Fabrikant, S. I. (2001) Visualizing Region and Scale in Information Spaces. *Proceedings*, The 20th International Cartographic Conference, ICC 2001, Beijing, China, Aug. 6-10, 2001, pp. 2522-2529.

VISUALIZING REGION AND SCALE IN INFORMATION SPACES

Sara Irina Fabrikant
Department of Geography
University of California Santa Barbara, Santa Barbara, CA 93106, U.S.A.
Phone: +1 (805) 893-5305, Fax: +1 (805) 893-3146, email: sara@geog.ucsb.edu

Abstract

The geographic concepts of region and scale can be preserved in semantic information spaces and depicted cartographically. Region and scale are fundamental to geographical analysis, and are also associated with cognitive and experiential properties of the real world. Scale is important when graphically representing a spatialization, as it affects the amount of detail that can be shown. Latent semantic indexing in conjunction with two different ordination techniques is employed to construct a semantic Reuters news wire space. Intramax, a hierarchical clustering algorithm, is applied to delineate semantic regions in the Reuters database based on a functional distance measure. This topological information helps to identify the appropriate levels of granularity at which the information space can be visually explored. Amplification of ordination techniques with the Intramax procedure is a useful strategy for creating scale-dependent information spaces that facilitate the exploration of abstract, complex data archives.

Introduction

Researchers in information visualization are concerned with graphically representing complex, abstract data domains to facilitate knowledge extraction from very large non-spatial data repositories. These depictions are often based on a geographic metaphor, and are commonly referred to as spatializations or information spaces. Most spatializations are generated by researchers outside of GIScience and cartography. Spatialization examples can be found in a recently published book on information visualization (Card et al., 1999) and at the Atlas of Cyberspaces website (Dodge, 2001). However, a structured approach based on solid theoretical foundations to formalize the underlying representational framework are scarce (Slocum et al., 2001). Information spaces are more than spatial depictions of non-spatial data. Spatialization includes both geometric and semantic generalization procedures that transform large complex data domains into basic information components (Fabrikant and Buttenfield, 2001).

The majority of spatialization examples rely on a narrow subset of formal, and typically Euclidean, spatial properties where the locations are nothing more than points in space and the distances between items are limited to straight lines. Geographic space is more than just Euclidean geometry. Entities and their relationships in space carry experiential and socially constructed meanings. Usable information spaces need sound semantic abstraction frameworks. Sound semiotic practices must complement these semantic abstractions. Current information space depictions also lack coherent representational frameworks easily provided by cartographic design principles. This paper forwards an ontological understanding of region and scale concepts for spatialization, outlines semantic and semiotic transformation approaches for representation, and then applies these procedures to a large text document archive as proof-of-concept of the theoretical framework employed.

The semantics of region and scale

Spatialization is based on a metaphorical mapping from physical space (source domain) into a conceptual space (target domain). Before the spatial metaphors of region and scale can be depicted in a semantic information space, it is necessary to formalize their appropriate source domains. Region and scale are very good candidates for spatialization, as they offer a rich array of sub metaphors. Increased attention needs to be paid to identifying the fundamental representational primitives associated with the source domains, to make the metaphorical mappings useful for information exploration. Cartography already has a long-standing scientific tradition on how to appropriately transform large heterogeneous data sets into visually accessible information displays for knowledge acquisition (Bertin, 1998). Cartographic design principles and geovisualization approaches offer the missing semiotic foundations to depict cognitively adequate semantic information spaces.

