Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents

James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, Vern Crow
Pacific Northwest Laboratory
Richland, Washington

Abstract

This paper describes an approach to IV that involves spatializing text content for enhanced visual browsing and analysis. The application arena is large text document corpora such as digital libraries, regulations and procedures, archived reports, etc. The basic idea is that text content from these sources may be transformed to a spatial representation that preserves informational characteristics from the documents. The spatial representation may then be visually browsed and analyzed in ways that avoid language processing and that reduce the analysts' mental workload. The result is an interaction with text that more nearly resembles perception and action with the natural world than with the abstractions of written language.

1: Introduction

Information Visualization (IV), extends traditional scientific visualization of physical phenomena to diverse types of information (e.g.text, video, sound, or photos) from large heterogenous data sources. It offers significant capability to different kinds of analysts who must identify, explore, discover, and develop understandings of complex situations.

IV has been studied for many centuries, integrating techniques from art and science in its approach [7, 9]. The information analyst's perspective illustrates that their process involves more than envisioning information. [4] It is both the visual representations and the resultant interactions with it that entail the analyst's work.

Current visualization approaches demonstrate effective methods for visualizing mostly structured and/or hierarchical information such as organization charts, directories, entity-attribute relationships, etc. [3,9]. Free text visualizations have remained relatively unexamined.

The idea that open text fields themselves or raw prose might be candidates for information visualization is not obvious. Some research in information retrieval utilized graph theory or figural displays as 'visual query' tools on document bases [5,8], but the information returned is documents in their text form--which the user still must read to cognitively process. The need to read and assess large amounts of text that is retrieved through even the most efficient means puts a severe upper limit on the amount of text information that can be processed by any analyst for any purpose.

At the same time, "Open Source" digital information--the kind available freely or through subscription over the Internet--is increasing exponentially. Whether the purpose be market analysis, environmental assessment, law enforcement or intelligence for national security, the task is to peruse large amounts of text to detect and recognize informational 'patterns' and pattern irregularities across the various sources. But modern information technologies have made so much text available that it overwhelms the traditional reading methods of inspection, sift and synthesis.

2: Visualizing text

True text visualizations that would overcome these time and attentional constraints must represent textual content and meaning to the analyst without them having to read it in the manner that text normally requires. These visualizations would instead result from a content abstraction andspatialization of the original text document that transforms it into a new visual representation that communicates by image instead of prose. Then the image could be understood in much the way that we explore our worldly visual constructions.

It is thus reasonable to hypothesize that across the purposes for perusing text, some might be better satisfied by transforming the text information to a spatial representation which may then be accessed and explored by visual processes alone. For any reader, the rather slow serial process of mentally encoding a text document is the motivation for providing a way for them to instead use their primarily preattentive, parallel processing powers of visual perception.

The goal of text visualization, then, is to spatially transform text information into a new visual

0-8186-7201-3/95 $04.00 © 1995 IEEE

See Color Plates, page 140.

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE

representation that reveals thematic patterns and relationships between documents in a manner similar if not identical to the way the natural world is perceived. This is because the perceptual processes involved are the results of millions of years of selective mammalian and primate evolution, and have become biologically tuned to seeing in the natural world. The human eye has it's own contrast and wavelength sensitivity functions. It has prewired retinal "textons", or primitive form elements used to quickly build up components of complex visual images. Much of this processing takes place in parallel on the retinal level, and so is relatively effortless, exceptionally fast, and not additive to cognitive workload. Even at the visual cortex, perception appears to rely on spatially distributed parallel construction processes in a topography that corresponds to the real physical world. The central conjecture behind the approach to text visualization described here is that the same spatial perceptual mechanisms that operate on the real world will respond to a synthetic one, if analogous cues are present and suitably integrated. The bottleneck in the human processing and understanding of information in large amounts of text can be overcome if the text is spatialized in a manner that takes advantage of common powers of perception.

