The Ecological Approach to Text Visualization

James A. Wise
Integral Visuals, Inc., 2620 Willowbrook Avenue, Richland, WA 99352. E-mail: JamesAWise@aol.com

"Words and rocks contain a language that follows a syntax of splits and ruptures. Look at any word long enough and you will see it open up into. . .a terrain of particles, each containing its own void. . ." Robert Smithson (1996)

This article presents both theoretical and technical bases on which to build a "science of text visualization." These conceptually produce "the ecological approach," which is rooted in ecological and evolutionary psychology. The basic idea is that humans are genetically selected from their species history to perceptually interpret certain informational aspects of natural environments. If information from text documents is visually spatialized in a manner conformal with these predilections, its meaningful interpretation to the user of a text visualization system becomes relatively intuitive and accurate. The SPIRE text visualization system, which images information from free text documents as natural terrains, serves as an example of the "ecological approach" in its visual metaphor, its text analysis, and its spatializing procedures.

This article both formalizes Smithson's evocative prose and responds to Steven Eick's recent challenge (Eick, l997) to proceed to a real "science of information visualization." It describes the theoretical rationale and technical basis of two years of research investigations at Pacific Northwest National Laboratory (operated for the Department of Energy by Battelle, Inc.) on the Spatial Paradigm for Information Retrieval and Exploration (SPIRE) project, which the author co-created and managed.

The SPIRE project is funded by the Department of Energy and the U.S. intelligence agencies to determine if it is possible and practicable to find a means of "visualizing text" in order to reduce information processing load and to improve productivity for intelligence analysis. Most intelligence information is in prose, in the form of cables, reports, and articles, and it is not unreasonable for 30,000 documents to cross the electronic desk of an analyst every week. There is no way that a person could read, retain, and synthesize even one-half of 1% of these. Clearly, another way was needed to both represent the documents and their contents, while permitting their rapid retrieval, categorization, abstraction, and comparison, without the requirement to read them all.

What became known as the SPIRE project began in January 1994, with the somewhat loquacious title "Multidimensional Visualization and Browsing." The first product software, "Galaxies" (described by Crow et al., 1994) was produced in August of that year and delivered to the Army's Pathfinder Program for field tests. The second product software, "ThemeScapes™,"[1] was demonstrated in an alpha version at the Automated Intelligence Processing and Analysis Symposium in March of l995 (Pennock & Lantrip, 1995; Wise & Thomas, 1995) and in a beta version at Information Visualization '95 in September of that year (Wise et al., 1995). Those short, descriptive publications focus on project and software performance, and hardly convey the breadth of theoretical rationale and depth of technical research that actually underlie the SPIRE project. This article attempts to redress that shortcoming, placing the SPIRE visualizations in the context of their full "ecological approach" to text visualizations, acknowledging their technical bases in previously published work (e.g., Chalmers & Chitson, 1992; Chalmers, 1993; Pazner, 1994), and responding to Eick's (1997) challenge to provide a scientific foundation for information visualization.

The "Ecology" of Text Visualizations

Ecology may seem like a strange term to use in reference to visual information retrieval interfaces (VIRIs), as it is customarily understood to refer to the "science of the relations of organisms to their environment." Yet it is crucial to understanding the distinctive viewpoint that was taken from the beginning of the SPIRE project, which from its inception sought a coherent and comprehensive approach to the analysis and visualization of text in a manner that best

[1] ThemeScape™ is now a trademarked term of Cartia, Inc.

utilized human analysts' native perceptual abilities. It posed the question: How should an analyst view and manipulate a visualization with features directly determined by text characteristics relevant to the analyst's task, and that appear in visual forms which everyone, both genetically and experientially, already knows how to interpret? The project's task was thus seen to be one of creating a "synthetic ecology" for prose that incorporated, analogously or literally, visual features of the natural world within which human visual perception has so successfully operated. This viewpoint on what text visualizations should be thus became the "ecological approach" in four ways:

- Steps from text analysis to visualization are tuned to reflect the needs of ecological vision.
- Visualizations are built as "emergent forms" in the manner that natural visual patterns originate.
- Visualizations access processes of human's "ecological perception" and are thus intuitively interpreted.
- The visualization and correspondent analyst's perception are co-determined or "enacted," within an ecological context of guided activity, analogous to processes of color perception (Thompson, Palacios, & Varela, l992).

According to the "ecological approach," then, one surprising implication is that successful visualization of text requires that text analysis—the process by which words are re-represented mathematically so as to provide the computational basis for visualizations—is best undertaken by working backwards from the visualization itself to see what is computationally optimal for its support. This forgoes the traditional "information retrieval" view of text analysis, and instead sets up the comparative analogy of how the retina of the eye begins to construct visions of the natural world. When text analysis is approached this way, some traditional "intractable" information retrieval problems simply vanish.

Another inference of the "ecological approach" is that text visualizations ought to take advantage of the visual appearances of natural forms that humans have learned to interpret visually as part of the biological heritage from their species' history on the Earth. It was not accidental that the "Galaxies" visualization invoked the metaphor of documents as stars in the night sky, or that ThemeScape™ s represented themes as sedimentary layers that together create the appearance of a natural landscape. These are fundamental visual experiences of our world that people have incorporated and responded to for eons. They both carry a natural interpretation that does not require instruction or prolonged training to appreciate and use.

The "ecological approach" also redirects attention away from seeing a visualization as an "illustration" of a preexistent semantic form, and correctly characterizes it as the end result of an interactive process between the observer and the informational content of the prose. As Stoner (1990) so tersely expressed it: "Structure represents the product of information interacting with matter." A text visualization's structure should then result from and reflect a process history, being intimately bound up with "how it got that way." In the "ecological approach," text visualization is not a process of "drawing pictures." It is a result of transferring to the spatial realm the results of computational processes that are themselves analogs of the means by which physical forms are produced.

