Measuring Similarities of Spatial Datasets

Douglas M. Flewelling

Abstract: *The rapidly increasing availability of digital spatial datasets through on-line media generates a new demand for assessing and comparing spatial datasets with computational tools. Current geographic database systems are tailored to retrieving configurations based on the spatial relations among individual entities. Such one-by-one measures, however, do not scale-up, neither computationally nor cognitively, when applied to entire datasets within digital libraries or data warehouses. A particular challenge for users of the Internet is to find spatial datasets that are similar to a given dataset. To enable the comparison of spatial datasets, new computational measures are needed. We discuss the required properties of such spatial similarity measures based on their adherence to rules for deductive databases.*

1. Introduction

The World-Wide Web as operational in the late 1990s is strongly text-based, with search engines supporting the retrieval of documents based on the frequency of keywords. Finding spatial datasets does not fit into this setting, because it makes little sense to search for or compare strings of coordinates in order to match a spatial target configuration with spatial datasets that are available on-line. The addition of metadata, often promoted as the mechanism essential to select the right datasets, moves what should be a spatial search into the domain of textual keywords. While metadata may provide ways to describe datasets concisely, they are often subjective descriptors given by a data provider that may not necessarily fit the various needs of the data users (Flewelling and Egenhofer 1999). In order to enable the design of new search engines that would take into consideration spatial criteria, it is necessary to gain a better understanding of how to assess analytically differences among spatial datasets.

This paper is concerned with the comparison of spatial datasets that are assumed to share common elements such that there is a direct mapping between objects in the sets. While some of the measures presented here could be used to compare sets with different types of elements, it is beyond the focus of this research. In order to assess similarity it is necessary to perform a difference operation over the set attribute measures for each pair of spatial datasets. The ability to assess similarity will provide significant advantages in searching very large data collections from multiple sources, such as data warehouses (Garcia-Molina *et al.* 1995) and digital libraries (Smith 1996), where the likelihood of finding equivalence to a query is low.

The development of computational methods for the assessment of spatial-set similarity differs from approaches that are based on the comparison of individual objects. Algorithms to retrieve objects with similar shapes (Mehrotra and Gray 1995) or similar spatial relations between lines (Gudivada and Raghavan 1995) or objects (Bruns and Egenhofer 1996) are tailored to extracting individuals or pairs of individuals from a spatial dataset. In computer vision and image processing, content-based image retrieval relies on similarity measures that represent histograms of representative values of spectral values (Flickner *et al.* 1995; Faloutsos *et al.* 1994). On the other hand, geographers have for a long time investigated methods to describe similarity of point sets for spatial analyses, addressing such properties as pattern, density, and dispersion (Unwin 1981).

This paper continues with definitions of a model for spatial objects, spatial sets, and their respective attributes (Section 2). In Section 3 a model for comparing spatial datasets through similarity, equivalence, and identity is defined. Section 4 discusses the classes of content in spatial datasets. Section 5 investigates scales of data and the limits those scales have on analysis of set attributes. It continues to define formal methods for assessing similarity for all the data measurement scales. A formal approach to the combination of individual similarity measures into a single *simi-*

Douglas Flewelling received his Ph.D. in Spatial Information Science and Engineering from the University of Maine in 1997, a MS in Geography from Oregon State University in 1984, and BA in Geography from Syracuse University in 1981. He is a Project Manager with the National Center for Geographic Information and Analysis at the University of Maine. His research interests are currently similarity-based spatial data retrieval and GIS interoperability with discrete and continuous data models.

larity index value is defined. The similarity index can be used for evaluating set similarity among three or more spatial datasets. Finally, the requirements for appropriate spatial measures as defined by the model presented here are summarized in Section 6.

2. A Model of Spatial Datasets

In order to compare spatial datasets, it is necessary to define a limited number of terms that form the components of a model of spatial datasets. In their work on types and data abstractions, Cardelli and Wegner (1985) laid out a terminology for sets and types. We expand on their definitions to introduce four key concepts for comparing spatial datasets: spatial object, spatial class, spatial set, and spatial ideal.

