Conf-9409278--2

PNL-SA-24870

MULTIDIMENSIONAL VISUALIZATION AND BROWSING FOR INTELLIGENCE ANALYSIS

V. Crow
M. Pottier
J. Thomas
D. Lantrip[a]
C. Struble
K. Pennock
A. Schur
J. Wise
T. Fiegel
J. York

September 1994

Presented at the
GVIS '94 Graphics and Visualization
Conference
September 8, 1994
Richland, Washington

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

(a) University of Michigan, Ann Arbor, Michigan

Multidimensional Visualization and Browsing for Intelligence Analysis

Vern Crow, Kelly Pennock, Marc Pottier, Anne Schur, Jim Thomas,
Jim Wise, David Lantrip, Tarra Fiegel, Craig Struble, Jeremy York

Applied Physics Center
Pacific Northwest Laboratory[1]
Richland, WA 99352

Abstract

Visualization tools have been invaluable in the process of scientific discovery by providing researchers with insights gained through graphical tools and techniques. At PNL, the Multidimensional Visualization and Advanced Browsing (MVAB) project is extending visualization technology to the problems of intelligence analysis of textual documents by creating spatial representations of textual information. By representing an entire corpus of documents as points in a coordinate space of two or more dimensions, the tools developed by the MVAB team give the analyst the ability to quickly browse the entire document base and determine relationships among documents and publication patterns not readily discernible through traditional lexical means.

1. Introduction

Today's intelligence analyst is forced to deal with vast amounts of information available as unstructured text in the form of news briefs, professional articles, records and miscellaneous documents. Up to 30,000 documents a day may electronically cross an analyst's desk. The sheer volume of these documents results in information overload, and causes traditional methods of analysis (scanning, reading, sifting and synthesis) to break down.

Instead of 'drinking from the information firehose', analysts are coming to rely on a variety of text extraction programs that search a corpus, and retrieve documents on the basis of text strings entered through boolean queries. The "hits" from their electronic searches are also presented in text form, and may easily number several hundred documents. But even this assisted retrieval approach requires prodigious effort to scan and locate interesting "nuggets" of information, to cross-correlate them, and to extract meaningful patterns relevant to questions of interest.

Working directly with text demands allocating time and mental effort to the sheer mechanical serial processing of information. This deflects the analyst's mental resources from the useful analysis of the information that is recovered. Information visualization is proposed as the means for creating tools for thought to change the analyst's current workload. If textual information can be transformed into spatial representations, the spatial forms may instead be more easily browsed and interpreted using the faster parallel processes of human vision. This results in reduced workload as well as increased analytical efficiency.

Information visualization under development at PNL is different from scientific visualization in two ways. First, it emphasizes the two-way interactive interface between humans and their information resources. Second, it is squarely aimed at melding the human's capacity for visual thinking with the machine's capability for analytical computing. Information visualization is neither just statistical graphics, data displays, nor physical simulations. It is a visually expressed embodiment of the ongoing interaction of analytical intentions and information resources.

[1]Pacific Northwest Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RLO 1830.

Information visualization does not have clear boundaries in terms of its relationships with other aspects of information handling. On the upstream side, the encoded extraction of text from documents can be performed in computationally different ways, which determine the nature of the 'signal' on which the visualization is based. On the downstream side, the visualization itself needs to be incorporated into an analytical briefing and presentation. Different kinds of visualizations are appropriate for different audiences and objectives. All of these considerations effect the nature and operation of the visualization itself.

This paper describes the results of the first year of a multi-year effort to apply visualization technology to the problems of intelligence analysis. It describes system requirements necessary for processing data, and core functionality provided to users to facilitate the visualization and analysis of large textual databases. The initial product of this R&D project is called the Galaxies Visualization Environment (GVE).

2. Text Visualization

The defining duty of GVE is text visualization. In order to visualize documents containing textual information, a meaningful transformation from words to images must be provided. This section describes the components which cooperate to produce visualizations from raw text.