3. Visualization transformations: from text to pictures

Four important technical considerations need to be addressed in the creation of useful visualizations from raw text. First, there must be a clear definition of what comprises text and how it can be distinguished from other symbolic representations of information. Second, there must be a way to transform raw text into a different visual form that retains much of the high dimensional invariants of natural language, yet better enables visual exploration and analysis by the individual. Third, suitable mathematical procedures and analytical measures must be defined as the foundation for meaningful visualizations. Finally, a database management system must be designed to store and manage text and all of its derivative forms of information.

For the purpose of this paper, text is a written alphabetical form of natural language. Diagrams, tables, and other symbolic representations of language are not considered text. Text has statistical and semantic attributes such as the frequency and context of individual words, and the combinations of words into topics or themes. The differences between text's statistical and semantic compositions provide much of the opportunity for the text visualizations described in this paper. For example, reading a text document to extract it's semantic meaning is different from learning that a document is of a certain relative size, type, or authorship, with particular content themes. But both semantic and content knowledge can be valuable to an analyst. Identifying publishing activity on particular subjects from particular authors at certain places and times is useful, especially if one does not have to read all of the documents to determine that pattern.

In digital form,written text can be treated statistically to extract information about it's content and context, if not semantic meaning. While this does not necessarily entail natural language processing algorithms, it does require a set of special purpose processes to convert text to an alternative spatial form that can be displayed and utilized without needing to read it.

The first component of a software architecture to visualize text is the document database or corpora. Documents contained within such databases are derived from messages, news articles, regulations, etc., but contain primarily textual material. The next component is the text processing engine, which transforms natural language from the document database to spatial data. The output from the text engine is either stored directly in a visualization database, or projected onto a low dimensional, visual representation. Other components of the architecture are the Graphical User Interface (GUI), the display software (such as visualization packages), the Applications Interface (API), and auxiliary tools.

3.1: Processing text

The primary requirements of a text processing engine for information visualization are: 1) the identification and extraction of essential descriptors or text features, 2) the efficient and flexible representation of documents in terms of these text features, and 3) subsequent support for information retrieval and visualization.

Text features are typically one of three general types, though any number of variations and hybrids are possible. The first type is frequency-based measures on words, utilizing only first order statistics. The presence and count of unique words in a document identifies those words as a feature set. The second type of feature is based on higher order statistics taken on the words or letter strings. Here, the occurrence, frequency, and context of individual words are used to characterize a set of explicit or implicitly (e.g. associations defined by a neural

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE

network) defined word classes. The third type of text feature is semantic in nature. The association between words is not defined through analysis of the word corpus, as with statistical features, but is defined a priori using knowledge of the language. Semantic approaches may utilize natural or quasi-natural language understanding algorithms, so that the semantic relationships (i.e., higher-order information) are obtained.

Text features are a "shorthand" representation of the original document, satisfying the need of a text engine to be an efficient and flexible representation of textual information. Instead of a complex and unwieldy string of words, feature sets become the efficient basis of document representations and manipulations. The feature set information must be complete enough to permit flexible use of these alternatives. Text engines support both efficiency and flexibility, though these criteria are often in opposition.

The third requirement of the text engine is to support information retrieval and visualization. The text processing engine must provide easy, intuitive access to the information contained within the corpus of documents. Information retrieval implies a query mechanism to support it. This may include a basic Boolean search, a high level query language, or the visual manipulation of spatialized text objects in a display. To provide efficient retrieval, the text processing engine must pre-process documents and efficiently implement an indexing scheme for individual words or letter strings.

The more visual aspect of information retrieval is known as information browsing. The specificity of querying has a counterpoint in the generality of browsing. The text processing engine or subsidiary algorithms can support browsing by providing composite or global measures which produce an intuitive index into topics or themes contained within the text corpus. A set of measures which characterize the text in meaningful ways provide for multiple perspectives of documents and their relationships to one another. One example of such a measure is "similarity". Based on the occurrence and the context of key words or other extracted features, measures of similarity can be computed which reflect the relatedness between documents. When similarity is represented as spatial proximity or congruity of form, it is easily visualized. A diversity of measures is essential, given that documents can be extraordinarily complex entities containing a large number of imprecise topics and subtopics. Clearly, no single visualizable snapshot of a document base can provide the whole picture.