2.1 Spatial Objects

In the domain of geography, scientists are concerned with instances of *spatial objects*. Spatial objects are things with an identity and identifiable characteristics, among which is location. Some spatial objects are tangible things, such as trees or buildings, while others are concepts such as counties or sales districts. Spatial objects, such as buildings, can be aggregates of other spatial objects (doors, windows, walls), but the atomic spatial objects in the domain of geography are those large enough to be observable by the human eye and smaller than planetary scale. These bounds exclude both the microscopic and the astronomic (Mark and Freundschuh 1995).

2.2 Spatial Classes

People have developed a set of cognitive processes that reduce the complexity of an unstructured world. From a very early age people begin to develop *schemata* for the important parts of our world (Pinker 1990a; Johnson 1987; Eastman 1985; Neisser 1976). These schemata or *class definitions* provide structures upon which to organize the phenomena they observe and form the basis of the *classification* of spatial objects. For instance, a class definition for trees might include: location, has_woody_trunk, can_burn, has_leaves, and has_branches. A book would satisfy three of the functions, but not the others and is, therefore, not a tree. A spatial class has a class definition that supports, among others, the *location* function.

Spatial classes may restrict the range of values that may apply to a particular class of spatial objects. When a class supports all of the functions of another class, but is more restrictive on values or adds other functions, we call it a *subclass*. While a pine tree is certainly a tree, the needles and cones make it clearly a conifer rather than a deciduous tree. All functions of the superclass must be supported by a subclass, but subclass definitions need not be mutually exclusive (Smith and Smith 1977). For instance, a spatial object in the Building class could also be a member of both the subclasses *home* and *office*.

A particular object may be classified into more than one class depending on the attribute values it possesses. Within limited domains it is possible to construct hierarchies of classes that exclusively classify a group of objects. These *taxonomies* provide rules to classify any individual into an indivisible group. The applicability of a taxonomy is limited to a specific user group. For instance, while a particular stand of trees might be classified as a *mixed-growth habitat* by a forester, a child may classify the same stand of trees as a *playground*.

2.3 Spatial Sets

If we gather all the spatial objects of interest together we have defined a geographic universe for our problem domain. In this paper, this universe of spatial objects is denoted as G. In the universe G it is possible to create groups of spatial objects, called *spatial sets*, by collecting one or more spatial objects together and identifying them with a symbol or name. For instance, the Eiffel Tower, Moosehead Lake, and a maple tree are elements of a spatial set called Green. A spatial set does not have to have specific rules of membership, however, without membership rules a very large, but finite number of sets could be generated within G.

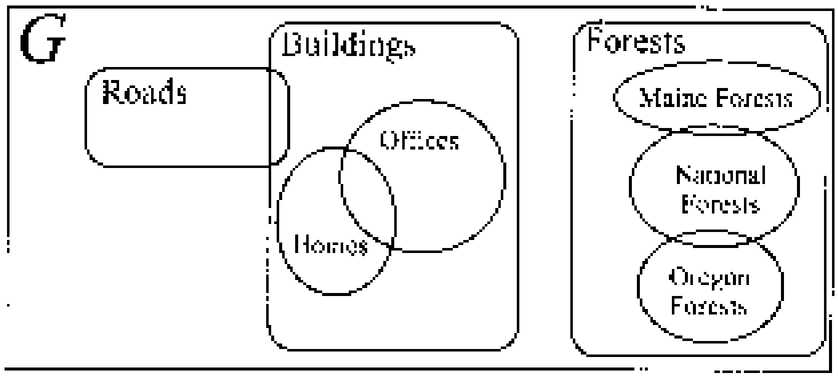
Creating a personal classification of the world with one's own symbols is of limited use. Fortunately, within a given culture people share class definitions with others and negotiate a common symbol set, which we call a language (Pinker 1990b; Zurif 1990). With a language, the connection between the symbol and the class is taught through a variety of means including example, simile, and in limited cases complete enumeration (Lakoff 1988; Johnson 1987).