Figure 1 illustrates the four principal components involved in the transformation. The source data component represents a generic store of documents to be visualized. It is the subject of the visualization to be produced. The original source of the text is varied - from memos and communiques to scientific journal articles to open source information retrieved from the Internet, wire services, etc. The second element in the illustration is the text engine. This important element processes the raw material of the data source and produces intermediate values, measures of similarities between all pairs of documents, from which the visualization will be constructed. The third element is comprised of a suite of n-D mapping algorithms. The mapping algorithms compress high dimensional values from the text engine into a smaller (currently two) dimensional display space. The final component identified in Figure 1 is the visual display and accompanying graphical user interface. Following are further descriptions of the components involved in the transformation from text to image.

3. Text Engine

The first functional component in the text to image transformation is the text engine. It is this component that processes documents and their constituent text into numeric quantities necessary for subsequent visualization.

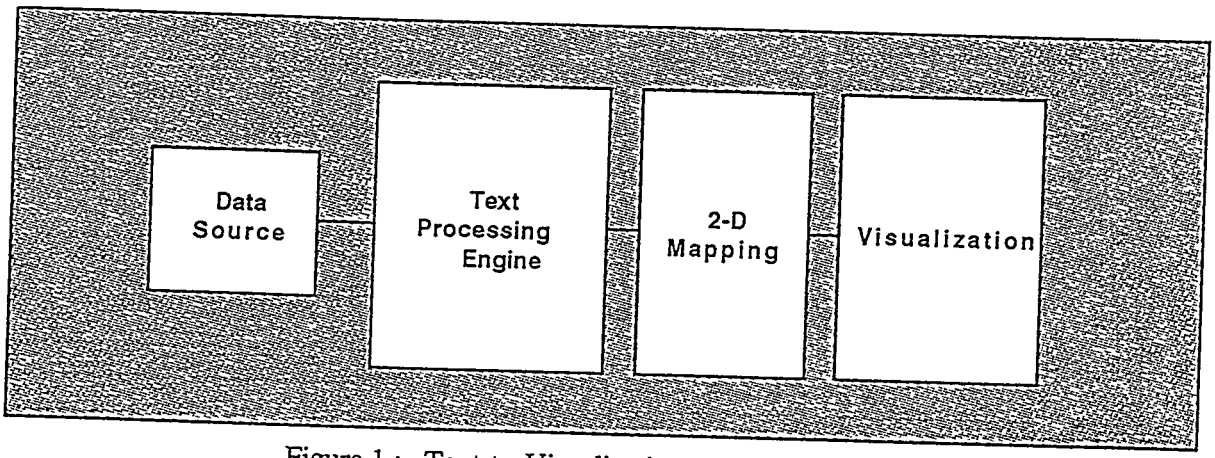

Figure 1 : Text-to-Visualization Transformation

A variety of potential text engines based on similarity measures have been, and continue to be, considered for use in the GVE application. GVE has three primary requirements of the text engine. First, it must provide specific document descriptors. These include such items as document identification tags, document names, and dates of publication, receipt, etc. Descriptors facilitate identification, access and filtering of the document corpus.

Second, GVE requires it's text engine to support queries. Querying the document base allows analysts to use a combination of traditional and novel visualization approaches to explore information. Several types of queries should be supported, the most common and most general being the text entry, keyword-based boolean query. Other query methods include "query by example," which processes a selection of text and searches for matches in the corpus, and "implicit query by proximity." Proximity queries allow similar documents to be retrieved through visual identification.

Third, GVE requires a text engine to provide similarity measures between each pair of documents. Similarity measures, to one degree or another, reflect the topical relatedness of documents. The scatter plot display of all documents relies on similarity data in that the position of any document point in the plot reflects its degree of similarity to all other documents in the display. Without similarity measures, the spatial relationships among documents, so critical to the GVE visualization, could not be derived. Further, the quality of the similarity measurements directly and profoundly affects the utility of the visual displays derived from them.