Regions are spatial taxonomies, based on perceived similarity/difference of a phenomenon's characteristics; that is, regions are partitions in space. Scale is the granularity (or 'mesh size') of the partitioning. The finer the partitioning scheme, the higher level of detail can be identified, and the more regions will be seen in an environment. Region and scale can be more formally described by axioms of mereotopology, a combination of topological and ontological theory (Smith, 1995). Discontinuities (i.e., boundaries) in the environment separate zones of relative homogeneity (i.e., regions). The partitioning of a landscape may be based on tangible, physical (or bona fide) entities which exist regardless of human cognitive agency such as continents, rivers, and forests (Smith, 1995), wherein bona fide *boundaries* such as coastlines, riverbanks, and forest edges delineate bona fide *regions*. *Functional* regions are semantic partitions of space, which only exist dependent on humans' cognitive capabilities of representation. Functional regions contain semantically similar entities (i.e., *fiat* objects) separated from different entities by *fiat* boundaries. Political and economic boundaries are examples of the fiat kind.

Generally defined, scale relates to levels of organization, or hierarchies. In GIScience, scale can be referred to as (1) the level of detail (resolution) of a phenomenon under study (e.g., 30m sampling interval of a digital elevation model), (2) as level of abstraction or spatial extent (e.g., a footprint at 1:200,000 scale), (3) as a level of human point of observation (i.e. body space, geographic space), or (4) as a purely semantic level, for example nested enumeration units, delineated by political boundaries. Scale operates in both the bona-fide and the fiat world, and it may be spatial, temporal or both (Ahl and Allen, 1996). The scale continuum operates within a logical geometric framework, or frame of reference. This reference frame is the higher order spatial concept of hierarchy. Hierarchy is composed of the spatial primitives identity, location, magnitude and connection (Golledge, 1995). In geography, a reference frame is usually based on a formalized coordinate system. Once the construction details of a chosen frame of reference are known, scale change can be identified and measured.

Hierarchies are a consequence of complexity of a system, and are the fundamental ordering principle of the physical environment and human nature (Salthe, 1985). Both nested and non-nested hierarchies exist. Containment is a key organizing principle of nested hierarchies, that is, entities on lower levels of detail are fully contained in the higher levels of organization. Features on the same level of the hierarchy have the same scale. Most bona fide objects are implicitly organized in nested hierarchies, simply by applying the size rule (spatial extent) and the containment principle. For example, a soil sample is part of a patch of land, the patch is part of a region, the region is part of a continent, and so on. Nested hierarchies in the human world only operate if the containment rule is explicitly defined and communicated (known) to all members of the hierarchy (Ahl and Allen, 1996). An organizational chart of a business corporation, or the line of command in an army are nested examples of semantic hierarchies.

Hierarchical order and categorization are also fundamental organizational principles of human cognition. Hierarchical order is an example of the cognitive 'part-whole' and 'scale' image schemata (Lakoff, 1987). These cognitive constructs match the ontological primitive 'part-whole' in mereotopology. Empirical evidence shows that hierarchical order is an important aspect of how humans learn about the environment (Golledge, 1999). Human cognition varies with scale, ranging from personal-scale space with direct sensory interactions, to larger-scale space, where direct sensory interaction might not be feasible. Different dominant and subordinate levels of detail are evident and stored in humans' cognitive maps (Golledge, 1999).

Humans' cognitive maps also include hierarchical organization and geometrical (Euclidean) knowledge structures. Configurational knowledge is organized with geometric primitives such as points, lines, areas, and surfaces (Golledge, 1999). Points encode landmarks or reference nodes; lines describe routes and paths for wayfinding; and areas encapsulate regions and neighborhoods, including topological containment and inclusion information. Surfaces cognitively afford support for activities in space and represent density estimates (Golledge, 1999). '*C*ontainer', 'center-periphery', and 'part-whole' image schemata are obvious source domains for the region metaphor (Lakoff, 1987). Landmarks (nodes) may be encoded as regional cores (centroids) to spatially partition functional regions on dominant and subordinate levels (Christaller, 1933). Paths connect landmarks and their associated neighborhoods (Lynch, 1960). Landmarks are also used in wayfinding for integration of routes and paths into a network configuration (Golledge, 1999). 'Object', 'collection', 'path', and 'link' are the relevant cognitive image schemata for representing nested functional regions (Lakoff, 1987).