2.4 Spatial Ideals

If all elements of a spatial set satisfy a specific class definition the set is called a *spatial ideal*. A spatial ideal does not necessarily contain *all* the instances in G that match the class definition. Therefore, any number of spatial ideals could share the same spatial class definition. The spatial ideal that contains all the spatial objects of that class is called the *universal spatial ideal* for that class. In Figure 1, the universal spatial ideal *forests* contains all the forests in G. The spatial ideals *Maine forests* and *Oregon forests* are both subclasses of *forests*, which have a spatial restriction in their respective class definitions. For practical purposes, it is often impossible to completely populate a universal spatial ideal with *all* the spatial objects that match a specific class definition. Without the universal spatial ideal fully populated it is impossible to reason absolutely about the objects in the ideal.

Gallaire et al. (1984) identified three basic assumptions upon which deductive reasoning in databases depend: (1) the closed world assumption, (2) the unique name assumption, and (3) the domain closure assumption. The *closed world*

Figure 1: *Spatial ideals and universal spatial ideals (Roads, Buildings and Forests). All spatial objects in a spatial ideal share a single class definition, but spatial ideals are not necessarily mutually exclusive.*



assumption states that facts that are not known to be true are assumed to be false. This assumption is required to support the use of Boolean logic in database systems. The *unique name assumption* states that objects with different names are different. This assumption is necessary to support uniqueness for entities in a database system. Finally, the *domain closure assumption* states that there are no other entities than those in the database. Without this assumption nothing definitive could be stated about the contents of the database because other objects might be found to invalidate the information derived from the database.

In practical terms, a spatial database system that operates with the domain closure assumption is working with universal spatial ideals. The use of universal spatial ideals has the advantage of permitting consistent reasoning over the broadest range of data scales. For instance, it is impossible to generate summary attributes about ordinal class attributes in spatial sets unless the sets are both subsets of the same universal spatial ideal.

3. Comparing Spatial Sets

In order to compare spatial sets, it is necessary to agree on the meaning of the term *equal*, which is often used in colloquial English for several different concepts about the relationship between two spatial datasets. For this purpose, three different forms of equal are defined called identical, equivalent, and similar.

3.1 Identity of Spatial Sets

Two sets A and B are *identical* if each and every member of A is a member of B and each and every member of B is a member of A. This is the traditional definition of set equality (Burington 1973). In addition, it is assumed that A and B are subsets of the same superset S. In effect, A is a true and faithful copy of B. It is possible for both A and B to be equal to S.

3.2 Equivalence of Spatial Sets

Equivalence is a mapping between two sets of objects such that there is a one-to-one relationship between all the members of set A and set B. There is no assumption of a common physical superset for two sets to be equivalent. While two identical sets are equivalent, the reverse is not necessarily true. Equivalence can also be achieved by transforming the elements of one set through a view function to be *view equivalent*. For instance, if set A is composed of cities with a location and population as class attributes and set B is composed of the same cities with location, population, and mean income for class attributes, a view B' could be generated over set B as a projection of the attributes location and population that is equivalent to set A. In this case set B is view equivalent to set A (Equation 1).

$$Given: A\{location.pop\}, B\{location.pop.income\}$$
$$A \cong B' \Leftarrow \pi_{location.pop}(B)$$

3.3 Similarity of Spatial Sets

When two datasets are neither identical nor equivalent they are different. Depending on how much they differ, they may expose various degrees of similarity. In order for two datasets to be similar there must be some shared properties between the two datasets, but not necessarily shared elements. For instance, a set of water treatment plants in Maine and a set of post offices in Maine may have similar patterns and dispersions within the State, yet there are no post offices that are also water treatment plants.

Any measure of similarity is in its essence a distance measure, that is, a measure of the difference between a dataset and a target dataset (Tversky 1977). This distance concept is counterintuitive to the normal usage of similarity. If two datasets have a high similarity, their difference is small. When the difference between two datasets is zero, as assessed by some criteria, they are as highly similar as possible (100% the same). When the datasets are "highly" similar (difference zero) the datasets are equivalent with regard to the attributes measured, as long as they have elements of the same type.

4. The Content of Spatial Datasets

When one considers a spatial dataset and its contents there are two groups of attributes to examine: class attributes and set attributes.

4.1 Class Attributes

Class attributes are the attributes associated with the spatial objects in the dataset. The values recorded for each of the class attributes are distinguishing characteristics between the individual members of the dataset. These values and the manipulation of them through queries and functions have been the focus of traditional database technology (Elmasri and Navathe 1994; Ullman 1982). Query and analysis of class attributes can be computation-intensive and cognitively overwhelming in very large datasets.