Various text engines produce similarity measures which broadly fall into three classes: frequency-based, semantic-based, and statistical-based similarity measures. In frequency-based methods, the frequency of unique words in a document is used to characterize that document numerically. Each document is represented as a two-dimensional digital "signal" with one axis representing separate words and the other axis the frequency of occurrence. Documents that are close in the normalized form of this signal are assumed to be close in meaning. (This is the method used in the first release of GVE.) The primary strength of this approach is its simplicity, with no unique computational requirements other than word search and count. The chief weakness of frequency-based similarity is that only first-order information, i.e., the presence or absence of keywords, enters into the numerical characterization of a document.

Semantic-based similarity measures interpret the meaning of the text. This method is frequently implemented using mechanisms such as semantic nets and spreading activation. Semantic approaches utilize quasi-natural language understanding techniques, so that the semantic relationship, i.e., higher-order information, is considered. The weakness of this approach is the magnitude of the task; the number and strength of relationships among all words must be maintained, expensive in both development time and computation.

The third major class of similarity measures is statistically based. Statistical measures of similarity rely on the extraction of features from documents in a corpus. Much as features are extracted from 2D digital images to compose objects, words and phrases are identified in context to determine the topic of a document. Words and word combinations standing as features in a feature space are weighted through the process of extraction to compose n-feature vectors in a high-dimensional feature space. The strength of statistical approaches to the derivation of similarity measures is the same as the semantic approach: higher-order correlations are derived. In addition, the approach is generally language-independent, since no semantic or *a priori* structural understanding of the language is required. The weakness is that semantic information is completely ignored.

4. n-D Mapping

Regardless of the method used by text engines to derive similarity measures, our required result is the same: GVE must have numerical measures of similarity for all documents. The GVE visualization requires a low-dimensional (2 to 5-D) representation of documents that inhabit a high-dimensional (hundreds of dimensions) space. It is the purpose of the n-D mapping algorithms to compress the high dimensional output of the text engines.

Two classes of algorithms were developed to perform the GVE dimensional compression. The first class of algorithms is principal components analysis. In this approach, the axes of the high dimensional space are selectable based on maximal variation. Intuitively, potential axis are analyzed until the those with the highest information content are identified. These axes become the coordinates of the display space, and all other dimensions are projected onto the selected coordinate system. For example, if the desired dimensionality for display is two, and the similarity measures represent a 50 dimensional hyperspace, then 2 principal axes would be derived and the other forty-eight dimensions would simply be projected onto the principle axis.

The second class of algorithms is multi-dimensional scaling (MDS). MDS algorithms require pairwise measures of distance or similarity and iteratively calculate the representation which most faithfully (least mean squared error) represents the distances in the selected lower dimensional space. Two types of MDS algorithms were studied: the metric MDS and the non-metric MDS. Metric MDS algorithms rely on similarity measures derived from a metric space. Statistical-based similarity engines, for example, may supply a feature space which is also a metric space. Metric multi-dimensional scaling algorithms can use pairwise distances from documents in the metric space to produce the low dimensional visualization. If the similarity measures are not derived from a metric space, then the non-metric MDS algorithm should be used to construct the lower dimensional representation.

5. GVE

Once a corpus has been processed by the text engine GVE provides the analyst with the ability to browse, query and analyze the database both visually and textually. As shown in Figure 2, the GVE interface consists of several areas. The visualization area to the left, the tool area to the right, and basic interface functions (message windows and grouping functions) along the bottom of the screen.

The purpose of these areas are to provide the analyst with the ability to understand the corpus and manipulate it to sift out relevant and interesting documents.

5.1 GVE Requirements

Basic user requirements for GVE were based on preliminary interviews with and observations of intelligence analysts at work. To successfully facilitate browsing, visualizations should provide a minimum of six basic capabilities to the information analyst. These capabilities are listed below:

- Represent the entire corpus visually and intuitively display the relationships among all the documents in the database.
- Provide simple browsing and querying capabilities.
- Provide methods for analysts to save observations and notes through annotations and "landmarks".
- Provide analysts with reference points while navigating through the corpus.
- Allow analysts to perform cognitive processes in parallel rather then serially.
- Allow analysts to detect and view change of document relationships as a function of time.

