SPATIAL METAPHORS FOR VISUALIZING INFORMATION SPACES

André Skupin
National Center for Geographic Information and Analysis
Department of Geography, 105 Wilkeson Quadrangle
State University of New York at Buffalo, Buffalo, NY 14261
voice: (716) 645-2722 ext. 32
email: skupin@geog.buffalo.edu

Barbara P. Buttenfield
Department of Geography, Campus Box 260
University of Colorado, Boulder, CO 80309
voice: (303) 492-3618
email: babs@colorado.edu

The growing volume and complexity of the World Wide Web creates a need for new forms of interaction with information. Spatial metaphors have been in the focus of interface research for a number of years. Recently, a related concept called *spatialization* has emerged as one possible strategy for dealing with modern information glut. However, the term remains ill-defined. We present a definition of spatialization that is based on the notion of *information spaces* that are non-spatial and high-dimensional. Through spatialization, they are projected into a low-dimensional form and made accessible for visual interpretation.

We implement this method to a body of about 100 newspaper articles. Following the extraction of keywords for each article, a multi-step process is applied. It involves the construction of a vector-space model, the computation of a proximity matrix and the projection into two dimensions via multidimensional scaling. The resulting coordinate configuration is imported into ArcView and linked with the keyword list. A number of visualization examples are shown, all based on a representation of each article as a point. One goal of this research is to investigate the feasibility of applying cartographic expertise to spatialized representations. Cartographic generalization is among the tools that could provide valuable inspiration for the visualization of large information spaces.

INFORMATION SPACES

Information is that which is inherent in a set of facts (Oxford Dictionary, 1996). An information space provides a well-defined strategy for organizing information. It can be formalized by logic, mnemonics, metric or nonmetric coordinates. The goal of the strategy is to facilitate navigation, browsing and

retrieval of items. One creates a structure to support access to the content, in effect.

Information spaces can be distinguished by the ways in which structure and content are interwoven into a specific physical and conceptual form. They are by no means new artifacts, introduced by late-20th century technology. Instead we have been surrounded by and interacted with information spaces for a long time. Newspapers are good examples. They contain chunks of information that is neatly organized into articles that are placed in physically defined locations on a page. To those that are familiar with the layout of a specific newspaper, it is an easy task to find and retrieve articles dealing with a certain topic, for instance the latest sports scores or developments in local politics. Given its relatively small volume, familiar organizational scheme, and physical nature, a conventional newspaper information space is relatively easy to navigate.

Other information spaces are more difficult to navigate. This may be due to the sheer volume of contained information, the non-physical nature of the storage or browsing medium, or to novel ways in which content is structured. Perfect examples are large hypermedia spaces, such as the World Wide Web.

SPATIALIZATION

Definition

In recent years it has been realized that new kinds of information spaces will require new methods for access. Among the most discussed strategic tools is the employment of spatial metaphors. Spatial metaphors are at the heart of a concept called *spatialization*. That term is applied in a variety of contexts, notably in digital audio processing where spatialization facilitates the identification of the location of sound sources in three-dimensional space.

Lakoff's (1980) use of the term spatialization is the most influential as far as the application of spatial metaphors in user-interface research is concerned. Kuhn (1992, 1996) introduced "spatialization" into the GIS interface jargon. Nevertheless, the term "spatialization" remains ill-defined. One common tendency is to use it synonymously with "the application of spatial metaphors". Spatialization can be defined more rigorously and literally, by establishing that formal spatial characteristics of distance, direction, arrangement, and pattern have or have not been achieved.

We define spatialization as

> a projection of elements of a high-dimensional information space into a low-dimensional, potentially experiential, representational space.

Information spaces are generally high-dimensional, given the complex, multifaceted character of their contents. Since the goal of spatialization is the creation of a cognizable representation, the latter has to involve fewer, typically two or three dimensions. It appears appropriate to use the term projection for the occurring transformation. Spatialization applies formal criteria to project a

view of contents into a reduced or simplified arrangement. Spatial metaphors provide natural strategies for orientation and navigation. Other aspects of spatial relations should also be established. These might include (for example) ascertaining that distances are commutative, that spatial autocorrelation applies to contents, or that changes in scale increase the level of apparent detail, in the spatialized solution.

Related Research

Hypertext and hypermedia have long espoused the idea of supporting navigation and retrieval through graphical representation of their structure and content. These representations have gained renewed attention with the advent and continuing growth of the World Wide Web. The majority of these visual representations is two-dimensional. Traditionally, they have been called *maps*. There are several principle ways in which visual representations of hypermedia spaces can be created. Some are merely manually created two-dimensional bookmark maps. Others, and those are the most common ones, are based on *structural characteristics* of the hypermedia space (Woods 1995, Mukherjea & Foley 1995). A third approach derives low-dimensional visualizations based on an analysis of the textual content of hypermedia spaces. It is mainly this approach that is being addressed by this paper.

