CHAPTER 1

Future Directions in Text Analysis, Retrieval and Understanding

METHODOLOGIES FOR INTELLIGENT TEXT PROCESSING AND INFORMATION RESOURCES MANAGEMENT: ACHIEVEMENTS AND OPPORTUNITIES

Report of the NSF Invitational Workshop on Future Directions in Text Analysis, Retrieval and Understanding

The Workshop was funded by NSF grant IRI-9114210 and co-sponsored by The University of Nebraska Center for Communication and Information Sciences.

Edited by

Jitender S. Deogun, Edward Fox, and Vijay V. Raghavan

Program Committee:

Abraham Bookstein
Christine Borgman
Jitender S. Deogun (Chair)

Susanne M. Humphrey
Alexa T. McCray
Vijay V. Raghavan
Clement Yu

Workshop Participants

Richard Belew, Univ. of California, San Diego, La Jolla, CA
Nicholas J. Belkin, Rutgers University, New Brunswick, NJ
Abraham Bookstein, University of Chicago, Chicago, IL
Peter Bollmann-Sdorra, Technical University of Berlin, Berlin, Germany
Christine Borgman, Univ. of California, Los Angeles, CA
W. Bruce Croft, Univ. of Massachusetts, Amherst, MA
Jitender S. Deogun, Univ. of Nebraska, Lincoln, NE
Edward A. Fox, Virginia Polytechnic Institute & State U., Blacksburg, VA
Mark E. Frisse, Washington University School of Medicine, St. Louis, MO
Ralph Grishman, New York U., New York, NY
Rick Hemphill, Naval Ocean Systems Center, San Diego, CA
Susanne M. Humphrey, National Library of Medicine, Bethesda, MD
Michael Lesk, Bellcore, Morristown, NJ
David Lewis, Univ. of Chicago, Chicago, IL
Alexa T. McCray, National Library of Medicine, Bethesda, MD
Michael McGill, Ameritech Information Systems, Dublin OH (Later U.Mich., Ann Arbor, MI)
Christopher C. Owens, Univ. of Chicago, Chicago, IL
Vijay V. Raghavan, U. of SW Louisiana, Lafayette, LA
Lisa F. Rau, GE Research & Development, Schenectady, NY
Gerard Salton, Cornell University, Ithaca, NY
Phil Smith, Ohio State University, Columbus, OH
Craig Stanfill, Thinking Machines Corp., Cambridge, MA
Tomek Strzalkowski, New York U., New York, NY
Beth Sundheim, Naval Ocean Systems Center, San Diego, CA
Kazem Taghva, Univ. of Nevada, Las Vegas, NV
Richard Tong, Advanced Decision Systems, Mountain View, CA
Irene Travis, The World Bank, Washington, D.C.
Carl Weir, Unisys, Paoli, PA
Michael Wong, Univ. of Regina, Regina, Canada
William A. Woods, Sun Microsystems Laboratories Inc., Billerica, MA
Clement Yu, Univ. of Illinois at Chicago, Chicago, IL

National Science Foundation Observer:
Maria Zemankova

The Report of the NSF invitational workshop on *Future Directions in Text Analysis, Retrieval and Understanding* is published in two volumes. Volume I, this document, summarizes discussions at the workshop and presents recommendations resulting from the workshop as well as related follow up activities. This volume consists of two parts. Part I presents the *Workshop Report* and Part II presents a proposal for a *National Electronic Library,* the ideas for which originated during the workshop. However, the details presented in this part were not actually discussed at the workshop, but were worked out via post-workshop meetings and discussions.

Volume II, a separate document, contains the collection of *Position Papers* prepared for and presented at the workshop. However, there are many places in Volume I where reference is made to papers that appear in Volume II. The two volumes are published as technical reports in the Department of Computer Science and Engineering at the University of Nebraska. Copies of both volumes can be obtained free of charge by ordering technical report number **UNL-CSE-93-006** for Volume I and technical report number **UNL-CSE-93-007** for Volume II.

Editor's Note: Part II, the proposal for a ***National Electronic Library*** *appears first here in Section A, followed by excerpts from Part I in Section B, the* ***Workshop Report****. Minor editing has enhanced the original versions that are available from University of Nebraska.*

Address for ordering Technical Reports:
Ms. Linda Endres
Department of Computer Science & Engineering
University of Nebraska,
Lincoln, NE 68588-0115

email: tech-reports@cse.unl.edu

Chapter 1, Section A: National Electronic Library

A National Electronic Science, Engineering, and Technology Library

Edited by

M. Lesk, E. Fox, and M. McGill

Executive Summary

The NREN is a new high-speed data communications network for the United States, and a key element of the High Performance Computing and Communications initiative HPCC). This initiative will bring significant benefits to the research and education communities of the U.S.; D. Allan Bromley, Director of the Office of Science and Technology Policy wrote that it "will lay the foundation for changes in education at all levels." In order to achieve these goals we must expand beyond the currently planned physical resources by adding to them a knowledge component -- providing in particular a basic scientific and engineering using high performance systems, extending applications of software algorithms, and supporting basic research and human resources training.

It is strongly recommended that the NSF should also support research on information and document management in the sciences and on the exploitation of electronic information resources. Research on scientific and technical information resources management would maximize value achieved from HPCC in general and this new on-line information resource in particular.

A prototype electronic library, accessible everywhere in the country, would protect our national leadership in the information industry, industrial competitiveness, and science and technology education. It is estimated that such a library could be created, and important related research funded, for a total of $50M over five years. As Vannevar Bush, the Director of the Office of Scientific Research and Development, wrote in 1945, ''instruments are at hand which ... will give man access to and command over the inherited knowledge of the ages. The perfection of these pacific instruments should be the first objective of our scientists as they emerge from their war work.'' As the Cold War ends, perhaps we should think again about this job, whose time has certainly come.

Information Retrieval

The inspiration for desktop access to large information resources was described in 1945 by President Roosevelt's science advisor, Vannevar Bush. His proposal for the ''Memex'' machine was ingeniously designed around bar-coded microfilm. The development of digital computers changed our conception of such machines, and beginning in the 1950s and 1960s a new field emerged, called ''information retrieval,'' a term invented by Calvin Mooers and popularized by Cyril Cleverdon. Computer typesetting of reference books made it possible to develop on-line systems as a by-product of conventional publishing. On-line systems now dominate some information use, for example, use of abstracts by professional literature searchers in libraries.

The information resources available include those from universities. For example, the University of California provides free access to its library book catalog through its MELVYL system, which handles 350K searches/week and is available on the Internet. Other resources come from governments, e.g., the National Library of Medicine's MEDLINE system of medical documents (with 70K searches made each week) or the catalog of publications from HMSO (Her Majesty's Stationery Office) in the United Kingdom. Many electronic resources now are privately developed and funded, such as the BRS/Saunders on-line full-text medical library.

The National Science Foundation played a key role in the development of this industry. Research funded in the 1960s by the Office of Science Information Services effectively started electronic information use. One of the largest projects was the computerization of Chemical Abstracts and the resulting development of CA files on-line and the STN (Science and Technology Network), with files relating to chemistry and other disciplines, which were funded originally by the NSF (although long since maintained, in this case, by funding from a professional society).

Other government agencies have been important to the development of the electronic information area. For example software developed by NIST (then called the National Bureau of Standards) was a model for commercial on-line retrieval systems. The National Technical Information Service (NTIS), the National Library of Medicine (NLM), National Aeronautics and Space Administration (NASA), the National Endowment for the Humanities (NEH), the National Agricultural Library (NAL), the Library of Congress (LC), and many other agencies have developed on-line information systems or machine-readable information resources. For example, none of the on-line library catalogs would have developed without the MARC tapes from the Library of Congress. These made possible a complete change in U. S.

library cataloging practice and stimulated organizations like the On-line Computer Library Center (OCLC) and the Research Libraries Group (RLG), and more recently the development of companies like Bibliofile. However, the NSF has a special responsibility to coordinate educational and human resources applications of the HPCC, and so logically should take the lead in any new science and technical information resources management programs using the NREN.