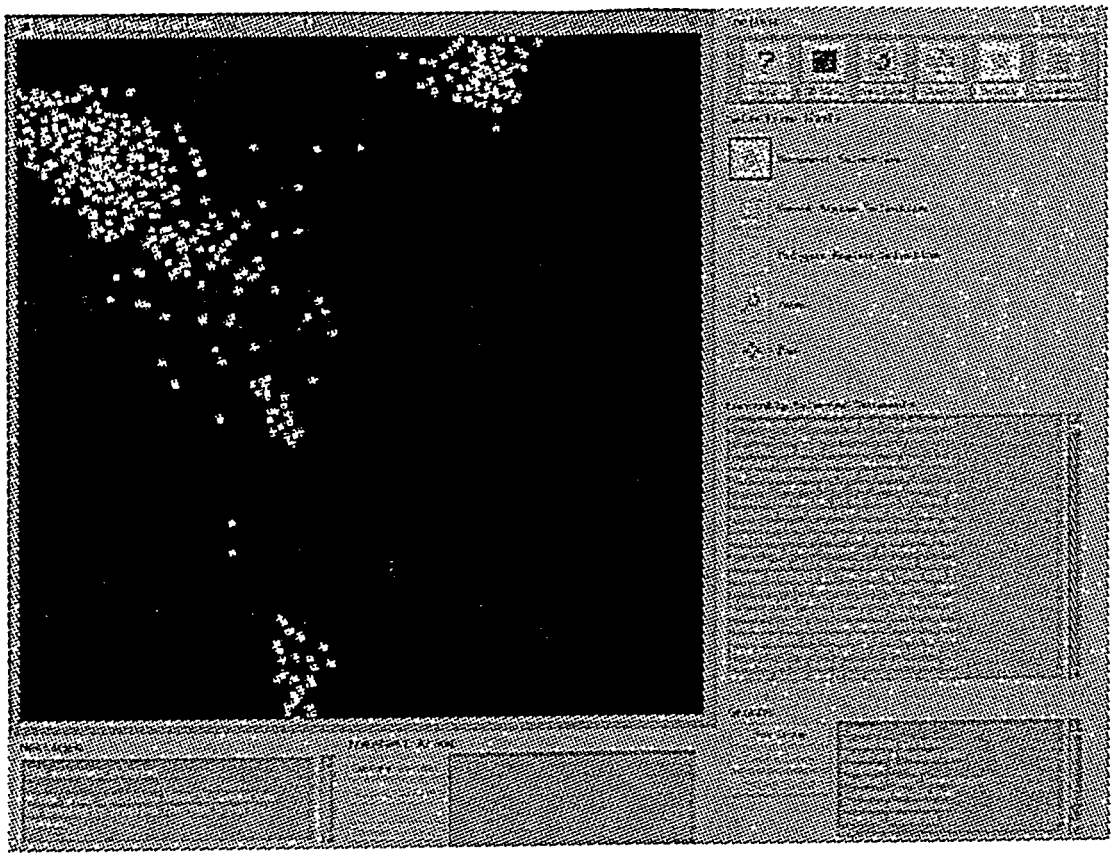

Figure 2: GVE Screen Layout

5.2 GVE Implementation

Visual Representation of the Corpus

Although simple, the 2D scaled scatter plot displayed in the visualization area provides a mechanism to present all elements of the information space at once. As displayed, the scatter plot looks very much like a universe of stars, with document clusters forming the visual effect of "galaxies". The properties of a scatter plot allow many units of information and their interrelationships to be shown simultaneously - interrelationships which might otherwise go undiscovered.

In the GVE application, the scatter plot is created by processing documents through a text engine to quantify their similarity. Subsequently, multidimensional scaling algorithms are applied to the relationship measures, and are presented in the visualization area. The points displayed in this area represent individual documents, and the degree of similarity between documents is expressed by the proximity of each point in the display. A group of documents which are closely related will cluster together in a tight group, while unrelated documents will be separated by large spaces.

