Abstract: Visualization is a promising technique for both enhancing users' perception of structure in the Internet and providing navigation facilities for its large information spaces. This paper describes an application of the Document Explorer to the visualization of WWW content structure. The system provides visualization, browsing, and query formulation mechanisms based on documents' semantic content. These mechanisms complement text and link based search by supplying a visual search and query formulation environment using semantic associations among documents. The user can view and interact with visual representations of WWW document relations to traverse this derived document space. The relationships among individual keywords in the documents are also represented visually to support query formulation by direct manipulation of content words in the document set. A suite of navigation and orientation tools is provided which focuses on orientation and navigation using the visual representations of document set and term collection.

1. Introduction

The explosive growth of the Internet has added to the need to filter and organize information so that users can efficiently and effectively identify relevant documents. Browsers such as Mosaic [Schatz & Hardin, 1994] for the World Wide Web and Harmony [Andrews & Kappe, 1994] for Hyper-G allow users to traverse the information space of documents through connections provided by document authors. Though this may account for much of the popularity of the Internet, it hides many of the difficulties of effective information access. Searching Internet resources is difficult due to size, diversity of data, and lack of common indexing scheme. The most widely used search tools are automatically generated search systems like Lycos and Harvest and manually organized systems like Yahoo and Internet Yellow pages. Though such tools are valuable assets to the Internet searcher, it seems likely that these tools alone will not solve the current problems of information access.

The challenges of information access common on the Internet are issues common to all forms of information retrieval. These issues include difficulties in using indexing vocabularies, indexing indeterminacy, and the user's inability to completely specify information needs [Ingwerson & Wormell, 1986]. Retrieving information that meets users' information needs is an iterative process, and techniques which explicitly incorporate users' judgments, such as relevance feedback [Maron & Kuhn, 1960], provide means to automate some aspects of user guided retrieval. It is also clear that mechanisms providing alternative paths of access to information can enhance retrieval effectiveness [Bates, 1986].

One promising approach for enhancing information retrieval through the Internet is visualization to facilitate users’ perception of document relation structure. A number of systems have been developed to provide visually based browsing mechanisms for traversing the link structure of Internet documents. McCahill and Erickson [McCahill & Erickson, 1995] describe designs for three-dimensional spatial interfaces for Internet Gopher which employ icons with various shapes and textures to represent documents arranged by file structure or search results. The Harmony browser for Hyper-G [Andrews, 1995] provides two-dimensional structure maps of a document’s link neighborhood, as well the three-dimensional Information Landscape visualization of link structure hierarchy. The Narcissus system [Hendley et al., 1995] creates a three-dimensional visual representation of WWW link connectivity with documents’ positions determined by user manipulable attractive and repulsive forces among document nodes. Munzner and Burchard [Munzner & Burchard, 1996] describe a system which displays WWW documents positioned in a hyperbolic space to increase display density compared to conventional display spaces.

Visualizations of document connectivity based on existing, author-created links provide an important component in facilitating orientation and navigation in the large WWW information space. However, for users it is typically docu-
ments’ content relationships which are of most interest, rather than link based relationships, and such semantic relationships are only partially reflected by link structure. Documents with similar content can be found throughout the WWW information space. Though it is likely that a user would find documents dealing with similar topics close together within a site, it is also likely that other documents dealing with similar topics are at other sites, and those related documents may be buried deep within the link structure of a site.

