Research in Information Access

Edward A. Fox
Dept. of Computer Science and Computing Center
Virginia Tech, Blacksburg, VA 24061-0106

October 20, 1993

1 Overview

The Research Department of the Virginia Tech Computing Center, and the Information Group in the Department of Computer Science, have emphasized new technologies and methodologies, and how they can apply to better services for the university community and beyond.

Important themes include:

- effective access to information — including searching, browsing, and linking — for large bibliographic, full-text, and multimedia databases;
- client-server and distributed computing approaches, based on low-cost workstations and fast RISC processors;
- object-oriented databases and object-oriented program development;
- integration of expert system and other AI methods with production processing;
- use of new, cost-effective storage technologies and efficient hashing algorithms;
- flexible human-computer interaction techniques and software for truly user-oriented systems;
- emerging standards for multimedia, hypermedia, and related information systems; and
- fast networking (e.g., FDDI, ATM).

2 Projects

2.1 MARIAN

MARIAN (Multiple Access Retrieval of Information with ANnotations) is a system being developed initially for library catalog record access, using almost 1 million records from the Virgina Tech Libraries. Its main design features include:

- Interactive access from a variety of client machines, operating on the Internet, using their standard interfaces, including VM/CMS, Motif, curses (for UNIX systems), and NeXTstep.
- Searching will be against the normal catalog data and/or annotations that will be added by members of the University community (with proper screening).

- Advanced search methods (e.g., pattern matching, relevance feedback, and extended Boolean) will be supported.
- The system will run on a small number of (NeXT or OSF1-compatible) workstations, so that incremental growth will be possible to meet an expanding user community.

After the library catalog version is complete, other versions can be built, using much of the same software, to provide access to a variety of bibliographic and full-text databases.

2.2 CODER

Work at Virginia Tech on the CODER (COmposite Document Expert/extended/effective Retrieval) system began in 1985 with initial support from the National Science Foundation. This has served as the foundation for much of the MARIAN system and for prototyping efforts related to full-text analysis, searching and browsing. CODER makes use of Prolog, C, and C++ modules, incorporating artificial intelligence, object-oriented programming, and object-oriented databases (through LEND). CODER uses a heterogeneous network of systems on the Internet running UNIX.

CODER has been applied to Navy messages, and to a collection of information about cardiology. The National Library of Medicine provided roughly 300,000 bibliographic records on that topic, along with the full Medical Subject Headings thesaurus. Columbia University provided a small text book on cardiology that could be search and browsed. Extensions were made to CODER by a number of graduate students to provide a Motif interface, to analyze the data, to support searching and following links, and to connect the cardiology information in with dictionary data taken from the *Collins English Dictionary*.

2.3 LEND

LEND, the Large External object-oriented Network Database, has been under development since 1989. It makes use of a family of newly developed algorithms for very efficient access to data, using perfect hashing, developed at Virginia Tech. There are routines for clustering data to further improve efficiency, and to support ad hoc queries against a graph or network view of the data. New classes for objects can be added, to complement the existing base, which can support semantic network, hypertext, hypermedia, and information retrieval applications. As such, LEND serves as an object-oriented database system with excellent performance for a variety of important applications.

LEND runs on NeXT computers and SPARCstations, and will be ported to RS/6000 systems. Requests for licensed copies have been received from companies in Virginia, Maryland, Texas, and California, as well as several universities including one in the Netherlands. Two corporations are using LEND for their internal development efforts, and one is developing commercially available information management services based on LEND.

2.4 Envision

The National Science Foundation awarded a 3 year grant of $1/2M, beginning 9/15/91, to create a user-centered database from the computer science literature. This involves use of hypertext, hypermedia, multimedia, object-oriented database, human-computer interaction, and other technologies. The Association for Computing Machinery, an educational and professional association founded in 1947 to serve the computing field, has agreed to provide access to all of its publications for this project. This ambitious effort to build an electronic library to help with education and research

in computing, that is making use of the CODER and LEND software developed at Virginia Tech, deals with electronic publishing methods and standards, visualization, task analysis of work by faculty and students, high volume scanning and OCRing, and networked databases of multimedia, full-text, page image, and bibliographic data. Work on digital video compression and usage, involving ACM's videotape archive, is planned.