Semantic transformations

Computational methods for creating semantic information spaces involve a two-step transformation process. First, a mathematical transformation creates a logically defined coordinate system to re-arrange a set of data items based on their content and functional relationships (Fabrikant and Buttenfield, 2001). Most mathematical procedures to generate information spaces are a combination of vector space modeling techniques (Salton, 1989) coupled with a variant of ordination such as multidimensional scaling (MDS). Vector space modeling is often the method of choice for semantic generalization to condense a database into a semantic proximity matrix typically characterized by word co-occurrences. Second, the spatialized transformation is graphically represented for subsequent data exploration and knowledge extraction. A proximity matrix is then subjected to an ordination technique for visualization. A major drawback of ordination is that all semantic (intrinsic) properties of data objects are collapsed into the extrinsic property of spatial proximity, usually based on Euclidean distance. Topology (functional distance) and multiple levels of detail (hierarchy, functional regions) that are important aspects of the experiential world are not preserved. Visual exploration of the information space at different levels of granularity is therefore hindered.

A cognitively plausible semantic space

This section describes the construction of a semantic news wire space that builds upon the theoretical concepts previously discussed. Latent semantic analysis (LSA), also known as latent semantic indexing (LSI), was chosen as the semantic generalization method (Deerwester et al., 1990). LSI is based on singular value decomposition (SVD), which allows the statistical construction of a large 'document-term' similarity space of chosen dimensionality. Documents and terms that are closely associated semantically will be placed near each other in a Euclidean vector space. LSI extends the keyword matching techniques of classical vector space modeling in that terms that do not actually appear in a text may still be close to a document if the relationship is consistent with the overall association pattern in the database (Deerwester et al., 1990).

Reuters news wire stories (n= 504) appearing February 9-10, 2001 were subjected to LSI. The semantic transformation is based on content similarity between news articles. The semiotic transformation includes a combination of two ordination techniques. First, a spring-node positioning algorithm (Kamada and Kawai, 1989) projects individual news articles as landmarks (regional centers) into a Euclidean information space. Figure 1 shows the Reuters news stories topology containing news stories as nodes (landmarks) and semantic relationships as links among them (e.g., cross-referencing).