Systems which provide facilities for organizing and viewing document organizations based on semantic content are less common than systems which focus on visualizing document link structure. Mukherjea, Foley, and Hudson [Mukherjea et al., 1995] describe a system which operates on the semantic content of WWW documents to form visually displayed hierarchies. In this system users can specify attributes which affect the organization of documents derived by the system and reconfigure the display to suit their needs. VR-VIBE [Benford et al., 1995] is a system which fixes WWW documents at locations in three-dimensional space by allowing users to interactively position keywords on a pyramid. Document locations are determined by calculating the distance for each document from keywords’ locations on the pyramid. In this way users create visually distinct clusters of documents for further inspection. LyberWorld’s Relevance Sphere [Hemmje et al., 1994] also uses this general approach for forming three-dimensional document cluster displays. Gershon et al. [Gershon et al., 1995] describe a system which allows users to view an organization of documents visited as a hierarchy of links and also to construct a separate hierarchy based on their own needs. Additionally, the system creates a co-occurrence map of terms in documents visited to help users identify word patterns useful in a more fine-grained search.

This paper describes an application of the Document Explorer system to the visualization of WWW content structure. The system bases the visualization, browsing, and query formulation mechanisms on the semantic content of documents. These mechanisms complement text and link based search by supplying a visual search environment using semantic associations among WWW documents. The system operates on sets of WWW documents to construct a visually displayed network of documents based on content similarity. The user can view and interact with the visual representations to traverse this derived document space. The relationships among individual keywords in the documents are also represented visually to support query formulation by direct manipulation of content words from the document set. A suite of navigation and orientation tools is provided which focuses on orientation and navigation in the visual representations of document set and term collection.

2. Visualizing WWW Semantic Content

Information visualization shares with scientific visualization the goal of providing understanding and insight using visual representations, yet is often faced with challenges not encountered in scientific visualization. For scientific visualization the visual representation is derived from a physical phenomenon, hence, there typically exists a natural visual representation based on spatial, temporal, or other properties of the phenomenon. Semantic relationships have no physical component which might naturally supply a basis for visualization [Gershon, 1994]. Rather, the elements to be visualized typically have only semantic properties with no inherently spatial analog from which to create a visual representation. The spatial ordering for semantic relations must be created as part of the information visualization process. From this perspective, visualization of semantic information spaces must provide: 1) spatialization of the abstract data, which may entail both data organization and derivation of a visual spatial representation of the data, 2) presentation of the spatial representation in a display space for user interaction and viewing, such as Euclidean two- or three-dimensional space, and 3) techniques and tools for user interaction with the visual representation.

The Document Explorer provides mechanisms for each of these three components of information visualization to display a semantic space of WWW documents and keywords. WWW document relationships and keyword relationships are organized as associative networks based on content similarity among documents and frequency of co-occurrence of keywords. The networks are presented in three-dimensional display space with network nodes positioned using a layout algorithm based on a mechanical spring metaphor. A variety of interaction and display tools are used to supply orientation cues and navigation mechanisms to assist users in browsing the document and keyword networks, as well as in the formulation of queries by direct manipulation of system objects.

2.1 Extracting and Organizing WWW Semantic Content

The systems principal visualizations are networks displays based on documents’ keyword lists. The lists can be provided by automatic content extraction tools, such as Harvest [Bowman et al., 1994], or derived from documents re-
trieved by the system. Keyword lists for each document are used to determine the associations among documents and among terms using a co-occurrence metric to derive similarity measures among documents and among keywords. Both visualizations of document space and term space for WWW document sets are available to the user.

The content based network of WWW documents is shown in Figure 1 below as the main window of the screen display. In this view documents are labeled by their content. Alternatively, the display can present the HTML title. Below are overview diagrams showing the location of the detailed view in the main window in the complete network. To the left and right of the overview is a series of visual bookmarks set while browsing which can be used to return to a previous viewpoint. The leftmost shows a view of the complete network. Other navigation and orientation tools are also available, such as anchors and signposts, which the user can leave and revisit at points traversed in the network. Change of viewpoint using navigation aids is always done by zooming to new viewpoints to maintain fluid motion and attenuate disorientation. In the upper right of the screen is a natural language query which has been transformed to a user-manipulable query graph. The query can be used to supply an entry point in the document network or be used for conventional weighted vector search to provide a list of documents from which to initiate browsing.