3.2: Visualizing output from text processing

Composing a spatial representation from the output of the text analysis engine is the next step to visualizing textual information. Spatialization itself is composed of several stages. The first involves representing the document, typically as a vector in a high dimensional feature space. The vector representation is the initial spatial expression of the document, and a variety of comparisons, filters, and transformations can be made from it directly.

To represent each of these documents, an initial visualization may consist of a scatter plot of points (one for each document), collocated according to a measure of similarity based on vector representations. Since visualization of the textual information requires a low-dimensional representation of documents that inhabit a high-dimensional space, projection is necessary. Typically, linear or non-linear Principal Components Analysis or metric Multi-Dimensional Scaling (MDS) can be used to reduce dimensionality to a visualizable subspace. One serious concern with these techniques, is their exponential order of complexity, requiring that dimensionality reduction and scaling be considered simultaneously since a large corpus may contain 20,000 or more documents. For large document corpora, alternatives to the projection of each document point are necessary. In these cases, clustering can be performed in the high-dimensional feature space and the cluster centroids become the objects to be visualized. For a review of clustering and metric issues, see [11].

3.3: Managing the representation

There are two basic classes of data that must be managed. The first is the raw text files for each document. The text itself as well as a variety of header information fall into this category. This first class of data is static in nature, simple in structure, and therefore easy to manage. The second broad class of data is the visual forms of the text. This class of data is derived from the numerous algorithms designed to cluster, structure, and visually present information, and is both extensive and dynamic.

For the current text visualization endeavor, an object-oriented database was selected for managing text and its various visual forms. This paradigm was chosen for its flexibility of data representation, the power of inheritance, and the ease of data access where complexity of the data structures to be managed is great. The structures contain both high and low dimensional spaces, substructures such as clusters and

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE

super clusters, entities such as documents and cluster centroids, and a variety of other components. The class structure of the database also permits the common elements to be shared (inherited) through the hierarchy of data classes, while the differences between the structures can be specified at lower levels. The selected object-oriented database also implements database entities as persistent objects, where the access and manipulation of the data are one and the same, eliminating the need for a query mechanism as such.

3.4: Interface design for text visualization

To achieve direct engagement for text visualization [6], the interface must provide 1) a preconscious visual form for information 2) interactions which sustain and enrich the process of knowledge building, 3) a fluid environment for reflective cognition and higher-order thought, and 4) a framework for temporal knowledge building.

Three primary display types are made available to the analyst. Tools are arranged along the perimeter of the display monitor and can be used as operators on the representations. Conversely, the representations or selected areas in the representations can be dragged and dropped onto the tools to spawn the appropriate action. The analyst can work on the primary information views [1] in an area known as the backdrop, which serves as a central display resource for visual information; alternatively, she can move the views to the workshop or the chronicle. The workshop is a grid where selected views or parts of views can be placed for work and/or visual review. The grid has a number of resizable windows to hold multiple views. The chronicle is a space where representations of more enduring interest can be placed. Views placed in this area can be linked to form a sequenced visual story where decision points are highlighted. The workshop and the chronicle take advantage of the phenomena of visual momentum; the ability to extract information across a set of successively viewed displays [10] that can be a series of static or dynamic images. The characteristics of the backdrop, workshop, and chronicle, known collectively as storylining, provide the ability to capture and visually organize situations across the time-past, present, and future. This endows the analyst with the ability to summarize their experience of knowledge building [2].

4: Examples from the MVAB Project

The Multidimensional Visualization and Advanced Browsing project is currently exploring a number of representations for the visualization and analysis of textual information. These approaches have been showcased in SPIRE™, the Spatial Paradigm for Information Retrieval and Exploration, which was developed to facilitate the browsing and selection of documents from large corpora (20,000+ documents). Described below are the two major visualization approaches or views which were developed in the first year of this project: Galaxies and Themescapes.

