Case Study
Visualizing Internet Resources

Nahum Gershon[1], Joshua LeVasseur[2], Joel Winstead[3],
James Croall[4], Ari Pernick[5], and William Ruh[6]

The MITRE Corporation
7525 Colshire Drive
McLean, VA 22102

[1]gershon@mitre.org, [2]jlevasse@mitre.org, [3]winstead@mitre.org, [4]jcroall@mitre.org, [5]apernick@mitre.org,
[6]war@mitre.org

Abstract

The goal of this work is to improve the ability of people from all walks of life and interests to access, search, and use the information distributed in Internet resources. The process of interacting with information resources starts with browsing, continues with digesting and assimilating pieces of information, terminates with generation of new information, and begins anew with analysis of pre-existing and new information. Our approach is user-centric- taking users' needs into account by allowing them to interact with the information contained in large arrays of documents. The visualization process is an integral part of the overall process.

We have covered three related categories in this methodology. The first one is browsing through the World-Wide Web (WWW) hyperspace without becoming lost, based on a visual representation of the hyperspace hierarchical structure (hyperspace view). The second category is overcoming the rigidity of the WWW by allowing the user to construct interactively and visually a personal hyperspace of information, linking the documents according to the application or problem domain, or to the user's own perception, experience, culture, or way of thinking. The third category includes discovery and analysis of new information and relationships in retrieved documents by aggregating relevant information and representing it visually.

1. Introduction

Linking numerous computers dispersed around the world through the Internet has created an exciting universe of information. This revolutionary capability has enabled us to explore this universe of distributed information resources from our desktop computers. If set up, managed, and presented correctly, these distributed information resources will enable us to:

- Find and obtain useful information from distributed information resources (*e.g.*, libraries) as efficiently as possible, and
- Enhance the creative thinking of the user to better solve his/her current and future problems.

Ordinarily, users would like to interact with distributed resources to find out:

- A topic of interest (it is not always clear to the user in the beginning of a search);
- Where the interesting information is located (e.g., in which documents);
- Where the relevant pieces of information reside in the documents. This step involves extracting and assimilating interesting pieces of information and creating new information from them; quite often, it is not a simple process.

In spite of the engaging nature of the Internet systems, we still have a long way to go before the use of this information universe is easy and intuitive. While interacting with information over the Internet, users feel at times lost, confused, and overwhelmed (justifiably so). To find required information or to browse through information, users currently need to conduct frustrating searches through arrays of (at times) debilitating menus and "belligerent" computer systems. Some of the remote sources are massive and once the user has the information, he or she needs to browse through large amounts of text, tables, and images. How can the user know where the sources of the relevant information reside, how to get them, and, once the sources are accessed, how to get the relevant information from them?

WWW, the World Wide Web or simply the Web, developed at CERN, Switzerland [T. Berners-Lee et al, 1994], and Hyper-G, developed at Graz University of Technology [Andrews and Kappe, 1994], allow the user

0-8186-7201-3/95 $04.00 © 1995 IEEE

See Color Plates, page 151.

to roam via menus and embedded links through information spaces of documents or images. Engaging browsers, such as the National Center for Supercomputing Applications (NCSA) Mosaic (for the WWW) [Schatz and Hardin, 1994] and also Harmony (for Hyper-G) have transformed the process of getting information from Internet servers. However, some major difficulties still remain.

To a large extent, users are not directly involved in the development of the Internet and its capabilities. If we do not involve the users in designing distributed information resources and their interfaces, we will create useless information systems. As long as there is a human being sitting in the front of the screen, the interface to Interenet information resources needs to be user-centric, taking the user needs into account. Users would like to interact with the information, preferably forgetting that there is a computer separating them from the information. A good human-computer interface (HCI) is a must, but it is not enough.