Information Technology Today

Until now the United States has led this industry. Most retrieval systems are based in the U.S., most technology was developed here, and the U.S. based industry here is larger than anywhere else. As we move further into the Information Age, the U.S.A. must retain its leadership in the information industries and must learn how to make better use of information resources.

Scientific and technical information resources management is critically tied to the HPCC. For example, parallel computers are now being used in advanced new text retrieval systems (such as the DowQuest service offered by Dow Jones that is implemented on two Connection Machines). New software algorithms for clustering and knowledge handling are part of modern retrieval systems. Networked and remote access is now typical, with many on-line catalogs, for example, available on the Internet both here and abroad. And, of course, neither basic research nor advanced education is possible without good information access, today increasingly meaning remote access to electronic databases.

Technology for Information Handling

The technology of information systems is diverse and changing rapidly. In the last thirty years the standard forms for distribution of machine-readable information have changed from paper tape, to punched cards, to magnetic tape, to optical disk (CD-ROM). Gigabyte stores (i.e., able to record 10^9 characters of text or other types of data) are affordable for workstations, and on-line terabyte stores (i.e., recording 10^{12} bytes) are commercially available. The information includes ASCII files, databases in various formats, images, video, sound, and combinations of all of these. The compactness of the data formats, in bytes per unit of weight, has increased by perhaps 300,000 fold since the days of punched cards.

Similarly, communications technology has greatly improved. Instead of the 110 baud teletypes of twenty years ago, we now have fibers carrying gigabits per second and experimental switches capable of the same speeds. The Arpanet of 1969 evolved into the NREN, and went from slow speed links to the present 45 Mbits/second and the future multi-Gbit backbone. Instead of only a few major research institutions, we now find tens of thousands of worldwide sites on the Internet.

Computational power has grown almost as fast. Even ten years ago a computer that performed one million operations per second was something to be proud of. There are now two manufacturers offering machines which promise one trillion operations per second. Supercomputers are routinely accessed over the Internet for remote operation. RISC architectures have made very fast workstations available for under $5000.

Fortunately, standards development has not slowed down this growth. The most important standards have been in static data coding: ASCII for character coding, and SGML for document representation. In interactive messaging, there is X.400 for email, X.500 for directories, and Z39.50 for formatting retrieval requests (at the heart of the Wide Area Information Servers network involving over 100 servers on the Internet). Z39.58 is the proposed Common Command Language to provide a consistent communication language for information systems and X.12 is the family of standards that provide for Electronic Data Interchange. For multimedia, two new standards will play an increasingly important role: the JPEG standard for gray scale and color images, and the MPEG standard for motion video and audio. All these standards are issued by ANSI (or its predecessors ASA and USASI) or by NISO or by international groups like ISO, IEC and CCITT.

A major change is coming from the need to include images and other forms of multimedia information. Image processing requires much larger storage, processing capacity, and communications speeds than does text handling. One page of text, for example, might contain 500 words or 3000 bytes.

A picture of this text (black and white, 300 dpi, Group IV FAX compressed) might be 50,000 bytes. If the 500 words were read aloud, which would take about 4 minutes, the corresponding sound signal, digitized and compressed (8KHz, 8 bits/sample) would contain 2,000,000 bytes. If the 4 minutes were stored as digital video, even with extreme compression to DS-1 (previously known as T-1) transmission rates (1.5 Mbit/second), more than 45,000,000 bytes would be needed. HDTV (high definition television) without compression is likely to require perhaps 45 Mbits/sec., another factor of 30 in information demand. Thus, the approximate expansion from text to image is 1:20; from text to speech is about 1:700, from text to compressed video is more than 1:15,000, and from text to HDTV 1:450,000.

The Information Industries

Electronic information access is an essential and rapidly growing industry. All ‘‘knowledge industries’’ (this includes all education) yield a third of the GNP; ‘‘information services,’’ broadly defined (including, for example, consultants and advertising agencies) contribute directly about 8% of the GNP. On-line information services now generate directly $6-7B per year and are growing at 20% a year [32].

The U.S. information industry in 1991 is estimated at about $10B, growing at 13% per year; the industry in Europe is $3.4B, growing at 19%; and in Japan it is $0.8B, growing at 30%. The largest company in the business is Dun & Bradstreet, with over $1B in revenues; other U.S. companies with more than $100M in revenues are ADP, Dow Jones, CDC, TRW, Mead Data Central, Citicorp, McGraw-Hill, Knight-Ridder, Trans-Union Credit, Equifax, and Computer Sciences Corp. Although the largest parts of the industry at the moment are financial services, scientific-technical-medical information in the U.S. is now a $600M industry, and is growing faster: 19% compared to 13% in financial and 9% in credit [11].

The largest general bibliographic on-line vendor in this area, Dialog Information Services, had revenues of around $100M/yr in 1989. By contrast, King Research estimates that the average scientist spends the equivalent of $8000 per year (in time and expenses) getting and reading information in all forms; the total market is estimated at $21B per year. Carnegie-Mellon University personnel have projected that an electronic offprint service would attract 6 million users, and represent a $7B market over the next twenty years.

There are now about one million CD-ROM readers in the world, and the number is doubling every year. About $3B worth of CD-ROM disks have been sold. In 1996 it is estimated that the amount of money the U.S. spends on electronic information will exceed the amount spent on paper information; at present publishers only get 10-20% of their revenue from electronic products. Multi-media information, it is predicted, will be another $10B business in 1996. As an example of how quickly these industries can grow, the 900 number business, which did not exist a few years ago, is now $700M per year.

Today the United States dominates the large text database industry. Vendors such as Dialog Information Services, Mead Data Central, Chemical Abstracts Service, the On-line Computer Library Center, and others are used by customers throughout the world. These databases grew out of a tradition of bibliographic and textual searching. Originally, they were intended for, and are still largely used by, professional searchers. Now end-user access is growing, which greatly increases the importance of information systems.

The information industry is extremely diverse. It includes government agencies (the National Library of Medicine and the National Agricultural Library), nonprofit private sector organizations (OCLC and RLG), and for-profit private firms (such as Dialog and Mead Data Central). Revenues for even the largest such firms are in the range of $400M per year. None of these organizations, inside or outside the government, compare in budget with, for example, the $6B printing/publishing companies in Japan or the almost as large companies in Europe. (Recall Robert Maxwell, of U.K., who was acclaimed for his ‘‘empire’’ in this area, which included a multi-media company and CD-ROM pressing plants.) Although the U.S. has thrived on this diversity, the size of some foreign competitors makes our situation dangerous.

The Information Challenge

Although the U. S. leads in the current on-line information industry, the changes in the industry leave future dominance at risk.

Foreign Initiatives

The French, for example, are very active in image processing and storage. At the Centre Pompidou a collection of 150,000 images is to be made available via ISDN throughout France. The French national library is proposing to scan 1,000,000 books, dwarfing any project elsewhere in the world. The first large distributed scientific image bank was the Adonis project (CD-ROMs of biomedical journals), run by a consortium of publishers out of Amsterdam. Information handling is a business with high start-up costs and low incremental costs; it is not going to be easy for competitors to enter this business if other countries get too much of a head start. Action in the U.S. is required to maintain our position in the information industry.

Foreign enterprises are also moving strongly into the U.S. information business. Foreign publishers, recording companies, and motion picture studios are buying U.S. companies in these businesses. The four largest printers in the world, for example, are all foreign. BRS and Orbit (two U.S. on-line system vendors) now belong to Maxwell (or a British bankruptcy court). Consumer electronics companies such as Philips and Sony are investing heavily in new electronic products and dominate these markets. For example, although the phonograph and videotape recorders were invented in the U.S. and the various formats of traditional records and tapes developed here, the VHS player market was taken over by the Far East, the CD was developed and introduced outside the U.S., and future products of this sort (the Sony Data Discman) continue to be developed by foreign companies. Major electronic information programs such as Adonis (a consortium of biomedical publishers) and TULIP (an Elsevier program for electronic archives in the crucial area of material science and engineering) are being run by foreign groups. Even in traditional areas, the increasing prices of journals and books, and the drop in the value of the dollar, have made U.S. research libraries unable to keep up their subscriptions to foreign journals [31].

