

Representing Information for Knowledge Discovery: Pattern Detection and Database Structure

Barbara P. Battenfield, University of Colorado - Boulder
babs@colorado.edu

Advances in information technologies have dramatically enhanced the volume of geospatial data that can be collected on a daily basis. It is paradoxical that as more data are collected, the more difficult it becomes to sift through, locate items, or identify latent patterns. Scientific and policy communities agree there is a need for new services that can intelligently extract useful information from massive amounts of data in support of decision-making, and to synthesize geographic knowledge.

Data mining comprises one of several knowledge discovery methods. Fayyad et al. (1996) provide a generally accepted definition: “...the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.” Although it is relatively straightforward to find pattern or structure in data, establishing its relevance and explaining its cause are both very difficult problems. Furthermore, much of what can be ‘discovered’ may be common knowledge already, to the expert. Finally, it is difficult to define “relevance” in advance of unanticipated or surprising events. Addressing these problematic issues requires synthesis of underlying theory from database science, statistics, and machine learning.

Visualization forms an important foundation for data mining research because the human visual system accounts for over 70 % of the neurons in the human brain. McCormick et al (1987) define visualization as “... a method of computing... a tool for both interpreting image data fed into a computer, and for generating images from complex multi-dimensional data sets.” The goal is “... to leverage existing scientific methods by providing new insight through visual methods.” MacEachren et al (1992) define it as “... a human ability to develop mental representations that allow us to identify patterns and create or impose order.” Visualization encompasses illustration, hypothesis formulation, pattern identification, knowledge construction, problem solving and decision support.

A summary of the knowledge discovery tasks of *finding* patterns, *reporting* and *representing* the findings, *validating* their significance and *optimizing* computational performance are cross-tabulated with visualization methods in Table 1. Middle columns give examples of relevant methods in database science, statistics, and machine learning. Emerging interest in the common ground between knowledge discovery and visualization is likely to lead to identification of additional techniques in the near future.

Data Mining	DATABASE SCIENCE	STATISTICS	MACHINE LEARNING	Visualization
Find	Association rules	Local pattern analysis, global inferential tests	Neural networks, decision trees	Exploratory visualization, Navigation
Report	Rule lists	Significance and power, Support, Confidence, Lift	Likelihood estimation, Information gain	Confirmatory visualization
Represent	Schema update Metadata and Uncertainty	Fitted statistical models, local or global	Conceptual graphs, Meta- models	Orientation, Visual reasoning, Information Design
Validate	Weak significance testing	Significance tests	Learning followed by verification	Human subjects testing, Usability analysis
Optimize	Reducing computational complexity	Data reduction & stratified sampling strategies	Stochastic search, gradient ascent methods	Hierarchical and adaptive methods, Grand tours

Table 1 (adapted from Yuan et al (2001))

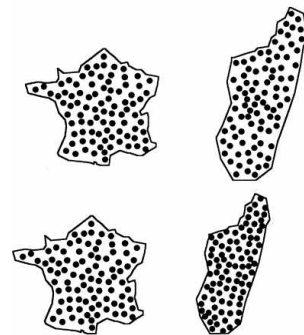


"The Window" Rene Magritte

Visualization synthesizes underlying theory from computational vision, cognitive science, graphical design and image processing. Ralph Waldo Emerson said, "We see only what we are prepared to see". This principle forms the basis of many optical illusions, and is used to advantage in designing contexts for display.

In the case of data visualization, it can be argued that what can be seen or detected is constrained in large part by the information structure, that is, how the representation is organized both in the display and in the database.

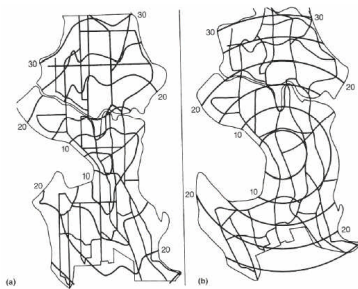
direction, distance from a point or line). It is an intractable problem to transform data and preserve distance and direction simultaneously. Thus the spatial relationships coded into coordinate systems can by themselves reveal or obscure the very patterns one hopes to detect. In the top illustration (taken from Robinson and Sale, 1969), France and Madagascar in an equivalent projection preserve relative size. Point density is comparable between countries.



On the bottom, the projection preserves shape but not size. The concentration of points in Madagascar appears more concentrated (relative to France) than is actually the case.

A database query on relative density based in projected coordinates would generate an incorrect result.

Cartographic practice also transforms attributes to elicit latent patterns. In a "cartogram", a transformation of Seattle Washington restructures coordinates from spatial distance ("as the crow flies") to temporal attributes (driving time from downtown). In the map on the left, street patterns are planimetrically correct (just as on a topographic map). Geographic relationships are linear, based on Cartesian geometry. On the right, travel times in minutes appear as circular arcs, to permit attribute analysis with geometric tools. In this information structure, temporal relationships are linear, based on polar geometry. (Tobler, 1961)



Other attribute transformations include spatialization, a reorganization of data items to redefine proximity on the basis of attribute similarities.

