Visualizing World-Wide Web Search Engine Results

Sougata Mukherjea and Yoshinori Hara,
C&C Research Laboratories,
NEC USA Inc.,
San Jose, Ca, USA
E-mail: {sougata,hara}@ccrl.sj.nec.com

Abstract

Most of the popular WWW search engines show the documents that match the users' queries as pages of scrolled lists. If lots of information are retrieved this is not very user-friendly. Moreover, there is no mechanism to easily determine documents linked to the retrieved documents or keywords related to the query terms. This paper presents a system that allows the user to visualize various related information for the search results. We also introduce a focus+context visualization technique for the search space of WWW queries.

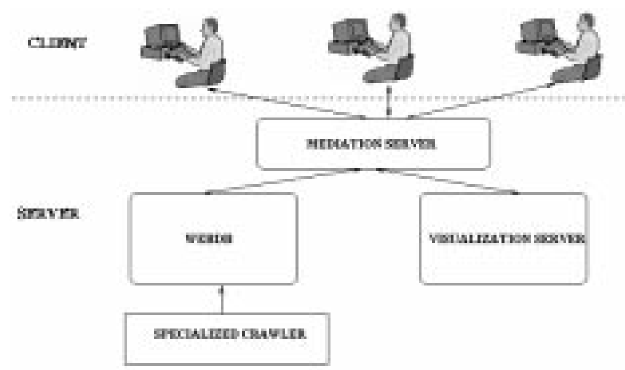


Figure 1. System Architecture

KEYWORDS: Query Result Visualization, World-Wide Web Search, Focus+Context Visualization.

1. Introduction

Although search engines are one of the best methods to retrieve information of interest from the World-Wide Web, most popular search engines only retrieve documents that match the user specified query terms. Sometimes the user may be interested in other related information. For example, they may be interested in the pages that are linked to the retrieved pages; they may also be interested in keywords that are related to the query terms.

Another problem with the search engines is that they show the results as pages of scrolled lists. With the explosive growth in the amount of information that is available on the World-Wide Web, many queries result in a large number of retrieved documents. It is very tedious for the users to go through pages of scrolled lists to find the relevant information.

This paper presents a system that allows the user to navigate through the result space of a World-Wide Web search to find the relevant information. The user can look at the documents connected by links to the retrieved documents. They can also look at keywords relevant to the query terms. Visualization is used to present the results because it allows people to use perceptual rather than cognitive reasoning in carrying out tasks [19].

This paper also introduces a visualization technique, **CardVis**, that is useful to show the search space to the user. The technique is based on the metaphor of a pack of playing cards and uses the *focus+context* technique to handle a large amount of information. For the focus or the section of the information space that is of interest to the user, we show the details. The 3rd dimension is used to show a large amount of contextual information.

The next section gives an overview of our system. Section 3 introduces the visualization technique by showing how it can be used to visualize World-Wide Web search engine results. Section 4 presents related work. Finally section 5 is the conclusion.

2. System Overview

Our system is based on client and server mediation architecture similar to [7]. A mediation server facilitates the interaction between users and various service providers. The me-

```
<query_result>
    <query_term>.....</query_term>
            ......
    <document>
       <url>...</url>
       <title>...</title>
       <linkin>...</linkin>
           ....
       <linkout>...</linkout>
           ....
      <keyword>....</keyword>
           ....
    </document>
       ....
    <semantically_similar_kwd>
      <keyword>.....</keyword>
      <document_url>...</document_url>
              .....
    </semantically_similar_kwd>
         .....
    <syntactically_related_kwd>
      <keyword>.....</keyword>
      <document_url>...</document_url>
         .....
    </syntactically_related__kwd>
         .....
</query_result>
```

Figure 2. Query Result Format for Visualization

diation server forwards users' requests to the corresponding servers based on users' preferences and functional requirements. An overview of the system architecture is illustrated in Figure 1. The components of the system are presented in this section.

2.1. Specialized Crawler

A specialized crawler is used to gather information from the World-Wide Web. Unlike traditional crawlers used by the search engines whose role is to gather as much information as possible, our crawler only gathers pages related to the users' interests. For example, we can gather all Web pages relevant to Visualization or all pages related to Computer Science conferences.