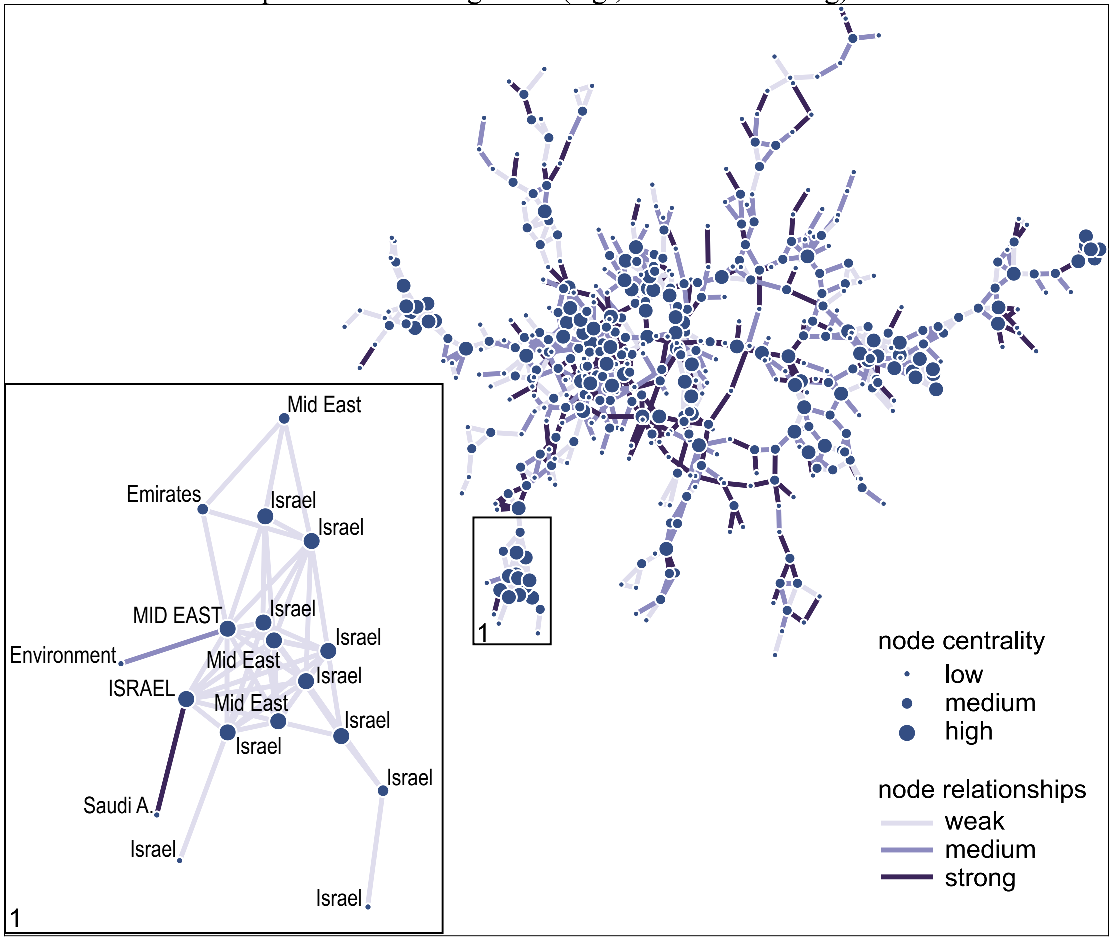

Figure 1: Reuters news topology

News stories (point symbols) that are more similar in content are placed closer to each other in the map than stories that are less similar. The Pathfinder scaling technique (Schvaneveldt, 1990) integrates the landmarks into a semantic network configuration. The nodes of the network are graphic representations of documents; the links show semantic relationships between documents. Pathfinder associative network scaling reveals semantic network structures of cognitive concepts based on psychological distance estimates. A Pathfinder network is a graph-theoretic model, in essence a minimum-spanning tree, where an undirected graph of minimal length is constructed that spans all the nodes of the network. Node connectedness (Equation 1) is one option to structure such information spaces:

$$L_i / N\text{-}1 \quad \textit{where:} \quad \begin{array}{l} L_i = \text{number of links for node } i \\ N = \text{number of nodes in the network} \end{array}$$

Equation 1: Node centrality (Wasserman and Faust, 1999)

This measure of node centrality (depicted by varying circle radii in Figure 1) reveals that certain nodes in the network are more connected than others. Better connected documents act as primary landmarks (i.e., semantic anchor points for information exploration). A Voronoi tessellation uses node locations as region centroids to partition the information space and to capture the cognitive concept of center-periphery. Areal representation provides a cognitively plausible depiction of semantic regions where individual documents aggregate to topical themes at higher levels of abstraction. Extending the surface metaphor into a third dimension provides opportunities for capturing the cognitive image schema of ‘more is up’ and ‘near-far’, as shown in Figure 2. Articles that are densely clustered tend to metaphorically ‘pile up’ into semantic mountains.