![Figure 1: Document Explorer screen. The content based document network is shown in the main window with overview diagrams below it. Windows to the left and right of the overview show visual bookmarks the user has set. At the upper right is the user’s natural language query and its visual representation.](image)

The representations underlying the system’s network displays are minimum cost networks derived from measures of term and document associations. The network of documents is based on interdocument similarity, as measured by co-occurrence of keywords between document pairs. For the network of terms, or associative term thesaurus, and the visual representation of the user’s query the associations are derived from text with association measured by keyword co-
occurrence and lexical distance within documents. The networks used in the system are Pathfinder networks (PfNets) [Dearholt & Schvaneveldt, 1990], which can be conceptualized as path length limited minimum cost networks. Algorithms to derive minimum cost spanning trees have only the constraints that the network is connected and cost, as measured by the sum of link weights, is a minimum. For PfNets, an additional constraint is added: Not only must the graph be connected and minimum cost, but also that the longest path length, as measured by number of links, is less than some criterion. To derive a PfNet the direct distances between each pair of nodes are compared with indirect distances, and a direct link between two nodes is included in the PfNet unless the data contain a shorter path satisfying the constraint of the maximum path length.

Reducing the complexity of network representations is a central objective in many efforts in visualizing Internet information structures. Complexity can be reduced by transforming the network of connectivity among documents to a hierarchy by removing links. Compared to the more general problem of representing directed graphs, visualization of hierarchies is relatively well developed. This sort of transformation is used by Mukherjea et al. [Mukherjea et al., 1995] for WWW documents. In the Document Explorer nodes in the complete network are identified for display and navigation in a fashion similar to the identification of cluster centroids in single link clustering. These document nodes are continuously displayed in the overview diagram and the three-dimensional space within which the user navigates. The location of the subnetwork displayed in the main viewing window is updated continuously in the overview diagrams as the user moves through the three-dimensional network display space.

Additionally, the system uses networks of differing densities to provide separate views of document and term interrelationships. The least dense network, which allows paths of any length in satisfying the minimum cost criterion, is useful for global navigation and orientation. Conceptually, it shows the strongest relations among elements. As such, it is effective in supplying views for global navigation and structure perception. To provide a more detailed view of relationships the system also maintains a PfNet of elements in which the maximum path length is relatively small, creating a complementary network with many more links. This more dense network display is most beneficial when viewing a small set of elements and serves as a sort of magnifying glass for revealing relations among elements not shown in the sparse network.

2.2 Spatial representation of the system’s networks

The spatial representation of the system’s networks are designed to visually reveal structure. Network nodes are positioned in three dimensions using a graph layout algorithm based on a spring metaphor [Kumar & Fowler, 1994] similar to Kamada and Kawai two dimensional network layout algorithm [Kamada & Kawai, 1989]. Nodes are treated as connectors and spring length and strength among connectors is derived from network link distances. Nodes are allowed to vary in three dimensions and iteratively positioned at the points which minimize energy in the system of springs. Varying spring length and strength allows layouts which are useful for user interaction and visually reveal clustering and connectivity among.

2.3 Display and interaction mechanisms

The size and density of the Internet document network requires network viewing and navigation tools that allow users to perceive the overall structure of document relations, explore smaller regions in detail, and select and view individual documents. Display and interaction mechanisms in the Document Explorer supply orientation and overview of the global structure of document associations, together with navigation and retrieval tools for exploring local detail. An overview diagram of the complete network is constructed using the nodes of highest degree in the network. As the user changes viewpoint in the main viewing window, and thus the portion of the network which is viewed in detail, the overview diagram tracks the overview nodes which are visible in the view volume of the detailed view. This helps attenuate disorientation by providing the context for the individual network nodes which are in view. We have also explored stereoscopic display of the networks. For the sorts of network displays in the Document Explorer there is a dramatic increase in the abilities to perceive structure, as found in a number of recent studies of stereoscopic network viewing [Ware & Franck, 1996].