2.2. WebDB

One essential server is the WebDB hypermedia database system [11] since it provides indexing of the gathered information and supports comprehensive database-like query functionalities. WebDB has two levels: a *logical* level consisting modules for logical Web document modeling and storage, the query language translator, and an XML document generator; and a *physical* level consisting modules for internal class representations, object depository, query processing, and an internal query result class generator.

In WebDB, we view and model the Web as a labeled directed graph $G_{web} = (V_{web}, E_{web})$, where the vertices (V) denote the pages and the edges (E) denotes the hyperlinks between these pages. The vertices have information, including title, keywords, and other metadata. The edges are links from source pages to destination pages. We represent the document contents and Web structure using an object-oriented data model and implement *WebDB* on top of NEC PERCIO OODBMS.

When a document is passed to *WebDB* by the crawler, link and keyword information is extracted and maintained in the database. For each keyword in the databases, we perform keyword co-occurrence analysis to derive syntactically related keywords (e.g. car and Toyota) and consult an online lexical dictionary to derive semantically similar keywords (e.g. car and automobile). These relationships are maintained using pointers.

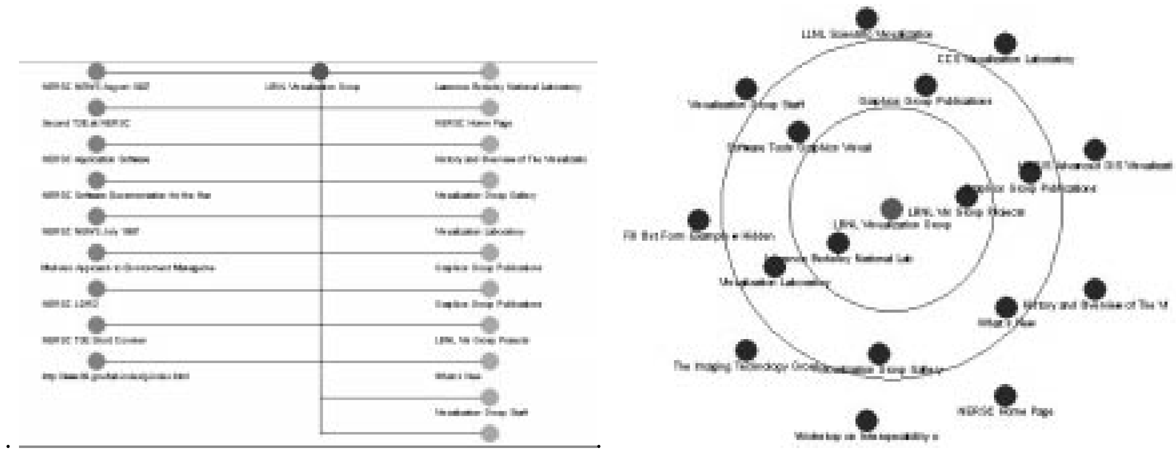
During querying, the system can traverse from the user specified keywords to related keywords and from the retrieved documents to linked documents easily. These information help the user to gain a better understanding of the search space. Instead of seeing the information in a series of HTML pages, the user can also visualize the data.

2.3. Visualization Engine

The visualization engine supports various languages for generating the visualizations. For example Java can be used and the resultant visualizations can be shown as an applet. We also allow the use of *Virtual Reality Markup Language (VRML)*, which can be shown in a VRML browser like SGI's CosmoPlayer.

To support various visualization functionalities, we define a protocol for data exchange between the visualization engine and *WebDB*. XML is used as the format in our implementation. The data exchange format is given in Figure 2. This allows the visualization engine to be integrated with other systems as long as they follow the same format.

With the augmented query results, the visualization engine has many options to display the query results based on users' preferences. For example, as shown in Figure 3, the user can visualize the neighborhood of any retrieved Web page. One of the retrieved pages for a specialized search engine for visualization was the homepage of the visualization group of the Lawrence Berkeley National Laboratory. The user can visualize the link neighborhood of (nodes linked to and from) the page as shown in Figure 3(a). The user can also determine pages that are semantically similar to this page.