U.S. technology in the information areas is in danger of becoming obsolete. Just as we have fuzzier television than Europe because we introduced the service at an early stage of technology, we have old software and systems in the information area. We focus on text when other media are becoming important; we have old fashioned intermediary-tailored search systems when end-user systems are becoming popular; and we have not integrated searching systems with information delivery systems. We rely on low-speed dial-up for most commercial services as medium speed (ISDN) and high-speed (NREN) networks are being built. Even in retrieval technology, commercial systems rely on simple Boolean search many years after the research community has demonstrated the value of term weighting, relevance feedback, and other techniques.

Information Competitiveness

Perhaps even more important than the status of the information industry itself, however, are the consequences to the U.S. of not having access to current scientific, engineering, and other technical information. With global competition in the marketplace, rapid technology transfer is essential to allow new products and services to be produced first by American corporations, and for maturing products to benefit as quickly as possible from new advances. It is critical for national competitiveness that technical information be made available rapidly and to the widespread domestic community. Presently, there are many places where the best technology is not used, because of failures in either education of staff or in technology transfer. Computer programmers use old languages, scientists use old measuring devices, and engineers lack the latest tools. The steel, automotive, and consumer electronics industries are well known examples where technology transfer failures have allowed other nations to supersede U.S. corporations as world leaders.

Delays in getting the best information to all workers arise from many reasons. Publications and patents are often delayed, sometimes for three or more years; electronic processing could yield significant speed-up. Access to publications is often too hard and/or too expensive for individuals or libraries. Indeed, the entire scientific publication process has been called into question regarding finances and effectiveness. Education is viewed as something that happens in school, where people simply retain information after graduation.

Even then, many students are not aware of the latest developments, and rarely have time or encouragement to browse or see how findings in disparate disciplines are inter-related. Further, we need to improve the ability of workers in industry to retrieve new information and keep at the cutting edge of new technology, i.e., engage in ‘‘lifelong learning.’’[18]

Other countries recognize these issues. In Japan, for example, a great many scientific conferences are run for the purpose of technology transfer. Academic conferences in the United States are normally to present new results; many corporations don’t pay for their staff to attend meetings unless they are giving a paper. By contrast, in Japan much less emphasis is placed on whether a conference describes new science, and more on whether many corporate employees are attending it. Often, major corporations in the U.S. recognize this problem, of course: IBM spends as much on training each year as the budget of Harvard.

But education in schools also needs help. Computer networks are being put into a great many colleges and universities, but all too often there is not very much in the way of information resources that is available. Generally an academic network mostly serves to transmit news, electronic mail and printer commands. An effort to provide technical information in electronic form will exploit networking to help students.

There have been many criticisms of the U.S. educational system. Our international competitiveness is slipping; and we are not able to make it all up with corporate libraries and training programs, which mostly exist only in large companies. We should be using electronic techniques to compensate for the inadequate resources of many public libraries, and to supplement what is now being done in college and university training. As students in K-12 (Kindergarten through 12th grade) gain access to computer networks at an increasingly rapid pace, education during those important years will also benefit from investment in making information available in electronic forms. Indeed, innovative efforts in Texas and other locations have begun to replace textbooks with information resources.

The capabilities being developed by the HPCC offer capacities to deal with these problems. Most obviously, the NREN would permit access throughout the country to advanced resources. In addition, the more powerful machines would allow greater flexibility and speed in answering user queries, and in extending information formats to new media. New software algorithms are needed to deal with the new media, particularly graphics, and to exploit better retrieval techniques. And, of course, human resources training is both the goal of this activity, and a supporting component to provide the people to do the work of building it. But the overall message is that we can not have an HPCC initiative that deals entirely with machines and networks; we can not build the systems without information. We will certainly use them to distribute information, and we definitely should use them in the most effective way to support science and technology in the United States.

Proposal for an Initiative in IR & DM Resources and Document Management

We recommend an initiative to build a basic science, engineering, and technology library on-line, available over the Internet/NREN, and for research regarding building and exploiting it. This would accelerate the development of the libraries of the next century, and improve the rate at which basic scientific knowledge is disseminated.

Our belief is that the library of the future will be based on electronic data. It will contain both text and graphics and be widely accessible via electronic networks. It is likely to be decentralized; even today, we are finding more and more libraries grouping into consortia to deal with the financial pressures of increasing prices for materials, and moving books to storage facilities in remote areas to deal with the pressures of increasing prices for land and buildings adjacent to today's facilities. As a result, it is already necessary for users to make use of collections which are in several places and owned by diverse managements; in the future, these problems can be alleviated by the ease with which information can be transmitted to the users' locations.

Which topic areas should be transformed into electronic form first? The Army, realizing the importance of optical digital information systems, has recently awarded contracts to scan over 100,000,000 pages of records; this may serve as a model for larger future DoD (Department of Defense) efforts of this kind. There are many non-military subjects to be treated. Also, we must support much more than just old records: the extension of electronic libraries, as soon as practicable, into all intellectual areas. Initially, however, we recommend scientific and engineering materials, for a variety of reasons:

- The users of such material are most likely to already possess computer workstations suitable for reading on-line, or to purchase such systems as the first generations of PCs are retired;
- This material has important impact on industry in the U.S.;
- Frequent updating, which electronics makes possible, is important in fast-moving areas such as science and engineering; and
- Science and engineering teaching is an area of U.S. education which has faced particular criticism, and in which a new initiative is thus appropriate.

We must use existing infrastructure to start with, wherever possible. The on-line resources of NLM in medicine, CAS in chemistry, UMI in dissertations, and OCLC and RLG in catalogs should be used as a base for whatever is built. Many major publishers are collecting their typesetting tapes; where possible, these should be used, with appropriate financial arrangements to safeguard authors, publishers and readers.

Similarly, existing technology must be exploited where possible. The most important resource is the NREN, and the U.S. base of installed computer workstations in corporations and colleges is also critical. New technology (cheaper disk drives, processors, parallel processors, and other areas) also is particularly important.

Collection Plan

We have the opportunity to make a major step forward in document handling, and to improve national scientific and educational productivity. This can be done by providing technical information in electronic form over the NREN. It will require putting together technologies in high performance computing, high performance networking, artificial intelligence, information retrieval, computer-human interaction, and other areas. The NSF should start an initiative to do it.

Why should we expect that electronics will help? Results with Bellcore's SuperBook and OWL's Guide program, for example, show that people using computers for text handling can get better answers, and find them faster than people using paper. Hypertext and hypermedia systems link related items to avoid the time-consuming labor involved in looking at related books or other media types. Over the last thirty years, research in information retrieval has shown how to improve systems to let people find information in text; and, in library after library, electronic search systems are replacing card catalogs and books of abstracts and indexes. The average U.S. scientist and engineer spends half time searching for and acquiring information; if, as in some SuperBook experiments, this could be done 25% faster with a gain in accuracy, the consequences for national engineering productivity would be immense. In addition, new modes of information access are possible only with an ‘‘active’’ electronic library.

A traditional library is passive; users come to it and search for what they want. In an electronic library it is possible to have pro-active elements, which send users messages when agents discover items that, based on earlier commands, it is known the users are looking for. In the electronic network of the future, we must compensate for the inability of the users to look around rooms of physical books. To facilitate this, we can build either asynchronously running programs which sit in the background, waiting for new information to arrive, as in SDI (selective dissemination of information) programs; or we can provide sophisticated ‘‘agent’’ programs, such as the ‘‘knowbots’’ described by Vinton Cerf, which are dispatched by users to search out and return desired documents. To truly serve a community of users these must learn about the needs, preferences, and capabilities of those users and consider that information during both search and presentation phases.