Finally, the "ecological approach" analogously resurrects and adopts Gibson's (1979) view of perception as being mediated and guided by the actions of the observer. This directly addresses the human-computer interaction (HCI) issues of text visualizations and suggests that the electronically mediated means of exploring visualizations should analogously reproduce, as far as possible, the ways we sensually explore objects in the natural world. This became the least well-developed aspect of the SPIRE software, although the ThemeScape™ 's probe tool (Wise et al., 1995) and other work on "intuitive user interfaces" (Wise, 1996; Lopresti & Harris, 1996) show how gesturing and audition can greatly enhance the usefulness and experience of information visualizations.

The SPIRE Process of Text Visualization

As implemented in the SPIRE project, there was a five step process to preparing a Text Visualization (Fig. 1):

1. The system received a corpus of unstructured, digitized text documents. There was no use of keywords, no dictionary, no preestablished topics or themes extracted, and no predefined structure to the text that would have tied the resultant visualizations to any particular text analysis.

   As prepared for use by the Intelligence Community, the visualizations were meant to optimally handle news stories, resumes, e-mail, letters, abstracts, short articles, communiques, etc. The upper limit of the number of documents that can be processed in such a corpus is set by a number of factors irrelevant to this article. These include the processing power of the computer, the screen space of the display (for the Galaxies visualization), the text analysis method and projection algorithm used, and the time demands of the analyst's need. But visualizations of corpora numbering up to 6K documents were routine in the project's investigations. While this is a relatively small number of documents with respect to other systems, the goal of the research was to develop new approaches to text visualization, and problems of scaling to large document sets were left for later study.
2. The digitized documents were then analyzed (via a resident text engine) to characterize them as high dimensional vectors. The SPIRE used the vector space model (see Salton, 1991) exclusively. In such a model, each text document is represented as a high-dimensional vector, which can be constructed using a variety of techniques. The exact means is not important as long as a statistically "rich," reasonable sized dimensional representation results. The two commercial text engines used in this research represent somewhat polar approaches to the same problem of vector construction.

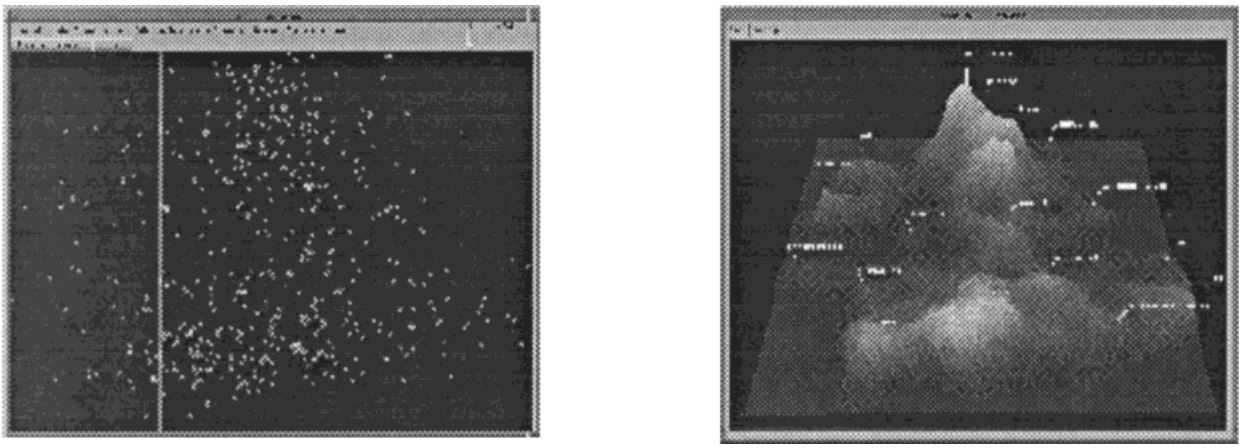

FIG. 1. The Galaxies and ThemeScape™ s text visualizations of the SPIRE system (1995).

The first Galaxies visualization in l994 relied on a commercially available text engine that utilized a dictionary of 200,000 words. If a word in the dictionary appeared in the document, a "1" appeared at that place in the vector. Otherwise a "0" was assigned. This resulted in each document being represented by a binary vector 200,000 (!) units long, which significantly restricted the kinds of manipulations one can perform (and subsequently the kind of visualization one can produce). The lesson is that text analysis approaches, which may be perfectly suitable for information retrieval, may be relatively unsuitable for text visualization.

When the ThemeScape™ 's visualization was prepared (early l995), the text analysis was performed by the Matchplus™ text engine created by HNC, Inc. This engine used a neural net trained on a document corpus within a given domain (general news or financial articles, etc.) The neural net had 280 output nodes, resulting in a 280-dimensional vector for each document. With their continuous numerical output, the 280 nodes also provided a much improved document representation with shorter vectors that allowed the ThemeScape™ visualization to be constructed. Using the experience gained with these text engines, the SPIRE team was later able to develop its own text engine, optimized for visualization purposes (see section: Text Analysis from a Visualization Basis).

Given any of these ways of constructing a vector space over the documents, a metric is then placed on that space to represent similarity in the content of the documents. The most common ones are the Euclidean distance or cosine measures (Salton & McGill, l983). The first is based on the sum of the squares of the differences between a pair of documents on every dimension. The second is based on the differences in the angles of the document's vectors from the origin of the space. Finally, the document vectors are normalized for the size of the document, and the next step of clustering the documents can begin.