(a) Link neighborhood

(b) Semantic neighborhood

Figure 3. Showing the neighborhood of a Web page (Url *http://www-vis.lbl.gov***)**

(Similarity is determined using the vector space model [18]). This semantic neighborhood can be shown in a **Bulls-Eye visualization** as shown in Figure 3(b). In this visualization the document of interest is at the center; the position of the other documents are determined by how similar they are to the main document. The next section introduces another visualization technique to show the entire search space.

3. Card Visualization

Sometimes the user is interested in getting an overview of the whole search space. The objective will be to show the structure of the retrieved Web pages as well as how these pages are relevant to the main keywords of the search space.

Considering the information space as a graph, we can first determine its *connected components*. Assuming that a link exists between two nodes if they are related, these connected components will have related nodes. A typical WWW search will result in many connected components. We have developed a visualization technique that is suitable for such an information space whose data organization is a group of disconnected graphs.

Figure 4 is an example of this visualization technique. It shows the results of a query to a database consisting of Computer Science conference Web pages indexed by WebDB. Each graph of the search space is shown in a card. One card is at the focus and shown at the top. The nodes and links comprising this graph are also shown. The nodes are shown as cubes and labeled by the HTML page titles. Moving the mouse over them shows their URLs at the status area of the VRML browser and clicking on them retrieves the corresponding URLs. The other graphs are shown in the context by arranging them in different planes in the 3rd dimension. The position in the 3rd dimension is determined by how relevant the nodes of the graph are to the original query terms. The titles of their main nodes are also shown. (The node with the maximum number of links is assumed to be the main node of a component).

The query terms are shown at the top as spheres. The other related keywords are also shown as cylinders. Their sizes indicate how closely they are related to the query terms. Initially one of the query terms as well as all the nodes which are relevant to the query term in the focus card are highlighted. The cards in the context are also highlighted if they have nodes relevant to the query term. Thus, in Figure 4 the query term *"web"* and nodes and graphs relevant to it are highlighted. The focus graph shows pages related with the WWW8 conference and the links between them.

The visualization allows two types of user interaction. (We use Javascript to program the interactions).

- The user can click on any of the related keywords. The nodes in the focus graph are highlighted if they are relevant; the other cards are also highlighted if they have relevant nodes.
- The user can click on any of the context cards to bring them into focus.

Thus in Figure 5 the user has clicked on the sphere for the keyword *database*; all the nodes and context cards relevant to the keyword were highlighted. The user has brought into focus a graph with pages related to the SIGMOD conference; note that a related conference (PODS) also belongs to this graph.

