity for ratio quantities: (a) as the inverse of logarithmic distance and (b) as the inverse of relative change.....	83
Figure 3.6:	Similarity for ratio magnitudes as the inverse of distance.....	86
Figure 3.7:	Values of ordinal rating scales: (a) as points and (b) as intervals.....	91
Figure 3.8:	Cyclic scales with uniform values: (a) angles and (b) cardinal directions.....	101
Figure 3.9:	Four different periods of land use including timbering, fishing, hunting, and fruit gathering (Hornsby <i>et al.</i> 1999). ....	103
Figure 4.1:	Similar results to a query involving relational operators.....	113
Figure 4.2:	Similar results to a conjunctive query using locally-better matching.....	116
Figure 4.3:	Combining two integral attributes to one that is separable (all weights are set to 1). ....	118
Figure 4.4:	Formulating multi-valued attribute similarity as (a) the problem of maximum-weight matching in a bipartite graph and (b) the assignment problem. ....	122
Figure 4.5:	Applications of a similarity measure for multi-valued attributes: (a) detailed topological relations and (b) disjoint temporal intervals.....	123
Figure 4.6:	Multi-valued similarity for sets of different cardinalities. ....	124
Figure 4.7:	Behavior of the value completeness parameter for similarity queries that involve sets of different cardinalities: (a) the query and database sets have the same number of values, (b) the query set has more values than the database set, and (c) the query set has fewer values than the database set. ....	126

Figure 4.8:	Establishing correspondences for multi-valued similarity of composite attributes. ....	127
Figure 4.9:	Similar results to a logical or-query involving one attribute. ....	129
Figure 5.1:	Psychological principles for spatial scene similarity assessments (Section 2.2.2.6). ....	135
Figure 5.2:	Forms of spatial scene queries. ....	137
Figure 5.3:	Representing a topological relation at progressively finer levels of detail. ....	139
Figure 5.4:	Representation of a spatial scene as a complete labeled pseudograph (road networks were omitted to avoid clutter). Properties of objects and relations that become constraints are denoted with a color-coding scheme, which is reused in subsequent chapter figures. ....	141
Figure 5.5:	Complete and incomplete solutions to spatial scene queries. ....	143
Figure 5.6:	Types of solutions to a relaxed CSP: (a) exact and complete, (b) exact and incomplete, (c) partial and complete, and (d) partial and incomplete. ....	144
Figure 5.7:	A low quality result produced by assigning the same significance to the class and the geometric attributes of the objects. ....	147
Figure 5.8:	Problems of coarse topological relations for scene similarity assessments: (a) reasoning for similarity based on distances in a conceptual graph may exclude highly similar matches in favor of others that are less similar and (b) the inability to discriminate among members of the same class treats relations, for which people may have distinct mental images, as equally similar. ....	150
Figure 5.9:	Two algorithmically different solutions (i.e., different assignments of database objects to query objects) to a scene query may be perceived from the users as a double retrieval of the same scene. ....	160

Figure 5.10:	Solving a CSP by creating the association graph for a query and a database scene and extracting the solutions.....	162
Figure 5.11:	Costs on efficiency and quality introduced by a careless relaxation. ....	163
Figure 5.12:	Creating the association graph for a relaxed query and a database scene and extracting the solutions.....	164
Figure 5.13:	The calculation of the similarities between objects and relations transforms the association graph into a weighted association graph. ....	165
Figure 6.1:	Overlap percentage $O$ and modified Spearman Rank Correlation coefficient $R'$ for the relevant portion of two ranking lists. ....	169
Figure 6.2:	Experiments ( $E_1$ - $E_4$ ) used to evaluate the hypothesis. ....	171
Figure 6.3:	Experiments for object-level queries: (a) compliant aggregation function, (b) deviant function that ignores integral attributes ( $E_1$ ), (c) deviant function that aggregates integral attributes with a Manhattan metric ( $E_{2A}$ ), (d) deviant function that aggregates separable attributes with a Euclidean metric ( $E_{2B}$ ), and (e) deviant function that ignores integral attributes and aggregates separable attributes with a Euclidean metric ( $E_3$ ). ....	172
Figure 6.4:	Experiment for scene-level queries: (a) three solutions to a scene query with dissimilarities computed for each node (i.e., object) and for each edge (i.e., relation), and the scene ranks produced when the dissimilarities were converted to similarities using (b) a linear function, (c) an exponential function, and (d) a Gaussian function.....	173
Figure 6.5:	Strict and generous non-linear conversion functions used in Experiment $E_4$ . ....	175
Figure 6.6:	Splitting integral attributes into groups using (a) an <i>optimal</i> and (b) a <i>worst</i> distribution policy. ....	179
Figure 6.7:	A 4-dimensional diagram depicting the measures (a) $O$ and (b) $R'$ . ....	180



Figure 6.8:	Experiment $E_1$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects. ....	182
Figure 6.9:	Experiment $E_1$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects. ....	183
Figure 6.10:	Experiment $E_1$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects. ....	184
Figure 6.11:	Experiment $E_1$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects. ....	185
Figure 6.12:	Overview of the results acquired from Experiment $E_1$ . ....	186
Figure 6.13:	Experiment $E_{2A}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects. ....	190
Figure 6.14:	Experiment $E_{2A}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects. ....	191
Figure 6.15:	Experiment $E_{2A}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects. ....	192
Figure 6.16:	Experiment $E_{2A}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects. ....	193
Figure 6.17:	Overview of the results acquired from Experiment $E_{2A}$ . ....	194
Figure 6.18:	Experiment $E_{2B}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects. ....	195
Figure 6.19:	Experiment $E_{2B}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects. ....	196
Figure 6.20:	Experiment $E_{2B}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects. ....	197

Figure 6.21:	Experiment $E_{2B}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects. ....	198
Figure 6.22:	Overview of the results acquired from Experiment $E_{2B}$ . ....	199
Figure 6.23:	Experiment $E_3$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects. ....	202
Figure 6.24:	Experiment $E_3$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects. ....	203
Figure 6.25:	Experiment $E_3$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects. ....	204
Figure 6.26:	Experiment $E_3$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects. ....	205
Figure 6.27:	Overview of the results acquired from Experiment $E_3$ . ....	206
Figure 6.28:	A sample diagram from Experiment $E_4$ , giving a rough estimate of the database size required to accommodate the tested query scenarios.....	210
Figure 6.29:	Experiment $E_4$ : averaged overlaps and correlations between the functions $L$ and $E_L$ . ....	211
Figure 6.30:	Experiment $E_4$ : averaged overlaps and correlations between the functions $L$ and $G_L$ . ....	211
Figure 6.31:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_L$ and $G_L$ . ....	212
Figure 6.32:	Experiment $E_4$ : averaged overlaps and correlations between the functions $L$ and $E_S$ . ....	212
Figure 6.33:	Experiment $E_4$ : averaged overlaps and correlations between the functions $L$ and $G_S$ . ....	213

Figure 6.34:	Experiment $E_4$ : averaged overlaps and correlations between the functions $L$ and $E_G$ .....	213
Figure 6.35:	Experiment $E_4$ : averaged overlaps and correlations between the functions $L$ and $G_G$ .....	214
Figure 6.36:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_L$ and $E_S$ .....	216
Figure 6.37:	Experiment $E_4$ : averaged overlaps and correlations between the functions $G_L$ and $G_S$ .....	217
Figure 6.38:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_L$ and $E_G$ .....	217
Figure 6.39:	Experiment $E_4$ : averaged overlaps and correlations between the functions $G_L$ and $G_G$ .....	218
Figure 6.40:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_S$ and $E_G$ .....	218
Figure 6.41:	Experiment $E_4$ : averaged overlaps and correlations between the functions $G_S$ and $G_G$ .....	219
Figure 6.42:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_S$ and $G_S$ .....	220
Figure 6.43:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_G$ and $G_G$ .....	220
Figure 6.44:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_L$ and $G_S$ .....	221
Figure 6.45:	Experiment $E_4$ : averaged overlaps and correlations between the functions $G_L$ and $E_S$ .....	221
Figure 6.46:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_L$ and $G_G$ .....	222

Figure 6.47:	Experiment $E_4$ : averaged overlaps and correlations between the functions $G_L$ and $E_G$ .....	222
Figure 6.48:	Experiment $E_4$ : averaged overlaps and correlations between the functions $E_S$ and $G_G$ .....	223
Figure 6.49:	Experiment $E_4$ : averaged overlaps and correlations between the functions $G_S$ and $E_G$ .....	223

# CHAPTER 1

## INTRODUCTION

*Similarity assessment* implies a conceptual process of judgment about the semantic proximity of two entities. In a rudimentary form, this process consists of a decomposition of the entities under comparison into elements in which they are the same, and elements in which they differ (James 1890). People are able to perform this task based on intuition and knowledge. Their judgments are usually subjective and display no strict mathematical models (Tversky 1977). Machines, however, must rely on mathematical formalisms if they are to reason accordingly. The challenge is to translate the cognitive process of a *qualitative* similarity assessment into the *quantitative* realm. Since human perceptions of similarity are also strongly influenced by situation as well as each individual's unique mental model (Goldstone *et al.* 1997), powerful yet flexible tools must be selected to guarantee a consistency between user-expected and system-generated results. This thesis explores such tools in the context of spatial database systems.

### 1.1 Terminology

Terminological confusion is often the culprit behind poor communication of ideas and lack of understanding, especially in scientific areas of multidisciplinary interest, such as those examined in this work. To avoid such problems, this section clarifies the meaning of several important terms that are used throughout the remainder of the thesis.

A *database* is a logically coherent collection of raw observations, called *data*. It is designed, built, and populated with data for a specific purpose and models some part of the real world, which is often called the *universe of discourse* or *miniworld*. A database is created and maintained with the help of a *database management system* (DBMS), that is, a system comprising a collection of software programs. A DBMS allows such tasks as constructing, manipulating, and querying databases for various applications (Elmasri and

Navathe 2000). An *information system* is a combination of one or more databases, managed by one or more DBMSs. In this thesis, the term *centralized database* denotes a single database managed by a single DBMS on the same computer system. A *multidatabase system*, in contrast, refers to a collection of multiple cooperating database systems (Sheth and Larson 1990). A *spatial information system* is an information system that contains, processes, analyzes, and displays spatially referenced data. When such data are limited to environmental-scale spaces (i.e., neighborhoods, street networks, cities) or to geographic-scale spaces (i.e., states, countries) (Freundschuh and Egenhofer 1997), a spatial information system is also called a *geographic information system* (GIS) (Laurini and Thompson 1992; Chrisman 2001; Worboys and Duckham 2004).

Spatial information systems store *data* about *entity instances* or simply, *entities*. These are real world objects or concepts that belong to *entity types* or *entity classes*. The latter are cognitive representations that people use to recognize and categorize entities or events in the real world (Dahlgren 1988). For example, *Rhodes* and *Greece* are entity instances of the entity types *island* and *country*, respectively. The database equivalents to entity types and instances are *classes* and *objects*, respectively. A *class* prescribes an intensional set of objects that are similarly structured and exhibit the same behavior (Dittrich and Geppert 1997). An *object* is the formal representation of a real-world entity in a *miniworld*. Objects of the same class in a database can be manipulated by common operators (Egenhofer and Frank 1992) and are described through a common set of properties. They are differentiated, however, by the different values they take for each property. In this sense, the properties may be viewed as functions that map specific qualities or quantities onto each object (Chen 1976). In the context of *relational databases*, classes, objects, and properties are also called *tables* (or *relations*), *tuples*, and *attributes*, respectively. Relational databases are based on the relational model for structuring data (Codd 1970) and account for the overwhelming majority of current

database implementations. In psychological terminology, sensory-identifiable entities are often referred to as *stimuli* and their perceived properties as *features* or *dimensions*.

People assess similarity among entity types and entity instances, whereas information systems perform the same task among classes and objects.

## **1.2 Information Retrieval in Geographic Information Systems**

Information is meaning extracted from the interpretation of data. The process of *information retrieval* from a database system typically comprises four steps. The first is the *query formulation*, when users employ the modalities of the system to specify a set of *constraints* (i.e., restrictions) on an *ideal* or *reference* object, which describes the entity they are looking for. Such an object may only incidentally exist in the database. During the second step, the DBMS searches through its database for objects that match the user's request. If matches are found, then the next task is their presentation to the user. The user's inspection and interpretation of the retrieved data, which results in the extraction of useful information, completes the process.

Traditional querying assumes that a user specifies exactly the constraints of valid results, and that the result set contains only those items that fulfill exactly the query constraints. These assumptions make it difficult for a user to always guess correctly the values stored, while exhaustive enumerations of acceptable alternatives to the ideal target would become a tedious process. Likewise, items that deviate somewhat from the query constraints should be part of a ranked result set as well, where items are ordered in ascending order based on a quantitative estimate of their deviation from the ideal object. A different paradigm, emphasizing similarity over equality, is of pivotal importance for information systems, and for geographic information systems in particular, for the following reasons:

- The *data provider-data user gap* is wide due to the differences between the nature of stored spatial data and the user's knowledge of these spatial data while querying.

People may know only approximately what they are looking for, so that they need to adopt an exploratory way of accessing spatial data (Schenkelaars and Egenhofer 1997). For example, in order to serve diverse user needs, GISs often employ a multi-resolution scheme (Buttenfield 1989; Bruegger and Kuhn 1991) that allows retrieval at varying levels of detail. Ideally, multiple representation databases should be derived from a single detailed representation by applying generalization algorithms (Beard 1989). Such algorithms, however, often encompass changes in the geometry of objects and the topological structure of their relations (Paiva 1998). Consequently, a query's geometric and topological specifications for a particular region of interest may differ from those that exist in the database for the same region.

- The *spatial-intuition gulf* between people who request spatial information and the models in spatial information systems becomes more apparent as spatial information systems are growing beyond the state of being tools of experts, and a wider and more diversified audience uses them on a daily basis. It is inconceivable that all GIS users share a common context and views about reality.
- The *lack of standard, cognitively-plausible formalizations* of spatial properties of geographic phenomena makes it even harder to support comprehensive, yet flexible methods for spatial information retrieval. Currently, only a few isolated efforts exist that capture how people interpret spatial properties and perceive spatial concepts (Lynch 1960; Mark and Egenhofer 1994; Worboys 2001; Worboys *et al.* 2004).
- The *diversity of background and expertise*, combined with the ill-defined spatial standards, are largely responsible for the *semantic, structural, and schematic heterogeneities* of cooperating database systems that model the same part of reality. In a recursive manner, the wide accessibility of these systems from a massive Internet audience—made possible by recent technological developments, such as the proliferation of web-scripting languages and web-enabled DBMSs—stresses further these problems and raises the requirements for effective information retrieval to a



whole new level. A recent study (Chang *et al.* 2004) estimated 450,000 online databases, a number that is likely to grow exponentially in the coming years.

- The *verbal-visual competition* of requesting spatial information verbally while presenting spatial query results graphically puts an undue cognitive load on users. Traditional spatial queries do not have a spatial expression *per se* as they are substituted by lexical or semantic equivalents. Thinking spatially is supported only in a very limited way at the query-formulation stage (Egenhofer 1994a), but alternative visual query modalities, such as sketching (Smith and Chang 1996; Egenhofer 1997; Haarslev and Wessel 1997a; Tversky *et al.* 2000), often help reveal a user's mental model of a spatial arrangement better than a verbal expression. By their very nature, however, such visual requests for spatial information retrieval are imprecise.

This sample is representative of the most significant problems affecting traditional methods for spatial information retrieval. It demonstrates why user-expressed queries may fail to coincide with—and consequently retrieve—any stored data.

### **1.3 A Framework for Similarity-Enhanced Retrieval in Spatial Information Systems**

Similarity-enhanced information retrieval goes beyond the determination of an exact match between queries and stored data. It provides the users with a range of possible answers, which are the most similar conceptually to the initial requests and, hence, the most likely to satisfy their queries. It also relieves users from the burden of reformulating their queries repeatedly until they find useful information. The results are ranked according to a similarity score associated with them, and the user has the possibility to choose any of the available answers. Thus, similarity becomes a tool for exploratory access to data. It resembles browsing, since users usually know only approximately what they are looking for. The advantage of browsing is that it is a highly interactive and familiar (i.e., web-browsing) procedure and leaves the final choice of what result to select to the user.

### 1.3.1 Concept

In terms of their dependency on one another, similarity assessments in a geographic information system can take place at three conceptually distinct levels so that any similarity assessment at a higher level of this framework implies prior similarity assessments within the lower levels. The building blocks of this schema are: (1) the spatial scene level, (2) the object (or relation) level, and (3) the attribute level (Figure 1.1).

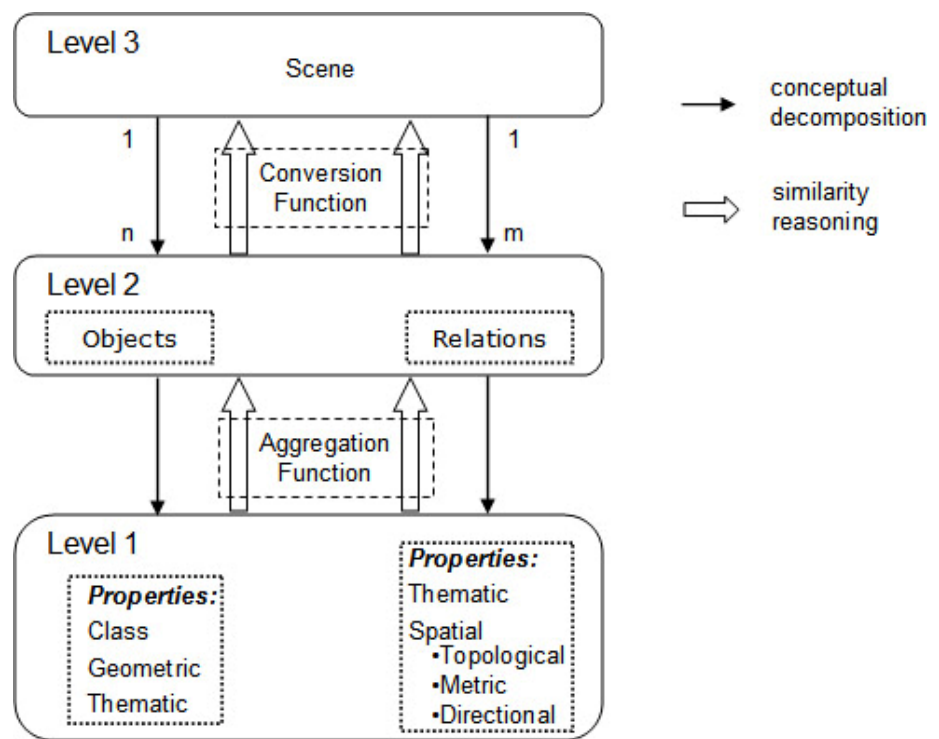


Figure 1.1: The three conceptual levels of a geographic information system: scenes, objects, and attributes.

A spatial scene is a collection of objects with spatial and potentially thematic relations among them. Images, sketches, maps, and even molecular structures of cells are types of scenes. In its most trivial form, a scene consists of a single object, whereas in an extreme setting it could be an entire large-scale geographic database with millions of objects.

Object characteristics consist of class information, as well as *geometric* and *thematic* attributes. Geometric attributes are associated with the geometry, shape, and size of the objects. Thematic attributes, on the other hand, capture non-spatial information. For example, the class of *Rhodes* is *island*, its *name* and *population* are thematic attributes, while a *shape description* or the *ratio of the major and minor axes of its minimum bounding rectangle* provide values for its geometric attributes. The class specification of an object determines its entity type in the real world. Sometimes, this information can be perceived as another thematic quality. Its special importance, however, is often reflected in the prominent position that it assumes within many DBMSs (e.g., object-oriented DBMSs (Atkinson *et al.* 1989)).

The same dichotomy of spatial and thematic characteristics carries on to relations, where the spatial component is typically subdivided into topological (i.e., pertaining to the connectivity relations of interiors, exteriors, and boundaries of spatial objects), metric, and directional parts. For example, *Rhodes*, which is *disjoint* from the *Greek mainland* and located *650km southeast* of *Thessaloniki*, has a *smaller* population than *Athens*. Conceptually, one can either talk about the existence of multiple relations between a pair of objects, or about a singular relation with topological, metric, directional, and thematic properties (Figure 1.1). This work adopts the latter view, which allows relations to be treated as objects and represented by a tuple; for example,  $\text{Relation}_{\text{Rhodes, Thessaloniki}} = (\text{"SE"}, \text{"650km"})$ .

Oftentimes, properties of objects and relations are distinguished as semantic and geometric, respectively (Blaser 2000; Rodríguez and Egenhofer 2004). This work refutes this segregation, because—as the previous examples demonstrated—objects may possess both geometric and thematic attributes and, reciprocally, relations are not exclusively spatial in nature. A query asking to retrieve two islands such that one has a larger population than the other still requires the retrieval of two objects with a definitive relation holding between them; this relation, however, is not spatial, but instead formed

by the difference in the value of a common thematic attribute of the objects. *Semantics*, on the other hand, is concerned with meaning on the large; it is an all-pervasive term relating to all kinds of measurements—whether of a geometric or a thematic nature—as well as people’s interpretation of such measurements (Wood 1975; Sheth 1995). Therefore, we abstain from such terminology and maintain the assertion that the main components of a scene are objects and relations and either of their attributes has thematic or geometric character. At times, *temporal* attributes are treated as a third type, which are then subject to typical temporal operations. We do not make this explicit distinction here, but rather include temporal as one special type of thematic attributes.

Within such a framework, the core of a similarity mechanism’s inferential ability is at the attribute level. By exploiting the differences among attribute values of objects and relations, a similarity algorithm can reason about the degree of difference or resemblance of a result to a query. When the query consists of a constraint on an atomic value of a single attribute, the process of similarity assessment takes place at the attribute level. When the query consists of multiple such constraints, a similarity assessment takes place at the object level. In both cases, the results are objects; the difference, however, is that in the latter case the individual similarity scores that were produced separately for each attribute must somehow be combined to a meaningful composite. In the same manner, a similarity assessment between two scenes requires an appropriate synthesis of the individual similarity measures derived separately for each pair of associated objects and relations.

### 1.3.2 Motivation

The establishment of methods for determining semantic similarity at the various levels of the framework has attracted an interdisciplinary interest. An important body of work originated within the field of natural language processing. These efforts established techniques that derive semantic similarity among concepts as a function of their distance

within a hierarchical structure and of their frequency of occurrence within large text corpora (Rada *et al.* 1989; Resnik 1995; Jiang and Conrath 1997; Leacock and Chodorow 1998). Psychology is another domain where the process of cognitive similarity assessments has been studied extensively and resulted in several proposals and models. Goldstone and Yun Son (2005) classify psychological models as *geometric*, *featural*, *transformational*, and *alignment-based*. The first three types are concerned with similarity assessments at the attribute and object levels, whereas the fourth category is interested in configuration similarity.

Scientists from the computer science and geographic information systems communities also yielded significant contributions. Dey *et al.* (2002) developed simple similarity measures for attribute values in order to identify double entries for the same entity in databases. Rodríguez and Egenhofer (2004) combined distinguishing features of entities with their semantic relations in a hierarchical network and created a model that evaluates similarity among spatial concepts (i.e., entity classes). Based on theories that were developed for representing and reasoning with topological, metric, and directional, relations (Egenhofer and Herring 1990; Randell *et al.* 1992; Egenhofer 1994c; Frank 1996; Shariff 1996), Egenhofer (1997), Egenhofer and Shariff (1998) and Goyal and Egenhofer (2001) developed, respectively, computational models that determine the similarity among values of such relations. Further studies integrated the results of these efforts and extended their scope to provide formalisms that incorporate all aspects of spatial relations during the comparison of spatial scenes. Some of these studies proposed qualitative similarity measures (Bruns and Egenhofer 1996; Li and Fonseca 2006), whereas others offered quantitative estimates (Gudivada and Raghavan 1995; Nabil *et al.* 1996; Petrakis and Faloutsos 1997; Stefanidis *et al.* 2002) of similarity for simple scenes, consisting of a small number of objects. Based on the idea of spatial-query-by-sketch (Egenhofer 1996), Blaser (2000) implemented a more elaborate prototype that assesses the similarity between a user-drawn sketch and a collection of spatial scenes stored in a

geographic database. This prototype relies heavily on geometric object attributes and spatial relations, but underestimates the thematic component. Further work enabled similarity evaluations between spatial scenes in the context of large-scale geographic databases, focusing primarily on relational similarity and efficient query processing (Papadias *et al.* 1999b; Papadias *et al.* 2001; Papadias *et al.* 2003).

Although all of these efforts have merit, each of them approaches the topic of semantic similarity from a different perspective. Some concentrate on a particular level within the overall framework, whereas others specialize on a specific aspect of a particular level. The outcome of such a fragmentary approach to similarity is a number of significant problems, such as:

- Inability to generalize or specialize the measures so that they apply to different levels. For example, many of the models for concept similarity cannot be readily applied to the task of attribute-level similarity assessments and vice versa.
- Restrictive or unrealistic assumptions justified for the sake of efficiency, or stemming from a narrow perception of the problem's extent, such as considering that the compared scenes have an equal numbers of objects, or that they have a relatively small number of labeled objects.
- Failure to accommodate different retrieval scenarios and to handle special cases, such as those arising when incomplete information is encountered.
- Incompatible measures (e.g., qualitative vs. quantitative) that are difficult to integrate and process together.

The large majority of the discussed proposals and prototypes share an additional disadvantage—with Rodríguez's (2000) work being a notable exception—neglecting the human factor. The similarity measures that they advocate are typically derived in an *ad-hoc* manner, guided by experience and observation, and serve practical retrieval needs. In this sense, they are concerned with similarity from a pragmatic rather than a cognitive

point of view. Findings from psychology about the way that people perceive the nature of similarity, its properties, and its relationship to peripheral notions, such as *difference* and *dissimilarity*, are largely ignored. The exclusive focus on the computational aspects and the dismissal of the cognitive elements render the plausibility of such approaches to human perception questionable. Context, which is another psychological factor with a profound influence on people's similarity judgments, is at best captured through the provision of a set of user-adjusted parameters that help finetune the produced similarity scores. Delegating context-specification entirely to users in this manner makes the process of information retrieval slow, tedious, and even abstruse in the case of complex similarity assessments.

### 1.3.3 Goal

The goal of this thesis is to create a comprehensive framework for supporting similarity queries in spatial information systems. The focus of this framework is primarily on conceptual aspects of similarity assessments. Its parts should include a sound theoretical foundation, solid computational formalisms that reflect people's similarity judgments, and a scalable architecture that allows similarity assessments at all three levels of attributes, objects, and scenes, in a consistent and coherent manner.

### 1.3.4 Hypothesis

A crucial component of the architecture of the framework is the interaction among its levels. The object level is primarily responsible for this interaction, because it provides the linkage between the attribute and the scene levels. In order to determine the similarity of two objects, a *distance* (i.e., *dissimilarity*) measure must first be defined between their formal representations. Since the end product comprises results that will be presented to people, this estimate must accord with human notions of object similarity (Gärdenfors 2000).

The choice of two functions becomes critical in achieving this objective: (1) the *aggregation function* and (2) the *conversion function*. Aggregation functions combine atomic judgments to an overall composite measure. These are the functions that connect the first and second levels of the framework. They are used when separate attribute dissimilarities must be combined to an overall measure, indicative of the global dissimilarity between a pair of objects, or a pair of relations. Conversion functions translate dissimilarity to similarity and vice versa. In typical approaches, the role of conversion functions is simply cosmetic; they perform a routine transformation because it is more enticing to present users with a similarity rather than a distance score. The role of these functions, however, is much more vital in this work because they are responsible for translating aggregate dissimilarity to *perceived* similarity (or dissimilarity).

A large body of intensive experimental and theoretical research in psychology during the last decades converged to a consensus on the desired form of such functions so that they reflect human similarity assessments (Attneave 1950; Torgerson 1965; Nosofsky 1986; Shepard 1987; Ennis 1988; Nosofsky 1992; Takane and Shibayama 1992; Hahn and Chater 1997; Gärdenfors 2000). The first part of this consensus pertains to the aggregation function, which should differ depending on whether the atomic judgments are made on *separable* or *integral* attributes. Separable attributes are those that are perceptually independent, that is, they refer to properties that are obvious, compelling, and clearly perceived as two different qualities or quantities that an entity possesses (Torgerson 1965). Conversely, a set of attributes creates an *integral*<sup>1</sup> group, when their values are conceptually correlated, and lack an obvious separability (Ashby and Townsend 1986; Ashby and Lee 1991). Conceptual correlation implies that the values of

---

<sup>1</sup> The term *integral* does not connote statistical or causal, but perceptual correlation. It is possible that two separable attributes have values that are causally correlated and, conversely, that the attributes of an integral group have values that are statistically independent.



these attributes are perceived as one property, regardless if the representational conventions in information systems model this property through a set of concomitant attributes. The second part of this consensus dictates that the perceived similarity and the aggregate distance do not have a complementary relationship, but rather that the former derives from the latter through nonlinear monotonically decreasing functions (Nosofsky 1986; Shepard 1987).

Both findings have repercussions for formalized similarity assessments if these processes are to comply with human reasoning. These repercussions become especially relevant in the setting of spatial information systems. Due to the monotonically decreasing relationship between perceived similarity and aggregate distance, the choice of the conversion function is rather indifferent for information systems where similarity retrieval is confined within the attribute and object levels. The similarity scores may vary, but the produced rankings for similar objects will be identical regardless of the conversion function chosen. This choice ceases to be indifferent and becomes essential in spatial information systems, however, where similarity assessments may be required at the higher level of spatial scenes. The similarity between two scenes depends on the perceived similarities of the associated object and relation pairs. The decision on the conversion function becomes, therefore, instrumental because it affects the ranking of the most similar database scenes to a scene query.

The situation is similar when it comes to segregating separable and integral attributes. General-purpose information systems employ primarily separable attributes. For example, the University of Maine's *personnel* database may contain such attributes as *age*, *job title*, *salary*, and *sex*, which are perceived as different things. A significant amount of integral attributes, however, may be hidden in the representational formalisms that GISs employ to model the complex topological relations of spatial objects (Egenhofer and Franzosa 1995; Clementini and di Felice 1998) (Figure 1.2a). The set of possible integral attributes may grow if one also considers that such topological

formalisms are often complemented with equally-complex metric refinement models (Shariff 1996; Nedas *et al.* in press), which introduce a large number of additional attributes in order to capture the metric aspects of topological relations (Figure 1.2b). The recognition of the integral attributes and the form of the aggregation function affect the rankings at the object level, and their influence also propagates to rankings at the scene level.

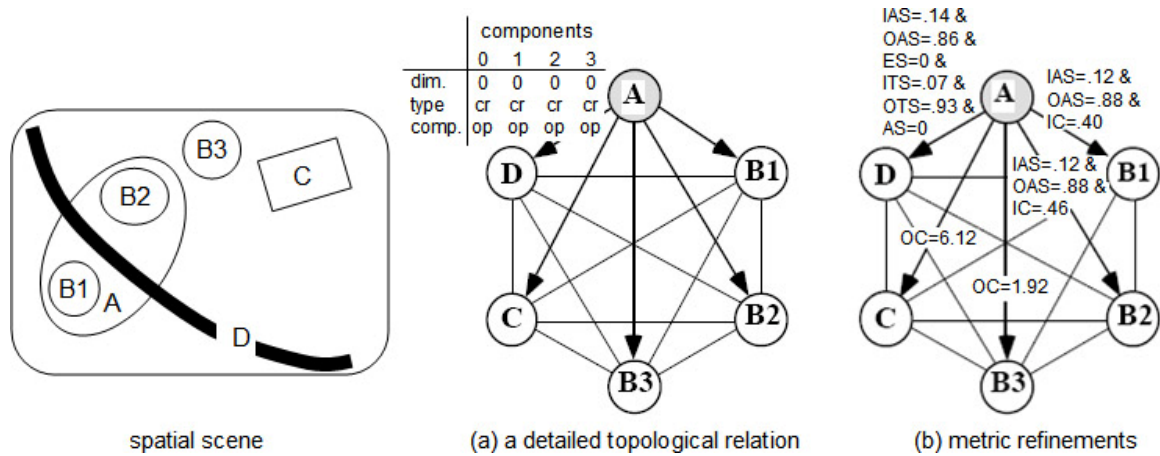


Figure 1.2: A simple configuration of spatial objects and the attributes used to capture (a) the topological properties of spatial relations and (b) the metric refinements that apply to the topological properties (modified from Egenhofer (1997)).

It becomes, therefore, apparent that a *psychologically compliant* model for similarity assessments within spatial information systems should (1) be aware of which attributes are integral and which are separable and (2) use *psychologically correct* aggregation and conversion functions to determine the similarity of a result to a query. Most of the current studies and prototypes in the literature typically ignore both requirements. Hence, it is relevant to determine whether the incorporation of these provisions into a formalized similarity assessment makes an essential difference or not. This observation leads to the following hypothesis:

*Psychologically deviant methods produce a set of results, in the relevant portion of the ranking list, dissimilar to that obtained by psychologically compliant methods.*

A *psychological deviant* method is one that deviates in some way from the highlighted psychological findings. Therefore, this hypothesis can be dissected to three testable statements (HS):

- HS<sub>1</sub>: A psychologically deviant method that fails to identify integral attributes and their groups produces a set of results, in the relevant portion of the ranking list, dissimilar to that obtained by a psychologically compliant method that recognizes such attributes and groups.
- HS<sub>2</sub>: A psychologically deviant aggregation function that deviates from the psychologically-suggested form produces a set of results, in the relevant portion of the ranking list, dissimilar to that obtained by the psychologically compliant aggregation function.
- HS<sub>3</sub>: A psychologically deviant conversion method that uses the linear function produces a set of results, in the relevant portion of the ranking list, dissimilar to that obtained by the non-linear psychologically compliant functions.

Proving these hypothesis statements requires comparing psychologically compliant functions against common psychologically deviant methods encountered in the literature and evaluating, through appropriate measures, their incompatibility with respect to the results that they retrieve for a given query. The focus of these comparisons is on the relevant portion of the ranking list, that is, the first few ranks of the results, because they capture the most similar items to a user's query. If the first testable statement (HS<sub>1</sub>) proves true, it will dictate the need for new research and human-subject testing in order to distinguish separable vs. integral attributes in spatial representational formalisms. Otherwise, research in this direction would be moot. If the second and third testable

statements ( $HS_2$  and  $HS_3$ ) prove true, they will provide a common grounding for the design of future prototypes and systems that are able to reason about similarity “intelligently.” A negative outcome, on the other hand, would imply that the criteria for choosing aggregation and conversion functions can be simply reduced to those that pertain to computational efficiency.

### 1.3.5 Research Questions

Four key questions drive the development of this thesis:

*Question 1: What are the psychological properties of similarity that a formal system should take into consideration?*

A successful similarity model for GISs would help eliminate the restrictions imposed by exact matches, thereby providing satisfactory reasoning mechanisms for semantically similar results. Satisfactory results imply a match of methods for spatial similarity retrieval with human perception and cognition. The major obstacle to this goal is the elusiveness and complexity of similarity, which is difficult to describe by formal logical theories or represent with mathematical models. Therefore, it is crucial to examine psychological findings on the nature of similarity, and isolate and formalize those that are relevant for semantic information retrieval.

*Question 2: How does one create a minimal set of generic algorithms that addresses similarity assessments for the majority of attributes typically encountered in spatial databases?*

We are interested in algorithms that yield results that are consistent with people’s judgments of similarity. Does each attribute require a unique algorithm or can one general algorithm achieve the stated objective equally-well for certain groups of attributes? If the latter is true, should we classify attributes into groups based on some structural characteristic, such as the specified data type in the database schema, or on a different criterion? Are there special cases of attributes that demand separate treatment,

and if so, what would the correct approach be for them? Finally, under what circumstances should one algorithm or model be preferred against another for the same group of attributes? Answers to these questions provide the theoretical foundation that is needed for formalized similarity assessments at the attribute level.

*Question 3: Which are the possible types of queries that a user may express at the object level?*

Queries at the object level may involve different kinds of constraints. For example the user may formulate a query using relational operators other than the basic equality (such as *greater than*, and *less than*) or logical operators (such as *and* and *not*), or a combination of both. It is important to examine the semantics of these queries and to develop methods that yield plausible similarity measures for assessments at this level. The combination of multiple constraints also suggests the need for an effective and intuitive weighting scheme that enables users to determine the relative salience of each constraint.

*Question 4: What are additional issues that emerge in scene similarity assessments?*

Assessing scene similarity can be a difficult problem. Its solution requires that one first identifies corresponding elements in the two compared scenes. This matching process can become increasingly complex and error-prone for large scenes as it is questionable how to choose one set of associations over another or how to account quantitatively when some of the elements remain unmatched. There are many additional requirements that scene similarity assessments introduce. We seek a comprehensive and theoretically sound methodology that simplifies the process and provides an organized approach to resolving such problems.

## **1.4 Approach**

This thesis aims at developing a framework for semantic information retrieval from spatial databases. The framework is strongly influenced by studies in cognitive

psychology. The results of those studies are based on numerous experiments that investigated the process of similarity assessment in human subjects and yielded significant findings about the intricacies of such mental processes. We do believe, therefore, that a reliance on these findings is likely to give desirable and commonly accepted measures of similarity.

The investigation starts with a systematic examination of the most important psychological insights about the nature of similarity, its properties, its relationship to related concepts such as *distance*, *dissimilarity*, and *difference*, and the different types of context that may influence similarity judgments. From models and studies that originated within the psychological discipline, we highlight and retain only those properties and theories that are relevant for the purposes of semantic information retrieval. A justification is provided for the properties that are deemed irrelevant. Part of the initial investigation is to assess the role and usefulness of ontologies in the framework. Ontologies are rich structures that capture a view of the world, provide an agreement on the meaning of terms used to describe this particular view, explicate the interrelationships between the concepts that these terms stand for, and distinguish semantics from data representation. Therefore, ontologies are semantic constructs that formalize meaning and are directly relevant to this work.

The results of the inquiry into the psychological domain provide the foundation for building a model that produces conceptually plausible similarity measures. We follow a bottom-up approach, starting from the attribute level and progressing systematically to the object and scene levels. To account for the diversity of attribute types we seek a classification scheme that segregates attributes into types that exhibit the same behavior so that generic classes of algorithms can be developed for each type.

The algorithms produced at the attribute level provide the basis for similarity assessments at higher levels. The developed set of methods for the object level contributes a consistent and comprehensive methodology for spatial similarity retrieval in

response to complex queries with combinations of logical operators. The focus is again on providing reliable similarity measures that are consistent with people's intuition, rather than conveniently conforming with theories that may have appealing mathematical properties, but contradict human similarity reasoning. We provide an exhaustive list of spatial query scenarios with conjunctions, disjunctions, and negation and present justified solutions for each case. In this way, this part of the thesis extends the seminal work of Salton *et al.* (1983), who first considered such issues in information retrieval. Research at this level also addresses cases of special attributes that require a customized approach, such as multi-valued and composite attributes, which extend beyond atomic value assessments. The interaction of multiple constraints raises the issue for a weighting model that allows specifying the relative prominence of some constraints over others so that different user objectives and preferences are reflected in the produced results. In this sense, weights capture a dynamic aspect of context. 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 the database (Hearst 1994).

The next step of the framework develops an infrastructure for handling similarity assessments between spatial scenes. This type of similarity assessment relies on a prior process of association that identifies the correspondences between elements of the compared scenes. This is a hard combinatorial problem and the solution that we advocate is dependent not only on the adoption of a sound and fitting computational formalism but also on an infusion into the process of a variety of knowledge related to the spatial domain.

For all three levels, a comprehensive suite of tools is provided for supporting similarity assessments in the scenario of incomplete information. Such information may be encountered at the attribute level in the case of null values, that is, values that introduce some degree of uncertainty in the specification of the object that they describe.

This scenario may also occur when comparing multi-valued attributes whose sets contain a different number of values, or spatial scenes with a different number of objects in them.

## 1.5 Scope

Although this thesis focuses explicitly on similarity in spatial information systems, its findings and contributions are expected to apply to information systems in general without requiring significant modifications. The differences between these two types of systems have largely disappeared in the last years, because spatial information systems that record spatial properties about shapes and spatial relations often include a large number of thematic attributes in their specification, while at the same time, traditional information systems are becoming increasingly spatially-aware (e.g., bank customer records getting joined with customer locations, or clinical records that are often geo-coded). Furthermore, both spatial and thematic properties are eventually stored and represented as quantitative or qualitative values (e.g., <Rhodes, *disjoint*, Greece>, <Rhodes, 650km, Thessaloniki>) so that a single approach suffices for both types of information systems up to the object level (Figure 1.1).

This work does not make the assumption that classes of objects or relations in the database should necessarily contain an identical set of features but assumes homogeneity under all other circumstances. A *homogeneous environment* is granted when objects are structured identically and represented through the same set of semantic and data specifications (e.g., same semantics, units, domains of values). In multidatabase systems, however, a similarity assessment must take place within a *heterogeneous environment*. Heterogeneity is the outcome of differences in the structure, schema, and semantics of the component database systems. Data integration studies have already reduced such conflicts to a large extent (Bishr 1998; Bernstein *et al.* 2004; Park and Ram 2004; Uschold and Gruninger 2004; Doan and Halevy 2005), albeit from the perspective of traditional information retrieval. It is possible that new requirements may need to be



imposed on data integration if similarity retrieval is to extend its scope beyond a homogeneous environment. Such issues are not investigated in this thesis. Furthermore, we assume that the homogeneous environment is *structured*. Structured data sources are those that adhere to a well-defined schema and their values are instances composed of simple atomic data types, like integer, real or character (Domenig and Dittrich 1999). Relational and object-oriented database systems are structured data sources.

Another special case occurs with similarity comparisons that involve binary large objects (BLOBs), such as images (Flickner *et al.* 1995; Carswell 2000), video clips (Sistla *et al.* 1997; Wu *et al.* 2000), or audio files (Kosugi *et al.* 2000; Liu and Huang 2000; Berenzweig *et al.* 2003), and character large objects (CLOBs), such as large text corpora and documents (Salton *et al.* 1975; Wong *et al.* 1987; Korfhage 1997). Unlike traditional databases dominated by retrieval with exact matches, the notion of similarity is inherent in retrieval of multimedia objects (Grosky 1997). The goal is to be able to direct queries against the actual objects themselves (i.e., querying-by-content), rather than querying their textual descriptions in the form of metadata. Users should be able to provide surrogates of the objects as inputs, against which the similarity of the stored objects would be compared. For example, a user may draw a sketch and retrieve digital images similar to the sketch (Blaser 2000). Due to the usually huge size, complex structure, and unique characteristics of each of these types of objects, the models to assess multi-media similarity expose a great variability. Deriving similarity among such objects is a separate field of research, with unique requirements and characteristics, and does not constitute part of this effort. Our work, however, is complementary to such efforts. For example, the methods in this thesis may be used to query the metadata associated with multimedia objects. They can also directly apply to the tasks of deriving and aggregating similarities for the attributes used to represent complex objects.

This work is concerned with similarity mostly from a conceptual rather than implementation point of view. Topics that pertain to computational optimization of the

algorithms and details of lower-level access to the data (e.g., similarity indexing techniques) are excluded.

The goal of this thesis is not to come up with a unique and single computational model that is capable of evaluating similarity under any situation or context. The choice of specific algorithms for a particular database and its attributes is at the discretion of the database administrator/designer or the users. Our task is to investigate the alternatives, provide the theory and methods, and pinpoint which of them should be preferred under different circumstances or contexts so that appropriate choices can be made.

## **1.6 Intended Audience**

This thesis is intended primarily for researchers and developers from the community of spatial databases. It may be of interest, however, to any person concerned with semantic information retrieval, similarity assessments, and the design of future geographic information systems. The audience also includes experts from the fields of computer science, cognitive science, human-computer interaction, linguistics, and artificial intelligence as it relates to the intelligent retrieval of semantic information and the design of intelligent search engines on the semantic web.

## **1.7 Organization of the Thesis**

The three conceptual levels in Figure 1.1 prescribe the organizational structure of this thesis. A chapter is devoted to each level of the framework. Each chapter builds on observations and findings of previous chapters. The assessment of previous research, the evaluation of the hypothesis, and the conclusions are each compiled in separate chapters. This leads to the following structure of the remainder of the thesis:

The second chapter embeds this thesis into the context of previous research efforts. It provides the necessary background in related fields of study and argues about the relevance and applicability of previous results to this work. This thesis uses terminology,

ideas, and findings from those fields. Therefore, a basic understanding of their main concepts is required from the reader, in order to understand our work.

The third chapter investigates similarity assessments at the attribute level. Its objective is to identify a functional classification of the most common attribute types such that generic algorithms that capture the similarity among the values of each type can be developed. An important set of categories is based on the four scales of measurement, referring to cognitive and structural commonalities that are typically found in captured data. The chapter includes in its beginning an argument for a unifying perspective of similarity, which aids in establishing reliable similarity measures, determining the suitability of previous similarity models and theories for each of the proposed attributes types, and capturing implicit aspects of context that may not be immediately obvious. A comprehensive rationale is also formulated for handling attributes that include null values.

The fourth chapter creates the transition from the attribute to the object level by extending similarity assessments beyond simple equality queries on atomic values. It is concerned instead with addressing the similarity requirements of more complex requests that involve a number of attributes, and where constraints may interact through alternate combinations of conditional and logical operators. Particular emphasis is put on the process of conjunction, and on developing a set of aggregation functions that best express its semantics.

The fifth chapter advances similarity assessments to the most difficult and complex level of spatial scenes, where all the findings of chapters 3 and 4 are integrated. The notion of an association graph is introduced, which consists of nodes and edges that represent matched objects and matched relations in the compared scenes. The approach is centered on the extraction of the maximal cliques from this graph, which are substructures corresponding to the most similar scenes specified in the query. This methodology is based on a graph-theoretic algorithm, originally introduced in the field of

computer vision, which is adapted to accommodate scene comparisons in a geographic-context. Parts of this adaptation include (1) an examination of the different types of databases and query modalities in spatial information systems and their effect on similarity retrieval for spatial scenes, (2) an analysis of different methods for relaxing the constraints of the original query so that similar matches can be found, (3) a set of considerations for evaluating the relative significance of object constraints, (4) a comprehensive investigation on the suitability and the role of different types of spatial relations in scene similarity assessments, and (5) a detailed and flexible model for handling incompleteness when the query and database scenes have a different number of objects. Issues relevant to result presentation and to computational efficiency are also addressed.

The sixth chapter evaluates the three testable hypothesis statements. The chapter starts with an overview of the experimental design and introduces the measures used to provide evidence for the support or the rejection of the hypothesis. Each hypothesis statement is evaluated through one or more experiments. Each experiment comprises a description of its setup, a graphical illustration of the obtained results, a comprehensive interpretation of the outcome, and the conclusion on the validity of the hypothesis statement that it tests.

The seventh chapter concludes this thesis. It offers a summary of the thesis, discusses the major results, and highlights the most important contributions of this study. It also speculates on future research activities that complement this research or were enabled through it.

## CHAPTER 2

### SEMANTIC SIMILARITY IN INFORMATION SYSTEMS

Addressing the problem of semantic similarity in information systems requires a combination of knowledge from fields as diverse as computer science, psychology, linguistics, and philosophy. The interdisciplinary efforts in some of the problems that we address suggest an issue-based rather than a discipline-based approach. Our overview is arranged in three sections. The first introduces ontologies, which are rapidly evolving as a central component of current information systems. The second section describes the most important properties of similarity as well as its relationship with context and the notion of *difference*. It also reviews models that were developed to assess similarity among objects, concepts, and spatial configurations. The third section presents definitions, formalisms, and concepts from fuzzy set theory and graph theory, which are prerequisites for developing and justifying the similarity framework of this work.

#### 2.1 Ontologies

The word *ontology* has lately become very popular within the knowledge engineering community (Staab and Studer 2004). Its interpretation, however, is still vague, since the term has occasionally been used under slightly different meanings. The notion was originally introduced by philosophers—ontology is a branch of philosophy—and its study dates back to Aristotle (350 B.C.-b). It is composed of the two Greek words *onto* (being) and *logos* (reasoning); therefore, one may say that ontology is the science of being that reasons about everything that exists (in Aristotle’s words “the science of being qua being”). Gruber (1992) states that ontology, in the philosophical sense, is a systematic account of existence. Its main goal is then the discovery of truth (Zuniga 2001). Guarino (1998) distinguishes between the Ontology (with a capital “o”), as the philosophical Ontology, and ontology (with a lowercase “o”), as the term originating from the

computer science community. He defines Ontology as a particular system of categories accounting for a certain vision of the world. According to this definition, there is only one philosophical Ontology independent of the language used to describe it.

Unlike this unique, global, and always true philosophical Ontology, every individual has a different understanding of reality and the surrounding world. This atomic view, which constitutes the individual's personal ontology, is commonly known in psychology as the individual's *mental model*. Such personal ontologies are mostly implicit and hidden within us (Farquhar 1997). Dissimilarities among such ontologies are a natural consequence of different experiences, needs, backgrounds, linguistic conventionalities, and cultures, which imply different viewpoints and assumptions (Goldstone 2003; Rosenthal *et al.* 2004). Although this natural divergence is valuable, it often leads to problems in people's interactions and understandings. The need of people, organizations, and especially software programs to communicate without ambiguity led to ontologies as defined and implemented from the knowledge engineering community.

#### 2.1.1 Defining an Ontology

The most frequently cited definition in the literature comes from Gruber (1992) who states that an ontology, in the context of computer science, is an explicit specification of a conceptualization. A *conceptualization* refers to an abstract model of how people think and organize concepts and things in the world, usually restricted to a particular area of interest. An *explicit specification*, on the other hand, means that the concepts and things of this abstract model are represented formally by explicit terms, relations, and definitions (Gruninger and Lee 2002). Guarino (1998) refined Gruber's original definition by distinguishing between an ontology and a conceptualization. For him an ontology is a logical theory, accounting for the intended meaning of a formal vocabulary (i.e., its ontological commitment to a particular conceptualization of the world); therefore, ontology is an engineering artifact. It is language-dependent and uses a specific

vocabulary to describe a part of reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary terms. On the other hand, a conceptualization is language-independent and equivalent to the philosophical Ontology. The definition from Guarino has also undergone some criticism. For example, Zuniga (2001) argues that what Guarino calls a conceptualization is distinct from philosophical Ontology. Alternative definitions of ontologies in the context of information systems are provided by Guarino and Giaretta (1995).

The multitude of the available definitions contrasts ironically with the purpose of ontologies in computer science, which is simply to provide an agreement on the meaning of the words. For this thesis, we use the definition from Mena *et al.* (1998), which states that “an ontology is a specification of a representational vocabulary for a shared domain of discourse, which may include definitions of classes, relations, functions, and other objects.” It names and describes the entities that may exist in that domain, their attributes, functions, as well as their relationships. Therefore, an ontology is roughly a synonym for an agreed-upon terminology. It provides an agreement on the meaning of a set of terms in order to represent a domain and to communicate knowledge about it (Farquhar *et al.* 1996).

A domain ontology stands somewhere in the middle between the philosophical Ontology and the mental models of individuals. It differs from Ontology, because it is interested only in one particular domain of knowledge and not in everything that exists; therefore, there is only one Ontology, but many domain ontologies (Fonseca 2001). A domain ontology also differs from implicit mental models by being explicitly structured and constructed and, most importantly, by being shared through the concept of ontological commitment. Multiple parties (e.g., persons, agents, software systems) agree to commit to a particular ontology when communicating about a common domain of interest, despite the fact that they do not necessarily share the same mental models (Holsapple and Joshi 2002).

### 2.1.2 Common Misconceptions about Ontologies

Ontologies are often erroneously equated with other constructs. The most common misconception seems to be the congruence with database schemas (Spyns *et al.* 2002). A database schema can be seen as an ontology as long as it is a conceptual schema (Gruber 1992; Guarino 1997). The main difference, however, is one of purpose. An ontology is developed to make clear the meaning of the terms used in a particular domain, whereas a database schema is developed to model some available data. The relations and attributes in a database schema have names carrying an implicit semantic, which is the concept they stand for; however, the schema carries only the names but not necessarily the concepts, because different people may interpret these names differently (Busse *et al.* 1999). A schema needs to be associated with an ontology in order to make the semantics of the data source clear (Cui *et al.* 2001); therefore, an ontology provides a domain theory and not the structure of a database. In addition, an ontology is concerned with the possibility, and not the actuality, of existence (Gangemi *et al.* 1998). It models all possible entity types that may exist in a domain, independently of whether information about entities belonging to these types exists and can be stored in a database (Fonseca 2001). Hence, an ontology is richer in its semantics and in its content than common database schemas.

Ontologies are also often equated with taxonomic hierarchies of classes. Hierarchies that specify classes and their subsumption relationships represent one structural means of building ontologies. Ontologies, however, need not be limited to these forms (Gruber 1993). They can be much more than simple taxonomies of concepts, involving constraints, axioms, and interrelations among concepts (Guarino 1997).

### 2.1.3 Ontology Types

One possible classification of ontologies according to their ontological depth (i.e., their level of explicitness and formalization) is the following synthesis from the classifications by Gangemi *et al.* (1998), Rodríguez (2000), Welty (2000), and Smith and Welty (2001):



- *Catalog*: A list of normalized terms without any axioms or glosses. A catalog can be the ontology of the products that a company sells.
- *Glossed catalog*: A catalog with natural language descriptions of the terms (e.g., the dictionary of biology).
- *Simple taxonomy*: A collection of concepts organized by a partial order induced by inclusion.
- *Thesaurus*: Description of terms, plus relations to other more general or more specific terms within a common hierarchy. An example of such an ontology is WordNet (Miller *et al.* 1998).
- *Characterized taxonomy*: A collection of concepts along with their relations and properties, such as the ontology for the (KA)2 community (Benjamins 1998).
- *Fully axiomatized taxonomy*: A collection of concepts, semantic relations, properties, and axioms, such as the GALEN project (Rector *et al.* 1993).
- *Context library*: A set of axiomatized taxonomies with relations among them, such as Cyc (Lenat 1995).

Another useful classification of ontologies is according to their levels of generality (Figure 2.1).

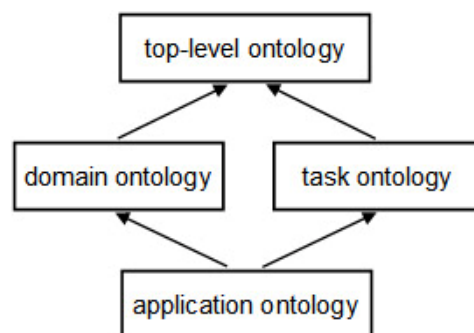


Figure 2.1: Types of ontology according to their level of generality (Guarino 1998).

- *Top-level ontologies* describe very general concepts, such as space and time, which are typically independent of a particular problem or domain. They are common-sense ontologies that may be accessed by large communities of users as well as from other ontologies. An example of a top-level ontology is SUMO (Niles and Pease 2001).
- *Domain ontologies* are the most commonly encountered and describe the vocabulary for a specific domain, such as cars or animals. In GIS such domains can be remote sensing or the urban environment (Fonseca *et al.* 2000).
- *Task ontologies* are more specific than domain ontologies as they describe a generic task or activity that occurs inside a domain. For example, a task ontology may describe noise pollution, an activity that occurs inside the urban environment.
- *Application ontologies* express concepts depending on both a particular domain and a task. They are often specializations of both of the related ontologies.

#### 2.1.4 WordNet

WordNet is a large semantic (and for the most part hierarchical) network for the English language that contains nouns, verbs, adjectives, and adverbs organized into sets of synonyms (synsets) (Miller 1995). The focus of WordNet is at the concept level (Lenat *et al.* 1995). Each synset is a node in the network corresponding to one concept, that is, a particular sense of an English word. WordNet encompasses both lexical and ontological information. Its lexical information is derived from the various word senses that it offers. In this sense, WordNet resembles a dictionary. It provides definitions of the words and includes sample sentences that demonstrate their use in natural language. The ontological information of WordNet is derived from the semantic relations that hold among the various word senses. From these relations our work considers synonymy, antonymy, hyponymy, hypernymy, and meronymy. A synonymy relationship between two terms holds when the terms have the same meaning (e.g., *building* and *edifice*). A hyponymy relationship holds when one term is less general than the other. A hypernymy relationship

is the inverse of hyponymy. For example, for the two terms *house* and *building*, the former is the hyponym and the latter the hypernym. Antonyms are terms that have opposite meaning (e.g., *lighted highway* vs. *unlighted highway*). The meronymy relation indicates the connection between parts (components) and wholes (e.g., *roof* is part-of a *building*). Although WordNet may be seen as an upper-level ontology, it can also be used as a domain-ontology building tool, allowing to pursue generality, identifying subtle differences in meaning between concepts, and enforcing readability and consistency by introducing linguistic discipline (Guarino 1997).

### 2.1.5 Problems of Ontologies

Although ontologies are becoming increasingly popular, ontological engineering—the discipline concerned with their development—is relatively novel and, hence, immature. One of the basic problems is the construction of poor-quality ontologies, often the result of unrestrained and erroneous use of the subsumption relationship (Guarino and Welty 2000). Although the representation of hierarchical knowledge is important in the design of formal ontology, there is little available advice on the problems that may be encountered during the ontology design process (Jones and Paton 1998). Ontoclean (Guarino and Welty 2002) is a methodology that provides guidance in validating taxonomies by exposing inappropriate modeling choices.

Another problem arises when the ontology users do not share the same assumptions and beliefs as the original designers. These differences result in ontologies that are not shared by many of the members of the community for which they were implemented. The ontological commitment may be very narrow, which in turn defeats the ontology's purpose for sharing and reusing of knowledge (Gruninger and Lee 2002). Holsapple and Joshi (2002) recommend a collaborative approach to ontology-design in order to overcome this problem. The only benchmark in evaluating the success or failure of an ontology with respect to its acceptance is its longevity and the extent to which it will be

adopted by the members of the community for which it was developed. Ontolingua (Farquhar *et al.* 1996) is an environment that allows an online collaborative approach to ontology modeling, editing, and reusing.

One last source of confusion is based on the different terms that are used to denote the various ontological elements. For example, people from the area of description logics use the terms *concepts*, *roles*, and *individuals* to refer to the ontological elements, whereas other scientists employ the frame-based terminology that uses *classes*, *slots*, *facets*, and *frames*. There are many other terminologies, an overview of which is presented in Kiryakov *et al.* (2001). In this thesis, we mainly use the terminology from the object-oriented and descriptions logics paradigms. *Classes* correspond to concepts, and *attributes* or *roles* to properties of the concepts. *Objects* are instances of a class and *relations* are the various relationships that hold among different concepts.

#### 2.1.6 The Role of Ontologies in Information Systems

Ontology usage is rapidly becoming widespread in many scientific fields, such as intelligent information integration (Hakimpour and Geppert 2001; Wache *et al.* 2001; Palopoli *et al.* 2003; Rodríguez and Egenhofer 2003; Doan and Halevy 2005), information retrieval (McGuinness 1998; Guarino *et al.* 1999; Jones *et al.* 2001; Biskup and Embley 2003), similarity assessment (Mena *et al.* 1998; Rodríguez and Egenhofer 2004), electronic commerce and web retrieval (Fensel 2000; Fensel *et al.* 2001; Doan *et al.* 2003; Dou *et al.* 2003; Embley *et al.* 2005), conceptual analysis (Burg and Van de Riet 1998; Guarino and Welty 2000; Bernstein 2003), and language engineering (Lang 1991). It has also attracted the interest of communities that bear a close relationship to computer science such as GIS (Coenen and Visser 1998; Fonseca *et al.* 2002), as well as from communities that are phenomenically unrelated, such as medicine (Gangemi *et al.* 1998; Mork and Bernstein 2004) and law (Bench-Capon and Visser 1997). This thesis

focuses on the use of ontologies in GIS and information systems in general, for the purpose of retrieving semantically similar information.

An ontology-based information retrieval is based on the concept of ontological commitment, which reveals the agreement between the user querying the database and the database administrator that made the information available (Kashyap and Sheth 1998). Database administrators map objects of the databases onto ontology terms, whereas users formulate their queries using the terms of an ontology that better corresponds to their view of one specific domain. Hence, consistency is guaranteed on the vocabulary used from both sides. An ontology-based retrieval of semantically similar results exploits the structure and content of an ontology in order to derive measures of similarity among concepts. For example, in the absence of information for a class specified in the user's query, the system may search for available information on the most similar classes in the ontology with respect to the original class that was specified in the query.

## **2.2 Modeling Similarity**

Similarity is ubiquitous in psychological theory and philosophy. It has also lately become an important area of investigation for computer scientists. Attempts to answer the question of "what makes things seem alike or seem different?" (Attneave 1950) have resulted in several suggestions and theories about the nature of similarity, as well as in a number of models that try to formalize and quantify it.

### **2.2.1 Properties of Similarity**

Similarity is often interpretable as proximity, which suggests a spatial structure (Shepard 1962a). For this reason, many studies favor a geometrical approach, where the objects compared are assumed to be points in a conceptual space, and dissimilarity is equated to the distance between the points. Similarity is then derived as a monotonically decreasing

function of the distance. Since the distance function is a metric, it satisfies for all points in the space the metric axioms of identity, symmetry, and triangle inequality, which translate for similarity to the properties of minimality, symmetry, and transitivity, respectively. The validity of these properties for similarity, however, has been the subject of an ongoing debate in the literature.

#### *2.2.1.1 Minimality*

The minimality axiom captures that the self-similarity of an entity to itself is always larger than the similarity of the entity to other entities. It also implies that the self-similarity between an entity and itself is the same for all entities. Tversky (1977), the main opponent of the spatial axioms of similarity, argued that the self-similarity measure is not the same for all entities and varies depending on the prototyping characteristics of an entity inside a domain. What matters, however, for the purpose of comparing two entities is that the self-similarity is always larger than the similarity between two different entities (Krumhansl 1978). In this thesis, we accept the property of minimality under all circumstances.

#### *2.2.1.2 Symmetry*

The symmetry axiom for similarity has been most heavily attacked in the literature. It was first questioned by Rosch (1975), who diagnosed, during an experiment, that categories are formed in terms of focal points or prototypes. According to Rosch, in sentences of the kind “*a* is essentially *b*” (e.g., “a robin is a bird”) the prototype appears in the second position and the variant in the first. This positioning in turn implies that the perceived distance from the prototype to the variant is greater than the distance from the variant to the prototype and, hence, the variant is more similar to the prototype than the prototype to the variant. For example, a robin is more similar to a bird than a bird to a robin. In other words, similarity varies depending on which stimulus is chosen as the

source and which as the target. The direction of asymmetry is determined by the relative salience of the stimuli (Tversky 1977).

These findings did not go unchallenged. In a more recent study, Rada *et al.* (1989) argued that asymmetry stems from the existence of another asymmetric relationship between the stimuli, such as the class-instance relationship, rather than being an intrinsic property of similarity. Asymmetry, however, is still manifested in comparisons of stimuli that are not characterized by a class-instance relationship. For example, in an experiment Tversky (1977) conducted, people judged that China is less similar to North Korea than North Korea is to China, although both of them are instances of the class *country*. Other researchers have argued that even the asymmetry detected among two instances of the same class says nothing about the truth or falsity of the symmetry relation, but that it is only concerned with its pragmatics (Richter 1992). One suggestion is that asymmetry in this case is the result of people's tendency to consider and emphasize different features when assessing the similarity of the prototype to the variant, rather than when assessing the similarity of the variant to the prototype (Gärdenfors 2000).

On a parallel argument, Nosofsky (1991) supported the idea that asymmetric proximities can be characterized in terms of symmetric similarity together with response bias. People may have prior biases to certain responses that involve a particular entity, because this entity is highly salient in their perception or memory, easily recognizable, encoded, and attended. These properties pertain to individual entities and not to relations between the entities; therefore, they may be better characterized as biases rather than similarities. For example, one may say that an actress looks like the president, but if the actress would eventually become the president, she would become the prototype and the people compared to her would become the variants. Hence, similarity is symmetric, but there is a change in the response bias.

It appears overall that similarity judgments are not always commutative and, therefore, the symmetry axiom can hardly be accepted as a universal principle of

similarity. It seems to hold when comparing entities along a few, specific, and well-defined dimensions. It fails, however, when we perform a broad assessment of similarity between two entities that involves a comparison along an arbitrary number of not so explicitly defined dimensions and when one entity occupies a more prominent position in our perceptions than another. Hence, in this thesis we accept or reject symmetry depending on the specific task at hand.

#### *2.2.1.3 Transitivity*

The transitivity property relates similarities among three elements. Opponents of the transitivity property for similarity argue that this geometric principle does not adapt well to the cognitive task of similarity assessment. For example, Tversky (1977) argued that if Jamaica is similar to Cuba (due to their geographic proximity) and Cuba to Russia (due to their political affinity) then Jamaica must also be quite similar to Russia, a statement hard to accept. Proponents of the property countered that the phenomenal failure of the principle in such examples is due to an inconsistent use of similarity, emphasizing different features and dimensions in successive comparisons (Rada *et al.* 1989; Richter 1992). Although Tversky's argument is logically inconclusive, transitivity may not always hold from an implementation point of view. For example, adhering to the convention that a computer-produced similarity score of 0 means that two entities are not similar at all, then depending on the specifics of the implemented similarity algorithm, for three entities  $a$ ,  $b$ , and  $c$ , it could be the case that  $S(a,b), S(b,c) > 0$  but  $S(a,c) = 0$ .

#### *2.2.1.4 The Relationship of Similarity to Difference, Dissimilarity, and Distance*

*Difference*, *dissimilarity*, and *distance* are all often used as logical opposites to similarity. There are, however, subtle differences of their meanings in the psychological literature, as well as of the functional relationships that tie these concepts with similarity. Difference and similarity are, undoubtedly, closely related. Mill (1829) stated that, "distinguishing differences and similarities is the same thing; a similarity being nothing



but a slight difference.” Therefore, differences are things that people observe. Dissimilarity, on the other hand, is simply an estimate; a judgment made based on the perceived differences of two entities. This estimate is typically abstracted as the psychological (i.e., perceived) distance between the representations of the two compared entities in a conceptual space. In this sense, dissimilarity and psychological distance coincide. The dimensionality of the space is determined by the conceptually distinct features of the instances (e.g., color and size), upon which differences have been observed.

Such a geometric view implies that similarity is related to dissimilarity, and, consequently, to distance and to differences, through an inverse function. Conventional wisdom suggests that the magnitudes of the two notions are complementary (Hosman and Kuennapas 1972); that is, the similarity  $S(i, j)$  (or for simplicity  $S_{ij}$ ) between two entity instances  $i$  and  $j$  is a linear function of their psychological distance  $D_{ij}$  with a slope -1 (Equation 2.1a) (Figure 2.2a). The prominent assumption in the psychological literature, however, is that similarity is related to distance via a non-linear decay function (Gärdenfors 2000). Some researchers (Shepard 1987; Goldstone 1999) supported the idea that this function has an exponential form (Equation 2.1b) (Figure 2.2b). Shepard (1987) baptized this exponential decay as the *universal law of generalization*. Nosofsky (1986) argued instead in favor of a Gaussian form (Equation 2.1c) (Figure 2.2c). Ennis (1988) showed that under certain circumstances it is difficult to discriminate which function yields better results with respect to human similarity judgments. Finally, Shepard (1988) and Takane and Shibayama (1992) concluded that the Gaussian form is most appropriate when the observers are highly practiced (i.e., they have familiarity with the objects being compared) and the exponential form otherwise.

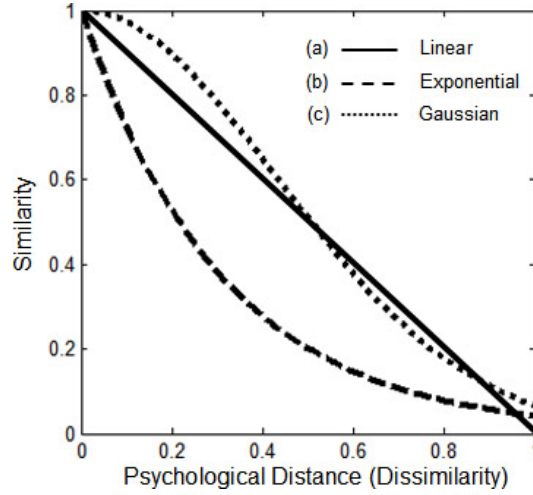


Figure 2.2: Similarity versus distance (dissimilarity) as expressed by (a) a linear, (b) an exponential, and (c) a Gaussian function.

$$S_{ij} = 1 - D_{ij} \quad (2.1a)$$

$$S_{ij} = e^{-c \cdot D_{ij}} \quad (2.1b)$$

$$S_{ij} = e^{-c \cdot D_{ij}^2} \quad (2.1c)$$

Equations 2.1b and c correspond to a family of functions, rather than a single function. The parameter  $c$  is used as a general *sensitivity* parameter to adjust the response of the functions. Regardless of the exact form, similarity has a value of 1 when the distance is zero, and decreases monotonically with the increase of distance. For the exponential and Gaussian family of functions, similarity between two entities decreases rapidly when their distance is relatively small, while it decreases more slowly when the distance is relatively large, such that it converges to, but never reaches zero. This behavior has an interesting analogy to Tobler's first law of geography (Tobler 1970), which states that “everything is related to everything else but near things are more related than distant things.” It also seems to be well-suited for the purposes of semantically-similar information retrieval: objects fairly distant from a user's query are practically of no interest, hence almost equally dissimilar, whereas objects closer to the query have a higher impact.

### 2.2.1.5 Similarity and Context

Similarity is a very flexible notion and strongly dependent on context (Goldstone 1994b). It has been argued that the flexibility that similarity exhibits is enough to doom it and that there is no such thing as overall similarity that can be universally measured. Indeed, similarity relations manifest themselves only if one has a point of view. Saying that two entities are similar means nothing, unless we define with *respect to what property or properties* they are similar (Goodman 1972; Popper 1972). Even if we delimit the scope of a comparison in this manner, there could still be present implicit or personal forms of context that influence similarity judgments. The quality of a similarity measure, however, relies critically on context, therefore, it is important to discuss how the different types of context can be captured and modeled for the purposes of information retrieval.

An *explicit context* exists when the relevant frame of reference is unambiguously identified. For example, one may ask the question of how similar two buildings are with respect to their height. The similarity between the two objects will then be evaluated only with respect to this attribute. Such a question is more specific than the question of how similar a museum is to a theater, where the two entity types compared may vary with respect to several properties. In a loose setting, this kind of similarity evaluation would be a hopelessly ambiguous task. In information systems, however, the chances for a cognitively accepted and coherent similarity measure increase through the use of ontologies. Ontologies narrow down the frame of reference by defining explicitly all the entities that may exist within a domain as well as the properties of these entities that are of interest to the domain community. In addition, ontologies eliminate cognitive heterogeneity through ontological commitment.

Another type of context is the *implicit context* often introduced by the set of stimuli under consideration. This context is responsible for several effects on the perceived distances between the stimuli. Tversky (1977) observed that people weight more heavily during a comparison those features of the stimuli that have a high diagnostic value. The

diagnosticity of a feature refers to its classificatory significance. For example, when comparing a clinic to a hospital, the property of providing health services has a small diagnostic value, because it is shared by both objects. It does, however, have a larger diagnostic value when comparing a hospital to a theater. Other effects originate from the spread and concentration of the stimuli within the conceptual space. The extension effect (Torgerson 1965; Tversky 1977) states that the addition of a new entity into the set of entities under consideration will alter the pre-existing similarity judgments. For example, assume a set of values, denoted as  $\{1,2,3,4\}$ . If we add to it the value 10, the similarities among the first four values will become larger than they were judged to be before the addition of the new value. The similarity relations that hold among the entities are different in the original and the extended context, because people tend to adjust their conceptual spaces depending on the pair of the two most dissimilar entities in the set that they have to compare. Similar effects were observed by Goldstone (1994b) and Krumhansl (1978). The latter also found evidence that similarity is sensitive to the density of the stimuli within a space. Two objects in a less spatially dense region of the stimulus domain will be judged more similar than two objects that differ an equivalent amount, but lie in a spatially denser region of the domain. For instance, if we also add the value 9 into the set of the previous example, the similarity between 9 and 10 will be judged larger than the similarity between 2 and 3. This effect implies that people attempt some form of distribution equalization, similar to the process of histogram equalization in digital imaging applications (Gonzalez and Woods 2002). They spread the objects in their perception, so that the new distribution comes closer to becoming uniform.

Although it is possible to account for such effects mathematically, from a pragmatic standpoint there is no reason to do so. The existence of these effects depends on prior observation of the stimuli and their characteristics within a domain and the ability to retain such knowledge. Information retrieval, on the other hand, is immune to such phenomena, because in databases no such sensory processes are involved and users are

typically unaware of the set of entities against which their queries will be directed. A single exception concerns occasions where the range of an attribute for a set of objects may be bounded within two extreme values that are conventionally perceived as opposites (e.g., black and white). In these cases the extension effect becomes relevant (i.e., similarities among other values must be judged relative to the extreme pair).

Other properties that are relevant for similarity comparisons may vary widely with age (Gentner 1988), expertise (Sjoberg 1972), environment (Harnad 1987), method of presentation (Gati and Tversky 1984), cerebral hemisphere of processing (Umiltà *et al.* 1978), and—most importantly for information retrieval—the individual comparison-maker’s goal and knowledge (Goldstone 1994b). All of these factors constitute a *personal context* that biases similarity estimates. Hence, even for the same set of entities and considering the same properties in the assessment, similarity judgments may vary among individuals. Although it is expected that people sharing backgrounds, interests, and experiences (i.e., the people who commit to the same ontology) will also share the same similarity assumptions and biases for the entities in a domain of interest, it is logical to expect slight deviations from individual to individual. The personal context is typically captured by letting users specify weights or other parameters in order to fine tune similarity assessments according to their needs and intentions.

#### *2.2.1.6 Similarity in Classification*

Another factor that may influence similarity judgments is classification. Similarity and classification bear a close relationship (Rips and Shoben 1973; Lakoff 1987; Rips and Collins 1993; Goldstone 1994b). People tend to group entities into clusters based on their similarities. This process also works reciprocally, that is, the existing classification will influence the insertion of an entity into a cluster. Thus, similarity arises as a consequence, but also influences classification.