3. Using the normalized document vectors, the documents were then clustered in the high-dimensional space (Frakes & Baeza-Yates, l992). This produced topical groups of documents where each document was assigned to only one cluster. The SPIRE visualizations were not designed to be dependent on any one particular clustering method. The K-Means and complete linkage hierarchical clustering approaches were both studied in-depth, and found to be satisfactory for up to ~5K document sets. Their algorithms are widely available.
4. The next step was to project the high dimensional document vectors and their cluster centroids down onto a two-dimensional plane. This plane provided the groundplan for both the Galaxies and ThemeScape™ 's visualizations.

   Again, different projection techniques are available. For small (up to 1.5K) document sets, multidimensional scaling analysis (MDS) (see Shepard 1962a,b) is sufficient. Larger document sets required development of the team's own projection algorithm, which we called "Anchored Least Stress" (ALS).
5. The final step was the construction and display of the Galaxies and ThemeScape™ 's visualizations based on the positions of the document in the second ground plane. While Galaxies represented the documents directly as blue-green "docustars" in a night sky with orange cluster centroids, ThemeScape™ 's used the document positions as points from which to build up a landscape representation when thematic terms taken from the documents were successively layered over the groundplane.

   The complete sequence of those five steps is schematically represented in Figure 2.

Clustering and Projection of Documents for Visualizations

Since similarity of document content equals document placement in the high-dimensional space, preserving high-dimensional spatial relations among documents in their clustering and projection to "visible" spaces is essential in this and similar schemes of text visualization. The simple requirement is that proximity of visible Euclidean distances in the plane be proportional to distances and topical similarities among the documents in the high dimensional representation.

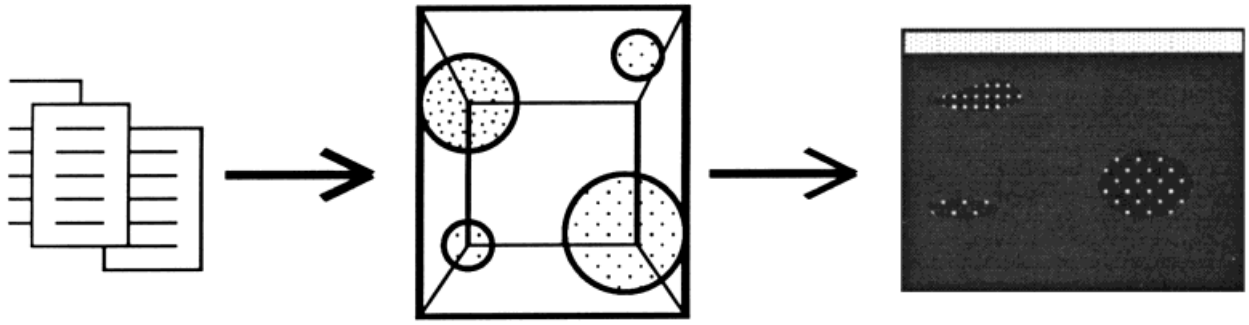

FIG. 2. The sequence of steps leading to the text visualizations of the SPIRE system.

The clustering approach designed to handle large document sets was developed primarily by Jeremy York of the SPIRE team. It was called "Fast Divisive Clustering."

The process begins with the analyst selecting the number of clusters he or she wants to contain all of the documents to be visualized. This number needs to be heuristically determined by the analyst's experience and prior knowledge of the document set, and could be the result of a more formal, Bayesian analysis in itself. This number sets the number of cluster "seeds" distributed in the high-dimensional space. The seeds are distributed randomly, and subspaces are then sampled to ensure that seeds have not inadvertently ended up too close to one another. Then non-overlapping hyperspheres are defined around each cluster seed, and all documents in the high-dimensional space whose coordinates fall within a cluster's hypersphere are assigned to that cluster. Through an iterative procedure, the center of mass defining a new centroid is calculated for each cluster, which shifts its corresponding hypersphere. As hyperspheres shift, documents drop out and are assigned to correspondingly new clusters for hyperspheres that now enfold them. After a few iterations, the cluster centroids stop shifting above an assigned threshold, and documents take their final cluster memberships.

This approach remained under refinement and was experimental until the end of government fiscal year l996, when SPIRE development was transferred to a privately funded company outside the laboratory.

While MDS is a "tried and true" technique in the psychometrics of information retrieval (see Rorvig, l988) it has severe shortcomings for dimensionality redirection as document sets become large. Multidimensional scaling analysis uses pairwise distances (Euclidean or cosine angle) and attempts to minimize a measure on differences in pairwise distances ("stress") between high- and low-dimensional document positions. The intent is to preserve distance relations between documents in the high-dimensional space as they are projected into the two-dimensional one. As the number of documents, n, grows, the number of pairwise distances to be considered grows as a simple quadratic, producing an exponential increase in computational complexity. This significantly increases the requisite processing time. The first Galaxies runs in l994 on Sparc 3 workstations, for a few hundred documents could take 12 hours when using binary vector representations.

York (1995) cleverly found a way around the MDS projection bottleneck through the ALS approach. Beginning with cluster centroids (that are two-dimensional) based on an initial clustering of the documents, a document's iterative projection and placement is based on a vector of its distances to the different cluster centroids, not its pairwise distances to all other documents. The document is ultimately placed in the 2-D plane so that its position reflects its similarity to every cluster, *not* every other document. This used a computationally simple linear regression solution that constructed a new vector for every document which contained the distances of that document to each cluster centroid, then minimized the squared differences between the observed and fitted distances to the centroids. Linear Principle Components Analysis was used to initially project the cluster centroids onto the 2-D plane. Its algorithms are widely available.