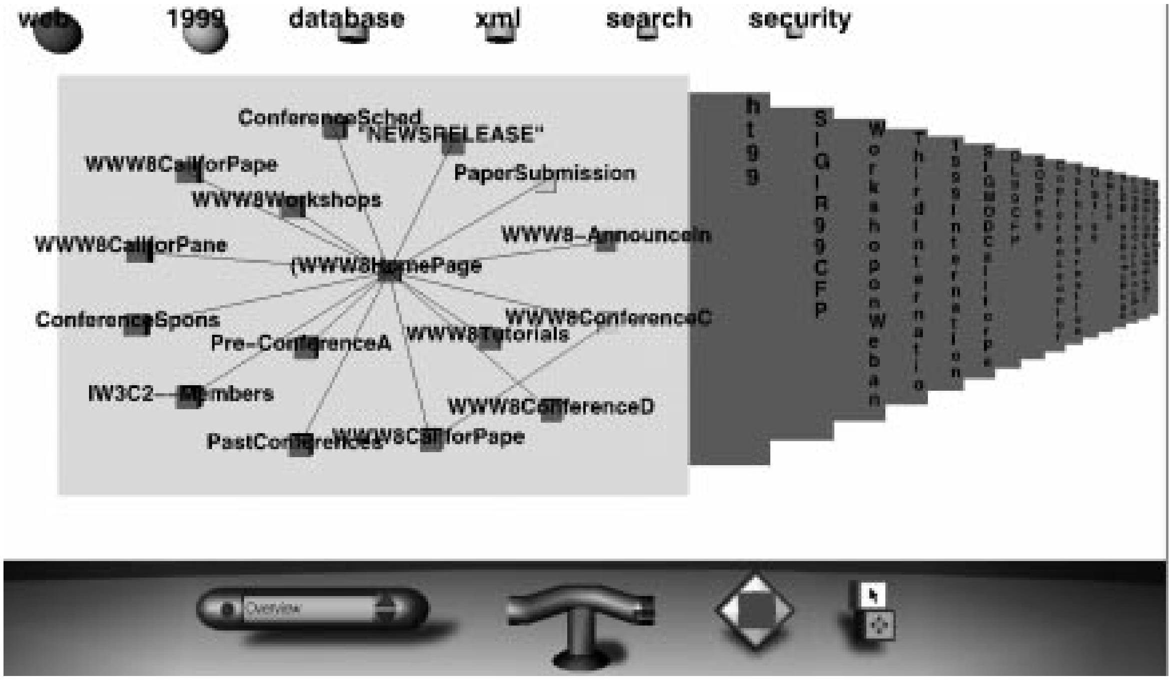


Figure 4. Visualizing the results of a query with the term “1999” and “web” to a WebDB database containing Computer Science conference Web pages. Initially, pages for the WWW8 conference is at the focus.

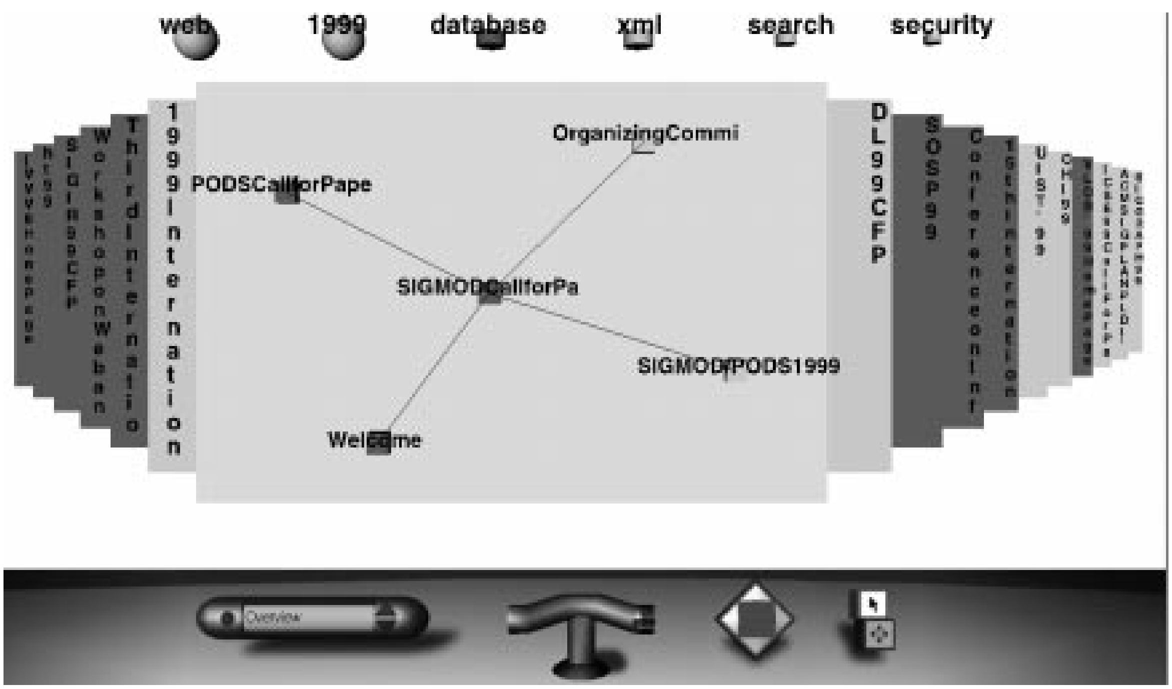


Figure 5. Visualization with *database* as the selected keyword. The user has also selected a database conference as the focus.