4.2 Set Attributes

Set attributes are distinctly different from, but in some cases dependent upon, the class attribute values of the members of the dataset. The concept behind set attributes is that information about a dataset can be helpful in finding and sorting among several datasets.

Descriptive Attributes. Some set attributes, such as the data's source agency or region of coverage, are specific to the dataset itself and have nothing to do with the class attribute values of the objects in the dataset. These are *descriptive attributes* and are used primarily for cataloging the datasets. Attempts to describe such attributes through metadata are currently promoted by international and agencies (FGDC 1997; Weibel *et al.* 1997) and by spatial digital library research (Beard and Smith 1997). The burden of creating the metadata is the data producer's and is often considered uneconomical when datasets are perceived to have limited use to external users. This often results in cursory compliance with regulations regarding metadata generation.

Summary Attributes. Other set attributes, called *summary attributes*, are summary statistics for the contents of a dataset and can be traced to a direct relationship to one or more class attributes of the objects in the set. Examples are total population or mean income. Because there is a functional relationship between an element's class attribute value and the value of the set's summary attribute, it is possible to quantify the relative contribution of each element to a summary attribute value. This functional relationship in turn permits the selection of the most important elements, defined as largest contribution, in the dataset in terms of a particular attribute. The selection of the most important elements has direct application to operations that attempt to simplify large collections of data, such as cartographic generalization and data visualization. For example, New York City is the largest city in the United States, contributing about 3% of the total 1990 U.S. population. With as few as ten cities, it is possible to give an impression of the distribution of population in the United States as long as the most important (i.e., largest) cities are used. These ten cities may not sufficiently represent the spatial qualities of the dataset, but the dataset's population characteristics are preserved.

Synoptic Attributes. A third type of set attributes, called *synoptic attributes*, can be defined as functionally related to class attributes, but having meaning only with relation to the collection of objects in the dataset. Synoptic attributes, such as data ranges, data frequency, density, dispersion, and pattern, refer to qualities inherent to the spatial dataset. They are either invalid or nonsensical when the set is empty. While each element of the dataset contributes to the resulting characteristic, the element's contribution is its relationships with the other elements and is not quantifiable in the absence of the other elements. Synoptic attributes are of primary interest in evaluating spatial datasets and summary attributes can be used to evaluate the non-spatial aspects of the datasets.

5. Measures of Spatial Dataset Similarity

As stated previously, similarity is a distance function (Tversky 1977) between two sets that share a group of attributes. The scales of these attributes-measured according to Stevens's (1946) categorization into nominal, ordinal, interval, and ratio-determine the appropriate methods for evaluating that distance. In order for these various measures to be combined it is necessary to standardize all of the measures. We use the convention that a *similarity* value of one means that there is no measurable *difference* between the two spatial sets. A similarity value of zero means that the two spatial sets are completely dissimilar-as distant as possible. Once similarity measures are standardized it is possible for a user to set a desired similarity level in a meaningful way. For instance, similarity ($\approx$) could be defined for all datasets (D) over a given set of attributes (A) as having a difference (δ) of less than a 0.1 from the target (T) (Equation 2).

$$T \approx D \text{ if } \delta_{[A_1 \ldots A_n]}(T, D) \le 0.1$$

Ultimately measures of similarity have the most value when two or more candidate datasets are being compared to a single target. The target could be a collection of attribute values from a known spatial set or may be a hypothetical spatial set whose values have been constructed in a query. In either case, the fundamental question is, "Which is closer in the evaluation space, the *target* and set X or the *target* and set Y?" (Equation 3a or 3b)

$$\delta_{[A,B]}(T,X) < \delta_{[A,B]}(T,Y)$$

$$(T,X) \gtrless (T,Y)$$

5.1 Measuring Similarity on Interval and Ratio Scales

Making distance measures on interval and ratio scales is straight forward, since the ability to support difference operations is part of their definition. The standardization of these distances is more of a concern, however. In both interval and ratio scales the zero point for distances between values is set by the set attribute values of the target dataset (T). For most summary attributes and some synoptic attributes reasonable maximum distances can be defined by knowing what the actual ranges of values are for all the class attributes. When there is a fully defined universal spatial ideal A these values should be known and stored in the set's metadata as summary attributes (X_{min}) (Equation 4). When the datasets (D) being evaluated do not belong to a defined universal spatial ideal the ranges must be calculated by combining the known ranges of the separate datasets.