Recent developments in visualization, interactive computer graphics, and mass storage have created new possibilities for information navigation, access, and retrieval in which visualization and user interface (UI) could play a central role. The question is how to exploit the advances in visualization and graphics technologies and experience while understanding how the human mind works in order to reduce the frustration, time, and cost of using information dispersed over the Internet.

The work described in this paper is concerned with improving the way users interact (visually and non-visually) with information embedded in distributed information resources. The process of interaction with the information starts with browsing, continues with digesting and assimilating pieces of information, terminates with generation of new information, and begins anew with analysis of pre-existing and new information. The methods developed are for browsing through hyperspace without becoming lost, overcoming the rigidity of the WWW, and aggregating and classifying relevant information. Moreover, in this work, the visualization process does not stand alone, but rather it is an integral part of the process of interacting with the information.

2. Hyperspace View: Browsing Through Hyperspace without Getting Lost

Using current browsers to "surf" over the Internet, users often traverse a multitude of documents via hyperlinks. After opening a number of linked documents on the WWW, users often do not know where they are in the information space and do not remember how they got there. In short, users feel lost or become disoriented. This is a major problem. One solution is to provide users with both a local and a global view of the information space. These views should be represented visually to promote quick perception and understanding of the hierarchical structure of the hyperspace and to help the users quickly locate where they are in hyperspace, i.e., to re-orient themselves.

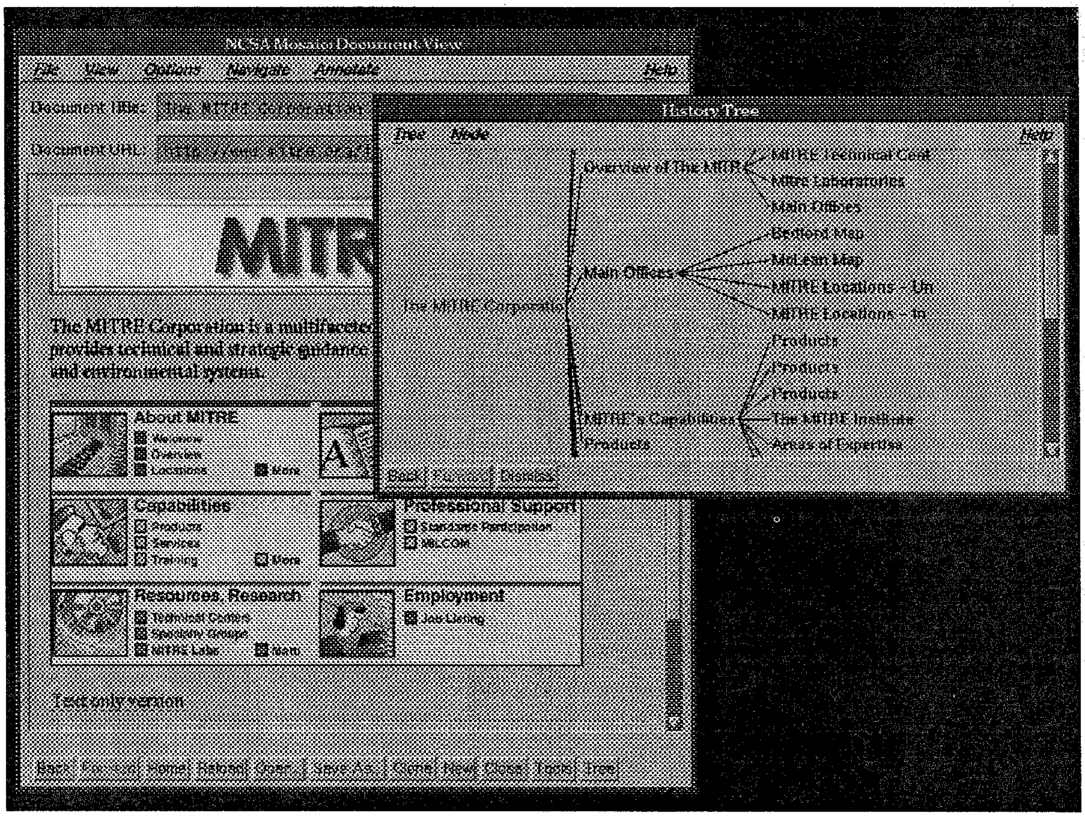
We developed an enhancement of NCSA Mosaic that allows the user to view the hyperspace depicted as a visual "tree" structure (see Figure 1). In addition to viewing, users can "jump" from one document to another by pointing and clicking the mouse without having to go back resource by resource or "page by page." Recently, two additional approaches for visualizing hyperspace structures were independently proposed [Mukherjea and Foley, 1995, and Wood, *et al*, 1995]. Both interesting approaches focus on how to make complicated hyperspace structure more comprehensible by letting the user view the hierarchy globally and in detail. The latter allows restructuring the view by introducing physical repulsive and attractive forces among the hierarchy elements.