The central focus of the initiative should be the creation of on-line resources in a variety of scientific areas, which then could be used in new and exciting ways.

Recommendation:

The NSF should solicit proposals from groups of researchers to create on-line libraries in key scientific areas, in accordance with the following guidelines:

- Each group should explain how they will create an archive that will include materials for at least basic scientific activities, at the level of a completed undergraduate degree, in their area. They must agree to do this in a way that the end result will be compatible with the archives for other disciplines.

- Each group should involve participants who are (a) scientific users of information; (b) computer and information researchers; (c) publishers and other information creators and distributors.

- Each proposal should look to national, rather than local, availability of on-line information.

- Each group should address how their information will be used.

- In each case, the economic basis by which copyright permissions will be managed (whether a royalty, license, or other scheme) should be explained.

- The sum of all the proposals should provide reasonable coverage of basic science and engineering, without duplication of effort, and should be inclusive rather than exclusive of information sources (e.g., all of textbook, monograph, technical report and journal literature are candidates for the libraries).

We do not believe that these proposals should be constrained in terms of the technology to be used. Each group may propose the use of scanning, keying, OCR, re-use of typesetting data, or other technology. However, an effort must be made to maintain compatibility between different systems, regarding the representations used in the archives.

Since the intent is to provide discipline-oriented projects, it is likely that the larger publishers and other information creators may be involved in more than one library creation group. For this reason, the number of projects is being kept relatively small, to minimize the effort in coordinating the information conversion work.

Recommendation:

The NSF should create two oversight boards.

A. The principal investigators of each project should form a board to manage document standards, interchange principles, and other necessary agreements to produce a national science and engineering network rather than just local or discipline-specific networks.

B. A second board should consist of one representative from each of the scientific agencies in the U. S. Government concerned with these areas (e.g., the National Science Foundation, the Department of Energy, the National Aeronautics and Space Administration, DARPA, NIST, and so on). These representatives should ensure that the right information is available, and that shared funding can be arranged where appropriate.

These two boards will determine at planning and coordination conferences the ‘‘standards’’ for coordinated development of systems, ensuring that the results of the development and research activities are available in their entirety to the research and educational communities.

We anticipate that the sum of the necessary collections would amount to perhaps 100,000 volumes, or 50,000,000 pages. If this were converted to machine-readable form by scanning and OCR, the cost might approach $10M. Correcting errors, coordinating figures with text, marking up text and other elements, storing and allowing access, and other management costs might bring the total to $25M. That seems like a relatively small price to pay for creating a valuable national resource in science and technology. For example, if the existence of this collection permitted a 1% reduction in teaching time spent with undergraduates, there would be on the order of $100M (i.e., 1M students * $10,000 tuition/student X .01 reduction) saved in equivalent tuition costs. In fact, the real expenses (which should be borne largely by the users, not the NSF) are likely to be the royalty or license payments to the creators or publishers of the material.

With an archive developed, authors would prepare future articles and manuscripts in a compatible format, so minimal expense would be incurred in keeping the archive current. The National Institute of Standards and Technology might play a role in working with the two boards, to ensure that the archive follows standards, making it an ongoing project that should become self-sustaining.

Work on the applications of the collections should of course go hand in hand with the creation of the digital collections. These should include both ways of using the material in education and using it for technology transfer, as well as for retrieval. Educational applications of documents, and projects that merge documents, teaching, and student research are particularly important to get maximum value from our efforts.

Current Related Research

The basic library materials needed for an electronic library database can be collected today. The creation of large databases of text is technically feasible. In addition, recent advances in the technology of document handling are enabling construction of substantially more powerful systems for information access. Among the major areas in which progress has been made and continues to be made, and in which exploitation in applications offers great promise, are:

- Text analysis. Technology now exists to do simple natural language analysis and to try conceptual rather than purely word-based representation of knowledge, fully automatically. Local context analysis, detection of phrases, and assignment of concept knowledge is all becoming possible. These areas, after many years of effort, are now starting to be practical and must be pursued, since they could make it easier for users of an electronic archive to describe their information need, and could help increase the thoroughness or ‘‘recall’’ of searching.

- Theoretical modeling. We can now begin to predict the performance of retrieval systems, rather than just measuring them after the fact. Studies in user behavior are teaching us how to adapt computer systems to particular user tasks. Modeling should be tried to improve the use of systems and accelerate the design of new systems. Systems could then handle more users, more effectively.

- Technology. Disks and display screens are both getting larger, faster and cheaper. Audio and video systems are going to be common. Transportable systems will be coming into use, using wireless networks. Most important is the ability of parallel computers to implement search algorithms that laboratory tests show improve effectiveness of searching by over 30%, but were previously thought too expensive, e.g., the use of relevance feedback on the Connection Machine.

- Intelligent systems. On the basis of user studies, knowledge representation, and text analysis, it is possible to design rules for improving searches for users. In this way, we can extend the range of people who find computer information systems easy and fruitful to use, so that end-users rather than intermediaries do most of the searches.

- Indexing and structuring information. With large collections of information we can exploit the structure and context of individual works and their inter-relationships. Citation and subject searching have been shown to be much more effective when used together. Hypertext and hypermedia facilitate the associative access recommended by Vannevar Bush and bring together multimedia information resources for easy manipulation with pointing devices and multisensory presentation. Knowledge-based techniques for indexing permit more controlled and more powerful ways for manual organization of information resources.

- Graphical displays. Two-dimensional information representations are now being developed to give better overviews of a collection of documents than can be presented with a traditional shelf-list. Knowledge viewers for both documents and conceptual dictionaries are possible. Work in this area holds great promise for decreasing the difficulties many users find in dealing with retrieval systems, just as scientific visualization systems can help scientists work with their data and models.

Complementary Research

In order to make the best use of progress to-date in research, most of which has been limited because of practical considerations dealing with technology and funding, the collection building initiative should be extended. In addition to building the collections, we need research in their exploitation.

Recommendation:

The NSF should fund an expanded program of research into information utilization. This should include projects in

- multimedia storage and retrieval, including audio, video and image processing and compression techniques;

- combined text and data storage and retrieval, tied into the scientific databases initiatives;

- knowledge representation, including frame-based and other representations of information conveyed in documents, and including indexing, classification and retrieval as possible with higher-level knowledge representation that support logical or probabilistic inferencing;

- user studies, involving the nature of retrieval requests, user models, and other topics related to what people of all types, and with a myriad different sorts of needs, want from retrieval systems;

- human interface design, including hypertext, graphical and audio display methodologies both to deal with non-textual material and also to assist in text retrieval;

- evaluation techniques, including both traditional relevance assessments and research on new ways of measuring utility and dealing with uncertainty in system performance;

- distributed systems, including information sharing and the administrative and economic complexity of shared systems;

- coordination systems, in which people work together on information tasks;

- uncertainty and “fuzzy” systems, including systems where retrieval, document description and relevance may be defined in probabilistic terms;

- “agents” or “knowbots”, which execute user tasks autonomously in a non-interactive environment;

- organization and retrieval (clustering, classifying), i.e., the grouping or linking of documents, rather than their retrieval from queries;

- natural language text analysis, including lexical, syntactic, semantic and pragmatic considerations, but focusing on how natural language processing can improve text retrieval;

- theoretical foundations of information retrieval, such as models of collection structure, retrieval techniques, and methods for predicting retrieval performance.

All of these research topics should be in collaboration with and in the context of the data collection projects. Similarly, the data collection work should be done so as to facilitate research and evaluation on the collections. Observational research and controlled experimentation on users are both necessary and encouraged.

The goals of this research include developing: algorithms and interfaces to serve directly the broad class of users in our nation, that make them want to explore the electronic library; databases that support integrated keyword, concept, natural language, citation, and visualization approaches to search and browsing; methods to adapt to a spectrum of requests, from those requesting a few of the very best passages to others calling for a ranked list of all related materials; techniques to prepare multimedia educational materials in a cost-effective fashion by drawing upon data in the archive; mechanisms to stimulate action by users, through providing current awareness or notification of new useful data or explanations; capabilities to clearly show the interconnections and inter-relationships between raw data, primary research, commentary, and secondary sources; approaches to solve the cost and space problems of current libraries and at the same time to replace the confusing and ineffective world of scientific publishing; and procedures to scientifically compare systems that vary in terms of functions, efficiency, and effectiveness of delivering the right results.