$$\Delta^{d}_{attribute} = 1 - \frac{|x_T - x_D|}{Max(x_{max} - x_T, x_T - x_{min})}$$

For example, local density of points in a dataset is often measured by calculating the mean of the distances of each point to its nearest neighbor. The maximum distance any two points can have from one another is the diagonal of the minimum bounding rectangle (MBR). If the MBR of the universal spatial ideal is not known then we can estimate the maximum mean nearest neighbor value by choosing the maximum MBR diagonal from the spatial datasets being evaluated. Because the calculation of maximum values for any particular set attribute is highly dependent on the semantics of the attribute it will be necessary to rely on the user or a domain specialist to define methods for the calculations of maxima.

5.2 Measuring Rank Similarity

There are several standard statistical measures used to compare two sets of ordinal values. However, measures such as Spearman's rs and Kendall's tau and gamma (Blalock 1972) are only useful for two separate variables and sets of equal size. Because there is an assumption that the two sets A and B share some elements one statistic is the number of shared objects (Equation 5), however, this measure loses the concept of rank similarity. For instance, two sets of cities that share Russell, Kansas and Hope, Arkansas would be assessed to be as similar as two sets that share New York and Los Angeles.

$$X = 1 - \frac{count(A \cap B)}{count(A)}$$

The rank similarity of a spatial set (A_1) to its universal spatial ideal (A) can be evaluated if the objects in the universal spatial ideal are ordered on a common attribute. The ρ-score (Equation 6) measures the degree to which a subset's members retain the rank importance of the original set. This assessment is done by comparing the sum of its n rank values to the sum of a hypothetical set containing the objects ranked 1 to n. The values for the ρ-score range between 0 and 1. A value of 0 indicates that the subset contains the n highest ranked members of the superset. If the n lowest ranked members of the superset are in the subset the ρ-score is 1. When n and the number (N) of elements in A are equal, the ρ-score is assumed to be 0 since the sets are identical by definition. A comparison of ρ-scores is only valid between two spatial sets if they belong to the same universal spatial ideal.

$$\kappa = 1 - \frac{\sum|f_o - f_e|}{\kappa_{max}}$$

$$\kappa_{max} = |N - f_{e,i}| + \sum|0 - f_e| = 2(N - f_{e,i})$$

The ρ-score is calculated in the following manner. Each subset member's rank (R_i) in the universal spatial ideal (A) is summed for the entire range of the dataset's (A_1) n members. The expected sum, if the first n members of A were chosen, is subtracted from A_1's actual rank sum. This value is standardized with the expected sum of the ranks of the last n members of A (where N is the total number in A).

5.3 Similarity on Nominal Data Scales

Nominal data scales by definition cannot be ordered. It is, however, possible to evaluate the similarity of the nominal values in a given spatial ideal to those in its universal spatial ideal. For this goal, we must know the frequency of each particular value in the spatial ideal and the universal spatial ideal. The κ–score sums the differences between the frequencies of nominal values (f_o) in a set with the expected frequencies (f_e) (Equation 7). The result is normalized by the maximum difference that could occur by substituting the number of occurrences (N) for (f_0) in the category with the lowest expected frequency (f_e) and a zero is substituted for all other observed frequencies (Equation 8). The standardized form permits comparison between subsets of different sizes. The κ–score relies on a set of theoretical frequencies which were expected, this requires knowledge of the character of the universal spatial ideal and therefore can only be used where the ideal is known.

$$N = n: \quad \rho' = 1$$

$$N > n: \quad \rho' = 1 - \frac{\sum_{i=1}^{n} R_i' - \frac{n(n+1)}{2}}{n(N' - n)}$$