The goal in linking knowledge discovery with visualization is to enable an information management strategy that converts information to knowledge through visual and computational means. Visual means support human acuities for detecting pattern; and these must be protected against missing details or context that are unprepared for. Computational means will establish import and relevance and thereby validate the visual interpretation. The information structure (the spatial referencing, attribute and spatial relationships stored in the database) underlies the validity of visual and computational representation. Just as a specific map projection can modify a cartographic information structure to elicit or distort specific geometric properties, the way information is organized in a database can restructure spatial relationships and/or modify attribute relationships.

To utilize information design and structure in data mining and knowledge discovery tasks to one's advantage at the database level, the research community faces enormous challenge. Information and

database structures must be implemented in such a way as to permit both geometric and semantic transformations. Three areas of research ongoing at University of Colorado come to particular attention.

One is the area of multiple representations, to accommodate sensitivity of geospatial information to spatial resolution. We know for example that spatial enumeration creates often dramatic variation in statistical summaries; this is referred to as the Modifiable Areal Unit Problem (MAUP). Current association rule mining software requires classed attributes for input. We need to systematically study the impacts of attribute granularity and of feature resolution on rule formation and confidence, to establish the influence and potential for scale dependence in data mining outcomes and validation. A project is ongoing at Colorado to study scale effects on association rule mining outcomes (Buttenfield, Mennis, and Liu, 2004). We additionally need to develop database objects encapsulated with scale-dependent behaviors, analogous to the 'smart maps' that add street labels depending on the scale of display. For example, a 'smart' database object should link multi-scale representations to engage seamlessly in local to global inference and reasoning, and to reduce the potential for error following database updates.

A second area for research is temporality. In the case of epidemic disease and other sudden but expected phenomena, events do not take place in regular increments of time. In the case of a disease outbreak, morbidity rates will be uneven and sparse, with many cases reported on one day, few or none the next. Customarily, we adopt a "thin slice" sampling frame. Sampling practice aggregates data in a coarse timeframe to insure a minimum of empty records. For sudden and unexpected events (e.g., wildland fires, tornados, toxic chemical spills), regular time sampling will result in a sparse matrix, and a coarse frame may miss an event altogether. We need to design sampling strategies that organize data within a varying resolution "thick slice" sampling frame, such that new data points are recorded only when events occur. A recent project at Colorado (Huff, 2001; Buttenfield and Huff, in preparation) created initial designs for thick-slice information structures. Varying resolution information structures will complicate statistical computations because of the need to compensate for the 'size' (duration) of heterogeneous sampling.

A third area for research is the representation of uncertainty. Current metadata summarizes positional and attribute uncertainty for a data layer as a whole. We know that uncertainty can vary within a dataset, and yet current commercially available geospatial data models (ranging from an entire Geodatabase down to a node in a TIN network) cannot accommodate uncertainty representation at the individual item level. Incorporation of uncertainty information at the item level would strengthen numerical capabilities for refining estimates of confidence and relevance in data mining outcomes. This is an area of the author's current research.

Cited Work

Buttenfield and Huff. Event-based temporal sampling and data modeling. (in preparation)

Buttenfield, Mennis, and Liu. Data Resolution as a Possible Bias in Association Rule Mining. In Preparation for AAG Annual Meetings, Philadelphia PA, March 2004.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. 1996 From Data Mining to Knowledge Discovery: An Overview, In Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R (eds.) *Advances in Knowledge Discovery and Data Mining*, (Cambridge, MA: AAAI/MIT Press), p. 1-34.

FGDC 2003 *Homeland Security and Geographic Information Systems*. Federal Geographic Data Committee. (<http://www.fgdc.gov/publications/homeland.html>)

Huff, R. 2001 *GIS Data Models and the Representation of Time*. Unpublished MA Thesis, University of Colorado.

MacEachren, A., B. Buttenfield, J. Campbell, D. DiBiase, and M. Monmonier. 1992. Visualization. In *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*, 101-137. New Jersey: Rutgers University Press.

McCormick, B., T. DeFanti, and M. Brown. 1987. Visualization. *Scientific Computing in Computer Graphics*: 21 i-E-8.

Robinson, A. and Sale, R. 1969 *Elements of Cartography*. 3rd Edition. New York: Wiley.

Tobler, W. R. 1961 *Map Transformations of Geographic Space*. Unpublished Ph.D. dissertation, Dept. Washington.

Yuan, M., Buttenfield, B., Gahegan, M. N., and Miller, H., 2001. *Geospatial Data Mining And Knowledge Discovery*, http://www.ucgis.org/priorities/research/research_white/2000whitepapersindex.htm