The Need For Metrics In Visual Information Analysis

Nancy Miller, Beth Hetzler, Grant Nakamura, and Paul Whitney
Pacific Northwest National Laboratory

1 ABSTRACT

This paper explores several methods for visualizing the thematic content of large document collections. As opposed to traditional query-driven document retrieval, these methods are used for exploring and gaining insight into document collections. For our experiments, we used 12,000 medical abstracts. The SPIRE™ system was used to create the mathematical signal from text and to project the documents into a universe of "docustars" and as a thematic contour map based on thematic proximity. A self-organizing map is used to project the documents onto a "Tree" fractal. A topic-based approach is used to align documents between concepts in the "Cosmic Tumbleweed" projection. In the 32-D Hypercube, documents are organized by cascading theme strengths. An argument is made for a new type of metric that would facilitate comparisons among the many methods for visualizing or browsing document collections. An initial organization is proposed for some of the relevant research that metrics for information visualization can draw upon.

1.1 Keywords

Information Visualization, Browsing, Multidimensional Information Space, Projections, Metrics, Clustering, Fractals.

Pacific Northwest National Laboratory
902 Battelle Boulevard; P.O. Box 999
Richland WA 99352
(509) 375-6979

E-mail:
Nancy Miller: ne_miller@pnl.gov
Beth Hetzler: eg_hetzler@pnl.gov
Grant Nakamura: gc_nakamura@pnl.gov
Paul Whitney: pd_whitney@pnl.gov

2 INTRODUCTION

Information visualization systems can enable users to quickly determine the general subject areas of hundreds to millions of documents. Users are then tasked with identifying which of these documents should be examined in greater detail. This kind of information analysis problem, which increasingly occurs, is not directly addressed by a system that retrieves relevant documents based on user-defined queries. Our approach is to create visualizations of document collections to help the user understand the collection as a whole, discover important hidden relationships, and formulate insights with a minimum of reading. One difficulty that we face is that the performance of these visualizations is not easily assessed using current information retrieval metrics.

In this paper, we first look at several methods of visual text analysis that have been developed by an interdisciplinary team of computer scientists, cognitive scientists, and mathematicians at the Pacific Northwest National Laboratory in response to such user needs. These systems include the SPIRE system [15], a fractal-projection system, the "Cosmic Tumbleweed," and the "32-D Hypercube." The SPIRE system is far more mature than the latter three, which are works in progress and have yet to be tested. While different, all four approaches share a common goal of helping the user discover connections that he or she might not make independently.

Second, we briefly review current metrics for information retrieval. And finally, we pose a question to the research community: how to assess the performance of such visual information analysis systems that may, in fact, more closely represent the way people really interact with vast information spaces, such as the World Wide Web [11].

3 VISUALIZATION APPROACHES

Each approach described herein has a method for creating and displaying images representing potentially large document collections. The input to each approach is a large collection of text documents in almost any format (for this paper, we have used the same collection of 12,000 MedLine abstracts to illustrate the approaches). The software automatically analyzes the textual content, identifies key topics and themes, and generates mathematical signals representing the various documents in the collection. These signals form an n-dimensional vector representation of the information space; the n-dimensional vector for each document represents how strongly that document is related to each of n key topics for the collection. Each system makes use of the n-dimensional vectors in different ways, providing alternative visualizations that are designed to provide different insights into the document collection.

3.1 SPIRE

SPIRE was designed to let users rapidly discover information relationships in a collection of documents by identifying the pertinent documents for reading, rather than wading through large volumes of text. The n-dimensional document signals are clustered into groups of related documents, which are then projected from n-space into 2-space and used to create the visual representations. SPIRE currently includes two visual representations: Galaxies [7] and Themescape [15]. Galaxies displays the documents to look like a universe of “docustars.” Closely related documents cluster together in a tight group while unrelated documents are separated by large spaces. In Themescape, themes within the document spaces appear as a relief map of natural terrain. The mountains in Themescape indicate dominant themes; valleys indicate weak themes. Themes close in content are visually close based on the many relationships within the text spaces. Figure 1 shows the Themescape visualization of the 12,000 MedLine abstracts. The labels supplied in the Themescape are automatically determined by the system. Different peaks represent a different emphasis in the unde