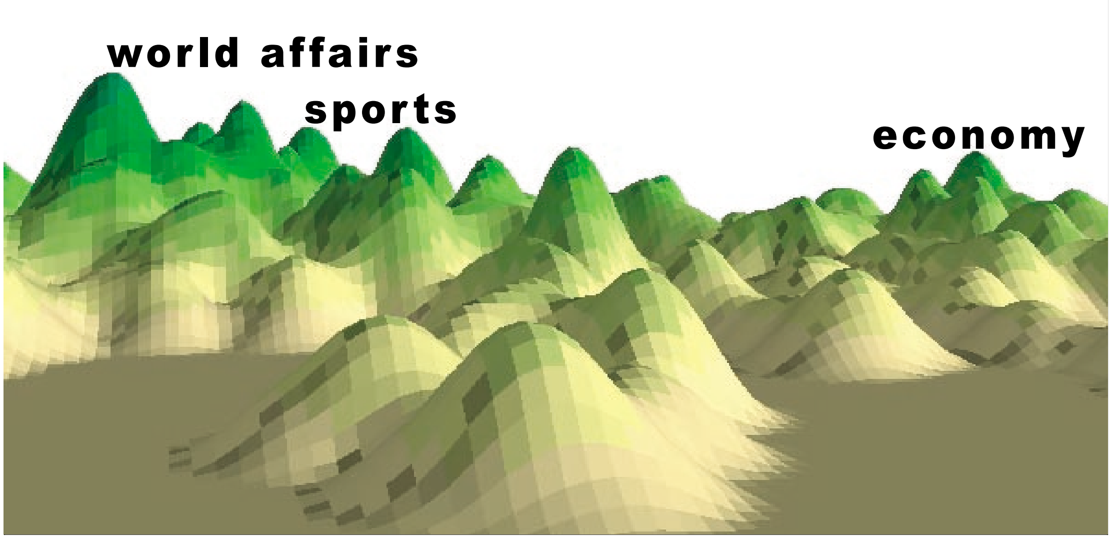

Figure 2: Topic density surface

Preserving the continuum of scale in a spatialization permits exploration of the information space at multiple levels of detail. One can imagine zooming in and out of a spatialization to view a document collection at different levels of granularity. Sometimes the breadth of topic themes might be of interest; other times one would like to see the density of individual documents pertaining to one subject in the data archive. The scale metaphor creates the potential for hierarchical grouping of items, nested regionalization, and other types of generalization. A proof-of-concept of the scale metaphor is achieved by subjecting the Reuters proximity matrix to Intramax, a hierarchical clustering algorithm (Masser and Brown, 1975). The Intramax procedure identifies nested functional regions in an information space by hierarchically aggregating spatial units on a functional distance measure (e.g., content similarity). The great advantage of Intramax over related hierarchical clustering techniques is that interaction effects between entities are analyzed (e.g., based on their functional proximity) and topological information among the spatial units are taken into account. A contiguity constraint is built into the algorithm to yield spatially more homogeneous clusters. The *x* and *y* coordinates of documents, which the ordination procedure generates, are input to Intramax along with the document similarity matrix to obtain nested hierarchical clusters. Intramax provides two outputs to evaluate how the regions are formed (de Jong et al., 1994). The fusion report shows the aggregation history of the clustering procedure. Break points, where large amounts of interactions become *intra*zonal, identify how documents are joined to clusters. Two, four and nine hierarchically embedded clusters optimally represent the Reuters news space. Similar to cartographic map scales, which represent slices through a tree of nested spatial footprints along the scale continuum, one could think of these hierarchically embedded cluster solutions as semantic levels of resolution in the information space. Figure 3 shows semantic generalization levels of the Reuters news space found by the Intramax procedure.