Anchored Least Stress also made a qualitative difference in the way it treated distances with respect to traditional MDS. In MDS, fitting all of the pairwise distances among documents means that small deviations among points are

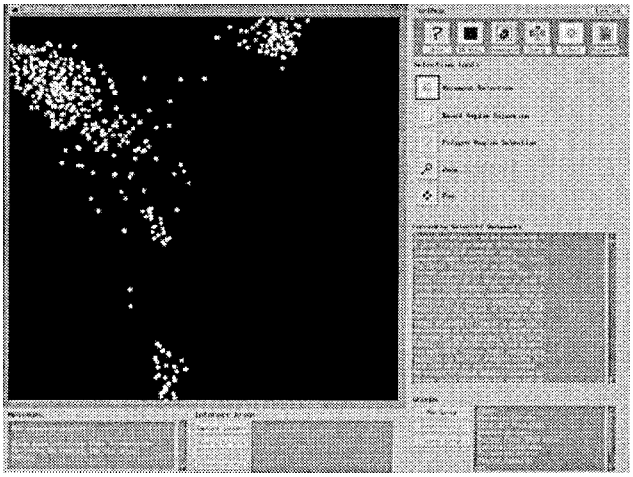

FIG. 3. The first "Galaxies" visualization software product displaying documents of a technology database.

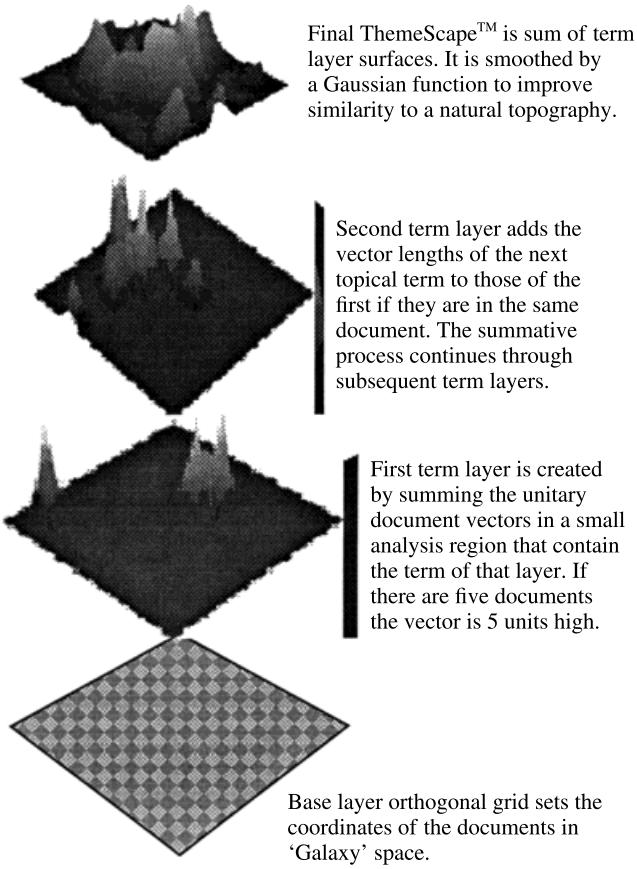

FIG. 4. The process of constructing a ThemeScape™ type representation that mimics sedimentary deposition.

placed at high relative importance. Under ALS, it is the large deviations that are considered important. Smallscale differences in the projected placement of points onto the second plane are sacrificed somewhat in order to arrive at a better and faster overall largescale solution. Since so much information is necessarily lost in compressing high-dimensional spaces down to the 2-D plane anyway, it seems as if this is a worthwhile price to pay to overcome a major computational bottleneck to the visualization process.

Construction of Visualizations

The original Galaxies visualization was essentially a "starfield" of documents in a type of display seen previously in visualizations like "Filmfinder" (Ahlberg & Schneiderman, l994) and IVEE (Ahlberg & Wistrand, l995) and now commercialized in a product called "Spotfire." The ubiquity and usefulness of scatterplot and starfield type displays demonstrate how well even a simple visualization can aid human problem solving, particularly when it is accompanied by an effective user interface for selective interactions. The enthusiastic reception given by the intelligence-community users to even the first generation Galaxies' visualization tool (Hendrickson, l995) demonstrates both the power of visualization and the value of an aesthetically rendered "ecological" visual metaphor that is intuitively apprehended by the analyst.

The particular value of Galaxies as a first software product for the PNNL research was that it demonstrated the usefulness of document visualization for analysts' tasks, and strengthened the resolve to seek further visualization metaphors derived from visual and cognitive processes that enable spatial interactions with the natural world. Within four months of the delivery of Galaxies, this effort resulted in a spatialization of unstructured document information derived from GIS techniques that formed a landscape representation we called a ThemeScape™ .

Construction of ThemeScape™ Type Text Visualization

A ThemeScape™ is a surface plot similar to Chalmers (1993) built up by successively layering computed contributions of recovered theme terms over underlying document positions (see Pazner, 1994). It is constructed directly from the distribution of documents in the Galaxies' two-dimensional plane.

First, the characteristic thematic terms that describe each cluster of documents (or the visualized corpus as a whole) are identified on the basis of their discriminability across regions of the high-dimensional document space. The metric is the common term frequency, inverse document frequency weighting scheme first proposed by Salton (1991).

$$\text{Term } N \text{ Value} = f_{\text{term } n/\text{cluster } i} * \sum_{j \text{ not } i} 1/f_{\text{term } n/\text{cluster } j}$$

Where

$f_{\text{term } n/\text{cluster } i}$ = frequency of term n in cluster i and

$\sum_{j \text{ not } i} 1/f_{\text{term } n/\text{cluster } j}$

= frequency of term n in all other clusters.

