# Presenting Web Site Search Results in Context: A Demonstration

Michael Chen
Computer Science Department
492 Soda IIall
University of California, Berkeley
Berkeley, CA 94720-1776
http://www.cs.berkeley.edu/~mikechen

Marti A. Hearst
School of Information Management & Systems
102 South Hall
University of California, Berkeley
Berkeley, CA 94720-4600
http://www.sims.berkeley.edu/~hearst

We address the case of search over large, heterogeneous web sites such as those found at universities and within corporate intranets. The goal is to make use of structure implicit within the site to provide context for the retrieved documents, even for those sites for which there is no centralized organization.

Most web site search, if available, presents results as a list of titles and URLs (and optionally, summary information and an opaque score). The context in which the pages appeared, and their relationship to one another, cannot be discerned from such a display. Our alternative is to place the retrieved pages within an automatically generated context. For each hit, we determine a root page (typically, the home page for a server within the target domain) and find and display the shortest path from the root to the retrieved page. Hits that share paths are grouped together. Thus, a hierarchical structure is dynamically imposed on the search results. We demonstrate our idea on web pages within the berkeley.edu domain. (Space does not permit inclusion of a screenshot.)

We have found this grouping by link path information, although simple, is a surprisingly effective way to organize what might otherwise be perceived as a disconnected jumble of pages. The system, which we call Cha-Cha, is written in Java. A web crawler gathers inlink, outlink, and title information from all the pages in the domain of interest and builds up a graph representation, which is stored in a disk-based database (from www.sleepycat.com).

Searches are sent to the Inktomi search engine via a specially set up connection, using Inktomi's IDP protocol (www.inktomi.com). The results returned are parsed into a set of hits. For each hit, the root node is the home page for the server on which the hit is located, and a bidirectional shortest path search is conducted from the root to each hit. A new graph is constructed that combines all the paths so that hits located close together in web space are shown near one another in the display. Hierarchially structured HTML is produced by traversing this graph.

This work is closely related to that of SuperBook [1], which demonstrated that showing hit search results in the context of the chapters and sections of the manual from which they are drawn can improve users' information access experiences. The AMIT system [5] indexed a web site covering a specific topic (sailing) in a similar manner, providing a Superbook-like focus-plus-context environment to place search results in context. Our project is of larger scale and broader functionality.

The WebTOC system [3] also imposes a hierarchical table of contents over a web site, but focuses on showing the number of pages within a subdirectory and emphasizes browsing at the expense of search. Also relevant is the Dynacat system [4] in which medical journal abstracts are organized according to what type of query they are found in response to and where they fall within a taxonomy of medical metadata.

In future we plan to incorporate user behavior into the interface. We envision a "learning" mode in which parts of the web site the user does not usually see are shown first, and a "familiar" mode in which hits found along the users' most often traveled paths are shown towards the top of the list of search results. We also plan to incorporate general domain knowledge into the ranking and organization algorithms (currently the domain is universities). In the next few weeks we plan to conduct user studies comparing the Cha-Cha approach to standard output produced by other search engines. We also plan to use Cheshire II [2] as our search backend.

**Acknowledgments** We thank Eric Brewer for arranging access to Inktomi, and Kevin Heard for general technical assistance.

## References

[1] Dennis E. Egan, Joel R. Remde, Thomas K. Landauer, Carol C. Lochbaum, and Louis M. Gomez. Behavioral evaluation and analysis of a hypertext browser. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 205–210, May 1989.

[2] Ray R. Larson, Ralph Moon, Jerome McDonough, Lucy Kuntz, and Paul O'Leary. Cheshire ii: Design and evaluation of a next-generation online catalog system. In *ASIS '95: Proceedings of the 58th ASIS Annual Meeting*, 1995.

[3] A. Nation. Visualizing websites using a hierarchical table of contents browser: Webtoc. In *Proceedings of the Third Conference on Human Factors and the Web*, Denver, CO, 1997.

[4] Wanda Pratt. Dynamic organization of search results using the umls. In *American Medical Informatics Association Fall Symposium*, 1997.

[5] Kent Wittenburg and Eric Sigman. Integration of browsing, searching, and filtering in an applet for web information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Late Breaking Track.* ACM, 1997.