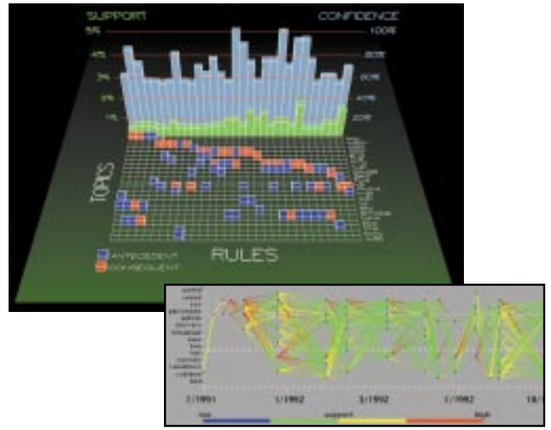
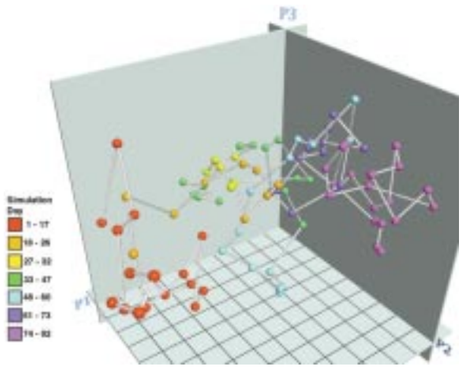
Visual Data Mining

Discovering associations and sequential patterns

The visual data mining system integrates visualization tools with more conventional text mining approaches—and users get the best of both worlds. They can discover more knowledge from a large collection of data and do so more quickly than if either method was used alone. The visual data mining system allows users to study many complex relationships in a large data set simultaneously. Incorporating a front-end visualization system with a data-mining engine's statistical pattern identification helps users find and understand associations and sequential patterns. This system presents a graphical representation of the various relationships, their strengths, and how they're distributed over time.





Data Signatures

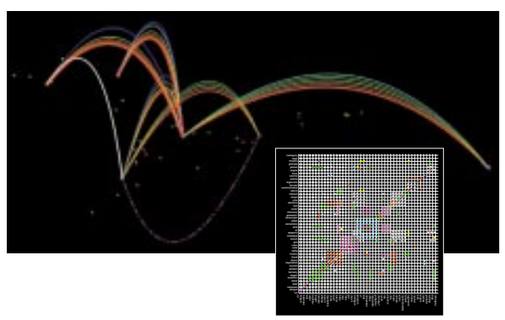
Simplifying very large data sets

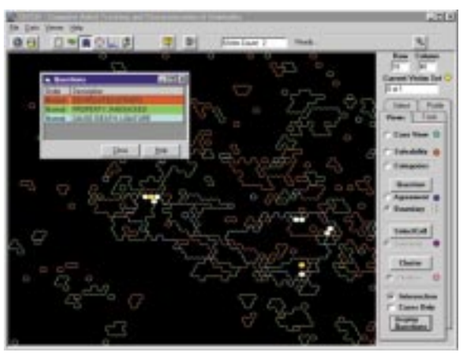
The concept of data signatures is being explored as a new way to analyze and understand large data sets such as the scientific computations associated with climate modeling or combustion simulations. A data signature is a mathematical data vector that captures the essence of a large data set in a small fraction of the size. Because conventional analysis techniques are not effective in dealing with such large quantities of data, this filtering technique could be used to eliminate the "noise" caused by less important features. These signatures could be used to conduct analysis by examining the essence of the data, yet still reflect the intended results as if the entire original data set were used.

Rainbows and Connex

Discovering relationships and connections

Rainbows and Connex are tools that allow users to explore multiple relationships among entities including documents, people or concepts. With a combination of dots and arcs, Rainbows can display progressively more detail about relationships among entities and identify places of interest where there is an absence of a relationship. The Connex visualization looks like a three-dimensional graph, where the x- and y-axes correspond to an ordered list of entities. Connex enables users to categorize relationships, highlight asymmetric relationships and discover combinations of relationships.





