The First Law of Cognitive Geography: Distance and Similarity in Semantic Spaces

Sara Irina Fabrikant, Marco Ruocco, Richard Middleton

Department of Geography, University of California Santa Barbara,
Santa Barbara, CA 93106, Email: {sara, ruocco, richard}@geog.ucsb.edu

Daniel R. Montello[1], Corinne Jörgensen[2]

[1]Department of Geography, University of California Santa Barbara,
Santa Barbara, CA 93106, Email: montello@geog.ucsb.edu

[2]School of Information Studies, Florida State University,
Tallahassee, FL 32306-2100, Email: cjorgens@lis.fsu.edu

As GIScientists we know that "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970: 236). Unfortunately this law does not tell the spatially aware researcher how to conceptualize "nearness" or which proximity measure to utilize to quantify "relatedness". One of the main challenges in GIScience is not only to formalize people's often fuzzy distance notions for distinguishing near features from far ones, but also to find the appropriate transformation rules to model and depict actual and perceived relatedness between real world phenomena within GISystems.

The spatial metaphor distance is also often utilized in information visualization to depict semantic similarities between data items in very large databases (Card et al., 1999). Conceptualization and formalization of proximity is even more critical in information visualization (also known as spatialization) where non-spatial data are projected into a logically (but arbitrarily) defined spatial reference frame based on semantic relatedness. Straight-line Euclidean distance is typically used in spatialization to represent relatedness in data content, such that semantically similar documents are placed closer to one another in an information space than less similar ones. As intuitive this mapping principle is, surprisingly the InfoVis community has never empirically validated its basic assumptions. The Euclidean distance–similarity mapping is also widely used in information science and content-based retrieval. A number of researchers in these communities have recognized that other relatedness measures may be more appropriate, but very little experimental research has been done to evaluate any of the proposed alternatives. The key question remains if and how people grasp the graphically encoded similarity–distance mapping when exploring a semantic information space. In addition, what kinds of proximity strategies do viewers employ when conflicting notions of relatedness are jointly shown in a display (e.g. metric vs. topologic proximity)? Are proximity judgments affected by specific space types (e.g. Cartesian vs. geographic space), or are they influenced by the display size (e.g. table top vs. large scale spaces)? How do people's background and training modify their proximity judgments, and thus alter metaphor comprehension? Finally, do semiotic generalization principles affect similarity judgments (e.g. the use of visual variables such as, color or size)?

This paper outlines an experimental design to answer above research questions. Empirical results are presented on how people decode three proximity types utilized to

represent semantic relatedness in a database of Reuters news stories. The construction of the semantic news wire space follows ontological modeling principles of generalization, association, and aggregation. Four display types were devised where content similarity between news stories is metaphorically mapped into (1) straight-line Euclidean distance between points (metric proximity), (2) linkages between points (network-topologic proximity), (3) metric linkages between points (network-metric proximity) and (4) membership of points belonging to thematic regions (hierarchical proximity).

The spatialized displays were constructed using Latent Semantic Indexing (semantic generalization) in conjunction with two scaling methods for depiction (semiotic generalization). A spring-node algorithm projected documents into a two-dimensional point-scatter, and the Pathfinder network scaling procedure based on minimum spanning trees was used to construct network maps. In addition, Intramax, a hierarchical clustering technique, was applied to group documents hierarchically into thematic regions. These topic regions are depicted as Voronoi polygon maps. Whereas the first three proximity types relate to associative ontological abstractions (e.g. metric proximity is-a kind of similarity), the region displays are examples of the ontological aggregation principle (e.g. a news story is-part-of a news topic).

In one within-subject factorial experiment participants were presented with a (counterbalanced) series of large-scale monochrome point-scatters, node-link networks, and region displays, each treated with aforementioned proximity measures, that is, metric, linkage, and group membership proximity (independent variables). The wall-sized spatialized displays were projected through a backlit screen projection system. Forty-four participants were asked to judge similarity between news stories depicted as points in the three display types. Similarity judgments (on a metric scale), and participants' response times were digitally recorded during the tests (dependent variables).

Results from this experiment suggests that people are indeed associating metric graphical inter-point distances with semantic similarity of text documents depicted in 2D. However, document similarity along links of a network override similarity judgments between documents based on straight-line Euclidean distance. In addition, a document's membership in a topical region alters people's relatedness judgments on straight-line metric distance. This effect is not as strong as for networks. Graphic complexity in displays does not necessarily slow people down in their similarity assessments. A follow-up study currently underway features a between-subject design with 48 participants to additionally investigate how display size (wall-sized vs. computer screen size) affects people's similarity judgments. This study also examines potential interaction effects that color may have on similarity judgments, specifically for the region displays. The outcomes of the experiments provide important feedback on several levels. First, empirical results will allow establishing spatialization redesign guidelines that will facilitate the generation of cognitively adequate spatializations for knowledge discovery in very large databases. An empirically validated spatialization framework, based on ontological principles and GIScience theory may be transferred to the explicit geographic domain as a basis to reduce current limitations of how geographic space is represented within GISystems. In this way we can generate geometric configurations that best match people's conceptualizations of proximity for the task at hand.

Acknowledgments

Funding by the National Imagery and Mapping Agency (NMA-201-00-1-2005) is greatly appreciated. Thanks are also due to David Mark for his insightful input, discussion and brainstorming throughout this project.

References

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

Tobler, W. R. (1970). A Computer Model Simulating Urban Growth in the Detroit Region. *Economic Geography,* 46: 234-240.