While browsing, users often would like to view the names of documents and how they are linked to each other without actually opening and reading each document. Our enhancement allows the user to let the hyperspace view grow automatically up to a specified number of levels. After visually observing the hyperspace structure and contents, the user could decide to open and read none, some, or all the documents represented on the "tree" or to save them in his/her own personal space. In cases where changes to the content of a document is an important piece of information, saving Web documents allows the user to compare old and new versions.

3. Making the WWW Non-Rigid-Building One's Own Information Hyperspace

The rigid organization of hypetext documents often makes it difficult to use information distributed over the World-Wide Web. Creators of documents placed on the Web link them to other documents. These links form a rigid structure where no changes are allowed. Quite often, these hyperlinks reflect the document creator's own point of view and current interests, or some arbitrary considerations.

Depending on the application, problem, personal way of thinking and perception, experience, or culture, these pieces of information could be related to each other in various ways. For information resources to be effective and to enhance problem solving and analysis, they should allow each user to construct his or her own information space with links and associations (among pieces of information and whole documents and images) that fit the personal problem, application, or ways of thinking and perception.



The MITRE Corporation is a multifacete
provides technical and strategic guidance
and environmental systems.

Figure 1. Hyperspace View: A graphical view of the hyperspace emerging from a document depicted as a "tree" structure. The user can "jump" from one document to another by pointing and clicking the mouse without having to go back one by one.

Our enhanced version of NCSA Mosaic enables the user to modify visually and interactively the links among the documents and images using a point-click-and-drag operation on the display of the hyperlink hierarchical structure. This enhancement allows the user to effectively generate new, personalized links and to (visually) view the new and "old" information space globally and locally (see Figure 2). The new hyperlinks are stored at the end of duplicates of the corresponding documents and could be saved for future viewing and sharing with other users. The user can add annotations accompanying the new links.

Another problematic aspect of current distributed information systems such as the World-Wide Web is that the smallest unit of information is a document (or HTML "page"). Users often find that a paragraph, a sentence, a word, or even a part of an image is a piece of information relevant to their problem. Our enhanced version of NCSA Mosaic allows the user to define new documents that contain fragments of existing documents and link them to other documents as he or she wishes. With these enhancements, the user could create his or her own version of the information space, thus reflecting the current problem or his or her own interests and view of the information.

4. Finding New Information in Retrieved Documents- Aggregating Relevant Information

Once the user has retrieved the documents that are related to his or her problem, he or she often needs to analyze the different pieces of information and find out new information by doing so. In many cases, the amount of retrieved information is very large, which makes access (getting the relevant pieces of information) and analysis difficult. We have developed a method and tool that could assist the user in