This yields terms which best discriminate the clusters from each other. In repeated term extractions of this kind,

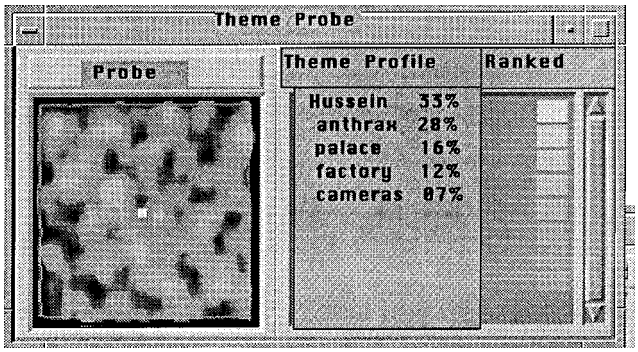

FIG. 5. A hypothetical view of a core probe from a point on a thematic landscape.

over 90% of recovered terms are usually nouns, which meets the goal of building the landscape from topical identities. The number of terms selected in this process can vary, and ThemeScape™s from as few as 50 terms were constructed in first efforts, but low numbers of terms degrade landscape detail differences considerably, and so from 150–300 terms are recommended.

A list of these thematic terms with their corresponding document coordinate pairs provides the basis for depositing each term's contribution to the height of the landscape, as shown in Figure 4. The terms are used like layers of sedimentary strata, wherein each term's layer will vary in thickness as the real probability of finding that term within a document at each point in the 2-D plane. The term layers are then summed and normalized to produce the composite thematic landscape visualization.

The summary vector at any point in the thematic landscape is equal to the sum of unitary document vectors within a selected analysis region. If there are, for example, 12 documents that contain a thematic term within the region, then the summary vector is 12 units high and placed at the center of the region. This placement assumes that the P of finding a thematic term, t_i in that region is $12/\Sigma t_i$ at the center and zero elsewhere in the region. As a final step, a smoothing filter is passed over the summary vectors of the different heights of the thematic terms to produce a more natural-appearing landscape form. This has the effect of decreasing the height of the central peak vector in each analysis region and distributing the probability according to the smoothing function employed. Different smoothing schemes can emphasize or decrease differentiation of the landscape, and can be adjusted to facilitate an analyst's tasks. A useful starting scheme is a variation of the standard Gaussian function that places it at the center of a document analysis region. Where the standard Gaussian function is

$$1/\sigma\sqrt{2}^{e-1/2(x-}\mu/\sigma^2$$

the adjusted one would be

$$N/\sigma\sqrt{2}^{e-1/2(x/}\mu\sigma^2.$$

This removes the -μ term, which centers the smoothing, while scaling the height to the number of documents found within the analysis region through the N term in the numerator. Placing the mean, μ, in the exponent denominator has the effect of leveling out the landscape in the region where the mean values are large, if many documents occur around the edges of the analysis region. Overall, distributed documents around the edges of a region will flatten it out (while raising the overall landscape height), and documents located near the center will tend to produce a peak or pinnacle in the landscape.

In a ThemeScape™, a term layer is thickest at the highest density of documents that carry that term because the probability of finding that term there is correspondingly greater. If the clustering and projections of documents onto the 2-D plane are accurate, documents containing same thematic terms should be in roughly the same place. As term layers accumulate, the highest elevations occur where the thickest layers overlay each other. Lower regions reflect places where there are fewer documents or where the documents are less thematically focused. When there is a sharp distinction from strong thematic term content to low content in the distributed documents, there will be a correspondingly sharp cliff in the ThemeScape™, while a ridgeline connecting two peaks indicates strong themes that are held in common by two different thematic concentrations.

The smoothed Y coordinate height for any x axis point is given by

$$y_x = \sum_{n-m}^{n+m} d_x + n * f(x + n)$$

where $dx + n = 1$ for a document at coordinate $x + n$, otherwise 0, $f(x + n)$ is the value of the smoothing function at x_n, and

2m = width of the smoothing function when centered on any x.

The final height, $z_{x,y}$, of any point on a thematic landscape is given by the sum of the heights of all of the term layers that correspond to their own "miniThemeScape™"s at that point. A final normalization is then usually added.

$$z_{x,y} = \sum_{j=1}^{\#\text{ of cluster terms}} \text{term layer } j_{x,y}$$

The result is a thematic landscape that has literally embodied the content information of a document corpus, and may be treated in most all respects like a sedimentary form, including taking a probe or "core samples" at any point in the landscape that reveal the corresponding terms and their % contribution to the ThemeScape™ at that location (Fig. 5).

However, a thematic landscape constructed in this fashion is different from a sedimentary landscape in one crucial way: The layering of terms has nothing to do with the age of the documents wherein the terms appear. Thus, a "core sample" of terms can be arranged to read from maximum to minimum contribution (the usual display method) or in any other sequence. Users can sometimes take the "sedimentary" deposition analog a bit too literally, and infer that the arrangement of terms in a probe corresponds to dates in documents. This is one instance where a naturalistic landscape metaphor can potentially mislead a viewer, but it demonstrates again the intuitive power of such a visualization.

A ThemeScape™ thus creates a full spatial representation of terms from documents, as opposed to the documents themselves as in Bead (Chalmers & Chitson, l992). Such a thematic terrain synthesizes a mimic of the natural physical

form-giving process of sedimentation. Other physical processes like accretion, condensation, and growth would seem to offer other useful bases on which to build spatializations of textual information.

Text Analysis From a Visualization Basis

The experience and insights gained from building landscape representations of text revealed fundamental shortcomings in the way that statistically based text analysis was undertaken by current software systems. All of these had their origins in the information retrieval (IR) research tradition, and none of them had been designed explicitly for purposes of visualizing their outputs. For example, the IR tradition uses Precision & Recall measures from a document corpus in response to a query as indicants of a text engine's effectiveness. Ideally, a query should acquire all relevant documents and no spurious ones, implying that a visualization should be equally selective. This is expensive in terms of time and computational resources. Why not simply visualize the entire corpus of meanings, and then select as needed?