5.4 A Mixed-Scale Similarity Measure

Most sets of spatial data have a broad range of summary attributes and synoptic attributes, which need to be combined to accurately measure the similarity of any spatial dataset to a target set of attribute values. Since it is assumed that all similarity measures are standardized to values between 0 and 1, we can combine the values in a simple n-dimensional distance measure. Equation 9 combines a number (n) of measures of mixed data scales in a single similarity measure (σ), where A identifies the class definition of the spatial ideal. Each σ also includes an attribute list, which identifies the attributes over which the similarity has been calculated. A σ -score of 1 means that the datasets have the highest degree of similarity and, therefore, are equivalent, while a σ -score of 0 means that the datasets are completely different.

$$\sigma^A_{|\text{[illegible]}|} = \sqrt{\frac{(\Delta^A)^2 - (\rho^A)^2 + \ldots + (\kappa^A)^2}{n}}$$

A similarity index can be created over a collection of two or more spatial datasets by using the σ values to order the datasets (Equation 10).

$$S^A_{\text{[illegible]}} = \{\sigma^A_{[\text{[illegible]}]_1}, \sigma^A_{[\text{[illegible]}]_2}, \ldots, \sigma^A_{[\text{[illegible]}]_k}\}$$

Because the σ values are created from summary and synoptic attributes, which can be stored as metadata, there is no need to access the data to evaluate similarity of the datasets. The open access to the metadata values makes it possible for individual domain experts to construct σ-functions, which express the particular semantics of their applications, rather than relying on predefined definitions of fitness for use explicitly stated in the metadata.

6. Conclusions

In this paper we distinguished the concepts of identity, equivalence, and similarity. In most digital spatial archives it is possible for each of these conditions to exist, but it is mostly likely that datasets will be similar, not equivalent, to one another. The ability to identify identical and equivalent datasets is relatively well defined and most database systems have utilities to identify redundant data items and datasets. Currently a difference of one item makes two datasets non-equivalent and results in both sets being stored. The elimination of redundant data items is a part of normalization in relational database theory (Ullman 1982), but we are focused on the dataset, with storage and retrieval of individual items. To better understand the degree to which two datasets are not equivalent it was necessary to define a model of datasets similarity.

A methodology for assessing similarity between a subset and a target dataset was specified to address the key question of how similar a subset is to its superset. This methodology has been defined in a general manner which handles dataset attributes in terms of their measurement scale rather than their semantics. Since the model of similarity can be applied in general to attributes on all scales of measurement and semantic content, it is necessary to examine and select a set of measures specific to spatial datasets. The values generated by these measures can in turn be used in the similarity model defined here to build a similarity index of the spatial similarity between a dataset and a target dataset. In most instances the target will be the universal spatial ideal. It is possible to define theoretical metadata for a desired dataset and to measure the similarity of all datasets to the theoretical target. This permits the user to query a digital spatial archive for datasets that are most similar to their needs.

When users request a sample of a dataset they can set a threshold level for similarity over a set of attributes. Using the model described here, it would be possible to determine when a subset meets that threshold. The sample can then be transferred to the users for analysis by whatever means they need to make a decision.

We will assume that the analysis of non-location class attributes and the generation of their summary and synoptic attributes can be handled by domain experts based on work that has been done in database technology (Mena et al. 1998; Bishr 1997) and data mining (Ng and Han 1994). The large body of research in spatial analysis and spatial statistics (Bailey and Gatrell 1995; Unwin 1981) suggests there may be measures of synoptic spatial attributes such as dispersion, density, and pattern that may be used to describe the spatial character of a dataset using the similarity methodology presented here.

7. Acknowledgements

This work was partially supported by the National Science Foundation through the National Center for Geographic Information and Analysis under NSF grants SBR-8810917 and SBR-9600465. This work has also been partially supported by the National Imagery and Mapping Agency under grant NMA-202-97-1-1023

REFERENCES

T. C. Bailey and A. C. Gatrell (1995) Interactive Spatial Data Analysis. Longman Scientific & Technical, Essex, England.

K. Beard and T. R. Smith, Ed. (1997) A Framework for Meta-Information in Digital Libraries. Mutlimedia Data Management: Using Metadata to Integrate and Apply Digital Media McGraw Hill, New York.