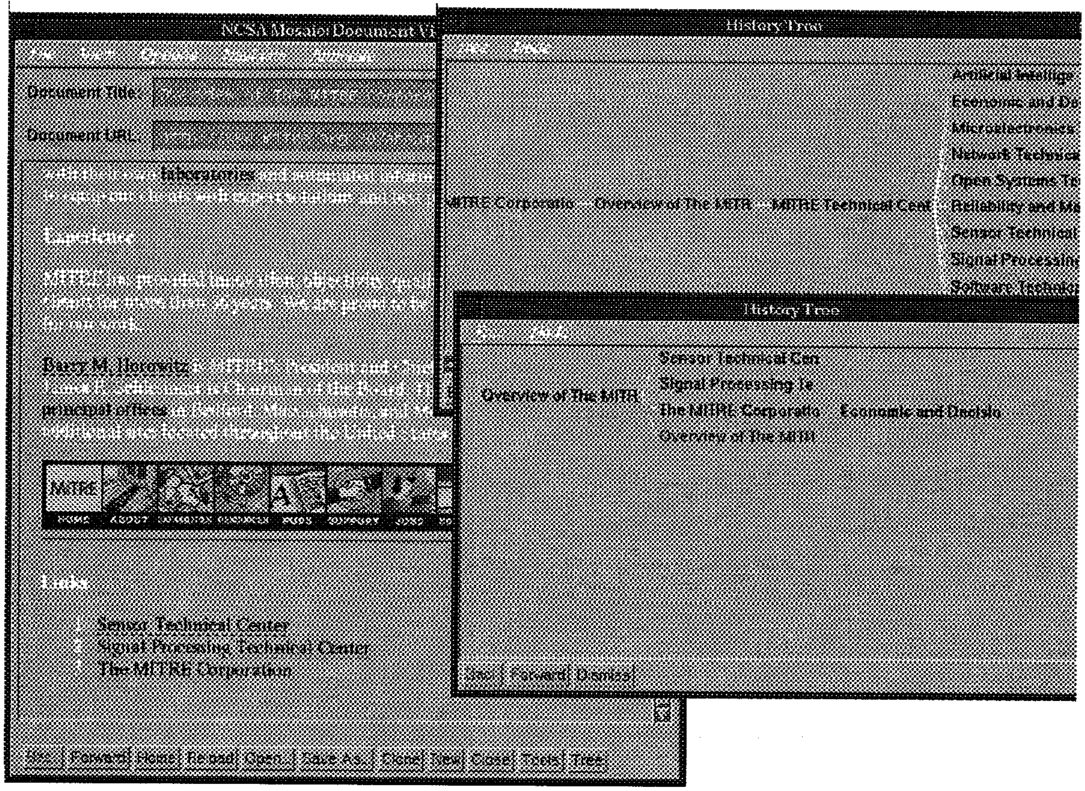


Figure 2. Making Hyperspace Flexible: The MITRE enhancement to NCSA Mosaic enables the user to interactively modify the links among the documents and images using a point-click-and-drag operation on the display of the hyperlink hierarchical structure. The user can generate new, personalized links and to view the new and "old" information space globally and locally. The new hyperlinks are stored at the end of the documents and could be saved for future viewing and sharing with other users.

performing these tasks based on the concept of aggregating relevant information.

We illustrate this technique of aggregation to create new information with word correlation. Documents are automatically analyzed and the proximity of any pair of words is calculated. Words are considered to be correlated if they are separated by no more than a (user-defined) number of other words.

The correlations among words in a document could be represented by a table where the rows and the columns are the list of words and each cell of the table contains the number of positive correlations in the document between two words in a pair. Such a table could be very large and the high correlations are generally scattered throughout all over the table.

Users find it very difficult to locate the highest correlations across a large table and to integrate and make sense of them. For example, the correlations of 4 words to 4 other words in a document (a subset of the full correlation table) could be represented by Table 1.

The information analyst wishing to identify the high hits of incidence needs to "fish" them individually from the various table cells. This is a slow process and could require the user to be very meticulous when the tables are large.