Efforts are now being made to unify Web visualization with a more general file space visualization, exemplified by Apple's HotSauce, based on the Meta-Content Format (MCF).

In recent years much research effort has been invested into the investigation of spatial metaphors for user interfaces (Mark 1992, Dieberger 1994, Kuhn & Blumenthal 1996). Much of this is relevant and related to our notion of spatialization. Refer to Kuhn & Blumenthal (1996) for an interesting overview of the subject in tutorial form.

Surveying Information Spaces

In order to meaningfully spatialize information it has to be broken down into meaningful units or 'chunks'. This can be illustrated by evoking the image of a topographic surveyor who chooses to measure those surface points that have geometric or semantic relevance.

What are the units into which information can be divided? Some information spaces might appear to have a structure with an inherent "sampling unit". Examples for natural sampling units could be chapters of a book, single Web pages or newspaper articles. All these are, however, meaningful only at a defined level of interest. For example, there are instances when the focus is on a whole web site instead of a single web page. One might want to compare and correlate all the books on the shelf rather than all the chapters of a single book. What we are dealing with is the concept of scale. Like the surveyor, we have to consider both the intended scale and the purpose of our future representations in choosing meaningful sampling units.

As mentioned before, the goal of spatialization is to project contents of an information space into an easily cognizable representational space. In order to make the process consistent, its criteria have to be well-defined. One important aspect to consider is that elements of information spaces can be related to each other in many different ways. For instance, web sites can be related through such factors as content, connectivity, lineage, or geographical location. The combination of these factors forms certain configurations in an high-dimensional information space. It is the assumption of the spatialization approach that the metric qualities of these configurations can be numerically expressed and projected into a low-dimensional geometric space.

SPATIALIZATION OF A NEWSPAPER INFORMATION SPACE

Sampling the Information Space

Two editions of the New York Times, dated November 7 and November 8 1995, were chosen as input. We applied the spatialization concept to the 96 articles contained in Section "A". That time was just after the assassination of Israeli Prime Minister Yitzhak Rabin and the news contained a large number of articles highlighting various aspects of that event. Prominence of these current events should be reflected in the final information space.

As sketched above, the choice of a sampling unit is one of the most critical early decisions. The single newspaper article appears to be the best candidate, considering its coherency and its limited size, relative to the newspaper as a whole. Some other choices, like a division of the newspaper into pages with even or odd page numbers, would appear quite meaningless. However, in a sufficiently large newspaper, it might make perfect sense to divide its content into pages, each of which is dedicated to a certain subject area, like "Baseball", "Football", or "Hockey".

Next, a criterion must be chosen to distinguish articles from each other. Theoretically, there are again many choices. One could refer to an article according to its physical location in the newspaper, e.g. "Page 6, upper right corner". In fact, such a scheme can be useful when locating updated articles within well-known structures. For instance, in a certain newspaper the baseball scores might always be found in a certain location. Other factors could include the length of an article or the number of photographs associated with it.

Our approach assumes that the newspaper information space is only a special case of a much larger group of information spaces, including the WWW, to which the chosen method should be applicable. It becomes obvious that one factor will take precedence, namely, the content of the articles. It makes indeed little sense to compare "page 6, upper right corner" with "http://www....", but a comparison of their actual content can bear useful results. The content of articles forms the basis for our spatialization.

Technical Concept

The spatialization performs a content-based projection from the newspaper information space into a map space. This requires the definition of two major factors:

(1) Configuration of articles in the information space

(2) Projection method

Configuration of articles in the information space. This refers to the location occupied by each article in the high-dimensional information space. Following the principles of vector-space modeling (Salton 1989), this idea can be taken quite literally. In a vector-space model the occurrence of keywords in each article determines its location in an n-dimensional information space (n = total number of unique keywords).

If we assume that the chosen keywords sufficiently express the content of all articles, then the distance between articles in the n-dimensional information space is equivalent to their similarity. This is the assumption behind the widespread use of the vector-space model in many search engines, for instance on the World Wide Web. The principle is to use one or more search terms as input, form a vector of terms, and compare it to a stored list of vectors, each of which represents the contents of a web page. The resulting numerical values are the similarities/distances between the search term[s] and the web pages.