CATCH

Neural networks for law enforcement

Computer Aided Tracking and Characterization of Homicides, or CATCH, helps crime analysts interpret large databases by locating and grouping similar cases. The Artificial Neural Network (ANN) employed in CATCH "learns" about existing crime cases in the database of the Homicide Investigation Tracking System Unit of the Washington State Attorney General's Office. The ANN groups cases with similar elements using parameters such as modus operandi and signature characteristics of offenders. Investigators can use CATCH to compare one case with others in the database.

PNNL's Visualization Website:
http://www.pnl.gov/infoviz

PNNL-SA-33653 September 2000

For more information contact:
Dennis McQuerry • *Email:* mcq@pnl.gov
Phone: (509) 375-2953 • *Fax:* (509) 375-3641
P.O. Box 999 MS: K7-28 • Richland, WA 99352

Science.

Technology.

Innovation.

From doctors interested in the latest cancer treatments to businesses examining technology trends, users can apply information visualization technologies to

- Sort through mounds of data
- Quickly identify themes, trends and relationships
- Interact with data to explore, discover and learn.

For more information, see http://www.pnl.gov/infoviz

Pacific Northwest National Laboratory

Operated by Battelle for the U.S. Department of Energy

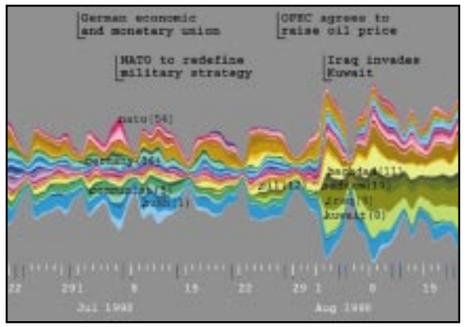
Visual Information Analysis

A picture is worth a thousand words

Pictures instantly convey information that the spoken or written word does not. Based on the ability of the human mind to rapidly perceive visual information, Pacific Northwest National Laboratory is developing technologies that use visual representations of data to help people discover hidden knowledge quickly and easily.

ThemeRiver™

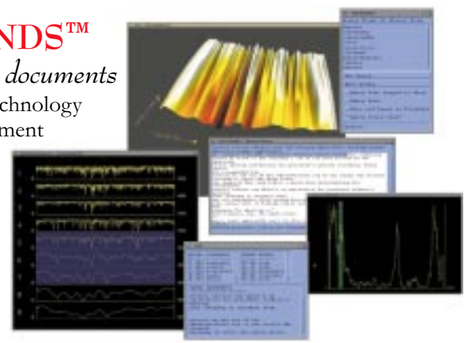
Discovering hidden trends



The ThemeRiver™ visualization helps users identify time-related patterns, trends and relationships across a large collection of documents. The collection of documents is represented as a "river" that flows through time. The river's overall width changes to depict changes in the collective strength of selected themes in underlying documents. Individual themes are represented as colored "currents" flowing within the river. These currents narrow or widen to indicate decreases or increases in the strength of individual themes at any point in time.

TOPIC ISLANDS™

Understanding large documents



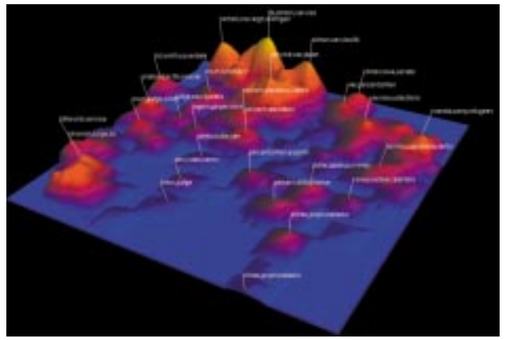
TOPIC ISLANDS™ technology transforms a large document into visualizations and excerpted summaries. Creating a hierarchy similar to an outline, this tool helps users determine what topics and themes are covered in a document. This novel approach to explore unstructured text allows users to browse a document, generate fuzzy document outlines and summarize text by levels of detail and according to their interests. Relying on an underlying technology called TOPIC-O-GRAPHY™, which applies wavelet technology to identify theme changes within a document, TOPIC ISLANDS™ allows users to filter content to meet their needs.