This visualization is based on the metaphor of how a player arranges a group of playing cards. One card is kept at the top and the other cards are arranged in such a manner that the main information (suit and color) of them are easily apparent. The user also has the ability of bringing any card to the top.

4. Related Work

4.1. Focus+Context Visualization

A popular approach for visualizing abstract information is to use focus+context techniques by which the information of interest to the user is shown in detail, smoothly integrated with other context information. By balancing local detail and global context, this technique can display information at multiple levels of abstraction simultaneously. An example of this technique is the fisheye-view concept [6]. The focus+context technique has been used to develop visualizations of different data structures. Examples include perspective walls [13] for visualizing linear data; document lens [17], a 3D visualization strategy for textual documents; table lens [16], a method for presenting tabular information and hyperbolic browser [10] for visualizing trees. A focus+context visualization for World-Wide Web nodes has also been developed [15]. In this paper we discuss a visualization technique that is suited for a collection of disconnected graphs.

4.2. Visualizing Web Sites

Several systems for visualizing WWW sites have been developed. Examples includes Navigational View Builder [14], Harmony Internet Browser [1], Narcissus [9] and WebCutter [12]. Since Web sites are large graphs and visualizing large graphs is difficult [3], these systems use various data simplification techniques like filtering and abstraction. The CardVis technique is not suited for visualizing large graphs. It is ideal for a collection of disconnected graphs and thus is a good technique to visualize Web search results.

4.3. Visualizing Search Results

Unique interfaces for viewing information retrieval results have been developed. Recent examples include SenseMaker [2] and Cat-a-Cone [8]. SenseMaker is an interface for information exploration across heterogeneous sources in a digital library. Cat-a-Cone is a 3D interface that integrates searching and browsing of very large category hierarchies with their associated text collections. Another interesting approach is the WebBook [4], which potentially allows the results of the search to be organized and manipulated in various ways in a 3D space. These systems show the usefulness of visualization for better understanding of the results of search engines.

Visualization techniques for World-Wide Web search results are also being developed. For example, the WebQuery system [5] visualizes the results of a search query along with all pages that link to or are linked to by any page in the original result set. On the other hand, Alta Vista *(http://www.altavista.com)* shows the keywords related to the original query terms. However, the World-Wide Web has two aspects: the contents of the Web pages as well as the links connecting different Web pages. Therefore, an effective visualization should show the user both the content and the structural information.

5. Conclusion

We have presented a system that provides specialized search on the World-Wide Web. It allows the user various techniques to visualize and navigate through the search results as well as other related information. We believe that the combination of visualization and searching will make it easier for the user to retrieve relevant information.

The paper also introduced a focus+context visualization technique called CardVis which is very suitable to visualize Web search results which can be organized as a collection of disconnected graphs. The visualization presents the details of the structure and contents of the focus graph. Moreover, it provides the user with an idea about the other graphs in the information space. The user can also bring any graph into focus. Therefore, we believe that the visualization is quite useful to gain an understanding of the search space.

Future work is planned along various direction:

- We believe that the CardVis technique can be generalized for visualizing information that can be clustered based on two criteria of similarity. Cards representing the clusters for one criteria of similarity can be arranged in different planes in the 3rd dimension. One of these cards will be in the focus and at the top. The details of the focus card will be shown to the user. To enable the user to get an understanding of the contents of the other clusters, their labels should be displayed. The labels of the clusters for the second criteria of similarity will be shown at the top. If one of these clusters is selected, all nodes in the focus card that belong to this cluster will be highlighted. The other cards will be also highlighted if they have relevant nodes.
- If the information space is really big, the visualizations may be too cluttered to be useful. Therefore, we are

exploring effective abstraction and filtering techniques.

- Usability studies are also essential to determine the actual usefulness of the system.