Through ecological vision, every creature extracts from its physical setting exactly what it needs for its immediate environmentally directed behaviors. Vision is highly selective (even in humans) from the retina of the eye onwards to more central processing centers. At the earliest stages of perception, assemblages of neurons in the retina are prewired to extract such things as corners, crossings, and edges from the visual field. These first order extractions are called "textons" (Treisman, 1986). They then become the primitives upon which visual perception is further constructed, without the need for a total internal computational duplication of the external environment.

The process of ecological perception provides clues as to how text analysis should proceed if its express purpose is to support text visualization. Specifically, text analysis should:

- mimic the sequential extraction of information that occurs in ecological vision rather than require holistic, upfront processing
- create a vector representation of the document that is "interpretable" by being constructed on the contributions of topical or thematic content in the document
- map the topical content of documents directly into a geographical landscape representation
- be scalable as larger document collections are processed
- provide a basis for altered visualizations of the information for different users and purposes (environments appear different to us under different behavioral intentions or "perceptual sets"; similarly, why should we preconceive that there is only one "correct" visualization of text information in a document corpus?)
- incrementally update the analysis with addition of new documents

This implies that the vector representation should be built directly on extracted topical terms in the documents, which become the "textons" of the system. It contrasts strongly with uses of neural net algorithms or dictionaries in other text visualization systems. Neural nets require extensive (and expensive) up-front training on the entire document corpus, do not give interpretable vectors from their output nodes, and tend to equally weight the output nodes in their contributions to the vectorization. They also require retraining as the document corpus is updated. Dictionary based vectorizations can be extremely limited in application, and require constant updating to remain current in technical fields.

Construction of a Text Engine Based on Ecological Vision

This effort was led by Kelly Pennock of the SPIRE team, from October of 1995 to October of 1996, when SPIRE was transferred to a privately funded company outside the laboratory. The successful research completed to that point seemed to fully justify the guiding principles, and had resulted in an intriguingly novel approach that we felt represented a paradigm shift in text analysis. This visually based analysis system centered on finding a limited set of topical terms in the document set for a vector representation that were frequent enough to span the documents, and yet discriminating enough to capture documents' distinctive content. It proceeded as follows:

Step 1 compressed the vocabulary for documents in the corpus by performing stop word removal and stemming. This step is similar to that taken in other text analysis approaches, and the algorithms for it are widely available.

Step 2 performed a band pass frequency filtering on the document corpus, eliminating high and low frequency terms. High frequency terms occur too often to discriminate among document content. Low frequency terms are also non-discriminating, and in addition produce unreliable statistics. Term frequencies depend upon the size of the corpus, but we found it useful with general news stories to filter so as to retain terms that occur $>$3–5 times in a document and that result in a reduction of the terms (after stemming and stop word removal) to a vocabulary 10–15% of its original size. This is based on our findings that in the $\sim$15% or so of middle frequency terms, there are $\sim$15% of these that show significant topical value.

Step 3 was the important filtering step that extracted significant topical terms (after Bookstein, Klein, & Raita, 1995) based on the condensation clustering value (CCV) of the terms surviving band pass filtering.

The CCV measures the degree of randomness in the appearance of a term in documents. Terms that are highly topical in their content (like nouns) tend to occur in bursts or serial clusters in language use, and are not randomly distributed throughout a document or documents. Terms that are less topical (like modifiers) tend to be more randomly distributed. The calculated CCV of any term (word) occurring in a document thus quantifies its potential topical

value. The CCV of a term in a document (or subdocument unit) is given by:

$$\text{CCV} = O_{ti}/E_{ti}$$

where the CCV is the ratio between the obtained number of occurrences of a term within a document or subdocument unit to the expected number of such occurrences. The expected number of occurrences is given by

$$E_{ti} = U[1 - (1 - 1/U)^T]$$

In the former, O_{ti} is the obtained frequency of term occurrence in a unit; E_{ti} is the expected frequency of term occurrence in a unit; U is the number of document or subdocument units in the corpus; T is the number of occurrences of term t_i in the corpus.

To avoid the considerable overhead that would be required in computing term occurrences in a large document corpus, the size and number of documents (or subdocument units) can be estimated from information in the inverted file index.

If the estimated size of the document in bytes is:

$$M_{\text{byte}} = \text{Mean doc size in bytes}$$

Then U (above) is approximated by:

$$U = \text{size of the corpus in bytes}/\text{Mean}_{\text{byte}}$$

And the number of documents with a certain term along with the estimated term frequency is approximated (assuming equal spread of terms in a document) by:

$$U_{\text{term } i} = \sum_{\text{term } i,\, doc\ j} \text{size doc } j/M_{\text{byte}}, \text{ and}$$

$$\text{frequency of } U_{\text{term } i,\text{doc } j} = \text{frequency}_{\text{doc } j}/U_{\text{term } i,\text{doc } j}$$

Under Condensation Clustering, highly topical terms are distinguished by CCVs < 1.00, with CCV becoming lower as topicality increases. Such terms are inevitably a list of nouns which effectively serve as a sort of surrogate list of contents for the document corpus. Note that this list has been established completely through first and second order statistical measures, without use of subject limiting dictionaries, preestablished training runs, or time consuming higher order processing algorithms.

