Text extraction: Developing techniques to automatically extract entities, attributes and relationships from text, and to integrate this capability with the retrieval techniques. This work involves more sophisticated natural language processing than is typically needed for information retrieval, and is more domain knowledge-intensive.

Integration with database systems: Developing approaches to integrating information retrieval and database system query languages, query processing, and object management. A current focus has been integration at the object management layer for consistency with regard to concurrency and recovery strategies, and to satisfy performance constraints of large-scale text-based applications.

A number of digital library projects involving collaboration between the CIIR and its members are underway. From our perspective, a digital library application is synonymous with a distributed, text-based, multimedia information system application, but the three projects described below also have the characteristic of providing access to a wide variety of end users, including the general public. The example projects involve the Library of Congress, the Department of Commerce, and a collection of organizations involved with environmental technology (the Envirotech project). The scope of these projects and the issues involved are discussed briefly below.

The Library of Congress is in the process of evaluating new technologies to provide on-line access to parts of their huge collections. As part of the "American Memory" project, the library has worked with the CIIR to provide text-based access to collections of photographs, speeches, and books. The initial prototype has used a Mosaic interface that incorporates some aspects of an advanced text retrieval interface, such as natural language queries, ranked output and relevance feedback. Future versions of the system will incorporate more advanced features of INQUERY such as probabilistic field-based retrieval and phrase-based representations. Some of the research issues that can be addressed in the context of this project include effective query processing for image-based queries (what types of queries do people ask in an image database?) and indexing strategies for MARC records, which have many short fields of varying importance.

The Department of Commerce is working with the CIIR to provide access to the National Trade Database and related information. As for the Library of Congress, an initial prototype has been made available using Mosaic. Because there are more than a hundred related databases with at least some textual content, the issue of effective merging of local retrieval results is important. The problem of heterogeneous information, and particularly the combination of tabular and textual data is also significant. We are carrying out experiments to test the effectiveness of indexing the textual information in tables, and how this type of retrieval can be merged with more conventional text retrieval.

In the Envirotech project, we will be providing access to a heterogeneous and multilingual collection of information. This information includes environmental regulations from the E.P.A. and its Mexican equivalent, other government-related environmental documents, environmental technology newsletters and journals, company databases, and possibly news articles. The challenge is to construct order from this collection, and to support a uniform way of accessing and browsing the information in it. In an initial prototype, we have automatically linked a pre-existing company and product database with text databases using simple extraction capabilities, and have automatically constructed a thesaurus based on the phrasal associations in the collection. We have also developed Spanish and Japanese versions of INQUERY and have begun work on techniques for cross-lingual retrieval (retrieve in different language than the query).

More details on the CIIR, including a bibliography and pointers to prototype systems, can be found on our WWW server, at http://ciir.cs.umass.edu/. C

W. Bruce Croft is professor and director of the CIIR, at the Computer Science Department of the University of Massachusetts, Amherst.

World-Wide Web and Computer Science Reports

Edward A. Fox

With the advent of the World-Wide Web, computing professionals have eagerly pursued the idea of moving from a paper-based technical report service to one that employs networked information systems. Many departments keep some version of their reports on an FTP server to help with this process. To facilitate access to CS reports a number of sites have set up lists of these archives (see sidebar).

This situation led to the seminal effort at Indiana University of setting up a Unified Computer Science

CS technical report sites lists and search services

URL: http://cs-tr.cs.cornell.edu/
Organization: ARPA CS-TR (esp. Dienst system at Cornell U.)
Content: CS technical reports of nine participating sites
Services: Local support software, browsable site lists, distributed search, report retrieval

URL: http://www.cs.cmu.edu:8001/afs/cs.cmu.edu/user/jblythe/
Mosaic/cs-reports.html
Organization: Carnegie-Mellon Univ.
Content: List of over 180 FTP or WWW archives, and CS TR search services

URL: http://harvest.cs.colorado.edu/brokers/cstech/query.html
Organization: IRTF-RD: Harvest system (Transarc, U. Arizona, U. Colorado, USC)
Content: Data extracted from 300 FTP archives
Services: Local Gatherer software, distributed search, report retrieval

URL: http://www.rdt.monash.edu.au/tr/siteslist.html
Organization: Monash Univ.
Content: List of over 280 FTP sites that appear to have CS reports
Services: List of FTP sites, central search

URL: http://fas.sfu.ca/1/projects/EPiCS/CS-TechReports
Organization: Simon Fraser Univ.
Content: List of over 120 FTP archives

URL: http://www.cs.indiana.edu/cstr/search
Organization: UCSTRI (at Indiana Univ.)
Content: Data extracted from 180 FTP archives
Services: FTP archive list, central search, report retrieval

URL: http://www.cs.odu.edu/WATERS/WATERS-GS.html
Organization: WATERS (at ODU, SUNY Buffalo, UVA, Virginia Tech)
Content: CS technical reports of 15 participating sites
Services: Local support software, browsable site lists, central search, report retrieval

Technical Report Index (UCSTRI) [4], which polls FTP sites, builds a centralized master index from extracted data, supports searching of that index, and then retrieves reports that appear interesting (based on bibliographic data or in many cases, an abstract). This service requires centralized efforts to keep the information correct, and can do only a partial job of finding and correctly analyzing data, since there are no standards or policies universally adopted for CS report archives.

The Internet Research Task Force Research Group on Resource Discovery (IRTF-RD) runs a CS report service [1] that differs from UCSTRI in that it encourages sites to run Gatherer software which has been specially prepared to analyze files, and uses more sophisticated retrieval and distributed processing programs. Thus, the Harvest system is designed as an integrated set of tools for managing information on the Internet, and though it has and can be tailored further to handle CS technical reports, its reliance on automatic content extraction means it yields inferior results when compared with manually generated indexes.

The two projects described below aim to: encourage development of CS report archives with bibliographic data in standard form, supply useful software to run at local sites to help this process, and provide tailored user interfaces. The WATERS effort [3] adopts the model of centralized search while the ARPA CS-TR Dienst software [2] searches each archive site in parallel.

Other proposals have suggested applying compression methods (see Bell et al. in this issue) to the data. For those interested in publishing, browsing, and searching CS reports it would be best for all of these ideas to be integrated into one (virtual) CS digital report library that becomes universally accepted. This matter will be explored at an NSF-hosted workshop involving several of the groups discussed in this article, to be held in Spring 1995. Perhaps, then, a similar service could be used by other disciplines, and for all other "gray literature."

References

1. Bowman, C.M., Danzig, P.B., Hardy, D.R., Manber, U., and Schwartz, M.F. The Harvest Information Discovery and Access System. In *Proceedings of the Second International WWW' 94: Mosaic and the Web, WWW'94,* 1994, pp. 763-771.
2. Davis, J. and Lagoze, C. Drop-In Publishing With the World-Wide Web. In *Proceedings of the Second International WWW'94: Mosaic and the Web, WWW'94,* 1994, pp. 749-759.
3. Maly, K., French, J., Selman, A., and Fox, E. Wide Area Technical Report Service. In *Proceedings of the Second International WWW'94: Mosaic and the Web, WWW'94,* 1994, pp. 523-533.
4. VanHeyningen, M. The Unified Computer Science Technical Report Index: Lessons in indexing diverse resources. In *Proceedings of the Second International WWW '94: Mosaic and the Web, WWW'94,* 1994 pp. 535-543.

Edward A. Fox *is associate professor at the Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Va.*