Projection Method. Tobler's first law of geography states that "everything is related to everything else, but near things are related more than distant things" (Tobler 1970). One of the primary assumptions of our approach is that a believable representation should be in accordance with that rule. Since the vector-space model already produced a configuration that expresses similarity through distance, a projection method is needed that strives to preserve these distance relationships.

One such method has been utilized by social scientists for many years: *Multidimensional Scaling* (MDS). It is a procedure that can be employed to transform a high-dimensional configuration, given in form of a proximity matrix, into low-dimensional coordinates. Over the course of several decades a variety of MDS algorithms were introduced and tested (Torgerson 1958, Kruskal 1964, Sammon 1969, Carroll & Chang 1970). The ALSCAL procedure (Takane, Young, and De Leeuw 1977) became the MDS method of choice for many statistical software packages, like SPSS and SAS.

Vector-Space Model. For each article a number of keywords was manually extracted. It was quite impossible to collect an equal number of keywords for each article, which varied in (a) the total length, i.e. word count, and (b) the level of generality. Some extremely short articles did not contain enough substance to produce more than five keywords. Other articles highlighted so many facets of a subject that even fifteen keywords hardly sufficed. On average, the essence of an article could be captured with the extraction of about ten keywords. Keyword extraction was performed independently by each co-author, and keyword sets compared to check for consistency.

A total of 415 unique keywords was extracted from 96 articles. They were merged to form a term vector T:

$$T = [t_1, t_2, \ldots, t_i] = ["AIDS", "advertisement", \ldots, "Rabin", \ldots, "Zyuganov"]$$

By matching vector T against each article individually, a term-article matrix (size 415x96) is created, with values of "1" indicating the presence of a keyword in an article and "0" indicating its absence. As a result, each article is characterized by a certain arrangement of "1" and "0" in one matrix column. That column vector describes the location of each article in the information space. The distance/dissimilarity of two articles can be computed by comparing column vectors. A variety of proximity coefficients exist to fulfill that purpose and the choice between them is somewhat arbitrary. After several tests, a Euclidean proximity measure was chosen:

$$\Delta_{jk} = \left[\sum_{i=1}^{n} (X_{ij} - X_{ik})^2 \right] \quad \text{(Sneath \& Sokal 1973)}$$

(n = 415; X = term-article matrix)

By applying this Euclidean measure to every pair of articles (n=415), a dissimilarity matrix is created (size 96 x 96). This matrix is input to the ALSCAL procedure in SPSS. The output is a two-dimensional configuration with coordinates for each article.

SPATIAL PRINCIPLES IN THE INFORMATION SPACE

Earlier in this paper, we argued that a rigorous spatialization must establish the presence of spatial metaphors such as distance, direction, arrangement and pattern. These four characteristics can be used to build up compound metaphors (autocorrelation, region building, intervisibility, etc.) The remainder of this paper will demonstrate two concepts, namely region definition and scale-dependence, in the newspaper information space. We begin with simple visualization and exploration.

Visualization

With the help of desktop mapping tools, the two-dimensional coordinates of each article can be linked with respective keywords. Figure 1 shows a point visualization in which a number of points have been labeled. Each point represents one article. The labels are created by accessing the first keyword identified for each article. Notice in the Figure that the keyword "Rabin" appears several times in the plot, thus we can determine that this visual display does not preserve spatial uniqueness: the same information "place" can appear in multiple locations. One might use this property to advantage, for example by linking a network between regions (article clusters, in the plot).