Starfields and topographical maps were selected as display metaphors because they offer a rich variety of cognitive spatial affordances that naturally address the problems of text visualization. Starfields create point clusters which suggest patterns of interest. Maps offer topographies of peaks and valley that can be easily detected based on contour patterns. Both these spatial arrangements allow overview and detail without a change of view. Each view, however, offers a different perspective of the same information and serves as the organizing points for knowledge construction.

4.1: Galaxies

The Galaxies visualization displays cluster and document interrelatedness by reducing a high dimensional representation of documents and clusters to a 2D scatterplot of 'docupoints' that appear as do stars in the night sky. Although the resulting visualization is simple, it provides a critical first cut at sifting information and determining how the contents of a document base are related. The key measurement for understanding this visualization is the notion of document "similarity". The more similar that clusters and documents are to one another in terms of their context and content, the closer or more proximate they are located in the 2D space. By exploring and animating this visualization, analysts can quickly gain an understanding of patterns and trends that underlie the documents within a corpus. At the highest level of representation, Galaxies displays corpus clusters and the gisting terms which describe them. **(Figure 1)**

A simple glance at this spatial representation reveals the fundamental topics found within the corpus, and provides an avenue of exploration which can be followed by simply clicking on a cluster of interest to reveal the documents within. These documents can then be grouped, gisted, annotated, or retrieved for more detailed analysis. In addition to simple point and click exploration of the document base, a number of sophisticated tools exist to facilitate more in-depth analysis. An example of such a tool is the temporal slicer. Designed to help tie document

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE

spatial patterns with temporal ones, this tool utilizes document timestamps to partition the document base into temporal units. The granularity of these units can be defined by the user as years, months, days, hours or minutes. Slicing a database entails moving a "temporal window" through the documents, and watching the visualization populate itself with documents. **(Figure 2)**

The resulting emergence of clusters can indicate temporal links that relate topics. When viewed in terms of known historical events and trends, these growing cluster patterns can provide insight into external causal relationships mirrored in the corpus.

4.2: ThemeScapes

ThemeScapes are abstract, three-dimensional landscapes of information that are constructed from document corpora (Figure 3). The complex surfaces are intended to convey relevant information about topics or themes found within the corpus, without the cognitive load encountered in reading such content. A thematic terrain simultaneously communicates both the primary themes of an arbitrarily large collection of documents and a measure of their relative prevalence in the corpus. Spatial relationships exhibited in the landscape reveal the intricate interconnection of themes, the transformation of themes across the whole of the document corpus, and the existence of information gap, or "negative information."

The ThemeScapes' visual representation has several advantages. First, a ThemeScape displays much of the complex content of a document database. Elevation depicts theme strength, while other features of the terrain map such as valleys, peaks, cliffs, and ranges represent detailed interrelationships among documents and their composite themes. At a glance, it provides a visual thematic summary of the whole corpus. The second major advantage of thematic terrain is that it utilizes innate human abilities for