The simplicity of the scatter plot provides intuitive feedback to the user about the relationships of documents across the entire corpus. This accelerates analysis by providing the user with an understanding of topical clusters and their relationships to each other. It also provides simple mechanisms for identifying redundant information, as well as topical outliers which may be of interest.

Querying and Browsing

Querying and browsing the corpus are the most basic capabilities needed in an analysis tool. GVE accomplishes these requirements through a combination of different approaches.

The query tool allows users to enter boolean queries as strings of text. This tool differs from traditional queries because the results are displayed visually. The whole document corpus is represented as a scatter plot where query results are highlighted allowing users to quickly identify patterns. Additionally, users can identify and create groups from the query results, and visually compare the contents of multiple groups for intersecting points.

Browsing in GVE can take place at many levels, such as corpus, topical and document, without requiring the analyst to switch views. The corpus can be browsed by simply examining the visualization area. The 2D scatter plot allows users to browse through clusters of interest and pick out trends in the data. At a topical level, groups of documents defined from the results of queries can be highlighted. Intersections between groups can be identified through color changes to illustrate where topical areas intersect. Finally, browsing can be performed at a document level by clicking on a document in the visualization and reading its contents in the tool area. This is especially important as the visualization alone cannot provide enough information for a thorough analysis.

Notes and Annotations

Analysts may typically spend hours or weeks examining the contents of a database. GVE goes beyond providing simple analysis tools by providing the ability for an analyst to capture their ideas and thoughts during the analysis process. The Annotation tool allows analysts to associate notes and colors to specified groups and individual documents for future reference. This is important, as it provides the analyst with the ability to attach their own interpretations to the actual data which led them to their conclusions. In a workgroup, annotations may be shared between individuals to enhance collective interpretation. This allows specialists in different areas to share results and gain a better understanding of the information spread through the entire corpus.

As users check document contents, they can highlight points in the scatter plot, form them into groups, label and annotate them. Analysts can create their own patterns based on their concepts and interpretation of the meaning of the documents. As they highlight individual documents and groups of documents, they soon create their own topography, or patterns of change of meaning, over time. These representations may continue to be clusters or linear pathways.

Reference Points

When dealing with large databases, navigating through a corpus by zooming and panning can easily become disorienting. To help prevent this, orientation and reference points are provided to the analyst.

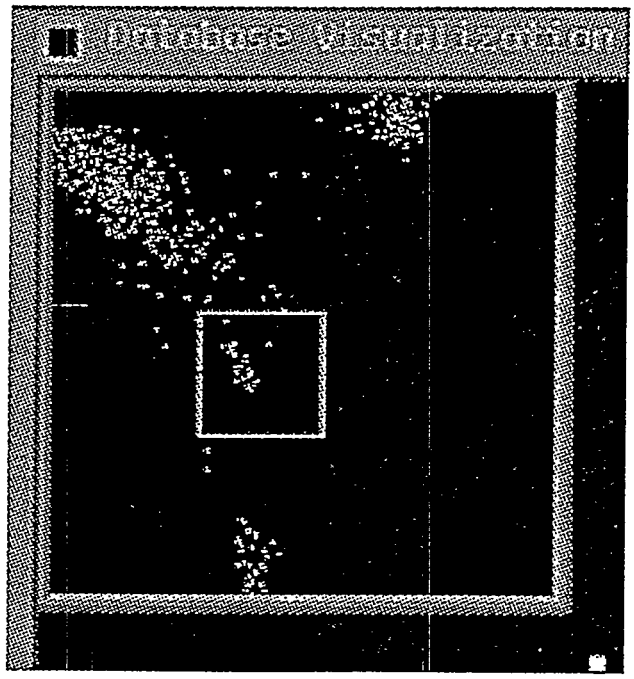

Figure 3: Corpus Overview

The patterns of document points themselves facilitates orientation. Distinctive groups such as very large clusters serve as a reference point or anchors which can be used as a basis to remember where other points and clusters are located.