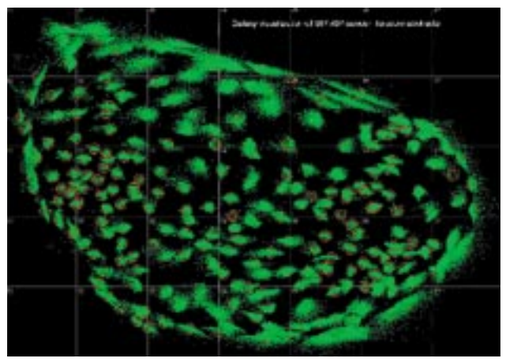
Spatial Paradigm for Information Retrieval and Exploration (SPIRE) *is a suite of tools for visual information analysis developed at Pacific Northwest National Laboratory. It automatically identifies key themes in a large set of documents and displays graphic images that illustrate document relationships based on these themes. Two visual representations within SPIRE are ThemeView™ and Galaxies.*

ThemeView™

Seeing the big picture

The ThemeView™ visualization provides a visual overview of major topics that appear in a set of documents. With this visualization, major concepts and themes are represented like relief maps used to show an area's topography. Documents with similar themes are grouped and appear together, while unrelated themes appear farther apart. Dominant themes rise from the surface like mountains, each with an automatically generated label that displays key topics. The higher the peak, the greater the relative strength of the topics within the documents in that region. Analysts using ThemeView™'s exploration tools can identify unanticipated relationships and examine how hot topics changed over time.





Galaxies

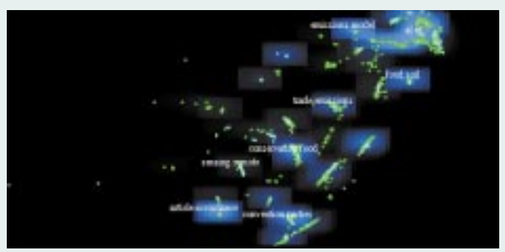
Detailed analysis of text

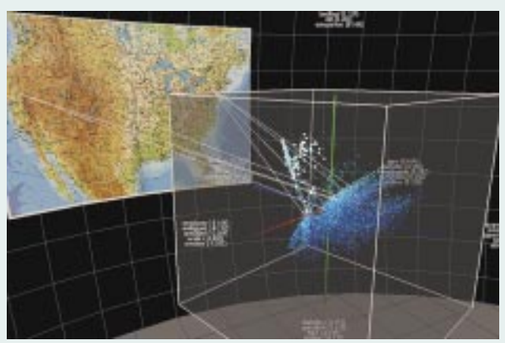
The Galaxies visualization depicts a set of documents organized by key themes. The sample shown here illustrates 567,436 cancer literature abstracts. Each document is displayed as a single "docustar." In the universe of the computer screen, closely related documents cluster together, like constellations, each labeled with key topics. Analysts can use Galaxies' exploration tools to investigate document groupings, query document contents and investigate time-based trends.

WebTheme

Discovery on the world wide web

The WebTheme visualization provides a new way to rapidly identify themes and relationships among thousands of pages of web-based text. It harvests data from the World Wide Web using search terms or by following links derived from user-specified web addresses. Unlike SPIRE, which is Unix-based, WebTheme can be operated on any computer with web browser software.



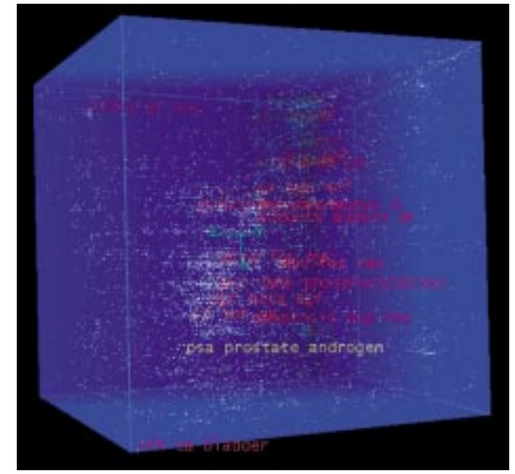


Starlight

Linking diverse databases

Originally developed for intelligence analysis applications, Starlight software is intended to support rapid, concurrent analysis of complex information, including structured and unstructured text, geographic information, and digital imagery. The system, which is Windows NT®-based, uses three-dimensional visualization techniques that interactively generate representations of explicit and implicit relationships in information collections of various types.