NSF Role

The National Science Foundation is the obvious organization to launch this initiative. It coordinates scientific knowledge creation, scientific information transmission, and the HPCC and thus the NREN. It is also a major funding organization for basic scientific research; and can coordinate with those other parts of the government that apply scientific knowledge in specific disciplines, such as the Departments of Energy and Defense. The NSF is also the organization which takes the leading role in setting scientific goals for the United States, even when such projects are funded by other routes.

We estimate that a reasonable scale for this initiative would be to devote $10M per year for each of 5 years, or $50M total; to be divided approximately equally between the creation of the library and the research on its exploitation. We hope that it will be possible to make this a cooperative initiative strongly supported by all of the NREN organizers.

The total initiative would be able to fund about 10 data creation projects, each costing approximately $2.5M, which should suffice to build the basic library; and $25M in research to use the electronic library for education, technology transfer, and other purposes. Thus, there could be approximately 4 major research centers receiving $1M for each of four years plus 12 additional smaller 3 year research projects at $250K per year, which would be able to exploit the data resources.

Our funding scale is based on the assumption that an electronic library would cover perhaps 25 key areas, each demanding a mixture of books and journals involving 2 million pages. Scanning the pages for one of these areas might cost $300,000 and with other tasks (OCR, staff time, editing typesetting tapes, mark up, storage, and so on) the total effort would require perhaps $1M. By distributing the work of archive creation, this initiative would benefit from competition at the proposal and project stages, with the net effect of having many groups knowledgeable regarding handling of electronic texts. We suggest that these areas be grouped into perhaps ten data conversion tasks, to avoid excessive start-up costs for small conversion projects.

The result of these ten electronic library construction (data conversion) tasks would be approximately 50 million pages of on-line information. This is not a huge library; it is only about 10% of that planned for the French national library; Dialog adds more than this much material each year. In addition to data creation, we suggest that for each data file several exploitation and research projects should be performed, to get the maximum value from the data; each of these projects should be at the scale of at least two full-time investigators.

With the cooperation of NIST, and the involvement of the boards and researchers in the various projects, as well as the professional associations, publishers, authors, and others who have been involved through this initiative, the resulting electronic library would have such an impact on standards of practice that it should become a self-perpetuating institution. New authors would prepare their works for inclusion, and publishing habits would be reshaped based on the existence of the library. In a decade, it would become the envy of the world, and permeate the worlds of science, engineering, and technology.

Implications for the Nation

Building an electronic science library will have impact on education, industry, research and the general state of science in the U.S. It will permit people throughout the U.S. to immediately answer scientific questions and keep abreast of new technology more rapidly and with less effort. It will permit U.S. scientists, whether in industry or academia, to deliver their results immediately to educational or commercial users. It will stimulate development of computer-based training systems, informal information circulation, and other new forms of information transfer. And it will provide a base on which the U.S. can maintain a leading position in the international information industry.

Educationally, we should try to make a reference library suitable for any undergraduate in the major scientific and engineering disciplines at any U.S. educational institution. This library should include textbooks, handbooks, monographs below the graduate level, and major journals in the areas. It should store several major lecture courses, in audio and in some cases video form; and provide search systems to retrieve based on full-text searching. We should also encourage access to this material in advanced secondary schools to accelerate the education of advanced students. We should fund research, in cooperation with the Department of Education, on the best ways to use this material in conjunction with conventional or non-conventional courses to improve student education. For example, teachers could use the library to rapidly select the best sources on each subject they cover, and use hypertext methods to synthesize a new integrated whole, rather than having to pick a single text.

Our goal should be to provide the equivalent of the best library at every school, combining the advantages of the resources at the biggest institutions with the environment and diversity of the many colleges and universities throughout the U.S.

In the information industry, we should become leaders in developing the new basic technology that will be most commonly used. We also should develop leadership in delivery of this information over fast networks, and in the use of audio and image material to supplement plain text. We should build on the NREN to establish an information industry that serves the needs of U.S. industry and education and does not become dependent on outside data sources. The structure of the industry may change, but it should be possible, in new technology as well as old, to have a leading and self-sufficient industry in the U.S.

In terms of national competitiveness, we should see to it that all locations of technological industry are connected to the NREN and thus to the national electronic library. Each industry should be able to get current scientific and technical information; and should, of course, also benefit from the better education given to its employees. What is most important is that we routinely apply the best technical information to each problem. One possibility is to harness methods of computer assisted problem solving and visualization, employing them for educational purposes in the large. Electronic communication and electronic reference works should provide the best way to deliver up-to-date scientific solutions to real problems, especially with the young scientists of American trained in these techniques during their university years.

Research in information delivery will be facilitated by such a data library and network. Once people become accustomed to getting answers in seconds rather than hours or days, once they become accustomed to doing their own searches, they will get their information from the best sources rather than from those easiest to find. Researchers will use a library and network to experiment with new techniques related to data, audio, image and text handling and retrieval.

We must stress the urgency of building such a system now. If we start soon, we can effectively provide a standard platform and format for electronic information archives. This will encourage other publishers to join in and make all national resources available in a compatible way. If we do nothing, various small incompatible efforts will be started, as in the early days of railroads when, for example, Ohio had its own track gauge. This is likely to hold up the development of the entire area. Furthermore, if foreign countries take an early lead in this area, it may not make sense for any U.S. databases to try to compete, and we will be dependent on other nations for our technical information.

The overall economic impact of an electronic library would be enormous. It is not possible to state precisely what would happen, any more than one could have predicted thirty years ago the present rise of companies like Mead Data Central or Dialog. However, imagine:

- that a million college students studying science, and representing perhaps $10,000 of tuition each, improve their learning by only one percent; that would save $100M per year.

- that a million engineers in industry, each representing costs of $100,000 per year, are able to work 1% more effectively - saving only two and a half days a year of information-seeking work; that would save $1B per year.

- that ten million workers in manufacturing, producing $100,000 per year each, are able to produce 1% more by employing more advanced technology; this would save $10B per year.

The United States must develop such a basic technological library; the implications of not doing so imply an inability to effectively compete in the information and knowledge areas.

In summary, we recommend that the NSF establish a national initiative to build a basic scientific and technological library for availability on the NREN. We also recommend that the NSF provide for research in the exploitation of this library. This would support U.S. industry, education, and research by providing rapid access to current information throughout our country.

Chapter 1, Section B: Workshop Report

Workshop Report

Edited by

Jitender S. Deogun, Edward Fox, and Vijay V. Raghavan

Executive Summary

In recent years, the Division of Information, Robotics and Intelligent systems (IRIS) at the National Science Foundation (NSF) has sponsored a number of workshops to identify future directions for research in a number of disciplines within Computer and Information Sciences. The Workshop on Future Directions in Text Analysis, Retrieval and Understanding was supported by the Database & Expert Systems Program of the NSF in the same spirit and was held on October 10-11, 1991 in Chicago just before the 14th International Conference on Research and Development in Information Retrieval. The focus of the Chicago workshop was on documents and text, user models and interfaces, and intelligent document management systems. The impact of these research activities, in a 5-10 year period, will be the development of document management systems and tools to: serve a broad class of applications and user types; support advanced approaches to searching and browsing; facilitate the generation of multimedia educational products; solve the cost and space problems of conventional libraries; and scientifically compare document management systems that vary in terms of functions, efficiency, and effectiveness.