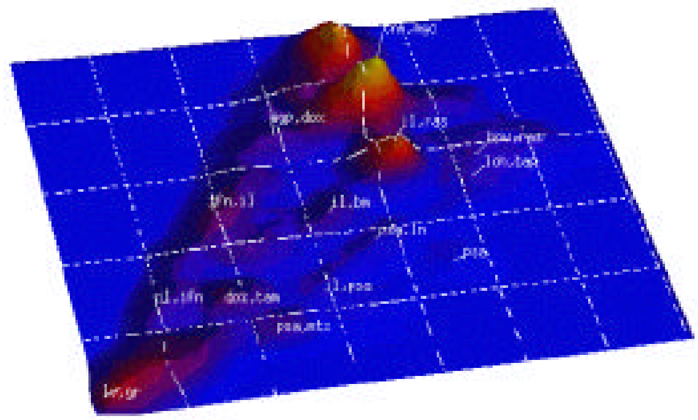

Figure 1: Themescape for 12,000 Cancer Literature Abstracts

Once the text analysis and projection tools have displayed the content similarities and themes in the documents, users can continue their exploration by using several built-in support functions.

3.2 Projecting Information into Fractal Spaces

Organizing and visualizing of information spaces by projecting onto fractal spaces is appealing for a number of reasons. Fractals have a different intrinsic measure of similarity from the Euclidean space and cosine similarity projections typically used in information visualization. Also, fractals have inherent compression properties that can be used to condense the corpus, collapsing outer nodes to a small set of predecessor nodes.

The fractal used in our initial studies is shown in Figure 2. The line segments show the structure of the fractal. Documents are projected onto the nodes (where the segments branch) of the fractal. In this figure, projected documents are shown as colored dots. This fractal is constructed by pasting together successively smaller “angles” and is a member of the class of iterated function system fractals [1]. In this “Tree” fractal, the distance between nodes is calculated by path distance within the fractal. Documents (or other information objects) are assigned to a node such that the distance between the information objects is similar to the distance between the fractal nodes (up to scale). A variant of self-organizing maps [8] is used to carry out the projection.

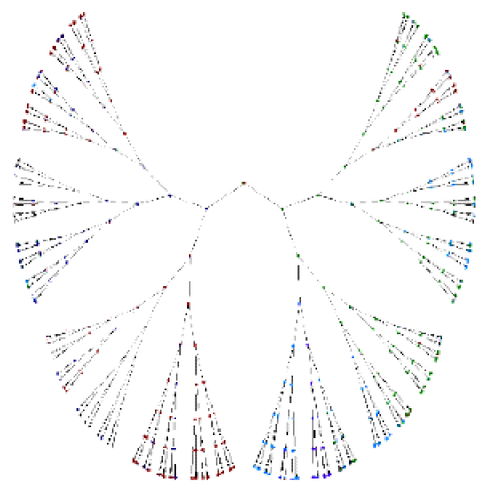

Figure 2: Fractal Projection of Cancer Document Vectors

3.3 Tumbleweed

The “Cosmic Tumbleweed” visualization is intended to quickly convey information about both the topical content of text documents and relationships among those topics. A model of the formative structure of the universe inspired the visual concept. The name “Cosmic Tumbleweed” is derived from having nodes distributed in a three-dimensional, tumble-weed-like structure of interconnected topics. Nodes in the tumbleweed represent topics and are placed so that related topics are close together. Threads between topical nodes represent relationships between those topics. Upon closer inspection, a thread between two topics is composed of individual representations of documents that are about those topics (see Figure 3 at the end of this document). This suggests that the visualization can be used interactively to progressively disclose detail. Other aspects of the visualization such as brightness and color can also be exploited. For example, brightness could convey how strongly topics are represented in the collection.

3.4 32D Hypercube

The 32D Hypercube visualization shows how a collection of documents relates to a selected ordered set of concepts, for example, the MeSH (Medical Subject Headings), the Dewey Decimal System, or an ordered version of the key concepts present in the collection itself. In the following explanation, we use this last option. Using a standard statistical correlation technique, we begin by ordering the key concepts. Then for each document, we determine which concept from the ordered set is most strongly represented in that document, which is second most strongly represented, and so on, up to an arbitrary limit of 32 total concepts. In this case, this step is achieved by examining the n-dimensional document vectors.

A three-dimensional visualization is created using the ordered set of concepts for each axis. The x-value for a document is the number of the concept that is most strongly represented in that document; the y- and z-values are the numbers of the second and third most strongly represented concepts. Similar visualizations can be shown for any trio of concept strengths desired. Our system provides rotation to help the user examine the visualization; even in the static image, certain patterns are clear (see Figure 4 at the end of this document). Concentrations of dots show where particular closely related concepts show up strongly in multiple documents. Selections can be made within each "cell" in a particular hypercube, and the documents can be reprojected using any other topic strength combinations — such as the fourth, fifth, and sixth. This recursive interaction is suggestive of World Within Worlds [3].