## 2.2.2 Models for Similarity Assessment

A classification of models for similarity assessment distinguishes between geometric, featural, transformational, network, and alignment models. Whereas network models were mainly developed by computer scientists, the remaining models were proposed from cognitive psychologists. Besides these categories, there also exist hybrid models that combine characteristics from the other approaches. Since similarity is not a unitary concept (Torgerson 1965; Goldstone 1994b), favoring the use of one model over another depends on the specific task, because each model carries different innate assumptions and emphasizes different properties of similarity. Selecting the appropriate model becomes a critical factor in improving the quality of a similarity measure.

### 2.2.2.1 Geometric Models

Geometric models have been amongst the most prevalent approaches in analyzing similarity. In these models, the entities under comparison are represented as points within a multi-dimensional metric space. A metric space is based on a distance function. The dimensions (i.e., axes) of the space represent features or properties that the entities possess. The coordinates of a point within the space represent specific (perceived) instances on each dimension; for example, a particular temperature or a particular length. Interpoint distances are perceived as measures of dissimilarity between the entities. They are typically computed by the  $r$ -Minkowski metric (Equation 2.2), where  $n$  is the number of dimensions and  $x_{ik}$ ,  $x_{jk}$  the values of entities  $i$  and  $j$  along dimension  $k$ . For  $r=1$  Equation 2.2 yields the city-block distances between the points, whereas for  $r=2$  it produces Euclidean distances. The latter means that one travels along the dimensions in order to get from one point of the space to another. These distances indicate the dissimilarity between  $i$  and  $j$ . The role of the weight coefficient  $w_k$  is to determine the salience of a particular dimension  $k$ . If it was omitted, then the scales of all dimensions would be identical and the distance measured along one of the axes would be the same as

the distance measured along another. Such an assumption is often violated, because in certain psychological contexts several dimensions are emphasized more than others (Attneave 1950; Torgerson 1965; Nosofsky 1992). Choosing the important dimensions depends on the knowledge, purpose, and interests of users who will perform the similarity assessment. The estimated distances can be converted to similarities through any of Equations 2.1a-c, however, non-linear functions are typically the norm in psychology (Ashby and Lee 1991).

$$Dissimilarity(i, j) = d_{ij} = \left[ \sum_{k=1}^n w_k |x_{ik} - x_{jk}|^r \right]^{1/r} \quad (2.2)$$

Geometric models are exemplified by the method of multi-dimensional scaling (MDS), which was originally implemented by Young and Householder (1938) and Torgerson (1952; 1958). Its conception, however, is attributed to Richardson (1938) who suggested that psychophysical judgments, such as similarity, involve more than one dimension for their representation. Since then, various researchers have improved the method (Klingberg 1941; Messick and Abelson 1956; Kruskal 1964; Nosofsky 1992) and provided the first computerized applications of it (Shepard 1962a; 1962b).

The objective of MDS techniques is to find  $n$  points whose interpoint distances match the experimentally obtained distances (i.e., dissimilarities) of  $n$  objects. The input to MDS routines may be similarity or dissimilarity judgments between a set of objects, whereas the output is a geometric model of the data in which each object of the set is represented as a point in a  $n$ -dimensional space. The intention is to come up with the space of the lowest possible dimensionality that will accurately reflect the original distances. Therefore, MDS does not aim at estimating similarity between objects; similarity judgments are only the input to the routine. It rather aims at revealing how the conceptual spaces of people are structured in dimensions, but the dimensions *per se* have no meaning. A secondary goal of MDS is the reduction of data. Substituting  $n^2$  implicit

distances for a set of  $n$  objects (i.e., the distance from each object to every other object of the set) with a set of  $k \cdot n$  coordinates, where  $k$  is the number of the dimensions (usually much less than  $n$ ), results in a reduction of data (Shepard 1962a). MDS is appropriate in complex situations when not all of the dimensions are known *a priori* (Torgerson 1952).

The extraction of similar results from a typical relational database is concerned with the converse problem. A relational table corresponds to a set of objects and the number of attributes of the relational table defines dimensionality of the space. With each object having a different placement along the dimensions, depending on its attribute values, the goal is to exploit the differences in the values in order to derive similarity measures among the objects.

Although geometric models can be modified to account for asymmetries in similarity judgments (Krumhansl 1978; Nosofsky 1991), these models typically adopt the view of a symmetric and transitive similarity. They perform better when the entities vary along attributes of a quantitative nature (Torgerson 1965; Tversky 1977), because values of quantitative dimensions represent points in a continuum. On the other hand, qualitative dimensions have a discrete structure such that the determination of a point with respect to a qualitative dimension presents difficulties in its placement. Hence, in such situations other models must be employed.

#### 2.2.2.2 Featural Models

Such an alternative approach is based on featural models, which have a qualitative foundation. Rather than estimating similarity as a function of distance, featural models infer the similarity between two objects as a function of their common and distinctive features. Common features increase similarity, whereas different features decrease it. Jaccard (1908) first suggested a simple mathematical formula that captures these ideas (Equation 2.3). His measure, known as the *Jaccard index* or the *coefficient of similarity*, determines similarity between two entities  $a$  and  $b$  with sets of features  $A$  and  $B$ ,



respectively, as the ratio of the cardinality of the intersection of their common features  $|A \cap B|$  divided through the cardinality of the union of their features  $|A \cup B|$ .

$$S(a, b) = \frac{|A \cap B|}{|A \cup B|} \quad (2.3)$$

Featural models are exemplified by the contrast model (Tversky 1977), which is a parameterized version of the Jaccard index. In this model, the similarity between two objects  $a$  and  $b$  (Equation 2.4a) is determined by three arguments:  $|A \cap B|$  the number of features that are common both to  $a$  and  $b$ ;  $|A - B|$ , the number of features that belong to  $a$  but not to  $b$ ; and  $|B - A|$ , the number of features possessed by  $b$  but not by  $a$ . The terms  $\theta$ ,  $\varphi$ , and  $\omega$  reflect the weights given to the one common and the two distinctive sets of features, respectively. Equation 2.4a defines a family of functions depending on the form of  $f$  and the values of the weights. The function  $f$  can be modified so that a particular common or distinctive feature will receive a larger or smaller weight. Usually, however, it is simply assumed to be additive (Equation 2.4b). The most interesting variation of the contrast model is the ratio model, where similarity is normalized and has values between 0 and 1 (Equation 2.4c). All these functions are called *matching functions*, because they measure the degree to which two objects match each other.

$$S(A, B) = \theta \cdot f|A \cap B| - \varphi \cdot f|A - B| - \omega \cdot f|B - A|, \quad \text{for } \theta, \varphi, \omega \geq 0 \quad (2.4a)$$

$$S(A, B) = \theta \cdot |A \cap B| - \varphi \cdot |A - B| - \omega \cdot |B - A|, \quad \text{for } \theta, \varphi, \omega \geq 0 \quad (2.4b)$$

$$S(A, B) = \frac{f|A \cap B|}{f|A \cap B| + \varphi \cdot f|A - B| + \omega \cdot f|B - A|}, \quad \text{for } \varphi, \omega \geq 0 \quad (2.4c)$$

Feature matching is a set-theoretic approach and, hence, is neither dimensional nor metric in nature. By modifying appropriately the weights and the form of function  $f$ , the contrast and ratio models may provide asymmetric measures of similarity when this is necessary. Features may correspond to components of an object (such as roof and

balcony for a house), concrete properties (such as having a square footage and construction date), or abstract attributes (such as quality of structure). It is obvious that the term *feature* in the parlance of the contrast model denotes the value of a binary or nominal variable; therefore, featural models are preferred when the available data for a similarity assessment consist of qualitative variables rather than values of the objects that can be mapped onto quantitative dimensions.

A criticism of the featural models is that under a relatively general context two entities may share an arbitrary number of properties and hence be arbitrarily similar (Goodman 1972; Gärdenfors 2000). For example, both Iraq and the US are countries, have mountains, are places where people live, exist in the same galaxy, and so forth. On the other extreme of a very narrow context, the number of available properties that will count as common and distinctive features may be quite small for these two entities. In this case, Equations 2.4a-c will yield very coarse similarity measures; therefore, the two basic assumptions for featural models to become conceptually operational are that (1) a relatively large number of features is associated with the objects, which may include functions, parts, and properties and (2) the features employed in the similarity assessment will be selected depending on the context, as it is specified within a particular domain of interest. The second assumption is crucial, because shifts of attention to other domains will result in the selection of different features for the similarity assessment and, hence, in shifts in overall similarity judgments. According to Tversky (1977) “the selection of features is viewed as a product of a prior process of extraction and compilation.”

Such extraction and compilation results in domain ontologies, which model all the properties and features of the entities as well as the relationships that hold among them depending on the context imposed by a particular universe of discourse. These properties and relationships may be counted as common or distinctive features of objects during a similarity assessment. Hence, ontologies satisfy both assumptions of featural models.

#### 2.2.2.3 Transformational Models

Another approach to similarity is based on the concept of transformational distance. The magnitude of a transformational distance measure is expressed by the number of operations that are required to transform one object into another (Imai 1977; Jagadish *et al.* 1995). For example, the sequence *XXO* requires one atomic operation to become *XXX*, whereas *XOO* requires two operations. Hence, *XXO* is more similar to *XXX* than *XOO* is. Similarity is assumed to decrease monotonically as the number of these operations increases. Transformational models are closely related to geometric models. Traditionally, it was thought that such models apply better to figures and visual configurations. Recent efforts (Hahn and Chater 1997; Hahn *et al.* 2001; Hahn *et al.* 2003), however, have resuscitated transformational models and made them applicable in a much broader context.

#### 2.2.2.4 Models Based on Semantic Networks

Unlike geometric and featural models, network models provide explicit support for similarity assessment among hierarchically organized concepts (Sattath and Tversky 1977). The main work in this area is based on semantic networks (Quillian 1968) and dates back to the theory of spreading activation (Collins and Loftus 1975). According to Lee *et al.* (1993), “a semantic network is broadly described as any representation interlinking nodes with arcs, where the nodes are concepts and the links are various kinds of relationships between concepts.” The closer two concepts are in the network, the more they are semantically similar.

Many of the network models aim at deriving the *semantic relatedness* rather than semantic similarity of two concepts (Hirst and Onge 1998; Banerjee and Pedersen 2003; Patwardhan 2003). The former is a term originating from studies in natural language processing and corresponds to a much broader notion that encompasses the latter. Semantic relatedness refers to the degree to which two concepts are related (or not). For

example, a *theater* is related to an *actor* because actors perform in theaters. Even concepts that are antonyms can be related to each other in this sense. Semantic similarity, on the other hand, is interpreted in this work as a measure that reflects the usefulness and suitability of a result to a user's query. Semantic similarity is, therefore, only a special case of semantic relatedness (Resnik 1995). This distinction is important, because semantic relatedness measures are inappropriate for measuring similarity.

The most basic network models are based on edge-counting techniques. The idea is straightforward: the shorter the path between two concepts, the more similar they are. Even such a simplistic measure has been found to perform surprisingly well with respect to people's judgments of similarity (Budanitsky 1999). Better results were obtained for networks that consider only *is-a* hierarchies (Figure 2.3) and where the concepts were restricted to a particular domain of interest, which ensures a relative homogeneity of the hierarchy (Rada *et al.* 1989). Both requirements are met by domain ontologies, therefore, this simple measure is a good candidate for such structures.

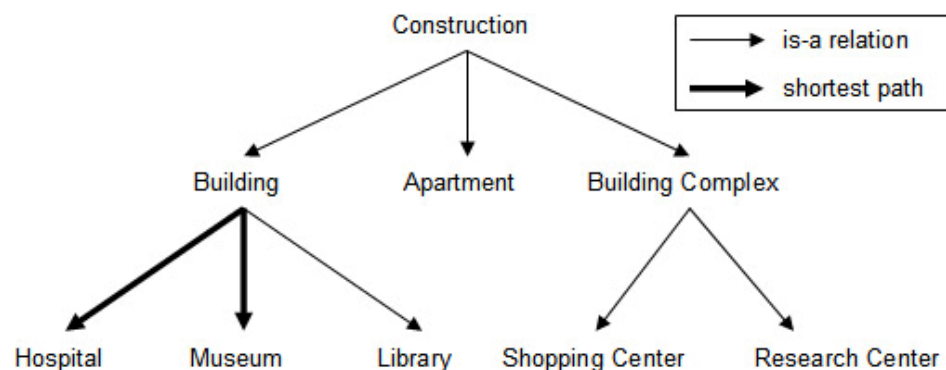


Figure 2.3: Shortest path and is-a relationships in a hierarchical network structure.

Leacock and Chodorow (1998) followed this edge-counting technique, but counted nodes instead of edges. Therefore, the distance for two synonyms is 1, rather than 0. Their measure was applied to measuring the similarity of nouns in WordNet. WordNet has several separate noun hierarchies, which were all combined into a single hierarchy by placing an imaginary root node on top, to ensure the existence of a path between any two

concepts. They converted their measure of distance to similarity with Equation 2.5, where  $c_1$  and  $c_2$  are the compared concepts,  $d(c_1, c_2)$  expresses the length of the shortest path between  $c_1$  and  $c_2$  in terms of the nodes counted, and  $D$  is the maximum depth of the WordNet hierarchy (also known as *height* in graph theory). Despite its simplicity, a problem with the edge-counting technique is the erroneous assumption that links in the hierarchy represent uniform distances. In a realistic scenario, the distances in a hierarchy shrink as one descends in depth, because the classifications are based on finer details. Another factor that is neglected is the density of concepts in the hierarchy. It is expected, that concepts in a dense part of the hierarchy should be ranked as conceptually closer than those in a sparser region.

$$S_{LC}(c_1, c_2) = -\log\left(\frac{d(c_1, c_2)}{2 \cdot D}\right) \quad (2.5)$$

To account for these additional factors, several researchers suggested other approaches (Sussna 1993; Wu and Palmer 1994). In one of them, Resnik (1995; 1999) combined the hierarchical structure of WordNet with the information content of concepts in order to derive similarity. His assumption was that the similarity of two concepts  $c_1$  and  $c_2$  is expressed by the information that they share, which is indicated in an *is-a* hierarchy by the information content (IC) of a concept  $lcs(c_1, c_2)$  that is the least common subsumer of  $c_1$  and  $c_2$  (Equation 2.6). According to Information Theory (Shannon 1948), the information content of a concept  $c$  is equal to  $-\log p(c)$ , where  $p(c)$  is the probability of the occurrence of  $c$  in a large text corpus (Ross 1976). This formula implies that the probability of a concept's occurrence in a corpus increases as the concept's informativeness decreases; therefore, abstract concepts are less informative than more concrete ones. For example, the information content of the concept *building* is less than the information content of more specific concepts, such as *hospital* and *schools*. The problem with this approach is that it underestimates the role of the hierarchical structure, which is used only for locating the immediate common superordinate of the compared

concepts. The measure depends completely on the information content of this lowest common subsumer, but the concepts themselves are not taken into consideration. Hence, in terms of semantic similarity, pairs of concepts that have the same lowest common subsumer are indistinguishable.

$$S_R(c_1, c_2) = IC(lcs(c_1, c_2)) \quad (2.6)$$

To address these limitations Jiang and Conrath (1997) developed a more sophisticated model that combines features from information content and from edge counting. Their measure is a distance measure (Equation 2.7a), but it can also be converted to a similarity measure by inverting the value of distance (Equation 2.7b) (Patwardhan 2003). Based on the same considerations, Lin (1998) proposed another similarity measure that uses the same constructs, but combines them differently.

$$d_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \cdot IC(lcs(c_1, c_2)) \quad (2.7a)$$

$$S_{JC}(c_1, c_2) = \frac{1}{d_{JC}(c_1, c_2)} \quad (2.7b)$$

An exhaustive survey of the majority of network similarity models can be found in Budanitsky (1999) and an evaluation of their performances in Budanitsky and Hirst (2001) and Patwardhan (2003).

#### 2.2.2.5 Integrated Approaches—The Matching Distance Model

An integrated approach to semantic similarity among concepts is the Matching Distance model (Rodríguez *et al.* 1999). This model combines elements from featural and network models by considering the number of common and different features of two classes along with their semantic distance in an ontology. The semantic relations used are hyponymy (is-a) and meronymy (part-whole). Features of a class are subdivided into attributes, parts, and functions. The focus is specifically on spatial concepts such as *building*, *highway*, and *park*. The spatial entities and their features, which were extracted from the

Spatial Data Transfer Standard (SDTS) (USGS 1998), were organized hierarchically based on their network representation in WordNet (Miller *et al.* 1998).

Similarity measures are obtained from Equation 2.8a, where the coefficients  $\omega_p$ ,  $\omega_f$ , and  $\omega_a$  represent weights. The global similarity function  $S(c_1, c_2)$  of two classes  $c_1$  and  $c_2$  is, therefore, a weighted sum of the similarity values for parts, functions, and attributes of two classes, denoted respectively as  $S_p(c_1, c_2)$ ,  $S_f(c_1, c_2)$ , and  $S_a(c_1, c_2)$ . Each of these values is evaluated separately by a formula based on Tversky's ratio model (Equation 2.8b).  $C_1$  and  $C_2$  are the respective sets of features of type  $t$  (parts, functions, attributes) for classes  $c_1$  and  $c_2$ , and  $|C_1 \cap C_2|$  and  $|C_1 - C_2|$ ,  $|C_2 - C_1|$  denote the cardinality of common and distinctive features, respectively. The coefficient  $a$ , is a function of the semantic distance between the two classes in the hierarchy, as well as of their distance to the immediate class that subsumes both of them (Equation 2.8c).

$$S(c_1, c_2) = \omega_p \cdot S_p(c_1, c_2) + \omega_f \cdot S_f(c_1, c_2) + \omega_a \cdot S_a(c_1, c_2) \quad (2.8a)$$

where,

$$S_t(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + a(c_1, c_2) \cdot |C_1 - C_2| + (1 - a(c_1, c_2)) \cdot |C_2 - C_1|} \quad (2.8b)$$

and

$$a(c_1, c_2) = \begin{cases} \frac{d(c_1, lcs)}{d(c_1, c_2)} & d(c_1, lcs) \leq d(c_2, lcs) \\ 1 - \frac{d(c_1, lcs)}{d(c_1, c_2)} & d(c_1, lcs) > d(c_2, lcs) \end{cases} \quad (2.8c)$$

The MD model presents several desirable properties. One important characteristic is that the formula that evaluates the coefficient  $a$  may account for an asymmetric evaluation of entity classes located at different levels in the hierarchical structure. Although such asymmetric scores are somewhat artificially generated, the model has been found to scale well with people's judgments of similarity (Rodríguez 2000). The assignment of different weights to the attributes, functions, and parts achieves context

flexibility. The consideration of the linguistic concepts of synonymy and polysemy (same word with multiple meanings) allows counting synonymous features as common, rather than as distinctive elements in the similarity assessment. The inclusion of meronymy relations and the consideration of parts in the assessment emphasize the spatial character of the model.

#### *2.2.2.6 Alignment Models and Configuration Similarity*

A configuration, such as a spatial scene, is a structurally rich description that comprises a collection of objects arranged in a specific manner. Geometric, featural, and network models cannot be readily applied to the task of configuration similarity assessments, because they rely on comparisons of isolated object (or concept) pairs and their attributes. Due to the multiplicity of objects, a similarity assessment between two configurations appears to be possible only after their objects have been placed in correspondence. The presence of a structure, encoded in the relationships that objects have with one another, also dictates that the quality of a match between two scenes is determined by the combined coherence of the correspondences created for objects and relations. Hence, the matching process should be governed interactively by both components. The dichotomy of a configuration into objects and relations suggests further that both of them should contribute to the similarity score between two such relational structures. Therefore, there is a need to assess and combine the similarity of the individual components within the scope of the more general comparison.

The validity of these intuitive claims and observations, as well as additional insights, can be traced back to psychological research. Goldstone *et al.* (1991) concluded that people construe spatial scenes in terms of objects and relations and both components were psychologically salient. Markman and Gentner (1993) discovered that when asked to assess the similarity of configurations, subjects preferred structurally sound object correspondences. Analogous findings were also reported by Goldstone (1994a) in a series



of experiments whose purpose was to evaluate the role of relations in scene similarity judgments. Based on previous work on analogical reasoning (Gentner 1983), he proposed an *alignment* model of similarity, where part of the comparison is to determine how elements correspond to, or align with, one another. Goldstone also elaborated further on the rationale that drives the creation of such correspondences. The most significant findings of his and of the other research efforts can be summarized as follows:

- *Finding 1:* People start scene comparisons by locating possible object-matches—whether exact or sufficiently similar—across two scenes. Very dissimilar objects are ignored rather than forced to fit (Aisbett and Gibbon 1994). Once the candidates for matching have been established, the process of object association takes place.
- *Finding 2:* Object association was done so as to also cause relations to be placed in correspondence. Subjects were reluctant to match similar objects that entailed correspondences of dissimilar relations (Markman and Gentner 1993).
- *Finding 3:* As the similarity between objects and relations gradually decreases (i.e., as the compared scenes start exhibiting large differences), subjects become confused, failing to report consistent rankings of similar data scenes to the input scene (Goldstone 1994a). This finding is analogous to those of Shepard (1987) and Nosofsky (1986) who formulated respectively the exponential and Gaussian functions that relate psychological distances to similarities for pairs of objects (Equations 2.1b and c). As in the case of objects, very different scenes become practically irrelevant and, in a sense, completely dissimilar to the query scene.
- *Finding 4:* When several mappings between objects and relations are possible, subjects choose the one that optimizes the overall fit (i.e., the one that maximizes similarity). While all previous findings are more or less inline with Ockham’s razor, this finding is also particularly reminiscent of the principle of *minimal change*, a

criterion often used in the field of default reasoning in order to revise a knowledge base's beliefs (i.e., internal logical propositions) about an application domain.

From a computational point of view, the problem of retrieving similar configurations has many commonalities with the problems of exact and inexact scene matching, which have been extensively studied in computer vision and pattern recognition (Shapiro and Haralick 1981; Ballard and Brown 1982). In these disciplines, configurations are interpreted as constraint systems. Typically, such systems are over-constrained; hence, most approaches relax the original constraints, retrieve solutions that satisfy the relaxed description, and rank them according to their similarity to the original scene. Retrieval is performed by algorithms operating on the graph representations of the scenes to be matched. Typically, scene-matching tasks translate to hard combinatorial problems of exponential complexity. Efforts from the community of multimedia databases adopted these techniques, but tried to incorporate aspects of domain knowledge in the process so that some of the complexity is reduced. For instance, image and video retrieval techniques focus primarily on aspects of visual content, that is, properties such as color, shape, and texture (Flickner *et al.* 1995; Santini and Jain 1996).

A number of proposals have also emerged within the context of spatial databases, albeit without much concern for the psychological findings that were outlined, and often based on simplifying assumptions that prevent their wider applicability. Some approaches for instance, create an easier version of the problem by neglecting the relational component during the matching process (Blaser 2000; Wang *et al.* 2004). Object pairs are formed so that object-to-object similarities are maximized, but this criterion alone does not necessarily yield the fittest assignment had the similarity of the relations been considered as well. Conversely, other approaches focus on the relational component, but underestimate or do not provide explicit treatment for the object component (Papadias *et al.* 1999b). Paiva (1998) addressed the problem from the perspective of topological equivalence, rather than similarity. Bruns and Egenhofer (1996) developed a systematic

methodology for constraint relaxation, measuring dissimilarity as the number of discrete gradual changes required to transform one scene to another. They assume, however, that object correspondences are known *a priori*. Furthermore, it is impossible to reason about the best match when several scenes require the same number of atomic changes in order to be transformed to the query scene. A number of methods are based on variations of *2D strings*, which encode object arrangements on each dimension using sequential structures (Lee and Hsu 1992; Chang and Jungert 1996; Papadias and Delis 1997). These methods restrict expressiveness since they rely on a restricted set of relations. Moreover, users are forced to specify queries by the schema of the relations according to which 2D strings are built.

A commonly encountered simplifying assumption relates to the size of the query and the database scenes to be compared. An uncompromising technique should allow for an arbitrary number of objects in both scenes. This ideal is rarely the case, however. Instead, it is usually hypothesized either that the compared scenes have the same number of objects (Gudivada and Raghavan 1995; Nabil *et al.* 1996) or that the number of objects is relatively small (i.e., fewer than ten) (Petrakis and Faloutsos 1997; Li and Fonseca 2006). Sometimes this difficulty is not explicitly stated, but the limitation practically applies due to the huge computational cost introduced when such techniques generalize to scenes of arbitrary sizes (Stefanidis *et al.* 2002).

In a series of papers, Papadias and colleagues improved on most of these issues (Papadias *et al.* 1998a; Papadias *et al.* 1998b; Papadias *et al.* 1999a; Papadias *et al.* 1999b). They are concerned with the efficient implementation of traditional constraint satisfaction algorithms, such as backtracking, forward checking, and branch and bound techniques (Kumar 1992). Their methods, which are customized to exploit R-trees (Guttman 1984) or similar indexing variants used in spatial databases, achieve significant performance gains. However, introducing thorough relaxation policies for spatial relations or objects is beyond the scope of their work. Some of these algorithms also

require a total matching for all objects, thereby dismissing incomplete but possibly useful solutions based on partially matched substructures of the compared scenes. In more recent work, the same team of authors considered using approximate algorithms (Papadias *et al.* 1999c; Papadias 2000; Papadias *et al.* 2003), which minimize retrieval time, but do so at the expense of several factors such as: (1) usability: users may often need to fine tune many of the algorithm's parameters (2) quality of output: the retrieved results cannot be guaranteed to be optimal and (3) quantity of output: some approximate algorithms retrieve a single match during each retrieval cycle.

The tremendous complexity of the scene-matching problem justifies many of the limitations that characterize previous efforts. Undoubtedly, some of the restrictions are an inevitable product of the exponential complexity inherent to the nature of the problem. Others arise, however, due to an underestimation of the problem's dimensions, neglect of provisions for accommodating different retrieval scenarios, and failure to incorporate geospatial domain knowledge and requirements into the approach. This thesis introduces a systematic methodology that considers such aspects and improves on previous work by addressing many of the difficulties of scene similarity assessments.

### **2.3 Mathematics for Similarity**

It is quite tempting and oftentimes useful to substitute ill-defined similarity and its derivative processes with compatible—to some extent—axiomatic theories. Fuzzy sets (Zadeh 1965) comprise such a theory. Considered by many a *lingua franca* for applications involving uncertainty, it provides for similarity what could be called, a formal coat. Being inherently vague, similarity finds a natural expression in fuzzy set theory, because many of its basic ideas and inference mechanisms can be elegantly captured through fuzzy concepts and operations, respectively. This formal coat however, may not always fit perfectly. The expressive plurality of fuzzy set theory can easily lead to unintended correspondences and produce outcomes that distort, rather than reflect,

human similarity perception. A presentation of the fundamental concepts of fuzzy set theory is, therefore, instrumental to guiding the correct correspondences between the two fields of science in the following chapters. The relationship of graph theory to similarity is also vital, but distinct. Graph theory (Harary 1969) provides a powerful layer of abstraction onto which spatial objects, spatial relations, and their attributes can be mapped. Formulating an often difficult and obscure similarity assessment through a graph-theoretic equivalent abstraction allows one to exploit the vast arsenal of algorithms and methodologies that have been developed in the graph domain in order to solve the original similarity problem. Informal references to graph and set theory concepts were already made in previous parts of this chapter. Here, we provide formal definitions for these and other concepts that are used throughout the remainder of the thesis.

### 2.3.1 Fuzzy Set Theory

Fuzzy set theory and fuzzy logic constructs are based on a generalization of their classic counterparts. Classic set theory considers elements of a domain as either members or nonmembers of a set. From this view, classic sets are *crisp* sets. The intersection  $A \cap B$  of two crisp sets  $A$  and  $B$  is the set containing only their common elements, their union  $A \cup B$  is the set containing all elements that belong to either  $A$ , or  $B$ , or both  $A$  and  $B$ , and the *relative complement* of  $A$  with respect to  $B$ , denoted by  $B - A$ , is the set comprising all members of  $B$  that are not also members of  $A$ . If  $B$  is the universal set  $U$ , then the complement of  $A$  in  $U$  is called the *absolute complement* or simply *complement* of  $A$  and denoted by  $\bar{A}$ . A *crisp relation* between  $n$  sets represents the presence or absence of association or interaction between the elements of the sets. Each crisp relation is a subset of the Cartesian product of the sets involved in the relation and can be written as a set of ordered tuples or more conveniently as a  $n$ -dimensional array. Each element of the first dimension of this array corresponds to one member of the first set, each element of the second dimension to one member of the second set, and so on. Crisp relations have a *characteristic function*, which assigns a value of 1 to every tuple of  $U$  belonging to the

relation and 0 to every tuple not belonging to it. This binary rationale follows traditional logic where conjunctive, disjunctive, and negating statements can be either true or false.

Extending the idea of a crisp set, a *fuzzy set*  $X$  is defined by assigning to each element in the universe of discourse  $U$  a value from the real interval  $[0,1]$ . This grade represents that element's membership to the fuzzy set and corresponds to the degree to which the element is similar or compatible to the concept represented by the *fuzzy set* (Klir and Yuan 1995). The function that performs this assignment is called the membership function  $\mu_x$  of a fuzzy set  $X$ , symbolized as  $\mu_x : U \rightarrow [0,1]$ .

In the same paradigm, fuzzy relations allow for various degrees of association or interaction among elements; therefore, the characteristic function of a fuzzy relation allows for degrees of membership of tuples in the relation. Thus, a fuzzy relation is typically represented as a  $n$ -dimensional membership array whose entries correspond to  $n$ -tuples in the universal set and each entry takes a value in the interval  $[0,1]$ . Of primary interest are the types of *fuzzy equivalence* and *fuzzy compatibility* relations. A fuzzy equivalence or *similarity* relation is a generalization of the well-known crisp equivalence relation, which is reflexive, symmetric, and transitive. A *fuzzy compatibility* relation is similar to a *similarity* relation, with the difference that it is not transitive.

The fuzzy theory concepts of intersections, unions, and complements correspond to the three fundamental scoring rules for conjunctions, disjunctions, and negations (Equations 2.9a-c), respectively in fuzzy reasoning. These functions are the most commonly used; however, a broad class of functions qualifies for the task of describing these operations. As Klir and Yuan (1995) point out, "since the fuzzy complement, intersection, and union are not unique operations, different functions may be appropriate to represent these operations in different contexts. The capability to determine appropriate membership functions and meaningful fuzzy operations in the context of each particular application is crucial for making fuzzy set theory particularly useful."

$$(A \cap B)(x) = \min[A(x), B(x)] \rightarrow \mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\} \quad (2.9a)$$

$$(A \cup B)(x) = \max[A(x), B(x)] \rightarrow \mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\} \quad (2.9b)$$

$$\bar{A}(x) = 1 - A(x) \rightarrow \mu_{\bar{A}}(x) = 1 - \mu_A(x) \quad (2.9c)$$

These concepts make apparent that similarity and fuzzy set theory demonstrate strong connections, because similarity is the basic idea underlying fuzzy set theory. When determining the similarity of several entities to a reference entity, it is of little use if not absurd to distinguish between those that are similar to it and those that are not. It is rather desirable to have a gradual transition among the entities, going from the most to the least similar. Therefore, the set of similar values to a reference value is a fuzzy set. Furthermore, similarity and fuzzy set theory are even more interrelated, because, in essence, the degree of membership in any fuzzy set can be interpreted as a measure of similarity. This measure expresses how similar or compatible an element of the set is to the basic concept that defines the set, whether that concept is vague (e.g., “*far*”) or crisp (e.g., “*3km*”). Thus, the retrieval of similar results can be viewed as a fuzzification of the classical information retrieval process that was until recently based on exact matches.

### 2.3.2 Graph Theory

A graph  $G = (V, E)$  of  $V$  nodes (or vertices) and  $E$  edges (or arcs) represents a structure, consisting of a set of elements related in a specific way. The *size* or *order*  $|V|$  of a graph  $G$  is defined as the number of vertices in  $G$ . An edge from node  $i$  to node  $j$  is said to *cover* or to *be incident* to these nodes and is represented as  $(i, j)$ . Conversely, the nodes are termed *adjacent*. An edge  $(i, i)$  that connects a node to itself is a *loop*. For multiple vertices, edges and loops generalize to *paths* and *cycles*: A *path* between two nodes  $u$  and  $v$ , is simply a non self-intersecting sequence of edges of the form  $(u, i_1), (i_1, i_2), \dots, (i_k, v)$ . When such a path exists, the nodes  $u$  and  $v$  are *connected*. A *cycle* is a path  $(u, i_1), (i_1, i_2), \dots, (i_k, u)$  containing at least one arc in which no node except  $u$  is repeated.

Based on these simple definitions it is possible to define several different kinds of graphs. A *complete* graph of  $n$  vertices, denoted by  $K_n$ , is a graph in which any two of its nodes are adjacent (Figure 2.4a). A *connected* graph has all pairs of nodes connected by a path of edges. A *directed graph* or *digraph* is a graph in which edges may be ordered pair of vertices, giving the direction from one vertex to another (Figure 2.4b). A *multigraph* is a graph or digraph with multiple edges between the same vertices, whereas a *pseudograph* is a *multigraph* that also contains loops (Figure 2.4c). Graphs containing additional information attached to their edges in the form of numerical or symbolic values (Figure 2.4d) are termed *labeled graphs* or *attributed relational graphs* (ARGs) (Ambler *et al.* 1973). An important subclass of ARGs are *weighted graphs*, which consist of a graph together with a function  $w$  from  $E$  to  $\mathbb{Z}$  or  $\mathbb{R}$ . The weight of an arc  $(i, j)$  can be denoted by  $w_{ij}$  or  $w(i, j)$ . A graph  $G$  is called *planar* if it can be drawn so that its nodes are points in the plane and each arc  $(i, j)$  is drawn so that it intersects no other arcs and passes through no other nodes except the ones that it covers. Otherwise, the graph is called *non-planar*. In a *bipartite graph* the vertices are partitioned into two disjoint sets  $A$  and  $B$  such that no two nodes in  $A$  or  $B$  are adjacent (Figure 2.4e). If  $A$  has  $a$  elements and  $B$  has  $b$  elements, the complete bipartite graph is denoted by  $K_{a,b}$  (Figure 2.4f).

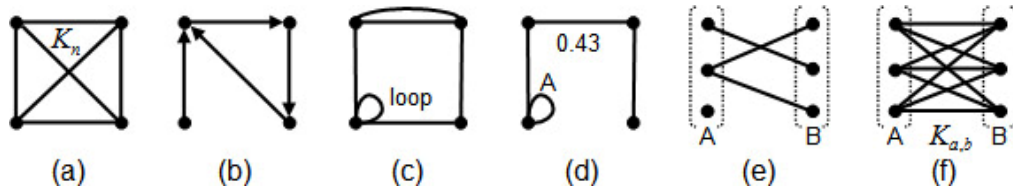


Figure 2.4: Various types of graphs: (a) complete graph, (b) digraph, (c) pseudograph, (d) ARG, (e) bipartite graph, and (f) complete bipartite graph.

A *matching* in a graph  $G = (V, E)$  is a subset  $M$  of the edges  $E$  such that no two edges in  $M$  share a common vertex (Figure 2.5a). Vertices that remain unmatched are called *free* or *exposed* vertices, whereas those that are incident to a matching edge are called *matched* or *covered*. A *maximum cardinality matching* is a matching with the maximum



number of edges. If the edges of the graph have associated weights, then a *maximum weight matching* is a matching for which the sum of the edge-weights is a maximum (Figure 2.5b). When the weights assume only positive values, then the maximum weight matching is always a maximum cardinality matching.

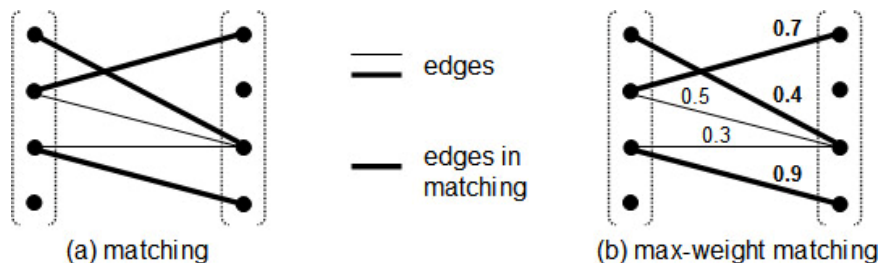


Figure 2.5: Matchings in bipartite graphs: (a) a simple matching and (b) a maximum-weight matching.

Before presenting additional graph concepts, a definition of the notion of *maximality* is required. A *power set*  $P(S)$  of a set  $S$  is the set of all subsets of  $S$ . The cardinality of the power set is  $|P(S)| = 2^n$  where  $n = |S|$  (the empty set  $\emptyset$  is also an element of  $P(S)$ ). Each element  $A$  in  $P(S)$  is a set. We say that  $A \in P(S)$  is minimal if there is no other set  $T \in P(S)$  such that  $T \subset A$ . Similarly, we say that  $A \in P(S)$  is maximal if there is no other  $T \in P(S)$  such that  $A \subset T$ . For example, the power set of a set  $S = \{1, 2, 3\}$  is  $P(S) = \{\{1, 2, 3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1\}, \{2\}, \{3\}, \emptyset\}$ , with  $\{1\}$  and  $\{1, 2, 3\}$  being examples of minimal and maximal elements, respectively.

A graph  $G' = (V', E')$  is called a *subgraph* of the graph  $G = (V, E)$  if  $V' \subseteq V \wedge E' \subseteq E$ , and a *proper subgraph* of  $G$  if  $V' \subset V \vee E' \subset E$ . If  $V' \subseteq V$  then the *subgraph of  $G$  induced by  $V'$*  has the node set  $V'$  and all edges  $(u, v)$  in  $E$  such that both  $u$  and  $v$  are in  $V'$ . A *complete subgraph* of  $G$  is called a *clique* and a maximal complete subgraph of  $G$  is called a *maximal clique*. A distinction is required between maximal and *maximum* cliques. Whereas a maximal clique is not a proper subset of any other clique, a maximum clique is a clique with largest cardinality. It follows that every maximum

clique is also maximal, but the converse does not always hold. The clique number of  $G$ , denoted by  $\omega(G)$ , is the size of the maximum clique. In the case of weighted graphs, the *maximum-weight clique* is the clique with the largest weight. The maximum-weight clique is always maximal, but it does not necessarily have the largest cardinality among other maximal cliques. A disconnected graph can be divided into *connected components*. A *component* is more formally defined as a maximal connected subgraph (i.e., it is not a subgraph of any other connected subgraph of the graph).

Figure 2.6 provides a comprehensive visualization of these concepts. The graph  $G$  consists of two connected components,  $A$  and  $B$ . The maximum (and maximal) clique is  $\{b, c, f, e\}$  with size 4. The maximum-weight (and maximal) clique is  $\{c, d, f\}$  with total weight 2.1. There is a total of five maximal cliques and twenty-one non-maximal cliques in the graph. For clarity, only four non-maximal cliques are shown.

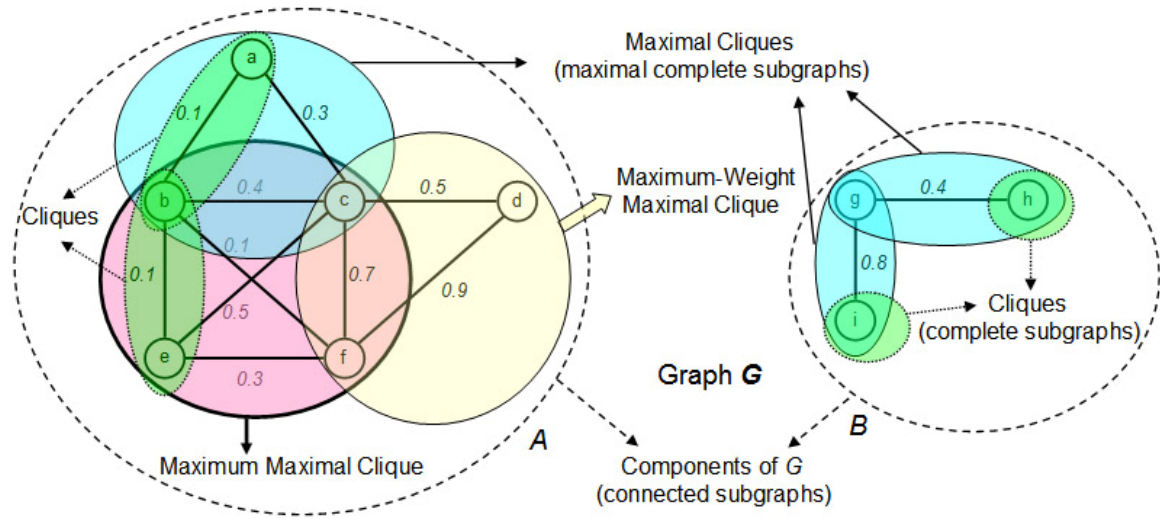


Figure 2.6: Demonstration of the concepts of component, maximum clique, maximum-weight clique, maximal clique, and clique.

Two graphs  $G$  and  $H$  are *isomorphic* if there exists a bijective mapping  $f$  between the vertices in  $G$  and the vertices in  $H$  such that the number of edges joining any two vertices in  $G$  is equal to the number of edges joining the corresponding two vertices in  $H$ ; that is,  $G \cong H$  iff  $f : G \rightarrow H : \forall (u_i, u_j) \in G \exists (f(u_i), f(u_j)) \in H$ . Informally, two graphs are

isomorphic if they contain the same number of vertices connected the same way. A *labeled graph* or *constrained isomorphism* also introduces the requirement that the bijective mapping is performed among edges of the same kind (i.e., the labeling of arcs and nodes must also be equivalent).

## **2.4 Summary**

Important characteristics of similarity are its often asymmetric behavior, its non-linear relationship to dissimilarity, and its dependence on various forms of context. Fuzzy set theory provides theoretical tools, which help model the complex behavior of similarity and complement traditional psychological models for similarity, such as geometric, featural, and network models. Geometric and featural models for similarity assessment have usually compared entities based on their quantitative and qualitative features, respectively, whereas network models consider the semantic relations among entities. These psychological models often need to rely on ontologies, which are explicit and axiomatized specifications of the vocabulary used to describe concepts and properties within a domain of interest. Ontologies provide a hierarchically organized structure of concepts that can be employed by a network model to assess similarity. In addition, they model qualitative features and properties of the entities that may be used as common or distinctive features by a featural model during a similarity comparison. Similarity comparisons of spatial scenes introduce additional requirements that traditional models cannot handle. Such comparisons are typically performed by specialized computational implementations that operate on the graph representations of the scenes. Graph theory concepts and algorithms can be thus exploited to provide more intuitive and efficient scene similarity assessments.

## CHAPTER 3

### SEMANTIC SIMILARITY AMONG ATOMIC ATTRIBUTE VALUES

Similarity among attribute values forms the foundation of the similarity framework developed in this thesis. Similarity is measured among atomic values of homogeneous entities that belong to a centralized or distributed GIS managed by a single DBMS. Entities could be either objects or relations. The term *homogeneous* implies that the entities conform to a common database schema, thus sharing attribute names and domains for each attribute. Numerical values are always expressed in the same units for each attribute or can be easily converted. Furthermore, the meanings of the same attribute names correspond to the same concepts in the universe of discourse for every entity. We also assume that there is no cognitive heterogeneity (Bishr 1998) among users of the database, meaning that they all interpret in the same way the concepts expressed by the attribute names and by the attribute values. If the database subscribes to a domain ontology, this assumption implies that all its local users also subscribe to the same ontology. In such a homogeneous environment, entities differ from each other only with respect to their attribute values, which assign to them specific qualities or quantities.

The approach to attribute-level similarity assessments consists of determining the nature of each individual attribute and discussing algorithms appropriate to resolve similarity for its values. Due to the different types of attributes that may exist in a GIS, we do not limit the approach by complying with specific similarity algorithms, but employ different models and accept different properties of similarity depending on the particular attribute type. The list of operations for similarity assessment is not exhaustive, but rather aims at creating a repository of well-defined operations that may be used as is or with slight modifications for the plethora of attributes typically encountered in databases. In support of similarity queries, a set of methods for reasoning over null values is developed. Although this thesis focuses on geographic attributes (USGS 1998), the

ideas developed in this chapter merit generic application since the main attribute types that are examined are common across all general-purpose information systems.

### **3.1 Similarity versus Change**

Retrieving and ranking similar results to a query on a single attribute is sometimes a simple task. An order of results to a query on a numerical attribute, for instance, could be determined through an operation as elementary as counting distances along the attribute scale. Values closer to the query would be more similar than values further apart. Similarity judgments at the attribute level, however, are prerequisites to inferring the similarity among higher-level representations, such as objects, relations, and spatial scenes. Hence, the choices involved in the quantification of similarity or dissimilarity at the attribute level will have a profound impact on the quality of the results obtained at the object and scene levels. Out of a set of alternative methods that work equally well at the attribute level (i.e., they produce identical ranks) the method deemed appropriate should be the one that ensures the scalability and coherence of the similarity framework as we ascend the levels. To avoid compromising the overall framework, the quantitative estimates must be derived based on a well-defined rationale that also takes into consideration psychological aspects of similarity.

The issue that must be resolved first is to understand what is being measured, or what exactly a quantitative similarity value represents. Failure to answer this question will render the measures devoid of significance (Caws 1959). Geometric, featural, and network models of similarity do not provide a clear answer to this question. Since most of these efforts originate from psychological studies, the usual approach is to hypothesize a model, conduct a series of experiments, and evaluate the goodness of fit between the outcomes of the model and the human judgments of similarity. A good performance of the models corroborates their validity for a specific domain, but does not elucidate what is being measured in the particular domain and why the models are valid. The tacit

assumption is that the models simulate psychological distances in people's mental spaces, but this claim does not answer the initial question. Instead, it shifts our effort in defining what psychological distances are and comprehending how humans arrive at their formation, both being issues open to interpretation; therefore, such models constitute *ad-hoc* methodologies—some performing better, some worse—because they lack a unifying conceptual base. Establishing such a base would provide a more basic and fundamental concept of similarity, able to glue competing alternatives under a set of primitive operations (Quine 1969). Furthermore, it would explain what current similarity models measure, thus providing the ability to make critical remarks on their performance or to suggest improvements.

Similarity is a relation between two things with respect to one or more perspectives (e.g., attributes). Chapter 2 emphasized that definitions and understandings of the similarity relation vary from researcher to researcher and from discipline to discipline (Holt 1999). Therefore, it seems appropriate to start the inquiry with a definition that leaves little room for dispute. Such a definition is provided by Bruns and Egenhofer (1996) who define similarity as “the assessment of deviation from equivalence.” Undoubtedly, equality is the one extreme of a similarity relation, since equal things are, in a way, totally similar. Any deviation from equivalence implies differences; therefore, an *assessment of deviation* means the assessment of differences, which is not surprising, since the term *similarity* frequently rides tandem with the term *difference*. It follows that when the differences between a pair of entities are equal to those of another pair, the similarity scores of the two pairs of entities should also be equal as well. Therefore, a successful measurement of similarity relies on the appropriate measurement of differences.

One way to assess the differences between two entities is in an absolute fashion. For instance, if two entities are compared with respect to their length, the absolute difference is the absolute value of the algebraic difference of their lengths. Similarly, in a set-based

interpretation, the absolute difference between two concepts—each associated with a set of features—is the cardinality of the symmetric difference of the sets. Despite much evidence to the contrary (Rosch 1975; Tversky 1977), a dissimilarity measure based on the absolute assessment of differences will always produce symmetric similarity measures and may frequently lead to counter-intuitive results. Consider, for instance, the example of the spatial scene query in Figure 3.1. If dissimilarity is calculated as absolute difference, then both database scenes will be judged equally similar to the query by the system. It should be evident, however, that Scene B is a better result, because the difference between the larger matched segments is very small compared to the actual length, whereas in Scene A the smaller street segment must double to coincide with the corresponding small segment in the query. Therefore, the difference between 10 and 20 does not mean the same thing as the difference between 1000 and 1010.

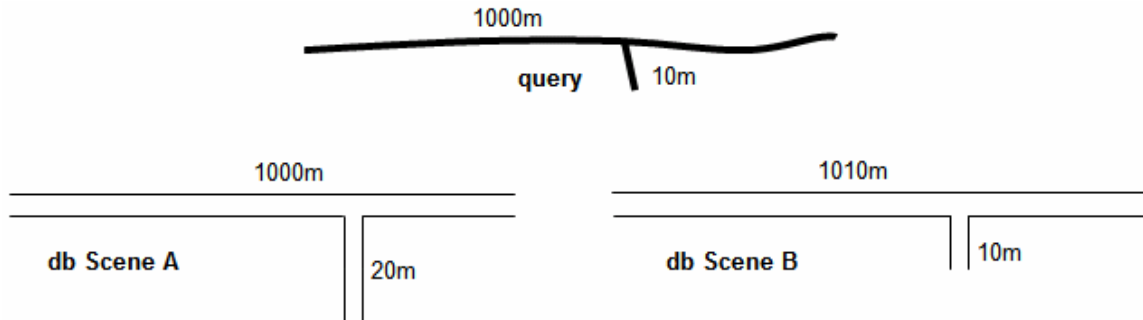


Figure 3.1: Assessing similarity based on absolute differences.

These difficulties can be alleviated if the deviation from equivalence is assessed based on *relative difference* or *change*. Adopting this paradigm implies that during a comparison between two objects we attribute to them the same identity and perceive one as the changed version of the other; that is, an entity retains its identity while altering in some respect. This assumption underlies semantic information retrieval, where an approximate match is a surrogate for an exact match. It represents an informed guess by the system for the item that we are looking for, only slightly changed. This assumption is also precisely what the phrase *deviation from equivalence* suggests. In this context,

change and dissimilarity become equivalent. A measure of dissimilarity expresses the degree of change that one entity must undergo in order to become identical to the entity that it is being compared. The similarity of the entities can then be derived as the inverse of that change.

Change does not always manifest in a like manner. Sometimes it may coincide with the distance of two values on the measurement scale that is being used, but in the general case it is just a function of this distance. The ability to measure change is predicated on a more analytical definition that allows recognizing the form of change that occurs and its properties. To accomplish this task, we partially rely on some simple yet powerful ideas that Aristotle developed in his work on *Physics* (Aristotle 350 B.C.-a; McKeon 2001). According to him, four generic types of change can be identified: (1) change in respect of substance or generation and destruction, (2) change in respect of quantity, (3) change in respect of quality, and (4) change in respect of place or movement. The same types of change are also recognized by contemporary researchers who either elaborate on one particular type of change (Galton 1995; Hornsby and Egenhofer 2000) or further subdivide these generic categories to apply better to their fields of study (Egenhofer and Al-Taha 1991; Claramunt *et al.* 1997; Yanwu and Claramunt 2003; Huang and Claramunt 2005).

Universally present across all types of change is the state *from* which the change proceeds and the state *to* which the change leads. Under the adopted interpretation of change for the task of similarity retrieval, the former corresponds to the entity characterized by the query value and the latter to the (supposedly same) entity characterized by a database value. Objects change between such states through transitions that maintain a temporal order. Excluding the change with respect to substance, another invariant is the object identity that persists through the change. Identity represents an object's individuality or uniqueness, independently of its attributes, values, and spatial characteristics (Khoshafian and Copeland 1986). A fourth variable, occasionally present,



is the existence of two extreme states that act as limits, bounding the potential for change. These states correspond to two values in the domain of the attribute that are considered opposites (e.g., north and south, black and white). When the two states of change coincide with the opposites the change is a maximum and similarity is zero. The extreme states are called *contraries* when intermediate states are possible between them and *contradictories* otherwise. Aristotle also observed that during similarity judgments people resort to a spatial metaphor, a fact for which Gärdenfors (2000) recently provided extensive evidence.

These insights and notions must be given a more practical translation, appropriate for the context of similar information retrieval in GISs. Change in respect of substance is *coming-into-being* and *going out of existence*. The philosophical debates on the meaning of terms such as *existence* and on whether or not an object retains its identity during this type of change are heated and plenty (Barnes 1995), but irrelevant for our practical purposes. In this work, existence and non-existence refer to the presence or absence of an entity, respectively. The entity can be a physical object (i.e., the Parthenon) or an entity created by human decree (e.g., the country of Switzerland) (Smith 1995). The two extreme states of a process of generation are non-existence and existence (Figure 3.2a). During the process of destruction (Hornsby and Egenhofer 2000), the two extreme states are the same, but occur in reverse order. These states are the only possible in this type of change and no intermediate state may exist between them; therefore, they are contradictories. This type of change is commonly implied in GISs (e.g., during similarity comparisons of spatial scenes with different numbers of objects in them).

Change in quantity is *growth or diminution* (Barnes 1995). We use the more casual terms *expansion* and *contraction* instead, but all of these terms imply the presence of magnitudes and quantities (Figure 3.2b). Such a change has to be assessed, for example, when comparing the similarity of street segments with respect to their lengths (Figure 3.1). A distinction is required here between the terms *magnitude* and *quantity*. Magnitude

is anything capable of being greater than or less than something else, whereas quantity is an instance of a particular magnitude (Russell 1938). For a street segment, its magnitude would be the point in the continuum of length that corresponds to the length of the segment. Quantity would refer to the length of the segment itself. The distinction is easily understood if one visualizes magnitude as a point and quantity as an interval. In quantitative change, the meaning of contrary states is undefined. There is no opposite to a length of two meters, a length of four meters, or a length of any extent. Therefore, the contrary states of change have to be implemented conventionally, by imposing a threshold. The purpose of the threshold is to express the maximum amount of change beyond which two values are considered completely dissimilar. The criterion for its specification should be based on the amount of deviation that will still yield useful results to a user's query. In the case of contraction, however, this threshold is limited by the value that results in complete loss of the quantity (i.e., 100% contraction). Infinite intermediate states are theoretically possible between the two extremes, but practically a finite number exists, determined by the precision of the system.

Change in quality, or alteration, is a broad category. It encompasses all cases where an entity differs from another by possessing or lacking a quality (i.e., a property) or by possessing the same quality, albeit in a greater or lesser degree. The characteristics of this type of change exhibit the largest variability and should be determined on a case-by-case basis. Sometimes the change is bounded by two opposite states that admit no intermediaries (at least in the miniworld being modeled). For example, one entity has the property of having a roof, whereas another does not (Figure 3.2c). This binary or Boolean interpretation of qualities is the foundation of featural models of similarity. Other times, the possession of a property is a matter of *degree*. For instance, the property of *having a black color* has *black* and *white* as contrary states and levels of grey as intermediate states (Figure 3.2d). Change to the lesser degree of the quality is change to the contrary of that quality, whereas change to the greater degree of a quality is change from the contrary of

the quality to the quality itself (Aristotle 350 B.C.-a). The presence or absence of a property, represented by its two opposite states, is a qualitative change and should not be confused with generation and destruction, which indicate presence or absence of the entity itself. For some qualities, opposite states are meaningless (e.g., construction date); therefore, they have to be artificially created as in the case of quantities.

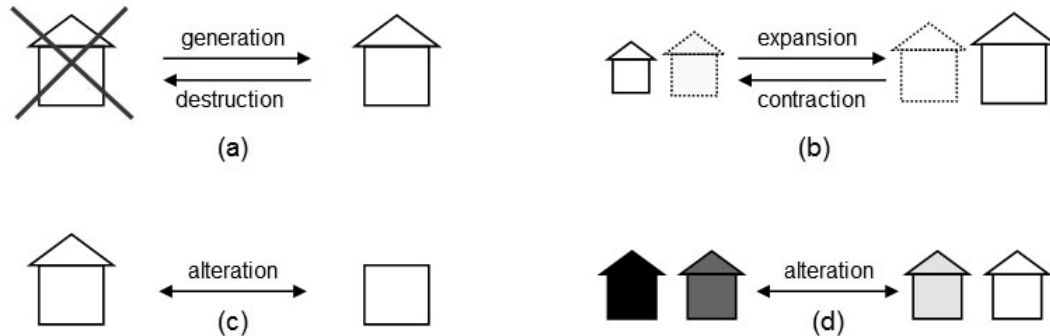


Figure 3.2: Forms of change: (a) generation and destruction, (b) expansion and contraction, (c) alteration with no intermediate states, and (d) alteration with intermediate states.

For static objects, change with respect to place, or movement, is irrelevant for attribute or object-level similarity comparisons because it does not affect the qualities or quantities attributed to an object. It becomes relevant in comparisons of spatial scenes, however, where the movement of an object changes the qualitative and quantitative properties of its spatial relations with other objects of the scene. Even there, however, if one perceives relations also to be entities with their own sets of attributes, then an object's movement may be alternatively registered as a quantitative or qualitative change on the relations of this object to the objects around it. Movement is also the change usually taking place in the abstract mental representations of the remaining forms of change. As Figure 3.2 demonstrates, all other types of change imply some motion from one opposite state to another in their metaphorical representation.

The perspective of similarity that we adopted is that of a theoretical entity (Gärdenfors 2000). Following Sneed's (1971) analysis on theoretical entities, similarity

can only be measured indirectly. The indirect measurement of the similarity between two entities is then provided by measuring the change required to make the entities identical. This perspective is close to the views of advocates of transformational models of similarity, who identify dissimilarity as *transformational distance* (Imai 1977; Hahn and Chater 1997; Hahn *et al.* 2001). It is also congruent with the argument, made in artificial intelligence, that objects are recognized by being aligned with visual descriptions stored or produced in memory (Ullman 2000). Defining similarity as the inverse of change does not rule out geometric or featural models. Instead, it provides the foundation for a more general theory that encompasses all accounts. Feature insertions and deletions as well as distance estimates along a continuous dimension are all *bona fide* expressions of change (Hahn *et al.* 2003); therefore, geometric and featural models measure change as well. Their weakness, however, is that they can only afford a restricted set of change types.

### 3.2 Similarity Functions

The conceptual definitions provided for similarity and dissimilarity highlight the meaning of these notions, but do not provide the specifics required to measure them. Therefore, they must be complemented with operational definitions, that is, algorithms that allow us to measure these concepts and quantify them.

#### 3.2.1 Specification and Properties

The purpose of a similarity function is to express similarity in the quantitative realm. This mapping into the domain of numbers enables an ordering with a value of 1 representing an exact match, and a value of 0 denoting a complete difference (i.e., no similarity at all). The attribution of these meanings to the numerals zero and one is standard practice, albeit not essential since it is only a matter of convention. If  $A$  is an attribute with a value of  $x$  then this is denoted by  $A(x)$  where  $x \in X$ , and  $X$  is the universal set containing all elements being considered in the domain of the attribute. A query with the value  $x$  on attribute  $A$  is denoted instead by  $*A(x)$ . The domain of the attribute may be infinite or

finite. For example,  $X$  may be the set of all integers  $\mathbb{Z}$  or the set of all real numbers  $\mathbb{R}$  or a list of alphanumeric values. The domain of a similarity function (Equation 3.1) is the Cartesian product of values in  $X$  and its codomain the real interval  $[0,1]$ . This is a dyadic (binary) function, because it accepts two arguments: the argument  $x_q$  represents the user input (i.e., the query value) for which similarity  $S$  is determined against every other value  $x_{db}$  that exists for attribute  $A$  in the database.

$$S_A(x_{db}, x_q) \rightarrow [0,1] \quad (3.1)$$

For an attribute with a finite domain, the results obtained by an exhaustive instantiation of the similarity function with all pairwise permutations of the values (i.e., the range of the function) produce a *similarity matrix*  $R$  (Table 3.1). The rows and columns of this matrix represent the elements of the attribute's domain and a cell coefficient  $R(x_{db}, x_q) : x_{db}, x_q \in X$  gives the similarity of element  $x_{db}$  to element  $x_q$ . Similarity matrices—also referred to as semantic distance matrices—have been used by psychologists in multi-dimensional scaling. These matrices served as input, from which the dimensions (i.e., features) involved in the cognitive similarity assessment of a set of stimuli were derived. In information retrieval from databases, however, the dimensions are known *a priori* (i.e., they are attributes themselves), and the similarity matrix is the end product that holds the similarity coefficients among all pairs of values.

	Disjoint	Meet	Overlap	Covers	Covered_by	Contains	Inside	Equal
Disjoint	1.00	0.75	0.50	0.25	0.25	0.00	0.00	0.25
Meet	0.75	1.00	0.75	0.50	0.50	0.25	0.25	0.50
Overlap	0.50	0.75	1.00	0.75	0.75	0.50	0.50	0.75
Covers	0.25	0.50	0.75	1.00	0.50	0.75	0.50	0.75
Covered_by	0.25	0.50	0.75	0.50	1.00	0.50	0.75	0.75
Contains	0.00	0.25	0.50	0.75	0.50	1.00	0.50	0.75
Inside	0.00	0.25	0.50	0.50	0.75	0.50	1.00	0.75
Equal	0.25	0.50	0.75	0.75	0.75	0.75	0.75	1.00

Table 3.1: A possible similarity matrix for the attribute *Topological\_Relation* (Egenhofer and Al-Taha 1992).

Although we always accept the notion that  $S_A(x_{db}, x_q) = 1$  iff  $x_{db} = x_q$  (i.e., identity), the property of symmetry in similarity (i.e.,  $S_A(x_{db}, x_q) = S_A(x_q, x_{db})$  for  $x_{db} \neq x_q$ ) might not always hold, because the change required for  $x_{db}$  to become  $x_q$  may not be the same as the change required for  $x_q$  to become  $x_{db}$ . For example, the function that we employ to assess similarity for an attribute named *Type\_of\_Structure* might be symmetric, yielding  $S(\text{building}, \text{house}) = S(\text{house}, \text{building})$ , or asymmetric, taking into account that a variant is more similar to the prototype than the opposite (Rosch 1975). Similar considerations are made for the properties of transitivity and connectedness. The latter applies when, given any two elements in the domain of the attribute, the relation holds either between the first and the second, or between the second and the first, or both (Russell 1920).

### 3.2.2 Mathematical Formalization

Similarity functions and similarity matrices accept formalizations in the context of fuzzy set theory and graph theory. A similarity function (Equation 3.1) is equivalent to the characteristic function (i.e., the intensional specification) of a binary fuzzy similarity relation on a single set (Section 2.3.1), symbolized as  $R(X, X)$  or  $R(X^2)$ . A similarity matrix, on the other hand, corresponds to the extensional representation of the binary fuzzy similarity relation. In fuzzy terminology, the cells of Table 3.1 represent the degree to which the topological relations in the columns are similar to those in the rows. Since a fuzzy similarity relation is a generalized equivalence relation, one may alternatively say that an arbitrary cell  $R(x_{db}, x_q)$  gives the degree of truth of the proposition  $x_{db}$  is  $x_q$ . For instance, the truth value of the proposition *disjoint is meet* is 0.75, or alternatively *disjoint is 0.75 similar to meet*. Fuzzy similarity relations are defined as strictly symmetric, whereas similarity can often be asymmetric. For this reason the produced  $n \times n$  similarity matrix is best described via a complete weighted multigraph (Figure 3.3).

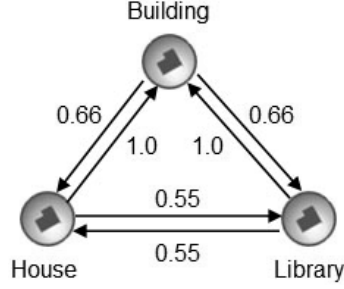


Figure 3.3: Graph representation of a similarity matrix.

While an equivalence relation groups elements into disjoint classes, a similarity relation groups elements into crisp sets whose members are similar to each other to some specified degree. The groups formed by the similarity relation are called similarity classes. For each  $x \in X$ , a similarity class can be defined as a fuzzy set in which the membership grade of any particular element represents the similarity of that element to the element  $x$ . The similarity class for each element is defined by the row of the membership matrix of  $R$  that corresponds to that element. For example, in Table 3.1 the similarity class of *disjoint* is given by the first row of the matrix; therefore, an instantiation of the generalized similarity function with a specific user input (i.e., query value) makes it the membership function of the fuzzy set defined by the values that are similar to some degree to that query value. For instance,  $S_R(x, \text{disjoint}) \rightarrow [0, 1]$ .

Assuming that the attribute values are crisp atomic values, then the fuzzy sets that are generated in this way are *normal* (i.e., there is at least one element that has a similarity value of 1 and, therefore, total membership in the fuzzy set) and the *core* (i.e., the set of attribute values with a similarity score of 1) consists of only one element (i.e., the query value). An  $\alpha$ -cut on this fuzzy set, with  $0 < \alpha < 1$ , determines the set of values that exhibit some similarity above  $\alpha$ . Setting a threshold on a similarity algorithm, such that only values within a certain range are considered similar, changes the *support* of the produced fuzzy set (i.e., the set of attribute values with a similarity score larger than zero).

### 3.2.3 Thresholds and Normalization

As long as the conversion function  $f$  that translates dissimilarity to similarity remains monotonically decreasing, its exact form is irrelevant at the attribute level. This observation entails a shift of attention to deriving appropriately the dissimilarity estimates. Since similarity is measured on a closed scale (i.e., a scale that also includes a fixed end in addition to a true origin), the same must hold for dissimilarity, which is its inverse. This effect can be achieved by using appropriately selected *similarity thresholds*. Their role is to define the meaning of maximum dissimilarity and they correspond to some amount of change beyond which two values become completely dissimilar. When the dissimilarity for a pair of values exceeds the threshold, the similarity is truncated to zero. A threshold, in this manner, defines the *semantic or conceptual neighborhood* of a query value, which delimits its potential for change. A conceptual neighborhood (Freksa 1991) was originally suggested as a graph connecting temporal or spatial relations, so that similar relations are closer to each other in terms of path distance than dissimilar ones. The semantic neighborhood is similar to that concept, however, depending on the implied type of change in the attribute that is being measured, two types of semantic neighborhoods can be distinguished: neighborhoods relative to the query value itself that do not necessarily span the entire attribute range and neighborhoods bounded by two values that are perceived as opposites (Figure 3.2c).

Normalizing by the threshold rescales similarity values in the closed interval  $[0,1]$ . The normalization of dissimilarities is also immaterial for the purposes of establishing similarity rankings within the level of an individual attribute. It becomes important, however, when dissimilarities with respect to multiple attributes must be summarized into a meaningful composite (Equation 2.2) in order to derive the similarity between pairs of objects or relations. In such cases, normalization enforces a common system of reference for dissimilarities across different dimensions (i.e., attributes), so that each of them



contributes equally to the aggregate similarity score. Otherwise, the attributes with the largest ranges will dominate the results.

### 3.3 Classifications of Attributes

The generic similarity function of Equation 3.1 must transmute differently depending on the type and the domain of the attribute to which it will apply. The semantics of the attribute type will be suggestive of the psychological properties of similarity that the function should incorporate, the form of change that should be measured, the thresholds that should be imposed on similarity neighborhoods, and the normalization techniques that should be applied.

Three common classifications for attributes are pertinent to the task of similarity assessments. At a conceptual view we employ the terminology of the extended Entity-Relationship Model (Hohenstein *et al.* 1986) to distinguish between properties such as atomic versus composite; single-valued versus multi-valued; and stored versus derived. In a perfectly normalized database, composite, multi-valued, and derived attributes should be eliminated. They often exist in typical databases, however. Composite and multi-valued attributes can contain several values and are, therefore, addressed in chapter 4. In this chapter, the focus is on single-valued, atomic attributes.

Another classification scheme is based on the domain of the attributes. The term *domain* in the context of an attribute embodies two concepts. The first is the enforcement of a data type. Although commercial DBMSs have a plethora of different data types to improve performance and save storage space, the main data types are bit or Boolean, integer and float for numbers, and alphanumeric or char for text. Specialized attributes also exist for date, time, currency, and large binary objects, such as images and sound files. The second notion of domain is associated with whether an attribute is defined by extension or intension. An extensional definition means that all possible values for the

attribute are listed explicitly (*enumerated data types*), whereas an intentional definition implies that the set of possible values is (theoretically) infinite.

Attributes have also been categorized as nominal, ordinal, interval, and ratio, depending on the type of measurement that their values perform (Stevens 1946). This is the highest semantic classification, since these scales indicate the meaning of measurement. Each of these scales is best characterized by its range of invariance under groups of transformations, meaning the kinds of transformations that leave the inherent structure of the scale undistorted. As the scales progress from nominal to ratio, the information one can extract from numerals and their relations increases, but the number of transformations that preserve the structure of the scale decreases. Besides these four standard scales, two extensions must be considered. The first regards cyclic phenomena. Many measures are bound within a range and repeat in a cyclical manner (e.g., angles or seasons) (Chrisman 1995). Those measurements do not strictly adhere to any of the four standard scales. The second extension is a higher level of measurement than ratio, called absolute (Ellis 1968), or as Stevens (1951) put it, the *numerosity scale*. Absolute scales are almost the same as ratio scales, but their units are discrete and non-arbitrary. These are the scales used to count things—the scales of *counts* (e.g., population)—where units are always perceived as a whole and are indivisible (e.g., one person). The distinction between ratio and count scales seems to dissolve at the atomic level, where quantum theory (Bohm 1951) reveals that many quantities occur in discrete units, or quanta.

The groupings based on the scales of measurement are ubiquitous in natural and social sciences. They also provide a convenient organizational structure for the definition of similarity algorithms customized to the type and the semantics of each scale. However, the correspondence between scale types and similarity algorithms is not one-to-one, but surjective. The reason for this discrepancy is that the scale type is not always an exclusive indicator of the form of change that is being assessed and, consequently, of the function appropriate to determine similarity or dissimilarity.

### 3.4 Similarity Assessment for Ratio Values

Ratio measurements are typically expressed in positive numbers that have a true origin and arbitrary values. The label of the attribute determines the meaning of the distance among values. Differences between ratio scale units correspond to equal intervals. In addition to subtraction or addition, operations such as multiplication and division are also meaningful. A ratio scale is invariant under the similarity group of transformations  $x' = a \cdot x$ , meaning that its numerical values can be transformed only by multiplying with a constant. In contrast, the only transformation that values on an absolute scale accept is the identity operation (i.e., multiplication by unity). This difference does not prevent the development of a uniform methodology to measure similarity for ratio and absolute values, because all permissible mathematical operations on ratios are also meaningful when applied to counts. An example of a ratio attribute is area. Its values may be placed on an axis isomorphic to the half-line of non-negative numbers and the origin is the zero point. Other examples include length, depth, and population. Attributes of a ratio nature are more commonly encountered in geographic databases than attributes that are interval or ordinal, since ratio is the predominant type of measurement for physical quantities.

Ratio values that are closer along the axis are naturally expected to be more similar than other values that are further apart. This intuitive assumption, which underlies geometric models of similarity, is also compatible with an interpretation based on change, because near values require less change than remote values. A dissimilarity measure for this type of measurement should, therefore, reflect the properties of identity and triangle inequality that hold for the actual distances among the values on the scale. These properties of distance impose the following postulates on the dissimilarity measure:

- *Postulate 1:*  $Distance(x, y) = 0$  implies that the values  $x$  and  $y$  are equal.
- *Postulate 2:*  $Distance(x, z) > Distance(x, y)$  means that the dissimilarity of  $x$  to  $z$  is larger than that of  $x$  to  $y$ .

A ratio scale is the most sophisticated level of scale, because it allows for the interpretation of one observation exceeding another, not only by a certain amount, as in interval measurement, but also by a certain ratio. Consequently, the interpretation of similarity or dissimilarity may sometimes be abstruse. The definition of a ratio scale was, in a sense, based on how much information about the property the numbers represented. In order to create meaningful dissimilarity measures, however, it is also necessary to distinguish between what might be called the *kinds* of information that the numbers represent. To do so, two broad approaches to the construction of a ratio scale must be recognized. The distinction corresponds roughly to the difference between fundamental measurements as used in physics, and measurements used in other disciplines, such as psychology, segregating ratio scales into two classes: (1) quantitative ratio scales and (2) qualitative ratio scales.

Under a loose interpretation, quantitative ratios refer to the entity itself, whereas qualitative ratios measure a property of the entity. Differentiating between these two kinds of ratio measurement is imperative for similarity, because different forms of change are implied in each occasion. The distinction between the two variations has also been noted elsewhere. Torgerson (1958) pointed out that the operation of central importance on quantitative ratio scales is that of addition, whereas what matters for the qualitative ratio scales is the relation of distance, or the difference between the values along the scale. This observation, incidentally, corresponds closely to Russell's (1938) formulation, that distinguished between "attributes whose quantities are divisible, and attributes whose quantities possess the relation distance." The same distinction—among other reasons—has also motivated several researchers to suggest alternative scale taxonomies (Mosteller and Tukey 1977), in which the two types of ratio scales are explicitly separated.

In quantitative ratio scales, the observable events in the real world are quantities (e.g., length, area); hence, the change that must be assessed is quantitative. Equal intervals along the scale do not indicate equal amounts of change for different pairs of values. The

absolute difference between two values informs about how far apart the values are on the scale, but it does not reveal the amount of relative change required for one value to transform into the other. For example, the difference between 10 and 20 meters is not the same as that between 100 and 200 meters, although each pair represents an expansion of 100%.

In qualitative ratio scales, the relevance of quantities dissolves, and the observable variable is the relative degree to which an entity possesses some property or quality. The relation is already innate in the values and the change that must be assessed is qualitative. What is of importance, in this scenario, is the distance relation between the magnitudes, since equal distances indicate equal amounts of change in the degree of possession of the property. Such ratio scales have inherent the notion of percentage, as there is a definitive limit implied on the degree to which an entity possesses the property. Percentage scales belong to the general class of ratio scales, although this has raised some criticism on Steven's classification (Velleman and Wilkinson 1993). Qualitative ratio scales are standard in psychology. It is also possible, however, to encounter them in GISs, but not necessarily in an explicit percentage format (Figure 3.4). As long as the scale has an origin that indicates complete lack of the property being measured, and the assumption of equal intervals applies, the scale under consideration is formally a ratio scale.

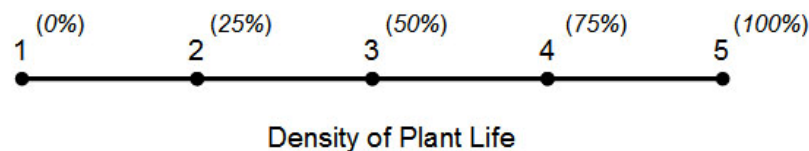


Figure 3.4: An ostensibly ordinal scale is ratio if the origin indicates absence of the property and the intervals between consecutive values represent equal amounts of change to the degree that the property is fulfilled.