Orientation cues are also provided to the user as they zoom into and out of clusters by an overview map of the database. This small inset map is located in the top left-hand corner of the visualization area, and contains a miniature view of the scatter plot (see Figure 3). When a user elects to zoom in or out of selected areas of the visualization, the overview map highlights the region which is currently visible. This provides the user with a firm anchor to the overall data, and allows the analyst to maintain a constant perspective during their procedures. Figure 3 illustrates how the overview provides the analyst with a global perspective.

Additionally, the annotation tool can be used to highlight and label documents and groups in the visualization area. This allows analysts to create their own visual markers by changing document colors to denote important features in the corpus. Each highlighted document or group can also be labeled, providing textual information about document contents.

Parallel Cognitive Processing

The scatter plot is an effective representation that stimulates spontaneous (preconscious) perception of structure and patterns in the data. It stimulates our natural capacity to perform several cognitive processing tasks in parallel. Examples include both lower order sensory and higher order analytical pattern recognition that occur in visual scanning.

Detecting Changes Over Time

Every document in a corpus contains a date associated with its release. This date can provide important information about a chain of events buried within the information space.

GVE provides a time-sliced MDS tool which allows users to display the contents of the database as a function of time.

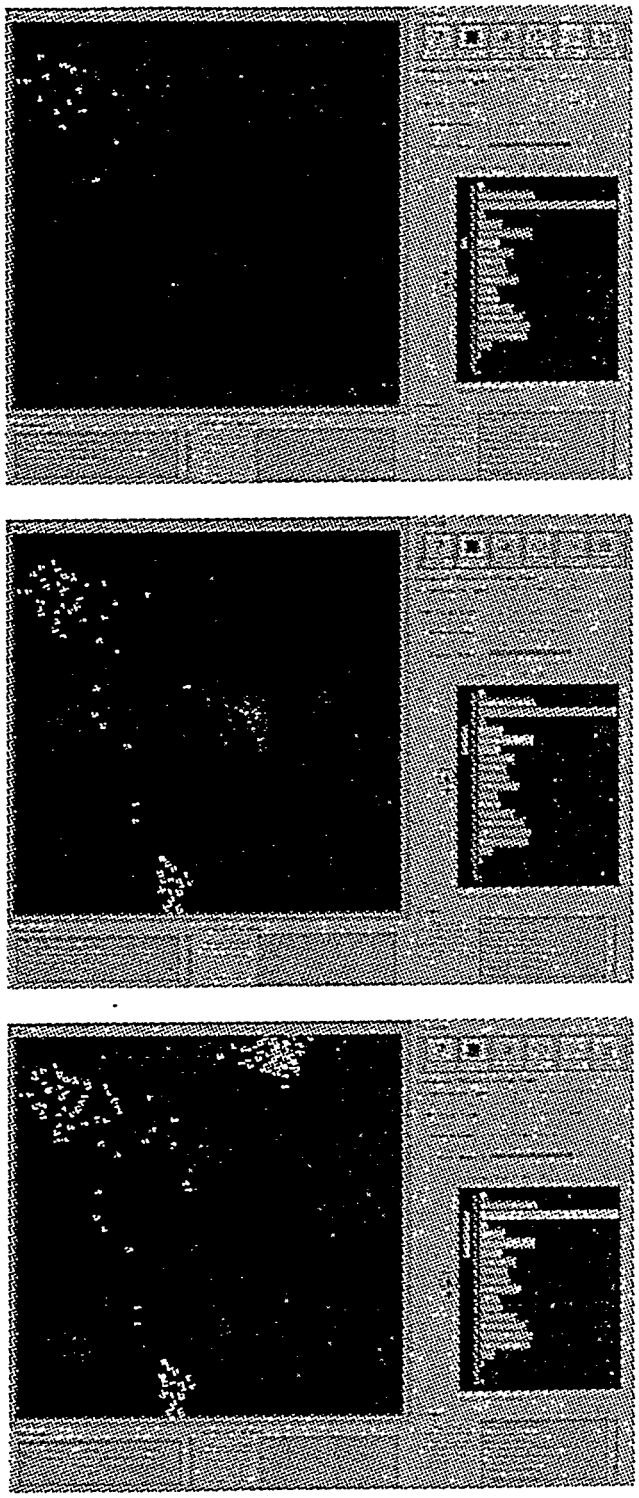

Figure 4: Temporally Slicing a Corpus

Time may be expressed as a range of years or a single year. Figure 4 illustrates a typical slicer display. The bars in this display represent the number of documents in a given year.

The MDS slicer provides the user with two approaches to analyzing the corpus temporally. The first is by slicing the corpus to visualize the documents relevant to a particular period of time (i.e. a single year). The second approach is to gradually build the visualization and watch the corpus populate itself over time. Figure 4 illustrates the effects of slicing through a database.

Slicing through a corpus temporally and showing this through animation is especially important for detecting trends across documents. Watching the corpus unfold, and clusters emerge over a period of time provides a mechanism to link topical patterns with temporal patterns. For example, the emergence of clusters indicating research into specific topical areas may outline a process for weapons development programs. When compared with current events, these trends can provide detailed insight into political motivations and help analysts formalize hypotheses explaining their occurrence.

6. Results and Conclusions

