

# Visual Text Analysis

The information revolution is upon us—and one of the most immediate indicators is information overload. Even the simplest of information retrievals can become overwhelming. Most of the information comes in text form, ranging from electronic mail and Internet documents, to scientific abstracts and papers, to newspaper articles. The text is often unformatted and written by many people, in different styles, for various purposes. The contents often have fuzzy and incomplete messages, are interrelated but not directly, and provide a literal pile of paper to read.

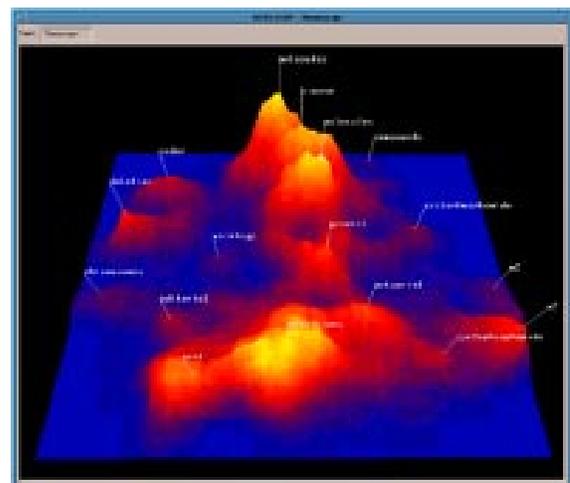
To efficiently locate relevant information, traditional text analysis approaches require the analyst to have knowledge of the information before retrieval. Once the information is collected, significant reading time still is required before deciding which documents are most relevant. Visual text analysis is a fundamentally new approach that enables the human mind to discover content within large text document sets with minimal required reading.

An interdisciplinary team of computer and cognitive scientists at the Pacific Northwest National Laboratory has developed a suite of information access and visual analysis tools that help clients quickly extract and use the information they need. The suite of new technologies is represented by the name SPIRE, or Spatial Paradigm for Information Retrieval and Exploration. SPIRE accepts large volumes of unformatted text, determines the dominant topics and relationships within the text, and presents them in a visual format that is natural for the human mind. This approach allows users to rapidly discover hidden information by reading only the pertinent documents.

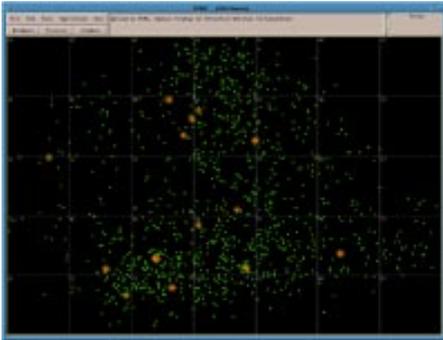
## What SPIRE Does

SPIRE is a new and exciting software application that allows users to explore complex themes and relationships found in written documents. A collection of visual and interactive tools, SPIRE graphically displays images based on word similarities and themes in text. No prior knowledge of the information or selection of themes or topics is required. SPIRE creates its visualizations by processing these similarities into the key topic and themes and organizing the data into visual representations that promote exploration and discovery.

Two visual representations within SPIRE, Galaxies and ThemeView™, provide natural visual metaphors that require little training to use. The Galaxies representation uses measures of document similarity, based on word usage, to produce a scatter plot of documents that looks like a universe of docustars.” Closely related documents will cluster together in a tight group, while unrelated documents will be separated by large spaces.



ThemeView™



*Galaxies*

In ThemeView™, themes within the document spaces appear on the computer screen as a relief map of natural terrain. The mountains in ThemeView™ indicate dominant themes; valleys indicate weak themes. Their shapes—a broad butte or high pinnacle—reflect how the thematic information is distributed and related across documents. Themes close in content will be close visually.

Once the visualization tools have displayed the content similarities and themes in the documents, users can refine their search by using several built-in support functions including a document and cluster characterization tool, a word search tool, a time analyzer, and an annotation tool. More tools are planned as SPIRE is further developed.

### **How SPIRE Can be Used**

SPIRE, originally developed for the United States intelligence community, has a wide variety of potential applications. Corporations researching competitive products, health care providers searching patient records, or attorneys reading through previous cases could all benefit from the SPIRE technology.

For example, a researcher could use SPIRE to find out in what direction the United States was heading in

breast cancer research. Drawing from a large, unstructured document base of information, the researcher uses SPIRE's visualization tools to automatically organize the documents into clusters according to their content similarities and into thematic terrains according to the themes in the text. Looking at these thematic spaces, over time, will enable the human mind to understand vast interrelated dynamic changes simply not possible to detect using traditional approaches.

This technology has been used on many topical and current issues of national security. Results have fielded insight into complex issues with enormous time savings. Discovering trends, finding interrelationships among topics, and rapidly identifying key issues are all benefits of using SPIRE.

---

For Further Information Contact:

Dennis McQuerry  
Pacific Northwest National Laboratory  
P.O. Box 999, MSIN: K7-22  
Richland, WA 99352  
Telephone: (509) 375-2953  
Facsimile: (509) 375-3641  
E-mail: mcq@pnl.gov