### 3.4.1 Similarity for Ratio Quantities

One approach to similarity of ratio quantities is based on the absolute difference of the logarithms of their magnitudes (Equation 3.2). On a logarithmic scale, equal differences in orders of magnitude are represented by equal distances. The coefficient  $C$  allows control over the amount of change that is required on the original scale, so that the distance between units becomes 1 on the logarithmic scale. For instance, if  $C$  is set to 1 (meaning 100% change), then the logarithmic distances between any two values, where one is the double of the other, will be 1. The logarithmic measure is symmetric, since it does not consider the direction of the change. The similarity of a pair of values will be the same regardless of which value becomes the query and which the target. Given a specific query value, the value that corresponds to its half will be equally similar to it as the value that corresponds to its double (Figure 3.5a). This behavior violates the second postulate about dissimilarity (Section 3.4). For example, if the query value is 100, a value of 40 would be less similar than a value of 200.

$$S(x_{db}, x_q) \left( \left| \log_{(1+C)} x_q - \log_{(1+C)} x_{db} \right| \right) = f_{inv} \left( \left| \log_{(1+C)} \left( \frac{x_q}{x_{db}} \right) \right| \right), C > 0 \quad (3.2)$$

This problem can be rectified by calculating dissimilarity based on the direct ratio of two values, rather than the logarithm of that ratio (Equation 3.4). Dissimilarity is defined as the relative change  $C$  that must be applied to the interval represented by the query value  $x_q$ , so that it coincides with the interval represented by the database value  $x_{db}$  (Equation 3.3). Similarity is then computed as the inverse of that change (Figure 3.5b).

$$x_{db} = x_q \pm C \cdot x_q \Rightarrow C = \frac{|x_{db} - x_q|}{x_q} \quad (3.3)$$

$$S(x_{db}, x_q) = \begin{cases} f_{inv} \left( \frac{C}{T} \right) & \text{if } C < T \text{ and } x_q \neq 0 \\ 0 & \text{if } C \geq T \text{ or } x_q = 0 \end{cases} \quad (3.4)$$

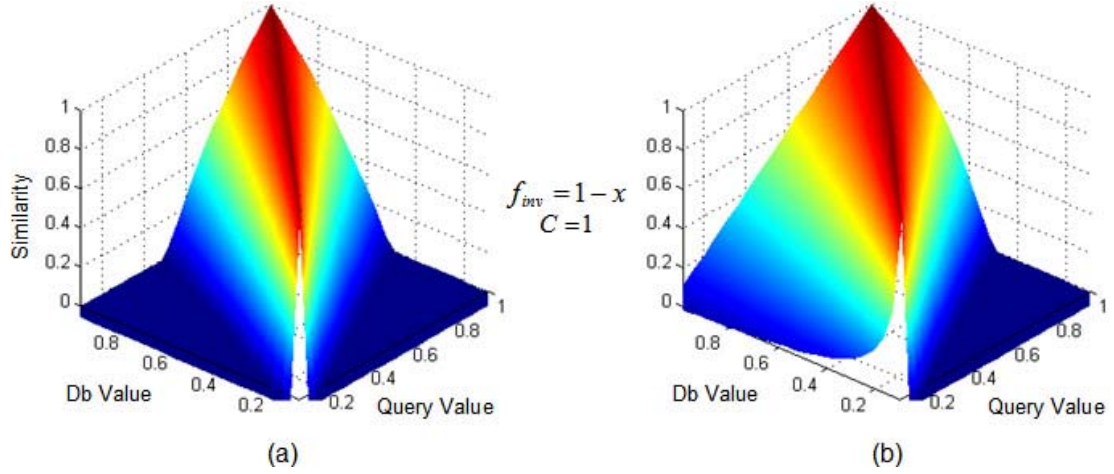


Figure 3.5: Similarity for ratio quantities: (a) as the inverse of logarithmic distance and (b) as the inverse of relative change.

The changes and similarities for ratio values, as computed by Equations 3.3 and 3.4, are asymmetric. They are both dimensionless quantities and cannot have a negative value. The type of change that must be applied to a query value so that it coincides with a database value can be either an expansion, if  $x_q < x_{db}$ , or a contraction, if  $x_q > x_{db}$ . In the former case, the value of the non-normalized dissimilarity is the interval  $(0, +\infty)$ , whereas in the latter it is in the interval  $(0, 1)$ . The parameter  $T$  in Equation 3.4 is a threshold that demarcates the conceptual neighborhood for each query value, serving at the same time as a normalizing constant. This threshold is specified as a percentage that expresses the maximum amount of change beyond which two values become completely dissimilar. It must be conventionally defined, since no quantity can be intuitively perceived as the opposite of another. If the dissimilarity for a pair of values exceeds this threshold, then the similarity is truncated to zero. For instance, if  $T$  is set to 2 (i.e., 200%) and the query value is 100, values equal to or larger than 300 will have a similarity of 0.

Defining the threshold in terms of change is conceptually and operationally simpler for the database users. If they know this threshold or set it at will, they may easily infer the permissible amount of fluctuation for any query value so that they adjust their mental representations of the similarity neighborhoods accordingly (i.e., extension effect). Thus,

an alignment is achieved between the objective and the perceived width of similarity neighborhoods. For multi-attribute queries (i.e. object-level queries), such a threshold results in a uniform creation of similarity neighborhoods for each attribute, because the extent of the neighborhood is tailored accordingly to the magnitude of the query value for each particular attribute. This is a preferred alternative to defining similarity neighborhoods with arbitrary assignments of ranges of permissible values on each attribute's scale. It allows different dissimilarities to be aggregated in a coherent manner, which approximates better the dimensions and the extent of a user's conceptual space.

An interesting observation can be made about the semantics of the zero point. Zero, on a ratio scale of quantities, has a particular physical meaning. If a query value is zero, change becomes infinite (Equation 3.3). This is no accident, since the non-arbitrary origin of the ratio scale is theoretical, rather than practical. For any physical and continuous property, an actual zero magnitude is unattainable, implying absence of the entity. There cannot be a physical object with zero length or zero area, as there cannot be an absolute temperature of zero degrees (i.e., in Kelvin) because this would require that even atoms stop their motion. Therefore, a possible explanation of why change becomes infinite is because this is the amount of change required to bring a non-existent object into existence, or simply, because there cannot be similar objects to an object that does not exist.

Such speculations, of course, do not prevent users from querying with a zero value. Furthermore, a zero value is also possible for counts. For example, there could be a deserted village with zero population. In addition, zeroes may exist in the database because of erroneous entries or due to the finite precision of measurement instruments.

A change-based framework can address such rarities in a theoretically sound manner. When the query involves a zero quantity, the type of change that takes place is not quantitative anymore. A transition from a state of zero quantity to a state of some quantity characterized by a positive magnitude is change with respect to substance. This



type of change has only two possible states: existence and non-existence. For this reason, the similarity of a zero quantity to any positive quantity is zero; hence, the provision in Equation 3.4. This strict interpretation is not always desirable. After all, a village with ten habitants is closer to becoming deserted than one with a thousand habitants. The key phrase here is “to be deserted.” This phrase defines a qualitative property for a place. Hence, from this standpoint, the question asked of the system is to find similar entities to an entity that has the quality of having a zero quantity. The change being assessed then is qualitative, with different distances from zero indicating different degrees of membership to that property. The methodology for ratio and interval similarity assessments where the form of change is qualitative is presented in the following sections.

### 3.4.2 Similarity for Ratio Magnitudes

When magnitudes on a ratio scale indicate the presence or absence of a quality to some degree, change and distance coincide. Equal distances indicate equal amounts of change to the degree to which the entity possesses the property (Figure 3.4). Opposite values or contrary states are those indicating complete absence or total presence of the property. Therefore, the similarity between two values is an inverse function of their distance on the scale (Equation 3.5). The normalizing parameter  $T$  should be set—under normal circumstances—equal to the distance between the two extreme values, because their opposite meaning is transparent and this setting is most likely what users would expect (i.e., extension effect); however, a smaller distance than the range could also serve as a threshold if so desired. The similarity scores, as computed by Equation 3.5, are symmetric and lie in the interval  $[0,1]$  (Figure 3.6).

$$S(x_{db}, x_q) = \begin{cases} f_{inv} \left( \frac{|x_{db} - x_q|}{T} \right) & \text{if } |x_{db} - x_q| < T \\ 0 & \text{if } |x_{db} - x_q| \geq T \end{cases} \quad (3.5)$$

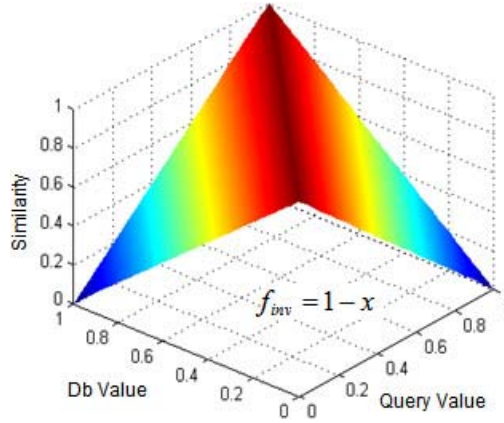


Figure 3.6: Similarity for ratio magnitudes as the inverse of distance.

### 3.5 Similarity Assessment for Interval Values

Interval values differ from ratio with respect to the existence of a zero point, which for interval scales is simply a matter of convenience. An additional difference compared to physical ratio quantities is that values can be negative. Measurements on one scale can be converted into values on another through any affine transformation of the form  $x' = ax + b$ , which highlights that the origin and the units of the scale are arbitrary. Examples of interval type values include the year date in various calendars, the common temperature scales, and energy (Neumann and Morgenstern 1947). Interval scales are commonplace in psychology, but scarce in physical sciences. On an interval scale multiplication and division have no meaning. Addition and subtraction, however, can be meaningful (i.e., adding energy, or subtracting dates to obtain time periods).

Interval values represent magnitudes, but the notion of quantity is inapplicable. Although such values are often referred to as quantitative, the transition from one value to another on an interval scale indicates change in quality. For example, consider a user who requests to find a building constructed in the year 2000. The quality (i.e., property) in this case is “constructed in the year 2000.” Buildings whose construction date differs from 2000 have varying degrees of membership to this property. What matters for similarity is the distance relation between the values. Therefore, the procedure is almost

identical to that described in Section 3.4.2. and similarity is computed by Equation 3.5. The only difference is the lack of a pair of values that act as logical opposites. Hence, the limit  $T$  that indicates complete absence of the property must be artificially created.

There are several approaches to defining the threshold  $T$ , thereby normalizing the distance. The easiest is to set  $T$  equal to the *range* of the attribute, determined as the absolute value of the difference between a maximum and a minimum value. This technique is known as *min-max normalization* (Korfhage 1997) and can be further classified into two scenarios. The first scenario arises when the range is dynamically specified from the maximum and minimum values that exist in the current database instance for the attribute in question. This specification is dynamic in the sense that a new entry in the database may alter the range. The second case arises when these values are statically defined *a priori* by a declarative constraint, such as those created with the “Create Assertion” and “Check” clauses of the SQL language (Groff and Weinberg 2002). Such statements restrict the values of the domain to a subrange of the data type. Division by the range usually fails at creating commensurate similarity measures, because it allows outliers (i.e., extreme values in the data) to have a profound effect on the contribution of an attribute to an aggregate score. For example, in presence of an extreme (or erroneous) value of 150, range normalization of an interval variable whose remaining values all fall within the interval (5,15) would make the values in this set appear almost identical to each other, since their distances would be trivial compared to the range.

To prevent this effect, measures other than the range can be used to describe the variation of values in an attribute’s domain (i.e., the *spread* of values). A more robust alternative is a linear transform that creates a normalized version of the scale of the variable, with the property that the mean  $\mu$  is 0 and the standard deviation  $\sigma$  is 1. This transformation is called *standardization* or *z-score reduction* (Equation 3.6). A zero mean avoids aggregation distortions stemming from differences among means of different attributes. The z-score of a value indicates how far the value is from the mean in standard

deviation units. The meaning of maximum dissimilarity is then defined by specifying some multiple of the standard deviation. Since in normal distributions approximately 95% of the values fall within two standard deviations from the mean, the difference between the values is divided by four standard deviations to scale each value into a range of width 1 (i.e.,  $T = 4 \cdot \sigma$ ) (Wilson and Martinez 1997). It is possible to use tighter thresholds by setting  $T$  equal to two standard deviations (i.e., 68.2% of the data) or equal to the interquartile range (i.e., 50% of the data). Whatever threshold is chosen, values exceeding it are mapped onto the minimum or maximum to avoid normalized values outside the range  $[0,1]$ , thus, trimming in-essence the tails of the attribute's distribution.

$$Z_x = \frac{x - \mu}{\sigma} \quad (3.6)$$

Clipping out-of-range values would be treating them as equivalent to the limits of the threshold range. Under rare circumstances, this may affect the correct sorting order of the list of similar results that are retrieved. For example, consider the query *\*Construction\_Date(1900)*. Two objects with construction dates of 1750 and 1450 might both be well off the threshold range and, therefore, have a similarity of zero. In absence of exact matches, however, it is still desirable to be able to sort such distant matches from most to least similar. This objective can be achieved with a logistic function (Equation 3.7), which can keep a specified range under a linear transform and still handle outliers without discarding them. Such a transformation is called *softmax* scaling (Pyle 1999). It transforms the range of  $[-\infty, +\infty]$  into the range  $[0,1]$ . The desired part of the range that should have a linear response  $r$  is defined in terms of standard deviations (e.g.,  $4 \cdot \sigma$ ).

$$x_{norm} = \left( 1 + e^{-\frac{(x-\mu)}{r(\sigma/2\pi)}} \right)^{-1} \quad (3.7)$$

All these normalization alternatives seek to automate the process of defining a maximum distance at which two values are considered opposite so that their similarity

becomes zero. The choice is important, because it determines the effect of the attribute on a composite similarity score and, hence, the produced rankings to multi-attribute queries. Successively stricter thresholds result in consecutively tighter similarity neighborhoods around a query value. Z-score and softmax scaling methods are superior to the range, because the latter is sensitive to outliers. Choosing among them depends on additional factors, such as the distribution of the values and their concentration around the mean. In information systems, however, it is unlikely that users have any knowledge about such information, especially in the case of interval and ratio values that may spread through very large ranges. Therefore, none of these system-imposed thresholds—despite the provisions they make—guarantees an alignment between the distance that they define as the maximum possible and the maximum possible distance that the users would expect.

Such an ideal situation is achieved only in three cases: (1) when the threshold is manually declared by the users, (2) when the attribute is enumerated and the domain contains a relatively small set of values with which users are familiar (e.g., ordinal rating scales), and (3) when the attribute's domain contains two opposite values (i.e., contrary states of change) that are unequivocally identified (Section 3.4.2). In the case of such interval values as temperatures or calendar dates, however, only the first option provides a viable alternative to an automated system threshold. As a last resort, weights can be used to calibrate the results. The scarcity of pure interval scales in GISs compensates for the unpredictable effects that the normalization of their values may entail for the quality of the produced similarity scores.

A final point of attention about similarity for interval type values relates to queries that use interval values in their expression, but do so in a way that transforms the type of the scale. For example, consider a spatial scene query where the user is interested in finding two buildings whose construction dates differ by 10 years. A difference of interval values becomes a ratio value (the subtraction gets rid of the additive constant  $b$  in the affine transformation equation); therefore, the period of time in this user's query

represents a ratio quantity (Section 3.4.1). Other queries may evoke the same kind of scale transformation, albeit in a more subtle manner. For instance, a user may be querying for a building constructed in the year 2000, but her actual intention is to find buildings that are 6 years old. Thus, the same attribute values are treated as measuring different things for different purposes. Such intentions cannot always be predicted automatically. Ultimately, the measurement level depends on the question asked and is not an immutable property of the data (Velleman and Wilkinson 1993).

### 3.6 Similarity Assessment for Ordinal Values

Attributes with ordinal values preserve the concept of ordering on a scale, but lack a numeric representation. The sequence of values is registered, but their positioning and spacing along the scale is not explicitly stated. Therefore, in addition to multiplication and division, addition and subtraction are also meaningless. An ordinal scale is invariant under the isotonic (i.e., order-preserving) group of transformations  $x' = f(x)$ , where  $f$  is any increasing monotonic function. Typically, ordinals are defined by extension. They can be grammatically expressed by adjectives, nouns, and adverbs, thus having assigned to them a variant of the data type *text*. Occasionally integers may be used; however, these integers should not be perceived as numbers, but rather as codes mapped onto the concepts or categories represented. They help resolve order-related ambiguity when the actual values are not intuitive for that purpose.

Ordinal scales can be divided into *rank-order* scales or *rating scales*. Rank-order scales represent the weakest form of ordinal measurement. They delineate nothing more than ordinal relationships. Values on a rank-order scale correspond to points. An example of a rank-order scale is a list of the most similar items to a query. A more widespread variation of the theme of ordinal measurement comprises rating scales. A classic—though not spatial—example of a rating scale is the grading system of U.S. universities from A to F. Examples of a spatial nature include the *Physical\_Condition\_Of\_Feature*, defined in

SDTS as, “the state of repair of a feature or the extent of deterioration,” or the *Density\_Of\_Growth*, defined as “the degree or measured degree to which the area is filled or occupied by plant life.” As these definitions suggest, the order of symbols in rating scales corresponds to successively increasing or decreasing degrees to which some property is fulfilled. Thus, the idea of intervals between successive values is somewhat more pronounced in this type of scale. The values themselves correspond either to points (Figure 3.7a) or to intervals (Figure 3.7b). In the latter case, the values have a fuzzy character, serving as groupings or sets of finer discriminations.

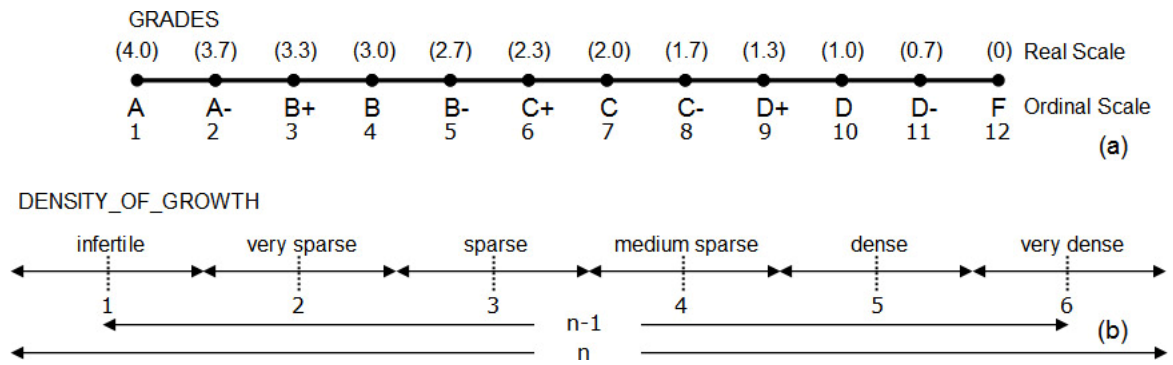


Figure 3.7: Values of ordinal rating scales: (a) as points and (b) as intervals.

Rating scales share many commonalities with qualitative ratio scales (Section 3.4.2), where transitions from one value to the next reflect qualitative change. Both scales are also bounded by two extreme values, which may be perceived as the origin and the end of the scale (Torgerson 1958). Unlike ratio or interval scales, however, consecutive ordinals are not intrinsically separated by equal intervals. The obvious implication is that, unless we promote ordinal scales to a higher level of measurement under certain assumptions, a quantitative similarity score between two values is impossible. Assuming equal intervals, similarity among the values may be derived by applying Equations 3.8a and 3.8b, where parameters  $j$  and  $i$  are integers onto which the  $n$  ordinal values have been mapped depending on their order of succession. These integers correspond to either points (Figure 3.7a) or the midpoints of the intervals defined by the ordinal values (Figures 3.7b).

The different denominators of the two equations point out a semantic distinction, which relates to the exact positioning of the two contrary states of change that indicate complete absence or total presence of the property being measured. Equation 3.8a regards the first and last values in the ordering as opposites, thus yielding a similarity of zero magnitude. Similarity among  $n$  ordinals is defined in the same way as for  $n$  integers on a closed ratio scale of length  $n-1$ . Equation 3.8b, on the other hand, treats the first and last values as the least semantically similar pair of values, rather than as opposites, yielding a slightly above zero positive similarity coefficient. In this case, the length of the scale is  $n$ . Favoring the use of one equation over the other is a matter of personal judgment. An intuitive decision can be taken based on the actual values at stake. For instance, it is more logical in the example of Figure 3.7b to apply Equation 3.8a, since an area with very dense plant life seems to be the opposite of an infertile area. If, however, the set of available values was {very dense, dense, medium dense, sparse, very sparse}, then Equation 3.8b should be preferred, because it would be exaggerated to consider an area with little vegetation as the exact opposite of an area with very dense plant life.

$$S(x_{db}, x_q) = f_{inv} \left( \frac{|j-i|}{n-1} \right) \quad (3.8a)$$

$$S(x_{db}, x_q) = f_{inv} \left( \frac{|j-i|}{n} \right) \quad (3.8b)$$

Both of these equations are based on the rather strong assumption of equal intervals. When ordinal values barely evolve from a nominal level, this speculation seems to be the only viable alternative in order to obtain approximate similarity measures. The surmise that people will most likely consent to an equal interval interpretation is also justified by the range-frequency theory (Parducci 1965), which states that people tend to divide their psychological ranges into a fixed number of sub-ranges of equal size and employ the alternative categories with equal frequency. The conclusion of this theory is not



surprising. Although purely illusory, the impression stimulated by ordinal values on the scale is often highly suggestive of equal intervals (Figure 3.7).

The same theory, however, explains many misunderstandings concerning ordinal scales that emerge from underlying interval or ratio models, for which they serve as crude—though convenient—surrogates. An example is the Richter scale of earthquake intensity. Because of its logarithmic basis, each ordinal magnitude increase on this scale represents a tenfold increase in earthquake intensity. When such additional information is available, it should be exploited so that similarity assessments can take place at a higher level of measurement. Unfortunately, people are often unaware of the mathematical basis supporting such ordinal scales, so that their mental representation of the scale may vary drastically from the real one, thus leading to preposterous conclusions. Similar problems may occur in other scales of measurement as well. Considering the interval scale of measurement, for instance, some users may believe that  $40^{\circ}$  means twice as warm as  $20^{\circ}$ . The argument demonstrates that approximating people's perceptions, although desirable (McCloskey 1983; Egenhofer and Mark 1995b), is not always a means to an end.

### **3.7 Similarity Assessment for Nominal Values**

A nominal scale is invariant under the permutation group of transformations  $x' = f(x)$ , where  $f$  is any one-to-one substitution. A nominal attribute type describes values that can be distinguished only by equality. Such attributes present the most challenging case of similarity assessment, because they perform a labeling on the entity instances for which no intuitive mapping onto a metric scale can be derived. With respect to this labeling, two types of nominal values are possible: (1) *classifiers*, which group entities into sets (i.e., many entities have the same label) and (2) *identifiers*, which distinguish each entity individually (i.e., each entity has a unique label). Identifiers may be viewed as a special case of classifiers where the sets are as many as the entities and, therefore, each entity is the only member of its class (Stevens 1946). An example of a classifier is the attribute

*Land\_Use\_Category*, defined in SDTS as “a broad classification of the use of land for planning and zone purposes.” Many entities in the database may belong to the same land-use category (e.g., *agricultural field*). A typical case of an identifier is the attribute *Name*, defined in SDTS as “a word or phrase that constitutes the distinctive designation of an occurrence of a feature.” In contrast to the *Land\_Use\_Category* example, each entity in the database will have its own unique name (e.g., *Boardman Hall*).

### 3.7.1 Similarity Assessment for Nominal Classifiers

The values of nominal classifiers group entities into disjoint classes and are often listed explicitly *a priori*. Grammatically, they are usually expressed by nouns. In contrast to ordinal, interval, and ratio attributes, which describe a single property and have values that vary along one dimension, nominal values represent concepts, which may vary with respect to multiple dimensions (Gärdenfors 2000). Similarity assessment in this case requires a more complex approach because one needs to compare the stimuli overall, and not with respect to a particular feature as was done with the other types of attributes. The global character of this comparison increases the cognitive factor and reduces the appropriateness of rigid geometric models (Torgerson 1965; Tversky 1977).

In order to find the similarity between two concepts one must first establish a *reliable* representation for them. Reliability in the representation implies modeling accurately the type of change that is required for one concept to transform into another. Since concepts can be described in terms of their qualities (i.e., possessed properties) (Sloman *et al.* 1998), measuring the similarity of nominal classifiers involves an enumeration of the properties that a concept must acquire, as well as those that it must discard, in order to become identical to the concept that it is being compared. This enumeration constitutes measurement of qualitative change along many dimensions. The Boolean interpretation of properties requires that change has only two contradictory states: full possession or total absence of the property. Ontologies that contain features of concepts, in addition to

the hierarchical relationships among them, can provide the representation upon which one operates to measure such qualitative differences of concepts. In this manner, an ontology becomes an organization of differences within similarity. The ability to measure change becomes, therefore, strongly dependent on the expressive plurality and the structural coherence of ontologies.

A key factor in the successful measurement of the qualitative change between two concepts through ontologies is the isolation of the core from the irrelevant properties for each concept (Lewis 1986). Redundancy or lack of relevant properties implies that the similarity measures will be overestimated or underrated, respectively. The quality of a similarity measure is also inextricably tied to the users' understanding and acceptance of an ontology (i.e., ontological commitment). The representational availability of a domain in terms of an ontology does not necessarily imply an explicit awareness of the ontology's structure and content from the users (Holsapple and Joshi 2002). The computed similarity scores will have greater fidelity for the users who have more expertise and familiarity with the domain of interest.

Assuming that these conditions are met, an appropriate model for the evaluation of semantic similarity of concepts is the MD model (Rodríguez 2000) (Section 2.2.2.5), which subdivides the features (properties) of spatial entity types into parts, functions, and attributes of the objects. By including all these components in the representation, the model considers simultaneously different descriptions of a spatial entity that capture diverse aspects of its use, purpose, and structure (Marr 1982). By separating these components for queries, users and database administrators can adjust the context of similarity assessments (Rodríguez and Egenhofer 1999). For example, a user may be interested in playing a sport, therefore, being interested mainly on similarity with respect to the *function* component of spatial entities and not with respect to the *parts* or *attributes* components. The consideration of meronymy relations, in addition to hyponymy, emphasizes the suitability of this model for spatial databases.

The treatment of the hyponymy relation in the MD model, as well as in other models that consider it, is problematic. This relation merits further comment if it is to be interpreted appropriately for the purposes of information retrieval. Usually, very general concepts, located at the top of the hierarchical structure, have very few distinguishing features compared to concepts that are more specific. Since set-theoretic models rely on a comparison of distinguishing features, the lack of such features in an entity's definition will produce a similarity value with respect to any other entity class in the ontology equal to zero. This is at odds with information retrieval, where a query value specifies a constraint. The more general a query value is, the less restrictive this constraint becomes. Therefore, if a user queries for an *entity* with some particular area, any subclass of *entity* (i.e., the most general concept in ontologies) will be an exact match to the query as long as the area constraint is also met. Similarly, a user who queries for a *building* is interested in anything that is a building and any subclass of *building* should be an exact match. The reverse does not hold, however; a *building* should not be an exact match to a *house* query. This asymmetry is due to the homonymic use of the word "is." A *house* is *actually* a building, hence, its similarity to *building* should be 1 (i.e., the house does not need to change to become a building). A *building* is *potentially* a house, therefore, its similarity to *house* should be less than 1 (i.e., a building may need to change to become a house). The conclusion is that whenever the database value is a subclass of the query value, the similarity score of this pair should be 1, indicating an exact match.

The application of any set-theoretic model is possible only if the ontology fully defines the features (i.e., attributes, functions, and parts) associated with the concepts. For ontologies that provide only a hierarchical structure, it is mandatory to resort to network models that rely on the concepts' information content and their distance in the hierarchy (Section 2.2.2.4). In a comprehensive comparison of these measures, Budanitsky (2001) concluded that the measure from Jiang and Conrath (1997) is the most reliable. His results have been independently confirmed by Patwardhan (2003). Regardless of what

network model one abides by, few—if any of them—are able to capture accurately the exact amount of qualitative change that one concept should undergo in order to coincide with another. Information-theoretic or edge-based models do not measure qualitative change directly. They are approximations, acting *in lieu* of featural models when the level of detail of the ontology prohibits the employment of the latter. Furthermore, some of the measures produced by network models may be hard to normalize (Patwardhan 2003). For instance, Resnik's (1995) measure does not have an upper bound. For both the MD model and the network models, retrieval time of similar results, can be optimized by computing *a priori* the minimum distances between the  $n$  concepts in the network, using a shortest path algorithm (Dijkstra 1959) and storing them in a  $n \times n$  matrix.

If a local database does not subscribe to an ontology, or it subscribes but the values of a nominal classifier attribute do not correspond to ontology classes, one must seek different methods to assess similarity among the nominal values. One solution is to resort back to geometric models. Since a concept is viewed as comprising several properties, the first step would be to identify the relevant properties and the second to examine how they can be represented geometrically. In this sense, the nominal value temporarily becomes an entity instance with its own set of ratio, interval, or ordinal attributes. This approach works well when the relevant prominent dimensions of the nominal values are relatively small in number, and can be easily recognized using common sense. A fitting example is color, for which several geometric models exist. Therefore, a nominal value such as *orange* can be mapped onto several concomitant attributes, whether these are levels of red, green, and blue, or hue, saturation, and brightness.

Custom geometric decompositions work sufficiently when the component dimensions are easy to obtain. Moreover, in all the scenarios discussed about nominal attributes, it was assumed that the set of values is defined by extension, or that a specialized spatial ontology exists. When none of these requirements is met, two simple alternatives are (1) to employ the hierarchical structure of WordNet (Miller 1995) or (2) to lookup for

synonym words. In the first case, any networking algorithm (section 2.2.2.4) may be used. WordNet, however, is a generic ontology and does not have the specificity of a domain ontology. The similarity measures obtained from pure network models are likely to vary widely and will be symmetric (Table 3.2). In the second case, when the user queries the information system for a nominal value, the value is passed by means of a module to WordNet, which returns a set of synonyms. A string matching may then be performed between the synonym words and the rest of the nominal values in the database. Nominal values that match one of the synonyms could be returned as similar results. Although synonym-lookup is a valid choice, it is a coarse approach to semantic similarity compared to the rest of the methodologies developed.

MODEL USED	Similarity of Building to Library	Similarity of Library to Building
Rada (Normalized Path Length)	0.5	0.5
Leacock and Chodorow (Eqn. 2.6)	2.8904	2.8904
Resnik (Eqn. 2.7)	5.1947	5.1947
Jiang and Conrath (Eqn. 2.8)	0.1106	0.1106
MD Model (original) (Eqn. 2.9)	0.557	0.666
MD Model (modified for hyponymy)	0.557	1.0

Table 3.2: Similarity measures obtained from WordNet with network models versus those obtained from a spatial ontology with the Matching Distance model.

### 3.7.2 Similarity Assessment for Boolean Attributes

For Boolean or binary variables, the classes of one-to-one, monotonically increasing, and affine transformations become identical; Therefore, it may be argued that Boolean variables are at least at the interval level. If the variable also implies presence or absence of a property then Boolean variables are at the ratio or absolute levels. For the purposes of similarity assessment it matters that the values divide the entities into two classes, with one being the negation (i.e., opposite) of the other. Such values can be *true* and *false*, *0* and *1*, or an arbitrary string and its antonym. In this sense, Boolean attributes always admit of a nominal interpretation, and the values imply a change between two

contradictory states. Similarity is 1 if the values are the same and 0 if they are different. An example of two Boolean values from the STDS is *Onshore* and *Offshore*.

### 3.7.3 Similarity Assessment for Nominal Identifiers

Nominal identifiers assign an unique value to each entity. Values of such attributes do not represent concepts and are not defined by extension. Numeric or text data types may be employed, however, the numerals, when used, are not subject to any valid arithmetical operations. If text is used, the values may consist of a single or multiple words.

Nominal identifiers do not represent entity classes; therefore, ontology-based or custom geometric approaches cannot be implemented. Although synonym-lookup may be a viable approach in certain situations, a similarity measure based on semantics is often undesirable. For example, assume that a user queries a lodging database for a hotel by providing part of the hotel's name as input. If the name of the hotel is *The Beacon*, then hotels whose name contains words such as *lighthouse*, *tower*, and *pharos* will be returned from WordNet as similar entries. It is highly unlikely, however, that hotel names containing these words have any association with the original hotel that the user was trying to retrieve. On the other hand, a string-matching algorithm (Aho and Corasick 1975; Boyer and Moore 1977) will behave more reliably in this scenario. Hence, a syntactic rather than semantic evaluation of similarity is preferred for nominal identifiers.

The most common string-matching algorithms are variants of approaches that operate in terms of transformational distances (Hamming 1950; Damerau 1964; Levenshtein 1965), thereby measuring implicitly the change required to transform one string into another in terms of insertion, deletion, substitution, and swapping of characters. Another possible approach—also based on transformations—is phonetic matching, which identifies strings of similar pronunciation (Zobel and Dart 1996) Table 3.3 summarizes the discussion on similarity among nominal values by presenting the possible and recommended methods for the attributes types discussed.

	Classifiers with Detailed Ontology	Classifiers with Basic Taxonomy	Classifiers with no Ontology	Identifiers
MD Model	++	—	—	—
Featural Models	++	—	—	—
Network Models	+	++	—	—
Custom Geometric	+	+	++	—
Synonym Lookup	+	+	+	+
String-Matching	+	+	+	++

Table 3.3: Alternative approaches to similarity assessment for the various cases of nominal attributes. A + represents a feasible approach for a case, whereas a ++ represents the recommended approach for that case. A – means that the approach does not apply.

### 3.8 Similarity Assessment for Cyclic Values

Ratio, interval, and ordinal values are typically thought of as being ordered along a straight line. Certain attributes, however, have values that are best conceptualized when positioned on a circle's perimeter. Such attributes are called *cyclic* (Chrisman 1995). Cyclic attributes order values such that the last element in a sequence coincides with the first element of the next round. The values can be either continuous or discrete, and can be represented by either points or intervals. The partitioning of the year into seasons or the week into days are examples where the values form discrete cyclic intervals. Although seasons and days can also be perceived as nominal values, sometimes their periodic order of succession is relevant. Examples of non-temporal cyclic attributes include angles (Isli and Cohn 1998) and the set of qualitative cardinal directions {N, S, E, W, NE, SE, NW, SW} (Frank 1996). The latter have been investigated as binary relations involving a reference and a target object (Goyal and Egenhofer 2001). The values of angles are continuous, whereas those of cardinal directions are discrete.

Cyclic scales are particularly interesting as they do not classify neatly within the ratio, interval, ordinal, and nominal scale typology, yet they are capable of exhibiting characteristics innate to each of these types of measurement. Hence, they are also capable



of implying different kinds of change. Cyclic values that can be represented as points or as non-overlapping intervals of an equal length are called *uniform* (Figure 3.8). In this case, the measurements resemble those on an interval scale: the values are ordered, they are separated by equal intervals, and the position of the zero point appears to be arbitrary. Although angles can be multiplied and divided, an angle of zero degrees does not indicate absence of an angle or absence of direction. Hence, the notion of quantity, as defined for ratio measurements, is inapplicable. The transition from one cyclic value to another represents a different kind of change than quantitative. For angles or cardinal relations, the change that is pertinent to the phenomenon being measured is change with respect to place, or movement, since an object moving cyclically changes its directional relation with the observer (i.e., the center of the cycle). If the values were of temporal nature instead, the change would have been of a qualitative nature as explained in Section 3.5 (although the perspective of movement through time would also be valid).

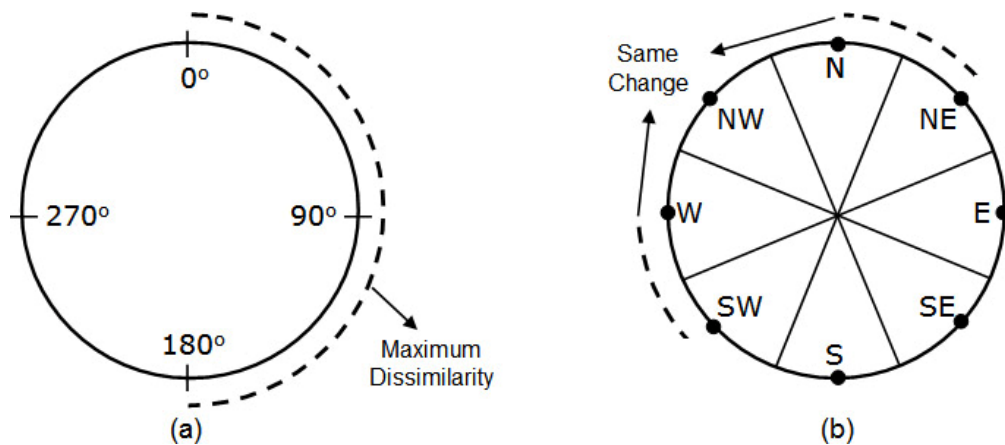


Figure 3.8: Cyclic scales with uniform values: (a) angles and (b) cardinal directions.

In both cases, equal distances along the perimeter indicate equal amounts of change. Dissimilarity can be taken equivalent to the length of the arc that must be traversed along the circle to join the two values under comparison. For any pair of values, there are two such arcs, one clockwise the other counter-clockwise. We choose the smaller arc, based on the criterion of minimum change (Section 2.2.2.6). When the values correspond to

intervals (Figure 3.8b), the length of the arc can be measured from the midpoints, starting points, or endpoints of the intervals, as long as the choice of the point to which the intervals are reduced remains consistent. Since the origin and the end of a cyclic scale coincide, the two contrary states of change that indicate maximum dissimilarity correspond to anti-diametrical points or intervals on the circle. Therefore, a unique characteristic of cyclic scales with uniform values is that each value has an exact opposite (although for values corresponding to intervals, this assertion holds only when the number of intervals is even). Similarity is computed from Equation 3.9, where  $P$  is the total length of the circle's perimeter. Both  $P$  and the absolute difference between the query and database values are expressed in the units of the cyclic attribute (e.g., for cardinal directions  $P = 8$  and  $(SW - NE) = 4$ ).

$$S(x_{db}, x_q) = f_{inv} \left( \frac{2 \cdot \min(|x_{db} - x_q|, (P - |x_{db} - x_q|))}{P} \right) \quad (3.9)$$

Uniform values do not exhaust all possibilities, since cyclic values may also correspond to intervals of unequal length (Figure 3.9). Movement alone is then insufficient to make the two values identical. The interval of the query value may also need to expand or contract by a certain amount, thereby undergoing also quantitative change. Under this setting, cyclic values can be viewed at a nominal level of measurement, differing along two constituent dimensions: position and size. Positional similarity is derived by Equation 3.9. The values along the second dimension represent quantities and are at a ratio level of measurement (e.g., the time periods of Figure 3.9). Similarity along the second dimension is computed by Equation 3.4. The overall similarity score between two cyclic values is then produced by combining the similarity scores in each dimension. Aggregating individual similarity scores is studied in chapter 4.

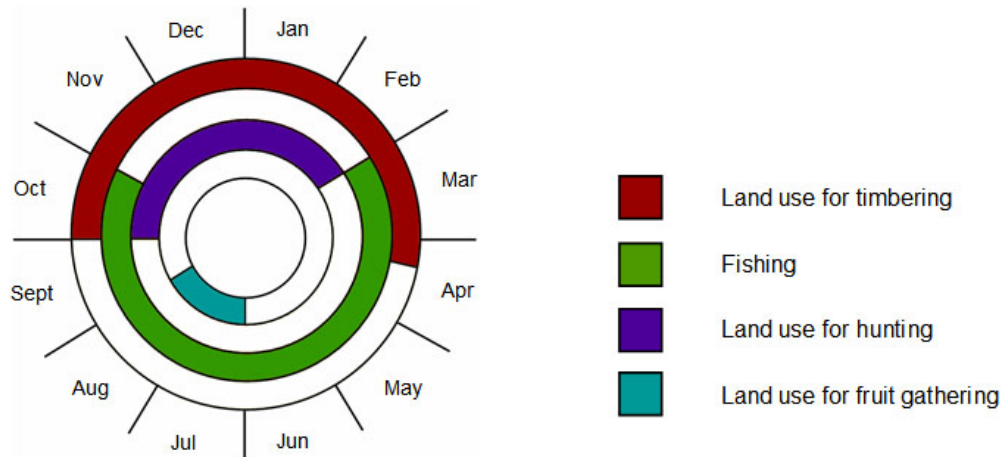


Figure 3.9: Four different periods of land use including timbering, fishing, hunting, and fruit gathering (Hornsby *et al.* 1999).

### 3.9 Attribute Considerations beyond the Five Levels of Measurement

The algorithms for the five levels of measurement (i.e., nominal, ordinal, interval, ratio, and cyclic) can accommodate the majority of attributes encountered in a database. The choice, however, may not always be intuitive. For some attributes, the classification into one of the measurement types may be abstruse or depend on each individual's interpretation of the data. Other attributes may require their own unique custom algorithm to be implemented. Capturing all attributes that exhibit such behavior is infeasible; therefore, we present characteristic examples of such cases that will serve as exemplars for similar situations. In most cases, the specialized similarity algorithms that need to be implemented for such attributes consist of a combination of primitive algorithms that were developed for the five levels of measurement.

Perhaps the most striking example of attributes that accept multiple interpretations is provided by temporal attributes. From a semantic perspective, two different views are usually adopted for temporal data types: the linear view and the cyclic view (Frank 1998). In the linear view, time events are points at some granularity and are represented on the dimension of time. For example, the purchases of land parcels or the construction

year of buildings may be respectively associated with the timestamps when the purchases occurred or the year the construction of a building was completed. These time values should be classified as interval. When the measurement pertains to duration and time values represent periods or are interpreted as such, then these time values should be classified as ratio, however. In the cyclic view, the values are associated with recurrent processes and, hence, should be treated as cyclic. For example, a weather database may record the day of the time that the maximum temperature occurred or the month of the most intense precipitation for each year. Cyclic events of arbitrary durations become nominal values differing not only with respect to their time of occurrence, but also with respect to their duration. The same is true for linear time intervals (Allen 1983).

There are additional examples of attributes with special innate semantics that prohibit their immediate classification under a scale of measurement. For example, a geographic database containing information about the lakes in the state of Maine may have an attribute *Average\_pH\_Value* for each lake. The pH value is a chemical term, which indicates the acidity of a liquid or a liquid body. It is measured on a closed scale, ranging from 0 to 14. This scale appears to be ratio, but these numbers are actually ordinals because the acidity is a logarithmic function of the amount of hydrogen ion concentration in the liquid body. Furthermore, the pH scale groups bodies of liquids into two classes: *acids*, if the pH value is less than 7, and *bases* or *alkalines* if the pH value is larger than 7. Acids present some general chemical properties that differ from the chemical properties of the bases. For example, most acids have a characteristic sour taste and act corrosively when they come in touch with the skin, whereas most bases have a bitter taste and produce a slippery or soapy feeling when applied to the skin. It might be desirable to reflect this difference of the properties of the two groups in the adopted similarity algorithm. From this perspective, a pH value becomes a nominal value. The similarity measure is not only a function of the distance of two values on the pH scale, but also dependent on whether the value classifies the water body as acidic or basic.

### 3.10 Null Values in Similarity Assessments

Null values refer to attributes that have no value stored. Database theory recommends the elimination of null values through proper database design and normalization (Elmasri and Navathe 2000); however, even in the most carefully designed systems null values are often unavoidable due to an inability to collect all required information about an entity, schema restructuring, or tradeoffs between performance and normalization. Null values waste space, lead to problems with relational *JOINS*, and database functions such as *COUNT* and *SUM*. Most importantly, they introduce ambiguities related to the meaning of the missing attribute values. The concern is to address the implications that derive from null values when such values are encountered in a similarity assessment.

A rudimentary way of dealing with null values is to assign a zero similarity measure between two values when one of them is null (Richter 1992). Another crude approach is to substitute a null with a precise extreme value, which is meaningless in the context of the attribute domain (Date 1982). This approach misses the different semantics that a null value may carry—for instance, up to 14 different types as reported in the ANSI/SPARC interim report (Bachman *et al.* 1975). Only a subset of three different interpretations, however, is vital for a formal treatment with respect to their meaning.

- *Unknown* null values were initially investigated by Codd (1979). An unknown value (*unk*) states that a precise value exists, but is currently missing. Specializations of *unk* nulls include *p-domains* and *p-ranges*. Both refer to a subset of the attribute's domain. A *p-domain* (Lipski 1979; Imielinski and Lipski 1984) implies that the unknown value, although missing, is restricted to a value in a subset of the attribute's domain; for instance, {2,4,7} is an example of a *p-domain* for a numeric attribute. In addition to *unk*, and *p-domains*, Morrissey (1990) also considered *p-ranges*, which state that the missing value is within a particular range. For example, a *p-range* of (20,50) means that the precise value is between 20 and 50. A *p-domain* applies better to attributes with an enumerated domain of finite elements, whereas a *p-range* is more

suited to attributes whose values vary along a continuum. In response to a query, one set of objects captures the exact matches, while another set captures objects with one or more null values that could possibly be exact matches. While this approach is concerned with the retrieval of possibly exact matches, this work is interested in finding similar results that, among others, encompass possible exact matches.

- *Non-applicable* (or *dne* for does not exist) nulls (Vassiliou 1979; Codd 1986), mean that the value is unavailable, because the specific attribute is not applicable for an object. Attributes that are applicable for several but not for all the entities of a class are called *partial* (Kusters and Borgida 2001).
- *No-information* nulls (*ni*) (Zaniolo 1982) are more generic, subsuming *unk* and *dne* types of nulls. They state that the value is missing either because it exists but is unknown or because it does not apply for that object. An *open ni* value includes the possibility of more than one existing but unknown values for a property of an object (i.e., a multi-valued property) (Gottlob and Zicari 1988). No-information nulls are conceptually simpler but less informative. Their use may result in loss of potentially useful information, since such nulls are unable to express the full spectrum of semantic interpretations that null values may have. For example, it is not possible to retrieve the set of objects for which an attribute does not apply.

Such different meanings imply that a successful treatment of null values relies on the simultaneous consideration of these types. To handle efficiently the different semantics of nulls, DBMSs must extend the domain of attributes in the system with the codes *unk*, *dne*, and *ni*, rather than using only the generic code *null*. Similarity between a null value and any other value of an attribute  $A$  may be derived from Equation 3.10, where  $a$  and  $b$  are the respective minimum and maximum values that define the range of  $A$ ,  $x_q$  is the query value, and  $S_A(a, x_q)$  and  $S_A(b, x_q)$  are measures of similarity between  $a$  and  $x_q$  and  $b$  and  $x_q$  respectively.

$$\forall x_q \mid (x_q \neq dne, unk, ni),$$

$$S_A(null, x_q) = \begin{cases} \min(S_A(a, x_q), S_A(b, x_q)) & \text{if } null = unk \\ 0 & \text{if } null = dne \\ 0 & \text{if } null = ni \end{cases} \quad (3.10)$$

A comparison between a *dne* and a query value is indeterminate, because *dne* does not exist, whereas the query value exists. This existence vs. non-existence of a value can be interpreted as the maximum possible dissimilarity and, therefore, a similarity measure of 0 is assigned to the pair of *dne* and any query value. *Dne* nulls are particularly useful in comparisons of objects that are not described through the same set of attributes. In such cases, missing attributes of one entity can be assumed to be present, and instantiated with *dne* nulls. An exception applies in some cases, where a *dne* value should be best substituted by a zero. For instance, when looking for employees that receive a low salary, volunteers could be considered as employees that receive a \$0 salary and, therefore, be retrieved as similar results to the query. In other cases, a *dne* specification would apply much better; for example, when comparing a lake to a building, and the lake has a *pH* value. In this scenario, any *pH* value for the building other than *dne* would be absurd. This distinction closely resembles the two different treatments of the zero value for ratio attributes (Section 3.4.1). Such issues constitute engineering choices that should be addressed during database design by the database administrator/designer.

Unlike *unk* and *ni*, a *dne* mark should always be treated by the database as a precise value, whether it is encountered in a stored object or used as a query. *Unk* and *ni* nulls, on the other hand, are treated as precise values only when a user queries the system by using them. Such queries are meaningful in the sense that the user may be looking for all missing values in the database in order to update them. In this case, a symbolic matching is necessary. In all other cases where *unk* and *ni* values are compared with precise query

values, they should be treated as placeholders instead and follow the substitutions (Equation 3.10). Here, the matching is semantic, rather than symbolic (Codd 1986).

The *unk* code represents knowledge that the actual value, although missing, belongs to the set of values that are allowed in the attribute range (Lipski 1979). Due to uncertainty, Equation 3.10 assumes minimum similarity and, therefore, substitutes *unk* with the domain value that maximizes the distance from the query value  $x_q$ . In cases of quantitative attributes, this value is logically either the minimum or the maximum as implied by the domain of the attribute or as specified via an explicit constraint. Hence, only two results need to be evaluated. For qualitative attributes, the algorithm may perform only when one deals with a finite domain of values. In this case, however, all values have to be checked in order to choose the one that minimizes similarity. If the user queries specifically for *unk* values, no substitution takes place and *unk* values are the only exact matches, followed by *ni* values.

The *ni* value is a lower-level placeholder for either *unk* or *dne* nulls and is the least informative. For any query where *ni* values are encountered (excluding the case when the query value is *dne*) a worst-case scenario is chosen, where *ni* values are treated as *dne* values and thus assigned zero similarity. During output presentation, however, tuples with *ni* values must be ranked higher than *dne* in terms of similarity, because they leave open the possibility of existence. If the query asks to retrieve specifically the tuples that have a *dne* value for the attribute instead, then the order is reversed, since *dne* values are exact matches and *ni* values the next best results, with everything else excluded. Such types of null-retrieving queries are typically performed by administrators for database maintenance purposes. In more realistic scenarios that account for the vast majority of database queries, users will enter precise values, and retrieve similar results, free of nulls.

For an example of queries involving null values, consider the relation in Table 3.4. Each record stores information about the type of the accommodation, the category of luxury, the total number of rooms, and the restaurant types within the establishments. Let



the range of possible rooms for accommodations vary from 5 to 70 and explicitly stated so by a constraint. The query *\*Type(hotel) and \*Restaurant\_Type(Greek) and \*Rooms(50)* requires similarity assessments with null values. *Dameia Palace* is a good result, because it is a hotel, the value for beds is relatively close to that of the query, and an *Italian* restaurant—also Mediterranean cuisine—exists on its premises. *Caldera Apartments* would be the second best match, followed by *Santorini Palace* and *Sun Rocks*. The reason for *Santorini Palace* being ranked so low is its *unk* value for rooms. This value will be substituted with number 5, since this is the value in the allowable range for *rooms* that minimizes similarity. If, however, there was a database constraint stating that hotels of category *A* must have between 40 and 70 rooms, then *unk* would be substituted by the number 70, yielding an ordering in which *Santorini Palace* is the most similar result, followed by *Dameia Palace* and then *Caldera Apartments*. *Sun Rocks* is the least similar match, because it is not a hotel and has no restaurants. The similarity between the query value for a *Greek* restaurant and the *dne* value would evaluate to zero.

Name	Type	Category	Restaurant_Type	Rooms
Sun Rocks	Apartments	B	<i>dne</i>	10
Dameia Palace	Hotel	A	Italian	70
Caldera Apartments	Apartments	A	Italian	30
Santorini Palace	Hotel	A	Greek	<i>unk</i>

Table 3.4: Relation *accommodations* with attributes that include null values.

This approach offers a semantically enhanced and elegant method when dealing with null values, especially when combined with consistency constraints that may be inserted as rules in the database and reduce the uncertainty for certain facts. Specifying the types of null values with different codes allows for more expressive power, both during the modeling of a database, as well as during the retrieval from it. The procedure adopts a pessimistic view when encountering *unk* values, by substituting *unk* with the most dissimilar value possible. Approaches based on probabilities, information content, or entropy (Morrissey 1990) do not apply for similarity assessments as they aim at locating

probable exact matches. For example, if the values of two tuples in the database are the *p*-domains {Greek, Chinese}, {Greek, Italian} and a query asks for a Greek restaurant. Since Italian cuisine is more similar to Greek cuisine than to Chinese, it is logically inferred that the second *p*-domain is always a better similarity match for the query. However, information content or entropy measures would yield equal estimates when assessing the probability of whether these two values are exact matches or not.

### 3.11 Summary

The relation of change to similarity is a close one. The similarity between two attribute values can be interpreted as the inverse of the change required to make the two values identical. Based on the ratio, interval, ordinal, nominal, and cyclic typology of measurements we described algorithms that yield a similarity measure between a query and a database value by assessing and measuring the type of change that the level of measurement implies (Table 3.5). In support of our methods, we also developed a rationale for reasoning with null values by denoting the semantics of different types of unavailable values with explicit identifiers that imply different degrees of uncertainty. Appropriate normalization techniques for each attribute type enable meaningful inferences when individual similarity scores need to be integrated. Complex attributes with rich semantics, such as nominal or cyclic values, may require a combination of similarity algorithms. Nominal values in particular are strongly dependent on the quality of the underlying representational structure.

<b>Attribute Types</b>	<b>Section</b>	<b>Equations for Similarity Comparisons</b>
Quantitative Ratios	3.4.1	Eqn. 3.2 or Eqn. 3.4
Qualitative Ratios	3.4.2	Eqn. 3.5
Interval Values	3.5	Eqn. 3.5
Ordinal Values	3.6	Eqn. 3.8a or Eqn. 3.8b
Classifiers (Detailed Ontology)	3.7.1	Eqn. 2.8 (modified for Hyponymy)
Classifiers (Basic Ontology)	3.7.1	Path Length, or Eqn. 2.5, or Eqn. 2.6, or Eqn. 2.7b (2.7b preferred)
Classifiers (no Ontology)	3.7.1	Geometric Decomposition, or Synonyms from WordNet
Boolean Classifiers	3.7.2	S=1 for same values S=0 for different values
Nominal Identifiers	3.7.3	String Matching Algorithms
Uniform Cyclic	3.8	Eqn. 3.9
Non-Uniform Cyclic	3.8	Eqn. 3.9 and Eqn. 3.4
Null Values	3.10	Eqn. 3.10

Table 3.5: The different attribute types, their corresponding chapter sections, and the recommended methods for performing similarity assessments between their values.

Under typical circumstances, the way in which the measurement was conducted will dictate the type of change being measured and, consequently, the level of measurement at which a similarity assessment occurs. This correspondence is not always clear. In some cases, the level of measurement depends not only on the data, but on the question asked and what one concludes from it. The solutions based on the notion of change contribute a sound framework for measuring similarity at the attribute level, reasoning about the appropriateness of existing similarity models, and capturing inherent properties of similarity, such as asymmetry.

## CHAPTER 4

### SEMANTIC SIMILARITY AMONG OBJECTS

The algorithms of Chapter 3 return similar results for equality-constrained queries on atomic values of a single attribute. Examples include a query to retrieve a spatial entity that occupies a certain area or a query for a particular lake. Queries, however, may link simultaneously a number of attributes through the combination of multiple constraints. This chapter develops a consistent and comprehensive methodology for spatial similarity retrieval in response to such complex queries formed by combinations of relational and logical operators. Relational operators refer to such predicates as *greater than* or *less than*, whereas logical operators combine separate spatial constraints using such connectives as *and*, *or*, and *not*. Multiple interacting constraints also raise the requirement for an effective weighting scheme that captures the users' personal intentions with minimal interaction, yet preserves the fidelity of the results to these intentions.

#### 4.1 Queries Expressed through Relational Operators

Relational operators extend the concept of an exact match to that of a range match. Besides the equality operator, relational operators determine whether one value is greater or less than another. They are denoted by the symbols  $>$  (i.e., greater than),  $\geq$  (i.e., greater than or equal to),  $<$  (i.e., less than), and  $\leq$  (i.e., less than or equal to). Specifying queries with relational operators is meaningful only on terms that have a natural order on a scale; therefore, their usage applies to ratio, interval, ordinal, and—in some cases—cyclic attributes.

The equality operator defines a single query value  $x_q$ , whereas a relational operator specifies a query range  $R_q$  with endpoints  $r_1$  and  $r_2$ . The range may be a closed or an open interval. For instance, in a query with  $x \geq 100$ ,  $r_1$  is the number 100 and  $r_2$  is plus infinity. Similarity between the range  $R_q$  specified by the user and any database value  $x_{db}$  of an

attribute  $A$  is derived by Equation 4.1, where  $S_A(x_{db}, r_1)$  and  $S_A(x_{db}, r_2)$  are measures of similarity between  $x_{db}$  and  $r_1$  and  $x_{db}$  and  $r_2$ , respectively. If an attribute value  $x_{db}$  is contained in the range  $R_q$ , then it is an exact match and, therefore, that attribute value receives a similarity measure of 1. If  $x_{db}$  is outside of the range, then its similarity is determined by the algorithm chosen for the attribute (Chapter 3). Relational operators are typically pertinent only to quantitative attributes where similarity is derived as a function of distance. In order to estimate the distance, we choose from the range of values that constitute exact matches the one that is closer to  $x_{db}$ . This value will logically be either the minimum or the maximum value of the range  $R_q$  (i.e., either  $r_1$  or  $r_2$ ).

$$S_A(x_{db}, R_q) = \begin{cases} \max(S_A(x_{db}, r_1), S_A(x_{db}, r_2)) & \text{if } x_{db} \notin R_q \\ 1 & \text{if } x_{db} \in R_q \end{cases} \quad (4.1)$$

For example, if a query requests all land parcels that occupy an area between 4,000 and 6,000 square feet (i.e.,  $r_1 = 4,000$  and  $r_2 = 6,000$ ), then every land parcel whose area is within the specified interval is an exact match. The similarity for land parcels with an area  $x_{db}$  outside of the interval is a function of the distance from  $x_{db}$  to  $r_1$  if  $x_{db}$  is less than 4,000, or from the distance of  $x_{db}$  to  $r_2$  if  $x_{db}$  is greater than 6,000 (Figure 4.1). In both cases, similarity is calculated by performing the appropriate substitutions in Equation 3.4.

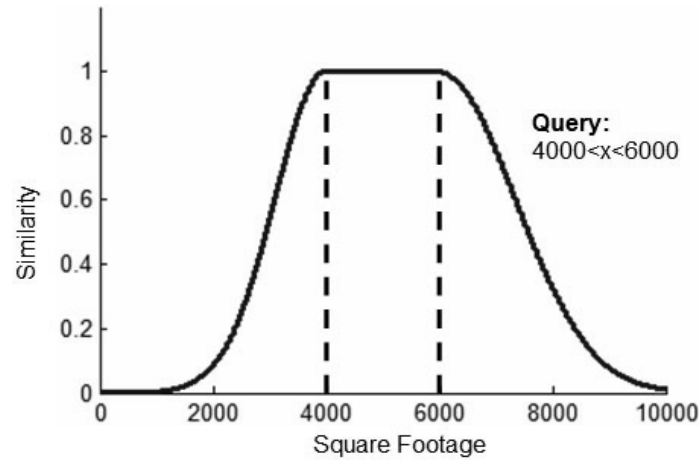


Figure 4.1: Similar results to a query involving relational operators.

## 4.2 Queries Expressed through Logical Operators

Querying a system with logical operators is based on concepts from Boolean algebra. Conjunctive queries refer to the combination of constraints using the logical operator *and*; for instance, “Find objects where attribute *A* has value *x* *and* attribute *B* has value *y*.” Similar combinations can be obtained with the use of disjunctions (*or*-operator) and negation (*not*-operator). The evaluation of each constraint yields a separate similarity value, so that the key issue becomes how to combine the similarity values.

### 4.2.1. Queries with *AND* on Different Attributes

The use of *and* requires that all the values that it connects be present in the results. Terms (i.e., constraints) joined by the *and*-operator are called conjuncts. In a typical *and*-query the conjuncts are values of two or more different attributes; therefore, the operator *and* is used to allow queries that simultaneously engage several attributes of an object. This usage of the *and* connective is particularly important, because it allows the extension of the similarity framework from the attribute to the object level, where two objects need to be compared globally with respect to multiple features. Furthermore, the manner in which constraints interact with one another and the order in which they are evaluated may vary depending on the tasks that users seek to accomplish. To guarantee the tractability of the framework, a detailed treatment is necessary that gives users the possibility of embedding diverse semantics into a conjunctive query.

A first step to an enhanced functionality is the separation of constraints (conjuncts) into those that are *required* or *hard* and those that are *preferential* or *soft*. Hard constraints accept only exact matches, whereas soft constraints can also accept similar results. The provision for the former is important since it maintains compatibility with standard database queries and allows the execution of tasks where similar results may be unacceptable. An example is the retrieval of buildings in violation of environmental regulations in order to be fined or demolished. In addition, hard constraints are more

efficient to process, because the similarity calculations are restricted only on the subset of database tuples that fully satisfies their union.

Additional semantics that facilitate the expression of diverse user objectives can be captured through different interaction modes among the constraints. Two variants are possible based on whether some constraints have total or partial dominance over others: (1) those that require *locally-better* results and (2) those that require *globally-better* results.

#### 4.2.1.1 Locally-Better Conjunctive Matching

Locally-better matching is based on the concept of *constraint hierarchies* (Borning *et al.* 1987; Borning *et al.* 1992). The constraints are organized by the user in a *constraint hierarchy* of depth  $n$ , where the different levels imply different degrees of preference. Constraints at a higher level are more important than constraints at a lower level. The levels of the hierarchy are assigned sequential integers with 0 denoting the highest level and  $n-1$  the lowest. Required constraints are placed at the zero level. Constraints at all other levels are preferential. If all constraints are placed at the zero level then only exact matches are acceptable. Otherwise, one database object is a locally-better match to a query than another, if for each of the constraints through some level  $k-1$  their values are identical and at level  $k$  the dissimilarity is strictly less for at least one constraint and less than or equal for all the rest. Hence, in locally-better matching, higher-level constraints have total dominance over lower-level constraints. Deviations from the query value at a certain level in the hierarchy are used to break the ties between results at the immediate higher level (Figure 4.2). Since sorting, rather than combining similarity values, plays the primary role in locally-better matching, the deviations can be calculated by the dissimilarity functions that are assigned to each attribute (Chapter 3) and the results can be ranked accordingly.

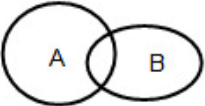
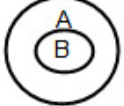
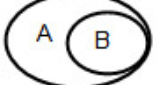

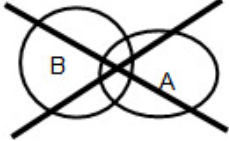
Query:	Required	Preferential	Preferential
	(level 0)	(level 1)	(level 2)
	$*area(A) > area(B)$	<b>and</b> $*commonArea(30m^2)$	<b>and</b> $*top\_Relation(A \text{ overlaps } B)$
<b>Results</b>			
1. 	$area(A) > area(B) := true$	$commonArea = 40m^2$	$top\_Relation(A, B) = inside$
2. 	$area(A) > area(B) := true$	$commonArea = 50m^2$	$top\_Relation(A, B) = covers$
3. 	$area(A) > area(B) := true$	$commonArea = 51m^2$	$top\_Relation(A, B) = overlaps$
	<b>rejected</b> $area(A) > area(B) := false$	$commonArea = 30m^2$	$top\_Relation(A, B) = overlaps$

Figure 4.2: Similar results to a conjunctive query using locally-better matching.

#### 4.2.1.2 Globally-Better Conjunctive Matching

Globally-better matching relies on a compensatory use of the *and* operator and follows principles from geometric models of similarity. According to such models, the similarity of one object to another is an inverse function of the distance between the objects in a conceptual space. The use of attribute weights that indicate each dimension's salience within the space offers a refinement of this process. The distance in a conceptual space indicates dissimilarity. A measure of the latter should be compatible with human judgments of overall dissimilarity and its correct calculation becomes, therefore, important. Following widely accepted psychological research (Attneave 1950; Torgerson 1965; Shepard 1987;1988; Ashby and Lee 1991; Nosofsky 1991;1992; Gärdenfors 2000), the perceived interpoint distances between the objects' point representations in the space should be computed either by Equation 4.2 or 4.3, where  $n$  is the number of dimensions and  $x_{ik}$ ,  $x_{jk}$  are the values of entities  $i$  and  $j$  on dimension  $k$ . Dividing by the sum of the



weights ensures that the final measure is bounded within 0 and 1. Equation 4.2 corresponds to a Euclidean metric. The distance is defined as the shortest path along a straight line between points  $i$  and  $j$ . Equation 4.3, on the other hand, corresponds to the city-block metric where the distance between the two points is defined as the sum of their distances on the individual dimensions.

$$Dissimilarity_E(i, j) = \frac{\sqrt{\sum_{k=1}^n w_k \cdot (x_{ik} - x_{jk})^2}}{\sqrt{\sum_{k=1}^n w_k}} \quad (4.2)$$

$$Dissimilarity_C(i, j) = \frac{\sum_{k=1}^n w_k \cdot |x_{ik} - x_{jk}|}{\sum_{k=1}^n w_k} \quad (4.3)$$

Whether one employs Equation 4.2 or 4.3 depends on whether one deals with integral or separable dimensions. Integral dimensions are strongly unanalyzable and typically perceived as a single stimulus. For instance, the proximity of two linear objects may be described with a number of measures that associate the boundaries and interiors of the objects (Nedas *et al.* in press), but the closeness relation may be perceived as one stimulus from the users that inspect the lines. Another example includes color, where one cannot assign a value for an object in one dimension (i.e., brightness) without doing so for the others (i.e., hue and saturation). Hence, a set of integral dimensions constitutes in essence one multi-dimensional attribute (Torgerson 1965). Separable dimensions, on the other hand, are different and distinct properties (e.g., length and height) that are perceptually independent (Ashby and Lee 1991). It has been suggested and experimentally confirmed (Attneave 1950; Torgerson 1965; Shepard 1987) that, with respect to human judgments for similarity, a Euclidean metric performs better with integral dimensions, whereas a city-block metric matches more closely separable dimensions.

Perceptually separable dimensions are expected to have a higher frequency of occurrence in databases; therefore, in the general case the composite dissimilarity indicator between two objects will be calculated by the weighted average of individual dissimilarities along each of the dimensions (Equation 4.3). For a group of  $n$  integral attributes, however, an Euclidean metric (Equation 4.2) should be adopted to derive the dissimilarity of the objects with respect to this integral group. Therefore, the combination of the  $n$  concomitant attributes of an integral group should yield one dissimilarity component rather than  $n$  individual components in the composite measure. (Figure 4.3).

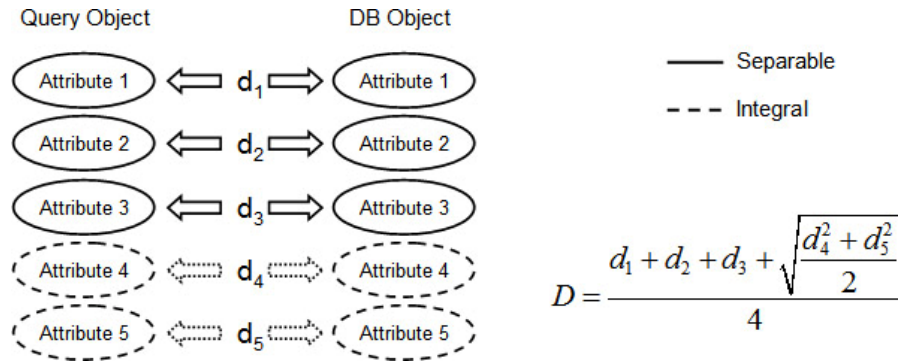


Figure 4.3: Combining two integral attributes to one that is separable (all weights are set to 1).

Converting composite dissimilarity to composite similarity can be done via any inverse monotonically decreasing function. As in the case for similarity assessments at the attribute level (Section 3.2.3), the choice of the conversion function also remains irrelevant at the object level. Different functions, such as linear (Equation 2.1a), exponential variants (Equation 2.1b), or Gaussian variants (Equation 2.1c), will affect the similarity scores for each tuple in the set of retrieved similar results, but the ordering from most similar to least similar object will be preserved. The choice of the aggregation function that yields dissimilarity matters, however, as different choices may produce divergent rankings. For instance, the employment of an Euclidean metric on separable dimensions or a city-block metric on groups of integral attributes is likely to distort the results. The extent of such distortions is investigated in detail in Chapter 6. The approach

of this thesis differs from other efforts in the literature (Motro 1988; Papadias *et al.* 1999b; Blaser 2000; Dey *et al.* 2002; Ortega-Binderberger *et al.* 2002; Stefanidis *et al.* 2002; Chakrabarti *et al.* 2003) in that it does not employ an Euclidean or city-block distance metric in an *ad-hoc* fashion, but introduces instead a *psychologically correct* dissimilarity measure that offers explicit treatment for separable and integral dimensions.

#### 4.2.1.3 Other Approaches to Conjunctive Matching

Additional methods for calculating the similarity to conjunctive queries include the productive combination (Ruttkay 1994) and approaches based on the fundamental scoring rule for fuzzy set intersections, which, for conjunctive queries, resorts to selecting the *minimum* of the similarity values produced for each attribute (Equation 2.9a). The problem with the productive combination is that it cannot differentiate between results that receive a zero similarity score for one of the conjuncts. The fuzzy-based approach that uses the minimum operator suffers from an even more compelling lack of discrimination among the retrieved output, because the rank of a retrieved item depends only on the lowest similarity measure (Santini and Ramesh 1997; Fagin 1998; Ramakrishna *et al.* 2002). Two objects *A* and *B*, for instance, would both score as 0.2 similar to a conjunctive query with three attributes if the similarities of object *A*'s and object *B*'s attributes to the query's attributes were (0.2,0.8,0.9) and (0.2,0.3,0.3), respectively. This seems counter-intuitive (Elkan 1993;2000), because object *A* is clearly a better match. In fact, most researchers who have used this measure seem to be somewhat troubled by their results. Santini and Ramesh (2000) report problems between judgments of similarity with their model and others that were experimentally obtained, and admit that the minimum is too restrictive for conjunction. The same is raised by Ortega *et al.* (1998), as well as Fagin (1998) who justifies the use of minimum because it has attractive properties that are useful in optimizing the algorithms for faster access to the database.

*Accuracy* and *correctness* of a computer-produced similarity measure and the suitability of an algorithm are only reflected in their fidelity to human behavior, perceptions, and intuition. In the realm of similarity it makes little sense to succumb to the niceties of a well-defined theory or model that does not comply with human reasoning. This argument is not to say, however, that fuzzy logic is flawed, but rather that the choice of minimum as a fuzzy intersection operator when reasoning for similarity is erroneous and counter-intuitive. As Goldstone (1994b) puts it “our most basic similarity computation appears not to be one of determining identity in a particular dimension, but one of determining proximity across many dimensions.” Klir and Yuan (1995) also stress this very point when urging for the careful selection of fuzzy operators so that they reflect appropriately the context of the application in which they are used. Under this perspective, the approach of this thesis is compatible with fuzzy logic, because it uses another valid function—the weighted average—as a fuzzy aggregation operator that combines in a desirable way several fuzzy sets to produce a single fuzzy set.

#### 4.2.2. Queries with *AND* on the Same Attribute

An alternative but rather unorthodox use of *and* occurs when the conjunction is used to connect values of the same attribute. If  $A$  is an attribute of a set of objects, the expression  $*A(x)$  and  $*A(y)$  means that the user wants to retrieve those objects for which the attribute  $A$  simultaneously attains the values  $x$  and  $y$ . This objective is not related to fuzzy variables to which an object may belong simultaneously with different degrees of membership (Cross and Sudkamp 2002), but rather it implies the presence of multi-valued attributes (i.e., attributes that have a set of values for an entity). Comparing a multi-valued property of two objects requires a different logic than comparing a single-valued property. The similarity measure in this case relates two sets of values, rather than two individual values, and describes how similar one set is to the other.

#### 4.2.2.1 Conjunctive Queries on Multi-valued Attributes

In order to calculate the similarity of two sets  $Q$  and  $H$ , the correspondences between compared values must be first established. Such correspondences can be based on the criterion of *optimum fit*, which seeks to maximize the sum of the individual similarity scores (or minimize the dissimilarities). This choice is justified by people's tendency to evaluate similarity from the perspective of the minimal change required to transform one of the compared things into the other (Section 2.2.2.6). In the case of multi-valued attributes, the criterion of optimum fit also captures indirectly a combination of principles from featural and geometric similarity models. Pairs of values that are common in the two sets have a similarity coefficient of 1 assigned to them so that they are likely to be included in the combination of pairs, which yields the maximum sum; therefore, common values are counted as common features and contribute significantly to the overall similarity of the sets compared. The remaining pairs, which consist of different values from each set, are not simply treated as distinctive features according to the binary logic of featural models, but are rather assigned a similarity score that indicates how different they are.

The two sets  $Q$  and  $H$  can be formally represented with a complete bipartite graph (Figure 2.4h), where each node in  $Q$  corresponds to a value of the query set, and each node in  $H$  to a value of the database set. A weighted edge from each node of  $Q$  to each node of  $H$  denotes the similarity for this pair of atomic values. The objective is to retrieve a maximum-weight matching from this graph (Figure 4.4a). If the edges of the bipartite graph indicate dissimilarities instead, then the objective becomes to minimize their sum. This alternative formulation is known as the assignment problem (Papadimitriou and Steiglitz 1998), which states: given a  $n \times m$  matrix, find a subset of the elements, exactly one element in each column and one in each row, such that the sum of the chosen elements is minimum (Figure 4.4b). For multi-valued attributes,  $n$  refers to the elements of the query set,  $m$  to the elements of the database set, and the  $n \times m$  matrix is the

dissimilarity matrix that contains the pairwise dissimilarities. The maximum-weight matching problem on a bipartite graph and the assignment problem are equivalent. One can easily formulate the former as the latter, simply by subtracting all edge weights (i.e., similarities) from a value larger than the larger weight. Although the possible permutations for sets of  $n$  cardinality are  $n!$ , efficient polynomial algorithms that can cope with multi-valued sets of reasonably large sizes (thousands of elements) exist (Goldberg and Kennedy 1995). The most famous is the *Hungarian algorithm* (Kuhn 1955) with a complexity of  $O(n \cdot (m + n \cdot \log n))$ . Once the pairs have been created, the overall dissimilarity of the sets can be computed (Equation 4.2 or 4.3) and converted to similarity through an inverse function (Equations 2.1a-c).

