Visualizing and Browsing WWW Semantic Content[1]

Richard H. Fowler[†], Aruna Kumar[‡], Jorge L. Williams[†]

[†]Department of Computer Science
University of Texas - Pan American
Edinburg, TX 78539
fowler@panam.edu

[‡]Architecture Labs: New Media Prototypes
Intel Corporation
Hillsboro, OR 97124

Abstract

In order to realize the potential of the World Wide Web (WWW) to supply information it will be necessary to significantly increase the effectiveness of existing search and retrieval mechanisms. One promising approach is visualization, which has been used successfully in an wide range of domains to provide insight into the organization of complex systems. The Document Explorer system creates a visual browsing environment for WWW documents based on semantic content, rather than on link structure such as used in other WWW browsing and visualization tools. Existing WWW indexing and content extraction tools are used together with new techniques for graph based abstraction and clustering to present dynamic, visualization of content based three dimensional space.

1 Introduction

The dramatic increase in information available throughout the Internet and through the World Wide Web in particular presents unparalleled opportunities for public information access. However, the very quantity of information coupled with the lack of universally accepted means for accessing the content of that information presents unparalleled challenges of information access. Content based search and retrieval systems have both been available for a long time, e.g., WAIS, and a number of efforts are underway to improve the effectiveness of indexing and information gathering systems [4], yet it seems likely that these tools alone will not solve the current problems of information access.

Central to the challenge of information access on the Internet are issues that are common to all forms of information retrieval. These issues include indexing indeterminacy, difficulties in using indexing vocabularies, and the user's inability to completely specify information needs [6]. It is widely recognized that retrieving information that meets users' information needs is an iterative process, and techniques which explicitly incorporated users' judgments, such as relevance feedback, have been developed to provide mechanisms guiding retrieval. It is also widely accepted that mechanisms providing alternative paths of access to information can enhance retrieval. One class of tools which seems particularly promising for augmenting the information retrieval process is visualization.

2 Information Visualization and Information Spaces

Scientific visualization presents a visual representation of a physical phenomenon to the user to enhance understanding and provide insight into the phenomenon. Information visualization shares this goal and is faced with challenges not encountered in scientific visualization. For scientific visualization the visual representation is derived

[1] This work was supported by NASA grant NAG9-551 and an Intel equipment donation.

0-8186-7585-3/96 $5.00 © 1996 IEEE

from a *physical* phenomenon, hence there typically exists a natural visual representation based on spatial, temporal, or other properties of the phenomenon to be visualized. For information visualization elements to be visualized typically have no physical component which might naturally supply a basis for visualization. Rather, the elements to be visualized often only have *semantic* properties which have no inherent spatial analog from which to create a visual representation. Any spatial ordering used in display must be created as part of the visualization process. Viewed in this way, information visualization requires three components: data organization, visual spatial representations of the data organization, and display and interaction mechanisms.

This paper reports on the application of the Document Explorer system [3] to visualization of the WWW. Other visualization systems have provided mechanisms for visualization of WWW documents based on HTML link structure. The approach we present bases the visualization and browsing mechanisms not on link structure, but on the semantic content of documents. This approach can serve as an adjunct to text and link based search by supplying a visual search environment based on semantic associations among WWW documents.

3 Visualizing and Navigating WWW Semantic Content with Document Explorer

In order to visualize a semantic space of WWW documents the Document Explorer provides mechanisms for each of the three components of information visualization listed above. Figure 1 below shows some of the components of the system.

3.1 Data organization: Extracting and Organizing WWW Semantic Content

Recently, attention has been directed toward search of the WWW and a number of content extraction tools have been developed [4]. One such tool is the Harvest system [1]. In this system indexing information is made available and can be used to perform searches on documents' keyword indexing. Additionally, Harvest can serve as a general purpose tool for filtering or connecting indexing information to other tools by making keyword lists and location information available.