The topical list now needs to be cut at a point so as to produce a short list of topical terms that can be used to construct a vectorization of the document corpus. In experiments performed on *Time* news story sets, it was determined that much less than 500 topical terms (from an initial vocabulary of over 20K terms) was needed to characterize the stories in a basis vectorization. After condensation clustering, eliminating topical words that tended to occur together (e.g., "Ford" and "Bronco"), stemming topics, or performing overlap filtering on topical terms that have the same discrimination ability, the list of topical terms comes down to 300 or less. This term list is short enough to permit rapid calculations in a vectorization, and yet powerful enough to discriminate among documents and thereby place them accurately in a visualization. It also provides a vectorization that is interpretable because documents can now be characterized on the basis of these canonical topical terms that actually carry meaning to any user.

Step 4 now disambiguates these remaining N topical term meanings by creating an M length association list of terms with both moderate CCVs and moderate frequencies in the document corpus. These are M words that tend to occur differentially with the N topical terms, thereby revealing the context in which the topical terms are being used. For example, the word "bank" in a topical term list would communicate something very different if it were associated either with terms like "bond," "trades," and "market" or if it occurred with "river," "steep," and "sandy."

So an N X $(N + M)$ term (rectangular) association matrix is then formed with the N topical terms and the M association terms. The cell entries are the conditional probabilities of the term occurrences in documents (or subdocument units) of the corpus, minus the base term's independent probability.

$$X_{ij} = P(t_i/t_j) - P(t_j)$$

where X_{ij} = the ith row and jth column of the conditional probability matrix $P(t_i/t_j)$ = the conditional probability of t_i given t_j, and $P(t_j)$ = the probability of t_j.

The matrix expresses the probabilistic associations among discriminating topical terms, and among topical terms and the disambiguating association terms.

topic 1, topic 2, . . . topic n, cross term 1,

cross term 2, . . . cross term m

topic 1I
topic 2I
. I
. I
. I
. I
topic nI

$$X_{ij} = P(\text{term } i/\text{term } j) - \beta \times P(\text{term } j)$$

The "Beta" parameter can be inserted and adjusted to weight the influence of the link to the base term, based on it's CCV or other (co)occurrence value.

The topical terms and cross terms comprise the "essential vocabulary of words of interest" in the document corpus, with the topical terms serving as basis dimensions for the vector space. Each word's vector is the column of matrix entries headed by that word in the matrix, where each row

of the vector is the (modified) conditional probability of that word's association with a topical basis term. The vectors for each word of interest are summed and these summations are then normalized so that the sum of all component magnitudes = 1.00. This effects a vectorization of documents along interpretable dimensions of topical basis terms that characterize meaning.

Deriving a Flexible View of Document Information

Now each document is represented by a high dimensional vector whose components indicate that document's discriminating words and how those words are connected to all other topics of interest that span and describe the document collection. Unlike, e.g., neural net vector models, the document dimensions are interpretable and meaningful. Unlike dictionary based text analysis, the vectorization is based on the document corpus currently retrieved.

The approach is scalable because the computations are based on no more than second-order statistics of the terms, and the vocabulary continues to be compressed through the filtering extraction of non-discriminating, nontopical terms.

The construction of basis vectors on topical terms themselves also avoids the bottleneck and inaccurate placement that occurs when a landscape visualization like a ThemeScape™ is based on an intervening document projection to the two-dimensional plane as occurred with Galaxies. Under this new text analysis procedure, *topics* themselves are projected onto the two-dimensional plane, where their placement is used to construct the landscape as before. Since there are far fewer discriminating topical terms than there are documents, their placement and the subsequent appearance of the landscape is more veridical with respect to actual thematic relationships than if these had depended on documents' placement.

Through bypassing the Galaxies view and going directly to the ThemeScape™ view, the severe limitation on screen space for displaying large numbers of documents as a starfield is also avoided. As documents grow much beyond a few thousand, the user's monitor view simply fills up with stars, and distribution patterns become indistinct. However, by reversing the viewing sequence and using the landscape view as the overall introduction to the document corpus, it now becomes possible to allow the user to select merely a portion of the landscape, and view a starfield document projection based on the documents containing those topics in that part of the landscape.

A topical term based vectorization also permits an easy way to alter the view of the visualization for users with different interests. Since topics are dimensions of the space, and any generalized distance measure can be used to represent similarities among topics (or documents), a linear transformation on any topical basis dimension will rescale the view and result in a landscape or starfield that is altered to (de)emphasize the contributions of selected topics. Through use of a slider control on each topical term, the user can select their relative weightings according to their interests and acquire correspondingly transformed visualizations of the text and documents. This brings both the analysis and visualization under user control in a means analogous to how, in ecological vision, directed attention increases environmental salience dependent upon the viewer's intent.

Conclusion

What the "ecological approach" to text visualization represents, of course, is a beginning, not an ending. It argues simply that information visualizations be developed from a perspective informed by the ways in which we perceptualize the real world. These are the neural computations set into our experiences and into our genes, and their extension into information space can be a natural one if we observe and adhere to the perceptual strategies that have taken us this far from our origins.

Seeing text visualization as part of an "ecological" whole itself, bound into the very ways that we analyze and characterize the text document, also resolves the perennial call for "new visualization metaphors." Without an analytical context within which to embed them, such ideas are little more than new fashion statements, mostly incomparable, and hardly building towards the "science" of information visualization that is our common goal. But within the "ecological approach" new metaphors are as numerous as the form-giving processes that surround us, and most of these have their own well-developed formulations that give them reason and meaning. These are the processes of accretion, condensation, erosion, branching, crystallization, and more that give rise to the shapes we are meant to see because they were valuable to our ancestors' lives. In the ways that bones and trees grow, that landscapes lay and rivers flow are the embodied means by which information is manifest in the natural world. These are recipes for how embodied information can be reasserted in digital forms, while still retaining the meaning that makes that form informative.