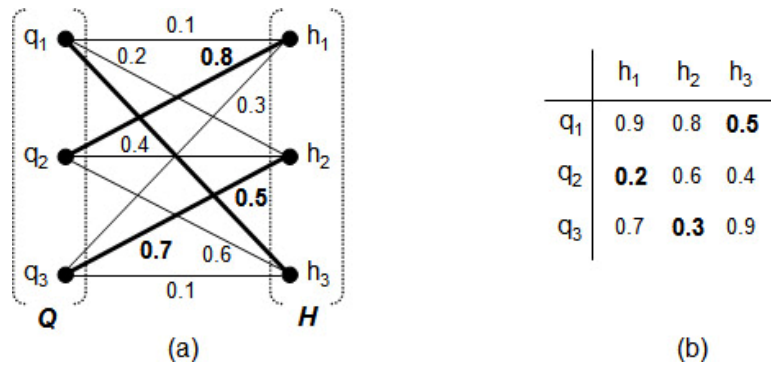


Figure 4.4: Formulating multi-valued attribute similarity as (a) the problem of maximum-weight matching in a bipartite graph and (b) the assignment problem.

This approach to multi-valued attribute similarity applies under all circumstances where two sets of values must be compared and the sequence of the elements in the sets is immaterial. This statement implies that the identity of the elements is irrelevant and, therefore, every element in one set may be matched with any element of the other set. Such comparisons may be necessary in a number of different scenarios in GISs, for instance, in the comparisons of detailed topological representations (Egenhofer and Franzosa 1995; Clementini and di Felice 1998). Detailed representations elaborate over their coarse counterparts (e.g., the 9-intersection) by describing a topological relation in

terms of multiple component intersections. Examples include the two possible ways to infer the similarity of the topological relation between the two configurations (Figure 4.5a) and the case of periods of time represented by disjoint temporal intervals (Figure 4.5b). To assess the similarity between the *hunting* period and the *fishing* period a multi-valued similarity assessment must be performed.

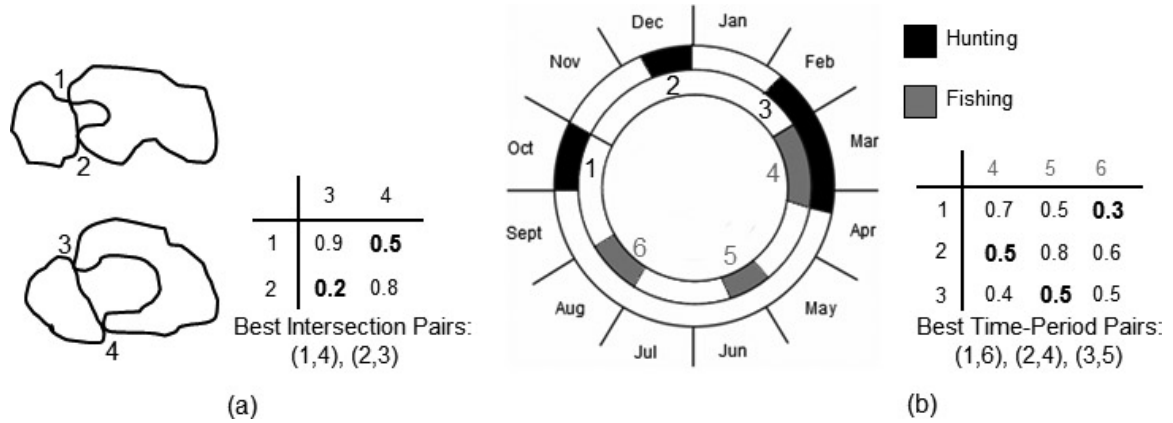


Figure 4.5: Applications of a similarity measure for multi-valued attributes: (a) detailed topological relations and (b) disjoint temporal intervals.

A problem arises when the sets have different cardinalities as it is questionable how to account quantitatively for the missing elements. For instance, if the second configuration of Figure 4.5a had only one intersection component (Figure 4.6), neglecting the additional elements of the set with the larger cardinality will lead to misleading similarity estimates. Therefore, a method is needed that accounts for the discrepancy in the number of values between the database and the query set. A simple approach to inflicting this penalty would be to extend the smaller set in the assessment with *dne* nulls, up to the cardinality of the larger set (Figure 4.6). The addition of the pair (1, *dne*) in the formula that yields the dissimilarity of the sets (Equation 4.2 or 4.3) reflects the existence of one additional intersection in *configuration 1* and produces a similarity estimate that corresponds better to the real-world situation.

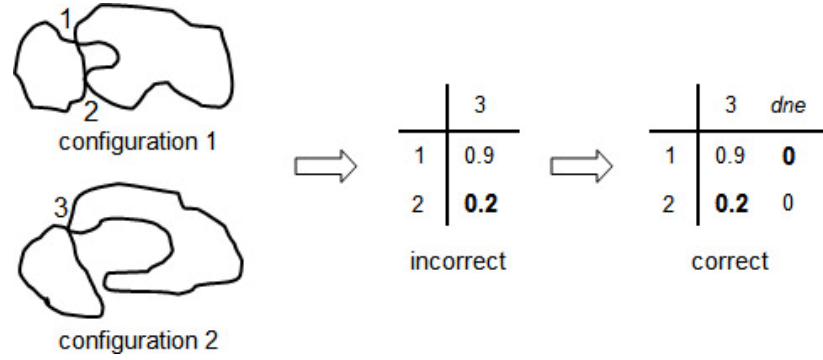


Figure 4.6: Multi-valued similarity for sets of different cardinalities.

Such cases of incomplete correspondences can be handled more flexibly and in a more general manner by introducing the *value completeness* parameter of two sets  $Q$  and  $H$ , denoted by  $V_{comp(H,Q)}$ . This parameter's value should be interpreted as the similarity of set  $H$  to set  $Q$  with respect to completeness. Its specification is based on the ratio contrast model, which allows expressing the completeness as a function of the matched and unmatched elements of the two sets, and its value is bounded in the interval  $[0,1]$ . The simpler approach considers each value of equal importance (Equation 4.4a), whereas a more elaborate version assigns a weight to each value (Equation 4.4b).

$$V_{Comp(H,Q)} = \frac{M}{M + \alpha \cdot (n - M) + \beta \cdot (m - M)} \quad (4.4a)$$

$$V_{Comp(H,Q)} = \frac{\sum_{i=1}^M w_{O_i}}{\sum_{i=1}^M w_{O_i} + \alpha \cdot \sum_{j=1}^{n-M} w_{O_j} + \beta \cdot (m - M)} \quad (4.4b)$$

where:

- $M$  : Number of matched elements
- $n$  : Number of elements in the query set  $Q$
- $m$  : Number of elements in the database set  $H$
- $\alpha$  : The weight of the subset of unmatched query elements
- $\beta$  : The weight of the subset of unmatched database elements



$w_{O_i}$  : The weight of the  $i$ -th matched query element

$w_{O_j}$  : The weight of the  $j$ -th unmatched query element

and  $\alpha, \beta \in [0,1]$

The value completeness measure is a flexible measure able to accommodate a number of different scenarios and to produce asymmetric similarities between two sets. Different user intentions can be captured by adjusting the weights  $\alpha$  and  $\beta$ . Three cases are possible:

- $m = n$  : When the cardinality of the sets is equal then all values in the query set will be associated with values in the database set (Figure 4.7a). In this case, the value completeness becomes 1, because the number of matched pairs  $M$  equals the cardinality of the sets.
- $n > m$  : When the cardinality of the query set  $n$  is larger than the cardinality of the database set  $m$ , then the number of associated value pairs  $M$  equals  $m$  (Figure 4.7b). The weight  $\beta$  plays no role in this case since the term  $\beta \cdot (m - M)$  is cancelled out. Setting  $\alpha$  to any value larger than 0 will inflict a penalty for completeness. Setting  $\alpha$  to 1 is equivalent to extending the cardinality of the database set with *dne* values up to the cardinality of the query set. For the bipartite description of the problem this setting translates to adding new nodes with edges of zero weight incident upon them.
- $n < m$  : When the cardinality of the query set  $n$  is smaller than the cardinality of the database set  $m$ , then the number of associated value pairs  $M$  equals  $n$  (Figure 4.7c). The weight  $\alpha$  plays no role in this case since the term  $\alpha \cdot (n - M)$  is cancelled out. Setting  $\beta$  to any value larger than 0 will inflict a penalty for completeness. A positive value in this case means that the user is interested in finding a database set that matches the query set exactly. A value of 0, in contrast, means that the user is interested in locating a database set with at least as many elements as those in the

query set. The interest is shifted only to the matched elements, while unmatched values in the database set are ignored.

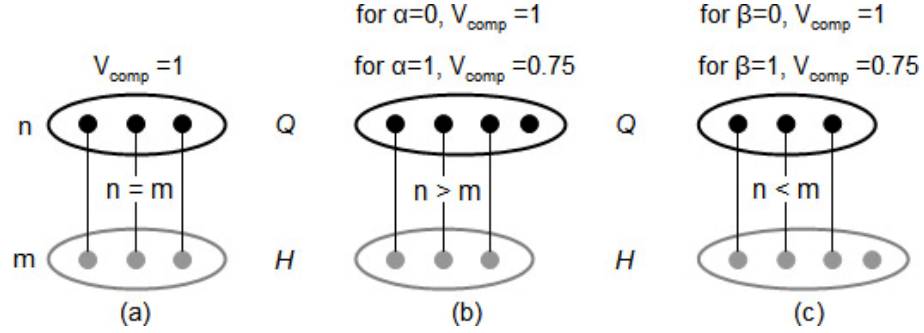


Figure 4.7: Behavior of the value completeness parameter for similarity queries that involve sets of different cardinalities: (a) the query and database sets have the same number of values, (b) the query set has more values than the database set, and (c) the query set has fewer values than the database set.

The final similarity between the two sets  $S_{(H,Q)}$  incorporates the completeness correction (Equation 4.5). The value completeness has a limiting influence on the similarity of the sets. If the weight of the value completeness is 1, then the set similarity cannot exceed the specification of the value completeness. The weight  $w_{Comp}$  for the completeness is distinct from the weights  $\alpha$  and  $\beta$  of Equations 4.4a and 4.4b. The former determines the degree to which completeness affects the final score, whereas the latter define what completeness means.

$$S_{(H,Q)} = S'_{(H,Q)} \cdot (w_{Comp} \cdot (V_{Comp(H,Q)} - 1) + 1) \quad (4.5)$$

where:  $S'_{(H,Q)}$ : The averaged similarity of the matched pairs only

$w_{Comp}$ : Weight of the value completeness parameter

#### 4.2.2.2 Conjunctive Queries on Composite Attributes

A special variation of the multi-valued attribute theme concerns composite attributes. A composite attribute can be divided into smaller subparts, which represent more basic

attributes with independent meanings (Elmasri and Navathe 2000). In this case, the identity of the elements becomes significant as there is an unambiguous correspondence between the elements of the compared sets. Hence, the methodology for composite multi-valued attributes is identical to that for globally-better conjunctive matching (Section 4.2.1.2), since such attributes can be treated as objects or as nominal values varying in several dimensions. Depending on whether these dimensions are integral or separable, Equations 4.2 or 4.3 can be respectively used to determine the dissimilarity of the sets.

An example is the *boundary closeness* measure (Nedas *et al.* in press), which applies to line-line relations, describing the remoteness of one line's boundary from the boundary of the other line. This attribute comprises a pair of normalized ratio values. The smaller value corresponds to the smallest realizable distance between boundary points of the two lines, whereas the larger value corresponds to the distance formed between the remaining boundary points. The distances are chosen such that the two sets of boundary points are mutually exclusive (Figure 4.8). The magnitude of the distances essentially prescribes an identity to each of the two values. The smaller distance can be thought of as the *minimum boundary closeness*, while the larger distance forms the *maximum boundary closeness*. Hence, when two pairs of lines are compared with respect to their boundary closeness, the correspondences during the similarity assessment become evident.

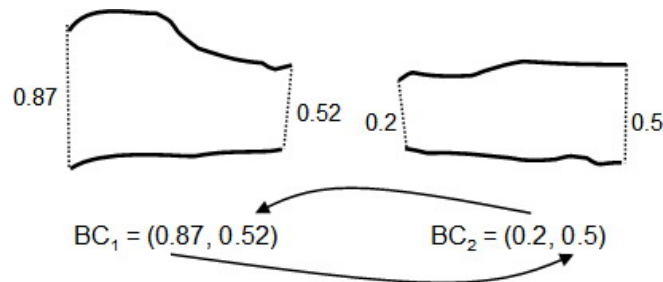


Figure 4.8: Establishing correspondences for multi-valued similarity of composite attributes.

### 4.2.3 Queries with OR on the Same Attribute

The use of the disjunction *or* requires that at least one of the values that it connects be present in the result. Terms joined by the *or*-operator are called disjuncts. In a typical *or*-query, the disjuncts are values of the same attribute. If  $A$  is an attribute of a class of objects, the expression  $*A(x)$  *or*  $*A(y)$  means that the user wishes to retrieve objects for which the value for attribute  $A$  is either  $x$  or  $y$ . As in the case with relational operators, there is not one query value, but a set of query values. The difference to queries expressed through relational operators is that the set of query terms is not represented by a range, but by a finite number of distinct values.

Similarity is derived from Equation 4.6, where  $Q = \{x_1, x_2, \dots, x_n\}$  is the set containing the  $n$  values that are connected by the *or*-operator in the query expression, and  $x_{db}$  is any stored value for attribute  $A$  in the database. If  $x_{db}$  coincides with any of the values in  $Q$  then it is an exact match and the similarity is 1. Otherwise, the process consists of examining the similarities between  $x_{db}$  and all the values that are elements in  $Q$ . Since all values in  $Q$  are exact matches, we choose the one that gives the largest similarity measure for  $x_{db}$ , when compared to it, that is, the similarity of  $x_{db}$  is determined by its distance from the closest exact match.

$$S_A(x_{db}, Q) = \begin{cases} \max(S_A(x_{db}, x_i), \text{ where } i = (1, \dots, n)) & \text{if } x_{db} \notin Q \\ 1 & \text{if } x_{db} \in Q \end{cases} \quad (4.6)$$

For example, for a query asking to retrieve buildings in downtown Bangor that occupy an area either of 400 or 600 square feet (i.e.,  $Q = \{400, 600\}$ ), every building whose area is 400 or 600 is an exact match. For buildings with a different area value  $x_{db}$  the maximum similarity measure obtained for the pairs  $(x_{db}, 400)$  and  $(x_{db}, 600)$  is chosen, as this is computed from the algorithm that has been assigned to attribute  $A$  (Figure 4.9).

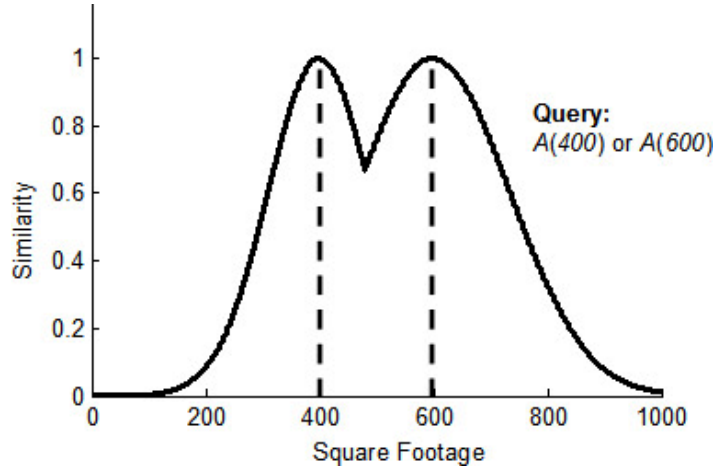


Figure 4.9: Similar results to a logical or-query involving one attribute.

This approach is equivalent to the standard scoring rule for fuzzy disjunction (Santini and Jain 1996; Ortega *et al.* 1998). The values of the query set can be interpreted as prototyping concepts of fuzzy sets, and the measures  $S_A(x_{db}, x_1), S_A(x_{db}, x_2), \dots, S_A(x_{db}, x_n)$  as indicators of the membership of the database values to these sets.

#### 4.2.4 Queries with OR on Different Attributes

Specifying queries with *or* where the disjuncts are values of different attributes constitutes uncommon practice, but is still a viable option for the database users. If  $A_1$  and  $A_2$  are two attributes of an object, the expression  $*A_1(x)$  *or*  $*A_2(y)$  means that the user wants to retrieve those objects for which the attribute  $A_1$  has the value of  $x$ , or those objects for which the attribute  $A_2$  has the value of  $y$ . The satisfaction of either of these constraints implies an exact match, therefore, the methodology is similar to that for disjunctive queries on the same attribute (Section 4.2.3).

If  $Q = \{ *A_1(x_q), *A_2(x_q), \dots, *A_n(x_q) \}$  is the set containing  $n$  query values ( $x_q$ ) of  $n$  different attributes connected by the *or*-operator, and  $H = \{ A_1(x_{db}), A_2(x_{db}), \dots, A_n(x_{db}) \}$  is the set containing the corresponding  $n$  database values for an object in the database, then Equation 4.7 yields the similarity score between the reference object  $O_q$ , characterized by

the query values of the user and any other object (i.e., record)  $O_{db}$  in the database. If a database object matches any of the values contained in  $Q$  for some attributes  $A_i$  then it is an exact match. Otherwise, we separately examine the similarity of all corresponding pairs  $(A_i(x_{db}), A_i(x_q))$  for all attributes  $A_i$  connected by the *or*-operator and the pair of maximum similarity is chosen.

$$S_A(O_{db}, O_q) = \begin{cases} \max(S_A(x_{db}, x_i), \text{ where } i = (1, \dots, n)) & \text{if } A_i(x_q) \notin Q \\ 1 & \text{if } A_i(x_q) \in Q \end{cases} \quad (4.7)$$

For example, consider the query *\*Relation(covers) or \*CommonArea(100)* (Table 4.1). Regardless of its value for the attribute *CommonArea*, configuration 1 is an exact match, because it matches the query value for the attribute *Relation*. Similarly, configuration 2 is also an exact match, because it matches the query value for the attribute *CommonArea*. For configurations 3 and 4 the similarities between their attribute values and the respective attributes of the query are calculated separately, and the larger score of each configuration is assigned as its overall similarity to the configuration specified by the query.

ID	Topological Relation		Common Area		Overall Similarity
Configuration 1	covers	100%	80	80%	100%
Configuration 2	contains	75%	100	100%	100%
Configuration 3	overlaps	75%	150	50%	75%
Configuration 4	meets	50%	0	0%	50%

Table 4.1: Similar results to a logical *or*-query involving two attributes.

#### 4.2.5 Queries with NOT

Values that the *not*-operator takes as arguments are missing in the results. If  $A$  is an attribute for a class of objects, the expression *\*not A(x)* means that the user wants to retrieve any object, except those that have a value of  $x$  for attribute  $A$ . The similarity between a database object  $O_{db}$  with a value of  $x_{db}$  for attribute  $A$  and the query object  $O_q$  characterized by the negation statement *\*notA(x<sub>q</sub>)* can be calculated by Equation 4.8.

$$S_A(O_{db}, O_q) = \begin{cases} 0 & \text{if } x_{db} = x_q \\ 1 & \text{if } x_{db} \neq x_q \end{cases} \quad (4.8)$$

Negations are another area where common fuzzy-based implementations of similarity (Section 2.3.1) to complex queries suffer. The effect of the standard fuzzy operator for negation (Equation 2.9c) is that it returns as most similar the objects that are the most dissimilar with respect to the value negated in the query. While such objectives may be best captured with different operator combinations, interpreting negation in this manner may not always align well to human reasoning, and may even return paradoxical results. For instance, if a traveler queries for a hotel, but not in the center of a city, then this query does not necessarily mean that she would like to find a hotel in the middle of the desert or on the top of a mountain, while one in the suburbs would be acceptable. Similarly, it is absurd to search for one land parcel containing another in response to user's request for finding non-disjoint land parcels. Therefore, the role of negations in information retrieval is to avoid undesirable associations or, in general, eliminate unwanted tuples from the set of retrieved results. Hence, it should be interpreted by a similarity query processor as it has always been interpreted traditionally in the classic logic paradigm.

An interesting situation occurs during the combination of a conjunction and a negation over the same constraint; for instance  $*A(x)$  and  $*not A(x)$ . This expression can be interpreted as “find the objects that simultaneously have and do not have the value  $x$  for attribute  $A$ .” Although in classic logic this is a contradiction, in a similarity setting it can be interpreted as a request to retrieve all similar results for a query, excluding those that are exact matches.

### 4.3 Attribute Weights

Whereas normalization removes the unintentional and persistent distance scale biases that are introduced in the data space from different attribute ranges, weights aim at reinserting biases—albeit deliberate and dynamic this time—so that the data space is aligned with

the user's conceptual space. Such an alignment is often required, because the central trait that influences similarity judgments is attention (Smith and Heise 1992). Selective attention to different properties changes the perceived similarity of two objects; therefore, the primary role of weight coefficients is to serve as context adjusters: a dimension weight determines the relative importance of that particular dimension on the composite score. A large value for the weight of a dimension stretches the space along that dimension, while a small value shrinks it. Hence, similarity is a function of both the magnitude of the difference between values of entities on the dimensions and of the dimension weights (Gärdenfors 2000).

Currently, no uniformly agreed-upon methodology exists to weighting individual dimensions before aggregating them into a composite measure. It is common practice to either assume equal weights on all attributes (Dawes and Corrigan 1974; Dawes 1979) or to rely on the users' explicit weight specification (Motro 1988; Blaser 2000). Translating, however, one's objective into a set of precise ratio values may be a challenging task as it assumes knowledge of what weights are and how they interact, but also mandates a precision that may be absent in the mind of the decision maker (Kirkwood and Sarin 1985; Borcherting *et al.* 1991). Weighting decisions may become even more abstruse and error-prone as the number of soft constraints increases (e.g., spatial scenes with multiple objects and attributes), forcing users to vacillate among their own judgments, or even worse, become unwilling to specify weights at all.

A better approach is offered by *rank-order* weighting methods (Barron and Barret 1996), which rely solely on ordinal information in order to derive ratio weights. The user's responsibility is reduced to ranking the constraints based on their importance. Providing ordinal preference is easier and more reliable than specifying exact values (Stillwell *et al.* 1981; Barron and Barret 1996). Hence, such an approach is more suitable for complex spatial information retrieval. The concept is peripheral to that of constraint



hierarchies (Borning *et al.* 1992) (Section 4.2.1.1), but differs in that higher-ranked constraints prevail, but do not dominate completely, their subordinates.

There exist several methods to convert rank information to ratio weights (Stillwell *et al.* 1981), however, the most effective and reliable is the *rank-order centroid* method (Barron 1992; Barron and Barret 1996; Jia *et al.* 1998), which interprets weights as defining the vertices of a simplex. For example, for two attributes the simplex is a straight line with coordinates (1,0) and (0,1). All points on this line have coordinate pairs whose sum is the unit value. Absence of knowledge about the weights is represented by a uniform probability density function on this line. The expected value of this distribution is the centroid of the line with coordinates (0.5, 0.5) and the values of this pair define the weights. Knowledge that the first attribute is more important than the second means that it should also receive a higher weight, therefore, we expect that  $0.5 \leq w_1 \leq 1$ . The expected value of the uniform probability density function over this interval is 0.75, therefore,  $w_1 = 0.75$ , which implies in turn that the value for  $w_2$  is 0.25. Equation 4.9, where  $w_k$  is the weight of the  $k$ -th dimension, generalizes this argument to  $n$  attributes.

$$w_k(ROC) = \frac{1}{n} \cdot \sum_{i=k}^n \frac{1}{i}, \quad k = 1, \dots, n \quad (4.9)$$

#### 4.4 Summary

This chapter elevated similarity comparisons from the attribute level to the object level, by developing a comprehensive model for dealing with complex similarity constraints expressed through relational and Boolean operators (Table 4.2). Current implementations of complex similarity assessments that use standard fuzzy logic operators have limitations, especially for conjunctions and negations. Although disjunctions perform realistically with a fuzzy logic interpretation of the *or*-operator, negations require a traditional logic interpretation. On the other hand, conjunctions require a pluralistic approach.

	Section	Methods for Similarity Comparisons
Relational Operators	4.1	Eqn. 4.1
Locally-Better Conjunction	4.2.1.1	Hierarchical Sorting
Globally-Better Conjunction	4.2.1.2	Eqn.4.2 (Integral), Eqn. 4.3 (Separable)
Multi-Valued Attributes	4.2.2.1	Eqn. 4.5 (Hungarian Algorithm and Eqn. 4.4)
Composite Attributes	4.2.2.2	Eqn.4.2 (Integral), Eqn. 4.3 (Separable)
Single-Attribute Disjunction	4.2.3	Eqn. 4.6
Multi-Attribute Disjunction	4.2.4	Eqn. 4.7
Negation	4.2.5	Eqn. 4.8
Weight Specification	4.3	Eqn. 4.9

Table 4.2: The different types of constraint connectives, their corresponding chapter sections, and the recommended methods for similarity assessments with each type.

Locally-better matching is useful for applications that demand absolute dominance of some constraints over others. Globally-better matching relies on a compensatory use of the *and* operator. The constraints can still be prioritized using weights, but all of the individual similarity estimates contribute to the final score. The theory for this type of conjunction was based on widely accepted psychological findings about similarity. The Euclidean aggregation function is appropriate for perceptually correlated attributes, whereas a Manhattan metric approximates more closely perceptually distinct properties. An interesting case of conjunction occurs when the aggregated terms refer to values of the same attribute. Such queries are possible in systems that allow storage of multi-valued attributes. A new set of methods was developed to support them. Weighting the constraints to reflect user preferences and goals constitutes an important component of the similarity process but the process may oftentimes be unintuitive. Ranked-weighting methods can address this problem because they rely on minimal user interaction and delegate the main computational details to the system.

## CHAPTER 5

### SEMANTIC SIMILARITY AMONG SPATIAL SCENES

This chapter extends similarity assessments to the scene level. Geographic scene-matching problems present several variations depending on the types of the scene query and the underlying database (Section 5.1), as well as on the different kinds of results that are possible (Section 5.2). An explicit awareness of such parameters exposes the intricacies of the problem within a geographic context, but must also be complemented with a plausible rationale for obtaining similar results to a scene query (Sections 5.3 and 5.4). The three key psychological principles (Section 2.2.2.6) that people: (1) match only sufficiently similar objects in a way that preserves the correspondences among relations (Figure 5.1a), (2) ignore entirely very dissimilar scenes (Figure 5.1b), and (3) choose among different solutions the one requiring the least amount of change (Figure 5.1c), can serve as loose guidelines for such a formalized rationale.

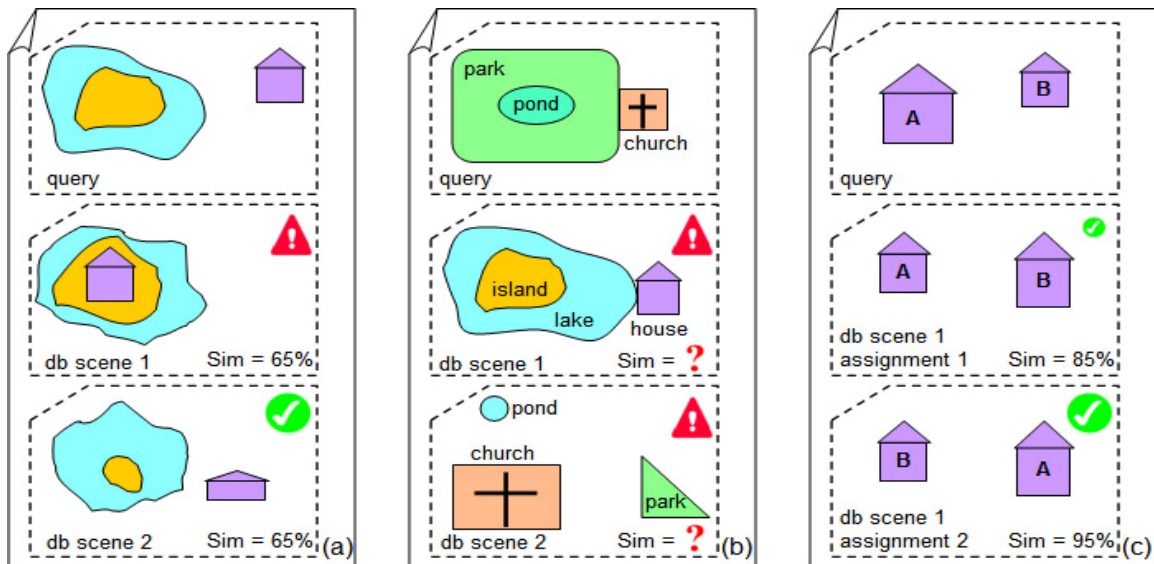


Figure 5.1: Psychological principles for spatial scene similarity assessments (Section 2.2.2.6).

The interpretation of these principles in the context of scene retrieval and the incorporation of knowledge unique to the spatial domain drive the choice for a systematic computational methodology that is able to obtain reliable similar results to a spatial scene query (Section 5.5). The final part of this chapter (Section 5.6) concludes with a detailed example that demonstrates how the presented concepts and methods apply in a practical retrieval scenario.

## 5.1 Spatial Scene Queries

Retrieving similar configurations crosses the boundaries of many disciplines and has stimulated considerable research due to its numerous applications in such fields as computer vision (Ballard and Brown 1982), multimedia databases (Flickner *et al.* 1995), medicine (Petrakis and Faloutsos 1997), and biology (Wang *et al.* 2004). The hard combinatorial nature of the problem often implies that its solution requires not only the adoption of appropriate computational techniques, but also their fusion with domain or application-specific knowledge. The central question that arises in scene comparisons, one that is virtually irrelevant for simpler levels of representation, is how to associate parts of one scene with corresponding parts of another scene. Therefore, it is important to recognize what those parts are under a geographic setting, and what the principles are that should guide the correspondences amongst them. Aspects of geographic domain knowledge are also implicit in the form in which a spatial query is expressed, since different forms of input may suggest alternative distributions of significance to the various components, thereby affecting similarity.

### 5.1.1 Types of Spatial Scene Queries

Spatial scene queries can be roughly divided into two categories based on their form of input: (1) *queries by expression*, which the user constructs using the modalities provided by the system and (b) *queries by selection*, where the query is set equal to a selected database configuration (Figure 5.2). Queries by expression can be syntactic (Chamberlin

*et al.* 1976), sketched (Egenhofer 1996), or a combination of both (Calcinelli and Mainguenaud 1994; Di Loreto *et al.* 1996; Agouris *et al.* 1999). They can be formulated via an appropriate command-oriented language (Egenhofer 1994a), or a graphical user interface that facilitates sketching on the screen (Gross 1996; Haarslev and Wessel 1997b; Blaser and Egenhofer 2000). Queries by selection, on the other hand, require minimal user intervention as users simply select a prototype scene and the system must retrieve other scenes that resemble it. This method of querying is popular in databases that contain collections of individual scenes (*collection databases*), such as image databases (Kelly *et al.* 1995), or databases of protein structures (Artymiuk *et al.* 1994). In the case of large continuous datasets (*continuous databases*), querying by selection could be carried out by selecting part of a map on the screen and requesting similar areas from the database. Current GISs, however, do not yet natively support such functions.

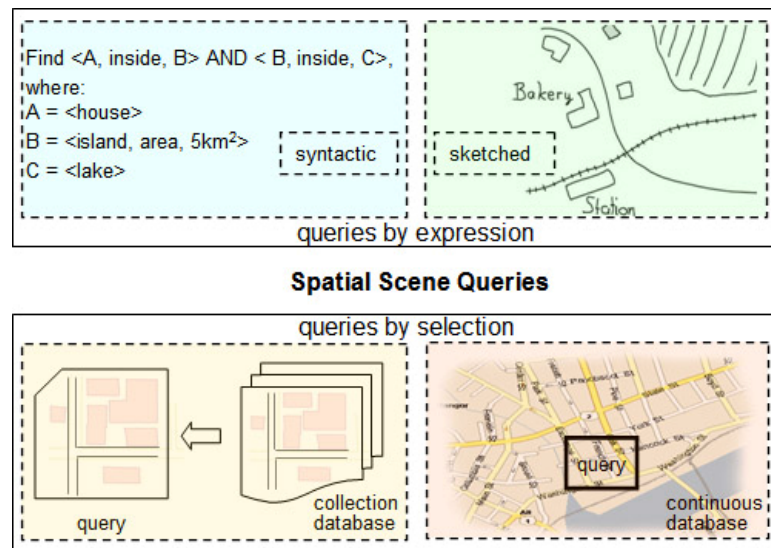


Figure 5.2: Forms of spatial scene queries.

The various forms of spatial scene queries spawn different considerations for their processing. Syntactic queries contain an explicitly stated and precisely specified set of constraints. However, there could be relational constraints that are missing but implied, or the set of existing constraints might be logically inconsistent, thus describing an

impossible configuration. Such anomalies should be detected during a preprocessing stage. Sketch queries are inherently approximate, although sometimes they also contain precise constraints (e.g., through optional hand-written annotations). A feature unique to syntactic and sketched queries is their often-exploratory character, since they need not necessarily correspond to real-world scenes reconstructed from memory.

Whereas queries by expression employ only a subset of the constraints that may be imposed, selection queries contain the exhaustive set of all possible constraints (i.e., each attribute value of objects and relations in the selected database scene becomes a constraint during selection). Considering all constraints might be undesirable, therefore, users should be able to shift the context of similarity to the dimensions of interest, either by appropriate weight allocation or by dismissing those constraints that are irrelevant for the purposes of the comparison. By definition, all selection queries correspond to real-world configurations.

#### 5.1.2 Components of a Spatial Scene Query

A spatial scene query comprises a number of objects, each with its own set of specifications. Furthermore, the objects must adhere to a certain structure, meaning that there may be several spatial (and potentially thematic) relations among them (i.e., Figure 1.1); therefore, such a query has two major *components*: objects and relationships among the objects (Figure 1.2). The characteristics of the objects (e.g., their class or a geometric attribute) form a set of unary constraints, while those of the relations (e.g., the topology or distance) form a set of binary constraints on the pairs of objects. An exact match to such a scene query is then any database scene that simultaneously satisfies both sets of constraints.

Typically, object constraints are specified by assigning an atomic value to an attribute (e.g., *\*class(house)* or *\*area(300m<sup>2</sup>)*). Cases of multivalued attributes are also possible. Spatial relational constraints are more complex, mainly for two reasons: (1) the lack of a

universally accepted representational scheme (Hernández 1994; Cohn and Hazarika 2001) and (2) the availability of representational structures at various levels of detail. For example, a topological constraint between two objects could be defined as coarsely as a simple topological relation classifier (i.e., *overlap* or *contains*), or as comprehensively as the resolution of the representational formalism allows, containing also additional topological invariants (Egenhofer and Franzosa 1995; Clementini and di Felice 1998) and metric refinements (Shariff 1996; Egenhofer and Shariff 1998; Godoy and Rodriguez 2002; Stefanidis *et al.* 2002; Nedas *et al.* in press) (Figure 5.3). Currently, all approaches in the literature that address geographic scene similarity deal exclusively with coarse spatial relations. For clarity of presentation, the examples in this chapter also use coarse relations; however, this issue is revisited in Section 5.4.2.

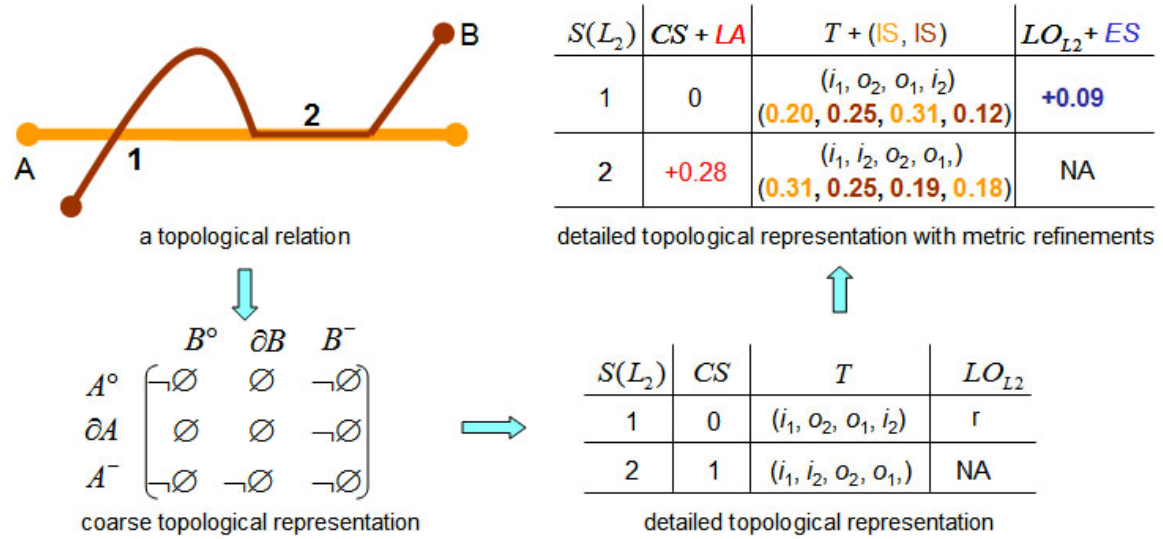


Figure 5.3: Representing a topological relation at progressively finer levels of detail.

### 5.1.3 Formulating Spatial Scene Queries

A spatial query can be formulated as a constraint satisfaction problem (CSP), (Kumar 1992) which consists of:

- A set of  $n$  variables  $V_1, V_2, \dots, V_n$  that correspond to the objects appearing in the query.

- For each variable  $V_i$ , a finite domain of  $N$  possible values  $D_i = \{O_1, O_2, \dots, O_N\}$  that correspond to the objects in the database.
- For each variable a  $k$ -tuple of  $k$  unary constraints  $P_i = (p_{1i}, p_{2i}, \dots, p_{ki})$ , where  $p_{1i}, p_{2i}, \dots, p_{ki}$  are specific instantiations of the attributes  $p_1, p_2, \dots, p_k$  for variable  $V_i$ . If the database contains  $j$  attributes for each object and the domain of each attribute is  $A_1, A_2, \dots, A_j$ , then  $P_i$  is a subset of an element of the Cartesian product  $\prod_{i=0}^j A_i$ , that is,  $P_i \subseteq x, x \in \prod_{i=0}^j A_j$ .
- For each pair of variables an  $m$ -tuple of  $m$  binary constraints  $R_{ij} = (r_{1ij}, r_{2ij}, \dots, r_{mij})$ , where  $r_{1ij}, r_{2ij}, \dots, r_{mij}$  are specific instantiations of the properties  $r_1, r_2, \dots, r_m$  of the relation between the variables  $V_i$  and  $V_j$ . If the database contains  $l$  properties for each relation and the domain of each property is  $B_1, B_2, \dots, B_l$ , then  $R_{ij}$  is a subset of an element of the Cartesian product  $\prod_{i=0}^l B_i$ , that is,  $R_{ij} \subseteq x, x \in \prod_{i=0}^l B_i$ .

A solution to the CSP is an assignment of values to variables (i.e., database objects to query objects) such that no constraint is violated. When no constraints exist on relations, the problem becomes conceptually identical to that of matching multi-valued attributes. In this case the CSP degenerates to the assignment problem, which can be solved with the methodology developed for multi-valued attributes (Section 4.2.2.1). In the general case, however, a different method is required that performs the matching in a manner that satisfies both the unary and the binary constraints.

#### 5.1.4 Representing a Spatial Scene as a Graph

A scene CSP can be abstracted as an attributed pseudograph (Figure 5.4). Objects and binary relations in the scene are abstracted as nodes and edges of the graph, respectively. The edges of the graph are directed if non-symmetric relations, such as containment and direction, are modeled in the scene. Numerical or symbolic attribute values attached to



the loops correspond to properties of the objects, and those attached to the remaining edges correspond to properties of the binary relations. Adhering to the view of a singular relation that encapsulates all relational properties for an object pair, only one edge is drawn between two nodes. If a relation exists between any two objects in the original scene, then the resulting graph is complete and its number of  $m$  edges is equal to  $n \cdot (n-1)/2$  (or  $n \cdot (n-1)$  for directed relations), where  $n$  is the number of nodes.

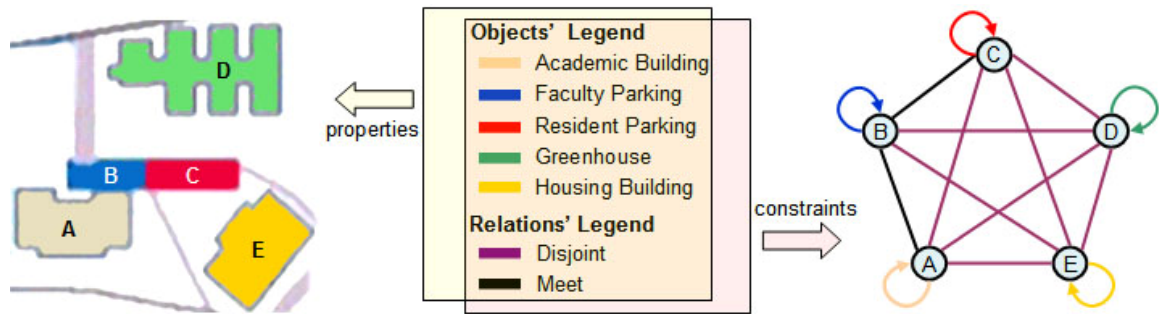


Figure 5.4: Representation of a spatial scene as a complete labeled pseudograph (road networks were omitted to avoid clutter). Properties of objects and relations that become constraints are denoted with a color-coding scheme, which is reused in subsequent chapter figures.

Graphs of spatial scenes can be automatically derived during a preprocessing stage. Although in some approaches this transformation is a prerequisite (Messmer and Bunke 1995), the method that we outline can operate either on the graphs of the scenes, or on the scenes themselves. A graph abstraction is helpful, however, in revealing the nature of the scene-querying problem and the different types of solutions that are possible.

## 5.2 Types of Solutions for a Scene Query

To test for simple isomorphism between the graph representations of a query and of a database scene would be insufficient to determine if the two scenes are equivalent, because the identity of the nodes and edges of their graphs must also be identical. This requirement introduces the need for a constrained isomorphism testing between the query

and the database scene (Section 2.3.2). Even then, establishing a constrained isomorphism between compared graphs is of little value, since the graphs of two spatial scenes will rarely contain the same number of nodes. There might be objects in the query scene missing from the database scene, or the other way around. Several approaches attempt to compensate for this disparity by defining a distance measure between two graphs in terms of node or edge deletions, insertions, or substitutions required to make the graphs isomorphic. Such methods have little practical merit for spatial scenes unless the graph sizes differ only slightly. Hence, their effectiveness on collection databases is questionable and their application to large continuous datasets, where the number of objects in the query is trivial compared to the number of objects in the database (i.e.,  $|n| \ll |N|$ ), is difficult.

In the general case, the graphs will contain a different number of nodes. A solution, therefore, is derived by a *constrained subgraph isomorphism*, not a constrained graph isomorphism. The objective is to match corresponding *substructures* of the two graphs, rather than match the graphs in their entirety. Depending on how such substructures align with one another, three types of solutions can be distinguished with respect to their completeness to a scene query:

- A subgraph  $H'$  of the database graph  $H$  is isomorphic to the query graph  $G$ . In this scenario,  $H'$  constitutes a *complete solution*, because all query objects and relations have a counterpart in the database (Figure 5.5a). It is likely—especially for large continuous databases—to have several complete solutions, meaning that there exist multiple proper subgraphs of  $H$  isomorphic to  $G$ . However, if  $H'$  is not a proper subgraph (i.e., if  $H' = H$ ) then  $H'$  is also the only complete solution. Obviously, having two isomorphic graphs is just a special case of the general problem.
- A subgraph  $H'$  of the data graph  $H$  is isomorphic to a proper subgraph  $G'$  of the query graph  $G$ . In this case,  $H'$  is an *incomplete solution* to  $G$ , because the solution

matches only a subset of the queried configuration (Figure 5.5b). As for the case of Figure 5.5a, there could be more than one incomplete solutions and if  $H' = H$ , then  $H'$  is the only incomplete solution (Figure 5.5c).

- No subgraph  $H'$  of the data graph  $H$  is isomorphic to any subgraph  $G'$  of the query graph  $G$ , therefore, no solution exists (Figure 5.5d). This last scenario is more likely to occur in collection databases.

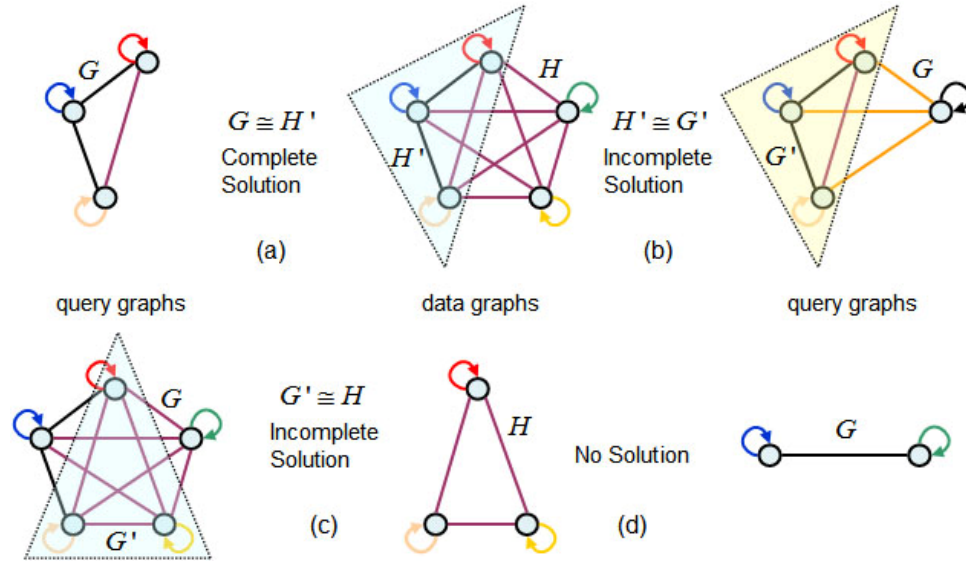


Figure 5.5: Complete and incomplete solutions to spatial scene queries.

Performing a constrained subgraph isomorphism yields solutions—whether complete or incomplete—that require an exact match between the corresponding objects and relations. In realistic scenarios, the approximate nature of spatial queries and the abundance of combined constraints make the existence of exact solutions unlikely. Therefore, it is desirable to relax some of the initial constraints in order to permit additional acceptable value combinations so as to retrieve similar results. The relaxation implies an *error-tolerant subgraph isomorphism* as well as a method for measuring the deviation (i.e., dissimilarity) from the ideal solution (i.e., the original CSP prior to constraint relaxation). Such methods are provided by the algorithms developed in Chapters 3 and 4.

Relaxing the constraints produces a weaker version of the original problem, which is known as *partial constraint satisfaction problem* (PCSP) (Freuder and Wallace 1992). The relaxation means that the original problem  $P$  is modified to a different problem  $P'$  such that the set of solutions to  $P$  is a proper subset of the set of solutions to  $P'$ . The new set will contain some additional approximate (or partial<sup>2</sup>) solutions, which can be ranked according to their similarity from the exact ones. Exact and approximate solutions could both be complete or incomplete (Figure 5.6).

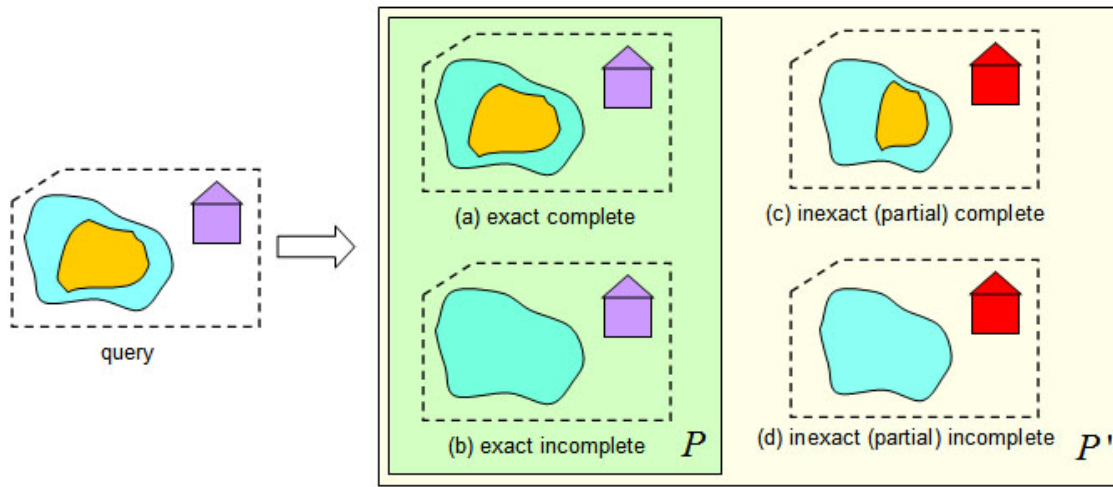


Figure 5.6: Types of solutions to a relaxed CSP: (a) exact and complete, (b) exact and incomplete, (c) partial and complete, and (d) partial and incomplete.

### 5.3 Types of Retrieval

An entirely unconstrained version of the relaxed problem  $P'$  (i.e. setting each constraint to equal its domain) would require generating all  $n$ -permutations for a query and a data scene of  $n$  and  $N$  objects respectively. Such a brute-force approach would need to

---

<sup>2</sup> The term partial, referring to constraints or solutions, should be taken as synonymous to inexact or approximate, not to be confused with the term incomplete, which was reserved to refer to solutions that evaluate only a subset of the query variables

consider  $N!/(N-n)!$  different valuations and assess the similarity of each to the ideal solution. For any practical application, exhaustive searches of this kind fail to complete within a reasonable amount of time. For instance, for a relatively small geographic dataset of 1,000 objects and a moderate size query of 5 objects,  $99 \cdot 10^{13}$  solutions would need to be tested. To ensure retrieval within realistic time bounds, the extent of relaxation must be controlled through thresholds, which define what objects and relations of the database scene can be matched to those of the query scene.

A *global threshold*  $T$  applies to the whole scene and represents the maximum acceptable dissimilarity of a result to the query scene. A database scene is considered a solution if its global dissimilarity  $D$  to the query scene is less than or equal to  $T$  and rejected otherwise. *Local thresholds*, on the other hand, can be imposed either at the component level (i.e., individual objects or relations) or at the attribute level (i.e., individual constraints on each object or each relation). They are defined as *component-local* and *constraint-local* thresholds and denoted with  $t$  and  $\tau$ , respectively. For a component-local threshold, an association between a pair  $i$  of objects or relations is valid if the dissimilarity  $d_i$  (Equations 4.2 and 4.3) of that pair is less than or equal to  $t_i$ . For a constraint-local threshold, an association between a pair  $i$  of objects or relations is valid if the dissimilarity  $\delta_j$  (Equations 3.2-3.5 and 3.8-3.10) of each individual constraint  $j$  on this pair is less than or equal to  $\tau_j$ . Alternatively, we say that the association partially violates (or satisfies) the constraints. The dissimilarity value  $\delta_j$  determines the degree of satisfaction for each constraint  $j$ . If  $\delta_j > \tau_j$ , a constraint is totally violated. A solution is acceptable if none of the individual constraints is totally violated and rejected otherwise. Obviously, if there is only one constraint on an object or a relation then  $d = \delta$  and  $t = \tau$ , and if the scene consists of a single object then  $D = d$  and  $T = t$ .

Deciding on the usage of a particular type of threshold has different repercussions on the efficiency and the semantics of the retrieval. A global threshold corresponds to a *soft retrieval strategy*, which finds solutions that are on average good. However, it is prone to

creating a few locally weak matches that make little sense, but whose effect on the global similarity score is insufficient to prevent such discrepancies from occurring. In this sense, it defies the first three psychological findings about how people perform scene similarity assessments (Section 2.2.2.6), which imply that: (1) the quality of local matches between individual pairs of objects and relations is more important than that of a global scene match that is highly similar on the average, but contains a few weak object associations or relationship correspondences (Figure 5.1a), and (2) unlikely solutions that create absurd object associations and relationship correspondences need not be considered and should be excluded from assessment early (Figure 5.1b). In terms of performance, a soft retrieval type also suffers, because all different solutions need to become partially instantiated in order to decide whether they should be rejected or not. In general, processing time increases with higher values for  $T$  and as the number of constraints in the query increases.

Using only local thresholds, on the other hand, corresponds to a *hard retrieval strategy*. A solution must then satisfy, either partially or totally, every individual constraint. If a single constraint is totally violated the solution is rejected. Hard retrieval on a relaxed CSP possesses several desirable properties: (1) it approximates human perception of similar scenes, because it maintains high quality local-matches consistently throughout the entire configuration (Figure 5.1a and b); (2) all local constraints could be automatically relaxed by the system and translated in the form of range queries, on the precondition that users simply input the number of desired candidate matches for each object or relation (Schumacher and Bergmann 2000), a property, which relieves users from the burden of manually specifying multiple thresholds; and (3) it guarantees a considerably more efficient query processing, because the DBMS can exploit database indexes to execute the range queries faster and the search space for each variable is pruned significantly.

## 5.4 Relaxation

The semantic and computational benefits of local thresholds come at the expense of a very strict retrieval policy that will reject a solution, if a single constraint threshold is slightly exceeded. Therefore, a hard retrieval strategy must also be complemented with a rationale that prevents arbitrary choices during the relaxation of the initial constraints. Since some alternatives may compromise the quality of the retrieval by producing results that deviate from the intentions and meaning of the original query and others might practically trivialize the problem by allowing thousands of new solutions, such a rationale must incorporate aspects of spatial domain knowledge, as they relate to spatial objects and relations.

### 5.4.1 Relaxation for Spatial Objects

Constraints that are more significant should be relaxed less than others in order to preserve the quality of the results. Among object-specific constraints, the class is the central element of an object's identity and a primary characteristic, since it conveys information about the possible attributes, parts, and functions of the object. The diagnostic effect of classes in object categorization is the highest among all object features (Tversky 1977). Therefore, when class constraints are present they should be relaxed conservatively compared to constraints on the remaining properties of objects. Failure to do so may produce incongruous matches (Figure 5.7).

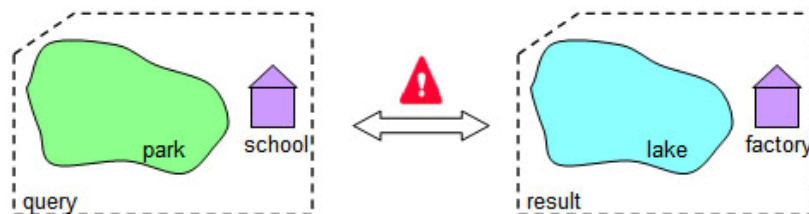


Figure 5.7: A low quality result produced by assigning the same significance to the class and the geometric attributes of the objects.

Establishing the relative significance among other attribute types of an object or properties of a relation, depends largely on the form of a spatial query. Sketched objects are typically crude approximations of their real-world counterpart, consisting mainly of simple boxes and lines. Furthermore, when sketching real scenes, people fail at capturing the metric relations between objects accurately, but are better at preserving the topological and directional structure (although this can be partially attributed to the dominance of the *disjoint* relation) (Blaser 2000). An automated transformation of a freehand sketch to a partial CSP should consider such evidence and take corrective action by relaxing object geometries and metric relations more than other constraints (Figure 5.7).

Another distinction with possible ties to significance is between explicit and implicit constraints. For instance, the syntactic statements in the query of Figure 5.2a introduce explicit constraints, but the missing relation  $\langle A, \textit{inside}, C \rangle$  is implicit. Similarly, in the sketch query of Figure 5.2b, the metric relation between the *bakery* and the *station* is explicit, in contrast to the other metric relations whose quantitative properties have to be extracted by the system. Choosing to assign more detail to a subset of the objects or relations in a query implies a pronounced interest in those components.

#### 5.4.2 Relaxation for Spatial Relations

Spatial relations distinguish the relative placement of objects in the embedding space. Excluding topological relations, which are inherently qualitative, directional and metric relations can be expressed either in the quantitative or in the qualitative realm. The relaxation of quantitative distance and angle relations can be treated in the same way as ratio and cyclic attributes, respectively. The deviation from the initial values can be delimited by specifying an amount of change that the relaxed values should not exceed (i.e., a maximum allowed percentage of fluctuation), or by entering a desired number of  $k$  to-be-retrieved matches and let the system infer the extent of relaxation (Schumacher and



Bergmann 2000). The only difference is that, in order to define the value domain of such relations, an exhaustive database search must first be performed to determine all  $N \cdot (N - 1)$  relations between the distinct object pairs.

Oftentimes, however, qualitative relations are required in a scene similarity assessment (e.g., a map or image without scale and orientation, or a sketch query). Such relations are graphically organized in terms of their conceptual neighborhoods based on the gradual changes required to derive one set of relations from another (Freksa 1991; Egenhofer and Mark 1995a). The concept of gradual change originates from the gradual deformation of objects until the spatial relation between them is changed. Conceptual neighborhoods allow measuring similarity between two spatial relations as a function of the length of the shortest path that joins them along the graph. If the shortest path has one edge then the relations are 1<sup>st</sup> degree neighbors, if it has two edges then they are 2<sup>nd</sup> degree neighbors, and so forth. Hence, the process of relaxing a qualitative binary constraint (relation) consists simply of gradually expanding its domain with its  $n$ -degree neighbors, where  $n$  is determined by the desired amount of relaxation.

Despite their simplicity and intuitive appeal, coarse topological and directional qualitative models have characteristics that render them unsuitable for a coherent relaxation of relational constraints. Choosing to represent relations in a continuous space with a number of discrete equivalence classes introduces two fundamental problems for scene similarity assessments. The first problem arises out of the implicit assumption of an equi-distance step between adjacent classes in the conceptual neighborhood graph. Because of this assumption, the relaxed version of an original constraint may dismiss potentially good matches, while introducing weak ones (Figure 5.8a). The second problem is inability to distinguish among members of the same class. As a result, all relations within the same category are considered equally similar (Figure 5.8b).

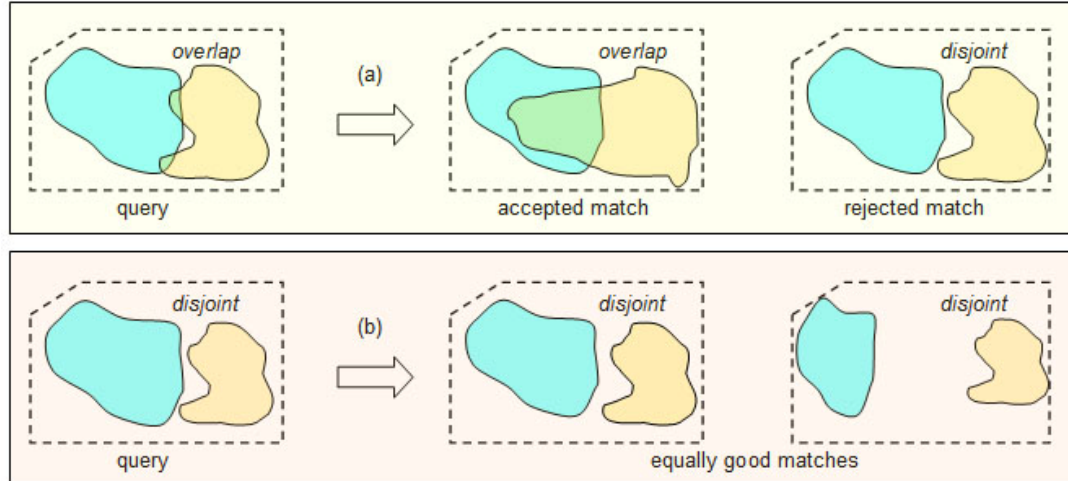


Figure 5.8: Problems of coarse topological relations for scene similarity assessments:  
 (a) reasoning for similarity based on distances in a conceptual graph may exclude highly similar matches in favor of others that are less similar and  
 (b) the inability to discriminate among members of the same class treats relations, for which people may have distinct mental images, as equally similar.

These problems persist even if one substitutes coarse relations with their detailed counterparts in the relaxation process. Detailed representations of topological relations rely on a number of invariants in order to establish topological equivalence, and assign a number of properties to each intersection component between interiors and boundaries of the objects (Egenhofer and Franzosa 1995; Clementini and di Felice 1998); therefore, the equi-distance assumption is only transposed at a finer level of detail and the lack of discrimination between topologically-equivalent relations persists. In fact, one may argue that employing detailed relations makes matters worse. Establishing a reliable relaxation process for them is a largely unintuitive and complex task, and it is questionable if there is value in relaxing constraints at such a fine level of granularity (i.e., if any additional matches will result out of the relaxation). Furthermore, detailed relations have performance ramifications because they largely increase the computational cost for similarity assessments between relations.

The difficulties associated with the coarse and detailed formalisms for spatial relations can be overcome by employing a representational scheme that comprises a number of semi-qualitative metrics. These measures are more appropriate for the relaxation of qualitative relational constraints because they are continuous. The core notion behind them is that of a normalized distance or angle. The normalization of these quantities can be achieved in numerous ways. For example, the distance between the centroids of the objects' MBRs could be divided by the total area of the MBRs; or the distance between the boundaries of the objects' could be divided with their perimeter. Therefore, one talks about a family of metrics because multiple measures are possible. Choosing the quantities to be normalized and those that they should be normalized by, are important choices that instill different qualities and weaknesses in the produced measures. Several research efforts have recently tried to introduce such measures, albeit with mixed success (Egenhofer and Shariff 1998; Goyal and Egenhofer 2001; Godoy and Rodriguez 2002; Stefanidis *et al.* 2002; Nedas *et al.* in press). For instance, some studies assume a-priori knowledge of the objects' identities, while others require that the two scenes contain the same number of objects or that the objects are of the same type (i.e., lines or regions).

The different assumptions underlying these approaches inflict different restrictions on the applicability of the qualitative metrics that they advocate. Deriving such metrics is beyond the scope of this thesis. However, it is relevant to provide a list of requirements that a family of such measures must comply with, in order to be useful for the purposes of constraint relaxation of spatial relations for scene similarity assessments. This list comprises five orthogonal preconditions: *continuity*, *scale-invariance*, *object identity-invariance*, *universality*, and *minimality*. The first two requirements are obvious, since continuity is the reason for introducing these measures in the first place, and scale-invariance is an indispensable characteristic for qualitative representations (Lindeberg 1993). Object identity-invariance implies that the choice of a reference object should be

immaterial for any relation between two objects. In other words, a different labeling of the objects should not change the measure that describes their relation. This precondition is necessary because in the general case of scene similarity queries there is no a-priori knowledge about the correspondences between objects of different scenes. Universality means that a measure must not be tied to any specific coarse topological or directional relation but apply to the full spectrum of relations to which it serves as a surrogate. The last requirement of *minimality* pertains to efficiency. A family of qualitative metrics should achieve the maximum descriptive ability with the fewest possible measures.

The importance of qualitative metrics for similarity purposes can be conceived by considering that, on average, over 95% of the topological relations in spatial datasets of normal density are disjoint. For such relations, these metrics are the only viable alternative for making similarity judgments. Qualitative metrics, however, should complement rather than replace models based on conceptual neighborhoods because in some cases the employment of the former might be impossible (e.g., Figure 5.2a). Furthermore, since similarity is goal-dependent, the user might insist that the similarity of relations is determined strictly with respect to topology.

## **5.5 Query Execution**

The relaxation process creates a weaker version of the original CSP (i.e., a PCSP). A methodology that identifies and extracts subgraphs of the database scene, which are constrained isomorphic to the graph representation of the PCSP, yields a set of similar solutions to the original CSP. The most elegant approach to solving the common subgraph problem is by extracting the maximal cliques of an association graph (Bomze *et al.* 1999).

### 5.5.1 Query Preprocessing

Before executing a spatial scene query, all relational constraints that are missing but implied must become explicit (e.g., Figure 5.2a). This process achieves better efficiency because it prunes the search space for the implicit relations and does not have to consider their entire domain. It also helps with the early detection of logically inconsistent queries that correspond to impossible configurations. The explication of implicit relations can be automated with *composition tables*, which encode the possible spatial relations between two variables  $V_i, V_j$  given the relations between variables  $V_i, V_k$  and  $V_k, V_j$ . Composition tables exist for topological relations (Egenhofer and Sharma 1993; Egenhofer 1994b), directional relations (Papadias and Egenhofer 1996), and combinations of directional and distance relations (Papadias *et al.* 1999b).

### 5.5.2 Creating the Association Graph and Extracting the Maximal Cliques

The solutions to a scene query can be given by extracting the maximal cliques of an association graph. An association graph (Ambler *et al.* 1973) captures the mutual dependencies between two relational structures. For a query graph  $G$  with node set  $(v_1, \dots, v_n)$  and a database graph  $H$  with node set  $(u_1, \dots, u_N)$  the nodes and edges of their association graph are created in two distinct steps, as follows: during the first step, a node of the association graph is created for each compatible pair of nodes between  $G$  and  $H$ . Specifically, if a node  $u_j$  of the database graph satisfies the relaxed unary constraints of a node  $v_i$  in the query graph, then an association graph node  $a_{ij} = (v_i, u_j)$  is created to register this possible correspondence. During the second step, the edges of the association graph are generated by joining nodes that have compatible relations; that is, an edge is inserted between nodes  $a_{ij}$  and  $a_{kl}$  of the association graph if the relationship between nodes  $u_j$  and  $u_l$  of the database graph satisfies the relaxed binary constraints explicated by the relationship between nodes  $v_i$  and  $v_k$  of the query graph.

Given the way that the association graph was constructed, the notions of *complete solution* and *incomplete solution* coincide with those of *maximum clique* and *maximal clique* (Section 2.3.2), respectively. Simple cliques amount to *redundant solutions*, that is, they are incomplete solutions already encapsulated within a larger complete or incomplete solution. The traversal of the association graph and the extraction of the maximal cliques can be done by clique-enumerating algorithms (Bron and Kerbosch 1973; Loukakis and Tsouros 1981; Tomita *et al.* 1988).

### 5.5.3 Post-Processing of Results

The stage of post-processing consists of evaluating the similarity of each retrieved scene to the query, filtering the results, and presenting the final set of solutions to the user.

#### 5.5.3.1 Component Similarity

The association graph, which was obtained for the graphs of the relaxed scene query and the database scene, is transformed into a weighted association graph by attaching a dissimilarity score to each node and each edge. The value at each node  $(V_i, O_j)$  represents the dissimilarity of object  $O_j$  of the database scene, with respect to object (variable)  $V_i$  of the query scene, whereas the value at each edge  $((V_i, O_k), (V_j, O_l))$  represents the dissimilarity of the relation  $(O_k, O_l)$  in the database scene, with respect to the relation  $(V_i, V_j)$  in the query scene. Since both relations and objects are modeled as tuples that contain several attribute values (i.e., their constraints), the similarity scores at each node and edge are calculated by performing the following steps: (1) For each attribute-level constraint a dissimilarity measure is calculated by the algorithms that were developed in Chapter 3. (2) The aggregation of the attribute-level dissimilarities yields the overall dissimilarity for a pair of objects or relations (Equation 4.3). This step takes into consideration the weights specified on constraints at the attribute level. Groups of integral attributes are also combined to form separable attributes before being aggregated

(Equation 4.2). (3) The dissimilarities for each pair of matched elements are converted to perceived similarities using either of Equations 2.1b-c.

The *object similarity component*  $S_{Obj}$  between the matched substructures of two spatial scenes is calculated based on the similarities of all their associated object pairs as described by the labeled nodes of the maximal clique of the scenes in the weighted association graph (Equation 5.1).

$$S_{Obj} = \frac{\sum_{i=1}^M w_{O_i} \cdot s_{O_i}}{\sum_{i=1}^M w_{O_i}} \quad (5.1)$$

where:  $s_{O_i}$  : Object similarity of an associated object pair  $i$   
 $w_{O_i}$  : Weight of the query object in the  $i$  th associated object pair  
 $M$  : Number of associated object pairs (matched objects)

The relational similarity component  $S_{Rel}$  between the matched substructures of two spatial scenes is computed based on the similarities of their corresponding binary relations, as described by the labeled edges of the maximal clique of the scenes in the weighted association graph (Equation 5.2).

$$S_{Rel} = \frac{\sum_{i=1}^{M \cdot (M-1)/2} w_{R_i} \cdot s_{R_i}}{\sum_{i=1}^{M \cdot (M-1)/2} w_{R_i}} \quad (5.2)$$

where:  $s_{R_i}$  : Relational similarity of an associated pair  $i$  of binary relations  
 $w_{R_i}$  : Weight of the query relation in the  $i$  th associated relation pair  
 $M$  : Number of associated object pairs (matched objects)

The weights  $w_{O_i}$  and  $w_{R_i}$  are global weights on each object and relation of the query, respectively, which should not be confused with attribute-level weights that apply to a

particular property of an object or a relation. Due to their dependency on algorithms operating at the attribute level, the measures  $S_{Obj}$  and  $S_{Rel}$  are not always symmetric, but depend on the order of the scenes in the assessment.

### 5.5.3.2 Scene Completeness

Each non-maximum maximal clique of the association graph corresponds to an incomplete solution that matches only a subset of the query objects. Under typical retrieval circumstances, the objects that remain unmatched should inflict a penalty to the incomplete scene's similarity score, which implies a reduced similarity value for that scene. The specification of this penalty is the purpose of the *scene completeness* parameter, which is analogous to the value completeness measure for multivalued attributes (Section 4.2.2.1) based on the ratio contrast model (Tversky 1977). The scene completeness  $S_{Comp(db,q)}$ , a directed measure that operates at the scene level, is a function of the matched (i.e., common) and unmatched (i.e., different) objects for two scenes, taking values between 0 and 1. Its value should be interpreted as *the similarity of the database scene to the query scene with respect to completeness*. The assessment of this type of similarity depends only on the existence or absence of corresponding object pairs and is invariant under all other parameters. The simpler approach considers each object in the query scene of equal importance (Equation 5.3a), whereas a more elaborate version considers the weight assigned to each object (Equation 5.3b).

$$S_{Comp(db,q)} = \frac{M}{M + \alpha \cdot (n - M) + \beta \cdot (N - M)} \quad (5.3a)$$

$$S_{Comp(db,q)} = \frac{\sum_{i=1}^M w_{O_i}}{\sum_{i=1}^M w_{O_i} + \alpha \cdot \sum_{j=1}^{n-M} w_{O_j} + \beta \cdot (N - M)} \quad (5.3b)$$



where:	$M$	: Number of matched objects
	$n$	: Number of objects in the query scene
	$N$	: Number of objects in the database scene
	$\alpha$	: The weight of the set of unmatched query objects
	$\beta$	: The weight of the set of unmatched database objects
	$w_{O_i}$	: The weight of the $i$ -th matched query object
	$w_{O_j}$	: The weight of the $j$ -th unmatched query object

This scene completeness measure is an extension of Blaser's (2000) measure. By explicitly accounting for the effect of unmatched objects in both scenes, Equations 5.3a and 5.3b embed more flexibility and expressive power to the scene completeness measure, allowing it to capture different retrieval objectives through the adjustment of weights  $\alpha$  and  $\beta$ . Three cases are of special interest:

- $\alpha = \beta = 1$ : setting both weights to 1, results in a strict penalty for scene similarity with respect to completeness. The completeness of one scene to another relies not only on the matched objects, but also on the symmetric difference of the sets of unmatched objects. In this case, scene completeness behaves symmetrically. Such an assignment is useful when comparing scenes of approximately equal cardinality and the interest is distributed evenly on elements that match, as well as those that are different in both scenes (e.g., two aerial photographs of the same area, taken at different dates).
- $\alpha = \beta = 0$ : setting both of these weights to 0 results in no penalty for completeness. This weight specification makes scene completeness symmetric, yielding 1 if pairs of matched objects exist and 0 otherwise. The similarity of the scenes depends only on the similarity of the corresponding elements in the matched substructures.
- $\alpha = 1, \beta = 0$ : the penalty for completeness depends only on the unmatched query objects. This weight assignment reflects the purpose of the most typical retrieval

scenario, which occurs when trying to locate a sub-scene in the database that matches best the query (e.g., a sketched query against a large continuous database). The interest is shifted to matched and unmatched query objects, but the unmatched objects in the database scene are ignored. In this case, the measure produces asymmetric values, depending on what scene becomes the query and what scene is the target.

### 5.5.3.3 Scene Similarity

The similarity between two scenes is called *scene similarity*. For a query and a database scene, the similarity of their matched substructures  $S'_{Scene}$  is computed as the weighted and averaged sum of the relational and the object components (Equation 5.4). The final scene similarity  $S_{Scene}$  between the query and the database scenes incorporates the completeness correction (Equation 5.5).

$$S'_{Scene(db, qry)} = \frac{(w_{Obj} \cdot S_{Obj}) + (w_{Rel} \cdot S_{Rel})}{w_{Obj} + w_{Rel}} \quad (5.4)$$

$$S_{Scene(db, qry)} = S'_{Scene(db, qry)} \cdot (w_{Comp} \cdot (S_{Comp} - 1) + 1) \quad (5.5)$$

where:

- $w_{Obj}$  : Weight of the object similarity component
- $w_{Rel}$  : Weight of the relational similarity component
- $w_{Comp}$  : Weight of the scene completeness parameter

The scene completeness has a limiting effect on the scene similarity: if the weight of the scene completeness is 1, then the scene similarity cannot exceed the value of the scene completeness. The weight for the completeness should not be confused with weights  $\alpha$  and  $\beta$  of Equation 5.3. The latter determine the type of completeness (i.e., what is meant by completeness) whereas the weight  $w_{Comp}$  in Equation 5.5 specifies the effect of the chosen completeness type on the scene similarity score.