The Visual Analysis and Synthesis Toolset (VAST) *is a software suite designed to process and visualize massive quantities of information in a Microsoft® Windows environment. As new approaches for processing and visualizing information are developed, its software framework will readily accommodate expansion and enhancement. The ultimate goal is to produce a flexible system that can visualize a variety of information types in different ways and provide an extensive suite of exploration tools to facilitate discovery. Initial components include the Hypercube and Starstruck visualizations.*



Hypercube

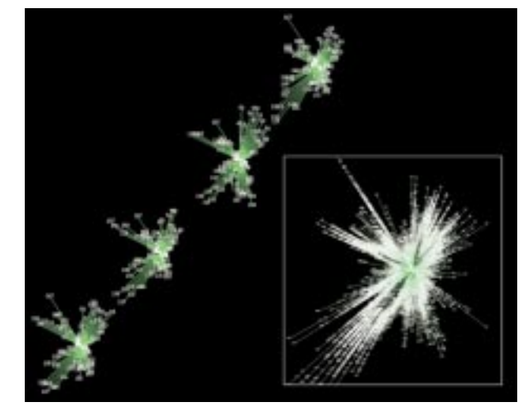
Taking advantage of taxonomies

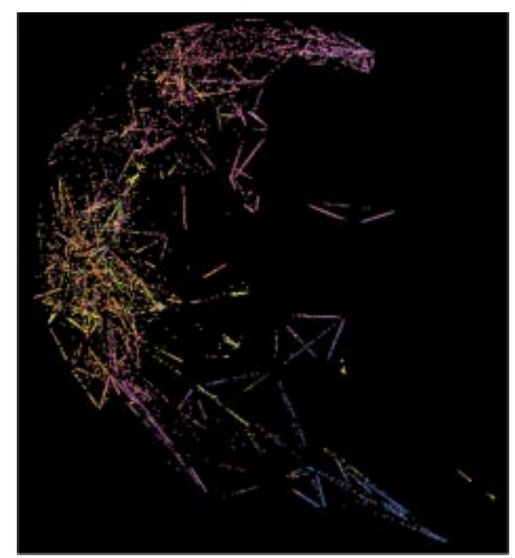
The Hypercube visualization shows how a collection of documents relates to selected ordered sets of concepts such as the Medical Subject Headings or the Dewey Decimal System. The benefit of this visualization is that analysts can quickly go to the same location to get information of interest regardless of changes in the data set being visualized. In contrast, Galaxies and ThemeView compare the relative strength of topics among the documents present so the locations of topics change as different data sets are analyzed. The Hypercube creates a semi-static framework from the most relevant concepts in each document, drawing from a set of standard terms.

Starstruck

Detailed document comparison

The Starstruck technology shows details about themes for a small group of documents. Each document is displayed as a planar starburst made of individual rays, where each ray represents a single topic or theme. The longer the ray, the stronger that topic is in the document. The position of the rays around each starburst is the same so users can look at individual stars in more detail to recognize specific patterns. By rotating the view, the stars stack on top of one another and the distribution of rays illustrates patterns in themes or topics for the entire group of documents.





Cosmic Tumbleweed

Identifying relationships among concepts

Analysts can determine how topics within large collections of data are interrelated with the Cosmic Tumbleweed visualization. Each topic is represented as a point in three-dimensional space. A dot representing each document falls between the two or three strongest topics it contains. When a large data set is analyzed, the dots form curved lines between related topics. By selecting points along the lines, users can find documents relating to topics of interest.