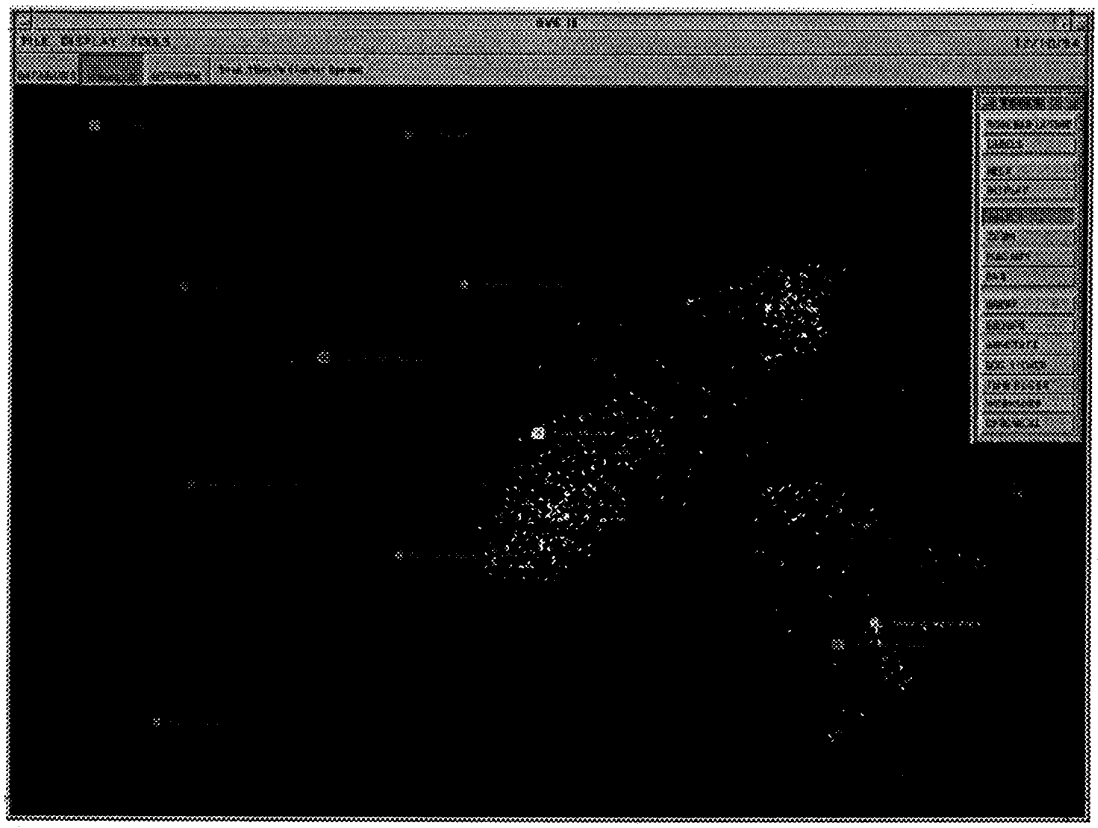

Figure 1: Galaxies visualization of documents and document clusters in a text database.