The weights of the object and relation similarity components allow an easy adjustment of the contribution of each parameter to the scene similarity. A simpler way to calculate the similarity of the solutions and to rank them might be to add up the similarity scores for each maximal clique in the weighted association graph. The maximum-weight clique would then represent the best match. The similarity of the relations, however, would then dominate the scene similarity score for larger cliques, because for  $n$  objects there are  $n \cdot (n-1)/2$  undirected or  $n \cdot (n-1)$  directed relations. In fact, this method of deriving the scores is just a special case of Equation 5.4 and its equivalent normalized case can be reproduced by specific values for the weights  $w_{Obj}$  and  $w_{Rel}$  (i.e., for  $w_{Obj} = n$  and  $w_{Rel} = m$ , with  $n$  and  $m$  being the number of objects and relations, respectively, in the query scene). Although it has been unequivocally established that both object and relational similarity contribute to the scene similarity score (Dubitzky *et al.* 1993; Goldstone 1994a), further research is required to determine the appropriate weight distribution for these two components.

Applying all calculations involved in Equations 5.1 to 5.5 for each maximal clique of the association graph produces a set of results that are ranked according to their similarity to the original scene query.

#### 5.5.3.4 Filtering and Presentation

Each of the ranked results represents a spatial scene. In a GIS environment, these scenes should typically be retrieved in visual form (e.g., by zooming in the part of the map that contains the match or returning the matched sub-scene in a new window). An artifact of the algorithmic approach to the scene retrieval problem is that, occasionally, what seems for the user to be the same scene is retrieved as two or more different solutions. This peculiarity occurs when the same subset of database scene objects are assigned differently to the query objects (Figure 5.9). Although mathematically justified, the multiple retrieval of the same scene would be redundant for the purposes of visual

inspection and analysis of the results, because users care for the combinations rather than the permutations of the matched objects. In such cases, only the solution that yields the maximum similarity score should be retained, as the criterion of minimal change purports (Figure 5.1c).

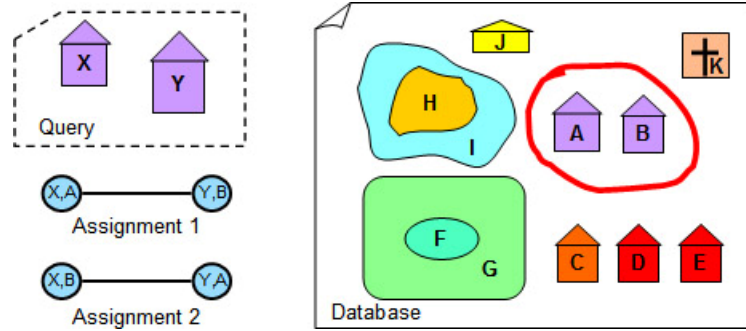


Figure 5.9: Two algorithmically different solutions (i.e., different assignments of database objects to query objects) to a scene query may be perceived from the users as a double retrieval of the same scene.

Another feature unique to the clique approach is that for a query of  $n$  objects, all incomplete solutions with  $n-1$  to  $1$  objects will be retrieved. The number of incomplete solutions is likely to increase as the number of matched objects decreases. To avoid presenting an overwhelming amount of results, or results of little value such as single object matches, the set of solutions may be filtered to include only maximal cliques (i.e., solutions) that exceed a certain size. Such a threshold may be specified as a percentage of the size of the maximum clique  $\omega(A)$  of the association graph  $A$ . An additional filtering option consists of returning only solutions whose scene similarity exceeds some similarity value  $S$ . Such a threshold, however, is not related to the process of constraint relaxation; it is simply cosmetic and serves presentation purposes.

## 5.6 An Example of Processing a Spatial Scene Query

To demonstrate how the concepts and methods of this chapter apply to a practical scene retrieval scenario, consider the example of the spatial scene query of Figure 5.10. The

database scene represents part of the campus of the University of Maine. The solutions to the CSP that corresponds to the user's query can be extracted as the maximal cliques of an association graph, which is formed by comparing the respective constraints between the query and the database scene. These solutions, which can be complete or incomplete, are the subgraph isomorphisms between the query graph and the database graph.

The construction of the association graph starts by selecting an arbitrary object in the query scene, for instance,  $X$ , and finding objects in the data scene that are compatible. Object  $X$  is an *academic building*; therefore, it can be matched with objects  $A$ ,  $I$ ,  $E$ , and  $N$  of the data scene that are also academic buildings. Thus, nodes  $(X, A)$ ,  $(X, I)$ ,  $(X, E)$  and  $(X, N)$  of the association graph are generated. The rest of the nodes are created accordingly, by matching variables  $Y$  and  $Z$  of the query scene with all objects of the data scene that are *faculty parking lots* and *resident parking lots*, respectively. To insert the edges of the association graph, all node pairs are examined sequentially. Nodes  $(Y, G)$  and  $(X, A)$  should not become adjacent, because  $Y$  *meets*  $X$ , whereas  $G$  is *disjoint from*  $A$ . However, nodes  $(Y, G)$  and  $(Z, H)$  should be joined by an edge, because the relation between  $Y$  and  $Z$  is the same as the relation between  $G$  and  $H$  (i.e., *meets*). The only pairs of nodes that are a priori excluded from this process are those that include the same variable in both nodes of the pair. For instance, the pair  $((X, E), (X, A))$  need not be examined at all, because variable  $X$  cannot correspond to objects  $E$  and  $A$  simultaneously. Differently expressed, the *uniqueness* requirement prevents solutions that assign multiple objects to one variable. Continuing this process for all nodes completes the creation of the association graph.

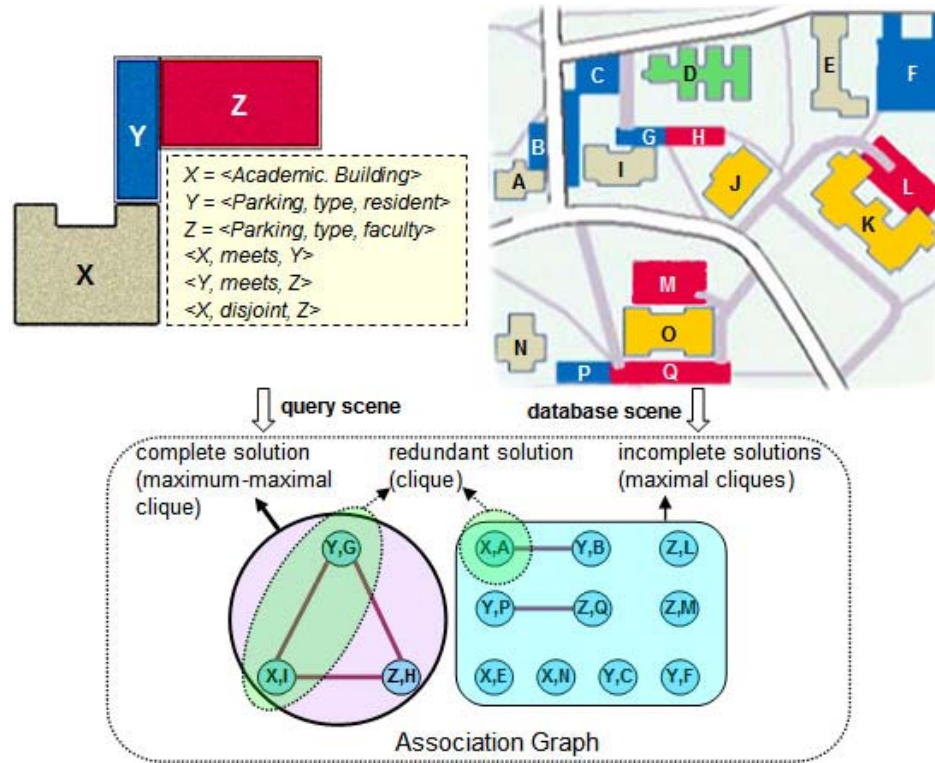


Figure 5.10: Solving a CSP by creating the association graph for a query and a database scene and extracting the solutions.

Maximum-maximal cliques in this graph correspond to complete solutions, maximal but not maximum cliques to incomplete solutions, while simple cliques correspond to redundant incomplete solutions already embedded into a larger solution. For example, the clique  $\{(X, I), (Y, G)\}$  is a redundant solution, because it is already contained within the maximum clique  $\{(X, I), (Y, G), (Z, H)\}$ . The latter is the only complete solution, yielding the object assignment  $(I, G, H)$  to variables  $X, Y$  and  $Z$ , respectively. The graph also contains two maximal cliques of size 2 and six maximal cliques of size 1, all of which constitute assignments that yield incomplete solutions.

The solutions to the original query, whether complete or incomplete, are all exact. To retrieve similar results, a relaxed version of the original CSP (Figure 5.6) must be generated by weakening the original constraints. The solutions to the relaxed version of the CSP (i.e., the PCSP) are obtained in exactly the same manner as those for the original

query, that is, by extracting the maximal cliques of an association graph. The nodes and edges of the new association graph, however, are formed this time with respect to the relaxed constraints. The extraction of the maximal cliques for the PCSP is equivalent to obtaining a set of approximate solutions to the original CSP, which also includes incidental exact matches. An arbitrary relaxation policy that relies on 1<sup>st</sup> neighbors of the coarse topological relations and enlarges the domain of the objects' class constraints degrades the speed of the retrieval by creating a complex association graph, as well as the quality of the results by retrieving many irrelevant solutions (Figure 5.11).

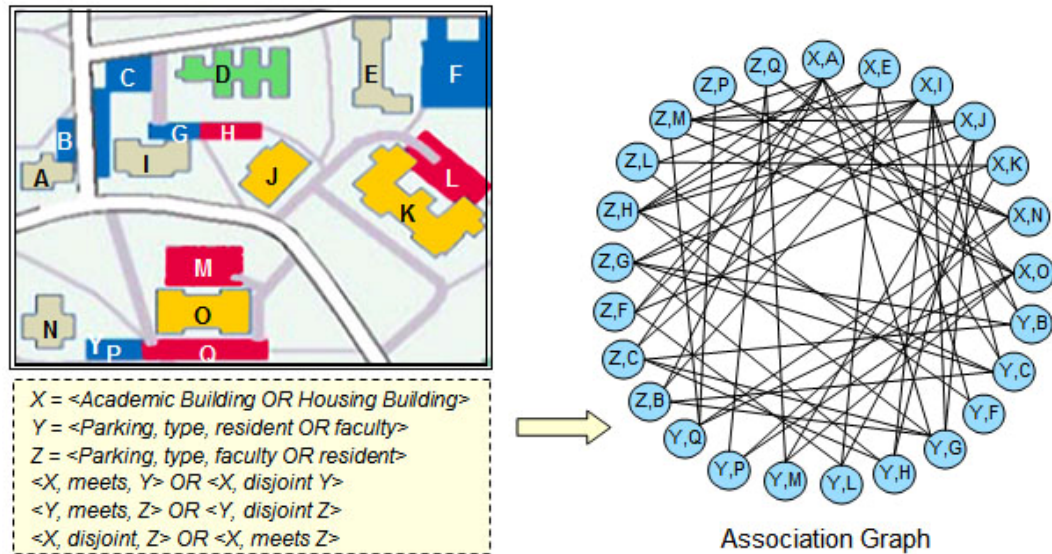


Figure 5.11: Costs on efficiency and quality introduced by a careless relaxation.

These problems are alleviated if the original topological constraints are substituted with semi-qualitative metrics and the class constraints are not relaxed (Figure 5.12). For example, the original *meet* constraint on the relation of object  $X$  to object  $Y$  is substituted with a normalized distance of  $\theta$ , which is then relaxed to allow matches with database distance relations in the range  $(0, 0.15)$ . In addition to the exact complete solution  $\{(X, I), (Y, G), (Z, H)\}$  in Figure 5.10, the solutions to this PCSP include two more complete, but approximate solutions, which are  $\{(X, E), (Y, F), (Z, L)\}$  and

$\{(X,N),(Y,P),(Z,Q)\}$ . There are four incomplete solutions, two of them being exact and two being approximate.

As the example demonstrates, the combination of quantitative or qualitative distance and object constraints is likely to return solutions that form local structures in the database scene. Such local structures correspond to disjoint components in the association graph (Figure 5.12). Hence, further efficiency can be achieved by operating an enumerating clique algorithm independently on each of these components, rather than on a larger graph consisting of a single connected component (Figure 5.11).

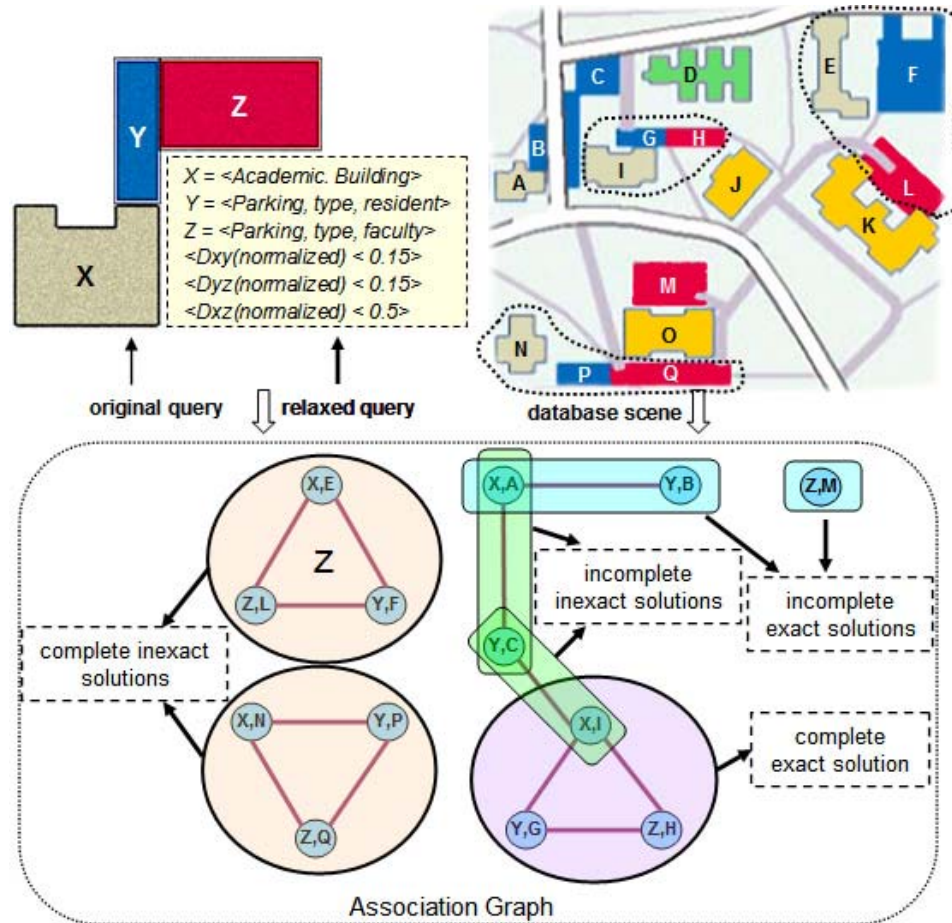


Figure 5.12: Creating the association graph for a relaxed query and a database scene and extracting the solutions.



Solutions of little value, such as single-object matches (i.e., isolated nodes in the association graph) may now be discarded by applying a filter as a percentage of the nodes of the maximum clique. For example, a filter of  $> 0.6 \cdot \omega(A)$  for Figure 5.12 omits from the results all maximal cliques consisting of a single node, thus eliminating node  $\{(Z, M)\}$ . The dissimilarities to the ideal values at each node and edge are converted to similarities by a non-linear monotonically-decreasing function (Equations 2.1b-c), thus transforming the association graph to a weighted association graph (Figure 5.13). By applying the sequence of Equations 5.1 to 5.5 on each maximal clique of the resulting graph, a scene similarity score is assigned to each maximal clique, and the results can be ranked and presented to the user.

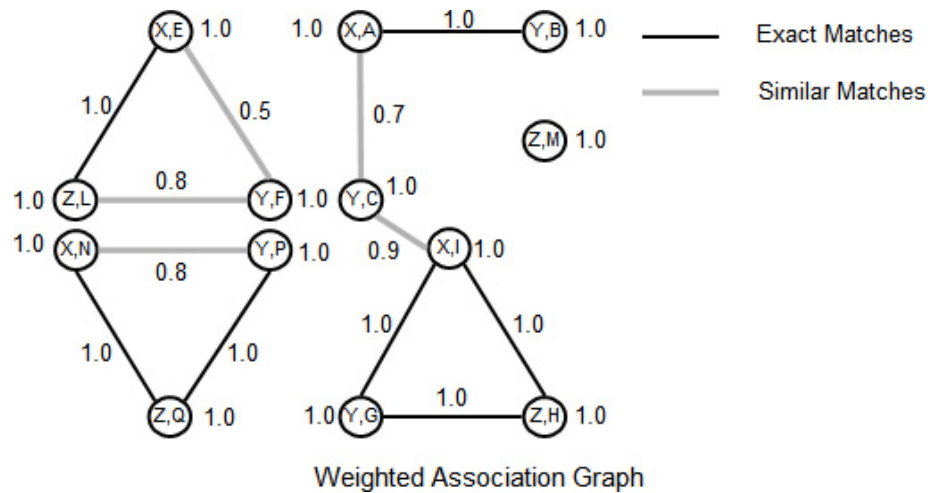


Figure 5.13: The calculation of the similarities between objects and relations transforms the association graph into a weighted association graph.

## 5.7 Summary

The often-exploratory character of a spatial query and the approximate expressions that it may take combined with the relatively large number of constraints that exist in it diminish the possibilities of retrieving exact matches. Whereas similarity retrieval at the level of attribute values and objects may be considered a welcome enhancement to current spatial information systems, similarity at the level of a scene becomes imperative.

Thus, the initial constraints should be seen as ideal starting points that should be approximated by some measure. Approximating is tantamount to relaxing the original constraints, thus substituting the original problem with a weaker version of it. Relaxation is a critical part of the solving process because it affects the efficiency of the retrieval and the quality of the solutions. Performing the relaxation is not simply a matter of expanding the set of acceptable attribute values to a constraint but requires the aggregation of a variety of knowledge specific to the spatial domain. Such knowledge can be captured by deciding on the relative importance of constraints based on the form of the query and by considering what spatial relations to employ in order to create a weaker version of the original problem. Solving the weaker problem yields a number of complete and incomplete solutions that may be exact or similar matches to the initial query. The solution process consists of extracting the maximal cliques of an association graph. The latter is constructed by matching objects and relations of the database scene, whose properties satisfy the relaxed constraints of the query scene. Each solution is assigned a similarity score based on the similarity of the matched relations and objects. Incomplete solutions are optionally penalized for their lack of completeness. Further filtering of the results is also possible based on several criteria. The maximal clique approach establishes the best possible correspondence between the inherent conceptual nature of the problem and its practical implementation and does not rely on simplifying assumptions that may restrict its applicability.

## CHAPTER 6

### MODEL EVALUATION

The hypothesis of this thesis stated that a psychologically compliant approach to similarity yields a set of results, in the relevant portion of the ranking list, dissimilar to that obtained by other commonly used methods. In this context, a *psychologically compliant* method produces the set of results that is consistent with people's judgments of similarity and, therefore, desirable. Any deviation from such an approach distorts this set. To evaluate the hypothesis we implemented SASA (Sensitivity Analyzer for Similarity Assessments), a software prototype used as a test bed for the examination of different processing strategies for an exhaustive set of similarity queries. Section 6.1 explains and justifies the measures chosen to evaluate the incompatibility between two result sets. Section 6.2 gives the general overview of the approach. Sections 6.3 and 6.4 describe the characteristics of the experiments aiming at object and scene-level similarity assessments, respectively, and discuss their results. Section 6.5 summarizes the findings of this study and concludes with the verdict on the hypothesis.

#### 6.1 Measures of Incompatibility

There exist several approaches to compute the deviations between two ranking lists (Mosteller and Rourke 1973; Gibbons 1996). Most rely on statistical tests, which consider the entire range of the lists. An evaluation of ranking lists produced from database queries or web search queries is different, however. The focus here is only on the first few ranks, because the relevance of retrieved items decreases rapidly for lower ranks. For the experiments in this study, the relevant portion of the ranking list was defined as that, which comprises the *ten best results*. This decision was partially based on the experimental outcomes that people retain no more than five to nine items in short term memory (Miller 1956). The rule of 7 +/- 2 items refers to unidimensional stimuli;

therefore, people are expected to be able to retain this number of results in short term memory only for very simple queries. This decision was also based on the typical strategy of current web-search engines, which present ten items per page, starting from the most relevant. Therefore, the set of the ten best results is not only easy to browse and inspect, but also convenient in the sense that users can memorize it to a large degree and perform swift comparative judgments about the relevance of each match to their query.

As the database size grows, the ranks of the ten best results are determined based on finer differences of their similarity values. If one also considers that psychologically compliant methods approximate better, but do not necessarily model human perception exactly, then a measure of incompatibility that relies only on rank differences would be strict. A more practical and objective indicator of the incompatibility between two methods considers instead the overlap of common objects within the relevant portion of the ranking lists. This measure, denoted  $O$ , expresses the percentage of the common items within the ten best results that the compared methods produce. The selection of this measure is also further justified by the fact that each of the items in the relevant portion is equally accessible to the users (i.e., ten results per page).

The actual rank differences are examined as a secondary and less crucial index of incompatibility. They are used as an additional criterion to support or reject the tested hypothesis when the overlap measure provides borderline evidence for that purpose. The rank differences are assessed using a *Spearman Rank Correlation* (SRC) test. This test is an appropriate statistic for ordinal data, provided that its resulting coefficient is used only to test a hypothesis about order (Stevens 1951). The SRC coefficient  $R$ , with  $x_i$  and  $y_i$  as the rank orders of item  $i$  in two compared samples that contain  $n$  items each (Equation 6.1), takes a value between  $-1$  and  $+1$ , where  $+1$  indicates perfect agreement between two samples (i.e., the elements are ranked identically), while  $-1$  signals complete disagreement (i.e., the elements are ranked in inverse order). A value of  $0$  means that

there is no association between the two samples, whereas other values than 0, 1, and -1 would indicate intermediate levels of correlation.

$$R = 1 - \frac{6 \cdot \sum_{i=1}^n (x_i - y_i)^2}{n \cdot (n^2 - 1)} \quad (6.1)$$

The SRC coefficient and similar statistics are designed for evaluations of ranking lists that contain exactly the same elements. Hence, it cannot be readily applied to tests that require a correlation value between a particular subsection of the ranking lists. This observation is essential, because the items in the relevant portion of the lists will only incidentally be the same for two different methods. To enable the comparison of lists with different numbers of entries, a modified SRC coefficient is computed as follows: first, the different elements in the two lists are eliminated and  $R$  (Equation 6.1) is computed for the common elements that remain. Second, the modified coefficient  $R'$  is calculated by multiplying  $R$  with the overlap percentage  $O$  (Figure 6.1). The second corrective step is necessary in order to avoid misleading results. For example, when among the top ten items only one common element exists,  $R = 1$ , but  $R' = 0.1$ .

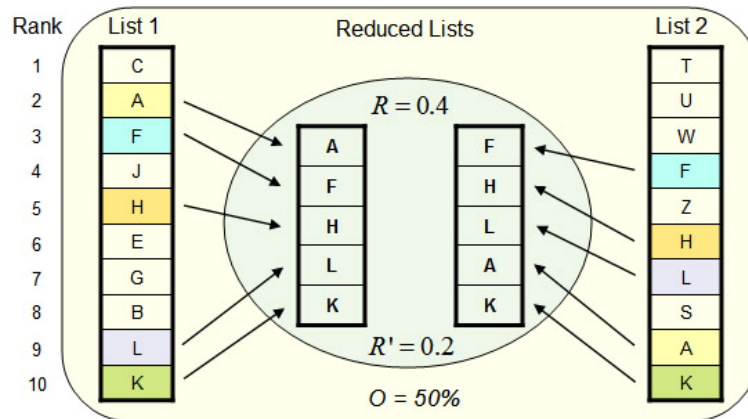


Figure 6.1: Overlap percentage  $O$  and modified Spearman Rank Correlation coefficient  $R'$  for the relevant portion of two ranking lists.

Methods that produce very similar results are characterized by positive values of the measures  $O$  and  $R'$ , close to 1, whereas methods that produce very dissimilar results are characterized by an overlap value close to 0 and by a modified SRC coefficient value close to 0 or negative.

## 6.2 Experimental Design

This thesis postulated that a psychologically compliant (or simply, *compliant*) approach to similarity has three crucial characteristics (Section 1.3.4):

- It identifies groups of integral attributes when they are present (testable hypothesis statement  $HS_1$ ).
- It aggregates these groups to form new separable attributes with a Euclidean metric (Equation 4.2) and, consequently, it combines these and other separable attributes to a total object or relation dissimilarity with the Manhattan metric (Equation 4.3) (testable hypothesis statement  $HS_2$ ).
- It translates the total dissimilarity scores obtained for each pair of objects or relations into similarity estimates using a non-linear conversion function (Equations 2.1b-c) (testable hypothesis statement  $HS_3$ ).

A *psychologically deviant* (or short: *deviant*) method is one that deviates in some way from the psychological findings. Any such deviation affects the similarity scores and may result in different ranks for a reference query. The evaluation consisted of four experiments, each highlighting the distortions on the desirable ranking list, which is produced by the compliant method, when one or several aspects of the hypothesis were violated (Figure 6.2).

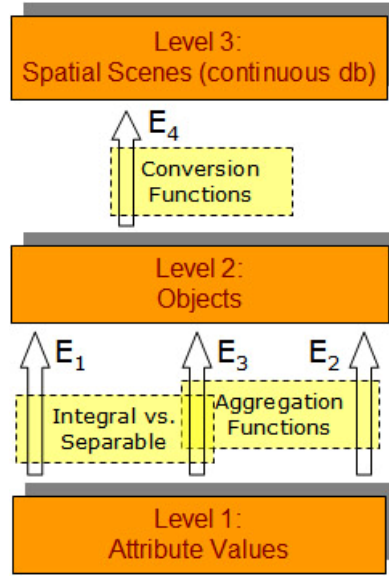


Figure 6.2: Experiments ( $E_1$ - $E_4$ ) used to evaluate the hypothesis.

A violation of the first two arguments of the hypothesis would distort the results for queries at the object level. The extent of such distortions is tested with Experiments  $E_1$ ,  $E_2$ , and  $E_3$ . Experiment  $E_1$  compares the compliant method (Figure 6.3a) with a deviant method that ignores, or does not recognize, possibly existing groups of integral attributes, thus treating each attribute as separable (Figure 6.3b). In Experiment  $E_2$ , the deviant method identifies correctly the groups of integral attributes. It uses, however, the same aggregation function throughout for both integral groups and separable attributes. This conduct is in contrast to the compliant method, which relies on a combination of functions. The variations tested are the single usage of the Manhattan (Experiment  $E_{2A}$ ) (Figure 6.3c) or the Euclidean function (Experiment  $E_{2B}$ ) (Figure 6.3d). Although additional aggregation functions have been proposed (Cross and Sudkamp 2002), the Manhattan and Euclidean metrics are predominant in existing similarity-enhanced information retrieval systems and current prototype implementations (Motro 1988; Petrakis and Faloutsos 1997; Papadias *et al.* 1999b; Dey *et al.* 2002; Ortega-Binderberger *et al.* 2002; Chakrabarti *et al.* 2003). Furthermore, these functions are the closest in form to the compliant method; therefore, proving the hypothesis for them is sufficient to justify

its validity for less similar aggregation functions. Whereas Experiments  $E_1$  and  $E_2$  concentrate exclusively on the integral attributes and the aggregation function hypotheses, respectively, Experiment  $E_3$  examines the combined effect of deviant choices for both of those premises on the results (Figure 6.3e).

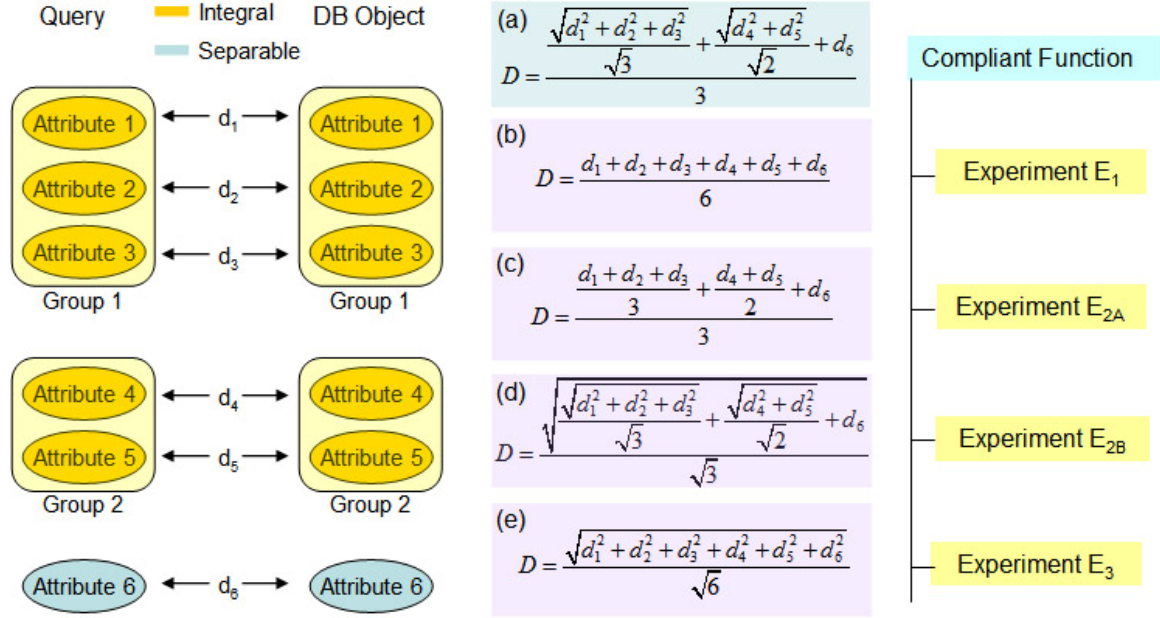


Figure 6.3: Experiments for object-level queries: (a) compliant aggregation function, (b) deviant function that ignores integral attributes ( $E_1$ ), (c) deviant function that aggregates integral attributes with a Manhattan metric ( $E_{2A}$ ), (d) deviant function that aggregates separable attributes with a Euclidean metric ( $E_{2B}$ ), and (e) deviant function that ignores integral attributes and aggregates separable attributes with a Euclidean metric ( $E_3$ ).

The third part of the hypothesis, which is concerned with results to queries at the scene level, is evaluated with Experiment  $E_4$ . To demonstrate the issue behind this section of the hypothesis consider the example in Figure 6.4. The association graph for scene queries is created by matching objects and relations below a certain dissimilarity threshold. The threshold used in this example is 0.6. The final similarity score of each solution (i.e., maximal clique) extracted from the association graph is computed with Equation 5.5. Excluding the completeness correction factor, this equation is a weighted



average of the similarities for the object (i.e., node similarities) and the relational (i.e., edge similarities) components of the association graph. Converting the dissimilarities to similarities at the nodes and edges with a linear, an exponential, and a Gaussian function yields three different rankings of the derived solutions to the submitted scene query. The linear function assigns equal importance to any match, whereas the non-linear functions promote highly similar pairs and disfavor highly dissimilar ones. The goal of Experiment E<sub>4</sub> is, therefore, to assess the extent of variation for scene results in the relevant portion of the ranking lists when different conversion functions are employed. The three types of functions considered in this experiment correspond to Equations 2.1a-c. In contrast to what psychologists have suggested, the linear function, which treats similarity and dissimilarity as complementary magnitudes, accounts for the majority of current systems' approach to similarity (Papadias *et al.* 1999b; Blaser 2000; Goyal and Egenhofer 2001).

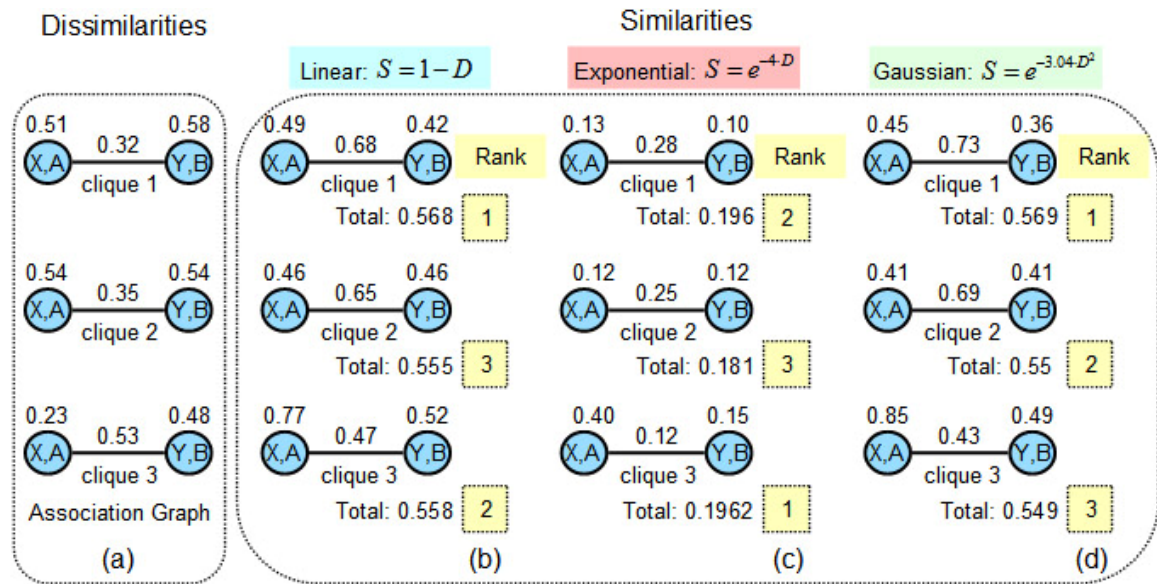


Figure 6.4: Experiment for scene-level queries: (a) three solutions to a scene query with dissimilarities computed for each node (i.e., object) and for each edge (i.e., relation), and the scene ranks produced when the dissimilarities were converted to similarities using (b) a linear function, (c) an exponential function, and (d) a Gaussian function.

A difficulty with the third postulate of the hypothesis is the lack of specificity in the psychological findings on which it was based. Although it is generally accepted that the conversion function should follow an exponential or Gaussian gradient, the exact form of such a function is not further elaborated, possibly because it may vary slightly depending on the stimuli under consideration. Mathematically, this uncertainty is represented by the coefficient  $c$  of Equations 2.1b-c, which is left unspecified. These equations describe, therefore, families of functions, rather than individual functions.

To compensate for this ambiguity, Experiment  $E_4$  compares the linear function  $L$ , with several different versions of the exponential and Gaussian alternatives (i.e., obtained for different values of the  $c$  parameter), abbreviated hereafter as  $E_i$  and  $G_i$ , respectively (Figure 6.5). The curves of  $E_L$  and  $G_L$  were made to fit the data of  $L$  with a regression technique. In this sense,  $E_L$  and  $G_L$  are the closest to the linear plot. The pairs  $(E_S, G_S)$  and  $(E_G, G_G)$  were defined such that they represent very strict and very generous functions of similarity, respectively. *Strict* means that similarity drops very fast as dissimilarity increases, whereas *generous* implies that similarity diminishes very slowly with a dissimilarity increase. These behaviors are also evident from inspecting Figure 6.5: the pair  $(E_S, G_S)$  is located, for the most part, to the left of the linear function, whereas the pair  $(E_G, G_G)$  lies mainly to its right. For both the strict and the generous pairs, the exponential function was first obtained empirically and the Gaussian was subsequently derived through regression.

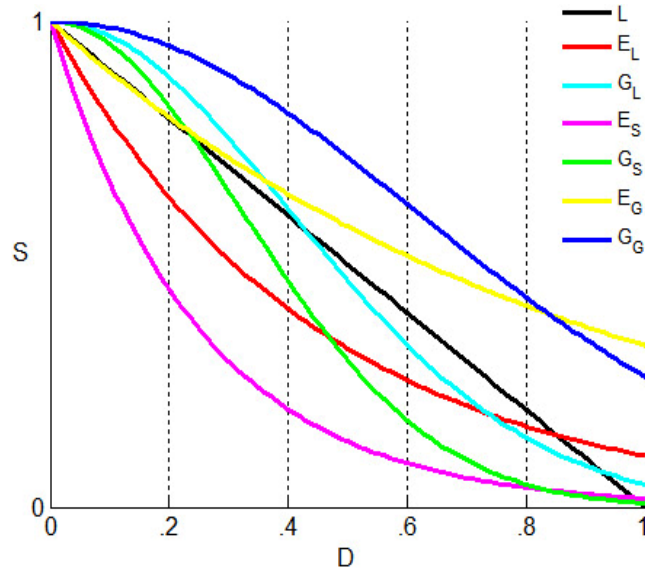


Figure 6.5: Strict and generous non-linear conversion functions used in Experiment E<sub>4</sub>.

Although in most psychological experiments the exponential and Gaussian curves fit better the points plotted from human similarity-dissimilarity judgments (Nosofsky 1986; Shepard 1987), the differences are often subtle (Ennis 1988). In some efforts, the slope of the regression lines obtained for such judgments was evaluated to be very close to -1 (Attneave 1950; Hosman and Kuennapas 1972; Tversky 1977), which is the slope of the linear function. Moreover, similarity and dissimilarity judgments mirrored each other closely in several MDS studies where, under some circumstances, they both produced almost identical results (Rapoport and Fillenbaum 1972). These observations suggest that the plots of the psychologically representative non-linear alternatives should not deviate significantly from the straight line of the linear method. For this reason, the emphasis for the validity of this section of the hypothesis is on comparisons between the three functions  $L$ ,  $E_L$ , and  $G_L$ . Experiments that involve additional pairs complement the investigation by revealing the relative behavior of members of the same (e.g.,  $E_L$  vs.  $E_G$ ) or different (e.g.,  $E_S$  vs.  $G_S$ ) families of functions, and by allowing inferences about the

repercussions on the results when more arbitrary choices are made for the conversion function.

The results of every experiment comprise two ranked lists, one obtained with each of the compared methods. The compatibility of these lists is then evaluated according to the value of the overlap  $O$  and the modified SRC coefficient  $R'$  (Section 6.1) for the relevant portion of the lists. In the following experiments, negative values of  $R'$  were rarely obtained and in all cases these values were only marginally below 0 (i.e., -0.05 in the worst case). To allow a uniform visualization scheme such values were truncated to 0, and the range used for both measures was delimited in the closed interval  $[0,1]$ .

The exhaustive character of the experiments was a prohibitive factor in locating real-world datasets that accommodate all of the tested scenarios. Hence, the assessment relies on simulations with synthetic datasets and queries, randomly generated within SASA. These synthetic constructs were originally populated with random values that followed different statistical distributions each time (e.g., uniform, normal). The underlying distribution of the data had a negligible effect on the final results. The distribution of random values is, therefore, kept constant and assumed to be uniform throughout this study. Likewise, a consideration of different attribute types in the simulated databases is immaterial for the purposes of the experiments, because all algorithms that perform atomic value assessments yield a dissimilarity measure between 0 and 1 regardless of the attribute type (Chapter 3). The focus of the experiments, however, is to examine how such atomic dissimilarities should be combined to create scores of aggregate dissimilarity and, consequently, how these scores should be converted to similarity values. Each experiment was conducted several thousand times and the results were averaged in order to make the measures  $O$  and  $R'$  converge to their medium values. The number of repetitions was determined empirically, such that successive executions of the experiments for that number of cycles yielded results with a deviation of less than 1%.

### 6.3 Experiments at the Object Level

This section describes the setup and discusses the results obtained from Experiments E<sub>1</sub>-E<sub>3</sub>. These experiments assume the existence of a user-submitted object query against which the similarity of the objects in the database is calculated with compliant and deviant approaches. All attributes of the query are weighted equally.

#### 6.3.1 Setup

The similarities or dissimilarities of the ranks obtained in response to an object query with different methods are captured through the incompatibility measures  $O$  and  $R'$ , which are each functions of five variables  $n$ ,  $m$ ,  $p$ ,  $g$ , and  $d$  (Equation 6.2):

$$O, R' = f(n, m, p, g, d) \quad (6.2)$$

- Variable  $n$  is the number of objects in the database, determining the database size. The experiments were conducted for the set  $N = \{1,000, 5,000, 25,000, 100,000\}$ , so that each database size increases approximately one order of magnitude over its predecessor. A dataset of 1,000 objects was adopted as a characteristic case of a small database, a dataset of 100,000 objects as a characteristic case of a large database, whereas datasets of 5,000 and 25,000 objects were used as representatives of medium-small and medium-large databases, respectively.
- Variable  $m$  is the number of attributes for each object, determining the number of attributes that participate in the similarity assessment of a database object to a query object. The set examined is  $M = \{2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$  and accounts for the most simple and complex modeled objects. The case of queries on a single attribute is omitted, because it is irrelevant for both hypotheses tested. For the integral-attributes hypothesis, one integral attribute is undefined because it essentially degenerates to one separable attribute. For the aggregation hypothesis, the rankings

produced by different aggregation functions become identical when the query involves a single attribute (i.e., no aggregation of dissimilarity measures takes place).

- Variable  $p$  is the percentage of integral attributes out of the total number of attributes  $m$ . The actual number of integral attributes is, therefore,  $p \cdot m$ . In this manner,  $p$  also indirectly determines the number of separable attributes. The percentages taken are  $P = \{0\%, 10\%, 20\%, 30\%, 40\%, 50\%, 70\%, 80\%, 90\%, 100\%\}$ . The two extreme values of 0% and 100% represent the cases where all attributes are separable and integral, respectively.
- Variable  $g$  is the number of integral groups in which the integral attributes are distributed. The possible values for this variable are constrained by the specific instantiations of the variables  $m$  and  $p$ . For example, when the objects have ten attributes ( $m=10$ ), four of which are integral ( $p=40\%$ ), then the number of integral groups  $g$  could either be 1 (i.e., one group of four attributes) or 2 (i.e., two groups of two attributes). For the experiments in this thesis,  $g$  has a range from 1 to 50. The smallest value occurs in various settings, starting with the case for  $m=2$  and  $p=100\%$ . The largest value occurs only if  $m=100$  and  $p=100\%$ .
- Variable  $d$  is the group distribution policy. This parameter describes how a number of integral attributes  $p \cdot m$  is distributed in a number  $g$  of integral groups. For some configurations there could be numerous such possibilities. For instance, when eight integral attributes must be distributed in two groups, there can be multiple allocations, such as 6-2, 5-3, and 4-4. Preliminary experimentation indicated that the results can be affected by the distribution policy, especially for larger percentages of integral attributes. This parameter is treated as a binary variable taking the values “optimal” and “worst.” An optimal distribution policy tries to distribute the integral attributes evenly, such that each integral group contains approximately the same number of attributes (Figure 6.6.a). A worst distribution policy will create disproportionately-

sized groups by assigning as many attributes as possible to one large integral group, while populating the remaining groups with the minimum required amount of attributes (Figure 6.6b). The binary treatment of the group distribution policy allows inferences about the behavior of this variable between its two extremes settings, while keeping the number of produced diagrams within realistic limits.

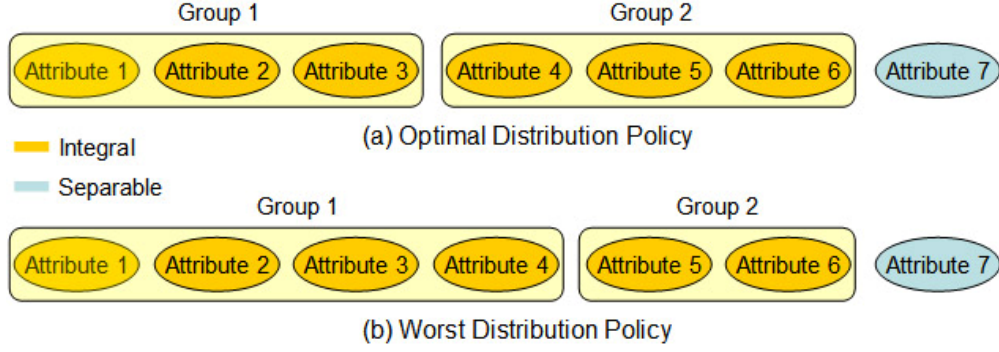


Figure 6.6: Splitting integral attributes into groups using (a) an *optimal* and (b) a *worst* distribution policy.

A specific instantiation of the variables  $n$ ,  $m$ ,  $p$ ,  $g$ , and  $d$  represents a possible database configuration and is referred to as a *db scenario*. The simultaneous interaction of all variables involved for such db scenarios and their effect on the ranks cannot be accommodated by the representational capabilities of typical 2-dimensional or 3-dimensional visualization techniques due to the large amount of diagrams that would have to be produced. In order to visualize the results effectively, while keeping the number of produced diagrams within acceptable bounds, a 4-dimensional visualization technique was employed. For each 4-dimensional diagram, the database size  $n$  and the distribution policy  $d$  are kept fixed, while the remaining variables are allowed to vary within a 3-dimensional cubic space. The axes  $X$ ,  $Y$ , and  $Z$  of this space correspond to the number of integral groups  $g$ , the number of attributes  $m$ , and the percentage of integral attributes  $p$ , respectively. Each point in the cubic space signifies, therefore, a db scenario determined by the instantiation of the triple  $(m, p, g)$  that defines the point, and the fixed values of  $n$  and  $d$ . The color assigned to a db scenario (i.e., point) embeds a fourth

dimension in the visualization, which represents the measurement of  $O$  or  $R'$  (i.e., the overlap or the modified SRC coefficient) between the two compared methods for that db scenario. Since there are two incompatibility measures, four database sizes, and two distribution policies, a total of sixteen diagrams was produced for each experiment.

As an example, consider the 4-dimensional diagrams of Figure 6.7. Point **A** in this figure corresponds to the scenario of a database of 1,000 objects, each having 40 attributes. There are 20 separable and 20 integral attributes. The latter are distributed in 10 groups through an optimal distribution policy, meaning that each group contains 2 attributes. For the db scenario of point **A**, the overlap measure is approximately 40%, whereas the value of  $R'$  is approximately 0.2.

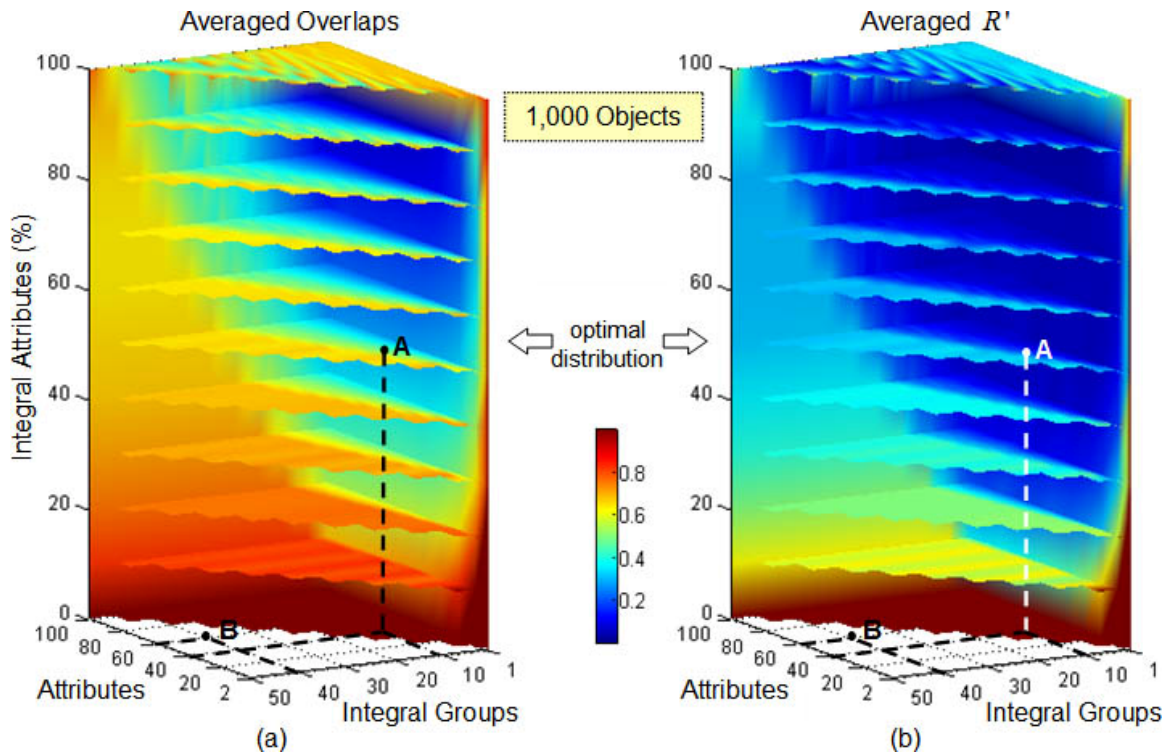


Figure 6.7: A 4-dimensional diagram depicting the measures (a)  $O$  and (b)  $R'$ .

A triangular half of the volumes of the produced cubes is not populated with measurements, because it corresponds to non-applicable db scenarios. For example, point **B** in Figure 6.7 is such a db scenario, because it is impossible to allocate 60 integral



attributes within 40 groups. Realizable db scenarios are located within the remaining half of the cube. Since the values of the variables  $m$ ,  $p$ , and  $g$  are discrete, the realizable db scenarios form a dense grid, rather than a continuous surface. The diagrams, however, use continuous color-rendered surfaces instead—produced by interpolating the grid values—in order to facilitate the interpretation of the results. Furthermore, the cube is sliced at regular intervals along the Z-axis to reveal the patterns in its interior.

### 6.3.2 Results and Discussion

The next sections present and discuss the results obtained from Experiments  $E_1$ ,  $E_{2A}$ ,  $E_{2B}$ , and  $E_3$  (Figure 6.3). Each experiment comprises 16 diagrams, accompanied by a summarizing figure, which reveals the overall trend of the results for different database sizes.

#### *6.3.2.1 Results of Experiment $E_1$ and Interpretation*

The results obtained for the first testable statement of the hypothesis  $HS_1$ , which evaluates how ignored integral attributes affect the results, are displayed for various dataset sizes in ascending order (Figures 6.8-6.11). Figure 6.12 provides the summarizing overview.

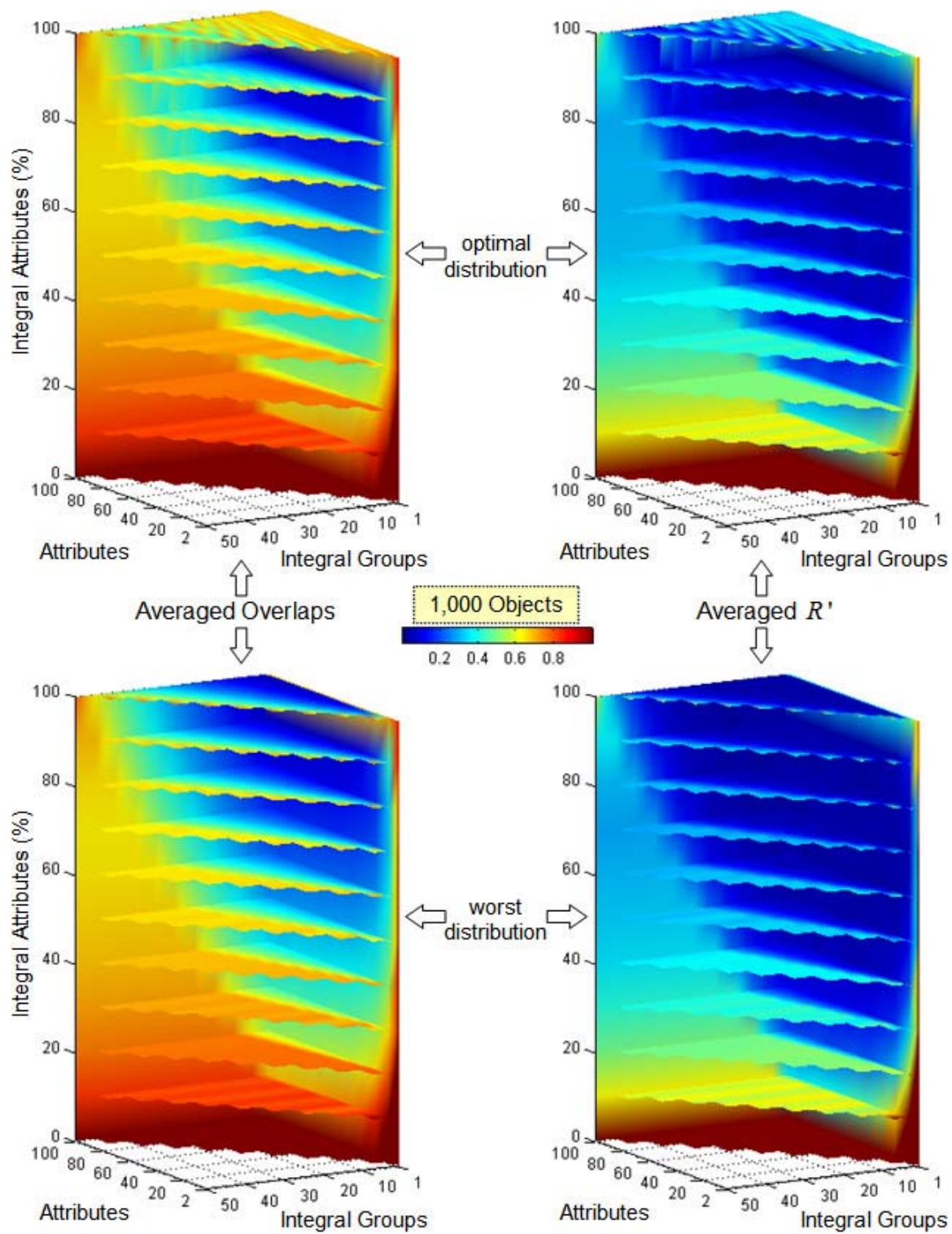


Figure 6.8: Experiment E<sub>1</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects.

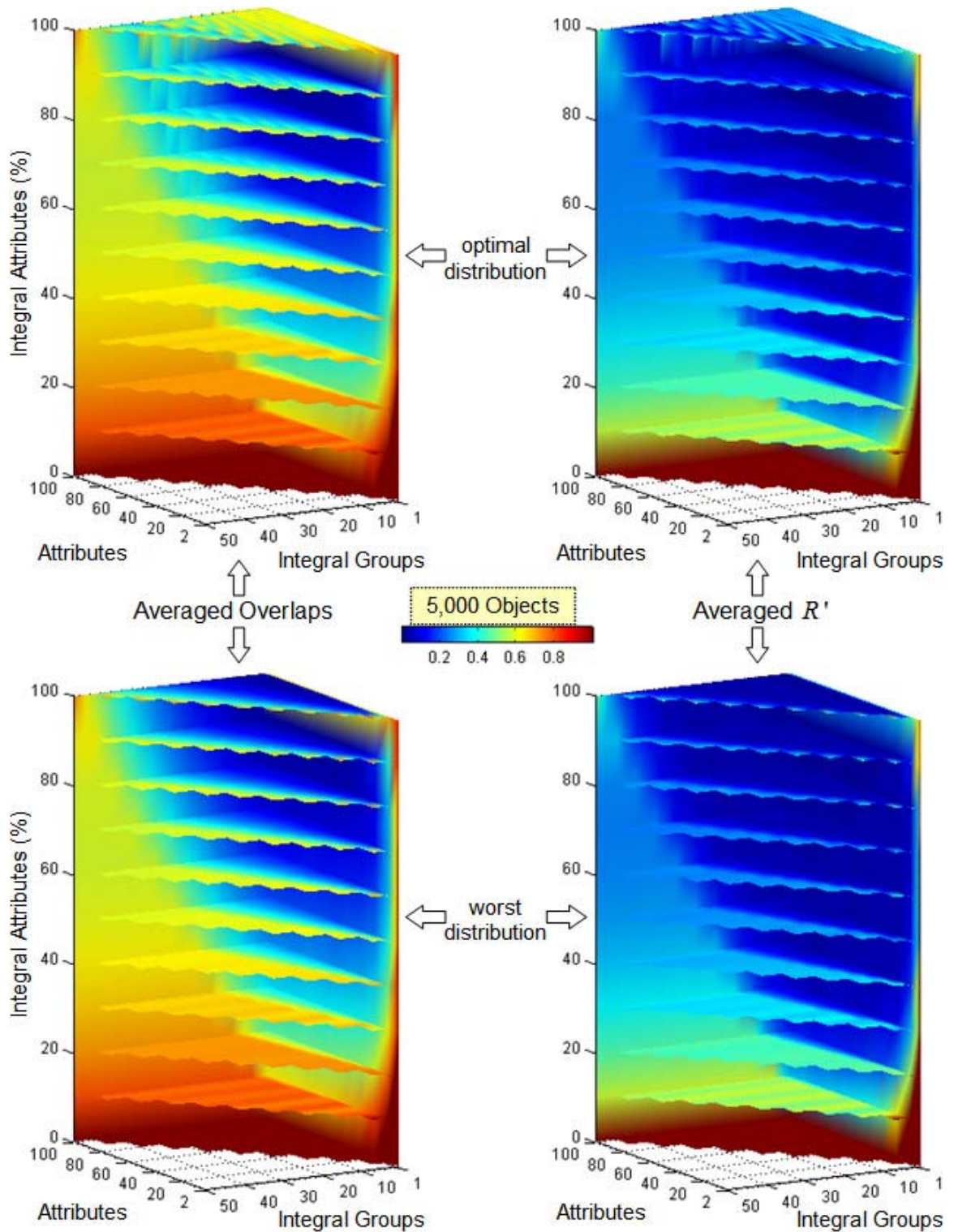


Figure 6.9: Experiment E<sub>i</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects.



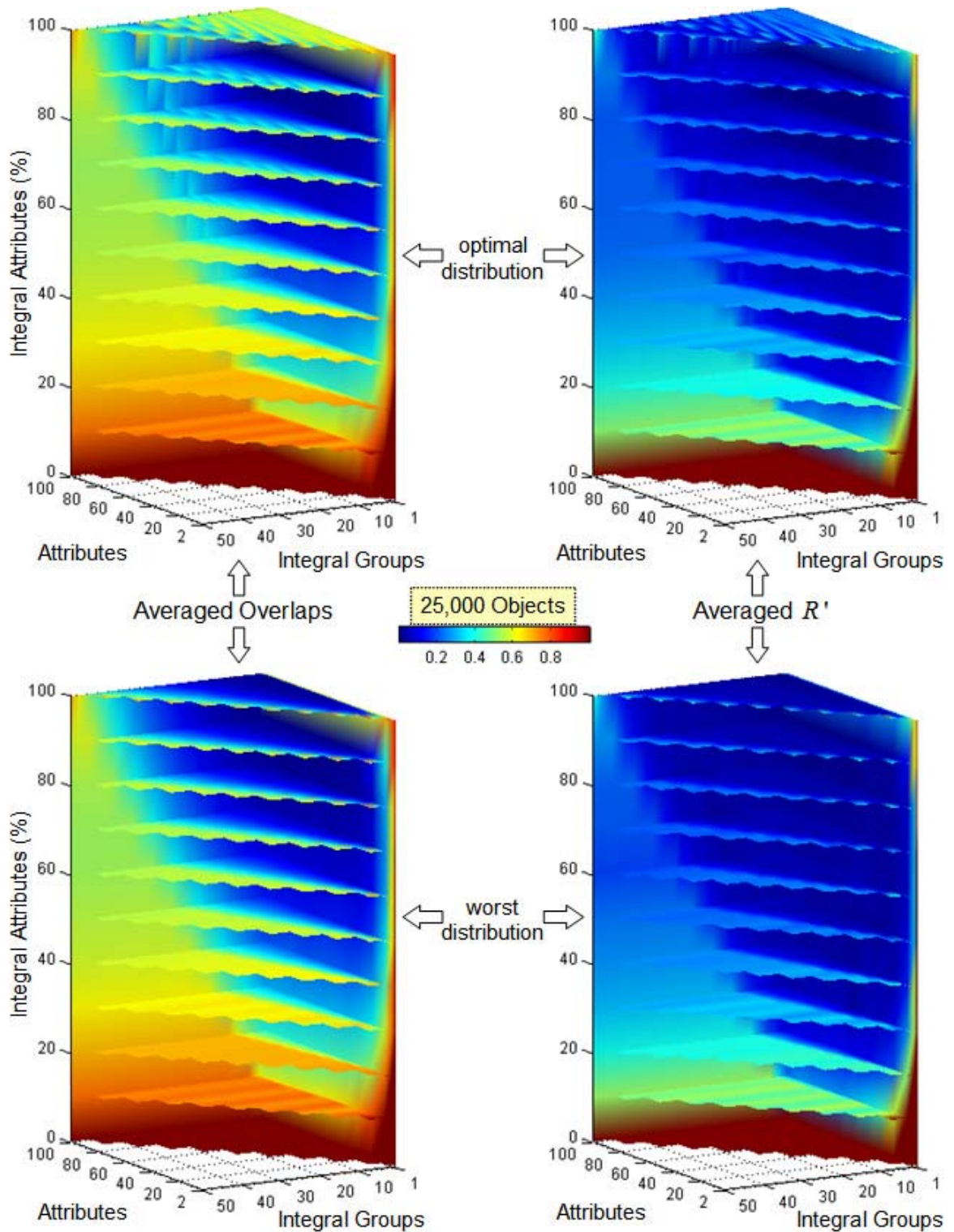


Figure 6.10: Experiment E<sub>i</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects.

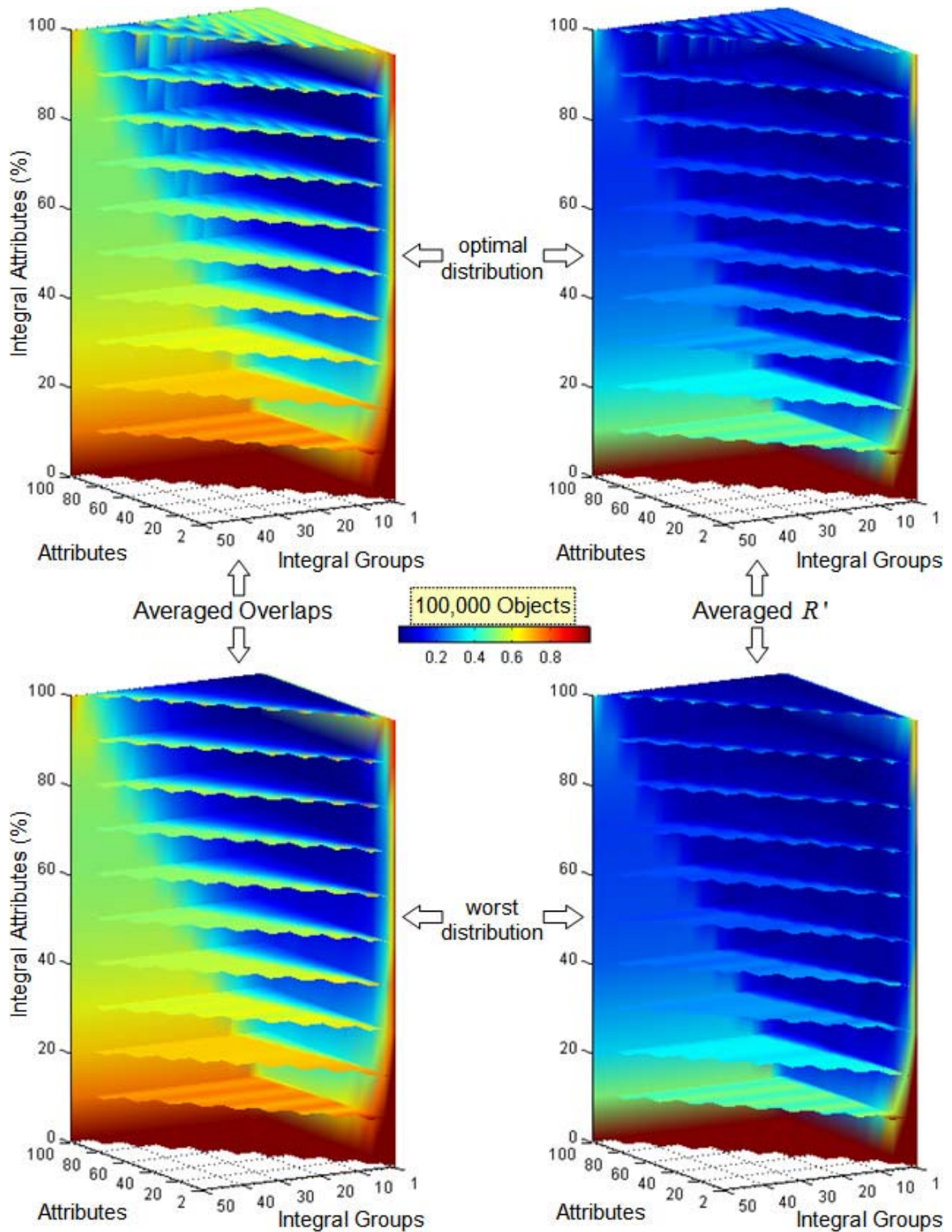


Figure 6.11: Experiment E<sub>i</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects.



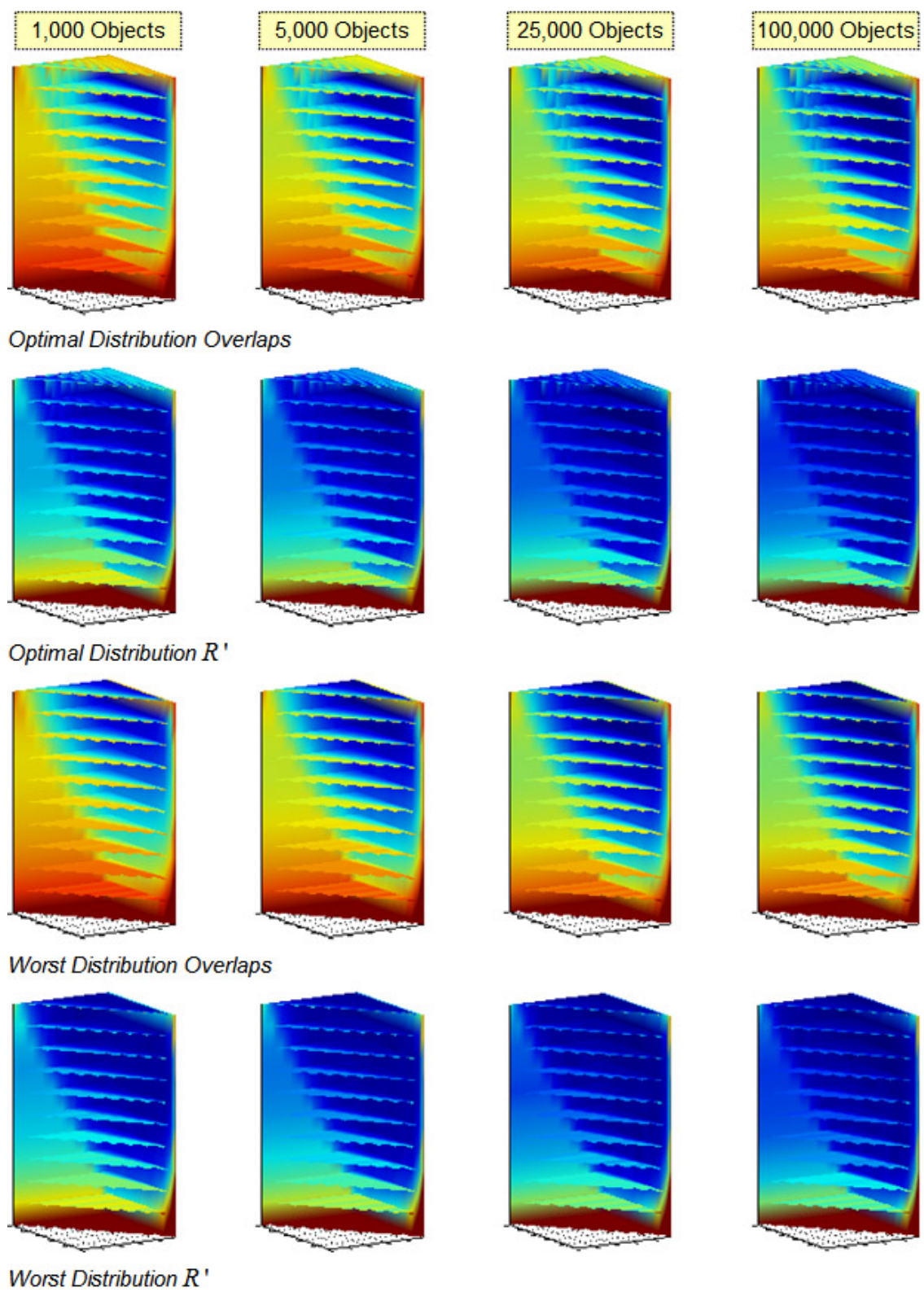


Figure 6.12: Overview of the results acquired from Experiment  $E_1$ .

These results indicate a definitive pattern of gradual variation. The deviant method in this experiment is a manifestation of the Manhattan distance function with no integral groups recognized. Hence, the number of aggregated terms is always equal to the total number of attributes  $m$ . Furthermore, each term contributes equally to the similarity score assigned to each object of the database. As the variables change, the form of the compliant method becomes more or less similar to the pattern of the deviant method. The interactions behind these deviations explain the outcome illustrated in the diagrams.

The main conclusion is that the measures  $O$  and  $R'$  become progressively worse as the percentage of integral attributes increases and the number of groups in which these integral attributes are distributed decreases. When either or both trends occur, the aggregated terms with the compliant method reduce to a number much less than  $m$ . For example, for one separable attribute, nine integral attributes, and three groups, the deviant method aggregates ten terms and the compliant four terms. Moreover, the effect of the one remaining separable attribute with the compliant method is disproportionate on the final score compared to that of the other attributes. As the number of groups increases, the measures have a greater concordance, because the impact of such isolated attributes on the final score diminishes.

This observation also explains the dissonance to the deterioration pattern observed at the highest layer of the optimal distribution policy diagrams, where such separable attributes disappear. The even distribution of integral attributes into groups makes the compliant method behave similarly to the deviant at this layer. For example, consider a query with ten attributes, all of which are integral and must be distributed in five groups. The deviant approach will aggregate all ten attributes as separable. The compliant will first separate the ten attributes in groups of two, aggregate each group, and combine the resulting five terms to derive the object's similarity. For a single group, the compliant method becomes identical to the Euclidean distance function. The trend of deterioration, however, is not interrupted at the highest layer of the diagrams for the worst distribution

policy because the group sizes with this policy differ drastically. In this case, the smaller integral groups continue to have a disproportionate influence on the final similarity score.

In general, the more uniform the distribution into groups is, the less significant the effects on the measures  $O$  and  $R'$  become. The wavy patterns at the higher layers of the diagrams that depict the optimal distribution measures are also related to this conclusion. Such effects are due to the alternating exact and approximate division of integral attributes into groups. For example, for nine integral attributes and three groups the division is exact with three attributes in each group. For ten or eleven integral attributes, the groups differ in size by necessity, whereas for twelve attributes, the groups contain again the same number of elements. In the diagrams of the worst distribution policy where group sizes remain consistently imbalanced, the small stripes of temporary improvements disappear. Excluding the wavy patterns and the case of all attributes being integral, the measures appear to be invariant to the group distribution policy elsewhere.

The results worsen slightly with an increase in the number of attributes; however, the influence of this variable is much more subtle compared to the others. When the attribute number is very small, and especially at its lowest setting (i.e., 2), the methods are often identical, because the attributes are insufficient to form integral groups (e.g., for two attributes and up to 50% percentage of integral attributes). This observation explains the cause for the very high values of  $O$  and  $R'$  detected at the rightmost edge of the diagrams.

The compared methods also yield progressively different outcomes as the database size increases (Figure 6.12). This was an anticipated result, because two functions are expected to demonstrate approximately the same degree of correlation regardless of the sample size with which they are tested. Hence, if the entire ranking lists were considered (i.e., if the lists contained all database objects), and assuming all other variables equal, the two compared methods would exhibit on average the same correlation, regardless of the database size. Increasing the number of objects in the database, while keeping the size



of the relevant portion constant leaves more potential for variations within the ten best results and explains why the overlaps and correlations decline for larger databases.

Both  $O$  and  $R'$  take a value of 1 at the lowest layer where all attributes are separable and the compared methods coincide. For all other db scenarios, the modified Spearman Rank Correlation coefficient  $R'$  has a lower value than the overlap  $O$ . This result is not surprising considering that  $R'$  is a stricter measure than  $O$ . The diagrams suggest that the correct recognition of integral attributes and groups is immaterial for smaller datasets as long as the percentage of integral attributes remains below 40%. For the largest database considered this limit drops to around 20%. At these percentages,  $O$  and  $R'$  have values of 0.5 and 0.2, respectively. Such values constitute borderline measurements for the acceptance of the first hypothesis statement  $HS_1$ , because they imply that only half of the retrieved objects in the relevant portion are the same and that these common objects are ranked very differently. Therefore, there is an approximate value for the percentage of integral attributes, which determines when this hypothesis should be accepted or rejected, and this value drops as the database size increases. Since there is no way, however, to know the percentage of integral attributes unless one identifies them first, hypothesis  $HS_1$  must be universally accepted. The validity of the first premise of the hypothesis is also corroborated by the fact that real-world geographic databases can often be much larger than the largest dataset in this experiment. The single exception, where the task of recognizing the integral attributes can be dismissed with certainty, is when there are no more than two or three attributes for the objects in the database.

#### *6.3.2.2 Results of Experiments $E_{2A}$ and $E_{2B}$ and Interpretation*

This section presents and discusses the results acquired for hypothesis statement  $HS_2$ , which is concerned with the choice of the aggregation function. The compliant aggregation function is compared to the Manhattan metric (Figures 6.13-17) and to the Euclidean metric (Figures 6.18-22).