The Document Explorer system operates on keyword lists to determine associations among documents using a co-occurence metric to derive similarity measures among documents. The associative networks used in the system are Pathfinder networks (PFNETs) [2]. In constructing a PFNET two parameters are incorporated: r determines path weight according to the Minkowski r-metric and q specifies the maximum number of edges considered in finding a minimum cost path between entities. As either parameter is manipulated, edges in a less complex network form a subset of the edges in a more complex network. Thus, the algorithm generates two orthogonal families of networks, controlled by r and q. The least complex network is obtained with $r = \infty$ and q = n-1, where n is the total number of nodes in the network and is used in the system.

The difficulty of abstraction, or reduction of the complexity, of networks is a central challenge to many efforts in visualizing the WWW. One approach is to transform the graph representing connectivity among document to a hierarchy by removing links to form a tree. Visual representations of hierarchies are relatively well developed compared to the more general problem of representing directed graphs. Mukherjea et al. [5] use this approach for WWW documents augmented with the ability to dynamically reform hierarchies based on user needs. The approach used in the Document Explorer is similar in the identification of nodes representing single documents to serve as distin-

guished points for display and navigation. Nodes with high degree are identified for use in display and navigation in a fashion similar to the identification of cluster centroids in single link clustering.

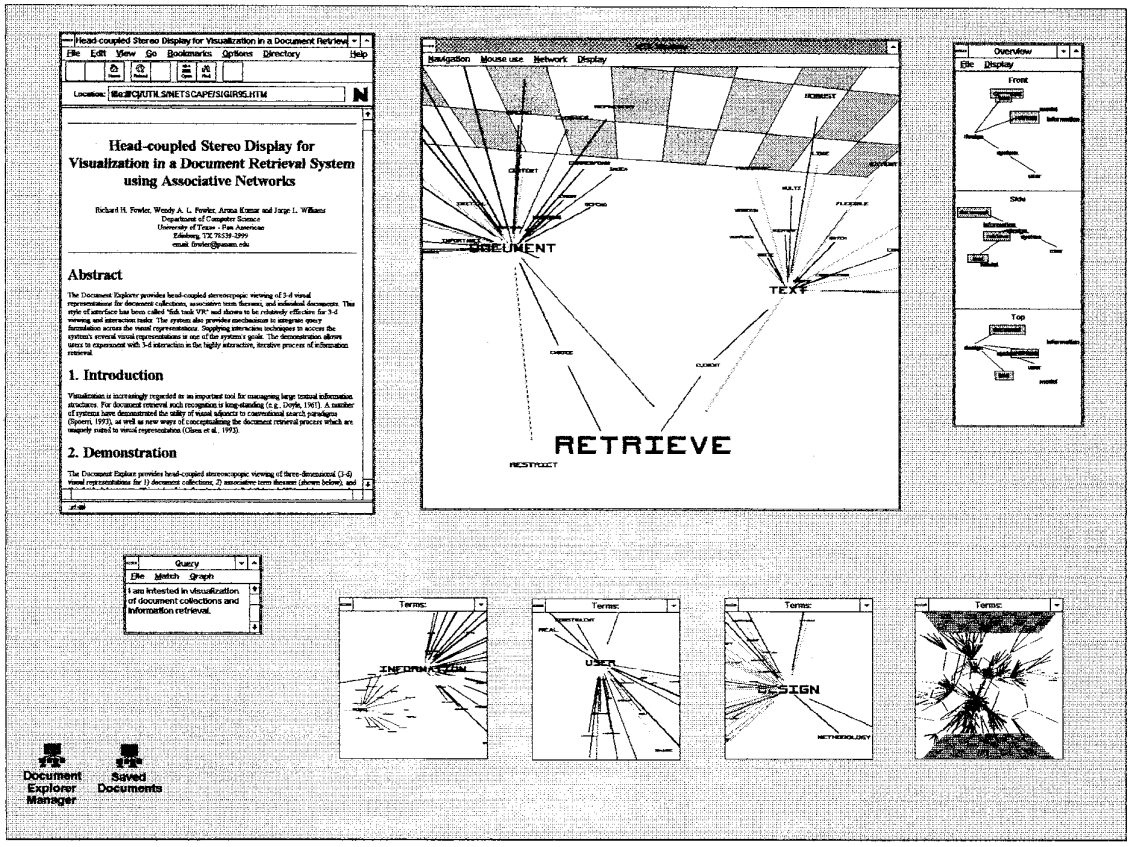

Figure 1: Document Explorer used with a web browser. The HTML viewer is the upper left window. The Document Explorer view of an associative network of documents is visible in the middle of the screen with overview diagrams to the right. In this view documents are labeled by content. Below the HTML viewer is the user's natural language query used to find an entry point for browsing in the document network. Lower windows show visual bookmarks the user has set. The rightmost is a view of the document set's global structure.

3.2 Spatial visual representations of the semantic data organization

The PFNET of WWW documents serves as the basis for the browsing and retrieval tools available in the system. To create a spatial representation a graph layout algorithm based on a spring metaphor to position nodes of the document network in three dimensional space is used. Nodes are considered as connectors and the length and strength of springs among connectors is derived from PFNET path distances. Nodes are positioned at the points which minimize total energy in the system of springs as nodes are allowed to vary in three dimensions. By manipulating parameters for spring length and strength it is possible to create layouts which are useful for user interaction and which visually reveal the clustering and

connectivity among documents, such as shown at the bottom right of Figure 1.

3.3 Display and interaction mechanisms