4 NEED FOR NEW METRICS

With several alternate visualization approaches, the question of which one is better for what set of user needs and under what circumstances naturally arises. Each visualization method is designed to emphasize parts of the discovery process; however, we still need some way to measure the actual results. It is likely that interacting with more than one visual paradigm/method offers great promise.

Metrics for query-based information retrieval are well established. The TIPSTER Text Program has ushered in truly major advances in document detection and information extraction through the Text Retrieval Conferences (TREC) and Message Understanding Conferences (MUC), respectively [URL: http://www.tipster.org/]. The querying and ranking methods of document detection can be characterized by the right answers to user-formulated retrievals. The metric of "recall" is used to measure what percentage of a predetermined set of "right answers" in the collection a system actually finds. The metric of "precision" is used to measure what percentage of answers returned by the system is actually correct. Since the start of the annual TREC competitions in November 1992, precision and recall performance has increased to focus the competitions on the fourth digit of accuracy. MUC has facilitated the growth in natural language processing technique to allow facile extraction of both proper and organization names and some scenario assessment in domain-specific areas. Our systems, though useful to our clients, would likely not shine in a TREC-type competition because the metrics do not actually measure what our system does best: make connections that are not known a priori. Note also that the measures of precision and recall do not require graphics.

Our work depends heavily on graphic presentation for the communication of content. Bertin distinguished several distinct functions of graphics; among them are the capabilities to display information already understood and to facilitate the process of understanding of the information [2]. Guidelines for the presentation of static information have been used in the statistical community for years and have been incorporated into many statistical exploratory data analysis packages [6]. Perhaps best known are the principles of visual presentation as outlined by Tufte in his various books over the years [12 - 14]. Brath's recent work proposes four metrics to assess the efficacy of static 3-D presentations: number of data points and data density; number of dimensions and cognitive overhead; percentage of occlusion; and reference context and percentage of identifiable points [4]. The advantage of these measures is that they are objective and fairly easy to measure. The disadvantage is that they are for static pictures and thus have not been extended to interactive models.

In the interactive use of graphics to gain insight into information, most researchers begin by a definition of data types. This area is more subjective and immature than the "static" metrics research but is very relevant to our work. Schneiderman has identified seven data types and seven tasks: overview, zoom, filter, details on demand, history, and extract [10]. These tasks fit very closely to our approach to information exploration. Other data types have been defined by others, such as Zhou [16] and Card [5]. Card proposes tables that can be used to track the processing for each variable used in the visualization and the interaction available for each. Data types and tables are important because they facilitate comparisons across very different visualization systems and can be used to succinctly identify similarities and differences.

An interesting area of further research would be to investigate whether we can

(1) predict successful visualizations based on objective quantities that can be easily measured, and then
(2) design the experiments to take the measurements and verify the predictions.

In our current experiment, the first informal effort was to compare our alternate visualizations for the 12,000-document corpus by importing the coordinates of each document into a common analysis package. Each document is represented in the figures below as an individual point. The points are colored according to an independent clustering so that the point representing a document is the same color in each visualization. The figures show that each display broadly classified documents similarly; the blues are separated from the reds and purples in all cases.

We pose a question to the research community: How can we measure the "goodness" of a particular or combined visualization? The recently added (1995) Interactive Task in TREC somewhat more closely measures what it is we try to do in visual information analysis. The Interactive Task is used to study both user interaction with text retrieval systems and how quickly a variety of "aspects" can be discovered about specific topics. Using a similar approach, some ideas for making this assessment might include

- contests to characterize main themes or topics addressed in a test corpus
- contests to identify hidden connections that can only be made by integrating information across documents
- variations on software usability tests.