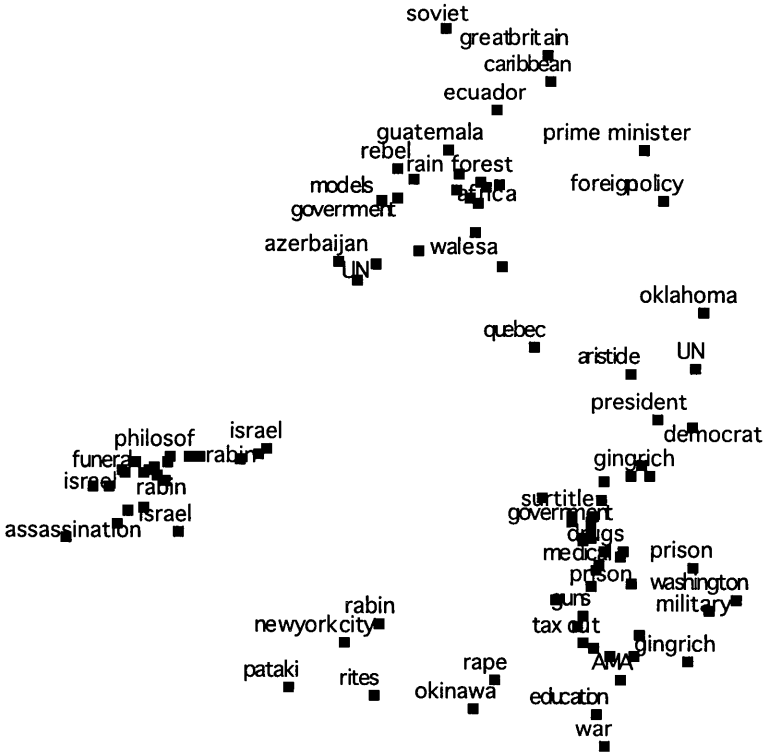

Figure 1. Spatialized Solution Represented in ArcView

One can perform simple queries that are common in desktop GIS. Figure 2a illustrates how the graphical selection of a number of articles corresponds to a highlighted section of the table. (The scatter plot in this Figure is a reduced version of Figure 1). In the selection rectangle, all articles refer to foreign events, from a U.S. perspective. In Figure 2b an attribute search is performed, searching for all articles with the keyword "assassination". Corresponding points are highlighted in the plot.

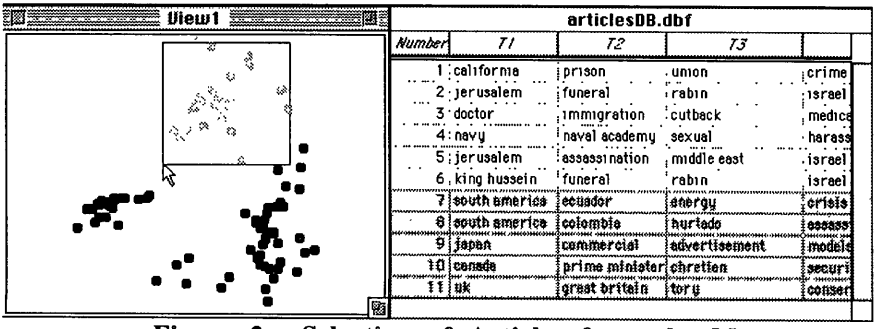

Figure 2a. Selection of Articles from the Map

These simple visualizations confirm three main clusters of articles: (1) domestic events in the lower right corner, (2) foreign events in the upper half, and (3) events in Israel on the left. The articles concerning the assassination and funeral of Yitzhak Rabin form a distinct and distant cluster. The exceptions are three points in the lower left of the map. These refer to articles surrounding Rabin's assassination that were related to the U.S.. Examples are the Jewish mourning in New York City and U.S. politicians attending the funeral in Jerusalem.

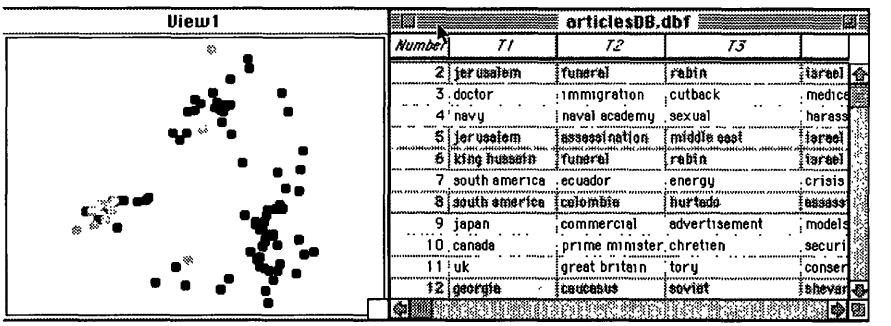

Figure 2b. Selection of Articles from the Table

Generalization and Scale-Dependence

As the focus of interest changes, e.g. from single books to single book shelves, so must the visual representation of information spaces change. Two options exist for obtaining more abstracted or more detailed representations.

One option is to initiate a new spatialization, with different sampling units. This would involve recomputing coordinates. Locations (and thus regions) in the new information space would not be comparable to those in the first. The other option involves a process that cartographers call generalization.

Its application to spatialized representations is most intriguing. Appropriate generalization permits exploration of the rate at which information densifies as scale changes, and will additionally preserve relative locations, permitting multi-scale analyses of the information space.