The display and interaction mechanisms provided in the Document Explorer focus on providing orientation and overview of the global structure of document associations together with navigation and retrieval tools for exploring local detail. One orientation technique is the overview diagram constructed using the nodes of highest degree in the network which tracks the view volume of the detailed view.

Several viewing tools are available to the user. Standard interaction techniques using six degree of freedom input are used to move through the network's three-dimensional space. Nodes in view at a particular time can be determined both by the standard viewing projection of the three-dimensional space and user-controlled tools that vary display density based on network structure. For example, the user can use fisheye viewing techniques to control density by specifying threshold values for the degree of interest or path distance from the focus node. Most simply, the user can select a node which represents a single document or a cluster centroid and expand or collapse connected nodes.

To facilitate orientation and navigation in the space tools are provided for fluid navigation in the space, such as zooming to documents or document clusters which are selected manually, or through the user's entry of a natural language question which is used to select a document cluster to serve as a starting point for user directed exploration. Link and node coloring are used to indicate the parts of the network which have been viewed or are of high similarity to the user's query.

4 Acknowledgments

Thanks to Xiannong Meng for work with the Harvest system and Zack Parry for work on the Sun implementation of the Document Explorer.

5 Conclusion

The Document Explorer provides a suite of visualization tools for aiding users in interactively navigating the WWW using semantic relationships among documents. This form of access can supply a useful adjunct to other forms of information access. In particular the system's visual representation of global semantic structures allows the user to maintain a view of high level structure while seamlessly exploring detailed information.

6 References

[1] Bowman, C. M et al. (1994). The Harvest information and discovery system. *Proceedings of the Second International WWW Conference, WWW '94*, 763-771.

[2] Dearholt, D. W. & Schvaneveldt, R. W. (1990). Properties of Pathfinder networks. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization*, 1-30. Norwood, NJ: Ablex.

[3] Fowler, R. H. Fowler, W. A. L., Kumar, A. & Williams, J. L. (1995, July). *Head-coupled stereo display for visualization in a document retrieval system using associative networks.* Demonstration presented at the ACM SIGIR Conference, Seattle, WA. (abstract in *Proceedings,* 360*), paper at http://www.cs.panam.edu /info-viz*

[4] Fox, E., Akscyn, R. M., Furuta, R. K., & Leggett, J. J. (1995). Digital Libraries, Communications of the ACM, *38(4).*

[5] Mukherjea, S. Foley, J. D. and Hudson, S. (1995). Visualizing complex hypermedia networks through multiple hierarchical views. *CHI '95*, 331-337.

[6] Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Addison-Wesley: New York.