The Electronic Library Project

Overview

- Initiative of the Corporation for National Research Initiatives (Robert Kahn and John Garrett)

- Five Universities

  Some combination of CS, Library, and Computer Center people.

  - UCB (Wilensky, Stonebraker, Larson, Buckland, Lynch)
  - CMU (William Arms)
  - MIT (Jerome Saltzer)
  - Stanford (Rebecca Lasher and Hector Garcia-Molina)
  - Cornell (Stuart Lynn and Dean Krafft)

- Two "advisory" organizations: ISI and Meridian

- Funding from DARPA (for 3 years, starting October 1992)

  Our budget? $150-300k/yr

The Electronic Library Project

Global Project Goals

Short-term focus: Capturing computer science technical reports and making them available electronically.

- A shared, electronically available bibliography of all CS tech reports at participating universities.
- Locally stored full-text documents in page image and text format.
- Real-time access to resources over the Internet.
- Standard protocols for search, retrieval, authentication, access control, bibliographic and rights and permissions management, image collection and storage.
- Instrumentation for monitoring and evaluation.

Longer-term focus: Build the foundation for a self-sustaining system for wide-spread networked access to technical documents.

Develop technology for

- Distributed, networked information access (Client/server models, distributed searching, issues of scaling, protocols)
- User-interfaces
- Intelligent indexing and retrieval

The Electronic Library Project

Our Goals

General

- Design, implement and evaluate the viability of various types of user interfaces and retrieval paradigms for the electronic library.
- Integrate a DBMS view of electronic libraries with a communication networking view
- Investigate methods that will allow cost-effective search over a large number of widely distributed library sites.
- Investigate exploiting natural language processing and artificial intelligence techniques to improve information access and retrieval.

Specific

- Create a database of UCB Computer Science and other technical reports
- Build a protocol-compatible information server (by building a protocol-to-POSTGRES converter)
- Build a text-oriented user interface (i.e., a protocol-compatible Chesire)
- Build a visually-oriented browser (the Sequoia Visual Browser)
- Build a Protocol-compatible Melvyl system client and server
- Create and implement a set of distributed information search methods
- Create and implement a document segmenter and access method
- Create and implement a semantically-oriented information browser

Tech Report Database Acquisition and Status

(as of March 2, 1993)

Mechanism:

- Email/ftp system set up to acquire new CS reports (preferably in PostScript). We examine and install.
- We requested electronic versions of recent reports.
- We're scanning in the rest.
- We create a bib record with abstract for each report.
- These are kept in ftp-accessible files on toe, and also loaded up into POSTGRES.

Status:

- Approx. 242 tech reports have been scanned, OCR'ed and installed (in `toe:pub/tech-reports`).
- 333 remain to be scanned. (65 are too fat for the scanner.)
- We're scanning at the rate of about 8 "normal" reports/day. (Probably will take 2 months to finish at current rate.)
- 52 reports have been deposited by authors (in PostScript, in `toe:pub/upload`; none moved to `tech-reports` yet, but most have been incorporated in a test version of Lassen).
- New reports are deposited at the rate of about one every two weeks.
- 37 Cognitive Science reports have been installed.
- There are 22 Sequoia reports on-line.

(Thanks to Julie Malork, Claire Mosher, Liza Gabato, Jon Forrest, Jim Frew and others!)

Information Retrieval

Versus

Document Retrieval

- State-of-the-art document retrieval leaves much to be desired.
- Moreover, most document retrieval has involved titles or abstracts, not full-text.
- Users want information, not just documents.
- Need new paradigms for retrieval of full-text documents and of facts available in document collections.
- Try to incorporate a little bit of NLP.

Document Segmenting and Topic-based Retrieval

Marti Hearst
Chris Plaunt
Anne Fountaine

- Divide a document into topical segments ("TextTiles").
  - Use a combination of standard statistic methods plus semantic category information to segment text.
  - Fully implemented, initial results seem good, many possible improvements.

- Use segments to enhance retrieval.

  One experiment just completed:
  - Segment texts, index as documents.
  - Retrieve against segmented collection.
  - Recombine results (in different ways) to determine relevance of parent document.
  - Compare against standard "whole document" retrieval.

  Used Ziff subset of TIPSTER test collection: 274 documents (of at least 1500 words), 43 queries. (Turns in 5,926 segment-documents.)

  Results: Best method (sum) showed improvement over standard approach (both in precision and recall) of between 18.9-28.2%.

Other Applications of Technology

- Implement "topic-subtopic"-based retrieval.

- Experiment with disambiguation algorithm based on semantic categories for retrieval.

- Characterize segments with topic descriptors; index in *topic lattice*; use for semantic-browsing of facts.

Searching Networked Resources

Michael Buckland

1. Central index – Impractical?
2. Simultaneous search of all nodes – Inefficient?
3. Sequential heuristic search – Slow?

Issues:

- Routing: Where? Where next? Stop now?
- Modify query heuristicly during search.
- Requirement, Preferences. Satisficing.

Economic model of optimal search strategies.
MELVYL library catalog database as virtual 100 node network.

Other Possible Collections

1. More tech reports (i.e., ERL)
2. DWR documents
3. DWR photo collection
4. Bancroft library photo collection
5. Videos of CS lectures
6. Scanning someone's office
7. Business card collections
8. (Your idea here)