InfoHarness: A System for Search and Retrieval of Heterogeneous Information

Leon Shklar *‡ Amit Sheth † Vipul Kashyap †‡ Satish Thatte *

1 Introduction

Enormous amounts of heterogeneous information have been accumulated within corporations, government organizations and universities. It is becoming increasingly easier to create new information, but the knowledge about the existence, location, and means of retrieval of information, have become so confusing as to give rise to the phenomenon of *write-only* databases.

2 The InfoHarness Prototype

The *InfoHarness*^{TM} [1] [3] prototype is aimed at providing rapid access to huge amounts of heterogeneous information from the World-Wide Web browsers. InfoHarness provides advanced search and browsing capabilities without restructuring, reformatting or relocating the original information. This is achieved through encapsulating new and existing information in objects that are encoded by metadata entities.

One of our main objectives is to provide a framework for utilizing existing storage and retrieval methodologies. To organize the processing and representation needs of different types of information, we have designed a stable hierarchy of abstract classes and an extensible hierarchy of terminal classes. The latter simplifies support for new information types and utilization of new indexing technologies.

The InfoHarness prototype is now operational and is on trial at Bellcore for a variety of applications It supports the largely automatic generation of InfoHarness repositories, and provides access to information from Mosaic through an HTTP gateway. We expect to make the system available on the Internet in 1995.

3 Current Work

The main directions of our current work are as follows:

1. The design of a declarative repository definition language [2] to attain the power and flexibility in controlling the metadata generation process
2. Enhancing the quality of the retrieval by combining results of querying attribute-based metadata and multiple heterogeneous indices that reference the same information.
3. Supporting scalable search by meaningfully combining results of querying homogeneous or heterogeneous indices on different collections of objects.
4. Automating the design of metadata extractors for both structured and un-structured information.

Additional information on the project may be obtained at *http : //athos.rutgers.edu/~shklar/ih.html* or *http : //www.cs.uga.edu/LSDIS*

References

[1] L Shklar, S. Thatte, H. Marcus, and A. Sheth. The "InfoHarness" Information Integration Platform. *http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/shklar/shklar.html.*

[2] L. Shklar, K. Shah, and C. Basu. Putting Legacy Data on the Web: A Repository Definition Language, *To appear in the Proceedings of the Third International WWW Conference'95, April 10-14, Darmstadt, Germany, http://www.igd.fhg.de/www/www95/www95.html.*

[3] L. Shklar, A. Sheth, V. Kashyap, and K. Shah. InfoHarness. Use of Automatically Generated Metadata for Search and Retrieval of Heterogeneous Information. *To appear in the Proceedings of CAiSE'95, June 12-16, Jyvaskyla, Finland.*

*Bell Communications Research, 444 Hoes Lane, Piscataway, NJ 08854-4182, leon@bellcore.com, thatte@ctt.bellcore.com

†LSDIS, Department of Computer Science, University of Georgia, Athens, GA, 30601, {amit,kashyap}@cs.uga.edu

‡Computer Science Department, Rutgers University, New Brunswick, NJ 08903, {shklar,kashyap}@cs.rutgers.edu

^{TM} InfoHarness is a Bellcore Trademark