OODB Support for WWW Applications: Disclosing the internal structure of Hyperdocuments

J. T. de Munk, A. T. M. Aerts, P. M. E. De Bra
Department of Mathematics and Computing Science
Eindhoven University of Technology
PO Box 513, 5600 MB Eindhoven
The Netherlands
{munk,wsinatma,debra}@win.tue.nl

1. Introduction

Most World Wide Web (WWW or Web) servers use the operating system's native file system for the storage of the HTML documents and their embedded images. This access mechanism works fine for direct access, but is ill suited for finding documents containing certain information. Even just finding all documents that exist on a server is difficult because of the complexity of the hypertext link structure. The structure may not be completely connected, meaning that some documents on a server may not even be reachable by following links from other documents.

Several attempts have been made to build additional structures, that provide search facilities for the information stored on a WWW-server or a cluster of servers. Glimpse [Manber & Wu 94] provides indexing at the server level. Harvest [Schwartz et al. 94] extends Glimpse to offer retrieval over a set of servers.

The existing index databases ignore most or all of the internal structure of the documents. Asking for information that appears in a "header" of certain levels, e. g. levels 1, 2 and 3, is not possible. Finding information in an <address> field is impossible as well.

The source of the problem is the lack of a sound access-mechanism for the information one is interested in. The flat file system approach taken by Web servers makes it easy to access a whole document, given its address, but makes associative retrieval difficult. An "inverted" access mechanism is needed, providing access to documents or parts of documents, given a description of their contents, their internal structure, or the link structure of their environment.

Instead of adding an index-database onto a file system based Web server, we propose a server based on an object-oriented database, delivering the documents and the answers to search requests from the same information source. Documents are stored as objects of which the internal structure represents the HTML structure, thus enabling querying for structural elements like headers, hypertext links, quotes, addresses, etc.

The most influential similar approach to enhancing the WWW is the Hyper-G project [Andrews et al. 95], which has a richer structure than WWW, but can reduce documents to HTML in order to serve them to WWW browsers. We take a different approach by keeping the WWW architecture, to the point where we integrate an existing WWW server with an object oriented database system.

2. Requirements and Properties

When we started the development of the new server architecture, the following requirements guided the process, and resulted in the corresponding properties in the prototype system :

- The new server had to be built using freely available technology wherever possible. This requirement resulted in the following choices :
  - Instead of building a completely new WWW server, the code for the CERN server was used. Two versions of the new server have been built: one that leaves the CERN server code completely intact, and one that uses a small modification to eliminate the need for an additional CGI-script (which slows things down).
  - The Ode database system was chosen [Agrawal & Gehani 89] [Gehani 91], because it was and is freely available to universities (after signing a license agreement with AT&T).

  - The HTML parser needed to be at least as forgiving to syntax errors in HTML documents as most browsers. For this reason we have developed our own HTML parser, instead of reusing an existing parser like SGMLS or SP [Clark 95].

- The WWW server of our department would be the first test case. Since we have many people serving documents off that server, the transition from the file system based server to the Ode based server had to be very smooth. We have therefore developed a "two headed" server, which can serve documents from the file system and from the database simultaneously and transparently. The user cannot deduce from the address (URL) of a document, or from its appearance, whether it is served from the file system or the database. In fact, with scripts like the common *imagemap* program, it is possible to use the existing (file system based) script to associate the coordinates of a mouse click to a document, and to subsequently serve the document from the database.
- The server needed to parse HTML documents to store them internally as structured objects, but needed to be able to reproduce the documents exactly as they were input, including details that are insignificant to HTML, like the use of upper- or lowercase in HTML commands, the order of attributes to commands, additional white space and the position of newlines. By realizing this goal through the object structure used in the database it becomes easy for users working on documents together to retrieve a document from the server and compare it (using *diff* or a similar program) to their local version in order to find modifications made by their coauthors.

3. Future Work

We have built a WWW server which stores the internal structure of documents in the object oriented database Ode. A primitive search facility has been added to this server. The internal index structures to make searching efficient will be added in the near future.

We also wish to integrate our document repository system DReSS [De Bra & Aerts 95] with this server. By doing so the basic unit for locking can be reduced from a whole document to the smallest structural HTML elements, enabling more concurrent authoring than with the current version of DReSS.

4. References

[Agrawal & Gehani 89] Agrawal, R., Gehani, N. H., (1989). Ode (Object Database and Environment) : The language and the datamodel. Proc. ACM SIGMOD Int. Conf. on Management of Data, 36–45.

[Andrews et al. 95] Andrews, K., Kappe, F., Maurer, H., (1995). Serving Information to the Web with Hyper-G. Third International WWW-Conference, Darmstadt. (http://www.igd.fhg.de/www/www95/papers/105/hgw3.html)

[De Bra & Aerts 95] De Bra, P.M.E., Aerts, A.T.M., (1995). Multi-User Publishing on the Web : DReSS, A Document Repository Service Station. Proc. NLUUG Autumn Conference, Publishing on the World Wide Web, Ede, 13-26. (http://www.win.tue.nl/win/cs/is/debra/dress/dress-paper.html)

[Clark 95] Clark, J., (1995). SP, An SGML System Conforming to International Standard ISO 8879 — Standard Generalized Markup Language. (http://www.jclark.com/sp/index.htm)

[Gehani 91] Gehani, N. H., (1991). The Ode Object-Oriented Database Management System : An Overview. Advances in object-oriented database systems, 355-387.

[Manber & Wu 94] U. Manber, S. Wu, (1994). Glimpse : A Tool to Search Through Entire File Systems. Proceedings of the USENIX Winter Conference. (http://glimpse.cs.arizona.edu:1994/)

[Schwartz et al. 94] Schwartz, M.F., Bowman, C.M., Danzig, P.B., Hardy D.R., Manber, U., (1994). The Harvest Information Discovery and Access System, Proceedings Second WWW Conference, Chicago, 763-772.
(http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/schwartz.harvest/schwartz.harvest.html)