2.4.1 Interactive Learning with a Digital Library

The National Science Foundation is expected to award in 1993 over $400K for a 3-year project entitled *Interactive Learning with a Digital Library in Computer Science*. With support from ACM (and its special interest groups in information retrieval and hypertext), IEEE Computer Society, Addison-Wesley Publishing Company, the Coalition for Networked Information, and other organizations, the Virginia Tech Department of Computer Science will enhance the undergraduate learning experience by integrating courses with usable and useful computerized materials comprising a comprehensive Digital Library. Building on work in Project Envision, this effort will engage the faculty and students in adding and improving a variety of courses, including: human-computer interaction, database management, data structures, algorithm analysis, and numerical analysis. Dissemination work will make the Digital Library, course modules, textbooks, and other results of the project widely available.

2.4.2 GeoSim

The National Science Foundation awarded a grant beginning 5/1/92, to improve instruction in Geography, especially at the introductory level, using simulation, multimedia, and geographic information system techniques. In collaboration with Computer Science and Geography, the Research Department is assisting in this effort, especially regarding multimedia and CD-ROM technology. In a related grant from Intel in November 1991, a large 386-PC based multimedia development system was received, with extensive software and special hardware for managing digital video. In 1992 and 1993, Intel has provided smaller additions to this grant. The Intel donations will be used in GeoSim for a module on orienteering (using compass and other tools in a simulated hike through fields and woods), and for online tutorials with full-motion video played back from a hard disk.

2.5 High Speed Networking and Digital Multimedia

Another NSF grant was awarded for research involving FDDI, a very high speed (100 Mbps) networking scheme that typically uses fiber connections. This effort involves Computer Science, Electrical Engineering, CNS, and the Research Department. Very high speed connections are being made between computers in the Computing Center and 3 other buildings, and several Intel multimedia systems will be connected, that should support video conferencing on PCs and/or workstations. A multimedia server in the Computing Center will support this and other investigations; see below under Equipment for more details.

With these capabilities, the Computing Center and Computer Science will work with IBM ActionMedia II products, supporting digital audio, images, and digital video. Building upon research that began in 1989, we will develop multimedia training, education, presentation, and other programs, and will support them with a central server and CD-ROM publishing operation. Extensions to include the MPEG standard for digital video are planned.

2.6 TREC

Grants from DARPA and PRC Inc. are supporting involvement of Computer Science and the research department in TREC, a Text REtrieval Conference, held in Fall 1992 and 1993. This effort involves development of information retrieval and routing (i.e., as in selective dissemination of information systems) methods to work with large collections. DARPA has provided over 2 gigabytes of data including about 1M articles, along with at least 100 questions and 100 topics for routing. Expert judgments of relevance of articles relative to the questions and topics will allow testing of various advanced retrieval methods. This effort will help with other campus efforts to develop more effective search and routing techniques.

2.7 TULIP

The University is involved in the TULIP Project, in conjunction with Elsevier, and about a dozen other universities. Elsevier will deliver an electronic version of 42 journals for 3 years, covering Material Science and Engineering. Several academic departments have expressed an interest in this effort, which will be coordinated by the Research Department, with support from the Library.

Roughly 10-15 gigabytes of data will be received each year, so a large jukebox and server system is needed (see section on Equipment below). Local software may be used, along with software for electronic journals from OCLC Inc., so that workstations and PCs on campus will have direct access to this large collection of page images, bibliographic information and ASCII text. Studies are planned regarding how such an electronic library can be used to support instruction and research.

2.8 Blacksburg Electronic Village

In addition to effort directly tied into the university, there are efforts of an outreach nature. In particular, there is the Blacksburg Electronic Village, where the entire community of 34,500 will have high speed data connections and will make use of a variety of services for education, commerce, government, and other applications.

3 Equipment

The Research Department has many computers. To support the abovementioned projects on high speed networking, Envision, TULIP, multimedia server activities (for GeoSim and other efforts), and other efforts, a high performance server machine has been installed: an RS/6000 Model 560, with 128M RAM and 8G disk. Two separate jukeboxes increase its storage capacity dramatically. One is for an optical disc subsystem to hold over 50G, and the other for a helical scan tape system managing at least 50G, to support unattended network backups and access to large files. This system acts as a multimedia server, and has an FDDI connection for high speed networking research. Also in the Research Department are 4 NeXTs and a DECstation, each with large amounts of storage.

In Computer Science there are over 100 workstations and PCs, as well as large parallel systems and servers. Directly involved in these research efforts are 2 DECstations, 2 NeXTs, an Okidata i860-based workstation, an Intel ActionMedia II development system, etc.