References

[1] K. Andrews. Visualizing Cyberspace: Information Visualization in the Harmony Internet Browser. In *Proceedings of the 1995 Information Visualization Symposium*, pages 97–104, Atlanta, GA, 1995.

[2] M.Q.W. Baldonado and T. Winograd. SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interest. In *Proceedings of the ACM SIGCHI '97 Conference on Human Factors in Computing Systems*, pages 11–18, Atlanta, Ga, March 1997.

[3] G.D. Battista, P. Eades, R. Tamassia, and I.G. Tollis. Algorithms for Drawing Graphs: an Annotated Bibliography. Technical report, Brown University, June 1993.

[4] S.K. Card, G.G. Robertson, and W. York. The WebBook and the Web Forager: An Information Workspace for the World-Wide Web. In *Proceedings of the ACM SIGCHI '96 Conference on Human Factors in Computing Systems*, pages 112–117, Vancouver, Canada, April 1996.

[5] J. Carriere and R. Kazman. Searching and Visualizing the Web through Connectivity. In *Proceedings of the Sixth International World-Wide Web Conference*, pages 701–711, Santa Clara, CA, April 1997.

[6] G.W. Furnas. Generalized Fisheye Views. In *Proceedings of the ACM SIGCHI '86 Conference on Human Factors in Computing Systems*, pages 16–23, Boston, MA, April 1986.

[7] Ashish Gupta. Junglee: Integrating Data of All Shapes and Sizes. In *Proceedings of the 14th International Conference on Data Engineering*, Orlando, Florida, USA, February 1998.

[8] M.A. Hearst and K. Chandu. Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy. In *Proceedings of the ACM SIGIR '97 Conference on Research and Development in Information Retrieval*, pages 246–255, Philadelphia, Pa, July 1997.

[9] R.J. Hendley, N.S. Drew, A.M. Wood, and R. Beale. Narcissus: Visualizing Information. In *Proceedings of the 1995 Information Visualization Symposium*, pages 90–96, Atlanta, GA, 1995.

[10] J. Lamping, R. Rao, and P. Pirolli. A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In *Proceedings of the ACM SIGCHI '95 Conference on Human Factors in Computing Systems*, pages 401–408, Denver, CO, May 1995.

[11] W. Li and J Shim. Facilitating Complex Web Queries through Visual Interfaces and Query Relaxation. *Computer Networks and ISDN Systems. Special Issue on the Seventh International World-Wide Web Conference, Brisbane, Australia*, 30(1-7):149–159, April 1998.

[12] Y. Maarek and I.Z.B. Shaul. WebCutter: A System for Dynamic and Tailorable Site Mapping. In *Proceedings of the Sixth International World-Wide Web Conference*, pages 713–722, Santa Clara, CA, April 1997.

[13] J. D. Mackinlay, S.K. Card, and G.G. Robertson. Perspective Wall: Detail and Context Smoothly Integrated. In *Proceedings of the ACM SIGCHI '91 Conference on Human Factors in Computing Systems*, pages 173–179, New Orleans, LA, April 1991.

[14] S. Mukherjea and J.D. Foley. Visualizing the World-Wide Web with the Navigational View Builder. *Computer Networks and ISDN Systems. Special Issue on the Third International World-Wide Web Conference, Darmstadt, Germany*, 27(6):1075–1087, April 1995.

[15] S. Mukherjea and Y. Hara. Focus+Context Views of World-Wide Web Nodes. In *Proceedings of the Eight ACM Conference on Hypertext*, pages 187–196, Southampton, England, April 1997.

[16] R. Rao and S.K. Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information. In *Proceedings of the ACM SIGCHI '94 Conference on Human Factors in Computing Systems*, pages 318–322, Boston, MA, April 1994.

[17] G.G. Robertson and J.D. Mackinlay. The Document Lens. In *Proceedings of ACM Symposium on User Interface Software and Technology '93*, pages 101–108, Atlanta, GA, 1993.

[18] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[19] K.J. Vincente and J. Rasmussen. The Ecology of Human Machine Systems: II. Mediating "Direct Perception" in Complex Domain. *Ecological Psychology*, 2:207–249, 1990.