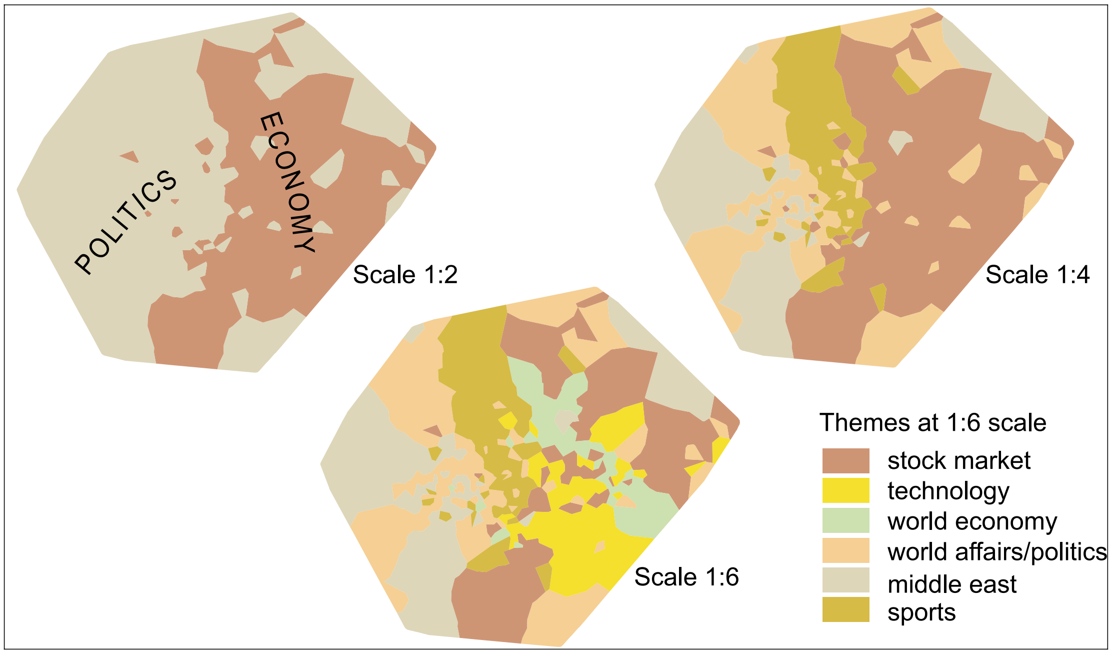

Figure 3: semantic levels of detail

At the lowest level of detail (Scale 1:2) one can see an overview of the entire news article space. Other more detailed levels ‘zoom in’ on more specific themes and documents. One can systematically modify the graphic density and complexity of the spatialization (i.e., cartographically generalize as in Scales 1:2 through 1:6). The scale-dependence of the view gives information seekers a cue about how ‘close’ they are to the information at which they are looking—ranging from the entire collection to individual news articles—and at what levels of granularity they will find desired pieces of information. The cluster analysis helps choose the appropriate levels of granularity at which the information space should be explored (semantic generalization). The Reuters news space should be ideally explored at scales 1:2, 1:4, and 1:9. Töpfer and Pillewizer's (1966) Radical Law can be used to further cartographically generalize the display density at each level of the semantic hierarchy. Other cartographic generalization aspects such as scale-dependent automatic feature labeling can also be considered (Skupin, 2000). Most spatialization examples in the Human Computer Interaction (HCI) community that are concerned with the scale issue typically deal only with the graphic complexity problem. Modifying graphic density in a spatialized display based on semantic levels of resolution (e.g. by Intramax clustering) in combination with semiotic approaches (e.g. Radical Law) makes a GIScience/cartography approach to spatialization so powerful and unique.

Summary and outlook

This investigation presents a theoretical framework grounded on the principles of GIScience and geovisualization to depict the geographic concepts of region and scale in cognitively plausible semantic information spaces. A proof-of-concept of a large Reuters news article space has been presented that allows information seekers to visually explore news stories at multiple semantic levels of detail. A spring algorithm and the pathfinder ordination procedures were utilized to transform a large news article archive into a map like representation for visual exploration and knowledge discovery. The news article spatialization was further enhanced with Intramax to preserve and depict scale-dependent topic regions in the news space. The Intramax procedure identified functional regions in the information space by hierarchically aggregating spatial units based on similarity of document content. This topological information not only helps to identify the appropriate levels of granularity at which the information space can be explored (semantic generalization), but also permits to systematically modify the graphic density that is visible in the spatialized view (cartographic generalization).