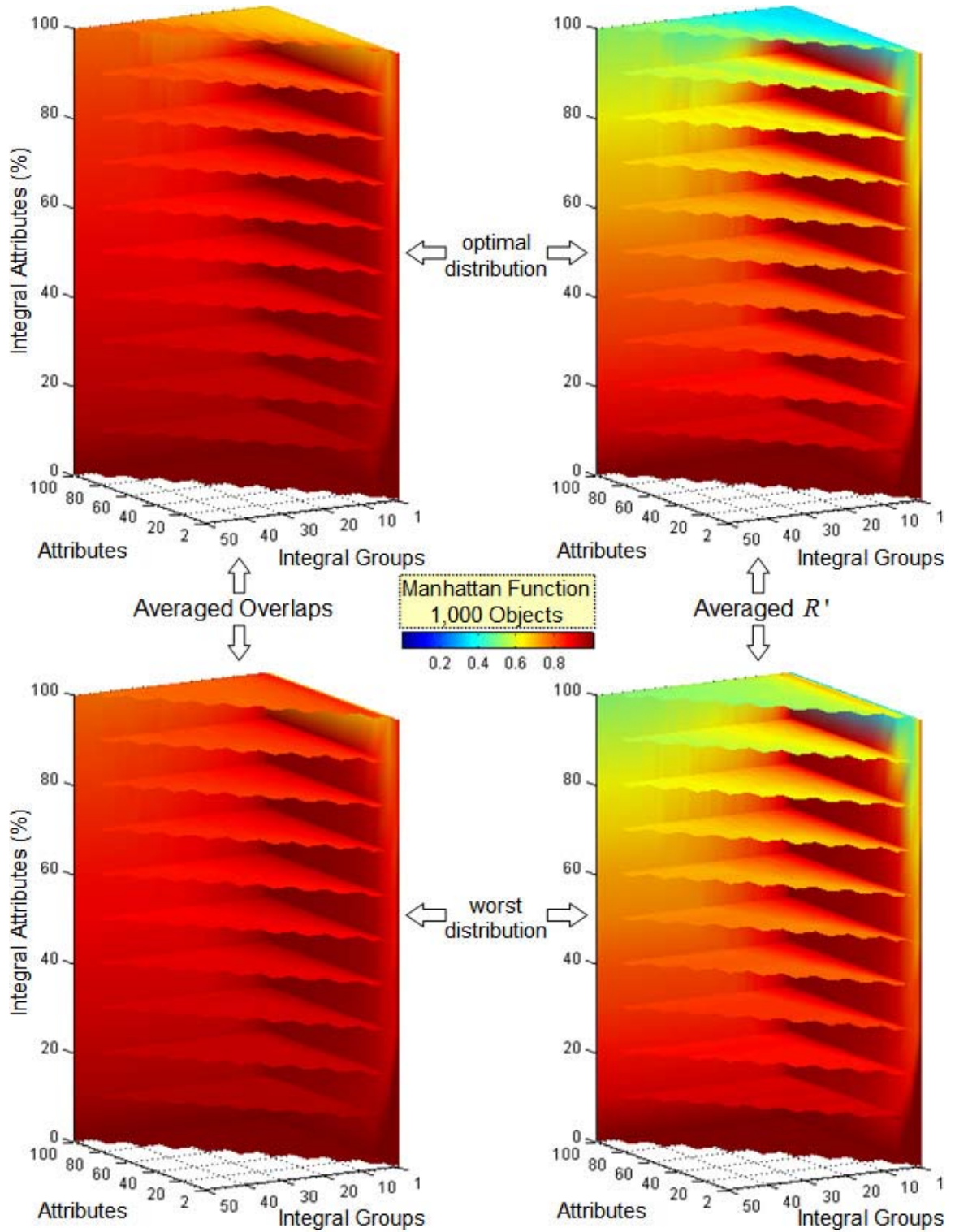


Figure 6.13: Experiment E<sub>2A</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects.

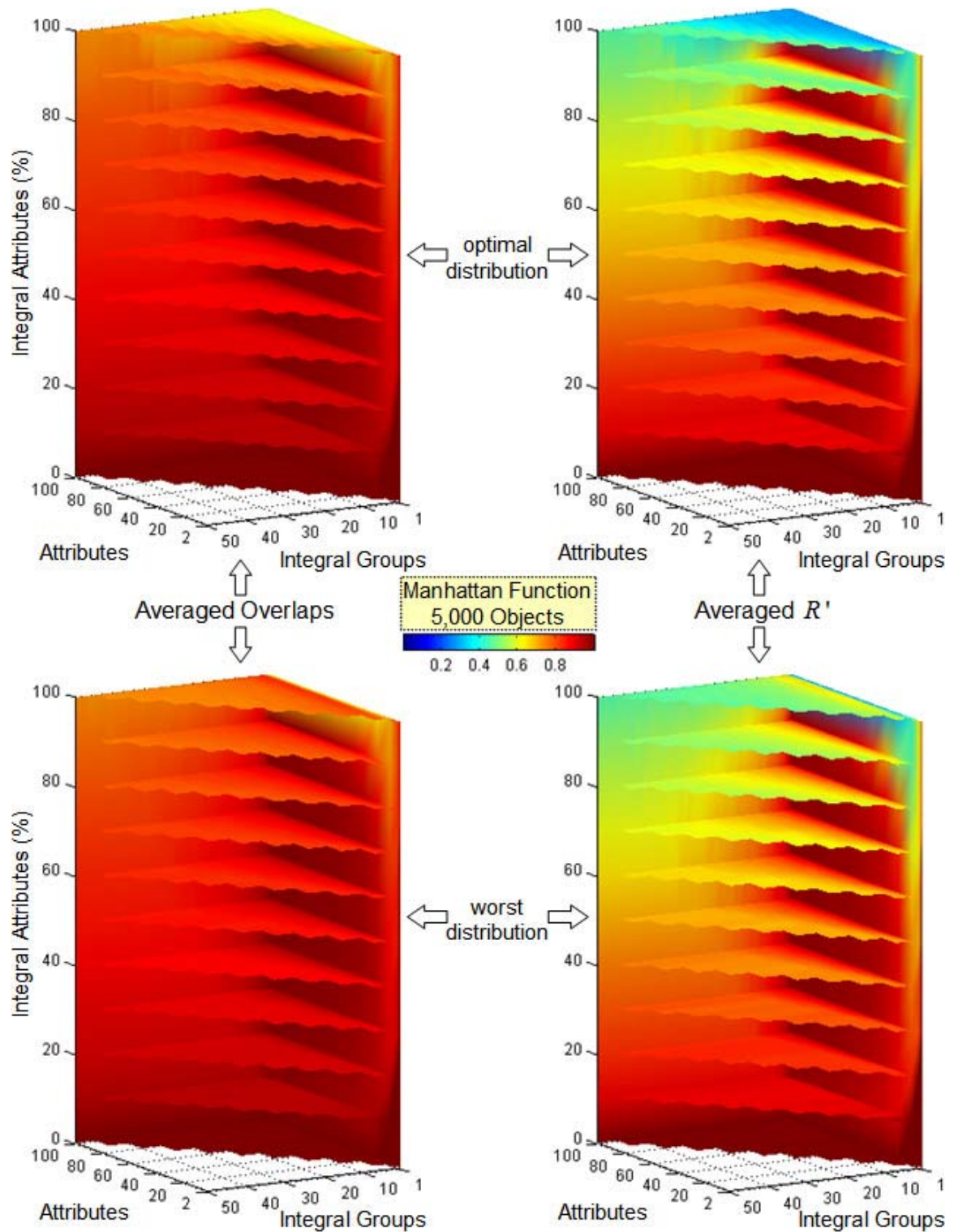


Figure 6.14: Experiment E<sub>2A</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects.

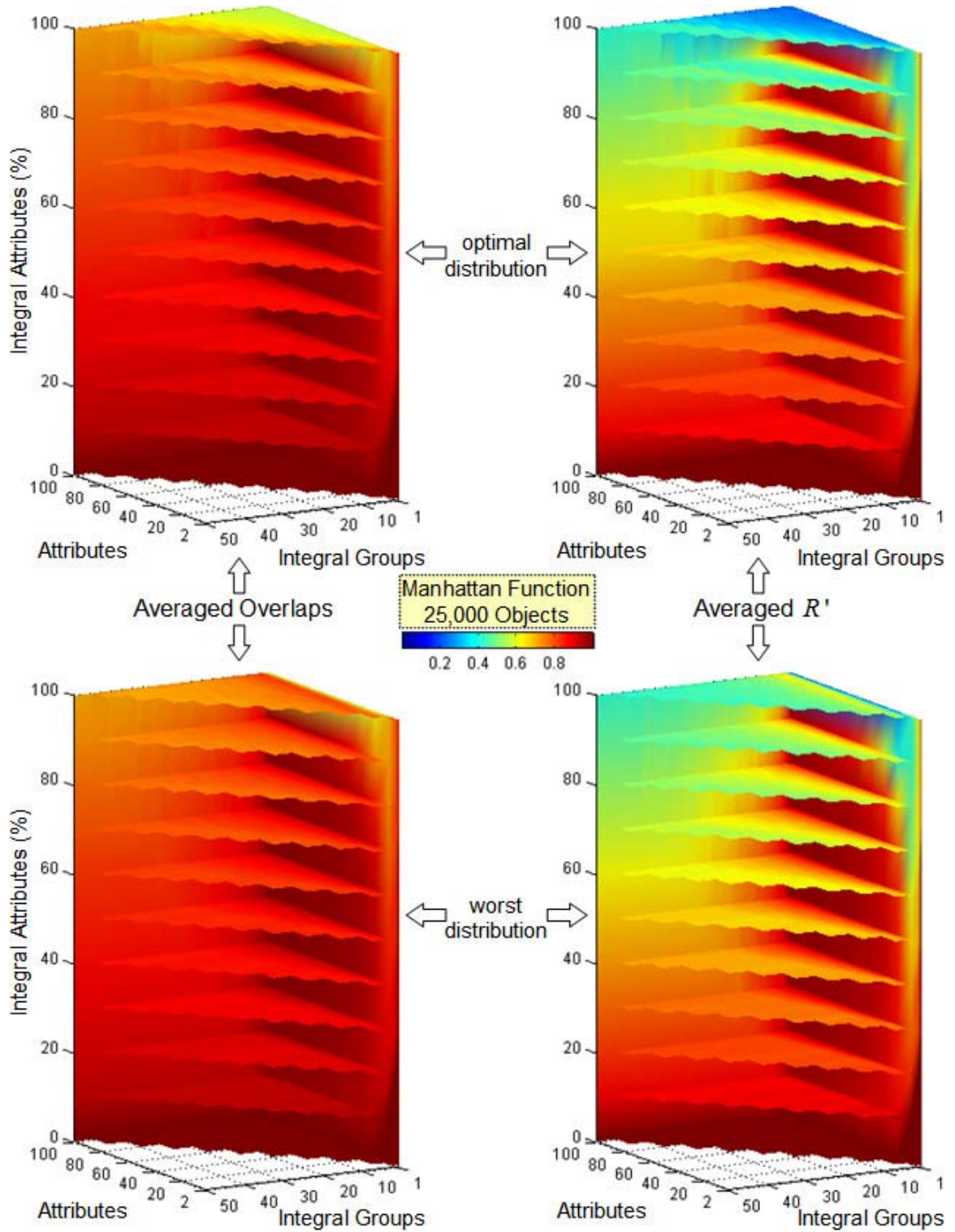


Figure 6.15: Experiment E<sub>2A</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects.



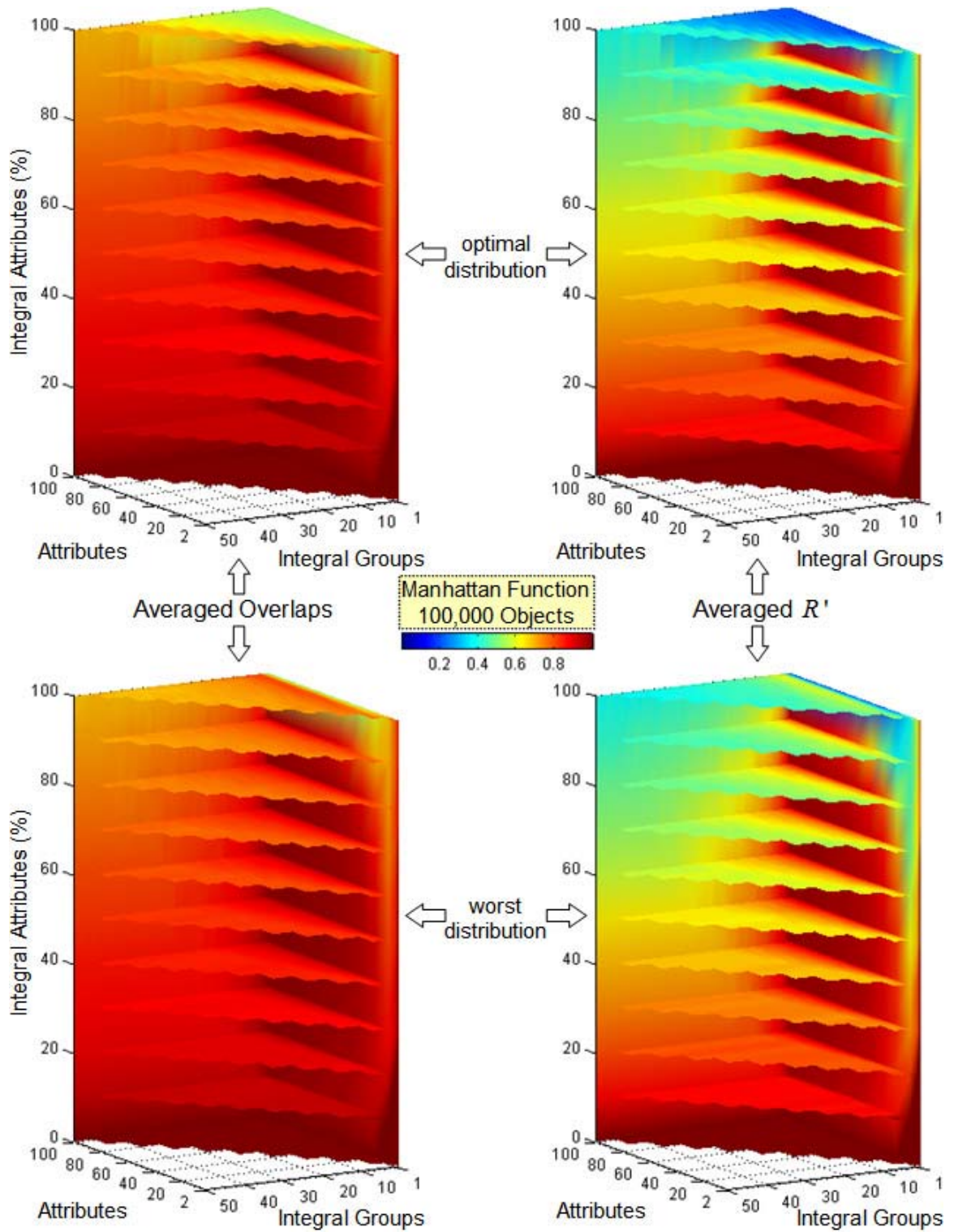


Figure 6.16: Experiment E<sub>2A</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects.

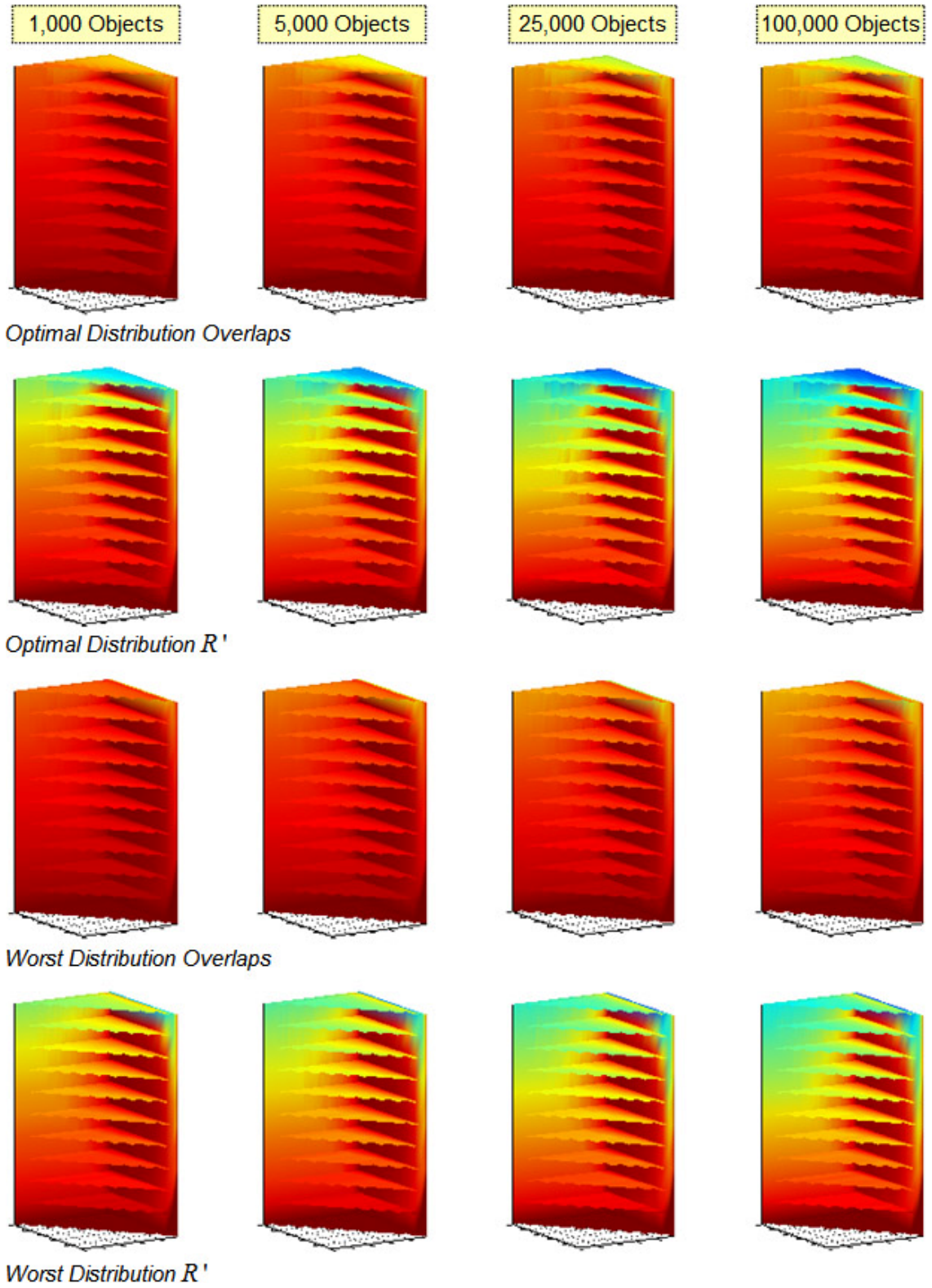


Figure 6.17: Overview of the results acquired from Experiment  $E_{2A}$ .

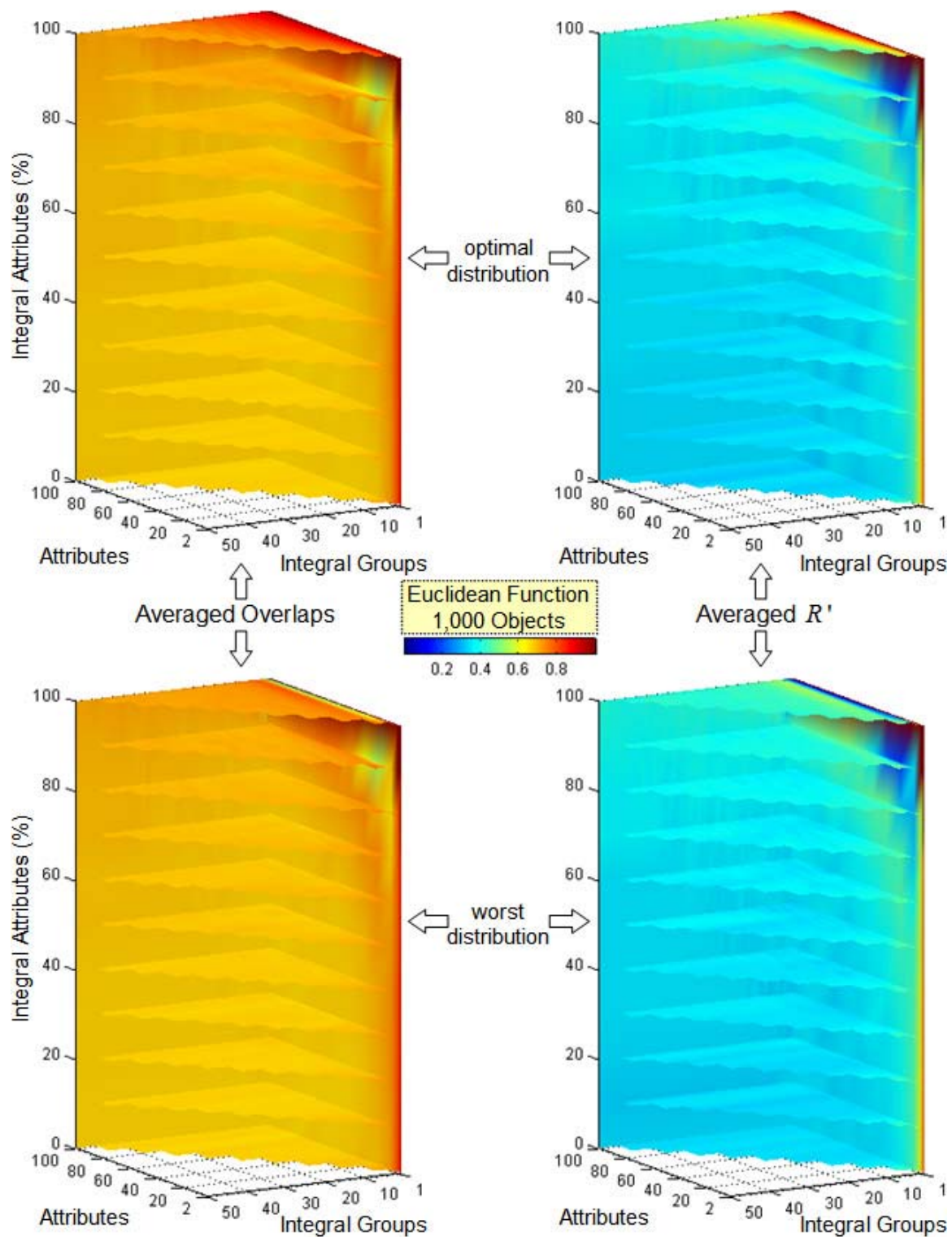


Figure 6.18: Experiment E<sub>2B</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects.



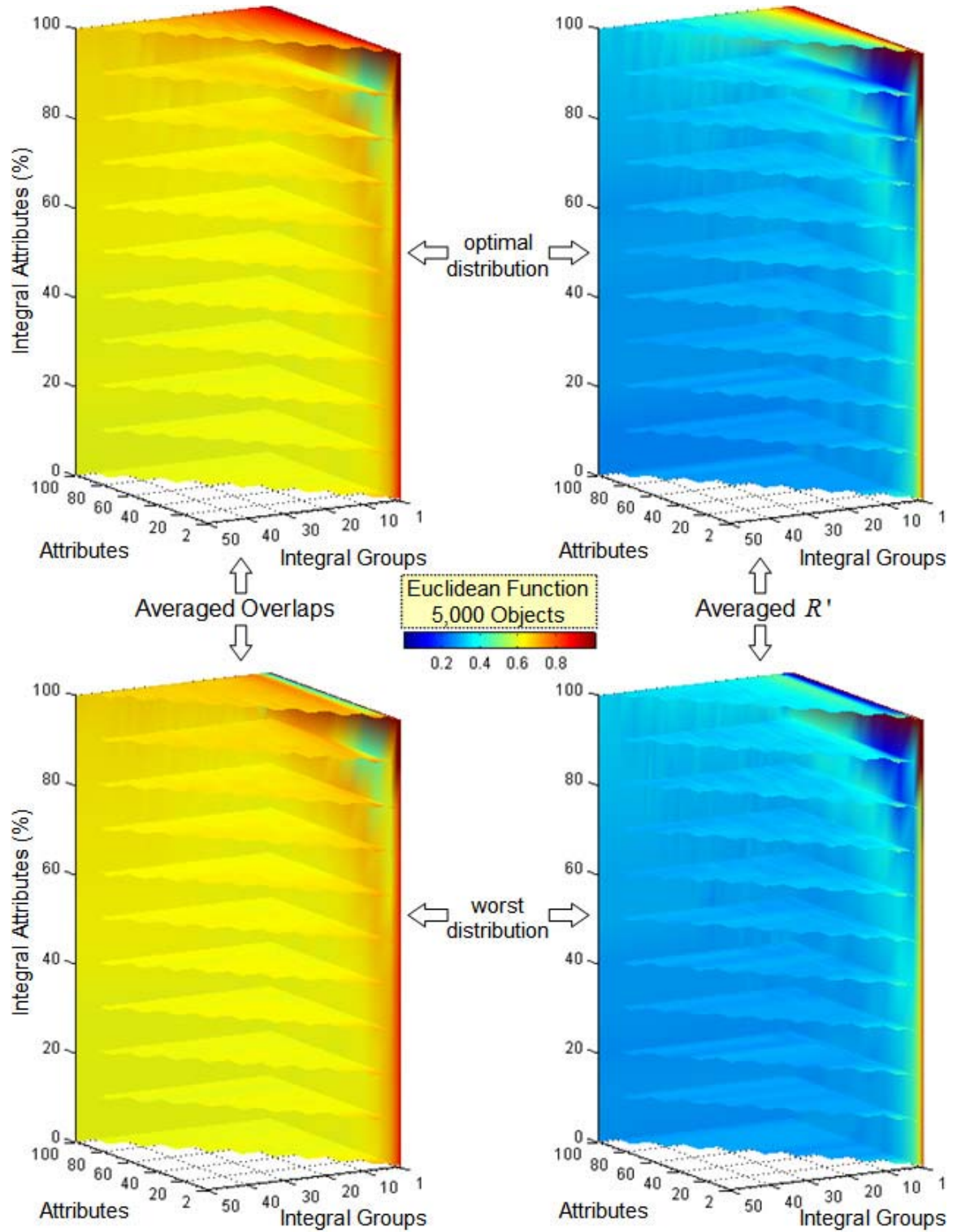


Figure 6.19: Experiment E<sub>2B</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects.



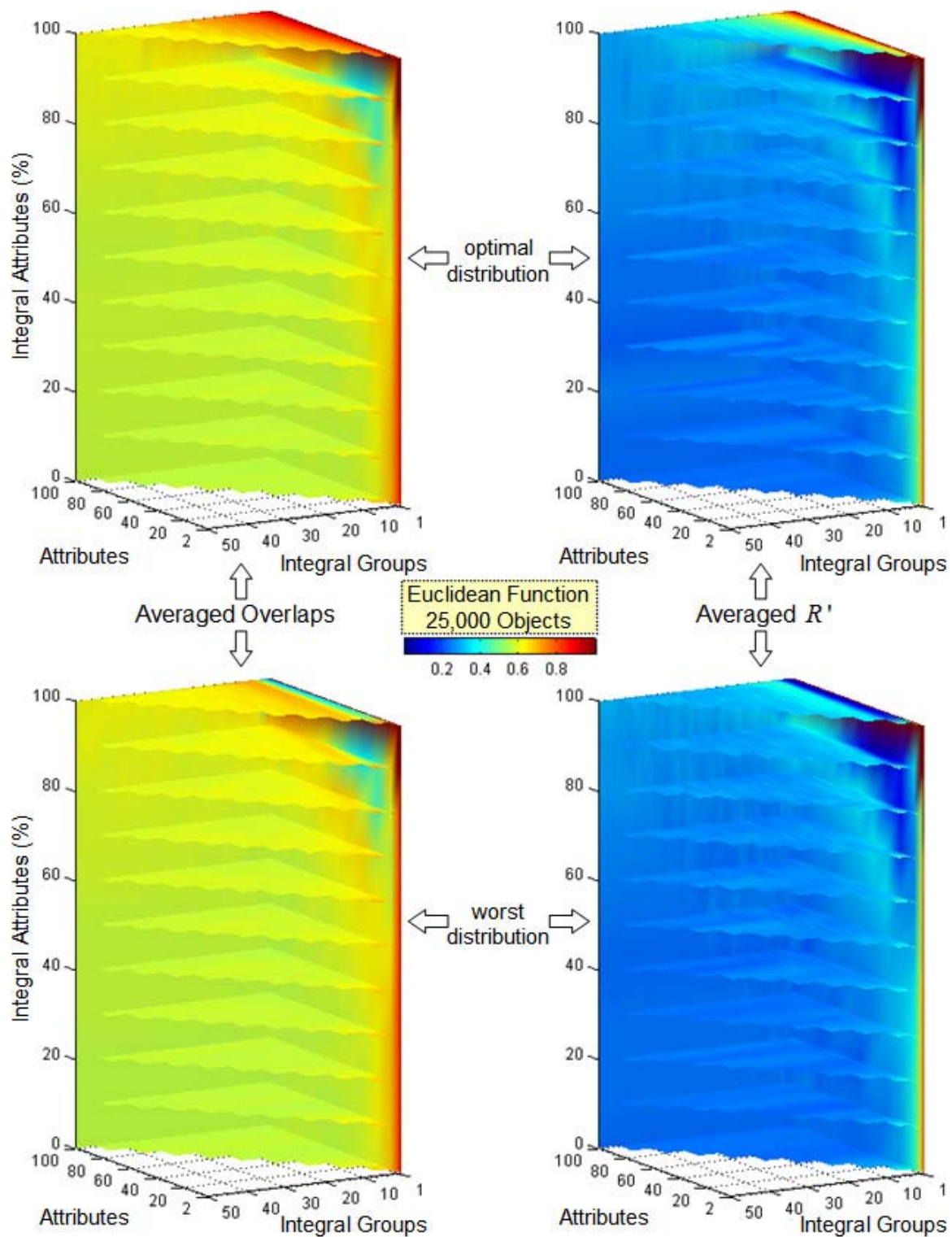


Figure 6.20: Experiment E<sub>2B</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects.

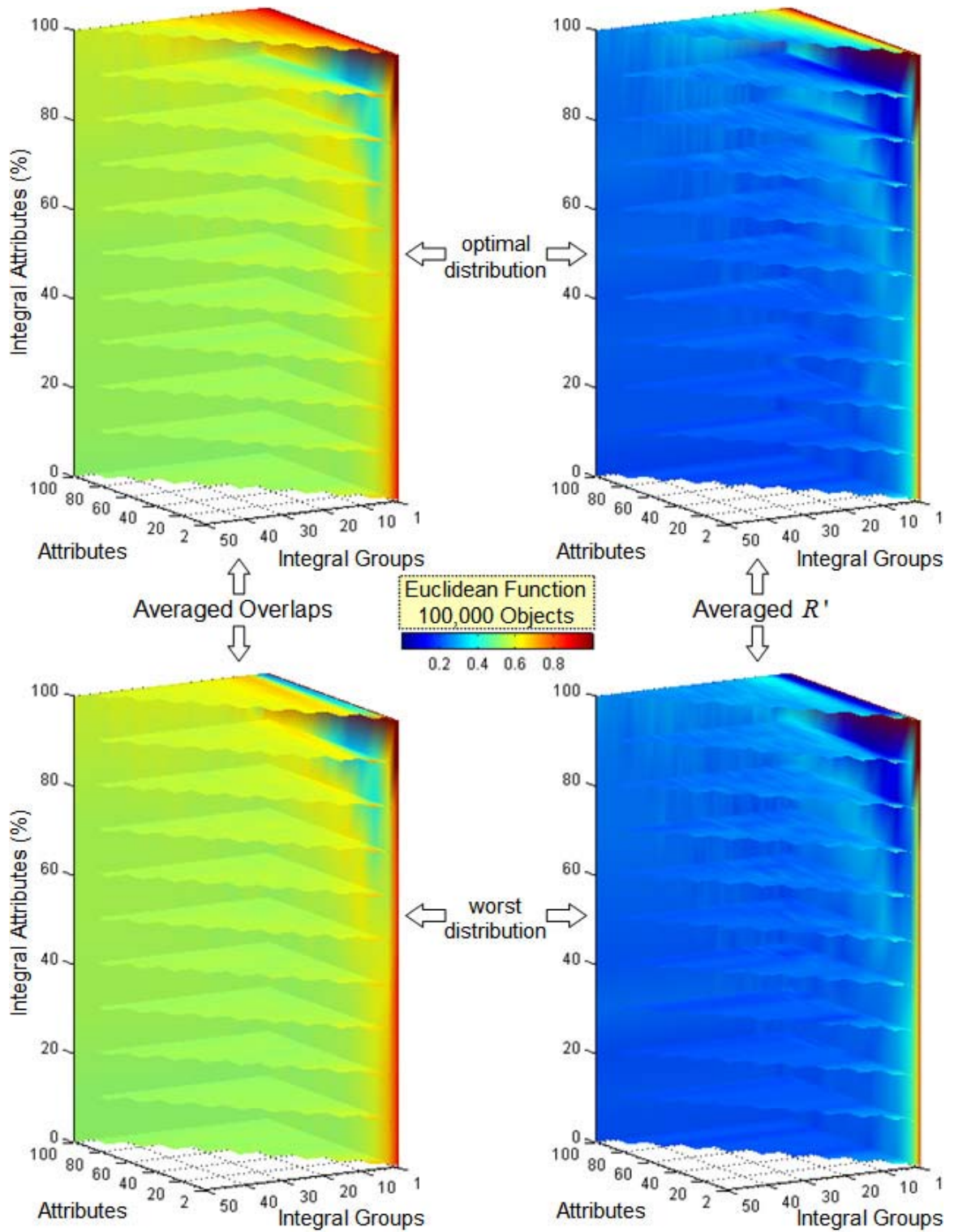


Figure 6.21: Experiment  $E_{2B}$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects.

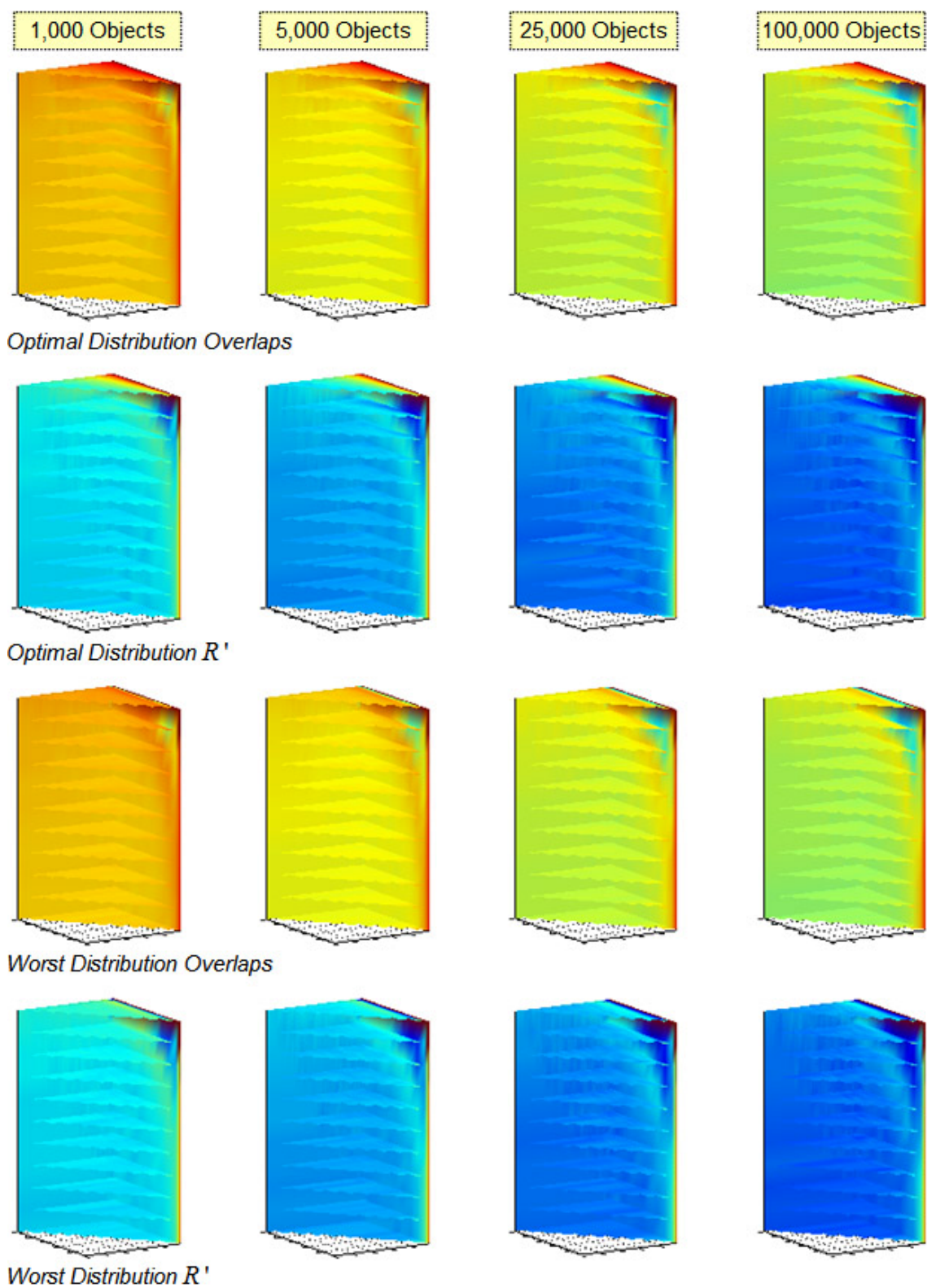


Figure 6.22: Overview of the results acquired from Experiment  $E_{2B}$ .

The results of Experiment  $E_2$  indicate again a gradual pattern of variation, although the pattern is considerably more subtle than that of Experiment  $E_1$ . The variation is more obvious in the diagrams of  $R'$ . The Manhattan aggregation function is identical to the compliant aggregation function with respect to the treatment of the separable attributes, whereas the Euclidean aggregation function is identical with respect to the treatment of the integral attributes. Hence, the two functions approach the compliant method from converse directions, an observation that explains many of the reverse trends that they demonstrate.

The more dominant reverse trend is evident along the  $Z$ -axis and pertains to the number of integral attributes. As this variable assumes higher values, the compliant method becomes progressively similar to the Euclidean metric; therefore, the results produced with the Euclidean function are worst at the lowest layer where no integral attributes exist, while they improve gradually for higher values of  $p$ . Conversely, the results produced with the Manhattan function are best at the lowest layer and deteriorate thereafter. The culmination of this trend occurs at the highest layer where no separable attributes remain. At the highest layer, the Manhattan function scores better with a worst distribution policy, whereas the Euclidean function yields more compatible results with an optimal distribution policy.

The two competitors also demonstrate a different behavior with respect to the number of integral groups. The Euclidean metric seems to be invariant to changes of this variable, whereas the Manhattan metric offers better results for fewer groups. The root of this phenomenon is that in the compliant approach the integral groups are aggregated with the Euclidean metric; therefore, more errors propagate to the final similarity score with the Manhattan function as the number of groups increases. The Euclidean metric, on the other hand, remains naturally unaffected.

Several edge effects appear in the diagrams. They take place for extreme values of the variables, for which the tested functions coincide with the compliant approach, or exhibit

the maximum deviation from it. Such db scenarios occur, for instance, in the case of only two attributes, where the overlaps and correlations have high values. They also occur in the case of one integral group where the Euclidean and the compliant functions coincide. For both functions, the overlaps and correlations deteriorate as the database size increases. The justification for this trend is the same as that given for the first experiment (i.e., increasing the database size while leaving the size of the relevant portion constant).

Experiment  $E_2$  gives unequivocal evidence that for the overwhelming majority of db scenarios the Manhattan function provides drastically better results than its Euclidean counterpart. The overlaps remain consistently high, occasionally reaching the maximum value of 1. The correlations also score highly, although to a somewhat lesser degree than the overlaps. These measurements imply not only that the results in the relevant portion are the same as those of the compliant approach, but also that they follow approximately the same order; therefore, the hypothesis statement  $HS_2$  about the aggregation function should be rejected for the Manhattan case. For the Euclidean case, the validity of this hypothesis is undecisive, since it could be accepted for larger datasets and rejected for smaller datasets. The interpretation of the hypothesis  $HS_2$  for the Euclidean function, however, becomes rather indifferent, as the Manhattan function can serve as a surrogate aggregator of higher fidelity to the compliant method. This is a welcome outcome, because the Manhattan metric, is simpler and usually more efficient than the compliant and the Euclidean aggregation functions.

#### *6.3.2.3 Results of Experiment $E_3$ and Interpretation*

The next diagrams (Figures 6.23 to 6.27) show the results of Experiment  $E_3$ . The diagrams depict the combined distortion on the desirable set of results when a deviant aggregation function is used and the groups of integral attributes are not identified.



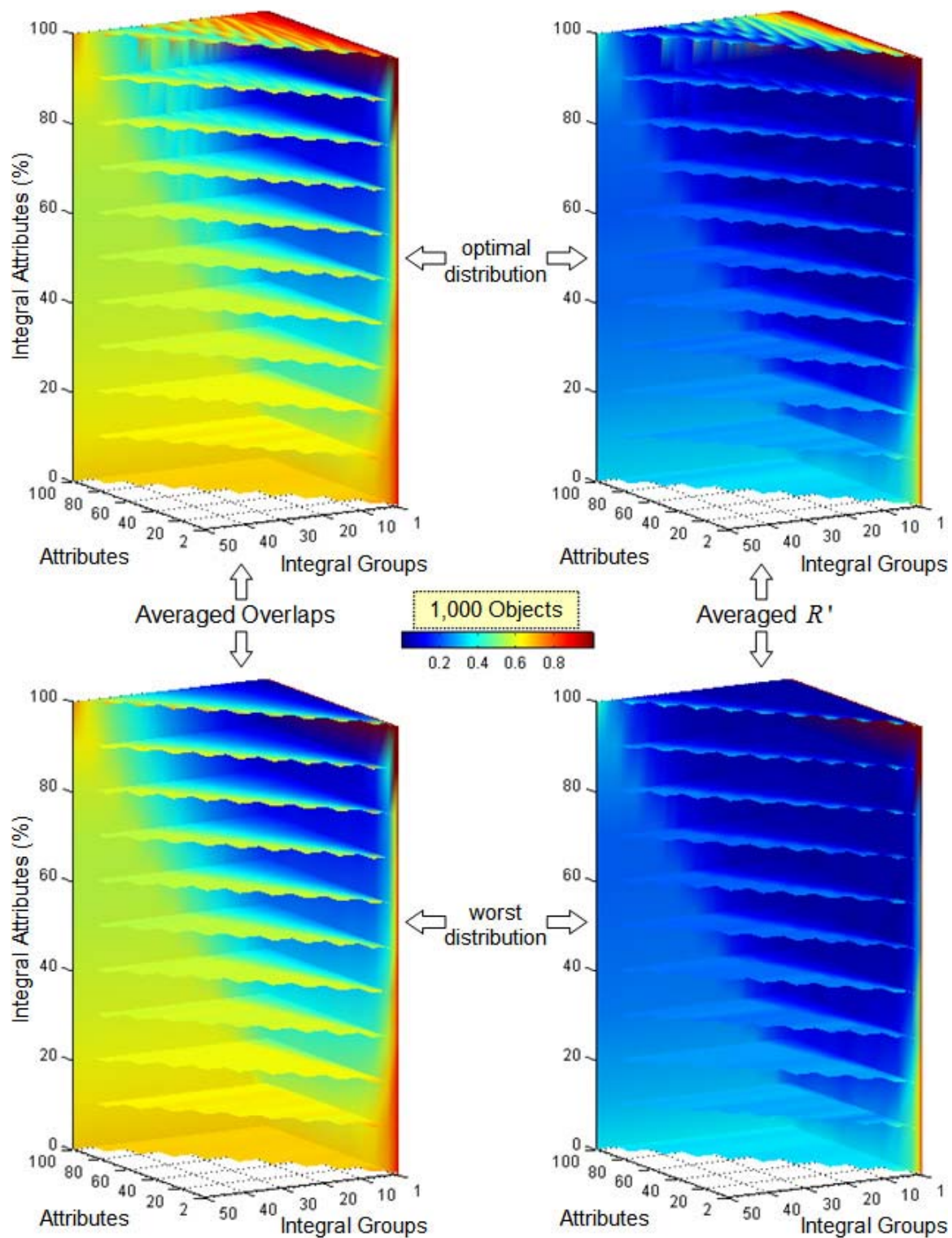


Figure 6.23: Experiment  $E_3$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 1,000 objects.

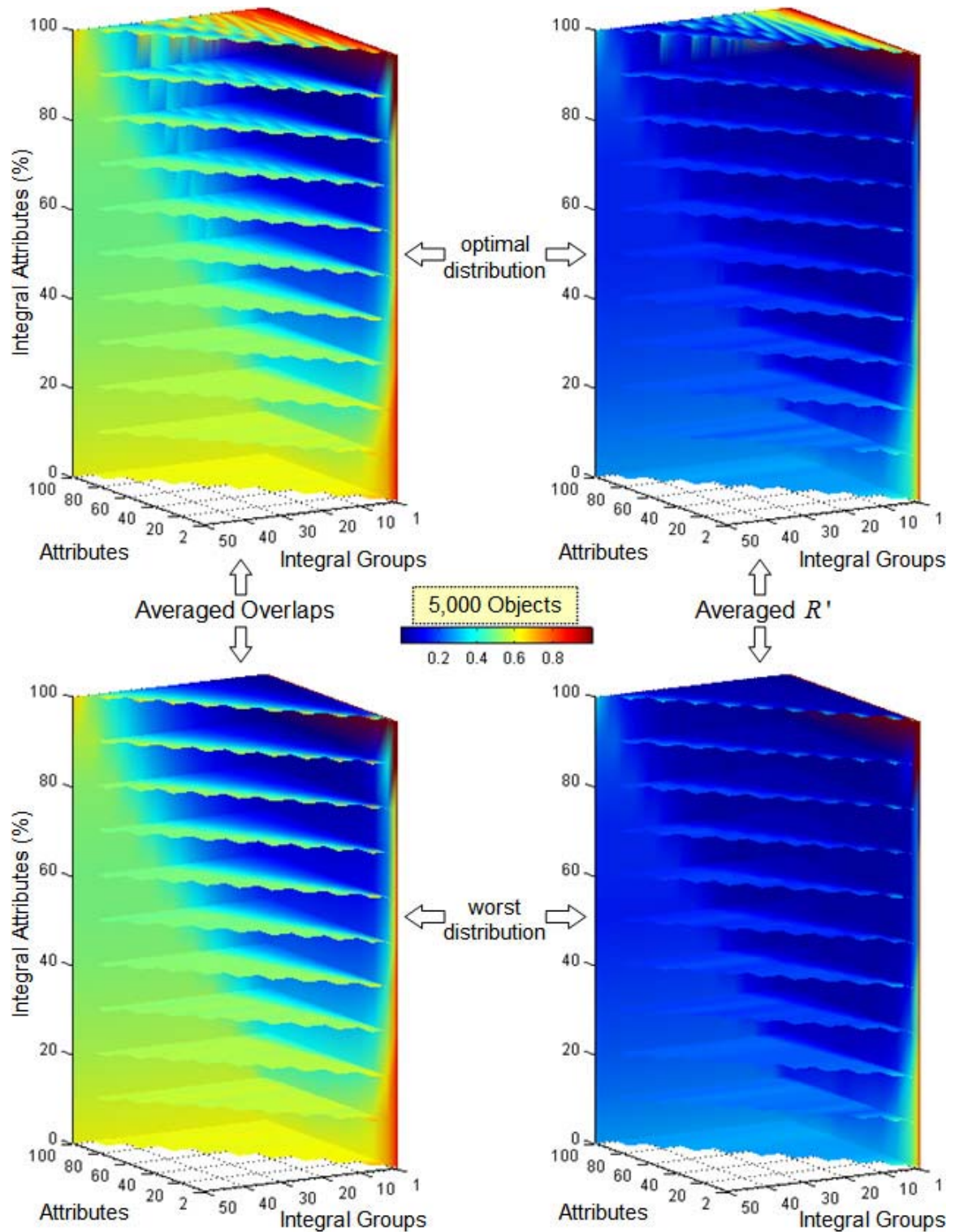


Figure 6.24: Experiment E<sub>3</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 5,000 objects.

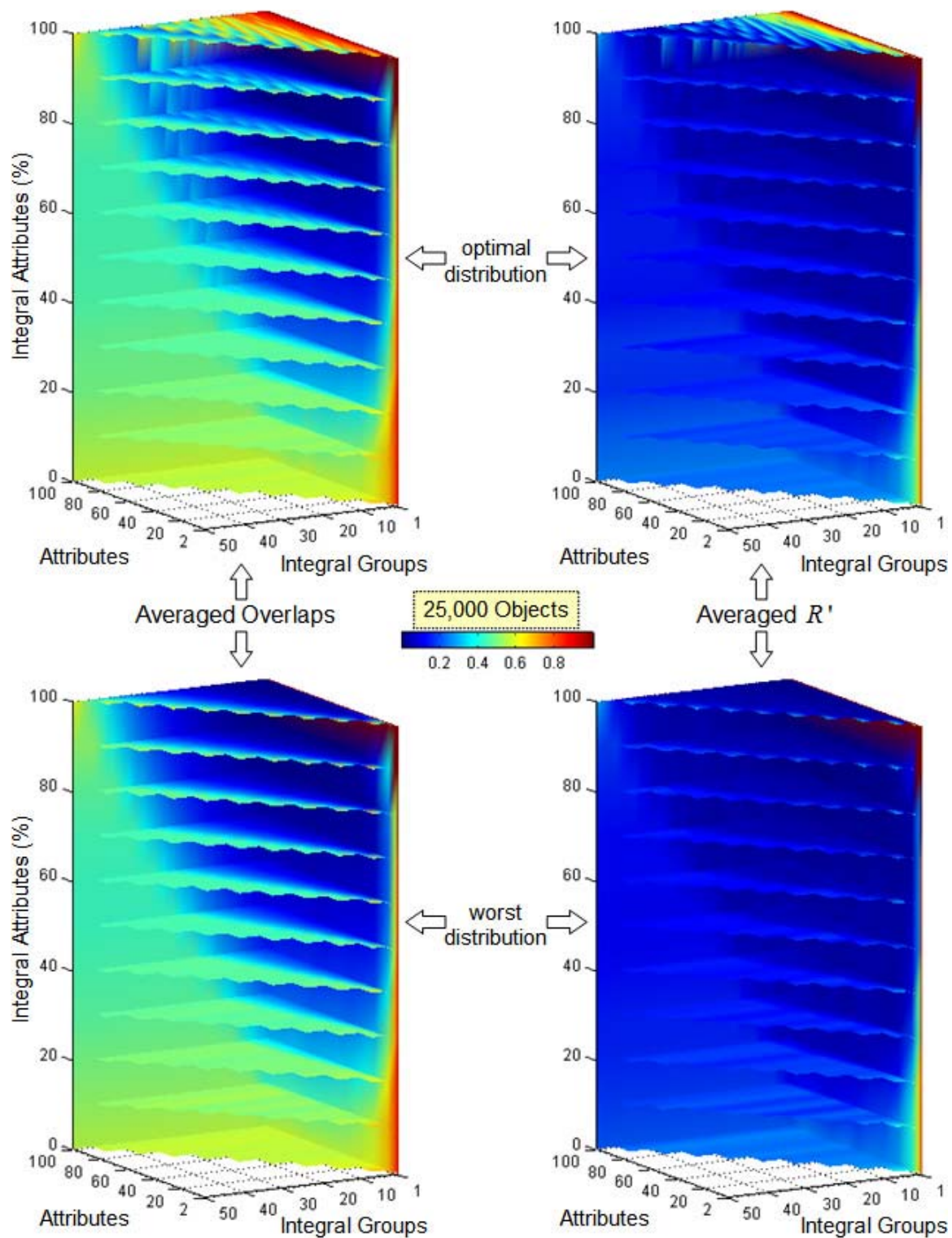


Figure 6.25: Experiment E<sub>3</sub>: averaged overlaps and correlations between the compliant and the deviant method for a database of 25,000 objects.



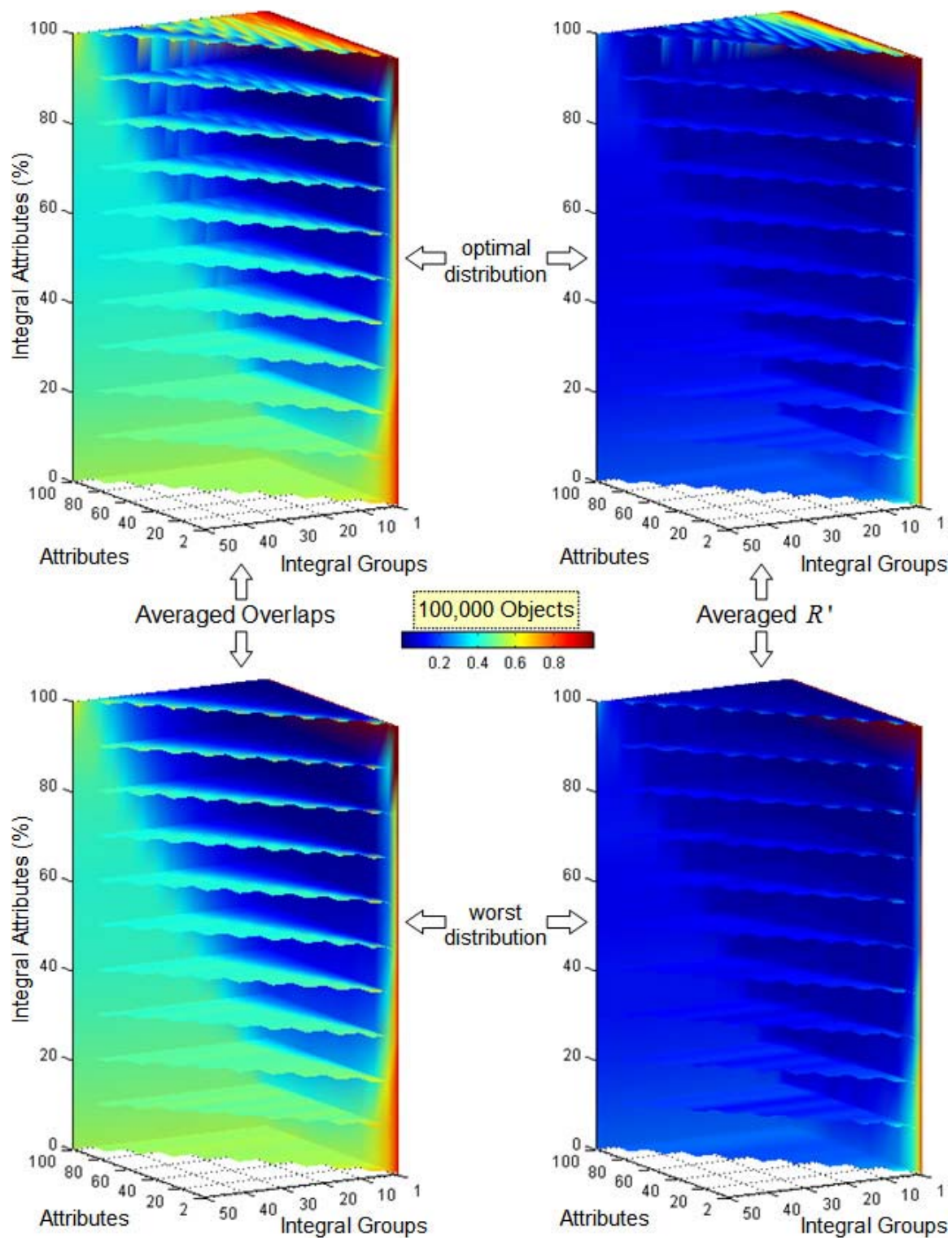


Figure 6.26: Experiment  $E_3$ : averaged overlaps and correlations between the compliant and the deviant method for a database of 100,000 objects.

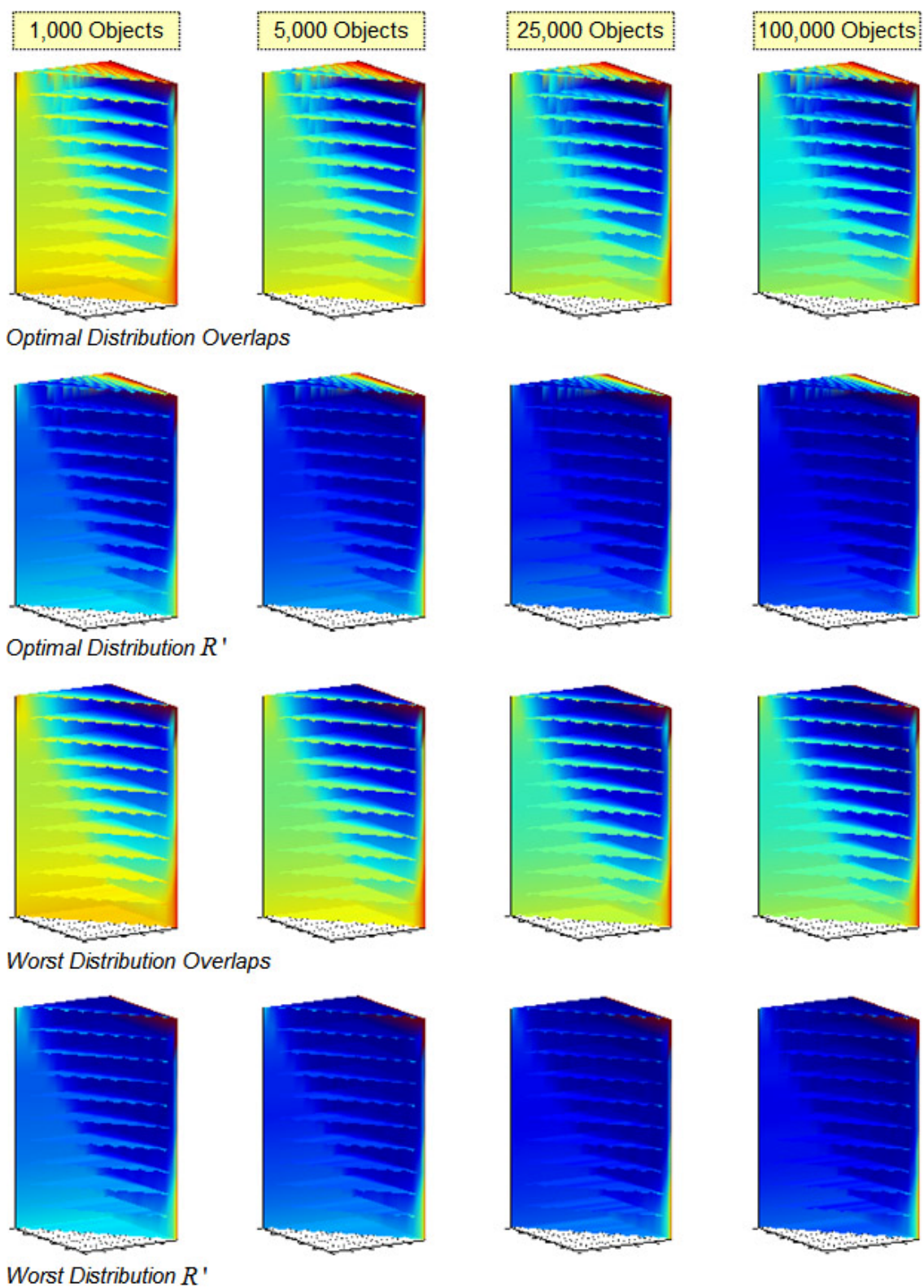


Figure 6.27: Overview of the results acquired from Experiment  $E_3$ .

As anticipated, the diagrams of Figures 6.23 to 6.27 interweave the diagrams of Experiment  $E_1$ , which illustrate the outcome when integral attributes are not recognized, and those of Experiment  $E_2$ , in which the Euclidean aggregation function is used to combine separable attributes. The overlaps and correlations increase only at the rightmost edges where the tried method converges to the compliant approach. For all other db scenarios the measures do not have a significant concordance; therefore, when both premises  $HS_1$  and  $HS_2$  are violated, the distortions in the desirable set of results are unacceptable. Experiment  $E_3$  confirms, therefore, the conclusions about the hypothesis statements  $HS_1$  and  $HS_2$  that were formulated in the commentary of Experiments  $E_1$  and  $E_2$ , respectively.

## 6.4 Experiments at the Scene Level

This section describes the setup and discusses the results obtained from Experiment  $E_4$ . This experiment relies on the prior existence of an association graph created in response to a scene query, where the aggregate dissimilarities at each node and edge have already been computed. The dissimilarity value of each element of the cliques is converted into a similarity value with each of the compared functions, and the final similarity score is then computed for the entire solution (Equation 5.5). All objects and relations are equally weighted. Furthermore, the object and relational components of each clique have an equal contribution to the similarity of each solution (Equation 5.4).

### 6.4.1 Setup

The incompatibility measures  $O$  and  $R'$  at the scene level are a function of three variables  $q$ ,  $t$ , and  $c$  (Equation 6.3):

$$O, R' = f(q, t, c) \quad (6.3)$$

- Variable  $q$  is the number of objects in the query scene, determining the query size.

The experiment was conducted for the set  $Q = \{2, 3, 4, 5, 7, 10, 20, 30, 40, 50\}$ .

Values between 2 and 10 were sampled more frequently and are considered of higher importance, because typical user-sketched queries contain a small number of objects (Blaser 2000). Larger query sizes of up to 50 objects are possible in cases of selection queries in collection databases, where users do not sketch or define the objects themselves, but rather select an existing scene that they use as the query. Single-object queries are omitted, because all functions rank the results identically in this case. The variable  $q$  also determines indirectly the number of relations present in the query (i.e.,  $q \cdot (q-1)/2$ ). This term, summed with  $q$ , gives the total number of elements in a scene query. The smallest and largest queries considered have, therefore, 3 and 1,275 elements, respectively.

- Variable  $t$  is the threshold used in the matching process during the creation of the association graph. This variable models the degree of the query's constrainedness. Database objects and relations are matched with those of the query only if their computed dissimilarity scores do not exceed the threshold (Figure 6.4). A threshold specification thus segments the functions and delimits their response within a particular subsection of their curves (Figure 6.5). The set of thresholds considered in this experiment is  $T = \{0.02, 0.2, 0.4, 0.6, 0.8, 1\}$ . The first element of this set was taken slightly above 0 to avoid trivializing the outcome of the experiment. If the lowest value had been set to 0, all results would have been exact matches, thus receiving a similarity of 1, which would render the ranking and comparison processes of the lists meaningless. Creating the association graph with a dissimilarity threshold of 1 is also poor practice, because such a specification implies no pruning of the search space and entails the retrieval of a huge number of solutions for large databases. In the controlled environment of the experiment, however, the maximum number of solutions was delimited to some maximum number, because the interest instead is in evaluating how the conversion functions react to severely under-constrained queries.

- Variable  $c$  is the number of cliques extracted from the association graph, determining how many solutions will be ranked after the similarities of their elements (i.e., objects and relations) have been computed with each of the compared functions. The set examined is  $C = \{10, 20, 50, 100, 200, 500, 1000, 2000, 3000, 4000, 5000\}$ . This variable depends on the underlying database size, because more solutions are anticipated from larger databases. It also depends on variables  $t$  and  $q$ , because the number of retrieved solutions is expected to increase with less constrained queries or queries that involve fewer objects.

A specific instantiation of the variables  $q$ ,  $c$ , and  $t$  is referred to as a *query scenario*. Experiment  $E_4$  uses the same visualization technique as that described for Experiments  $E_1$ - $E_3$ , with colored 3-dimensional diagrams sliced along the  $Z$ -axis. The axes  $X$ ,  $Y$ , and  $Z$  of the diagrams correspond to the number of objects in the query  $q$ , the number of solutions  $c$ , and the threshold value  $t$ , respectively. The different value combinations of the triple  $(q, c, t)$  create a grid in the cubic space, where each point represents a particular query scenario. The only difference with the diagrams of the previous experiments is that the entire cubic space is utilized this time, as all of the query scenarios in it are—at least theoretically—possible (Figure 6.28). The seven conversion functions that are considered amount to 21 pairwise comparisons. Since there are two incompatibility measures for each pair, a total of 42 diagrams was produced for this experiment.

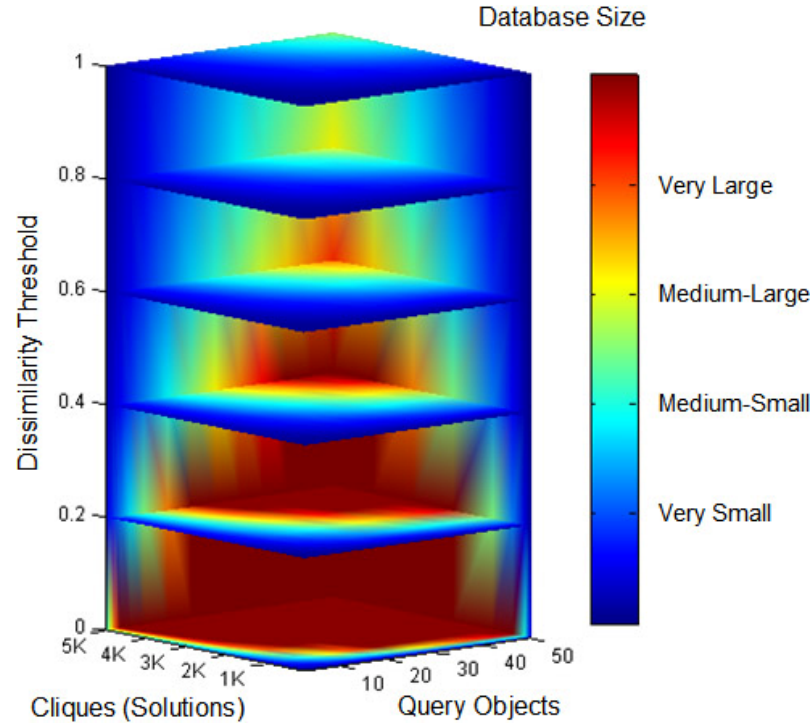


Figure 6.28: A sample diagram from Experiment E<sub>4</sub>, giving a rough estimate of the database size required to accommodate the tested query scenarios.

#### 6.4.2 Results of Experiment E<sub>4</sub> and Interpretation

The next figures show the agreement in the results of different pairs of conversion functions. The first set of diagrams (Figures 6.29-35) concentrates on comparisons of the linear function with the non-linear alternatives. The pair of functions  $(E_L, G_L)$  is representative of non-linear curves that are relatively close to the linear slope, a proximity, which psychological research suggests should hold in most situations (Section 6.2). Particular emphasis for the assessment of the third part of the hypothesis is, therefore, attributed to the interpretation of Figures 6.29-31, which depict how  $E_L$  and  $G_L$  compare to  $L$ , and to each other.



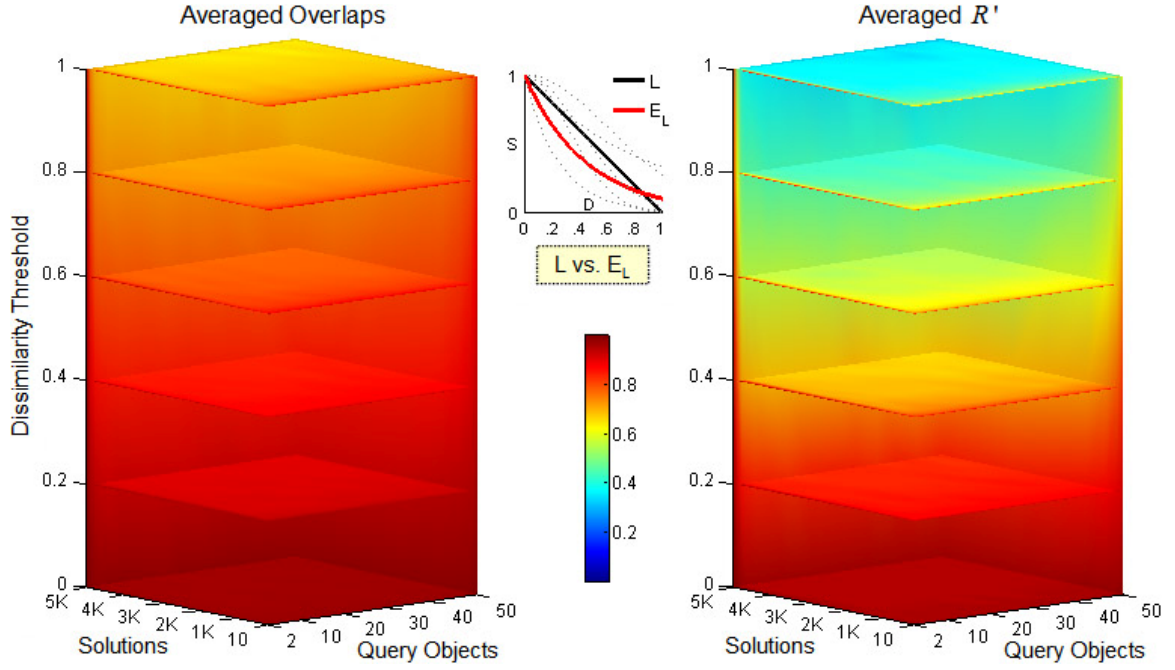


Figure 6.29: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $L$  and  $E_L$ .

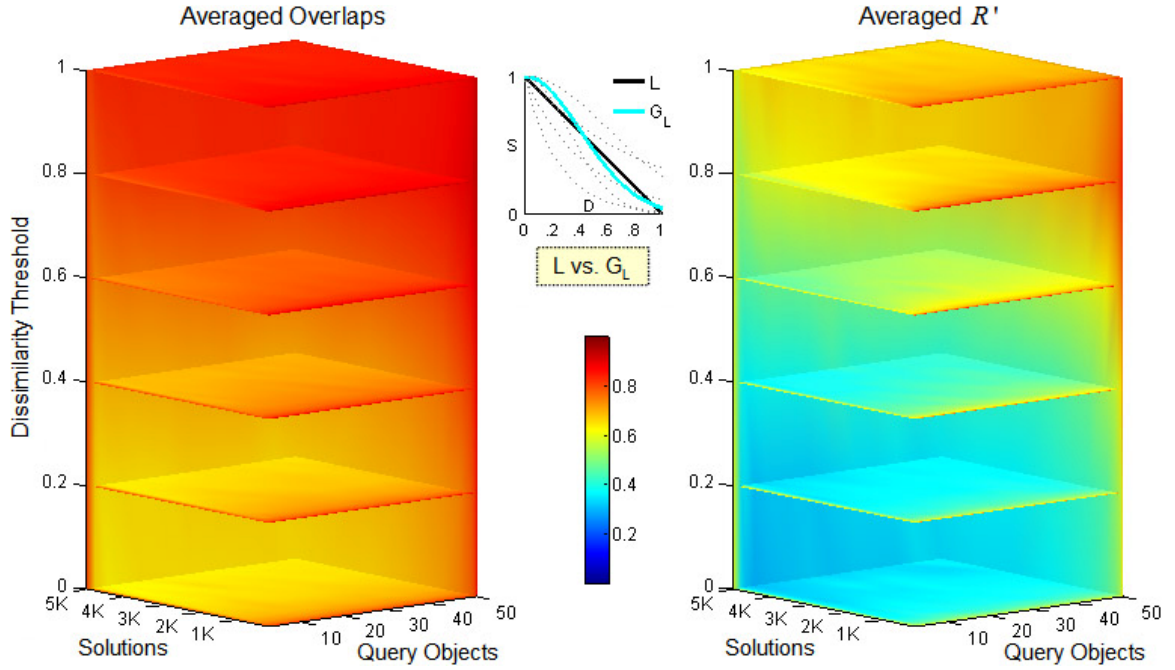


Figure 6.30: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $L$  and  $G_L$ .

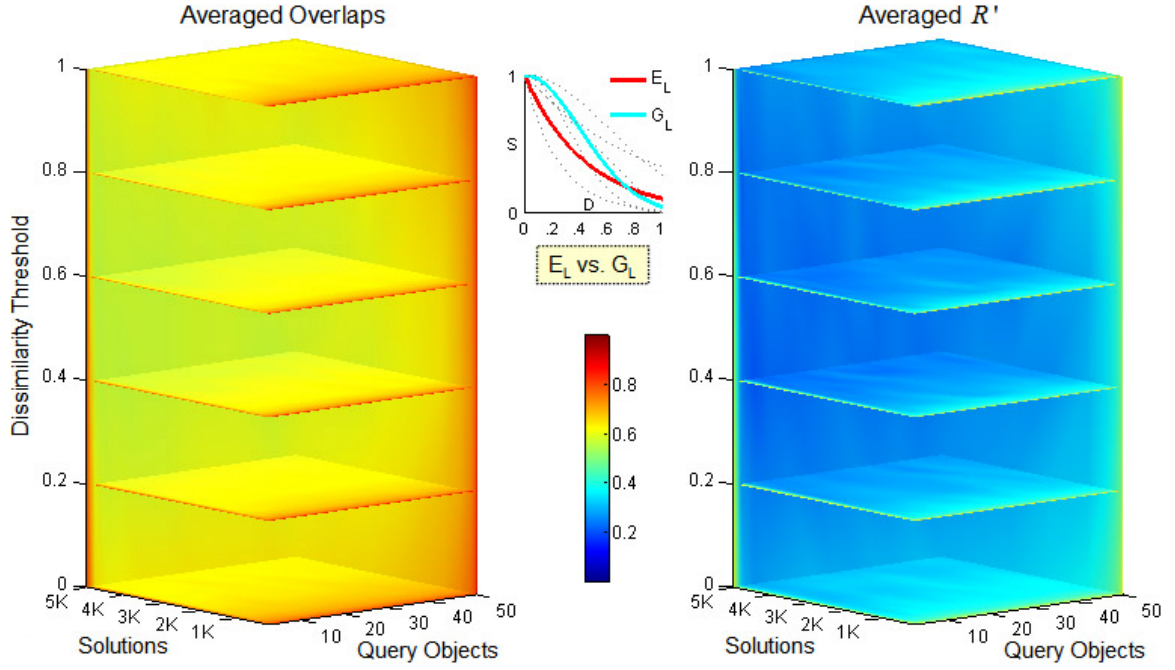


Figure 6.31: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_L$  and  $G_L$ .