Y. Bishr (1997) Semantic Aspects of Interoperable GIS. Ph.D. thesis, ITC, Enschede, The Netherlands.

H. M. Blalock Jr. (1972) Social Statistics. McGraw-Hill, New York.

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

R. S. Burington (1973) Handbook of Mathematical Tables and Formulas. McGraw-Hill Book Company, New York.

L. Cardelli and P. Wegner (1985) On Understanding Types, Data Abstractions, and Polymorphism. Computing Surveys 17(4): 471-522.

J. R. Eastman (1985) Cognitive models and cartographic design research. Cartographic Journal 22(2): 95-101.

R. Elmasri and S. B. Navathe (1994) Fundamentals of Database Systems. The Benjamin/Cummings Publishing Company, Redwood City, CA.

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

FGDC (1997) Content standard for digital geospatial metadata (revised April, 1997). Federal Geographic Data Committee. Washington, D.C., Technical Report .

D. M. Flewelling and M. J. Egenhofer (1999) Using Digital Spatial Archives Effectively. International Journal of Geographic Information Science 13(1): 1-8.

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.

H. Gallaire, J. Minker, and J.-M. Nicolas (1984) Logic and Databases: A Deductive Approach. Computing Surveys 16(2): 153-185.

H. Garcia-Molina, J. Widom, J. Wiener, W. Labio, B. Lent, and Y. Zhuge (1995) A Warehousing Approach to Data and Knowledge Integration. Stanford University, Technical Report.

V. Gudivada and V. Raghavan (1995) Design and Evaluation of Algorithms for Image Retrieval by Spatial Similarity. ACM Transactions on Information Systems 13(2): 115-144.

M. Johnson (1987) The Body In The Mind: The Bodily Basis of Reason and Imagination. University of Chicago Press, Chicago, IL.

G. Lakoff (1988) Cognative Semantics. in: U. Eco, M. Santambrogio, and P. Violi (Ed.), Meaning and Mental Representations. Advances in Semiotics pp. 237, Indiana University Press, Bloomington, IN.

D. Mark and S. Freundschuh (1995) Spatial Concepts and Cognitive Models for Geographic Information Use. in: T. Nyerges, D. Mark, R. Laurini, and M. Egenhofer (Ed.), Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems. pp. 21-28, Kluwer, Dordrecht, The Netherlands.

R. Mehrotra and J. Gray (1995) Similar-Shape Retrieval in Shape Data Management. IEEE Computer 28(9): 57-62.

E. Mena, V. Kashyap, A. Illarramendi, and A. Sheth (1998) Domain Specific Ontologies for Semantic Information Brokering on the Global Information Infrastructure. in: International Conference on Formal Ontology Information Systems, Trento, Italy, June 1998.

U. Neisser (1976) Cognition and Reality: Principles and Implications of Cognitive Psychology. W. H. Freeman, San Francisco, CA.

R. Ng and J. Han (1994) Efficient and Effective Clustering Methods for Spatial Data Mining. in: J. Bocca, M. Jarke, and C. Zaniolo (Ed.), 20th International Conference on Very Large Data Bases (VLDB '94), Santiago, Chile, pp. 144-155.

S. Pinker (1990a) Language Acquisition. in: D. N. Osherson and H. Lasnik (Ed.), Language. 1, pp. 199-241, The MIT Press, Cambridge, MA.

S. Pinker (1990b) A theory of graph comprehension. in: R. Friedle (Ed.), Artificial Intelligence and the Future of Testing. pp. 73-126, Ablex, Norwood, NJ.

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

T. R. Smith (1996) A Digital Library for Geographically Referenced Materials. Computer 29(5): 54-60.

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

J. D. Ullman (1982) Principles of Database Systems. Computer Science Press, Rockville, MD.

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

S. Weibel, W. Cathro, and R. Iannella (1997) The 4th Dublin Core Metadata Workshop Report. URL http://www.dlib.org/dlib/june97/metadata/06weibel.html (Last accessed April 2, 1999),

E. B. Zurif (1990) Language and the Brain. in: D. N. Osherson and H. Lasnik (Eds.), Language. pp. 177-198, The MIT Press, Cambridge, MA.