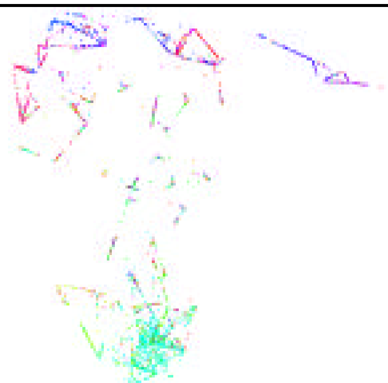

Figure 3: Cosmic Tumbleweed Projection

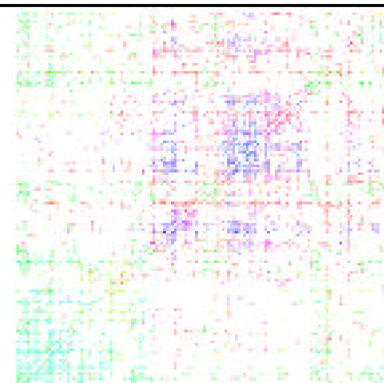

Figure 4: 32D Hypercube

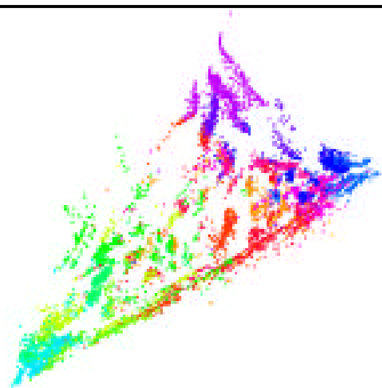

Figure 5: Spire Galaxy Projection

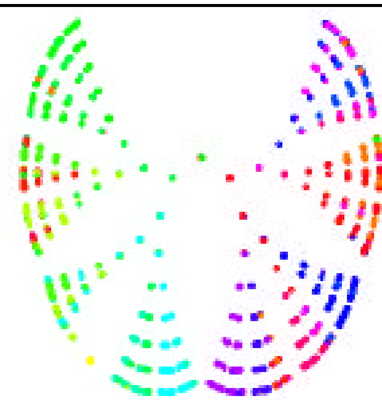

Figure 6: Fractal Projection

5 REFERENCES

1. Barnsley, Michael. *Fractals Everywhere*. Academic Press, 1993.

2. Bertin, J. *Graphics And Graphic And Graphic Information-Processing*. Berlin. Walter de Gruyter, 1981.

3. Beshers, C., Feiner S. AutoVisual: Rule-Based Design Of Interactive Multivariate Visualizations. *IEEE Computer Graphics And Application*, 13(4), pages 41-49, 1993.

4. Brath, Richard. Concept Demonstration: Metrics For Effective Information Visualization, in *Proceedings For IEEE Symposium On Information Visualization*, pages 108-111. IEEE Service Center, Phoenix, AZ, Oct 20-21, 1997.

5. Card, S., and Mackinlay, J. The Structure Of The Information Visualization Design Space, in *Proceedings for IEEE Symposium on Information Visualization*, pages 92-99. IEEE Service Center, Phoenix, AZ, Oct 20-21, 1997.

6. Chamber, J., Cleveland, W., Kleiner, B., and Tukey P. *Graphical Methods For Data Analysis*, Wadsworth International Group, Belmont, CA, 1983.

7. Crow, V.L., Thomas, J.J., et. al. Multidimensional Visualization And Browsing For Intelligence Analysis. *Gviz '94: Graphics And Visualization Conference*, Richland, WA, September 1994.

8. Kohonen, Teuvo. *Self-Organizing Maps*. Springer-Verlag, New York, 1995.

9. Over, P. Interactive Track Report, in *TREC-5 Proceedings*, Department of Commerce and National Institute of Standards, URL: http://trec.nist.gov/pubs/trec5/t5_proceedings.html (file: trackreport.ps).

10. Shneiderman, Ben. The Eyes Have It: A Task by Data Type Taxonomy For Information Visualizations, in *Proceedings for IEEE Symposium on Visual Languages*, pages 336-343. IEEE Service Center, Sep 3-6, 1996.

11. Thomas, J.J., McVeety, R.C., Miller, N.E. Beyond Data Mining: A New Human Information Discourse. *Advanced Information Processing And Analysis Symposium* (AIPA 97), March 25-27, 1997.

12. Tufte, Edward R. *Visual Explanations*. Graphic Press, 1997.

13. Tufte, Edward R. *The Visual Display Of Quantitative Information*. Graphic Press, 1993.

14. Tufte, Edward R. *Envisioning Information*. Graphic Press, 1990.

15. Wise, J.A., Thomas, J.J., et. al. Visualizing The Non-Visual: Spatial Analysis And Interaction With Information From Text Documents, in *Proceedings of IEEE 95 Information Visualization*, pages 51-58. IEEE Service Center, Atlanta, GA, October 1995.

16. Zhou, M. Z., and Feiner, S. Data Characterization For Automatically Visualizing Heterogeneous Information, in *Proceedings For Information Visualization 96*, pages 13-20. IEEE Service Center, 1996.