In scientific data, where the rows and columns are numerical coordinates, they are naturally ordered by increasing values of the coordinates. With words rather than numbers representing the rows and columns in tables of information, the order of the entities is arbitrary and thus one could permute them to yield the following representation depicted in Table 2.

	fat	smoking	snacks	sedentary
fatigue	22	3	47	2
aches	10	3	11	4
ailments	2	4	4	3
nausea	15	5	33	1

Table 1. Correlations of 4 words to 4 other words in a document.

	snacks	fat	smoking	sedentary
fatigue	47	22	3	2
nausea	33	15	5	1
aches	11	10	3	4
ailments	4	2	4	3

Table 2. Correlations of 4 words to 4 other words in a document after aggregation (the high incidences are aggregated in the upper left corner).

In this representation, the highest correlations are aggregated and it is easy to see which pairs of words are highly correlated. The analyst can then investigate why particular words are correlated and whether there is any significance in real life.

In small tables it is almost straightforward to permute the rows and columns and aggregate the high hits. However, in a large table of a few thousand rows and columns it is hard to do it manually. We have implemented a fast algorithm[1] that can aggregate information in large tables of n dimensions and the results are represented graphically on the computer display for visual inspection and analysis (humans still tend to perform pattern recognition faster and better than any existing computer). The user can then zoom on the interesting part of the aggregated table and observe the words that are highly correlated. The development of the visualization tools for representing unaggregated and aggregated correlation information was based on extensions of AVS (Advanced Visualization System). This tool is invoked from our enhanced Mosaic interface. Examples of a typical aggregated table generated from a document is given in Figure 3. This rather large table contains clusters of aggregated information. The users can now magnify these regions and analyze them visually. A magnified view of some of the clusters of aggregated information is given in Figure 4. From this figure, the information analyst can, for example, deduce from the color of the different regions that the words "invitation" and "duchess" (red) and also "sixteenth" and "dormouse" (yellow) are highly correlated and consquently, ask himself/herself why.

[1] The algorithm has been developed by Brian Dickens and further refined by Joel Winstead.

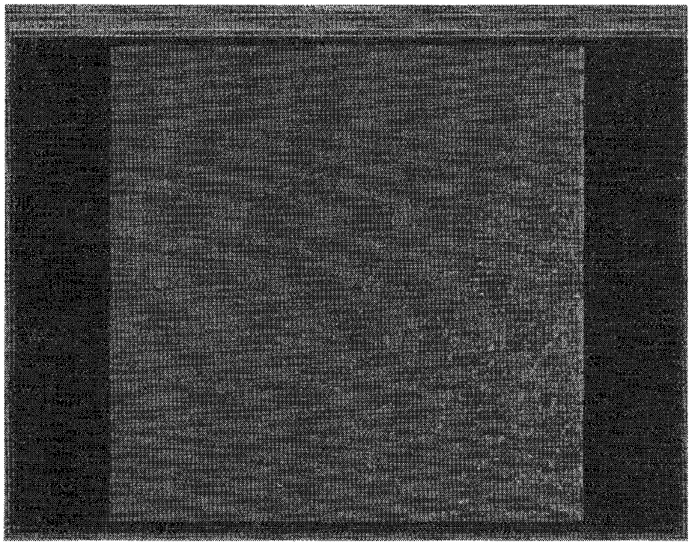


Figure 3. Information representation after aggregation. The correlation of words (excluding simple articles and propositions, such as "the," "by," and "and,"etc.) in a document was computed (by proximity) and the resulting table was aggregated. The axes are the string of words in the document.