Introduction

On October 10-11, 1991, just prior to the 14th International Conference on Research and Development in Information Retrieval, 30 researchers representing academia, industry and government met in Chicago for an NSF sponsored workshop dealing with the broad areas of information retrieval, document analysis, text understanding, and intelligent document management systems. This workshop was unique in its focus on information as opposed to data, on supporting humans' fundamental needs to organize document databases and find documents in terms of their information content, and on the use of language processing and knowledge representation techniques to help in these processes. This report summarizes the discussion, highlights the research issues identified, and makes recommendations regarding the NSF's future involvement in the area of Information Resources and Document Management.

The remainder of the workshop report (i.e., Part I) is divided into two sections and four appendices. In Section 2 we describe the main research areas that are relevant to national needs and thus are recommended as candidates for sponsored research. In Section 3 specific recommendations and concluding remarks are presented. Appendix A.1 provides an overview of related NSF sponsored workshops and Appendix A.2 describes the relationships of this report's recommendations to the HPCC effort and the NREN. Appendix A.3 summarizes the accomplishments of the last two decades in Information Retrieval and related disciplines and Appendix A.4 identifies emerging technologies that are relevant to the information industry.

Further Background

In 1945, Vannevar Bush, President Roosevelt's science advisor, who played a key role in the development of the National Science Foundation, wrote of the need to develop and use tools to manage mankind's accumulated knowledge. [7] Although Bush's 'memex' idea didn't turn out to be of much practical value, his vision inspired many researchers in the areas of information retrieval, information science, and hypertext. As a result, in part, of his insight, the first text retrieval systems were developed in the 1950s.

The NSF was given a broad directive in its enabling act of 1950 ''to foster the interchange of scientific information among scientists.'' The Office of Science Information Service (OSIS) at the NSF was responsible for providing research funding toward improved methods and techniques for handling scientific information, including mechanized systems, through the Documentation Research Program. The activities of the program was expanded considerably around 1956 under the directorship of Mrs. Helen Brownson. The National Defense Education Act of 1958 directed the foundation to undertake projects of a fundamental or general nature that may produce new insights, knowledge, or techniques applicable to information systems and services. [6] The projects supported at this time were featured in the semiannual reports on *Current Research and Development in Scientific Documentation*. [30]

During those early days, a number of conferences were sponsored by the NSF to chart a plan for research activities in the field of scientific documentation. National publicity prior to and following a conference on *Coordination of scientific and technical information relating to federal R&D activities* (Cleveland, February 3-4, 1958) attracted the attention of Senator Hubert H. Humphrey, who then arranged for the subcommittee of Senate Committee on Government Operations that he chaired to hold hearings on this topic. [44] An NSF sponsored conference, held in Washington, D.C. (October 2-3, 1964), was intended to encourage research on *Evaluation of Document Searching Systems and Procedures*. Active research programs at the Chemical Abstracts Service and the American Society of Physics were supported in large part by the NSF. OSIS supported research activities in information retrieval not only in the U.S., but on a worldwide basis. For example, the funds for the famous Aslib-Cranfield project was provided by the NSF to Aslib in England and, in turn, Cleverdon received the funds to execute the project. Agencies such as the Ford Foundation as well as the American Society of Metals initiated and supported many studies of library problems. [26] There is a general consensus among those doing research on scientific documentation at that time that ‘‘the 1960s were really a good time for IR research.’’ In the 1970s, the significance of the area was further emphasized at the NSF by the establishment of, for example, the Division for Information Science and Technology. In the current structure at the NSF, IR and related areas come under the Database and Expert System Program (Director: Dr. M. Zemankova), which is located within the IRIS Division of the Directorate for Computer & Information Science & Engineering.

That research led to the growth of a vast industry, where the current systems support searching for relevant items in large collections, and where the items may correspond to library catalog entries, bibliographic records, citations with abstracts and/or keywords, or full-text documents. In addition to searching, users can browse the documents or indexes. Some classes of retrieval systems, such as those for Selective Dissemination of Information (SDI), route new documents to those users who have filed a profile that defines their information needs. These alert systems send notifications or documents to interested parties, upon the arrival of news or messages. Yet another class of retrieval systems is based on employing the, already mature, Database Management System (DBMS) technology. In these cases, a two step approach may be appropriate, where a *fact extraction system* can fill a database automatically from textual descriptions (such as from a patient record in a hospital) and then a database of these extracted facts is organized for retrieval via a DBMS. Any of the above mentioned classes of retrieval systems may be used either directly by users, or in connection with embedded applications. Important domains of applicability include: legal, industrial and academic research, libraries, consumer services, news, publishing, financial, marketing, and military logistics and intelligence.

Clearly, the support of the NSF, industries and professional societies, from the 1960s through the 1980s, provided the impetus for the development of a multi-billion dollar information industry, which was founded on early research findings. More recently, as the result of keen competition for limited research dollars as well as changes in the responsibility of the NSF vis-a-vis Federal Science & Technology Information activities, it appears that the support for basic research in this area is not at the levels that it previously was. This workshop identifies new research directions and proposes that the NSF supports a bold new research initiative in the area of managing information resources and document databases and pave the way for a National Electronic Library of the 21st Century.

Impact of the Proposed Research

In keeping with these challenges and opportunities that the information industry will face in the 1990s and beyond, we propose a broad program of research be initiated that will, in the next 5 to 10 years, lead to the development of:

- query languages and interfaces to directly serve the broad class of users in our nation, thus making them want to explore various information resources including a National Electronic Library;
- methods to adapt Information Resources and Document Management Systems (IRDMS) to a spectrum of requests, from those wanting a few of the most relevant passages to others calling for a ranked list of all related text materials;

- alert systems that stimulate action by users, through providing current awareness or notification of new and useful text or extracted facts that have just become available;

- IRDMSs that support integrated keyword, concept, natural language, citation, and visualization approaches to search and browsing;

- capabilities to clearly show the interconnections and inter-relationships between raw data, primary research, commentary, and secondary sources, as well as other types of metadata;

- electronic approaches to solve the cost and space problems of current libraries, by bringing together heterogeneous sources of information, and for streamlining the confusing and ineffective world of scientific publishing;

- procedures to scientifically compare IRDMSs that vary in terms of functions, efficiency, and effectiveness; and

- techniques to prepare multimedia educational materials in a cost-effective fashion by drawing upon information in document databases.

Research Agenda

Given the importance of the need for having access to information in our research, education, technology transfer, and other endeavors, it is crucial that a coherent plan emerge for NSF funding of basic research in this area. We call for a significant increase in the level of support for basic research efforts, believing it essential if we are to succeed in the goals of the High Performance Computing and Communications (HPCC) initiative, remain competitive in the world marketplace and improve our educational system to where it becomes more effective and affordable.

To ensure that sponsored research is relevant to national needs, we call for the new funding to be in the context of a Nationwide Electronic Library in Science, Engineering and Technology. In particular, we recommend research in the following areas.

Retrieval Theories and Models

As in any scientific discipline, new theories and models emerge to accommodate new observations and approaches. Thus, formal theories and models are required that address the following key issues and related questions:

- IR is fundamentally an interactive process, where a user's information need or the knowledge state prompts some type of dialog. What are suitable ways to model the processes of browsing and query formulation?

- The process is iterative, where the computer and human provide each other feedback of various types. What methods of learning should be employed in order to improve system performance relative to future interactions?

- Uncertainty is a fundamental issue, which must relate the degree of clarity of an information need, the expressive power of the (formal) query language, weights used in representations of queries and/or documents, ambiguity and relevance. Fuzzy set theory, probability theory, and other approaches, which facilitate the handling of possibly conflicting evidence, have been applied and each seem to have some validity. How can one choose one of these alternatives for a specific problem, or how can they be improved or combined with each other?

- A number of types of entities are involved, including documents, terms, and concepts. In certain applications there are others, such as geographic entities, graphical illustrations, historical dates, and time-lines. How can systems cope with such a variety of entities?

- Documents are in reality two dimensional objects, although in many systems they may be handled as a stream of characters. Thus, in decisions made about relevance, an understanding must be developed of the role of structure and context of document components, both physical (e.g., a component's coordinate position in the bit map image of a page) and logical (chapters, sections, section headings, etc.). For example, how can documents that have sections and subsections at various levels be marked in order to make explicit the semantic connotations of the position of headings?