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE

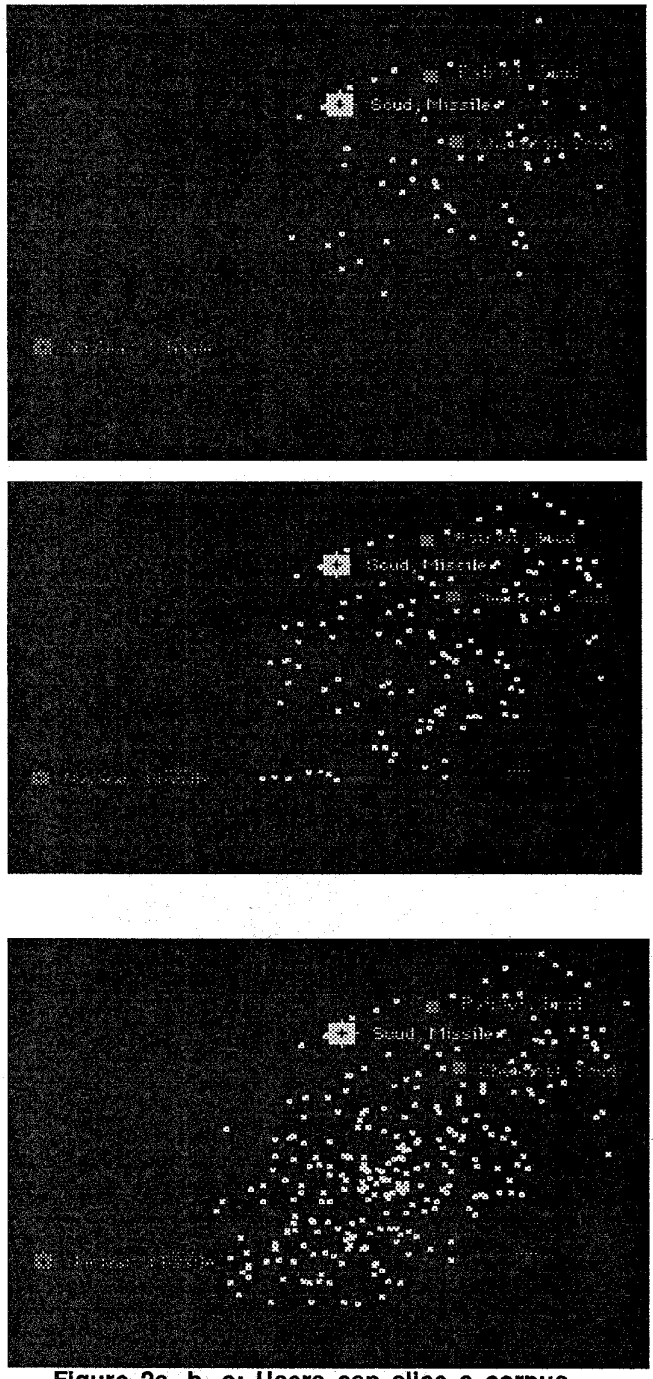

Figure 2a, b, c: Users can slice a corpus to relate document patterns with temporal ones.

pattern recognition and spatial reasoning. The complexity of the terrain is perceived and analyzed with parallel and preattentive processing which do not tax serial, attentional resources. This greatly expands the bandwidth of communication between the tool and the user. A third major advantage of the terrain implementation is its communicative invariance across levels of textual scale. An entire document corpus, a cluster of documents, individual documents, or even document components such as paragraphs or sentences can all be equally well visualized in a ThemeScape. This feature allows the ThemeScape to be used for automated document summarization as well as summarization of the whole document base, explicitly displaying the multitude of topics in a single image. Finally, ThemeScapes promote analysis by promoting exploration of the document space. Utilizing the metaphor of the landscape, associated tools allow the analyst to take 'core samples' and 'slices' through the thematic terrain to see its composition and to understand how thematic topics come to relate to one another in the underlying documents.

5: Conclusions and directions for future research and development

The MVAB project has started with a visualization (Galaxies) that provides a simplified universal view of the relationships among documents in an entire corpus. We have then proceeded to the Themescape visualization, which does the same for the thematic content expressed in those documents. In doing so, we have gone from a metaphor of points in space to one of a landscape. We are now pursuing development of a third visualization that would handle specific entity-attribute relationships found in the documents, such as treated by Link Analysis. This layering of informational detail and abstraction appears necessary for large document bases where investigating global structure is a primary concern yet relationships between individual objects are also important. It gives real meaning and form to the notion of 'data mining'.

So far, the R &D efforts on the MVAB project have shown that there appears to be substantial justification to the idea that text visualizations can overcome much of the user limitations that results from accessing and trying to read from large document bases. Even with the relatively simple first Galaxies visualization of documents as stars in a 2-D space, analysts have returned reports of enhanced insight and time savings such as "discovering in 35 minutes what would have taken two weeks otherwise." Analysts