Recently, the GVE, packaged as an integrated tool within the Pathfinder 7.0 application, was rolled out to an audience of 50 intelligence analysts and other interested parties. (Project Pathfinder is sponsored by the US Army National Ground Intelligence Center for the US intelligence community.) A test problem, involving a 500-document corpus pertaining to a specific country, was visualized using GVE's scatter plot and immediately produced three obvious clusters. The clustering was subsequently shown to be due to documents being related through three different technologies. At this point, boolean queries pertaining to other countries were used to indicate, through GVE's annotation and highlighting features, where the subject country was obtaining technology in each of these three subject areas. Thus, GVE provided information and insight over and above that obtainable through traditional means. This showed that a single, global view of a document corpus could, indeed add value to the analysis process.

Through a simple visualization we have demonstrated the beginnings of a new method that integrates multiple analysis activities in a single context. Preliminary results indicate that the scatter plot visualization enhances the analyst's ability to browse through a large corpus and perform high-order analysis tasks. In addition, the ability to visually index and annotate a large number of documents in a single display window appears to ease the burden of traditional search and query operations.

7. The Future

The capabilities provided thus far are clearly rudimentary. The amount and degree to which our initial efforts have received positive feedback from the user community indicates that the potential for future work is indeed ripe. Our next steps are to evaluate what we have created and, based on the results of this evaluation, expand it and develop new capabilities that assist analysis. For example, we have yet to enable the analyst to perform complex analysis, query by example, or to move seamlessly between functional spaces. We plan to extend the "Galaxies" concept to allow seamless transformation to different views that can show tightness of correlation between documents and allow the analyst to create their own information landscapes that represent their views of knowledge and concepts. One visualization concept we plan to explore is the presentation of documents in the form of information topographies where, like a geographic map, peaks and valleys can be easily detected based on contour patterns. These views may be both two dimensional with respect to space with further richness provided by adding other visual dimensions such as transparency, layering, dynamic interaction and animation. These techniques can take advantage of the human's innate perceptual and cognitive abilities to function in natural landscapes. Additionally, we plan to enhance the interaction paradigms we have already created to allow the exploration of change in specific individual or combined parameters over time.

Acknowledgments

The authors would like to thank our client, the Joint National Intelligence Development Staff and specifically, Mr. David Lutz, for funding and support of this first phase of the MVAB Project. We also thank the staff of Project Pathfinder and specifically, Mr. Tim Hendrickson, for insight and suggestions from the perspective of the user community. Finally, we wish to thank the Pathfinder development staff of Presearch, Inc. for their continuing cooperation and support in integrating the GVE with Pathfinder 7.0.