- Techniques for matching and determining similarity between texts or portions of texts are of particular interest. An understanding of the distinctions between retrieval of sets and ranking of results should be further developed. Given the long standing experience of the retrieval community with such matching problems, it is of interest to extend these to "emerging" applications. For example, how can the key concepts used for text matching apply to important problems of genome mapping and chemical structure matching? What is required to assess the ability of text matching methods in "Functioning as intellectual amplifiers to detect similarities and differences in molecules whose size and complexity are too vast for the unaided human mind ..." [Grand Challenges, p. 40]?

- IR models should be developed to predict the relevance, relative to some user need, of newly acquired documents and the behavior of users, depending on their educational background, level of experience, etc. In particular, they should lead to accurate predictions in new domains. IR system models may involve concepts from object-oriented databases or file structures (e.g., inverted files, signature files), and must be applicable for efficient real-time implementations and when scaling up to large environments. Performance issues relate, and depend on data and file structures, as well as searching and clustering algorithms. How can current models be integrated and generalized to carry-over, from information retrieval to more general problems of document management?

Structures of Documents and Text

What is meant by a document, and which view (form) of document we adopt, are at the core of this field. How can we formalize these structural aspects?

- Instead of paper or sound waves or light, a digital representation of documents and their components is used. What are the effects of this conversion?

- Given documents that have multiple media types, with links and with timing and sequencing and synchronization, how can they be indexed? What attributes or features should be considered?

- Documents include data, maps, tables, images, chemical symbols, equations, especially when part of knowledge repositories, in general, and scientific databases, in particular. How do these relate, logically as well as physically? How can we improve the "development of scientific databases; and distribution of integrated system building kits and toolsets."?[Grand Challenges, p. 21]

- Document structure may be arbitrarily complex, though regularities are captured with SGML, HyTime, and other markup schemes. How can structure be used to advantage? How can passage retrieval be defined and improved in quality? How can automatic document analysis methods identify logical structure, and use that to help improve subject indexing?

- Issues relating to the relationships, groupings, or links that may be provided between the types and instances of the entities are critical. For example, a user may express a preference among retrieved documents, a thesaurus indicates various types of semantic relationships between concepts, concept networks frequently include domain specific relation types, a classification system will impose committee defined groupings, indexing adds associations between documents and the indexing vocabulary, citation data imposes a variety of link types based on scholarly studies, statistical dependencies may be estimated to quantify the degree of association, and clustering will impose some grouping or organization based on feature sets. In hypertext and hypermedia, unlike the case in text and business databases, the meaning of links is not unique. What kinds of theories can aid with the modeling of links in hyperbases, where not only multiple media, but also multiple link types, are involved? Further, what are the trade-offs between links specified in a database schema versus those whose existence is only known at the time of navigation?

User Models and Interfaces

Since IRDMSs are a type of interactive system, what results can we apply from the area of human-computer interaction to improve these systems and their performance? Key issues in this regard include:

- How do we relate the user's world (of people, places, events, and things) to the worlds of information objects, in particular: terms, concepts, classification systems, and documents?

- How can we design and understand various types of systems, where some operations occur autonomously (e.g., automatic indexing of documents), and others are fundamentally interactive (e.g., browsing)? How do those parts of systems relate?

- What are the benefits of graphical displays? How far can we push fast browsing of documents (including multimedia) by using advanced presentation methods? To what extent can graphical interfaces compensate for difficulties that seem inherent in one-dimensional query formulation approaches?

- How can we perform task analysis to most efficiently gather data needed for the system design? How can we really specify, design, and develop user-centered interfaces and systems? How do we deal with individualized handling of user needs, communication of those needs over time, and the complexity of information seeking behaviors? What approaches are needed to model problem solving that takes place over long periods of time including a sequence of sessions, where multiple related or un-related problems may be confronted in an interleaved fashion? How do we integrate users' access of reference works, passages, and other document components?

- What types of user models can be built and maintained with sufficient reliability to facilitate system tailoring or adaptation? ([9]: pp 75-79) What are the relative importances of facilities for browsing and exploring vs. searching, recognition of words vs. recalling them, and moving about in logical hierarchies? How can our understanding of users be enhanced through an analysis of the use of help and navigation assistance aids?

- What extensions and generalizations must be made when multiple humans and multiple computers all work together, for cooperative problem solving and information interchange? How can we better support: 'knowledge transfer exchange programs at national laboratories, centers, universities, and industry; and software dissemination through national databases and libraries."? [Grand Challenges, p. 21]

- Given these interactive systems, what are the best evaluation methods? What aspects and measures of the process should be recorded?

Intelligent Document Management Systems

Since human intermediaries often provide effective access to information and experts may provide service par excellence, won't approaches for building intelligent systems lead to better information access? Key aspects of this problem follow:

- What basic features should we identify about documents? What statistical knowledge is most important and why? Can we integrate statistical knowledge, the results of natural language processing, and knowledge of a domain and its terminology?

- Depending on whether our goal is finding facts, extracting important features, or summarizing content, what methods are most effective? How can the results be best used for common tasks within document management systems?

- How useful is document understanding to support information retrieval requirements? How accurate, complete and deep should parsing be? What type and level of discourse analysis is of value? Does in-depth text analysis help most if applied to queries, to documents, or to both? How can these be done in a robust way for heterogeneous document collections?

- What have we learned from natural language processing, in English as well as multilingual environments, to help understand documents? How does morphology relate to truncation or syntax relate to phrase identification? What characteristics of corpora are important in determining statistical and linguistic insights? What can be learned from machine-readable dictionaries?

- What media and knowledge representation methods are best suited to various types of retrieval needs? What are the roles of data encoding, frames, semantic networks, rules, neural nets, etc.? How do these support inference and concept-based retrieval? What methods for combining evidence are best suited to each of these contexts?

- Given the vast quantities of documents, how can they be best used to help with automatic knowledge acquisition?

- How can we fully develop the idea of agents or knowbots, so they accurately represent human interests and needs? What knowledge about users, in terms of their information seeking behavior and plans, are needed? What kinds of computer-to-computer protocols are needed to ensure proper, cooperative functioning of agents?

Conclusion and Recommendations

To tie together all of the above research problems, to relate them to important needs, to allow study of very large systems, and to solve many of the needs of our society, we recommend implementation of the plan proposed in the white paper submitted to the NSF for a Nationwide Electronic Library in Science, Engineering and Technology (see Part II). In the following we summarize the key benefits of such a library as well as the proposed initiative to fund basic research in the area of information resources and document database management.

- Researchers will have ready access to at least one very large information system, with at least 100 gigabytes of text, and many gigabytes of multimedia data, to study usability and performance issues, and other topics related to scale.

- Research progress will be speeded up, since researchers will have ready access to repositories (or test beds) for software, text resources, interfaces, and retrieval engines. In particular, re-use of existing techniques and knowledge will be easier.

- With large scale information resources and systems, the true potential of recent technological breakthroughs, such as special-purpose hardware architectures and storage systems, can be studied. As computers move to gigaops and then teraops, scalable parallel designs can be tested and techniques to apply them can be investigated.

- Distributed systems, at the national, state, local, and personal system levels, are being built. Many important issues arising in this context should be considered, regarding effectiveness, mix of local and remote storage options, interoperability, sharing and cooperation, applications to education, and use of agents.

- Human resource and behavioral issues regarding learning should be studied, as the "information engineer" profession emerges and as "knowledge workers" increase in number, in the context of libraries, publishing, research, education, and technology transfer.

References

[1] S.K. Bhatia, J.S. Deogun and V.V. Raghavan, Automatic Rule-Base Generation for User-oriented Information Retrieval, *Proc. of the Fifth International Symposium on Methodologies for Intelligent Systems*, Knoxville, TN., pp 118-125, Oct. 1990.