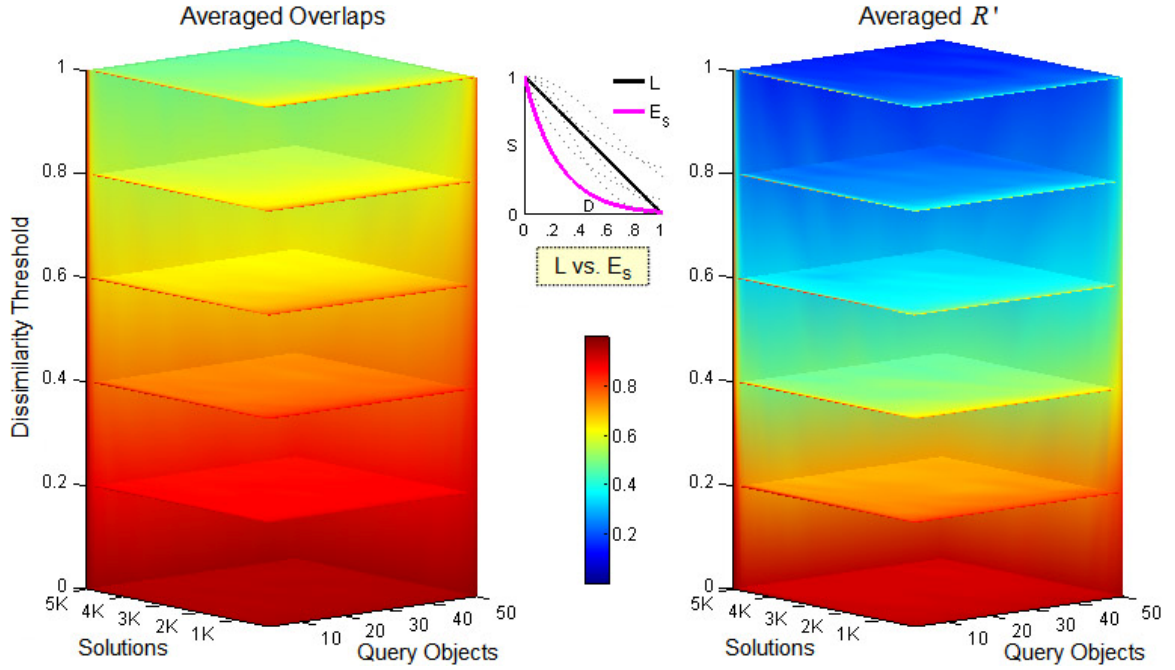


Figure 6.32: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $L$  and  $E_S$ .



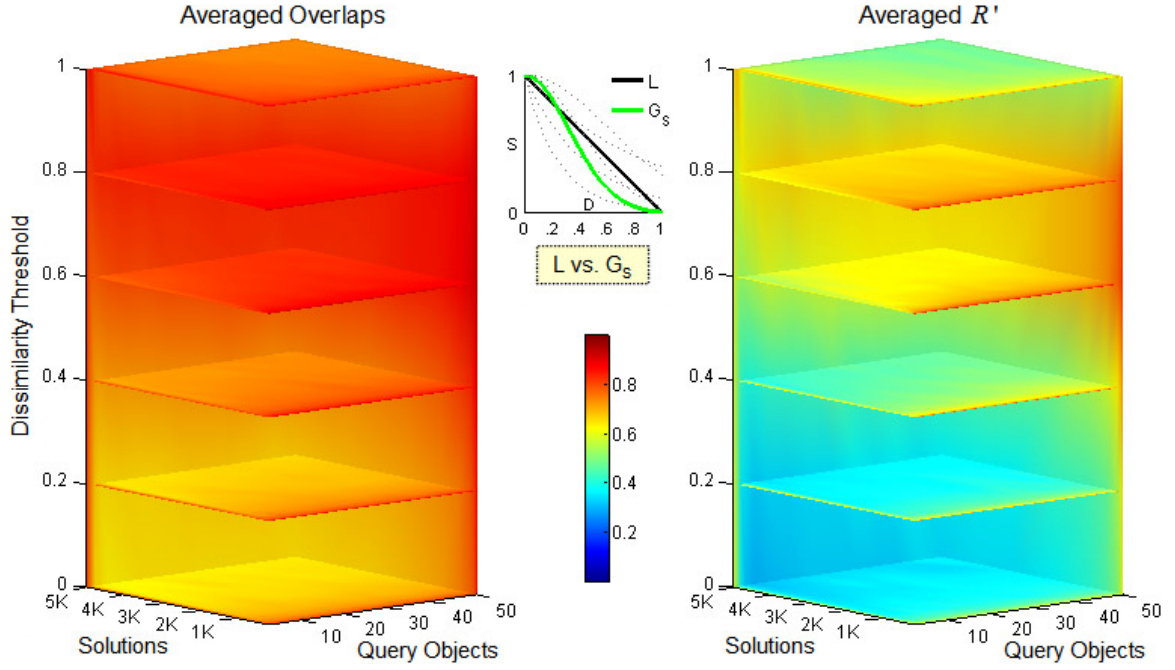


Figure 6.33: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $L$  and  $G_S$ .

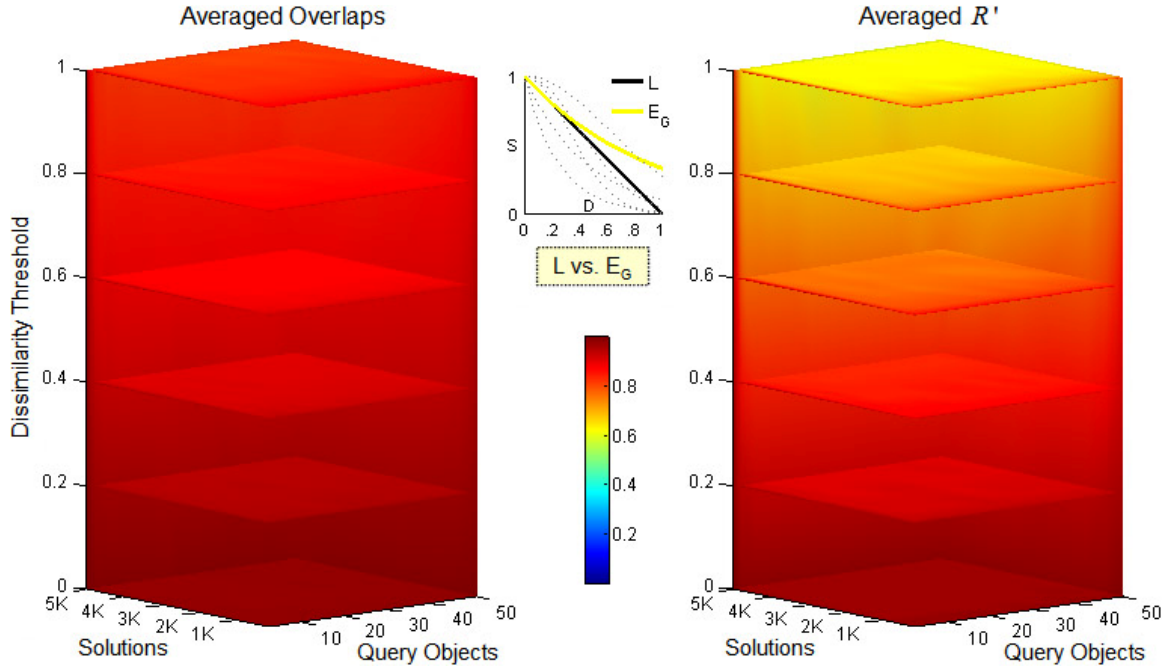


Figure 6.34: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $L$  and  $E_G$ .

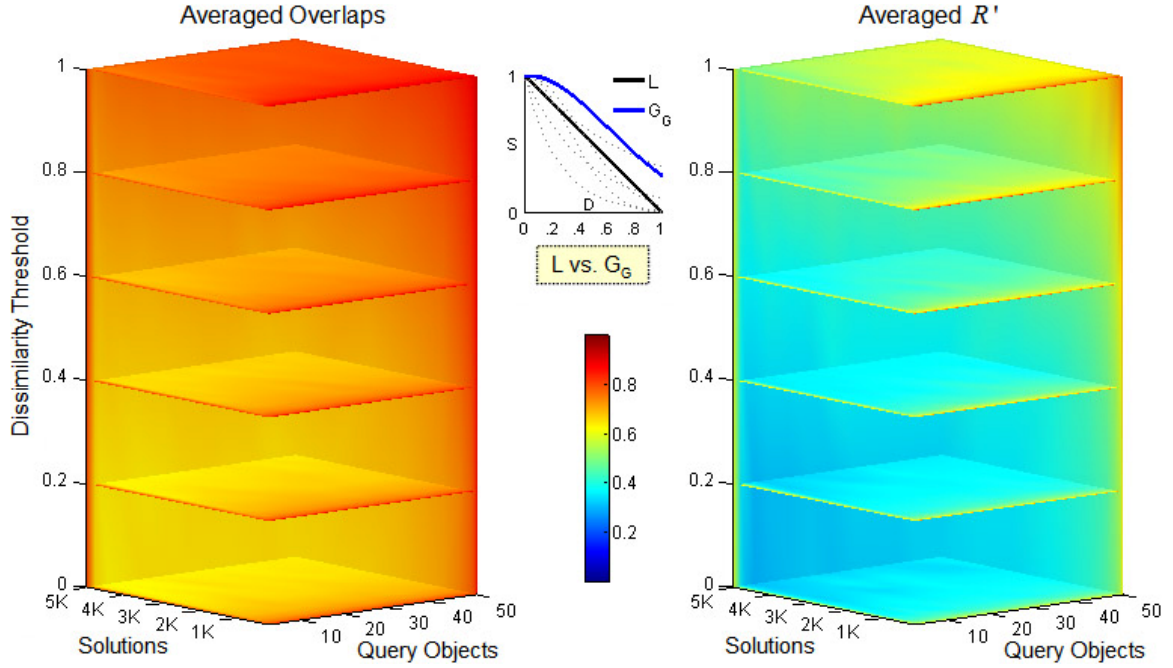


Figure 6.35: Experiment E<sub>4</sub>: averaged overlaps and correlations between the functions  $L$  and  $G_G$ .

The diagrams indicate that the non-linear functions produce very similar results to the linear function, but less similar results to one another. The congruence between  $L$  and the exponential functions is very high for low dissimilarity thresholds (i.e., over-constrained queries) and deteriorates slightly for higher dissimilarity thresholds (i.e., under-constrained queries). The interaction between  $L$  and the Gaussian functions is exactly the opposite, with the concordance of the results being less at the lowest layers and increasing for higher layers. A plausible interpretation for this reverse trend is that it is due to the different shapes of the curves. When the shape is convex, the results improve for higher threshold values. When it is concave, the results are relatively stable and independent of the threshold value. This speculation is substantiated from the diagrams in Figures 6.33 and 6.35. In Figure 6.33, the measures  $O$  and  $R'$  start improving at the tipping point where the shape of the Gaussian function changes from concave to convex. This improvement climaxes at the threshold value of 0.8. At this point, the measures start deteriorating again, following the same trend as that exhibited by the exponential

functions. In Figure 6.35, where the tipping point of the generous Gaussian function is shifted much further, the results are almost the same at the first three layers and start improving slowly thereafter.

As anticipated, the results are also affected by the distance between the graphical representations of the functions. The further apart two plots are, the worse the acquired measures become. For example, the strict exponential function  $E_S$  (Figure 6.32) gives worse results when compared to the linear, than the closer exponential function  $E_L$  does (Figure 6.29). The effect of the distance appears to be less significant than that of the shape. For example, even though the curve of  $G_L$  is closer to the straight line for threshold values below 0.4, (Figure 6.30) its concave shape produces worse results compared to the convex form of the more distant  $E_L$  (Figure 6.29). For larger threshold values, however, both curves become convex and  $G_L$  correlates better due to its smaller distance from the linear function.

Excluding the beginnings of the  $X$  and  $Y$  axes, the results stabilize shortly thereafter and remain invariant to the variables  $q$  and  $c$ , which correspond to query size and the number of cliques extracted from the association graph, respectively. Smaller-sized queries (i.e.,  $X$ -axis) have a positive effect on the results, which becomes evident at the front-left fringes of the layers and the left edge of the diagram along the  $Z$ -axis. A slightly more pronounced improvement is also observed when the number of cliques extracted from the association graph is relatively small (i.e., below 300). This improvement manifests at the red-colored front-right fringes of the layers and at the right edge of the diagram along the  $Z$ -axis. In general, for typical user queries that are reasonably constrained and contain a few objects only, the agreement between the measures is high.

The choice of a stricter or a more generous conversion function does not alter the results radically. The overlap measure, which was deemed of primary importance, maintains high values for the overwhelming majority of query scenarios. In some cases

(e.g., Figure 6.34) the results are practically identical throughout the cubic space. The worst deviations are observed in the performance of the linear function versus the strict exponential for severely under-constrained queries (Figure 6.32, highest layer). Even there, the overlaps are high around the edges, which correspond to more typical retrieval scenarios. In all other cases, the overlaps consistently exceed the value of 0.6. The diagrams strongly suggest, therefore, that the third hypothesis statement  $HS_3$  should be rejected.

The next set of diagrams (Figures 6.36-41) reveals how different non-linear functions of the same family compare to one another.

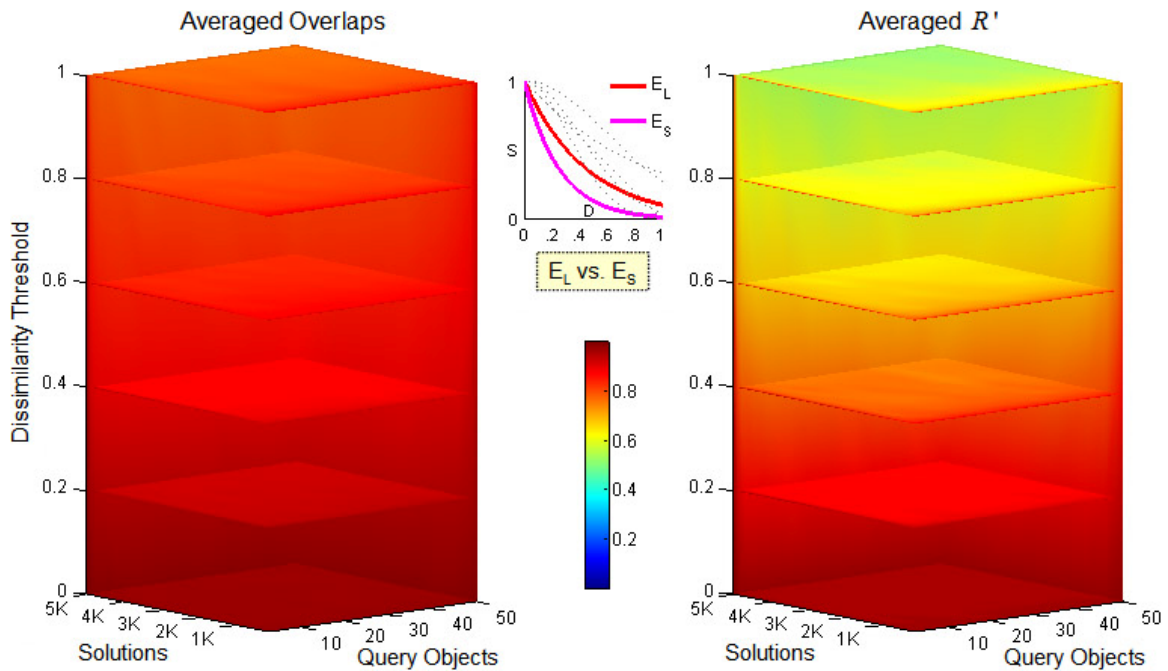


Figure 6.36: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_L$  and  $E_S$ .

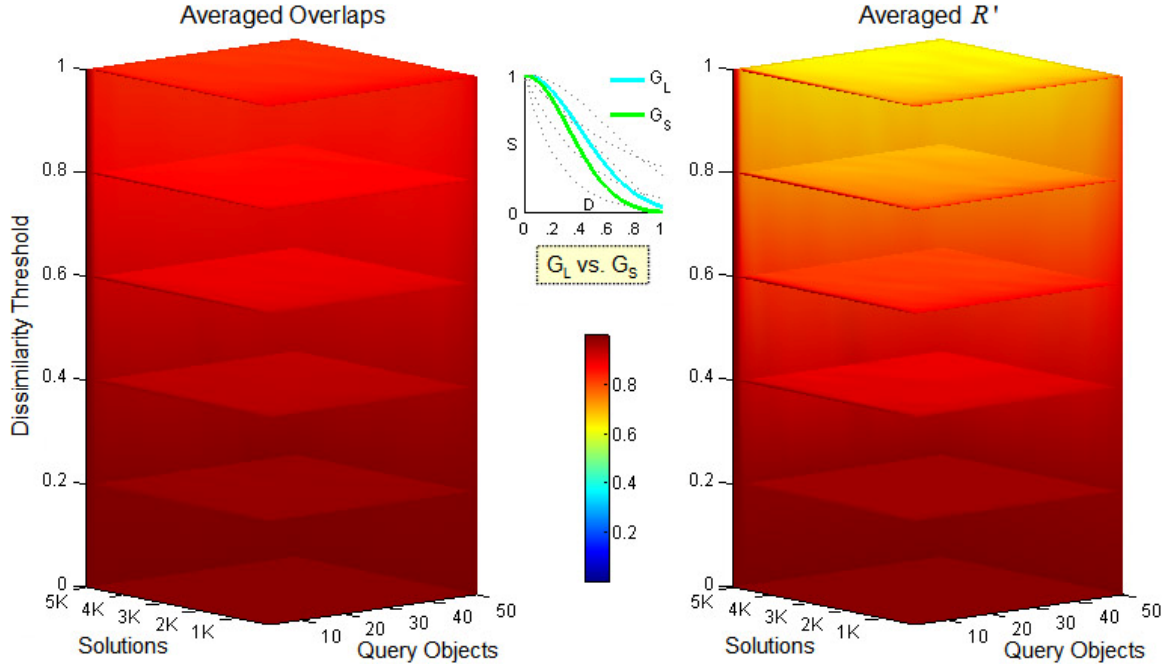


Figure 6.37: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $G_L$  and  $G_S$ .

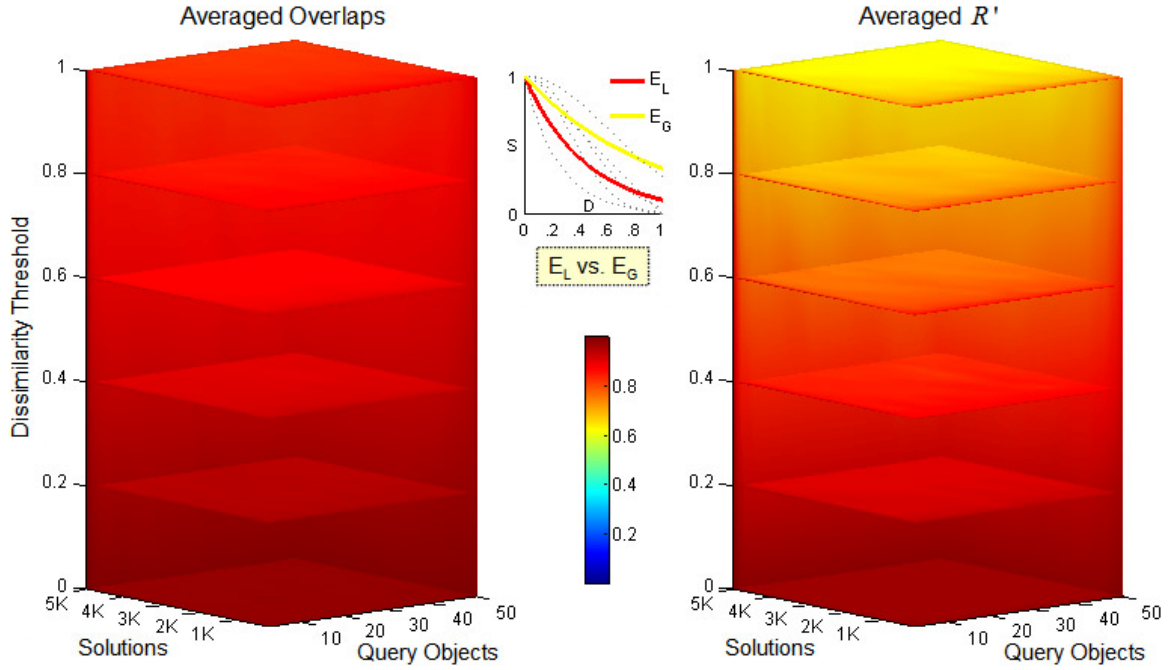


Figure 6.38: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_L$  and  $E_G$ .

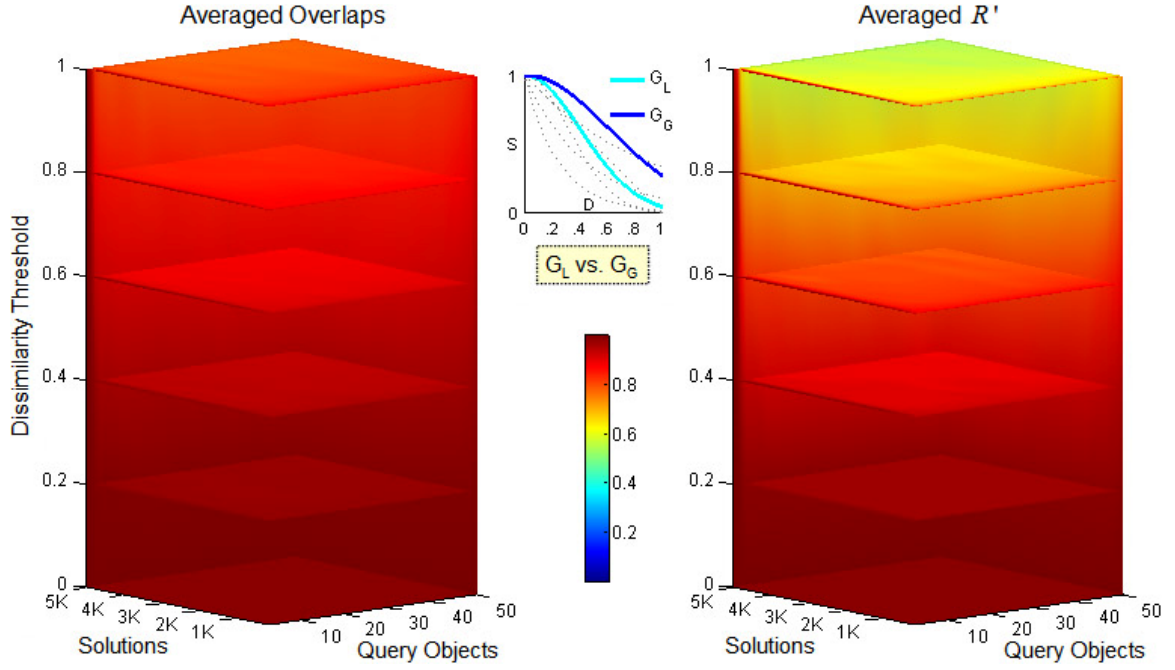


Figure 6.39: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $G_L$  and  $G_G$ .

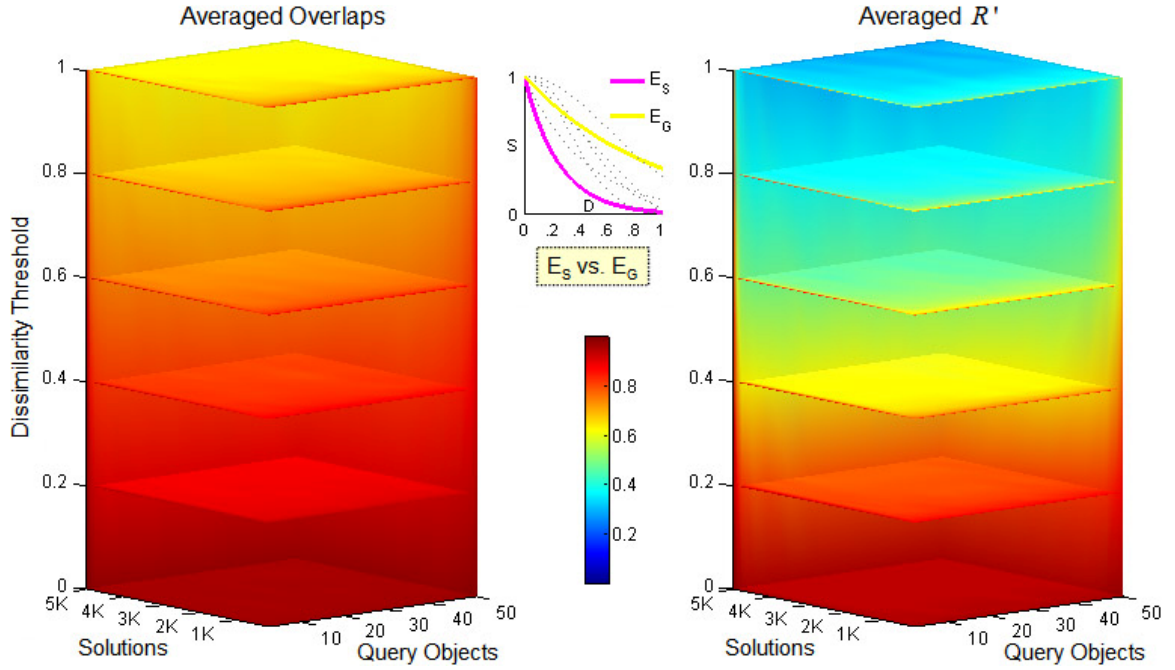


Figure 6.40 Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_S$  and  $E_G$ .



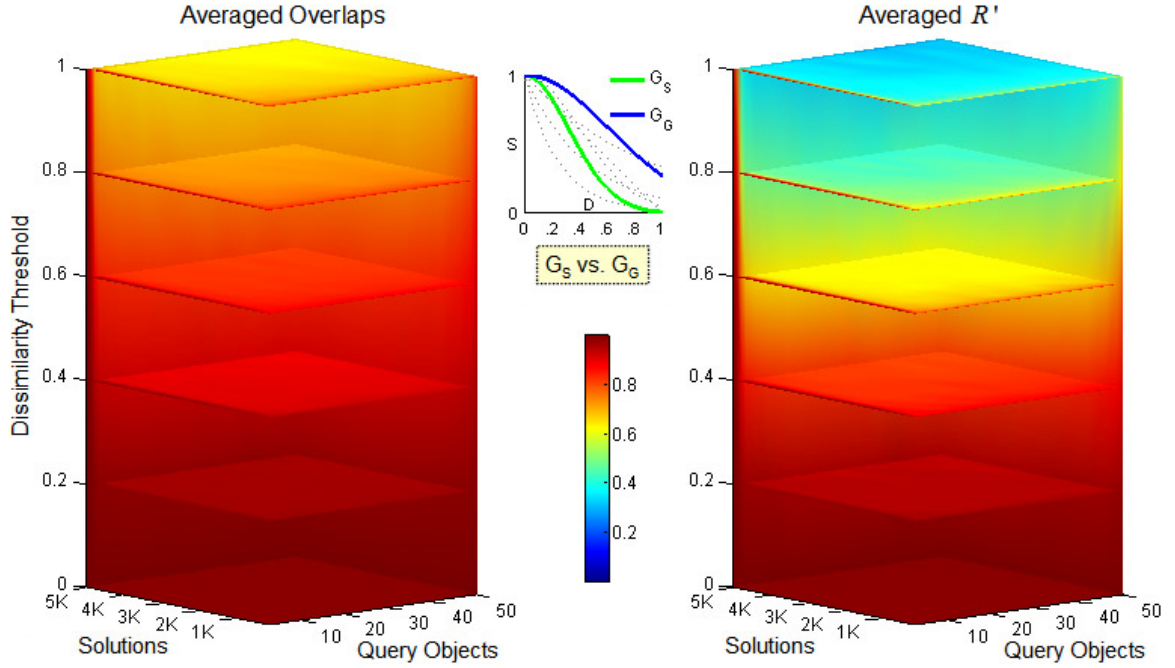


Figure 6.41: Experiment E<sub>4</sub>: averaged overlaps and correlations between the functions  $G_S$  and  $G_G$ .

These diagrams provide further evidence for the rejection of hypothesis statement HS<sub>3</sub>. The overlaps are very high and close to 1 in all query scenarios. The correlations also remain reasonably high for the most part. The diagrams corroborate the initial speculation about the dominant effect of the shape of the functions on the results. Functions with a relatively large distance among their curves (e.g., Figures 6.40-41) still yield results of high concordance as long as their shapes are similar. The variation of the measures along the three axes and the edge and fringe effects are the same as those detected in the previous set of diagrams (i.e., Figures 6.29-6.35).

The last set of diagrams (Figures 6.42-49) demonstrates the relative performance between non-linear functions of different families.

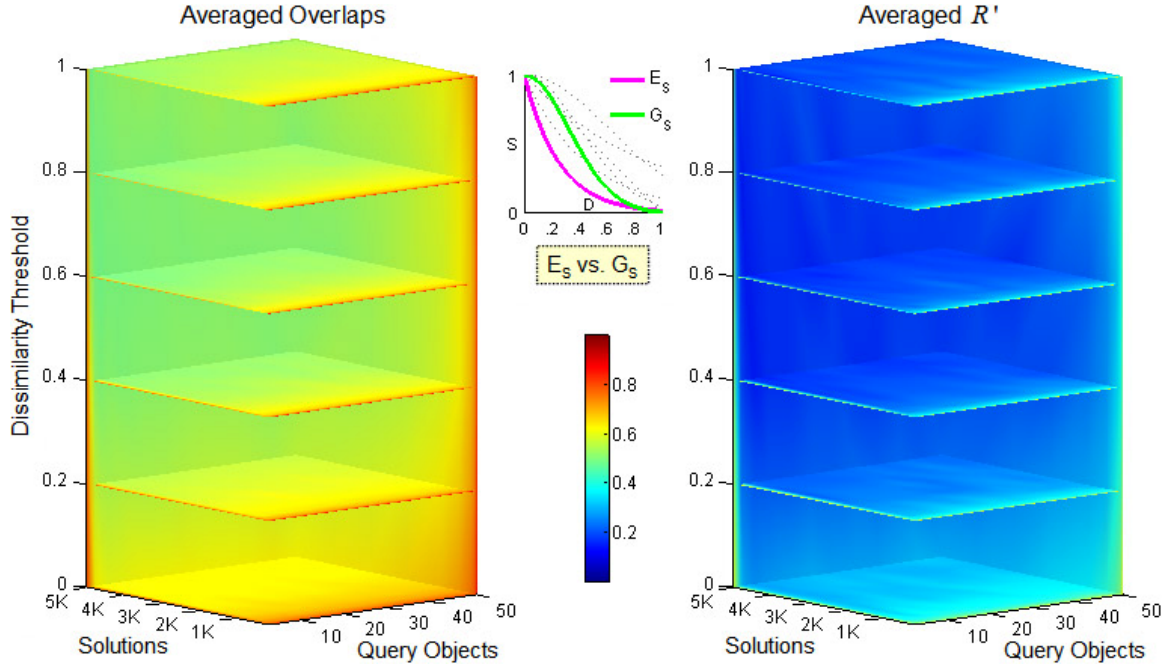


Figure 6.42: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_S$  and  $G_S$ .

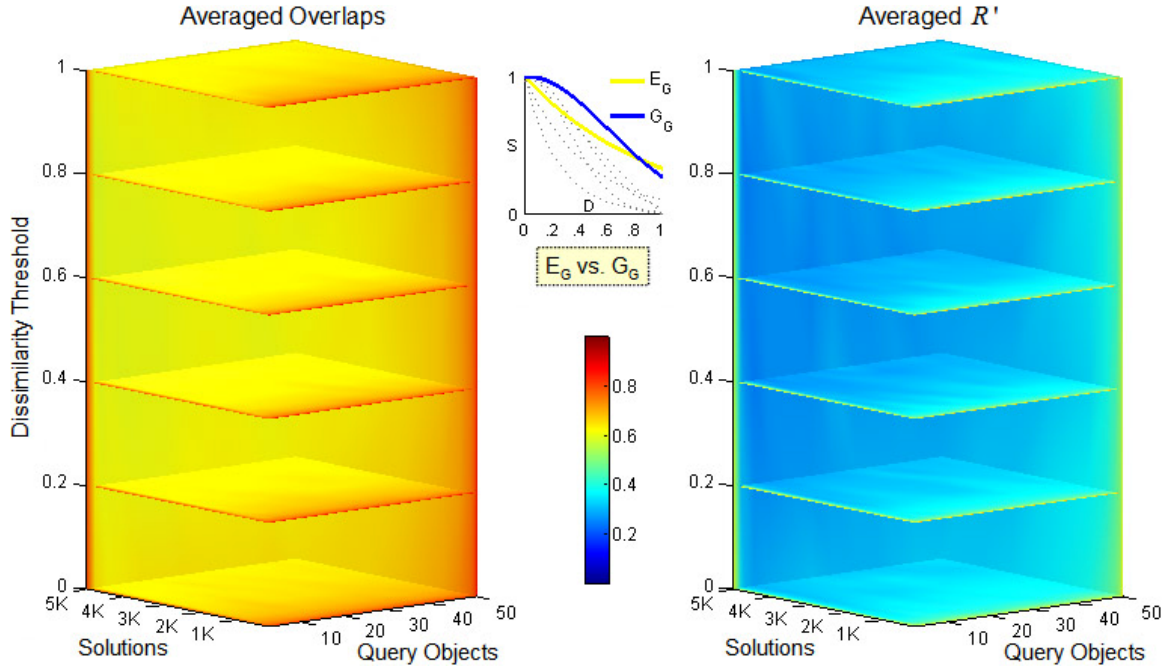


Figure 6.43: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_G$  and  $G_G$ .



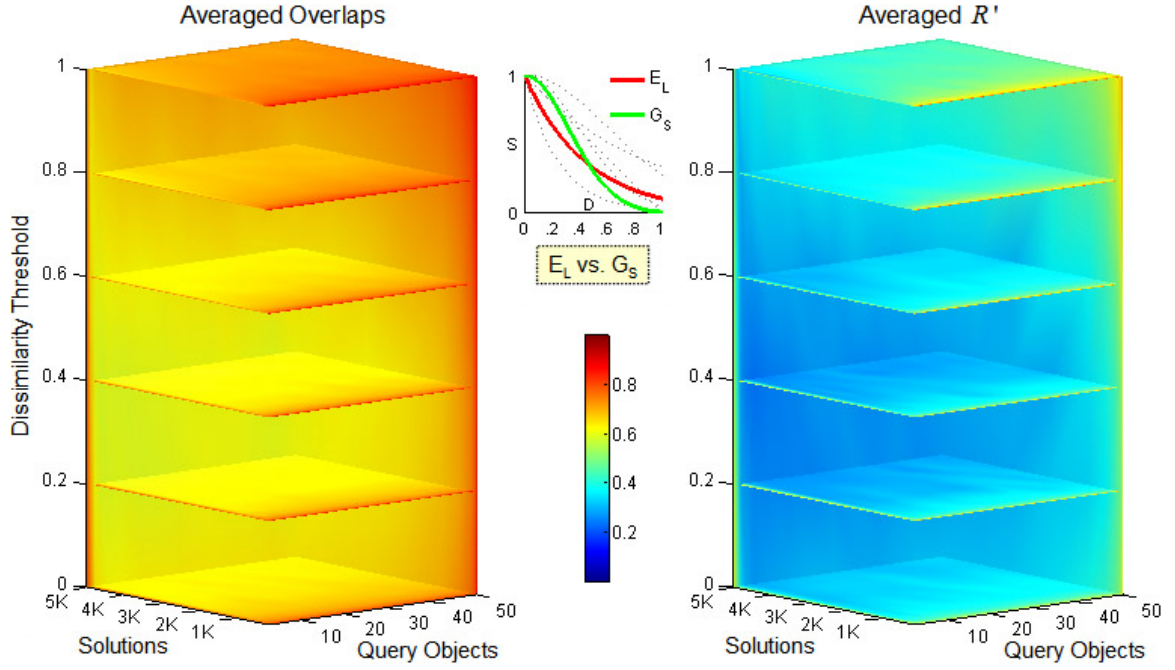


Figure 6.44: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_L$  and  $G_S$ .

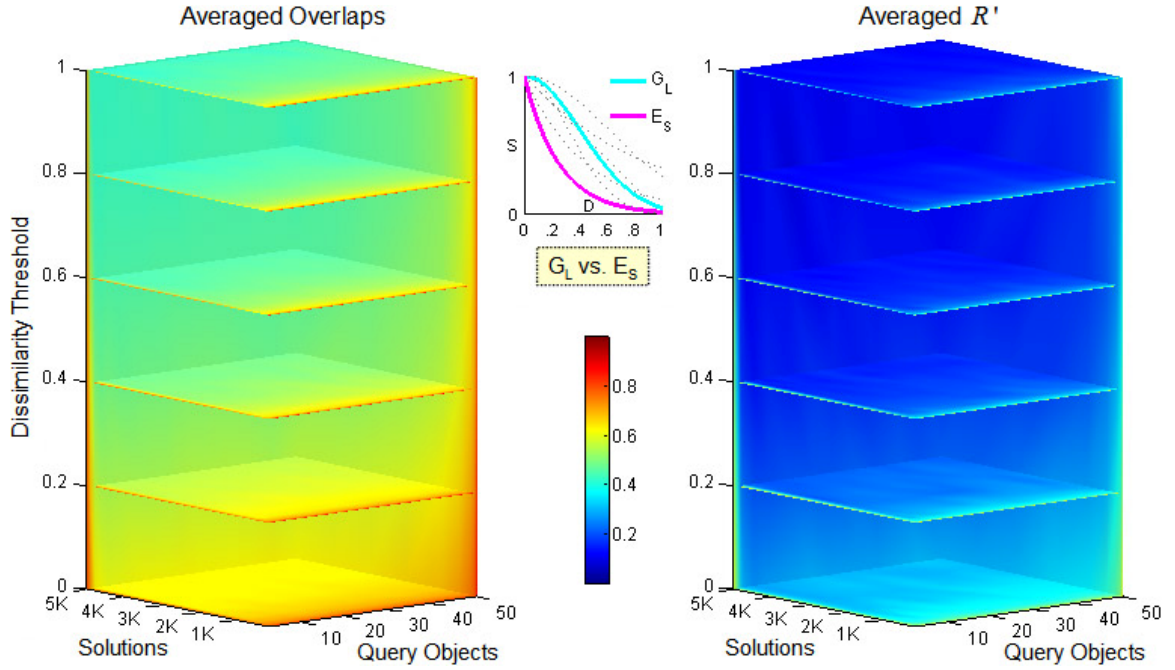


Figure 6.45: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $G_L$  and  $E_S$ .

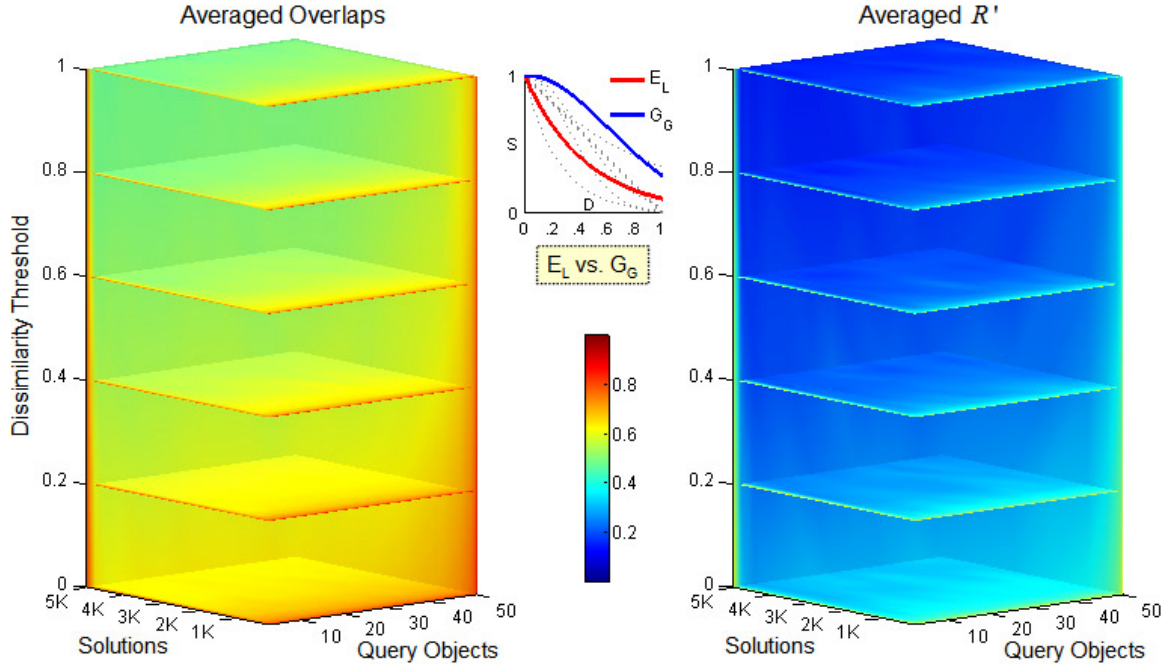


Figure 6.46: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $E_L$  and  $G_G$ .

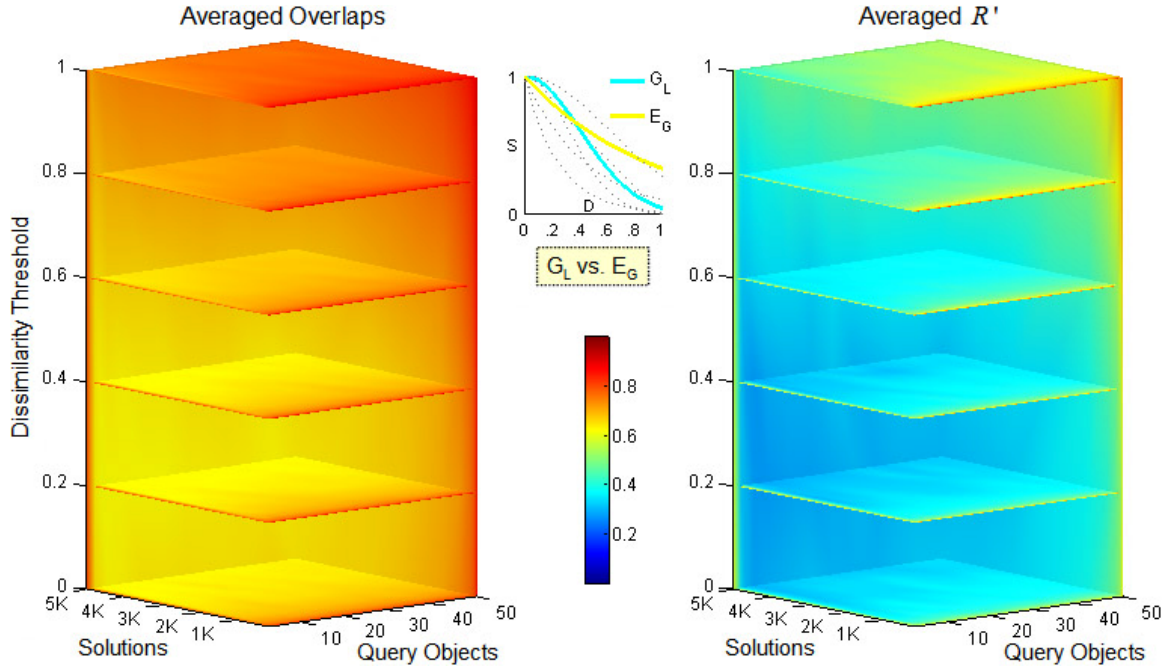


Figure 6.47: Experiment  $E_4$ : averaged overlaps and correlations between the functions  $G_L$  and  $E_G$ .

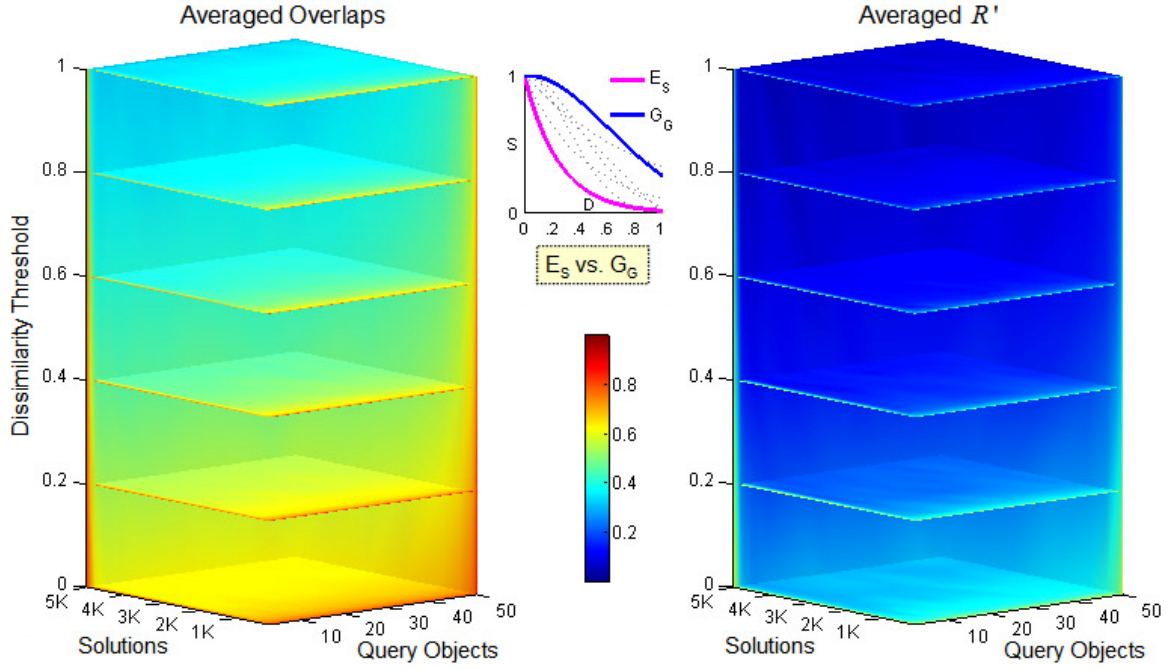


Figure 6.48: Experiment E<sub>4</sub>: averaged overlaps and correlations between the functions  $E_S$  and  $G_G$ .

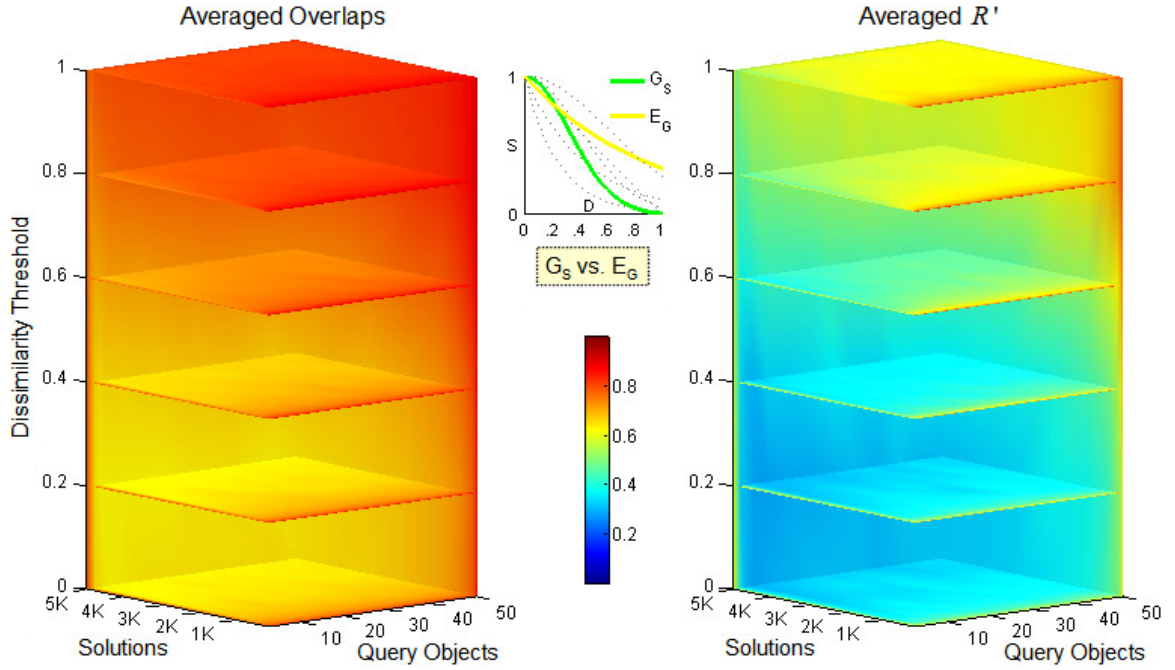


Figure 6.49: Experiment E<sub>4</sub>: averaged overlaps and correlations between the functions  $G_S$  and  $E_G$ .

The conclusion drawn from this last set of diagrams is analogous to that inferred from Figure 6.31. Exponential and Gaussian functions correlate less well to each other than each of these types does with the linear function. The diagrams verify again the initial conjecture that shape is more important than distance for the congruence of the results. For example,  $G_s$ , where the change from a concave to a convex form occurs very early along the curve, gives for the most part highly compatible results to the exponential functions. Nevertheless, it is evident that the distance factor can also become significant, particularly when its effect is propagated to that of dominantly different shapes (e.g., Figure 6.42). For well-constrained queries of a small size the measures still exhibit a high compatibility. Moreover, less arbitrary choices of the non-linear functions that do not deviate drastically from the linear function continue to produce results of high agreement (i.e., Figures 6.44, 6.47, and 6.49).

The results of Experiment E<sub>4</sub> demonstrate that, for practical applications, the choice of the conversion function does not affect seriously the results of a scene query. Therefore, the major conclusion is that the third postulate of the hypothesis (i.e., statement HS<sub>3</sub>) should be rejected. A corollary from this conclusion is that in all cases the linear function, which is simpler to calculate, can be used to convert dissimilarities to similarities.

## 6.5 Summary

This chapter evaluated the relative performance of a psychologically compliant similarity framework versus commonly encountered approaches in the literature that do not consider psychological principles about the nature and behavior of similarity. The evaluation was based on a comparison of the relevant portion of the ranking lists produced with the compliant and the deviant methods. The first three experiments focused on the distortions in the desirable set of results for queries at the object level. The fourth experiment examined the distortions for scene queries. From the three statements

of the hypothesis, only the one that pertains to the recognition of the integral attributes and groups was confirmed ( $HS_1$ ). The results point out that the distortions in the desirable set of retrieved objects are negligible when the Manhattan aggregation function is used. The distortions in the set of retrieved scenes are also acceptable for different dissimilarity to similarity conversion functions. These outcomes imply that the second and third premises of the hypothesis must be rejected. The second premise  $HS_2$ , which is concerned with the choice of the aggregation function, can be rejected only as long as a Manhattan metric is employed as a substitute to the compliant approach, whereas the third premise  $HS_3$  is rejected universally. An important implication is that the Manhattan aggregation function and the linear conversion function, both of which are simpler than their rivals, are reliable surrogates of their compliant counterparts, and able to provide psychologically trustworthy estimates of similarity.

## **CHAPTER 7**

### **CONCLUSIONS**

Relying on established psychological findings about the nature and behavior of similarity, this thesis developed a scalable framework for assessing the similarity among attribute values, objects, and spatial scenes. The framework addressed explicitly aspects of similarity that are unique to the spatial domain, but the approach is versatile enough to accommodate generic information retrieval scenarios. The formalization of the semantic aspects that are involved in the volatile and subjective task of similarity assessments is expected to contribute significantly to the design of future geographic information systems and spatial search engines that will be able to compare and process information on a semantic basis and, therefore, escape the narrow interpretation of a match to a query. This chapter provides a summary of the dissertation, highlights major contributions and findings, and discusses possible future research directions.

#### **7.1 Summary of the Thesis**

People's estimates of similarity are intuitive, qualitative, and subjective. To reliably enable corresponding comparisons in information systems, the qualitative needs to become quantitative, the subjective needs to become objective, and the comparison needs to be performed not directly on the real-world instances, but on their representations in a database. In order to perform this task computers depend on what is known and stored for the real-world entities in an information system. Such information can be encoded at different levels of abstraction. This work separated the conceptual structure of spatial information systems into the three levels of attribute values, objects, and scenes, each corresponding to user queries of successively increasing complexity. It then adopted a bottom-up approach for similarity assessments. Thus, complex assessments are simplified by breaking down the process into more simple comparisons, which involve a pair of

attribute values at a time, and then merging those individual scores to find the similarity of objects, relations, and spatial scenes.

Attributes in a database are rarely of the same type, however, and the nature of their values exhibits wide diversity. A careful inspection of different similarity models revealed that each model makes unique assumptions, emphasizes different aspects of similarity, and performs better with specific attribute types. Since none of these approaches applies globally, we did not comply with a specific model, but employed elements from each depending on the task at hand. A functional classification of attribute types was provided based on the scales of measurement (Stevens 1946; Chrisman 1995). Ratio, interval, ordinal, and cyclic values can be represented as points on a scale; therefore, a geometric approach is implied, where similarity is a function of the distance between values. For nominal values, the selection of a similarity algorithm is driven by whether such values correspond to ontological classes or not. In the first case, variations in the level of detail of the ontology will evoke the use of alternative similarity models. Detailed ontologies allow for the employment of more sophisticated models and are, therefore, capable of providing better measures of similarity than coarse ontologies. If the nominal values do not correspond to ontological classes, then a custom geometric approach can be implemented, where a nominal value is analyzed to a number of constituent ratio and ordinal dimensions. Special cases, such as counts, nominal identifiers, Boolean values, cyclic intervals, and temporal attributes were thoroughly addressed. An algorithm based on denotational semantics was also created for handling the uncertainty that null values introduce into similarity assessments.

In addition to this classification, a similarity score among attribute values required the specification of a similarity neighborhood, which divides the continuum of values into those that are similar and those that are not. Establishing fitting similarity neighborhoods and employing appropriate normalization techniques for each attribute type are important

factors for the fidelity of the computed scores to human perception, as they capture an implicit aspect of context.

The initial set of algorithms is adequate only for standard equality queries on atomic attributes. This set was expanded with a comprehensive model for handling more complex inquiries that involve the interaction of several constraints, expressed through relational and Boolean operators. A combination of such operators defines an ideal or reference object so that the objective becomes to retrieve objects similar to it. Negations require a traditional interpretation of the *not* operator, whereas the similarity for disjunctions depends entirely on the score of the most similar disjunct. Two semantically different modes of conjunction were identified: (1) locally-better and (2) globally-better matching. The former is appropriate for applications where higher-ranked constraints should dominate completely their subordinates in the constraint hierarchy. In this manner, locally-better matching resembles a multi-level sorting process. In globally-better matching, on the other hand, similarity is a weighted average of all the conjuncts. A psychologically informed approach mandates that the form of the aggregation function should be predicated on the perceptual nature of the attributes. Integral attributes are those that are perceptually correlated and perceived as one quality. When the dimensions are obvious and compelling instead, the attributes are separable. A group of integral attributes becomes a separable attribute with a Euclidean dissimilarity function, whereas separable attributes are aggregated with a city-block dissimilarity metric.

In the context of object-level queries the treatment of special cases and the specification of a weighting scheme were investigated as well. A methodology based on the assignment problem was developed to support similarity assessments among multivalued attributes. Such attributes are especially common in the representational formalisms of detailed topological and directional qualitative relations. In this case, similarity involves a comparison between two sets of values. The sets may be of equal or unequal cardinality. A flexible and intuitive weighting scheme is of paramount



importance for well-accepted similarity results because it allows users to inflict a dynamic and personal context on the assessment. Such a scheme can be based on rank-order centroid method that relies on an ordinal specification of the weighting coefficients in order of significance. The transformation of the ordinal preferences into ratio values is delegated to the information system.

Spatial scenes comprise objects arranged in a particular structure. In this sense, spatial scenes are conglomerations of objects, relations, and their attribute values. The retrieval of similar scenes to a spatial scene query was performed in three stages: (1) the relaxation stage, (2) the matching stage, and (3) the actual assessment and ranking stage. The relaxation phase consists of enlarging the initial constraints of the scene query to permit additional acceptable value combinations. Arbitrary relaxation policies may compromise the quality of the similar results or trivialize the problem by retrieving a large number of irrelevant solutions. Successful relaxation strategies, however, are strongly application-dependent and domain-dependent. Part of the domain knowledge is captured by deciding on the relative significance of the different constraints. Important constraints should be relaxed conservatively to prevent absurd matches. For spatial queries, the significance of a constraint depends on its type, its explicit or implicit specification, and the form of the query in which it is present. A key aspect of relaxation relates to the kind of spatial relations that can be used to create a weaker version of the problem. Scale-independent semi-qualitative metrics are particularly fitting for this task, as they strike a balance between strictly quantitative and qualitative approaches. They absorb much of the quantitative detail, but maintain the discriminative ability that qualitative relations lack.

During the matching phase, objects and relations of the query scene are placed in correspondence with those of the database, provided that their respective dissimilarities are within the relaxed set of values. This interactive process ensures that the quality of the matches is determined based on the combined coherence of the correspondences generated for both objects and relations. The outcome of the matching stage was an

association graph, where maximal cliques give the set of solutions to the scene query. The extraction of the maximal cliques can be performed with an exact or an approximate algorithm. Many of the maximal cliques reduce to single object solutions, which can be of little value. Criteria for discarding such suboptimal solutions, while retaining the most useful ones for presentation, were also presented.

The ranking and assessment stage consists of computing a similarity score for each clique and ranking the solutions. A scene completeness coefficient was specified as a method that can be optionally used to inflict a penalty for incomplete solutions, where the cardinality of objects in the database scene does not coincide with that of the query scene. The final similarity score of a clique is a weighted average of its object and relational components. The dissimilarities at each node and edge can be converted to similarities with linear or non-linear functions. Linear functions view similarity and dissimilarity as complementary and have a constant slope. In non-linear functions, the slope varies such that similarity decreases more rapidly with an increase of dissimilarity. Hence, values closer to a user's query are weighted more heavily, whereas those that are fairly distant are practically ignored. Psychological research concluded that exponential and Gaussian functions that do not deviate significantly from the linear plot are likely to approximate better human perceptions of similarity.

This statement was part of the hypothesis of this thesis, which asserted that the ranks of the results to a similarity query differ for psychologically compliant and psychologically deviant approaches. Besides the form of the conversion function, key aspects of a compliant process are the recognition of integral attributes and groups, and the choice of the aggregation function used for the composition of atomic assessments. The hypothesis was evaluated within SASA, a prototype software application that examined the relative performance of compliant and deviant methods for an extensive set of different database and query scenarios.

## 7.2 Major Results

The major result of this thesis comprises the findings obtained from the evaluation of the hypothesis. A central tenet of this work was that a seemingly complex similarity assessment between any two things could be segregated into conceptually simpler operations on their parts or components. In an information system that reasons about similarity in such a bottom-up fashion, the methods for acquiring dissimilarities at the lower levels, aggregating them, and converting them to similarities in order to serve the needs of higher-level assessments become important. Negligible deviations from the psychologically compliant processes in the simpler assessments may propagate at higher levels, thus introducing considerable distortions in the set of results that are consistent with people's judgments of similarities and, therefore, desirable. The evaluation of the hypothesis separated psychological aspects with a major impact on the cognitive plausibility of the results from those that are immaterial for practical retrieval purposes.

An experimental comparison between a psychologically compliant approach that recognizes groups of integral groups and a psychologically deviant approach that fails to detect such groups showed that the rankings produced with each method are dissimilar to one another. Even for a modest amount of integral attributes within the total set of attributes considered, the dissimilarities are pronounced, particularly in the presence of a single integral group or a small number of them. This trend worsens for large-scale databases. Both scenarios correspond closely to spatial representations and geographic databases. The structure of the current formalisms used to represent detailed topological, directional, and metric relations is often based on criteria other than a one-to-one correspondence between the representational primitives employed and human perception. Such formalisms are likely to contain one or few integral groups within their representation. Furthermore, geographic databases are typically large, in the order of  $10^5$  or  $10^6$  objects. This result is, therefore, significant, because it suggests that existing

similarity models may need to be revised such that new similarity algorithms must consider the possible presence of perceptually correlated attributes.

The experiments revealed, however, that the differences between the Manhattan aggregator and the compliant function in the relevant portion of the rankings are negligible. Similarly, the experiments proved that the form of the conversion function is immaterial as long as non-linear functions do not deviate extremely from the linear plot, according to what psychological research suggests. These results are important for two reasons: (1) they suggest that current similarity implementations should rely on a city-block rather than an Euclidean metric and (2) they indicate that the Manhattan metric and the interpretation of similarity and dissimilarity as complementary magnitudes still produce results of high fidelity to human perception. The second finding could also help reduce the cost that similarity computations, since both the Manhattan aggregation function and the linear conversion function are typically more efficient computationally than their psychologically compliant counterparts.

An additional contribution from the hypothesis testing is that the significance of the effect of different choices on the results can be judged on a per-application basis. The experiments simulated a large number of alternative scenarios; therefore, the produced diagrams can be consulted for specific database configurations or expected query sizes and types. More sensitive applications, for instance, may require not only high overlaps, but also identical ranks in the relevant portion. For less crucial applications, on the other hand, even a small number of overlaps may be satisfactory.

The second major contribution of this thesis is the definition of a similarity-reasoning framework for spatial information systems. The framework introduced many novel ideas and methods, while at the same time it consolidated previous efforts on similarity into a single mechanism that discarded many of their incompatible characteristics and enabled their harmonious integration. Part of the consolidation process was to assess the relative performance and suitability of different models and algorithms for specific tasks and to

suggest extensions and corrections when necessary. New contributions include among others: (1) the algorithms implemented for different attribute types, (2) a model to support similarity assessments for multivalued and composite attributes, and (3) a rationale for the relaxation of spatial queries, and particularly the relaxation of qualitative spatial relations.

The analysis demonstrated that similarity assessments become feasible for any attribute through a relatively small and well-defined set of functions. The assignment of functions to attributes is facilitated by classifying the possible attribute types based on some criterion. The benefit of providing such an abstraction is that all attributes falling under a specific category can be assigned the same generic similarity algorithm. The criterion upon which classification was based was the type of measurement that the values of an attribute perform as well as the type of change that these values imply. Ratio, interval, ordinal, nominal and cyclic types of attributes were distinguished. This is a highly semantic classification, since these scales indicate the meaning of measurement.

An aspect of similarity assessments that has been largely neglected or only inadequately treated pertains to the handling of uncertainty and incompleteness. This thesis explicitly addressed these topics when they arose. It was concluded that their proper treatment relies on a combination of featural and geometric models. The former account for elements in the source that do not have a correspondence in the target of a similarity assessment. The latter produce a similarity measure between 0 and 1 for the corresponding elements, instead of adopting the binary perspective that considers them simply as common or distinctive elements. Joint application of these models might be required at several levels, for instance, at the attribute level when values are missing or when some entities comprise more attributes than others in their specification, at the object level among multivalued attributes, and at the scene level when the compared scenes contain a different number of objects. Instead of providing a generic formula for all these cases, each topic was addressed separately to accommodate the particularities

that it manifests. For example, an enhanced approach for null values and missing attributes is possible through the introduction of different identifiers that imply varying degrees of uncertainty. An effective treatment of incompleteness for multivalued attributes and scene queries relies instead on the introduction of complete and incomplete types of solutions, and the specification of special corrective coefficients.

A key characteristic of the current framework, and distinguishing feature from previous efforts, is its independence from simplifying assumptions that may restrict its wider applicability. Every methodology eventually reduces to comparisons among attribute values, which are universal primitives across all representational structures. Reliance on this framework expedites, therefore, the process of assembling similarity models for any attribute-based representation. Conversely, the need to resort to specialized and often incompatible models that are tailored to perform with spatial relations that must belong in a finite set of predefined classes (Chang and Jungert 1996; Papadias and Delis 1997) is avoided. The independence of the framework also persists over different types of databases and queries. The methods can apply to both continuous or collection databases, as well as sketched or syntactic queries. The graph theoretical approach for scene similarity assessments addresses the very essence of scene retrieval problem and presents many desirable properties such as: (1) object identity invariance, that is, no prior knowledge of the objects' identities is required, (2) derivation of solutions drawing not only on the similarity of objects or relations, but on the combined influence of both, (3) ability to retrieve more than one solution, (4) ability to retrieve incomplete solutions.

Another attractive aspect of the current implementation is that it can—to a large degree—be implemented on top of existing database systems, a considerable advantage when such systems cannot be modified (e.g., legacy databases). The similarity algorithms implemented in this work are only limited by the level of detail in the underlying representation. This is a pragmatic limitation, since the discriminative power of a

similarity algorithm that operates on a representation cannot exceed the discriminative ability of that representation.

Similarity is resistant to many theories and models that try to formalize it. On the cognitive side, this thesis contributed towards a unifying theory of similarity, which accounts for much of its volatile and flexible behavior and alleviates many of the inefficiencies of conventional models. Such a unifying perspective was based on the idea that similarity can be measured through change. Simple philosophical principles about the nature of change and the forms in which it can be manifested provided the foundation for this novel view and guided its computational implementation. Within this context, the acquisition of a cognitively plausible similarity score is predicated on the successful measurement of the amount of change required to transform one of the compared things into the other, whether such things are attribute values, objects, or spatial scenes. In the light of this interpretation, much of the asymmetric behavior of similarity judgments finds satisfying explanation, since the amount of change required for one entity to coincide with another is not necessarily the same as when the reverse process is followed. Asymmetries, in this context, can arise naturally, without resorting to corrective factors that artificially generate them (Nosofsky 1991; Rodríguez 2000). Moreover, it is possible for asymmetric measures of similarity to be produced not only in comparisons of instances that belong to classes at different levels of abstraction (i.e., superclass-subclass relationships), but also in comparisons between instances of the same class.

Interpreting change and similarity as inverses was also helpful throughout this study, as it assisted in: (1) making the subtle distinction between two conceptually different kinds of ratio attributes, (2) addressing anomalies or rare cases in a theoretically sound and consistent manner, (3) establishing appropriate similarity neighborhoods and defining the meaning of zero similarity, (4) detecting the strengths and weaknesses of existing similarity models and reasoning about the suitability of one similarity model over another for a particular task, and (5) developing a thorough rationale for handling incompleteness.

### 7.3 Future Research

The formalization and optimization of similarity operations in information systems is a field that encompasses many possible variations and extensions. The following compilation of topics highlights issues complementary to the work presented in this thesis, as well as others that were raised during this research. Each topic includes a short introduction that highlights the extent and significance of the issue to be addressed, followed by suggestions on how it could be approached.

#### 7.3.1 Similarity Models for Detailed Spatial Relations

The 9-intersection (Egenhofer and Herring 1990) and the set of the basic cardinal directions (Frank 1996) are effective tools to reason about qualitative topological and directional relations, respectively. Such formalisms are theoretically sound yet simple, therefore, attractive both for modeling as well as for querying purposes. The caveat of using these models in spatial querying is that they are too generic and cannot distinguish among situations for which people may have distinct mental images. Complex topological, directional, and metric formalisms were developed in an effort to establish equivalence between a spatial configuration and its representation (Egenhofer and Franzosa 1995; Clementini and di Felice 1998). They model a spatial relation through a number of intersection components, each described by several topological and, optionally, some metric properties (Figure 1.2) (Shariff 1996; Nedas *et al.* in press). For example, an overlap relation between two regions may have several interior-interior intersections, and each of these encompasses a set of attribute values in its description. The problem with complex relations is that they can be overwhelmingly detailed, and usually succeed only in creating a surjective, rather than a bijective, mapping from a spatial configuration to a representational structure (i.e., one configuration may correspond to multiple representations).



To alleviate the difficulties that both coarse and detailed relations entail for similarity assessments (Section 5.4.2) this thesis advocated instead the use of simpler semi-qualitative metrics for scene queries. Such metrics may perform fine for most practical retrieval scenarios. Occasionally, however, the focus of a query may be strictly on topological or directional similarity. On the other hand, the employment of semi-qualitative metrics may not always be possible and the ability to establish similarity among detailed relations may be further needed in order to break ties among retrieved solutions. Current similarity models mostly apply to coarse relations and yield crude estimates based on simple conceptual neighborhood graphs (Freksa 1991; Egenhofer and Mark 1995a; Blaser 2000). Models for detailed relations are scarce (Goyal and Egenhofer 2001). These arguments stress the need to establish effective similarity models for detailed spatial relations.

During the course of this thesis it was realized that the current framework is a good candidate for this task if one only transposes the level of abstraction. Within the context of a topological relation for instance, the detailed relation itself can be thought of as a spatial scene. Intersection components correspond to objects, and the only relation among these “objects” is their sequence. The parameters that are used to describe the intersection components and their values correspond to attributes and attribute values, respectively. This one-to-one correspondence suggests that the methodology employed for scene similarity can be recursively applied to assess the similarity of detailed spatial relations. The suitability of the current framework to establishing similarities of detailed relations is further emphasized by its ability to handle multivalued attributes because many of the parameters used to describe the intersection components can accept multiple values. The provisions made to account for incompleteness are also vital because the number of intersections between two compared relations may differ.

Although this thesis provides the foundation for reasoning about the similarity of detailed topological relations, there is room for differences in the approach, which future

research must detect and address. For example, a relaxation process may not be necessary because the anticipated number of intersection components is relatively small. Simplifications may also be possible because the only relation of interest among intersection components is their order. The results from the hypothesis testing also point out that future efforts on the same topic should also concentrate on the detection of integral groups within the representational formalisms used for complex relations. This task can be accomplished by combining human-subject experiments with multi-dimensional scaling techniques that reveal the prominent dimensions in similarity judgments. Maddox (1992) provides a survey and analysis of tests that can be used to decide the separability or integrality of sets of attributes. Besides contributing useful similarity models, research in this direction could also be reciprocally beneficial. It may discover, for instance, that simpler representations perform equally well, or derive new criteria about how future formalisms for representing detailed spatial relations should be structured.

### 7.3.2 Automated Weight Calibration and Constraint Significance

Understanding how people prioritize individual components (e.g., geometric vs. thematic specifications, completeness vs. topology vs. direction, a scene's relational vs. the object component) in a similarity assessment would assist in establishing the relative significance of constraints in spatial object or scene queries and improving the current framework in two significant aspects. First, it would help outline a more informed relaxation strategy, which is key to retrieving better results and speeding up the retrieval process. Second, it would enhance the user-system interaction, contributing to the vision of a naive geography environment where user involvement in the specifics of the system is expected to be minimal. For instance, users could simply query by selecting an object or a scene. This type of querying is more intuitive as it removes the burden of creating SQL statements, forming Boolean expressions, and worrying about weight specifications. For geographic information systems in particular, this querying technique would be even

more advantageous, because such systems provide inherent support for visual inspection and selection of objects.

In contrast to thresholds, however, an automated weight assignment by the system is a considerably more perplexed issue, because it depends on a multitude of factors, not all of which can be *a priori* known. Such factors are the form of the query, the context of the comparison that mirrors the intents and purposes of the user, and even the proficiency of each user in expressing the query using the constructs provided by the system. Relevant research has only contributed peripheral solutions, rather than addressing the core of the problem. For example, some efforts rely on a “more like this” criterion, where users indicate the result closest to their expectations, and the weights are fine-tuned accordingly for the next retrieval cycle (Ortega-Binderberger *et al.* 2002; Chakrabarti *et al.* 2003). An excessive repetition of the querying process, however, may become frustrating. Other methods, such as the ones adopted in this thesis (Section 4.3), aim at reducing the cognitive load through the assignment of ordinal preferences; however, the reliance on the user’s explicit instructions remains a prerequisite. Moreover, the process may become unfathomable for spatial scene comparisons due to the plethora of existing constraints, their presence at different levels of abstraction (i.e., scene, object, and attribute levels), and the complex interactions among them. On the other hand, a default equal-weighting scheme in absence of any user feedback is more like adopting the ostrich’s behavior to danger. It has been observed that in many contexts several dimensions are implicitly highlighted more than others (Attneave 1950; Torgerson 1965; Nosofsky 1992).

Recognizing these dimensions may be difficult. An automated weight calibration for all circumstances and users is an elusive and probably unrealistic goal. Future research should first establish whether the assessment of relevance of individual components is consistent for different users and tasks. If the outcome is affirmative, the next task would be to provide generic guidelines about the prominence of several components over others

and to incorporate them in weight templates for typical query cases. This should be done in conjunction with the development of algorithms or agents that monitor the users' querying patterns over time and create dynamic personalized user-profiles.

An obvious route to these objectives is through human-subject testing. An alternative approach is through statistical techniques and stems from the observation that in many cases, several attributes can have a functional dependency on others. A functional dependency between two attributes  $A_i$  and  $A_j$  holds when the value of  $A_i$  for a tuple  $t$  uniquely determines the value of  $A_j$  for the same tuple. Considering both of these attributes equally weighted through an automated process introduces a "double-counting" bias in the similarity assessment. This argument can be generalized to different degrees of correlation between attributes. It is in this area, therefore, that causal, rather than perceptual, correlation becomes relevant for similarity. Methods have to be found that assess the degree of correlation between attributes and derive the ratio values of weights accordingly. In the general case, a slight positive or negative correlation even between practically independent attributes will exist. Hence, such methods should also need to decide on the thresholds beyond which correlation entails a bias introduction.

### 7.3.3 Efficient Execution of Similarity Queries

This thesis focused primarily on the conceptual level of performing similarity operations in a database. The results of the hypothesis evaluation also contributed to more efficient query processing by justifying the use of simpler equations in the assessments. Many issues, however, still remain, which must be resolved in order to complement this work and provide efficient mechanisms and algorithms for the faster execution of similarity queries.

Traditionally, similarity operations have been in the realm of software engineering. Further efficiency can be achieved if similarity becomes an integral component of future system architectures. This integration will contribute to the trend that states that the

disciplines of information retrieval and database management should become more tightly joined (Elmasri and Navathe 2000). Many commercial products have already adopted this paradigm. Examples include the data blade feature of Informix Universal Server, which makes use of the WordNet thesaurus, and the specification of the *SIMILAR* function introduced within later versions of the query language SQL. Such extensions are still crude and unable to deal with the full spectrum of similarity in a database. Hence, further research is required on the language and architectural extensions needed to enhance current DBMSs with semantic capabilities.