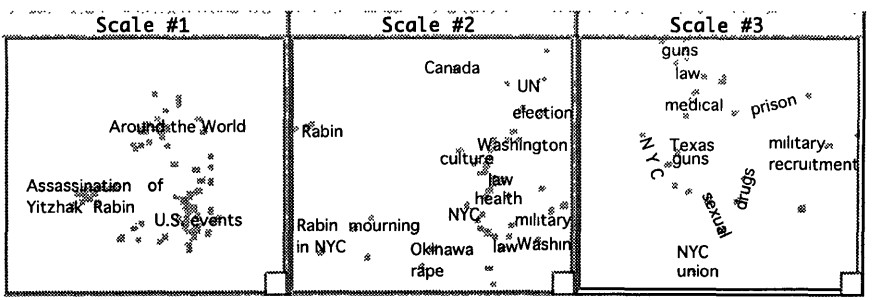

Figure 3. Visualization of spatialized data at three scales

Figure 3 shows a simple example of generalization applied to the New York Times information space. Three maps are shown, each at a different scale, with more specific content revealed as we zoom in. The center point is the cluster of U.S. events; as new keywords are resolved we can better define the fabric of the local information space.

SUMMARY AND SEARCH FOR MEANING

To the casual observer, the spatialization example may appear quite simple and almost trivial. One has to bear in mind that the complexity and sophistication of spatialized representations has to be in tune with the degree to which we can attribute meaning to each of the graphical components.

The processes leading up to the geometric configuration are complex. We must be careful to understand the mathematical and statistical assumptions underlying the geometry before reliable and meaningful interpretations of spatial relationships can be made. This paper is a 'proof-of-concept' demonstrating that existing statistical tools can be applied to generate information spaces, and that the presence or absence of simple spatial metaphors can be established to explore collections of information. As we move towards larger information collections, and towards more complex representations, one can envision three-dimensional models of information 'terrain analysis'. New questions might be asked, about the meaning of slope, intervisibility as a metaphor for indexing or cross-referencing. Other spatial analytic tools might be applied with varying degrees of effectiveness.

The spatialization of information spaces is an important application for geography. Early efforts for the mapping hypermedia structures were frustrated by the complexity of large hypermedia documents. In the late 1980's many hypermedia researchers even concluded that complexity stood in the way of navigating such documents and that spatial metaphors were unfeasible. What they ignored was that there existed a field of science and technology that had a wealth of experience in dealing with graphic complexity: cartography. Skupin & Wieshofer (1995) point out cases in which proven cartographic principles are being basically reinvented. Nielsen's (1990) ideas of "clustering" and "link inheritance" are examples. With the growth of the WWW, spatialized representations of hypermedia have found a renewed interest, but such notions as scale and region building remain virtually unknown in the hypermedia research community. This is a wide open field and cartographers have yet to discover it.

REFERENCES

Carroll, J.D., Chang, J.J., 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 35: 238-319.

Dieberger, A., 1994. Navigation in Textual Virtual Environments using a City Metaphor. Doctoral Thesis. Vienna University of Technology. Vienna, Austria.

Kruskal, J.B., 1964. Nonmetric Multidimensional Scaling. Psychometrika. 29(2): 115-129.

Kuhn, W., 1992. Paradigms of GIS Use. Proceedings 5th International Symposium on Spatial Data Handling, Charleston. IGU Commission on GIS.

Kuhn, W., Blumenthal, B., 1996. Spatialization: Spatial Metaphors for User Interfaces. Department of Geoinformation, Technical University Vienna.

Mark, D., 1992. Spatial Metaphors for Human-Computer Interaction. Spatial Data Handling. Proceedings 5th International Symposium on Spatial Data Handling, Charleston. IGU Commission on GIS.

Mukherjea, S., Foley, J.D. , 1995. Visualizing the World-Wide Web with the Navigational View Builder. Computer Networks and ISDN System, Special Issue on the Third International Conference on the World Wide Web '95, April 1995, Darmstadt, Germany.

Nielsen, J., 1990. Hypertext and Hypermedia. San Diego: Academic Press.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Publishing Company.

Sammon, J.W., 1969. A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers. C-18 (5): 401-409.

Skupin, A., Wieshofer, M., 1995. Cartography at the Hypermedia Frontier: Animation and Visualization of Hypertext Structures as two Examples. In: Mayer, F. (ed.) Wiener Schriften zur Geographie und Kartographie, Band 5.

Sneath, P., Sokal, R., 1973. Numerical Taxonomy. San Francisco: W. H. Freeman and Company.

Takane, Y., Young, F., De Leeuw, J., 1977. Nonmetric Individual Difference Scaling: An Alternating Least Squares Method with Optimal Scaling Features. Psychometrika, 42: 7-67.

Tobler, W. , 1970. A Computer Model Simulating Urban Growth in the Detroit Region. Economic Geography . 46 (2): 234-240.

Torgerson, W.S., 1958. Theory and Methods of Scaling. New York: John Wiley.