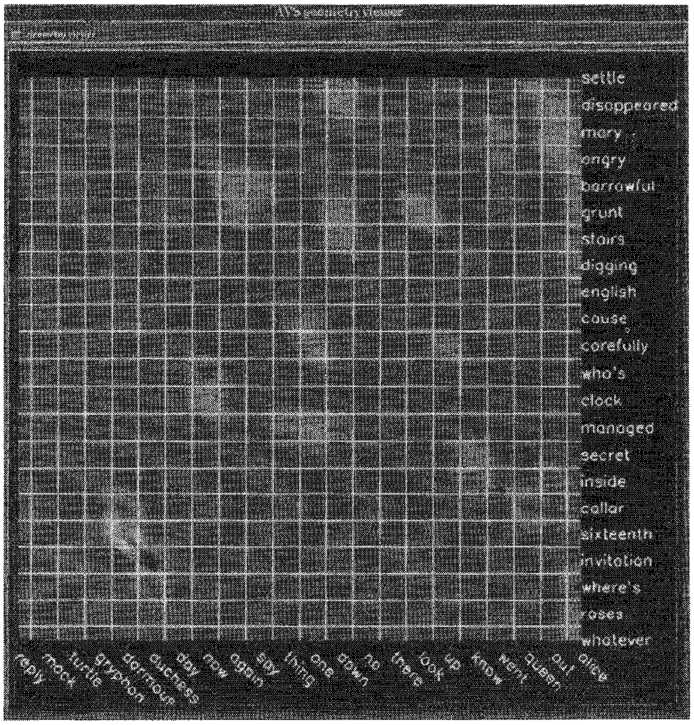


Figure 4. Aggregated information clusters magnified. The word coordinates are displayed allowing the user to find out what words are correlated in the analyzed document.

When the information is represented in 3-D, similar analysis could be done using volume visualization.

In Conclusion

The use of computers and networks as information resources is still a new concept involving new enabling technologies and permitting new types of interactions between humans and information resources (This is in contrast with traditional human inteactions with information contained in books). The situation has similarities with the one prevailing when the movie camera was invented. At that time, people tried to imitate theater via a stationary camera. They later discovered that these two media have different characteristics and were able then to better use the advantages of the new medium. Learning the new medium of computers and networks as information resources and further understanding how humans process and interact with information will enable us to provide new methods for discovery and searching of resources, enhancing creativity and thinking.

The developments reported in this paper make use of advances in visualization and interactive computer graphics technologies as well as the understanding of how humans search and process information. In this work, the visualization process does not stand alone, but rather is an integral part of the process of interacting with the information. Advances along these lines could improve the ability of people from all walks of life and interests to access, search, and use the information distributed in Internet resources. This will enable full use of the Internet's information universe from our desktops.

Acknowledgments

The authors would like to express their appreciation for the many exciting discussions with Brian Dickens and for his work on the information aggregation algorithm and for the creative advice on design and color given to us by Elaine Mullen.

References

Andrews, K. and F. Kappe, Soaring through Hyper Space: A Snapshot of Hyper-G and its Harmony Client, in *Proc. of Eurographics Symposium of multimedia/Hypermedia in Open Distributed Environments, Graz, Austria,* W. Hezner and F. Kappe, (editors), pages 181-191, Springer Verlag, June 1994.

Berners-Lee, T., R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret, The World-Wide Web, *Communications ACM*, **37**, No. 8, 76-82, 1994.

E.A. Fox (ed.), *Source Book on Digital Libraries*, Version 1.0, December 6, 1993.

Gershon, N.D., W.A. Ruh, J. LeVasseur, J. Winstead, and A. Kleiboemer, Searching and Discovery of Resources in Digital Libraries, in *Advances in Digital Libraries,* Nabil A. Adam, Bharat K. Bhargava, Milton Halem, and Yelena Yesha, Springer-Verlag, NY 1995, in press.

Mukherjea, S., and J.D. Foley, Visualizing the World-wide Web with the Navigational View Builder, *Computer Networks and ISDN Systems*, **27**, 1075-1087, 1995

Schatz, B.R, and J.B. Hardin, NCSA Mosaic and the World-Wide Web: Global Hypermedia Protocols for the Internet, *Science*, **265**, 895-901, 1994.

Wood, A., R. Beale, N. Drew, and R. Hendley, HyperSpace: A World-Wide Web Visualizer and its Implications for Collaborative Browsing and Software Agents, submitted to HCI '95, UK.