Other display mechanisms for changing the user’s views of the network are also designed to assist the user in remaining oriented to the overall structure while examining local detail, such as is the availability of two separate networks which differ in density, or the number of links. The most sparse is essentially a tree, thus having the fewest links necessary to have all nodes connected. It is this sparse network which is used to provide the spring analogs used to position the nodes in 3 space. The arrangement of nodes on this basis in practice supplies a characteristic clustering of nodes in which structure is relatively easy to perceive, such as shown in the leftmost bookmark of Figure 1. Yet, much of the utili-
ty of browsing in the derived semantic space comes from exploring relations which are relatively weak. These relations are captured in the more dense network. The user typically identifies a single document or document region by browsing traversing links from a known document in the sparse network, and then selects to have the weaker links of a node or node set become visible. These are identified by a differing color and the point at which the selection is made is identified by a marker or signpost in the network which is relatively large and can be easily zoomed back to during the course of exploration. Most simply, the user can select a node which represents a single document or term and expand or collapse connected nodes. Again, color is used to mark the course of browsing and network exploration.

Several navigational tools are available to the user providing various mechanisms to return to previous viewing points, or provide information about the context of the current position. Visual bookmarks can be set by the user at any point while moving through the network. Simply selecting the bookmark returns the user to the viewpoint at that time the bookmark was set. As mentioned above, visual anchors can be set which remain visible from long viewing distances and can be returned to at later times. Other display mechanisms include link and node colorings indicating parts of the networks of high similarity to the user’s query and user selected display of navigational signposts pointing to the closest overview nodes or bookmarks. All navigation maintains fluid movement in the space, always zooming, rather than jumping, to new viewpoints.

3. Query Formulation

Although the system’s most novel features center on its visual representations to support browsing and structure perception, it is useful for information retrieval systems to supply multiple paths of access to information items. For the type of minimum cost networks used in this system, there is close relation to various clustering algorithms. Using the visual document space representation to browse documents can be characterized as a form of user directed, cluster based search. The system also supplies conventional vector space retrieval as an adjunct, supported by direct manipulation techniques for forming queries.

In a typical use of the system the user enters a natural language statement of information need to begin the browsing and retrieval process. The system converts the natural language to a weighted vector representation and uses conventional weighted vector matching to form a sequence of documents matching a vector representation of the query. The user can then select a document from this list to serve as the entry point in the network of documents, i.e., the viewpoint is positioned near that document and the document is distinguished by color.

The availability of an association map of keyword relations, a term space, assists the user in formulating queries. The relations among terms as portrayed in an association map are typically quite different than found in a conventional thesaurus and serve as an alternative ordering of term relations. To support query formulation the system provides direct manipulation facilities for constructing queries. A visual, manipulable representation of the user’s query is displayed together with the natural language from which it was originally derived. This visual representation can be manipulated by the user by moving terms from the term network into the query window and connecting them to the query graph. The vector space retrieval is based on the query graph and is performed by converting it to a weighted vector representation. This functionality could certainly be provided by other interaction mechanisms, yet the visual graph manipulation technique encourages active interaction generally, and allows the use of network representations throughout the system.

4. Acknowledgments

Work on this project has been supported by NASA grants NAG9-551 and NAG9-842.

5. Conclusion

The Document Explorer supplies visualization, browsing, and query formulation mechanisms based on the semantic content of WWW documents. Users can view and interact with a visually displayed network of documents based on content similarity in a WWW information space that is an alternative to link based representations. Relationships among individual keywords in the documents are also displayed visually to support query formulation by direct manipulation and convey information about the keyword set. Navigation and orientation tools facilitate interaction and enhance perception of the document set and term collection structures through visual representations which facilitate the exploration of local
detail while remaining oriented to the global context. These facilities can serve as useful additions to textual representations and visual representations based on document link structure.

6. References