[2] P. Bollmann, V. V. Raghavan, G.-S. Jung, L.-C. Shu, On Probabilistic Notions of Precision as a Function of Recall, *IP&M*, v 28, no 3, pp 291-315, 1992.

[3] C. Borgman, D. Case, C. Meadow, The Design and Evaluation of A Front-End User Interface for Energy Researchers, *JASIS*, v 40, no 2, pp 99-109, Mar. 1989.

[4] Y. Breitbart, H. Garcia-Molina, W. Litwin, N. Roussopoulos M. Rusinkiewicz, G. Thompson and G. Wiederhold (Eds.), Multidatabases and Semantic Interoperability, Technical Report No. 188-91, Dept. of Computer Science, University of Kentucky, Lexington, KY.

[5] M.W. Bright, A.R. Hurson, S.H. Pakzad, A taxonomy and current issues in multidatabase systems, *IEEE Computer*, v 25, pp 50-60, 1992.

[6] H. L. Brownson, Research on Handling Scientific Information, *Science*, v 132, no 3444, pp 1922- 1931, 30 December 1960.

[7] V. Bush, As We May Think, *Atlantic Monthly*, v 176, pp 101-108, July 1945.

[8] W.B. Croft, Clustering large files of documents using the single-link method, *JASIS*, v 28, pp 341-344, 1977.

[9] J.S. Deogun, ed., Working Proceedings, *Workshop on Future Directions in Text Analysis, Retrieval and Understanding*, Chicago, IL., Oct. 1991, University of Nebraska-Lincoln Press, 1991, 160 pages, (forms Part III of this report).

[10] D.E. Egan, J.R. Remde, T.K.Landauer, C.C. Lochbaum, L.M. Gomez, Acquiring information in books and SuperBooks, *Machine Mediated Learning*, v 3, pp 259-277, 1990.

[11] M. Fischer, J. Fedder, The electronic information industry and forecast: in Europe, North America, and Japan, *On-line Information 90*, pp 515-525, 1990.

[12] R. Fidel, Towards Expert Systems for The Selection of Search Keys, *JASIS*, v 37, no 1, pp 37-44, 1986.

[13] E.A. Fox, G. Nunn, W. Lee, Coefficients for Combining Concept Classes in a Collection, *Proc. 11th Int'l Conf. on Research and Development in Information Retrieval*, Grenoble, France, pp 291-307, June 1988.

[14] E.A. Fox, Advances in Interactive Digital Multimedia Systems, *IEEE Computer*, v 24, no 10, pp 9-21, Oct. 1991.

[15] J.C. French, A.K. Jones and J.L. Pfaltz (Eds.), Scientific Database Management, Technical Report 90-21, Dept. of Computer Science, University of Virginia, Charlottesville, VA.

[16] J.C. French, A.K. Jones and J. L. Pfaltz (Eds.) Scientific Database Management -Panel Reports, Technical Report 90-22, Dept. of Computer Science, University of Virginia, Charlottesville, VA.

[17] G.W. Furnas, T.K. Landauer, L.M. Gomez, S.T. Dumais, The vocabulary problem in human-system communication, *CACM*, v 30, no 11, pp 964-971, 1987.

[18] W. Gardner, The Electronic Archive: Scientific Publishing for the 1990s, *Psychological Science*, v 1, n 6, pp 333-341, Nov. 1990.

[19] R.H. Giering, This is data central, Mead Data Central Inc., Dayton, Ohio, 1972, Report DTN-72-2.

[20] P. Hayes, S. Weinstein, CONSTRUE/TIS: a system for content-based indexing of a database of news stories, *Second Annual Conference on Innovative Applications of Artificial Intelligence*, 1990.

[21] IBM, Stairs VS -- A tool for the end user, IBM Scientific and Cross Industry Center, Uithoorn, Netherlands, 1978.

[22] C. Johnson, DARPA Eyes smart text sifters, *Electronic Engineering Times*, pp 35-39, February 1992.

[23] M. Lesk, The CORE Electronic Chemistry Library, *Proc. of the 1991 ACM SIGIR Conference*, Chicago, IL., Oct. 1991, ACM Press, pp 93-112, 1991.

[24] M. Liberman, Text on Tap: the ACL Data Collection Initiative, *Proc. of a DARPA Workshop on Speech and Natural Language*, Oct. 1989, Morgan Kaufmann, San Mateo, CA., pp 173-188, 1989 (ISBN 1-55860-112-0).

[25] R.S. Marcus, J.F. Reintjes, A Translating Computer Interface for End-User Operation of Heterogeneous Retrieval Systems; Part I: Design, Part II: Evaluations, *JASIS*, v 32, no 4, pp 286-303, pp 304-317, July 1981.

[26] B.P. McCune, R.M. Tong, J.S. Dean, D.G. Shapiro, RUBRIC: A system for rule-based information retrieval, *IEEE Trans. Software Engineering*, v SE-11, no 9, pp 939-945, 1985.

[27] E.J. McKinin, M.E. Sievert, E.D. Johnson, J.A. Mitchell, The Medline/full-text research project, *JASIS*, v 42, no 4, pp 297-307, 1991.

[28] I.A. McLeod, R.G. Crawford, Document retrieval as a database application, *Information Technology, Research and Development*, v 2, no 1, pp 43-60, 1983.

[29] J. Nielsen, *Hypertext & Hypermedia*, Academic Press, Inc., 1990, Boston, MA.

[30] NSF, *Current Research and Development in Scientific Documentation*, Rept. No. 6, Washington, D.C., 1960.

[31] A. Okerson, K. Stubbs, The Library "Doomsday Machine", *Publishers Weekly*, pp 36-37, February 8, 1991.

[32] A. Prozes, The Electronic Information Age, *Business Quarterly*, v 55, pp 80-84, 1990.

[33] S.E. Robertson, K. Sparck-Jones, Relevance weighting of search terms, *JASIS*, v 27, pp 129-146, 1976.

[34] G. Salton, *The SMART Retrieval System -- Experiments in Automatic Document Processing*, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[35] G. Salton, A blueprint for automatic indexing, *ACM SIGIR Forum*, v 16, pp 22-38, Fall 1981.

[36] G. Salton, Developments in Automatic Information Retrieval, *Science*, v 253, pp 974-980, August 1991.

[37] G. Salton, E. Fox, H. Wu, Extended Boolean Information Retrieval, *CACM*, v 26, pp 1022-1036, 1983.

[38] T. Saracevic, P. Kantor, A. Chamis, D. Trivison. A Study of Information Seeking and Retrieving. I. Background and Methodology, *JASIS*, v 39, no 3, pp 161-176, May 1988.

[39] H.-J. Schek, P. Pistor, Data structures for an integrated database management and information retrieval system, *Proc. VLDB Conference*, Mexico City, pp 197-207, 1982.

[40] A. Silberschatz, M. Stonebraker and J.D. Ullman (Eds.), Database Systems: Achievements and Opportunities, Technical Report TR-90-22, Dept. of Computer Sciences, The University of Texas at Austin, Austin, TX.

[41] A.F. Smeaton, Information Retrieval and Natural Language Processing, in: *Prospects for Intelligent Retrieval, Informatics 10* (Ed. K.P. Jones), pp 1-14, 1990, Aslib, London.

[42] C. Stanfill, R. Thau, Information retrieval on the Connection Machine, *IP&M*, v 27, no 4, pp 285-310, 1991.

[43] H. Turtle, W.B. Croft, Evaluation of an Inference Network-Based Retrieval Model, *ACM Trans. on Inf. Systems*, v 9, no 3, pp 187-222, July 1991.

[44] U.S. Congress Senate, *Documentation, Indexing and Retrieval of Scientific Information*, Senate Committee on Government Operations, Senate Doc. 113: 1960.

[45] C.J. Van Rijsbergen, Foundations of evaluation, *Journal of Documentation*, v 30, pp 365-373, 1974.

[46] E.M. Voorhees, *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval*, Ph. D. Dissertation, Cornell University, Ithaca, NY, 1986.