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE

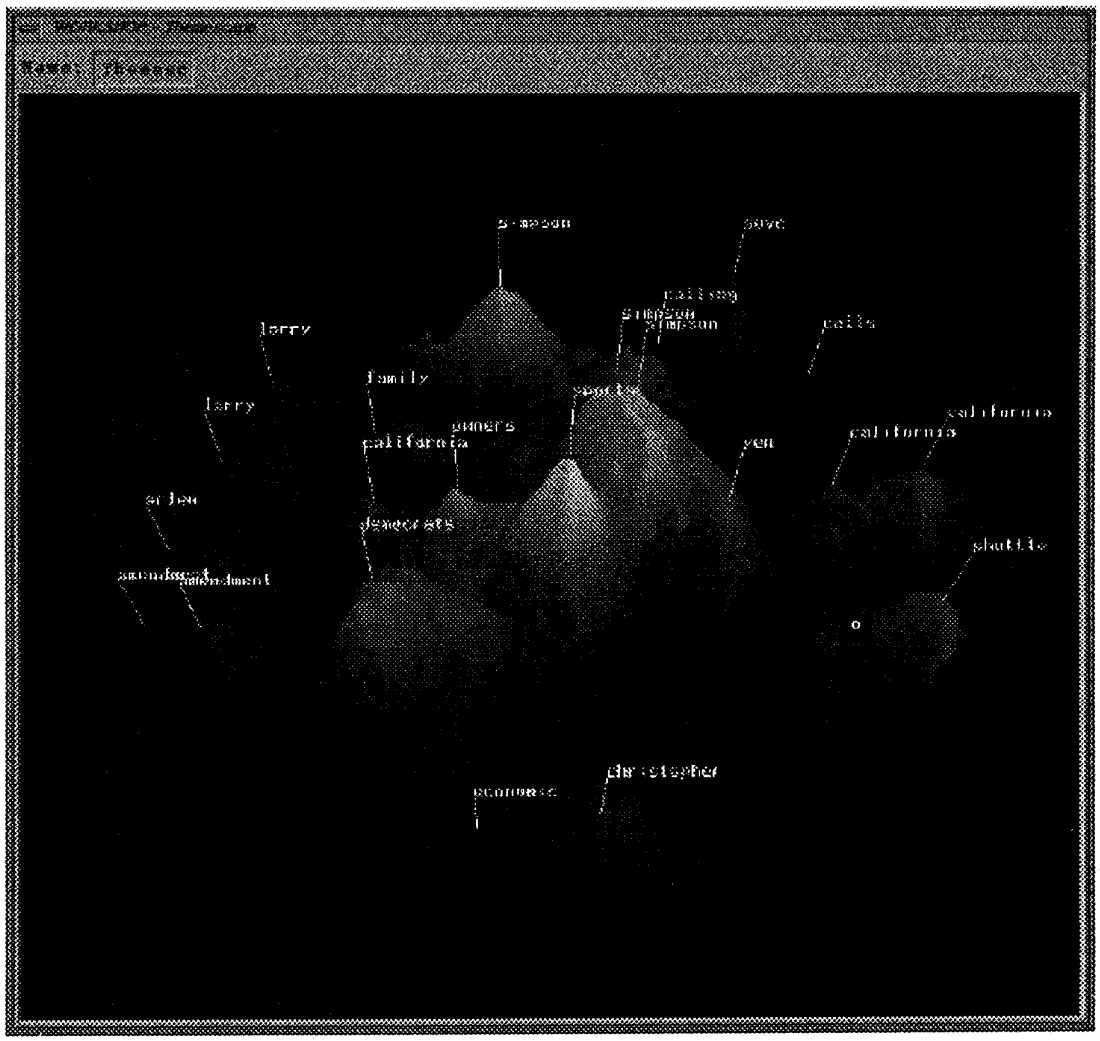

Figure 3: A ThemeScape of an entire week of CNN newstories comprising a document corpus

have also been quite creative with the tool, using the time slicer to do pattern recognition and comparison on evolving and historically documented situations. This kind of adaptation of what is essentially a visual browser has encouraged us to pursue analytical visualizations and to drop some distinctions among analysts' tasks that exist in the purely textual realm. For example, querying a document base by means of a Boolean text string is different from reading the documents returned from the query. However, the Themescape described above, although meant for visual analysis, also permits a different kind of visual querying than was ev er possible under the all-or-none choices of Boolean logic. Touching a Themescape can be a way to initiate a weighted query in terms of the themes that proportionally compose the elevation of the terrain. This allows the analyst to seek documents that talk about combinations of topics in a selected relative abundance based on the analyst's interests. Querying and analytical manipulation come together in a single visualization.