In relational databases, for example, the similarity functions could be implemented as system-stored procedures. A one-to-many relationship can exist between one of these procedures and some of the attributes in the database. These mappings could be registered in the system catalog or the data dictionary. Similar methods could be followed for object-oriented DBMSs where the similarity functions may be implemented as internal functions of objects—whether such objects are classes or attributes. In such systems, objects may contain more than one function, or make use of polymorphism to account for similarity comparisons with objects whose values use different data types. Part of the research should focus exclusively on the physical level to provide sophisticated indexing methods for similarity queries (Roussopoulos *et al.* 1995; White and Jain 1996), or investigate how such indexing structures as R-trees (Guttman 1984) can be fully exploited. Other topics for research include language extensions and interface design that will assist users in interacting more efficiently and customizing their queries during a similarity retrieval session.

Another set of future research questions, related to efficiency, deals with the implementation of approximate algorithms for scene matching (Papadias *et al.* 2003; Rodríguez and Jarur 2005), particularly the task of extracting maximal cliques from an association graph. Although there can be no formal estimates on the performance of such algorithms, they are able to demonstrate a remarkable improvement in efficiency

compared to their exact counterparts (Bomze *et al.* 1999). Some of the approximate algorithms, however, return only one solution and none guarantees the retrieval of the optimal solutions. Another problem is that the performance of several of these methodologies (e.g., genetic algorithms) is heavily dependent on a high number of input parameters that must be defined prior to query execution. Therefore, in order to tune such algorithms correctly and to obtain satisfactory results, the user must be thoroughly acquainted with the algorithms' internal operation. Examples of approximate algorithms include *DNA-Computing* (Zhang and Shin 1998), *simulated annealing* (Aarts and Korst 1989), *tabu search* (Battiti and Protasi 1995), and *genetic algorithms* (Marchiori 1998). Such algorithms should be evaluated to assess their relative performance, fine-tuned for the problem of spatial scene queries through the embedding of knowledge particular to the spatial domain, and modified, if possible, to require little or no user input. An additional challenging topic with efficiency repercussions is the development of better semi-qualitative metrics that comply with the requirements of *continuity*, *scale-invariance*, *object identity-invariance*, *universality*, and *minimality* that were outlined and analyzed in Section 5.4.2.

#### 7.3.4 Extension to Heterogeneous Database Systems

Previous work on multidatabase systems from the computer science (Doan and Halevy 2005) and the geographic information communities (Duckham and Worboys 2005; Lutz and Klien 2006) concentrated primarily on data integration, that is, the process by which the schematic, structural, and semantic heterogeneities among such systems are resolved. The ultimate goal is to ensure location, schema, and language transparency for the users, thus giving them the illusion of accessing a single centralized database (Busse *et al.* 1999; Uschold and Gruninger 2004). After the integration has taken place, users can retrieve information by querying the heterogeneous system in the same way that they would query a centralized DBMS.

Within this field, similarity was used mainly from the perspective of information integration rather than that of information retrieval. Thus, it was employed as a tool for identifying and matching corresponding structural elements among different systems that model related application domains (Rahm and Bernstein 2001; Maedche and Staab 2002; Kalfoglou and Schorlemmer 2003; Noy and Musen 2003; Palopoli *et al.* 2003; Rodríguez and Egenhofer 2003; Noy 2004). Little emphasis, however, was placed on the requirements on data integration so that similarity retrieval becomes feasible in such systems. The few approaches that considered the issue (Mena *et al.* 1996) provided coarse similarity measures, but that usually came as a welcome side-effect of the proposed data integration architecture, and not as a result of a thorough and explicit treatment.

Some of the basic assumptions for determining similarity in a homogeneous environment, however, could be violated in a heterogeneous setting. A logical extension of this work is, therefore, to investigate the various impediments in the retrieval of similar results from heterogeneous data sources and suggest ways to address them. The approach should follow a detailed compilation of possible heterogeneity problems (Batini *et al.* 1986; Sheth and Larson 1990; Kim and Seo 1991), examine each in isolation, and suggest extensions to the existing data integration architectures, when they are not adequate to enable the types of similarity assessments developed in this thesis. For some of the problems it is possible that the methods of this thesis will create less stringent requirements than those imposed by traditional retrieval, because similarity itself could be the means by which they could be resolved (e.g., missing attributes can be addressed with *dne* or *ni* types of nulls).

### 7.3.5 Formalizing Similarity in Ontologies for the Semantic Web

Prominent commercial GIS packages rely on a relational or object-oriented architecture to organize data. Throughout this study, it was assumed that results to similarity queries

were obtained from such structured data sources. It is interesting to investigate how the current framework should be adapted to apply on the semantic web (Berners-Lee *et al.* 2001) where the structure of data sources is less rigid, and how it can be combined with emerging standards and technologies such as XML (i.e., eXtensible Markup Language), RDF (Resource Description Framework) (Decker *et al.* 2000), and the web ontology language, OWL. This exploration could lead to a semantic web that is also able to reason about similarity.

The building blocks of the semantic web are domain ontologies. Research on this topic should assess the possibility of creating similarity-enhanced ontologies. These could be ontologies that, in addition to explicating the meaning and formalizing the relationships between concepts and properties for a specific domain of interest, also provide an agreement on the meaning of similarity between such concepts and properties for the domain community. This meaning can be captured by embedding similarity algorithms in an ontology as ontological functions. Each role (i.e., attribute) could be associated with one similarity function appropriate to assess the semantic proximity among its values. The similarity model for concepts would be a global function in the ontology and not unique to each concept. It would exist at a meta-ontological level, because the arguments passed to such an algorithm are the concepts themselves. These functions could have a suggestive character and need only be specified in their most generic form in the ontology. Their role could be to formalize the context of similarity, but not necessarily elaborate on its exact quantification details.

Similarity relations fit well into an ontological framework, because it is expected that people who commit to the same ontology perceive identically not only the concepts that are important in their domain of interest, but also the similarity relations that hold among these concepts. This alignment of individual similarity views towards a common one is emphasized by the fact that ontologies already have inherent a notion of qualitative similarity relations among the concepts that they model. This notion is reflected in their



structure (i.e., in the way they specify classes and subclasses) and in the properties and roles that are attributed to each concept. Furthermore, ontologies typically include restrictions on the allowed ranges for such properties and roles, thus demarcating explicitly the allowed ranges for values within a domain. Some of the similarity algorithms in this work exploit exactly those features to derive a quantitative similarity measure. Geometric models use range-related information, network models use the ontology structure, and featural models use the common and different properties of the entities. Formalizing similarity within ontologies would be a step forward in the employment of ontologies not only as means for semantic integration, but also as tools for semantic management (Rosenthal *et al.* 2004), and would help their transition from symbolic to conceptual constructs.

Another related topic of larger scope is the investigation of a more scalable architecture, by implementing a top-level similarity ontology and then provide mappings from the models and functions of that ontology to the concepts and roles of other domain ontologies. The feasibility of the construction of a top-level similarity ontology, as well as the precise details of its implementation, are interesting areas for future research.

#### 7.3.6 Discovering Additional Applications of Similarity

The motivation of this thesis was the enhancement of spatial information systems for semantic information retrieval. The formalization of similarity, however, can open up a world of additional exciting possibilities where similarity can be exploited in a variety of ways and as a tool that will facilitate and automate many diverse tasks. Some of these tasks are related to user-interface improvements and extensions in the functionality of existing GISs. Examples include:

- The identification and removal of duplicate entries in a database or during the process of merging different databases into a larger one (Chatterjee and Segev 1991; Monge and Elkan 1997; Dey *et al.* 2002).

- The use of similarity as a predictive tool. For instance, in the case of failure of a pipe, the sewer company may search its GIS for pipes with characteristics similar to the failed pipe, and place them under inspection or perform maintenance on them.
- Replacement of expert operations with simpler alternatives. Answering a query such as “find all locations that are within 0.5km of a major road, not in a built-up area and on a sand/gravel deposit” (Worboys and Duckham 2004) requires a layer-based analysis with current GIS tools, which will only yield exact matches. In contrast, a similarity-based approach would not force the users to cope with sequences of complex operations of buffering and overlays. The answer to the above query could be provided immediately by applying the methods presented in this thesis.
- Landmark determination. This area is another promising field for similarity, or rather, dissimilarity. Since landmarks are entities that stand out from their surroundings, they are expected to be the most dissimilar from other entities nearby. A four step approach could then be to: (1) decide on the conceptually salient characteristics (i.e., attributes that are important for a landmark’s determination) (Winter 2003; Nothegger *et al.* 2004; Klippel and Winter 2005), (2) define the desired extent of the spatial neighborhood from which landmarks will be extracted, (3) calculate for each object in this neighborhood the sum of its dissimilarities to other objects, and (4) based on some threshold extract the objects with the largest sum as possible landmarks. Of course, these are only crude guidelines that would also need to be combined with methods for space partitioning and principles from psychological theories of attention. Another important aspect to consider is the distribution of objects (Haken and Portugali 2003). For example, many prominent buildings are often grouped together in the center of cities; therefore, landmark determination for such areas should be based on finer differences or, perhaps other criteria.

This list of examples is only a small subset of the enhancements that become feasible with semantic similarity assessments. Future studies should explore the theory and

methods required to make these ideas concrete and incorporate them in the functionality of next-generation GISs. Such enhanced modes of interaction have the potential to maximize expressiveness, and to change the conventional ways of thinking about query formulation.

#### 7.3.7 Evolution of Similarity Models

Defining a flawless computational implementation of a notion as abstract as similarity would probably require total decryption of the processes of the human mind. The problem encompasses many aspects and questions for some of which no definitive answers yet exist. This thesis strived to keep a balance by incorporating the findings of theoretical disciplines, such as philosophy and psychology, while still maintaining practicality and versatility. The evolution of the field of semantically similar information retrieval needs to proceed hand-in-hand along with the latest developments in those disciplines. New theories and interpretations of change must be considered and new findings about the nature and properties of similarity must be integrated into the future similarity algorithms as we make advances in our understanding of the human brain.

## REFERENCES

- E. Aarts and J. Korst (1989) *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley and Sons, Inc., New York, NY.
- P. Agouris, J. Carswell, and A. Stefanidis (1999) An Environment for Content-Based Image Retrieval from Large Spatial Databases. *Journal of Photogrammetry and Remote Sensing* 54(4): 263-272.
- A. Aho and M. Corasick (1975) Efficient String Matching: an Aid to Bibliographic Search. *Communications of the ACM* 18(6): 333-340.
- J. Aisbett and G. Gibbon (1994) A Tunable Distance Measure for Coloured Solid Models. *Artificial Intelligence* 65(1): 143-164.
- J. Allen (1983) Maintaining Knowledge about Temporal Intervals. *Communications of the ACM* 26(11): 832-843.
- A. Ambler, H. Barrow, C. Brown, R. Burstall, and R. Popplestone (1973) A Versatile Computer-Controlled Assembly System. in: *Proceedings of the 3rd International Joint Conference on Artificial Intelligence (IJCAI '73)*, Stanford, CA, pp. 298-307.
- Aristotle (350 B.C.-a) *Physics*.
- Aristotle (350 B.C.-b) *Metaphysics*.
- P. Artymiuk, A. Poirrette, H. Grindley, D. Rice, and P. Willett (1994) A Graph-Theoretic Approach to the Identification of Three-Dimensional Patterns of Amino Acid Side-Chains in Protein Structures. *Journal of Molecular Biology* 243(2): 327-344.
- F. Ashby and J. Townsend (1986) Varieties of Perceptual Independence. *Psychological Review* 93(2): 154-179.

- F. Ashby and W. Lee (1991) Predicting Similarity and Categorization from Identification. *Journal of Experimental Psychology: General* 120(2): 150-172.
- M. Atkinson, F. Banchilhon, D. DeWitt, D. Maier, K. Dittrich, and S. Zdonik (1989) The Object-Oriented Database System Manifesto. in: *Proceedings of the First International Conference on Deductive and Object-Oriented Databases*, Kyoto, Japan, pp. 223-240, Elsevier Science Publishers.
- F. Attneave (1950) Dimensions of Similarity. *American Journal of Psychology* 63(4): 516-556.
- C. Bachman, L. Cohn, W. Florance, F. Kirshenbaum, H. Kuneke, C. Mairet, E. Scott, E. Sibley, D. Smith, T. Steel, J. Turner, and B. Yormark (1975) Interim Report: ANSI/X3/SPARC Study Group on Data Base Management Systems. *ACM SIGMOD Record* 7(2): 1-140.
- D. Ballard and C. Brown (1982) *Computer Vision*. Prentice-Hall, Englewood Cliffs, NJ.
- S. Banerjee and T. Pedersen (2003) Extended Gloss Overlaps as a Measure of Semantic Relatedness. in: *Proceedings of the Eighteenth International Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, pp. 805-810.
- J. Barnes (Ed.) (1995) *The Cambridge Companion to Aristotle* (10th Edition). Cambridge University Press, New York, NY.
- F. Barron (1992) Selecting a Best Multiattribute Alternative with Partial Information about Attribute Weights. *Acta Psychologica* 80(1-3): 91-103.
- F. Barron and B. Barret (1996) Decision Quality Using Ranked Attribute Weights. *Management Science* 42(11): 1515-1523.
- C. Batini, M. Lenzerini, and S. Navathe (1986) A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys* 18(4): 323-364.

- R. Battiti and M. Protasi (1995) Reactive Local Search for the Maximum Clique Problem. International Computer Science Institute, Berkeley, CA, Technical Report TR-95-052.
- K. Beard (1989) Design Criteria for Automated Generalization. in: International Cartographic Association Conference (ICA), Budapest, Hungary, pp. 32-40.
- T. Bench-Capon and P. Visser (1997) Ontologies in Legal Information Systems; The Need for Explicit Specifications of Domain Conceptualizations. in: Proceedings of the Sixth International Conference on Artificial Intelligence and Law, Melbourne, Australia, pp. 132-141, ACM Press, New York, NY.
- R. Benjamins (1998) The Ontological Engineering Initiative (KA)<sup>2</sup>. in: N. Guarino (Ed.) Proceedings of Formal Ontology in Information Systems 1998 (FOIS '98), Trento, Italy, pp. 287-301, IOS Press, Amsterdam, The Netherlands.
- A. Berenzweig, B. Logan, D. Ellis, and B. Whitman (2003) A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures. in: Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR 2003), Washington, DC.
- T. Berners-Lee, J. Hendler, and O. Lassila (2001) The Semantic Web. *Scientific American* 284(5): 34-43.
- P. Bernstein (2003) Applying Model Management to Classical Meta Data Problems. in: Proceedings of the Conference on Innovative Database Research (CIDR), pp. 209-220.
- P. Bernstein, S. Melnik, M. Petropoulos, and C. Quix (2004) Industrial-Strength Schema Matching. *ACM SIGMOD Record* 33(4): 38-43.
- Y. Bishr (1998) Overcoming the Semantic and Other Barriers to GIS Interoperability. *International Journal of Geographical Information Science* 12(4): 299-314.

- J. Biskup and D. Embley (2003) Extracting Information from Heterogeneous Information Sources using Ontologically Specified Target Views. *Information Systems* 28(3): 169-212.
- A. Blaser (2000) Sketching Spatial Queries. Ph.D. Thesis, University of Maine, Orono, ME.
- A. Blaser and M. Egenhofer (2000) A Visual Tool for Querying Geographic Databases. in: V. Di Gesù, S. Levialdi, and L. Tarantino (Eds.), *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '00)*, Palermo, Italy, pp. 211-216, ACM Press, New York, NY.
- D. Bohm (1951) *Quantum Theory*. Dover Publications, New York, NY.
- I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo (1999) The Maximum Clique Problem. in: D.-Z. Du and P. Pardalos (Eds.), *Handbook of Combinatorial Optimization (Supplement Volume A)*, pp. 1-74, Kluwer Academic Publishers, Boston, MA.
- K. Borchering, T. Eppel, and D. von Winterfeldt (1991) Comparison of Weighting Judgments in Multiattribute Utility Measurement. *Management Science* 37(12): 1603-1619.
- A. Borning, R. Duisberg, B. Freeman-Benson, A. Kramer, and M. Woolf (1987) Constraint Hierarchies. in: N. Meyrowitz (Ed.) *Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA '87)*, Orlando, FL, pp. 48-60, ACM Press, New York, NY.
- A. Borning, B. Freeman-Benson, and M. Wilson (1992) Constraint Hierarchies. *Lisp and Symbolic Computation* 5(3): 221-268.
- R. Boyer and J. Moore (1977) A Fast String Searching Algorithm. *Communications of the ACM* 20(10): 762-772.

- C. Bron and J. Kerbosch (1973) Algorithm 457: Finding All Cliques of an Undirected Graph. *Communications of the ACM* 16(9): 575-577.
- B. Bruegger and W. Kuhn (1991) Multiple Topological Representations. National Center for Geographic Information and Analysis, University of Maine, Orono, ME, Technical Report 91-17.
- T. Bruns and M. Egenhofer (1996) Similarity of Spatial Scenes. in: M.-J. Kraak and M. Molenaar (Eds.), *Seventh International Symposium on Spatial Data Handling (SDH '96)*, Delft, The Netherlands, pp. 173-184, Taylor & Francis, London, U.K.
- A. Budanitsky (1999) Lexical Semantic Relatedness and its Application in Natural Language Processing. Computer Systems Research Group, University of Toronto, Toronto, Canada, Technical Report CSRG-390.
- A. Budanitsky and G. Hirst (2001) Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures. in: *Workshop on WordNet and Other Lexical Resources*, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000), Pittsburgh, PA, USA.
- J. Burg and R. Van de Riet (1998) COLOR-X: Using Knowledge from WordNet for Conceptual Modeling. in: C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 353-377, The MIT Press, Cambridge, MA.
- S. Busse, R.-D. Kutsche, U. Leser, and H. Weber (1999) Federated Information Systems: Concepts, Terminology and Architectures. Technische Universität Berlin, Berlin, Germany, Technical Report Forschungsberichte des Fachbereichs Informatik Nr. 99-9.
- B. Buttenfield (1989) Multiple Representations: Initiative 3 Specialist Meeting Report. National Center for Geographic Information and Analysis, Santa Barbara, CA, Technical Report 89-3.



- D. Calcinelli and M. Mainguenaud (1994) Cigales, a Visual Query Language for a Geographical Information System: the User Interface. *Journal of Visual Languages and Computing* 5(2): 113-132.
- J. Carswell (2000) Using Raster Sketches for Digital Image Retrieval. Ph.D. Thesis, University of Maine, Orono, ME.
- P. Caws (1959) Definition and Measurement in Physics. in: W. Churchman and P. Ratoosh (Eds.), *Measurement: Definitions and Theories*, pp. 3-17, John Wiley & Sons, Inc., New York, NY.
- K. Chakrabarti, M. Ortega-Binderberger, S. Mehrotra, and K. Porkaew (2003) Evaluating Refined Queries in Top-k Retrieval Systems. *IEEE Transactions on Knowledge and Data Engineering* 16(2): 256-270.
- D. Chamberlin, M. Astrahan, K. Eswaran, P. Griffiths, R. Lorie, J. Mehl, P. Reisner, and B. Wade (1976) SEQUEL 2: A Unified Approach to Data Definition, Manipulation, and Control. *IBM Journal of Research and Development* 20(6): 560-575.
- K. Chang, B. He, C. Li, M. Patel, and Z. Zhang (2004) Structured Databases on the Web: Observations and Implications. *ACM SIGMOD Record* 33(3): 61-70.
- S.-K. Chang and E. Jungert (1996) *Symbolic Projection for Image Information Retrieval and Spatial Reasoning*. Academic Press, New York, NY.
- A. Chatterjee and A. Segev (1991) Data Manipulation in Heterogeneous Databases. *ACM SIGMOD Record* 20(4): 64-68.
- P. Chen (1976) The Entity Relationship Model-Toward a Unified View of Data. *ACM Transactions on Database Systems* 1(1): 9-36.
- N. Chrisman (1995) Beyond Stevens: A Revised Approach to Measurement for Geographic Information. in: *Twelfth International Symposium on Computer-Assisted Cartography, Auto-Carto 12*, Charlotte, NC, pp. 327-336.

- N. Chrisman (2001) *Exploring Geographic Information Systems* (2nd Edition). John Wiley & Sons, Inc., New York, NY.
- C. Claramunt, M. Thériault, and C. Parent (1997) A Qualitative Representation of Evolving Spatial Entities in Two-Dimensional Spaces. in: S. Carver (Ed.), *Innovations in GIS V*, pp. 119-129, Taylor & Francis.
- E. Clementini and P. di Felice (1998) Topological Invariants for Lines. *IEEE Transactions on Knowledge and Data Engineering* 10(1): 38-54.
- E. Codd (1970) A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM* 13(6): 377-387.
- E. Codd (1979) Extending the Database Relational Model to Capture More Meaning. *ACM Transactions on Database Systems* 4(4): 397-434.
- E. Codd (1986) Missing Information (Applicable and Inapplicable) in Relational Databases. *ACM SIGMOD Record* 15(4): 53-78.
- F. Coenen and P. Visser (1998) A General Ontology for Spatial Reasoning. in: R. Miles, M. Moulton, and M. Bramer (Eds.), *Research and Development in Expert Systems XV: Proceedings of ES '98, the Eighteenth Annual International Conference of the British Computer Society*, London, U.K., pp. 44-57, Springer.
- A. Cohn and S. Hazarika (2001) Qualitative Spatial Representation and Reasoning: An Overview. *Fundamenta Informaticae* 46(1-2): 1-29.
- A. Collins and E. Loftus (1975) A Spreading Activation Theory of Semantic Processing. *Psychological Review* 82(6): 407-428.
- V. Cross and T. Sudkamp (2002) *Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications* (2nd Edition). *Studies in Fuzziness and Soft Computing*, Vol. 93. Physica-Verlag Heidelberg, New York, NY.

- Z. Cui, D. Jones, and P. O'Brien (2001) Issues in Ontology-Based Information Integration. in: Seventeenth International Joint Conference on Artificial Intelligence, Seattle, WA, pp. 141-146.
- K. Dahlgren (1988) Naive Semantics for Natural Language Understanding. Kluwer Academic Publishers, Norwell, MA.
- F. Damerau (1964) A Technique for Computer Detection and Correcting of Spelling Errors. *Communications of the ACM* 7(3): 171-176.
- C. Date (1982) Null Values in Database Management. in: Proceedings of the 2nd British National Conference on Databases (BNCOD-2), Bristol, U.K., pp. 147-166.
- R. Dawes and B. Corrigan (1974) Linear Models in Decision Making. *Psychological Bulletin* 81(2): 95-106.
- R. Dawes (1979) The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist* 34(7): 571-582.
- S. Decker, S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks (2000) The Semantic Web: The Roles of XML and RDF. *IEEE Internet Computing* 4(5): 63-74.
- D. Dey, S. Sarkar, and P. De (2002) A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering* 14(3): 567-582.
- F. Di Loreto, F. Ferri, F. Massari, and M. Rafanelli (1996) A Pictorial Query Language for Geographical Databases. in: T. Catarci, M. Costabile, S. Levialdi, and G. Santucci (Eds.), *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '96)*, Gubbio, Italy, pp. 233-244.
- E. Dijkstra (1959) A Note on Two Problems in Connection with Graphs. *Numerische Mathematik* 1: 269-271.

- K. Dittrich and A. Geppert (1997) Object-Oriented DBMS and Beyond. in: Conference on Current Trends in Theory and Practice of Informatics, pp. 275-294.
- A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy (2003) Learning to Match Ontologies on the Semantic Web. *The VLDB Journal* 12(4): 303-319.
- A. Doan and A. Halevy (2005) Semantic Integration Research in the Database Community. *AI Magazine* 26(1): 83-94.
- R. Domenig and K. Dittrich (1999) An Overview and Classification of Mediated Query Systems. *ACM SIGMOD Record* 28(3): 63-72.
- D. Dou, D. McDermott, and P. Qi (2003) Ontology Translation on the Semantic Web. in: Proceedings of International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003), Catania, Italy, Lecture Notes in Computer Science, Vol. 2888, pp. 952-969, Springer-Verlag, Berlin, Germany.
- W. Dubitzky, F. Carville, and J. Hughes (1993) Case-Level Knowledge Modelling in CBR. *Irish Journal of Psychology* 14(3): 478-479.
- M. Duckham and M. Worboys (2005) An Algebraic Approach to Automated Information Fusion. *International Journal of Geographical Information Science* 19(5): 537-557.
- M. Egenhofer and J. Herring (1990) Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. Department of Surveying Engineering, University of Maine, Orono, ME, Technical Report.
- M. Egenhofer and K. Al-Taha (1991) Reasoning about Gradual Changes of Topological Relationships. in: U. Formentini (Ed.), *Theory and Methods of Spatio-Temporal Reasoning in Geographic Space*, Lecture Notes in Computer Science, Vol. 639, pp. 196-219, Springer-Verlag.

- M. Egenhofer and K. Al-Taha (1992) Reasoning about Gradual Changes of Topological Relationships. in: A. Frank, I. Campari, and U. Formentini (Eds.), Proceedings of the International Conference GIS - From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Pisa, Italy, Lecture Notes in Computer Science, Vol. 639, pp. 196-219, Springer-Verlag.
- M. Egenhofer and A. Frank (1992) Object-Oriented Modeling for GIS. Journal of the Urban and Regional Information Systems Association 4(2): 3-19.
- M. Egenhofer and J. Sharma (1993) Assessing the Consistency of Complete and Incomplete Topological Information. Geographical Systems 1(1): 47-68.
- M. Egenhofer (1994a) Spatial SQL: A Query and Presentation Language. IEEE Transactions on Knowledge and Data Engineering 6(1): 86-95.
- M. Egenhofer (1994b) Deriving the Composition of Binary Topological Relations. Journal of Visual Languages and Computing 5(2): 133-149.
- M. Egenhofer (1994c) Definitions of Line-Line Relations for Geographic Databases. Data Engineering 16(11): 479-481.
- M. Egenhofer and R. Franzosa (1995) On the Equivalence of Topological Relations. International Journal of Geographical Information Systems 9(2): 133-152.
- M. Egenhofer and D. Mark (1995a) Modeling Conceptual Neighborhoods of Topological Line-Region Relations. International Journal of Geographical Information Science 9(5): 555-565.
- M. Egenhofer and D. Mark (1995b) Naive Geography. in: A. Frank and W. Kuhn (Eds.), Spatial Information Theory: A Theoretical Basis for GIS, International Conference COSIT '95, Semmering, Austria, Lecture Notes in Computer Science, Vol. 988, pp. 1-15, Springer-Verlag.
- M. Egenhofer (1996) Spatial-Query-by-Sketch. in: M. Burnett and W. Citrin (Eds.), VL'96: IEEE Symposium on Visual Languages, Boulder, CO, pp. 60-67, IEEE Press.

- M. Egenhofer (1997) Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages and Computing* 8(4): 403-424.
- M. Egenhofer and R. Shariff (1998) Metric Details for Natural-Language Spatial Relations. *ACM Transactions on Information Systems* 16(4): 295-321.
- C. Elkan (1993) The Paradoxical Success of Fuzzy Logic. in: R. Fikes and W. Lehnert (Eds.), *Proceedings of the Eleventh National Conference on Artificial Intelligence*, Menlo Park, CA, pp. 698-703, AAAI Press.
- C. Elkan (2000) Paradoxes of Fuzzy Logic, Revisited. <http://citeseer.ist.psu.edu/elkan00paradoxes.html> Accessed: 04/29/2006.
- B. Ellis (1968) *Basic Concepts of Measurement*. Cambridge University Press, Cambridge, U.K.
- R. Elmasri and S. Navathe (2000) *Fundamentals of Database Systems* (3rd Edition). Addison Wesley Longman Inc., Reading, MA.
- D. Embley, C. Tao, and S. Liddle (2005) Automating the Extraction of Data from HTML Tables with Unknown Structure. *Data Knowledge Engineering* 54(1): 3-28.
- D. Ennis (1988) Confusable and Discriminable Stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General* 117(4): 408-411.
- R. Fagin (1998) Fuzzy Queries in Multimedia Database Systems. in: *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Seattle, Washington.
- A. Farquhar, R. Fikes, and J. Rice (1996) *The Ontolingua Server: A Tool for Collaborative Ontology Construction*. Knowledge Systems Laboratory, Stanford University, Stanford, CA, Technical Report KSL 96-26.
- A. Farquhar (1997) *Ontolingua Tutorial*. <http://www-ksl.stanford.edu/people/axf/tutorial.pdf> Accessed: 4/29/2006.

- D. Fensel (2000) *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, Berlin, Germany.
- D. Fensel, F. van Harmelen, I. Horrocks, D. McGuinness, and P. Patel-Schneider (2001) OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems* 16(2): 38-44.
- M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker (1995) Query by Image and Video Content: The QBIC System. *IEEE Computer* 28(9): 23-32.
- F. Fonseca, M. Egenhofer, C. Davis, and K. Borges (2000) Ontologies and Knowledge Sharing in Urban GIS. *Computer, Environment and Urban Systems* 24(3): 251-272.
- F. Fonseca (2001) *Ontology-Driven Geographic Information Systems*. Ph.D. Thesis, University of Maine, Orono, ME.
- F. Fonseca, M. Egenhofer, P. Agouris, and G. Camara (2002) Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS* 6(3): 231-257.
- A. Frank (1996) Qualitative Spatial Reasoning: Cardinal Directions as an Example. *International Journal of Geographical Information Systems* 10(3): 269-290.
- A. Frank (1998) Different Types of "Times" in GIS. in: R. Golledge and M. Egenhofer (Eds.), *Spatial and Temporal Reasoning in Geographic Information Systems*, pp. 40-62, Oxford University Press, New York, NY.
- C. Freksa (1991) Conceptual Neighborhood and its Role in Temporal and Spatial Reasoning. in: M. Singh and L. Travé-Massuyès (Eds.), *Proceedings of the IMACS Workshop on Decision Support Systems and Qualitative Reasoning*, North-Holland, Amsterdam, pp. 181-187, Elsevier Science Publishers.
- E. Freuder and R. Wallace (1992) Partial Constraint Satisfaction. *Artificial Intelligence* 58: 21-70.

- S. Freundschuh and M. Egenhofer (1997) Human Conceptions of Spaces: Implications for GIS. *Transactions in GIS* 2(4): 361-375.
- A. Galton (1995) Towards a Qualitative Theory of Movement. in: A. Frank and W. Kuhn (Eds.), *Spatial Information Theory: A Theoretical Basis for GIS*, International Conference COSIT '95, Semmering, Austria, *Lecture Notes in Computer Science*, Vol. 988, pp. 377-396, Springer-Verlag.
- A. Gangemi, D. Pisanelli, and G. Steve (1998) Ontology Integration: Experiences with Medical Terminologies. in: N. Guarino (Ed.) *Proceedings of Formal Ontology in Information Systems 1998 (FOIS '98)*, Trento, Italy, pp. 163-178, IOS Press, Amsterdam, The Netherlands.
- P. Gärdénfors (2000) *Conceptual Spaces: The Geometry of Thought*. The MIT Press, Cambridge, MA.
- I. Gati and A. Tversky (1984) Weighting Common and Distinctive Features in Perceptual and Conceptual Judgments. *Cognitive Psychology* 16(3): 341-370.
- D. Gentner (1983) Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science* 7(2): 155-170.
- D. Gentner (1988) Metaphor as Structure Mapping: The Relational Shift. *Child Development* 59(1): 47-59.
- J. Gibbons (1996) *Nonparametric Methods for Quantitative Analysis* (3rd Edition). American Sciences Press, Inc., Syracuse, NY.
- F. Godoy and A. Rodriguez (2002) A Quantitative Description of Spatial Configurations. in: P. van Oosterom (Ed.), *Spatial Data Handling*, pp. 299-311, Springer-Verlag, Ottawa, Canada 2002.
- A. Goldberg and R. Kennedy (1995) An Efficient Cost Scaling Algorithm for the Assignment Problem. *Mathematical Programming* 71(2): 153-177.



- R. Goldstone, D. Medin, and D. Gentner (1991) Relational Similarity and the NonIndependence of Features in Similarity Judgments. *Cognitive Psychology* 23(2): 222-262.
- R. Goldstone (1994a) Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(1): 3-28.
- R. Goldstone (1994b) The Role of Similarity in Categorization: Providing a Groundwork. *Cognition* 52(2): 125-157.
- R. Goldstone (1999) Similarity. in: R. Wilson and F. Keil (Eds.), *MIT Encyclopedia of the Cognitive Sciences*, pp. 763-765, The MIT Press, Cambridge, MA.
- R. Goldstone (2003) Learning to Perceive while Perceiving to Learn. in: R. Kimchi, M. Behrmann, and C. Olson (Eds.), *Perceptual Organization in Vision: Behavioral and Neural Perspectives*, Carnegie Mellon Symposia on Cognition Series, pp. 233-278, Lawrence Erlbaum Associates, Mahwah, NJ.
- R. Goldstone and J. Yun Son (2005) Similarity. in: K. Holyoak and R. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*, pp. 13-36, Cambridge University Press.
- R. L. Goldstone, D. Medin, and J. Halberstadt (1997) Similarity in Context. *Memory & Cognition* 25(2): 237-255.
- R. Gongalez and R. Woods (2002) *Digital Image Processing (2nd Edition)*. Prentice Hall.
- N. Goodman (1972) Seven Strictures on Similarity. in: N. Goodman (Ed.), *Problems and Projects*, pp. 23-32, Bobbs-Merrill, New York, NY.
- G. Gottlob and R. Zicari (1988) Closed World Databases Opened Through Null Values. in: F. Banchilhon and D. DeWitt (Eds.), *Proceedings of the Fourteenth International Conference on Very Large Data Bases*, Los Angeles, CA, pp. 50-61, Morgan Kaufmann.

- R. Goyal and M. Egenhofer (2001) Similarity of Cardinal Directions. in: C. Jensen, M. Schneider, B. Seeger, and V. Tsotras (Eds.), Proceedings of the Seventh International Symposium on Spatial and Temporal Databases, Los Angeles, CA, Lecture Notes in Computer Science, Vol. 2121, pp. 36-55, Springer-Verlag.
- J. Groff and P. Weinberg (2002) SQL: The Complete Reference (2nd Edition). McGraw-Hill Osborne Media.
- W. Grosky (1997) Managing Multimedia Information in Database Systems. Communications of the ACM 40(12): 72-80.
- M. Gross (1996) The Electronic Cocktail Napkin-Computer Support for Working with Diagrams. Design Studies 17(1): 53-69.
- T. Gruber (1992) A Translation Approach to Portable Ontology Specifications. Knowledge Systems Laboratory, Stanford University, Stanford, CA, Technical Report KSL 92-71.
- T. Gruber (1993) Toward Principles for the Design of Ontologies Used for Knowledge Sharing. in: N. Guarino and R. Poli (Eds.), Formal Ontology in Conceptual Analysis and Knowledge Representation, pp. 907-928, Kluwer Academic Publishers.
- M. Gruninger and J. Lee (2002) Ontology: Applications and Design. Communications of the ACM 45(2): 39-41.
- N. Guarino and P. Giaretta (1995) Ontologies and Knowledge Bases: Towards a Terminological Clarification. in: N. Mars (Ed.), Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, pp. 25-32, IOS Press, Amsterdam, The Netherlands.
- N. Guarino (1997) Understanding, Building, and Using Ontologies: A Commentary to "Using Explicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga. International Journal of Human and Computer Studies 46(2-3): 293-310.

- N. Guarino (1998) Formal Ontology and Information Systems. in: N. Guarino (Ed.) Proceedings of Formal Ontology in Information Systems 1998 (FOIS '98), Trento, Italy, pp. 3-15, IOS Press, Amsterdam, The Netherlands.
- N. Guarino, C. Masolo, and G. Vetere (1999) Ontoseek: Content-Based Access to the Web. *IEEE Intelligent Systems* 14(3): 70-80.
- N. Guarino and C. Welty (2000) A Formal Ontology of Properties. in: R. Dieng and O. Corby (Eds.), *Knowledge Engineering and Knowledge Management: Methods, Models and Tools: 12th International Conference, EKAW 2000*, Juan-les-Pins, France, *Lecture Notes in Artificial Intelligence*, Vol. 1937, pp. 97-112, Springer-Verlag.
- N. Guarino and C. Welty (2002) Evaluating Ontological Decisions with ONTOCLEAN. *Communications of the ACM* 45(2): 61-65.
- V. Gudivada and V. Raghavan (1995) Design and Evaluation of Algorithms for Image Retrieval by Spatial Similarity. *ACM Transactions on Information Systems* 13(1): 115-144.
- A. Guttman (1984) R-Trees: a Dynamic Index Structure for Spatial Searching. *ACM SIGMOD Record* 14(2): 47-57.
- V. Haarslev and M. Wessel (1997a) Querying GIS with Animated Spatial Sketches. in: *IEEE Symposium on Visual Languages*, Capri, Italy, pp. 197-204.
- V. Haarslev and M. Wessel (1997b) Querying GIS with Animated Spatial Sketches. in: *Proceedings of Symposium on Visual Languages (VL '97)*, Capri, Italy, pp. 197-204, IEEE Computer Society Press, Los Alamitos.
- U. Hahn and N. Chater (1997) Concepts and Similarity. in: K. Lamberts and D. Shanks (Eds.), *Knowledge, Concepts, and Categories*, pp. 43-92, The MIT Press.

- U. Hahn, L. Richardson, and N. Chater (2001) Similarity: a Transformational Approach. in: Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Edinburgh, Scotland.
- U. Hahn, N. Chater, and L. Richardson (2003) Similarity as Transformation. *Cognition* 87(1): 1-32.
- H. Haken and J. Portugali (2003) The Face of the City is its Information. *Journal of Environmental Psychology* 23(4): 385-408.
- F. Hakimpour and A. Geppert (2001) Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach. in: C. Welty and B. Smith (Eds.), *Proceedings of Formal Ontology in Information Systems 2001 (FOIS '01)*, Ogunquit, ME, pp. 297-308, ACM Press.
- R. Hamming (1950) Error Detecting and Error Correcting Codes. *Bell System Technical Journal* 26(2): 147-160.
- F. Harary (1969) *Graph Theory*. Addison-Wesley, Reading, MA.
- S. Harnad (Ed.) (1987) *Categorical Perception*. Cambridge University Press.
- M. Hearst (1994) Context and Structure in Automated Full-Text Information Access. Ph.D. Thesis, University of California at Berkeley, Berkeley, CA.
- D. Hernández (1994) Qualitative Representation of Spatial Knowledge. *Lecture Notes in Artificial Intelligence*, Vol. 804. Springer-Verlag, New York, NY.
- G. Hirst and D. Onge (1998) Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. in: C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 305-332, The MIT Press, Cambridge, MA.
- U. Hohenstein, L. Neugebauer, and G. Saake (1986) An Extended Entity-Relationship Model for Non-Standard Databases. Institut für Informatik, Technische Universität Clausthal, Lessach, Austria, Technical Report 3-86.

- C. Holsapple and K. Joshi (2002) A Collaborative Approach to Ontology Design. *Communications of the ACM* 45(2): 42-47.
- A. Holt (1999) Spatial Similarity and GIS: The Grouping of Spatial Kinds. in: P. Whigham (Ed.) *Proceedings of the Eleventh Annual Colloquium of the Spatial Information Research Centre (SIRC '99)*, Dunedin, New Zealand, pp. 241-250, University of Otago.
- K. Hornsby, M. Egenhofer, and P. Hayes (1999) Modeling Cyclic Change. in: P. Chen, D. Embley, J. Kouloumdjian, S. Liddle, and J. Roddick (Eds.), *Advances in Conceptual Modeling: Proceedings of the ER '99 Workshops on Evolution and Change in Data Management, Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling*, Paris, France, *Lecture Notes in Computer Science*, Vol. 1727, pp. 98-109, Springer-Verlag.
- K. Hornsby and M. Egenhofer (2000) Identity-Based Change: a Foundation for Spatio-Temporal Knowledge Representation. *International Journal of Geographical Information Science* 14(3): 207-224.
- J. Hosman and T. Kuennapas (1972) On the Relation between Similarity and Dissimilarity Estimates. University of Stockholm, Psychological Laboratories, Stockholm, Sweden Report No. 354.
- B. Huang and C. Claramunt (2005) Spatiotemporal Data Model and Query Language for Tracking Land Use Change. *Transportation Research Record*: 107-113.
- S. Imai (1977) Pattern Similarity and Cognitive Transformations. *Acta Psychologica* 41(6): 433-447.
- T. Imielinski and W. Lipski (1984) Incomplete Information in Relational Databases. *Journal of the Association for Computing Machinery* 31(4): 761-791.
- A. Isli and A. Cohn (1998) An Algebra for Cyclic Ordering of 2D Orientation. in: *Proceedings of the Fifteenth American Conference on Artificial Intelligence (AAAI)*, Madison, WI, pp. 643-649, AAAI Press/The MIT Press.

- P. Jaccard (1908) *Nouvelles Recherches sur la Distribution Florale*. Bulletin de la Societe de Vaud des Sciences Naturelles 44: 223-270.
- H. Jagadish, A. Mendelzon, and T. Milo (1995) *Similarity-Based Queries*. in: *Proceedings of Fourteenth Symposium on Principles of Database Systems (PODS'95)*, San Jose, CA, pp. 36-45, ACM Press.
- W. James (1890) *The Principles of Psychology*. Holt, New York, NY.
- J. Jia, G. Fischer, and J. Dyer (1998) *Attribute Weighting Methods and Decision Quality in the Presence of Response Error: a Simulation Study*. *Journal of Behavioral Decision Making* 11(2): 85-105.
- J. Jiang and D. Conrath (1997) *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. in: *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING X)*, Tapei, Taiwan, pp. 19-33.
- C. Jones, H. Alani, and D. Tudhope (2001) *Geographical Information Retrieval with Ontologies of Place*. in: D. Montello (Ed.) *Spatial Information Theory: Foundations of Geographic Information Science*, International Conference, COSIT 2001, Morro Bay, CA, *Lecture Notes in Computer Science*, Vol. 2205, pp. 322-335, Springer.
- D. Jones and R. Paton (1998) *Some Problems in the Formal Representation of Hierarchical Knowledge*. in: N. Guarino (Ed.) *Proceedings of Formal Ontology in Information Systems 1998 (FOIS '98)*, Trento, Italy, pp. 135-147, IOS Press, Amsterdam, The Netherlands.
- Y. Kalfoglou and M. Schorlemmer (2003) *Ontology Mapping: The State of the Art*. *The Knowledge Engineering Review* 18(1): 1-31.
- 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: Current Trends and Directions*, pp. 139-178, Academic Press, London, U.K.

- P. Kelly, M. Cannon, and D. Hush (1995) Query by Image Example: The CANDID Approach. in: W. Niblack and R. Jain (Eds.), SPIE Storage and Retrieval for Image and Video Databases III, San Jose, CA, Vol. 2420, pp. 238-248.
- S. Khoshafian and G. Copeland (1986) Object Identity. in: N. Meyrowitz (Ed.) Conference Proceedings on Object-Oriented Programming Systems, Languages and Applications, Portland, OR, pp. 406-416, ACM Press, New York, NY.
- W. Kim and J. Seo (1991) Classifying Schematic and Data Heterogeneity in Multidatabase Systems. IEEE Computer 24(12): 12-18.
- C. Kirkwood and R. Sarin (1985) Ranking with Partial Information: A Method and an Application. Operations Research 33(1): 38-48.
- A. Kiryakov, K. Simov, and M. Dimitrov (2001) Ontomap: Portal for Upper-Level Ontologies. in: C. Welty and B. Smith (Eds.), Proceedings of Formal Ontology in Information Systems 2001 (FOIS '01), Ogunquit, ME, pp. 47-58, ACM Press.
- F. Klingberg (1941) Studies in Measurement of the Relations among Sovereign States. Psychometrika 6(6): 335-352.
- A. Klippel and S. Winter (2005) Structural Saliency of Landmarks for Route Directions. in: A. Cohn and D. Mark (Eds.), Spatial Information Theory, International Conference, COSIT 2005, Ellicottville, NY, Lecture Notes in Computer Science, Vol. 3693, pp. 347-362, Springer.
- G. Klir and B. Yuan (1995) Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall.
- R. Korfhage (1997) Information Storage and Retrieval. John Wiley & Sons, Inc.
- N. Kosugi, Y. Nishihara, T. Sakata, M. Yamamuro, and K. Kushima (2000) A Practical Query-by-Humming System for a Large Music Database. in: Proceedings of the 8th ACM International Conference on Multimedia, Marina del Rey, CA, pp. 333-342, ACM Press.

- C. Krumhansl (1978) Concerning the Applicability of Geometric Models to Similarity Data: The Interrelationship Between Similarity and Spatial Density. *Psychological Review* 85(5): 445-463.
- J. Kruskal (1964) Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* 29(1): 1-27.
- H. Kuhn (1955) The Hungarian Method for the Assignment Problem. *Naval Research Logistic Quarterly* 2(1): 83-97.
- V. Kumar (1992) Algorithms for Constraint-Satisfaction Problems: A Survey. *AI Magazine* 13(1): 32-44.
- R. Kusters and A. Borgida (2001) What's in an Attribute? Consequences for the Least Common Subsumer. *Journal of Artificial Intelligence Research* 14: 167-203.
- 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.-R. Rollinger (Eds.), *Text Understanding in LILOG, Lecture Notes in Computer Science*, Vol. 546, pp. 464-481, Springer.
- R. Laurini and D. Thompson (1992) *Fundamentals of Spatial Information Systems*. A.P.I.C. Series, Vol. 37. Academic Press, San Diego, CA.
- 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.
- 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.
- S.-Y. Lee and F.-J. Hsu (1992) Spatial Reasoning and Similarity Retrieval of Images using 2D C-String Knowledge Representation. *Pattern Recognition* 25(3): 305-318.



- D. Lenat (1995) CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38(11): 32-38.
- D. Lenat, G. Miller, and T. Yokoi (1995) CYC, WordNet, and EDT: Critiques and Responses. *Communications of the ACM* 38(11): 45-48.
- V. Levenshtein (1965) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Akademii Nauk SSSR* 163(4): 845-848.
- D. Lewis (1986) *On the Plurality of Words*. Blackwell Publishers, Oxford, U.K.
- B. Li and F. Fonseca (2006) TDD: A Comprehensive Model for Qualitative Similarity Assessment. *Spatial Cognition and Computation* 6(1): 31-62.
- D. Lin (1998) An Information-Theoretic Definition of Similarity. in: J. Shavlik (Ed.) *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, Madison, WI, pp. 296-304, Morgan Kaufmann, San Francisco, CA.
- T. Lindeberg (1993) *Scale-Space Theory in Computer Vision*. The International Series in Engineering and Computer Science. Springer.
- W. Lipski (1979) On Semantic Issues Connected with Incomplete Information Databases. *ACM Transactions on Database Systems* 4(3): 262-296.
- Z. Liu and Q. Huang (2000) Content-Based Indexing and Retrieval-by-Example in Audio. in: *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo (ICME 2000)*, New York, NY, pp. 877-880.
- E. Loukakis and C. Tsouros (1981) A Depth First Search Algorithm to Generate the Family of Maximal Independent Sets of a Graph Lexicographically. *Journal of Computing* 27(4): 349-366.
- M. Lutz and E. Klien (2006) Ontology-Based Retrieval of Geographic Information. *International Journal of Geographical Information Science* 20(3): 233-260.
- K. Lynch (1960) *The Image of a City*. The MIT Press, Cambridge, MA.

- W. Maddox (1992) Perceptual and Decisional Separability. in: F. Ashby (Ed.), *Multidimensional Models of Perception and Cognition*, pp. 147-180, Lawrence Erlbaum Associates, Hillsdale, NJ.
- A. Maedche and S. Staab (2002) Measuring Similarity between Ontologies. in: A. Gómez-Pérez and R. Benjamins (Eds.), *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, 13th International Conference, EKAW 2002, Madrid, Spain, *Lecture Notes in Computer Science*, Vol. 2473, pp. 251-263, Springer.
- E. Marchiori (1998) A Simple Heuristic Based Genetic Algorithm for the Maximum Clique Problem. in: J. Carroll (Ed.) *Proceedings of the 1998 ACM symposium on Applied Computing*, Atlanta, GA, pp. 366-373, ACM Press, New York, NY.
- D. Mark and M. Egenhofer (1994) Calibrating the Meanings of Spatial Predicates from Natural Language: Line-Region Relations. in: T. Waugh and R. Healey (Eds.), *Proceedings of the Sixth International Symposium on Spatial Data Handling (SDH '94)*, Edinburgh, Scotland, pp. 538-553.
- A. Markman and D. Gentner (1993) Structural Alignment during Similarity Comparisons. *Cognitive Psychology* 25(4): 431-467.
- D. Marr (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt & Company.
- M. McCloskey (1983) Intuitive Physics. *Scientific American* 248(4): 122-130.
- D. McGuinness (1998) Ontological Issues for Knowledge-Enhanced Search. in: N. Guarino (Ed.) *Proceedings of Formal Ontology in Information Systems 1998 (FOIS '98)*, Trento, Italy, pp. 302-316, IOS Press, Amsterdam, The Netherlands.
- R. McKeon (Ed.) (2001) *The Basic Works of Aristotle*. Modern Library, New York, NY.

- E. Mena, A. Illarramendi, V. Kashyap, and A. Sheth (1996) OBSERVER: An approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. in: Proceedings of the First IFCIS International Conference on Cooperative Information Systems (CoopIS '96), Brussels, Belgium, pp. 14-25, IEEE Computer Society.
- 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 in Information Systems, Trento, Italy, pp. 269-283, IOS Press, Amsterdam, The Netherlands.
- S. Messick and R. Abelson (1956) The Additive Constant Problem in Multidimensional Scaling. *Psychometrika* 21(1): 1-15.
- B. Messmer and H. Bunke (1995) Subgraph Isomorphism in Polynomial Time. Institute of Computer Science and Applied Mathematics, University of Bern, Bern, Switzerland, Technical Report IAM 95-003.
- J. Mill (1829) Analysis of the Phenomenon of the Human Mind. Vol. 2. Baldwin and Cradock, London, U.K.
- G. Miller (1956) The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63(1): 81-97.
- G. Miller (1995) WordNet: A Lexical Database for English. *Communications of the ACM* 38(11): 39-41.
- G. Miller, K. Miller, C. Fellbaum, R. Teng, M. Hearst, K. Kohl, J. Douglas, R. Berwick, N. Nomura, U. Priss, S. Landes, C. Leacock, J. Grabowski, P. Resnik, M. Chodorow, E. Voorhees, G. Hirst, D. St-Onge, R. Al-Halimi, R. Kazman, J. Burg, S. Harabagiu, and D. Moldovan (1998) WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA.

- A. Monge and C. Elkan (1997) An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. in: Proceedings of the 1997 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD '97), Tucson, Arizona, pp. 23-29.
- P. Mork and P. Bernstein (2004) Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. in: Proceedings of the 20th International Conference on Data Engineering (ICDE 2004), Boston, MA, pp. 787-790, IEEE Computer Society.
- J. Morrissey (1990) Imprecise Information and Uncertainty in Information Systems. *ACM Transactions on Information Systems* 8(2): 159-180.
- F. Mosteller and R. Rourke (1973) *Sturdy Statistics: Nonparametric & Order Statistics*. Addison-Wesley, Menlo Park, CA.
- F. Mosteller and J. Tukey (1977) *Data Analysis and Regression*. Addison-Wesley.
- A. Motro (1988) VAGUE: A User Interface to Relational Databases that Permits Vague Queries. *ACM Transactions on Office Information Systems* 6(3): 187-214.
- M. Nabil, A. Ngu, and J. Shepherd (1996) Picture Similarity Retrieval using the 2D Projection Interval Representation. *IEEE Transactions on Knowledge and Data Engineering* 8(4): 533-539.
- K. Nedas, M. Egenhofer, and D. Wilmsen (in press) Metric Details for Topological Line-Line Relations. *International Journal of Geographical Information Science*.
- J. v. Neumann and O. Morgenstern (1947) *Theory of Games and Economic Behavior*. Princeton University Press.
- I. Niles and A. Pease (2001) Towards a Standard Upper Ontology. in: C. Welty and B. Smith (Eds.), *Proceedings of Formal Ontology in Information Systems 2001 (FOIS'01)*, Ogunquit, ME, pp. 2-9, ACM Press.

- R. Nosofsky (1986) Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General* 115(1): 39-57.
- R. Nosofsky (1991) Stimulus Bias, Asymmetric Similarity, and Classification. *Cognitive Psychology* 23(1): 94-140.
- R. Nosofsky (1992) Similarity Scaling and Cognitive Process Models. *Annual Review of Psychology* 43(1): 25-53.
- C. Nothegger, S. Winter, and M. Raubal (2004) Computation of the Saliency of Features. *Spatial Cognition and Computation* 4(2): 113-136.
- N. Noy and M. Musen (2003) The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping. *International Journal of Human-Computer Studies* 59(6): 983-1024.
- N. Noy (2004) Semantic Integration: A Survey of Ontology-Based Approaches. *ACM SIGMOD Record* 33(4): 65-70.
- M. Ortega, Y. Rui, K. Chakrabarti, K. Porkaew, S. Mehrotra, and T. S. Huang (1998) Supporting Ranked Boolean Similarity Queries in MARS. *IEEE Transactions on Knowledge and Data Engineering* 10(6): 905-925.
- M. Ortega-Binderberger, K. Chakrabarti, and S. Mehrotra (2002) An Approach to Integrating Query Refinement in SQL. in: C. Jensen, K. Jeffery, J. Pokorn, S. Saltenis, E. Bertino, K. Böhm, and M. Jarke (Eds.), *Advances in Database Technology - EDBT 2002, 8th International Conference on Extending Database Technology, Prague, Czech Republic, Lecture Notes in Computer Science, Vol. 2287*, pp. 15-33, Springer.
- J. Paiva (1998) Topological Equivalence and Similarity in Multiple Representation Geographic Databases. Ph.D. Thesis, University of Maine, Orono, ME.

- L. Palopoli, D. Saccà, G. Terracina, and D. Ursino (2003) Uniform Techniques for Deriving Similarities of Objects and Subschemes in Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering* 15(2): 271-294.
- D. Papadias and M. Egenhofer (1996) Algorithms for Hierarchical Spatial Reasoning. *GeoInformatica* 1(3): 251-273.
- D. Papadias and V. Delis (1997) Relation-Based Similarity. in: *Proceedings of the 5th International Workshop on Advances in Geographic Information Systems (ACM GIS '97)*, Las Vegas, NV, pp. 1-4, ACM Press.
- D. Papadias, N. Mamoulis, and V. Delis (1998a) Algorithms for Querying by Spatial Structure. in: A. Gupta, O. Shmueli, and J. Widom (Eds.), *VLDB '98, Proceedings of the 24th International Conference on Very Large Data Bases*, New York, NY, pp. 546-557, Morgan Kaufmann.
- D. Papadias, N. Mamoulis, and D. Meretakis (1998b) Image Similarity Retrieval by Spatial Constraints. in: G. Gardarin, J. French, N. Pissinou, K. Makki, and L. Bouganim (Eds.), *Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management*, Bethesda, MD, pp. 289-296, ACM Press.
- D. Papadias, P. Kalnis, and N. Mamoulis (1999a) Hierarchical Constraint Satisfaction in Spatial Databases. in: *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI '99)*, Orlando, FL, pp. 142-147, AAAI Press, Menlo Park, CA.
- D. Papadias, N. Karacapilidis, and D. Arkoumanis (1999b) Processing Fuzzy Spatial Queries: A Configuration Similarity Approach. *International Journal of Geographical Information Science* 13(2): 93-128.
- D. Papadias, M. Mantzourogianis, P. Kalnis, N. Mamoulis, and I. Ahmad (1999c) Content-Based Retrieval Using Heuristic Search. in: *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, pp. 168-175, ACM Press.

- D. Papadias (2000) Hill Climbing Algorithms for Content-based Retrieval of Similar Configurations. in: N. Belkin, P. Ingwersen, and M.-K. Leong (Eds.), SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 240-247, ACM Press.
- D. Papadias, N. Mamoulis, and V. Delis (2001) Approximate Spatio-Temporal Retrieval. ACM Transactions on Information Systems 19(1): 53-96.
- D. Papadias, M. Mantzourogiannis, and I. Ahmad (2003) Fast Retrieval of Similar Configurations. IEEE Transactions on Multimedia 5(2): 210-222.
- C. Papadimitriou and K. Steiglitz (1998) Combinatorial Optimization: Algorithms and Complexity. Dover Publications.
- A. Parducci (1965) Category Judgment: A Range-Frequency Model. Psychological Review 72(6): 407-418.
- J. Park and S. Ram (2004) Information Systems Interoperability: What Lies Beneath? ACM Transactions on Information Systems 22(4): 595-632.
- S. Patwardhan (2003) Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. M.S. Thesis, University of Minnesota, Duluth, MN.
- E. Petrakis and C. Faloutsos (1997) Similarity Searching in Medical Image Databases. IEEE Transactions on Knowledge and Data Engineering 9(3): 435-447.
- K. Popper (1972) The Logic of Scientific Discovery. Hutchinson, London, U.K.
- D. Pyle (1999) Data Preparation for Data Mining. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.
- R. Quillian (1968) Semantic Memory. in: M. Minsky (Ed.), Semantic Information Processing, pp. 216-270, The MIT Press, Cambridge, MA.

- W. Quine (1969) *Ontological Relativity & Other Essays*. Columbia University Press, New York, NY.
- R. Rada, H. Mili, E. Bicknell, and M. Blettner (1989) Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics* 19(1): 17-30.
- E. Rahm and P. Bernstein (2001) A Survey of Approaches to Automatic Schema Mapping. *The VLDB Journal* 10(4): 334-350.
- M. V. Ramakrishna, S. Nepal, and D. Srivastava (2002) A Heuristic for Combining Fuzzy Results in Multimedia Databases. in: *Proceedings of the Thirteenth Australasian Conference on Database Technologies-Volume 5*, Melbourne, Victoria, Australia, pp. 141 - 144, Australian Computer Society, Inc.
- D. Randell, Z. Cui, and A. Cohn (1992) A Spatial Logic based on Regions and Connection. in: B. Nebel, C. Rich, and W. Swartout (Eds.), *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR '92)*, Cambridge, MA, pp. 165-176, Morgan Kaufmann.
- A. Rapoport and S. Fillenbaum (1972) Experimental Studies of Semantic Structure. in: R. Shepard, A. Romney, and S. Nerlove (Eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. 2, pp. 93-131, Seminar Press, New York, NY.
- A. Rector, W. Nowlan, and A. Glowinski (1993) Goals for Concept Representation in the GALEN Project. in: *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, Washington, DC, pp. 414-418, McGraw-Hill.
- P. Resnik (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. in: C. Mellish (Ed.) *IJCAI '95. Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, pp. 448-453, Morgan Kaufmann.



- P. Resnik (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11: 95-130.
- M. Richardson (1938) Multidimensional Psychophysics. *Psychological Bulletin* 35: 659-660.
- M. Richter (1992) Classification and Learning of Similarity Measures. in: C. Opiz and L. Klar (Eds.), *Proceedings of the 16th Annual Meeting of the German Society for Classification, Studies in Classification, Data Analysis and Knowledge Organization*, Kaiserslautern, Germany, pp. 1-8, Springer-Verlag.
- L. Rips and E. Shoben (1973) Semantic Distance and the Verification of Semantic Relations. *Journal of Verbal Learning and Verbal Behavior* 12: 1-20.
- L. Rips and A. Collins (1993) Categories and Resemblance. *Journal of Experimental Psychology: General* 122(4): 468-486.
- A. Rodríguez, Egenhofer, and R. Rugg (1999) Assessing Semantic Similarities among Geospatial Feature Class Definitions. in: A. Vckovski, K. Brassel, and H.-J. Schek (Eds.), *Proceedings of the Second International Conference on Interoperating Geographic Information Systems*, Zurich, Switzerland, *Lecture Notes in Computer Science*, Vol. 1580, pp. 189-202, Springer-Verlag, London, U.K.
- A. Rodríguez and M. Egenhofer (1999) Putting Similarity Assessments into Context: Matching Functions with the User's Intended Operations. in: P. Bouquet, L. Serafini, P. Brezillon, and F. Castellani (Eds.), *Modeling and Using Context, Second International and Interdisciplinary Conference, CONTEXT '99*, Trento, Italy, *Lecture Notes in Artificial Intelligence*, Vol. 1688, pp. 310-323, Springer-Verlag.
- A. Rodríguez (2000) Assessing Semantic Similarity Among Spatial Entity Classes. Ph.D. Thesis, University of Maine, Orono, ME.

- A. Rodríguez and M. Egenhofer (2003) Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering* 15(2): 442-456.
- A. Rodríguez and M. Egenhofer (2004) Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science* 18(3): 229-256.
- A. Rodríguez and M. Jarur (2005) A Genetic Algorithm for Searching Spatial Configurations. *IEEE Transactions on Evolutionary Computation* 9(3): 252-270.
- E. Rosch (1975) Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General* 104(3): 192-233.
- A. Rosenthal, L. Seligman, and S. Renner (2004) From Semantic Integration to Semantics Management: Case Studies and a Way Forward. *ACM SIGMOD Record* 33(4): 44-50.
- S. Ross (1976) *A First Course in Probability*. Macmillan.
- N. Roussopoulos, S. Kelley, and F. Vincent (1995) Nearest Neighbor Queries. in: M. Carey and D. Schneider (Eds.), *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, San Jose, CA, pp. 71-79, ACM Press.
- B. Russell (1920) *Introduction to Mathematical Philosophy*. Dover Publications.
- B. Russell (1938) *Principles of Mathematics*. W. W. Norton & Company, New York, NY.
- Z. Ruttkay (1994) Fuzzy Constraint Satisfaction. in: *Proceedings of the Third IEEE Conference on Fuzzy Systems: IEEE World Congress on Computational Intelligence*, Orlando, FL, pp. 542-547.
- G. Salton, A. Wong, and C. Yang (1975) A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18(11): 613-620.

- G. Salton, E. Fox, and H. Wu (1983) Extended Boolean Information Retrieval. *Communications of the ACM* 26(11): 1022-1036.
- S. Santini and R. Jain (1996) Similarity Queries in Image Databases. in: *Proceedings of the International IEEE Computer Vision and Pattern Recognition Conference (CVPR '96)*, San Francisco, CA, pp. 646-651.
- S. Santini and J. Ramesh (1997) The Graphical Specification of Similarity Queries. *Journal of Visual Languages and Computing* 7(4): 403-421.
- S. Santini and J. Ramesh (2000) Integrated Browsing and Querying for Image Databases. *IEEE Multimedia* 7(3): 26-39.
- S. Sattath and A. Tversky (1977) Additive Similarity Trees. *Psychometrika* 42(3): 319-345.
- V. Schenkelaars and M. Egenhofer (1997) Exploratory Access to Geographic Libraries. in: *Autocarto 13*, Seattle, WA.
- J. Schumacher and R. Bergmann (2000) An Efficient Approach to Similarity-Based Retrieval on Top of Relational Databases. in: E. Blanzieri and L. Portinale (Eds.), *Advances in Case-Based Reasoning, 5th European Workshop, EWCBR 2000*, Trento, Italy, Vol. 1898, pp. 273-284, Springer.
- C. Shannon (1948) A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379-423 & 623-656.
- L. Shapiro and R. Haralick (1981) Structural Descriptions and Inexact Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3(9): 504-519.
- R. Shariff (1996) Natural-Language Spatial Relations: Metric Refinements of Topological Properties. Ph.D. Thesis, University of Maine, Orono, ME.
- R. Shepard (1962a) The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. *Psychometrika* 27(2): 125-140.

- R. Shepard (1962b) The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II. *Psychometrika* 27(3): 219-246.
- R. Shepard (1987) Toward a Universal Law of Generalization for Psychological Science. *Journal of Science* 237(4820): 1317-1323.
- R. Shepard (1988) Time and Distance in Generalization and Discrimination: Comment on Ennis (1988). *Journal of Experimental Psychology: General* 117(4): 415-416.
- A. Sheth and J. Larson (1990) Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys* 22(3): 183-236.
- A. Sheth (1995) Data Semantics: What, Where and How? Department of Computer Science, University of Georgia, Athens, GA, Technical Report TR-CS-95-003.
- P. Sistla, C. Yu, and R. Venkatasubrahmanian (1997) Similarity Based Retrieval of Videos. in: W. Gray and P.-Å. Larson (Eds.), *Proceedings of the Thirteenth International Conference on Data Engineering (ICDE '97)*, Birmingham, U.K., pp. 181-190, IEEE Computer Society.
- L. Sjöberg (1972) A Cognitive Theory of Similarity. *Psykologiska Institutionen, Göteborg University, Göteborg, Sweden, Göteborg Psychological Reports* 2(10).
- S. Sloman, B. Love, and W.-K. Ahn (1998) Feature Centrality and Conceptual Coherence. *Cognitive Science* 22(2): 189-228.
- B. Smith (1995) On Drawing Lines on a Map. in: A. Frank and W. Kuhn (Eds.), *Spatial Information Theory: A Theoretical Basis for GIS*, International Conference COSIT '95, Semmering, Austria, *Lecture Notes in Computer Science*, Vol. 988, pp. 475-484, Springer-Verlag.
- B. Smith and C. Welty (2001) Ontology: Towards a New Synthesis. in: C. Welty and B. Smith (Eds.), *Proceedings of Formal Ontology in Information Systems 2001 (FOIS '01)*, Ogunquit, ME, pp. 3-9, ACM Press.

- J. Smith and S.-F. Chang (1996) VisualSEEK: A Fully Automated Content-Based Image Query System. in: Proceedings of the Forth ACM International Conference on Multimedia, Boston, MA, pp. 87-98, ACM Press.
- L. Smith and D. Heise (1992) Perceptual Similarity and Conceptual Structure. in: B. Burns (Ed.), Percepts, Concepts, and Categories: Representation and Processing of Information, Advances in Psychology, Vol. 93, pp. 233-272, Elsevier.
- J. Sneed (1971) The Logical Structure of Mathematical Physics. Reidel, Dordrecht, The Netherlands.
- P. Spyns, R. Meersman, and M. Jarrar (2002) Data Modelling versus Ontology Engineering. ACM SIGMOD Record 31(4): 12-17.
- S. Staab and R. Studer (Eds.) (2004) Handbook on Ontologies. in International Handbooks on Information Systems. Springer.
- A. Stefanidis, P. Agouris, M. Bertolotto, J. Carswell, and C. Georgiadis (2002) Scale and Orientation-Invariant Scene Similarity Metrics for Image Queries. International Journal of Geographical Information Science 16(8): 749-772.
- S. Stevens (1946) On the Theory of Scales of Measurement. Journal of Science 103(2684): 677-680.
- S. Stevens (1951) Mathematics, Measurement, and Psychophysics. in: S. Stevens (Ed.), Handbook of Experimental Psychology, pp. 1-49, John Wiley & Sons, Inc., New York, NY.
- W. Stillwell, D. Seaver, and W. Edwards (1981) A Comparison of Weight Approximation Techniques in Multiattribute Utility Decision Making. Organizational Behavior and Human Performance 28: 62-77.

- M. Sussna (1993) Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network. in: B. Bhargava, T. Finin, and Y. Yesha (Eds.), *CIKM 93, Proceedings of the Second International Conference on Information and Knowledge Management*, Washington, DC, pp. 67-74, ACM Press.
- Y. Takane and T. Shibayama (1992) Structures in Stimulus Identification Data. in: F. Ashby (Ed.), *Probabilistic Multidimensional Models of Perception and Cognition*, pp. 335-362, Earlbaum, Hillsdale, NJ.
- W. Tobler (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46(Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods): 234-240.
- E. Tomita, A. Tanaka, and H. Takahashi (1988) The Worst-Case Time Complexity for Generating All Maximal Cliques. in: K.-Y. Chwa and J. Munro (Eds.), *Computing and Combinatorics: 10th Annual International Conference, COCOON 2004*, Jeju Island, Korea, *Lecture Notes in Computer Science*, Vol. 3106, pp. 161-170, Springer.
- W. Torgerson (1952) Multidimensional Scaling: I. Theory and Method. *Psychometrika* 17(4): 401-419.
- W. Torgerson (1958) *Theory and Methods of Scaling* (4th Edition). John Wiley & Sons, Inc., New York, NY.
- W. Torgerson (1965) Multidimensional Scaling of Similarity. *Psychometrika* 30(4): 379-393.
- A. Tversky (1977) Features of Similarity. *Psychological Review* 84(4): 327-352.
- B. Tversky, J. Zacks, P. Lee, and J. Heiser (2000) Lines, Blobs, Crosses and Arrows: Diagrammatic Communication with Schematic Figures. in: V. Haarslev (Ed.), *Theory and Application of Diagrams*, *Lecture Notes in Computer Science*, Vol. 1889, pp. 221-300, Springer, Berlin.

- S. Ullman (2000) *High-Level Vision: Object Recognition and Visual Cognition*. The MIT Press, Cambridge, MA.
- C. Umiltà, S. Bagnara, and F. Simion (1978) Laterality Effects for Simple and Complex Geometrical Figures, and Nonsense Patterns. *Neuropsychologica* 16(1): 43-49.
- M. Uschold and M. Gruninger (2004) Ontologies and Semantics for Seamless Connectivity. *ACM SIGMOD Record* 33(4): 58-64.
- USGS (1998) View of the Spatial Data Transfer Standard (SDTS). [http://mcmcweb.er.usgs.gov/sdts/SDTS\\_standard\\_nov97/p2toc.html](http://mcmcweb.er.usgs.gov/sdts/SDTS_standard_nov97/p2toc.html) Accessed: 05/03/2006.
- Y. Vassiliou (1979) Null Values in Data Base Management: A Denotational Semantics Approach. in: P. Bernstein (Ed.) *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data*, Boston, MA, pp. 162-169, ACM Press, New York, NY.
- P. Velleman and L. Wilkinson (1993) Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. *The American Statistician* 47(1): 65-72.
- H. Wache, T. Voegelé, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Huebner (2001) Ontology-Based Integration of Information: A Survey of Existing Approaches. in: A. Gómez-Pérez, M. Gruninger, H. Stuckenschmidt, and M. Uschold (Eds.), *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, Seattle, WA, pp. 108-118, CEUR-WS.org.
- Y. Wang, F. Makedon, J. Ford, and H. Huang (2004) A Bipartite Graph Matching Framework for Finding Correspondences between Structural Elements in Two Proteins. in: *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, pp. 2972-2975.
- C. Welty (2000) Towards a Semantics for the Web. in: *Dagstuhl Symposium on Semantics for the Web*, Dagstuhl, Germany.

- D. White and R. Jain (1996) Similarity Indexing with the SS-tree. in: S. Su (Ed.) Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, LA, pp. 516-523, IEEE Computer Society.
- R. Wilson and T. Martinez (1997) Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research 6: 1-34.
- S. Winter (2003) Route Adaptive Selection of Salient Features. in: W. Kuhn, M. Worboys, and S. Timpf (Eds.), Spatial Information Theory. Foundations of Geographic Information Science, International Conference, COSIT 2003, Ittingen, Switzerland, Lecture Notes in Computer Science, Vol. 2825, pp. 349-361, Springer, Berlin, Germany.
- S. Wong, W. Ziarko, V. Raghavan, and P. Wong (1987) On Modeling of Information Retrieval Concepts in Vector Spaces. ACM Transactions on Database Systems 12(2): 299-321.
- W. Wood (1975) What's in a Link: Foundations for Semantic Networks. in: D. Bobrow and A. Collins (Eds.), Representation and Understanding Studies in Cognitive Science, pp. 35-82, Academic Press, New York, NY.
- M. Worboys (2001) Nearness Relations in Environmental Space. International Journal of Geographical Information Science 15(7): 633-652.
- M. Worboys and M. Duckham (2004) GIS: A Computing Perspective (2nd Edition). CRC Press, Boca Raton, FL.
- M. Worboys, M. Duckham, and L. Kulik (2004) Commonsense Notions of Proximity and Direction in Environmental Space. Spatial Cognition and Computation 4(4): 285-312.
- Y. Wu, Y. Zhuang, and Y. Pan (2000) Content-Based Video Similarity Model. in: Proceedings of the 8th ACM International Conference on Multimedia, Marina del Rey, CA, pp. 465-467, ACM Press.



- Z. Wu and M. Palmer (1994) Verb Semantics and Lexical Selection. in: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, pp. 133-138, Morgan Kaufmann.
- Y. Yanwu and C. Claramunt (2003) A Process-Oriented Multi-Representation of Gradual Changes. *Journal of Geographic Information and Decision Analysis* 7(1): 1-13.
- G. Young and A. Householder (1938) Discussion of a Set of Points in terms of their Mutual Distances. *Psychometrika* 3(1): 19-22.
- L. Zadeh (1965) Fuzzy Sets. *Information and Control* 8(3): 338-353.
- C. Zaniolo (1982) Database Relations with Null Values. in: Proceedings of the ACM Symposium on Principles of Database Systems, Los Angeles, CA, pp. 27-33, ACM Press, New York, NY.
- B.-T. Zhang and S.-Y. Shin (1998) Code Optimization for DNA Computing of Maximal Cliques. in: J. Benitez, O. Cordon, F. Hoffmann, and R. Roy (Eds.), *Advances in Soft Computing: Engineering Design and Manufacturing*, pp. 735-742, Springer.
- J. Zobel and P. Dart (1996) Phonetic String Matching: Lessons from Information Retrieval. in: H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, Zurich, Switzerland*, pp. 166-172, ACM Press.
- G. Zuniga (2001) Ontology: Its Transformation from Philosophy to Information Systems. in: C. Welty and B. Smith (Eds.), *Proceedings of Formal Ontology in Information Systems 2001 (FOIS '01)*, Ogunquit, ME, pp. 187-197, ACM Press.

## **BIOGRAPHY OF THE AUTHOR**

Konstantinos A. Nedas was born in Thessaloniki, Greece on June 2, 1976. He was raised in Thessaloniki and Rhodes, both in Greece. Konstantinos entered the Department of Rural and Surveying Engineering of the Polytechnic School of the Aristotle University of Thessaloniki in the fall of 1994, by ranking first in the panhellenic exams. He obtained his diploma in 2000, graduating with high distinction. During his studies, Konstantinos worked for one year as a surveyor in private companies, while in parallel performing duties as website administrator for the Department of Rural and Surveying Engineering. In the fall of 2000, he entered the Ph.D. program in Spatial Information Science and Engineering at the University of Maine. He is currently a graduate research assistant with the National Center for Geographic Information and Analysis. Konstantinos is a candidate for the Doctor of Philosophy degree in Spatial Information Science and Engineering from the University of Maine in August, 2006.