The amplification of standard ordination techniques with the Intramax procedure has great potential for creating cognitively adequate information spaces which facilitate the visual exploration of abstract, complex data domains archived in rapidly growing digital repositories.

Acknowledgements

Funding by the National Imagery and Mapping Agency (NMA-201-00-1-2005) is greatly appreciated. Thanks are due to Bruce Rex of the Starlight Project at the Pacific Northwest Laboratory for sharing an extract of the Starlight's Reuters database, and to Tom de Jong at the University of Utrecht for the support on Intramax with the Flowmap software.

References

Ahl, V. and Allen, T. F. H. (1996) *Hierarchy Theory*. New York, NY, Columbia University Press.

Bertin, J.(1998) *Sémiologie Graphique: Les Diagrammes – les Réseaux – les Cartes*. Paris, France, Éditions de L'École Pratique des Haute Études.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999) *Readings in Information Visualization. Using Vision to Think*. San Francisco, CA, Morgan Kaufmann.

Christaller, W. (1933) *Die Zentralen Orte in Süddeutschland*. Jena, Germany, Gustav Fischer.

de Jong, T., van Eck, J. R., and Floor, H., 1994, *Using Flowmap Version 4.1. A Program for the Display and Analysis of Interaction Data*, Faculty of Geographical Sciences, University of Utrecht, Utrecht, The Netherlands. URL: http://flowmap.geog.uu.nl/ (Apr., 2001)

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harschman, R. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science,* 41: 391-407.

Dodge, M.(2001) *An Atlas of Cyberspaces*. Centre for Advanced Spatial Analysis, University College, London, U.K. URL: http://www.cybergeography.org/atlas/atlas.html (Apr., 2001)

Fabrikant, S. I. and Buttenfield, B. P. (2001) Formalizing Semantic Spaces For Information Access. *Annals of the Association of American Geographers,* 91: 263-280.

Golledge, R. G. (1995) Primitives of Spatial Knowledge. In *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems,* edited by Nyerges, T. L., Mark, D. M., Laurini, R., and Egenhofer, M. J. Dordrecht, Kluwer Academic: 29-44.

Golledge, R. G. (1999) *Wayfinding Behavior: Cognitive Mapping And Other Spatial Processes.* Baltimore, MD, Johns Hopkins University Press.

Kamada T. and Kawai, S. (1989) An Algorithm for Drawing General Undirected Graphs. *Information Processing Letters,* 31: 7-15.

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

Lynch, K. (1960) *The Image of the City*. Cambridge, MA, MIT Press.

Masser, I. and Brown, J. (1975) Hierarchical Aggregation Procedures for Interaction Data. *Environment and Planning A,* 7: 509-523.

Salthe, S. N. (1985) *Evolving Hierarchical Systems*. New York, Guilford Press.

Salton, G. (1989) Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer. Reading, MA, Addison-Wesley.

Schvaneveldt, R. W. (1990) Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood, NJ, Ablex.

Skupin, A. (2000) From Metaphor to Method: Cartographic Perspectives on Information Visualization. *Proceedings,* IEEE Symposium on Information Visualization (InfoVis 2000), Oct. 9-10, 2000. Salt Lake City, UT,: 91-97.

Slocum, T. A., Blok, C., Jiang, B., Koussoulakou, A., Montello, D. R., Fuhrmann, S., and Hedley, N. (2001) Cognitive and Usability Issues in Geovisualization. *Cartography and Geographic Information Science,* 28: 61-75.

Smith, B. (1995) On Drawing Lines on a Map. In *Spatial Information Theory. A Theoretical Basis for GIS (COSIT 1995),* edited by Frank, A. U., Kuhn, W., and Mark, D. M. Berlin, Springer: 475–484.

Töpfer, F. and Pillewizer, W. (1966) The Principles of Selection. *Cartographic Journal,* 3: 10-16.

Wasserman, S. and Faust, K. (1999) *Social Network Analysis*. Cambridge, U.K., Cambridge University Press.