Users' experience also justifies the initial conviction that text visualizations will have to access and utilize the cognitive and visual processes that enable our spatial interactions with the natural world. This suggests visual metaphors that recapitulate

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE

experiences of viewing the night sky and traversing landscapes.

Another observation echoed in the growing popularity of Visual Data Analysis (VDA) programs is that perception and action are provocative complements to one another. An image must be acted on in some way, which in turn suggests new facets of its character that stimulate further visual inspection. Galaxies success with analysts is in no small part due to the abilities to pan, group and timeslice the docupoints in the display. The success of other text visualizations will likely be determined by whether the user can manipulate them along the lines of their analytical intuitions.

Future efforts will elaborate the visual metaphors described above, as well as new ones that effectively capture how concepts and decisions 'come into form'. Much of the analyst's world is a dynamic changing information terrain. Seeking coherance and patterns in this environment carries a high price in time and effort. Capturing the development of a story or the threads of a concept communicated in prose is a high order for text visualizations. But there appears to be no formal reason why at least some of these aspects cannot be captured as well in image as they can in words.

Other extensions of this research are suggested by the addition of sensory modalities like sound to the text visualization. If text content or connections can be captured in three dimensional solid forms, then those forms might also be given other properties, like density, that characterize their appearance and behavior in the real world. Through enhanced means of 'virtual interaction', these properties could reinforce and extend the impressions gained by visual inspection alone, and start to give much more of the affective content and tone that well written prose conveys.

It is evident that the potentials of text visualization are just beginning to be explored and realized. With them, the incredible diversity and volumes of written information available around the world may yet be made more accessible and comprehensible through this perceptual restructuring. And the limitations of an Information Age will not be set by the speed with which a human mind can read.

References

[1] Bannon, L. J., and Bodker, S., *Beyond the Interface: Encountering Artifacts in Use*. In Carroll J. M. (Ed.) Designing Interaction: Psychology at the Human Computer Interface pages 227-253. Cambridge, Cambridge University Press, 1991.

[2] Henniger, S., Belkin, N., *Interfaces Issues and Interation Stratigies for Information Retireval Systems*. ACM Computer Interaction Tutorial Workbook #19, April 1994.

[3] Johnson, J. A., Nardi, B. A., Zarmer, C. L., and Miller, J. R., 1993. Information Visualization Using 3D Interactive Animation. *Communications of the ACM*, 36(4):40-56.

[4] Keller, P. R., and M. M. Keller. *Visual Cues: Practical Data Visualization*. IEEE Computer Society Press, Los Alamitos, California. 1993.

[5] Korfhage, Robert R. To See, or Not to See--Is That the Query? *Communications of the ACM*, 34, pages 134-141, 1991.

[6] Laurel, B. *Computers as Theatre*. Addison-Wesley, Reading, Massachusetts, 1993.

[7] Robertson, G. C., Card, S. K., and Mackinlay, J.D. 1993. Information Visualization Using 3D Interactive Animation. *Communications of the ACM*, 36(4):56-72

[8] Spoerri, Anselm. InfoCrystal: A visual tool for information retrieval. *Proceedings of Visualization '93*, pages 150-157. IEEE Computer Society Press, Los Alamitos, California, 1993.

[9] Tufte, E. R. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, 1990.

[10] Woods, D. D., *Visual Momentum: a Concept to Improve Cognitive Coupling of Person and Computer*. International Journal of Man-Machine Studies 21: 229-244. 1984.

[11] York, J. and Bohn, S. *Clustering and Dimensionality Reduction in SPIRE*. Presented at the Automated Intelligence Processing and Analysis Symposium, Mar 28-30, 1995, Tysons Corner, VA.

Proceedings of the Proceedings on Information Visualization (INFOVIS '95)
0-8186-7201-3/95 $10.00 © 1995 IEEE