The late SPIRE investigations demonstrated that it is possible to visualize text as natural forms built on well-established techniques, and that text analysis and text visualization gain substantial power and effectiveness when they are conceived together, and document analysis takes on a task analogous to retinal and lateral geniculate processing in our visual pathway. The use of condensation clustering to extract topical terms from text is itself an analog of edge detection in biological vision systems.

These were small accomplishments with large implications for the field of information science if they are seen and understood in the context of their conception: as demonstration examples of how to think about visualizing the nonvisual through means analogous to the ways visual structures are naturally formed. As Leonardo da Vinci repeatedly exhorted, we must but "open our eyes" to the possibilities for invention that the world around us persistently reveals.

Acknowledgment

The author sincerely thanks the SPIRE research team for the discussions and briefings that made this article possible. In particular, David Lantrip, Kelly Pennock, and Jeremy York provided key contributions, as did Quintin Congdon and Timothy Hendrickson of the Pathfinder Program. And deep appreciation to Dr. Russell Rose for his support, trust, and leadership that initiated and sustained the quest for text visualization.

References

Ahlberg, C. & Schneiderman, B. (1994). Visual information seeking: Tight coupling of dynamic query filters with starfield displays. Proceedings ACM CHI'94: Human factors in computing systems (pp. 313–317), New York: ACM Press.

Ahlberg, C. & Wistrand, E. (1995). IVEE: An information visualization and exploration environment. In N. Gershon & S. Eick (Eds.), Proceedings IEEE Visualization 95 (pp. 66–73), Los Alamitos, CA: IEEE Computer Society Press.

Bartell, T.T., Cottrell, G.W., & Belew, R.K. (1992). Latent semantic indexing is an optimal special case of multidimensional scaling. Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 161–167), New York: ACM Press.

Bookstein, A., Klein, S.T., & Raita, T. (1995). Detecting content-bearing words by serial clustering. Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 319–327), New York: ACM Press.

Chalmers, M. & Chitson, P. (1992). Bead: Explorations in information visualization. In Proceedings of SIGIR '92 (pp. 330–337), New York: ACM Press.

Chalmers, M. (1993). Using a landscape metaphor to represent a corpus of documents. In A. Frank & I. Campari (Eds.), Spatial information theory (pp. 377–390), London: Springer-Verlag LNCS 716.

Crow, V., Wise J.A., Thomas, J.J., Lantrip, D., Pottier, M., Schur, A. (1994). Multidimensional visualization and browsing for intelligence analysis. Proceedings of GVIZ '94.

Eick, S. (l997). Engineering effective visualizations. Keynote Address, CODATA Euro-American Workshop on Data and Information Visualization: Where are we and where do we go from here?

Frakes, W., & Baeza-Yates, R. (1992). Information retrieval: Data structures and algorithms. Englewood Cliffs, NJ: Prentice Hall.

Gibson, J.J. (1979). The ecological approach to visual perception. NY: Houghton Mifflin Co.

Hendrickson, T. (1995). Exploring the galaxies. Proceedings of the Symposium on Advanced Intelligence Processing and Analysis (p. 74), Washington DC: Office of Research and Development.

Lopresti, E. & Harris, M. (1996). LoudSPIRE, an auditory display for the SPIRE system. Proceedings of the International Conference on Auditory Display (ICAD 96).

Pazner, M. (1994). GIS analysis and modeling of nonspatial data. In Proceedings of the 6th Canadian Conference on Geographic Information Systems (pp. 1056–1069), June 1994, Ottawa, Canada.

Pennock, K. & Lantrip, D. (1995). A landscape representation of themes in text. In AIPA Steering Group (Eds.), Proceedings of the Symposium on Advanced Intelligence Processing and Analysis, (p. 47), Washington DC: Office of Research and Development.

Rorvig, M. (1988). Psychometric measurement and information retrieval. In M.E. Williams (Ed.), Annual review of information science and technology (Vol. 23, pp. 157–189).

Salton, G. (1991). Developments in automatic text retrieval. Science, 253, 974–980.

Salton, G. & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Shepherd, R.N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function I. Psychometrika, 27, 3, 125–140.

Shepherd, R.N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function II. Psychometrika, 27, 2, 219–246.

Smithson, R. (1996). In W.R. Morrish (Au.), Civilizing terrains (p. i), San Francisco, CA: William Stout Publishers.

Stoner, T. (1990). Information and the internal structure of the universe. (p. 74), London, UK: Springer-Verlag.

Thompson, E., Palacios, A., & Varela, F. (l992). Ways of coloring: Comparative color vision as a case study for cognitive science. Behavioral and Brain Sciences, 15, 1–74.

Treisman, A. (1986). Features and objects in visual processing. Scientific American, 255, 114–125.

Wise, J.A. & Thomas, J.J. (1995). The spatial representation of textual information: Lessons learned on the MVAB project. In AIPA Steering Group (Eds.), Proceedings of the Symposium on Advanced Intelligence Processing and Analysis, (p. 75), Washington DC: Office of Research and Development.

Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In N. Gershon & S. Eick (eds.), Proceedings IEEE Visualization 95 (pp. 51–58), Los Alamitos, CA: IEEE Computer Society Press.

Wise, J.A. (1996) Doing what comes naturally: The new look in decision support systems. Global financial review (No.1, pp. 24–25), Blue Bell, PA: UNISYS.

York, J., Bohn, S., Pennock, K., & Lantrip, D. (1995). Clustering and dimensionality reduction in SPIRE. In AIPA Steering Group (Eds.), Proceedings of the Symposium on Advanced Intelligence Processing and Analysis (p. 73), Washington DC: Office of Research and Development.