Electronic Reference Systems in the Year 2000: The Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992

by Lee A. Gladwin[1], Center for Electronic Records (NNX)
National Archives and Records Administration, Washington, DC 20408

"I'm drowning in open source information!" is the cry of intelligence analysts, a cry not unfamiliar to many other researchers. How does one navigate through a turbulent paper sea? Confronted with a wealth of new open source material (e.g. newspapers, television, technical journals, wire service bulletins, etc.), analysts, who should be interpreting incoming information, currently spend most of their time either sifting and reading documents or writing reports based upon them. Contributing to their problems are those of inefficient technological transfer, shrinking resources (funds and people), overlapping R&D efforts, and the lack of system integration. Unlike most researchers, however, intelligence community has an organization, the Advanced Information Processing & Analysis Steering Group (AIPASG), through which to present its needs to potential contractors and enough funding to attract bidders. AIPASG held its annual meeting March 24 - 26, 1992 in Reston, Virginia.

As with other researchers, analysts need assistance in scanning the data, selecting and organizing documents relevant to their problem, and printing the results. What they do not need is to spend time battling a recalcitrant retrieval system, reading system manuals, attending software workshops, or talking with customer service about software problems.

Through panel presentations, contractors were acquainted with technical developments in the five areas of need identified by AIPASG:

- Intuitive user interfaces
- Document processing, organization and management
- Transparent access to multiple data bases and sources
- Collaborative communications
- Automated data understanding

The first need is for a powerful but easily used interface which does not require exhaustive efforts to accomplish simple procedures. Any system should be developed in close consultation with the users, not in a vacuum.

Document processing addresses the need to organize, sort and link documents relevant to an intelligence problem before routing them to the analyst studying those problems. Central to the solution of the problem of coping with massive amounts of material is to shift the focus from document retrieval to managing discrete pieces of information, using visual indicators to point to other possibly relevant sources.

Transparent access to multiple databases addresses the need to know what data is out there and how to retrieve it.

Collaborative communications concerns the institutional barriers to sharing information; i.e., security, organizational territoriality, etc.

Symposium topics addressed these five needs and such key technologies as continuous speech recognition, natural language and graphical user interfaces, on-line tutors with user modelling capabilities, optical character recognition, automated data extraction, and multiple database correlation. Some of the technologies relating to this future system are described in the concluding portion of this report.

Natural Language/Text Processing

Papers presented in this session addressed the needs to translate text from foreign languages and then extract information and present it to the user in a meaningful manner. Programs were described for translating news items written in Spanish and Japanese, parsing the text syntactically and semantically, and displaying information in an on-screen template. Adrian Kleiboemer, MITRE, called for creation of reusable environments which could be ported easily to any number of applications without having to build a new natural language frontend (NLF) for every new application as is currently being done.

Natural language frontends are currently available for searching large databases without forcing the user to learn SQL. Natural Language Inc. developed a frontend for ORACLE which takes short phrases and even sentences, parses them into SQL statements, and runs them against the database. They may be used in conjunction with graphical user interfaces (GUIs).

Optical Characters Recognition and Neural Networks

The CIA currently is engaged in a five - six year research program aimed at translating source documents, in varying conditions and all languages, into machine-readable format. There are two forms of OCR enhancement: Digital and repair. Digital enhancement clarifies the image using bit mapping and a gray scale. Since it simply prints what is there, letters may be broken or run together. Digital enhancement cannot recognize letters or words. Repair enhancement techniques seek to reconstruct letters and words from faded, damaged or crooked images (i.e., paper orientation). An IBM program was described which uses neural networks to segment pages into regions (e.g., return address, recipient address, stamp, logo, signature), identifies and classifies these segments, and deduces whether the document is a letter, form, article, etc. Neural networks (computer simulated biological neurons) are used in pattern recognition tasks to identify printed or written text images. Applications are currently underway at the US Post Office to read handwritten addresses.

Document Processing, Organization & Management

Papers presented in this session dealt with the "derivation and use of statistical procedures for retrieval and data extraction". Richard M. Tong presented a paper entitled "Automatic Document Retrieval Using CART [Classification and Regression Tree]" which classifies and retrieves documents containing at least one of 15 specific sub-concepts of "civil unrest" e.g., "labor union". Their initial results show better retrieval results for concept-based search than by using standard key word searches in a test involving one-thousand news items. Relevant to this finding was a remark made by Paul Thompson, an attendee, who observed that Boolean or probabilistic ranking approaches to document retrieval are insufficient to assure a document's relevance. Simply adding terms to an SQL query only serves to increase errors.

Data Bases & Information Retrieval

Large scale information retrieval (LSIR) addresses the need to retrieve information from "data sources" that are complex and distributed globally or through different departments of an organization. Potential technologies for dealing with an "indexless encyclopedia" are object-oriented databases, hypermedia, natural language processing, parallel processing, expert systems, and information visualization.

In addition to meeting the five needs, it was suggested that intelligent user interfaces be developed which could model user expertise, and, in the event of error, infer what the researcher was attempting to do and provide greater assistance in information extraction; i.e., an electronic reference guide sensitive to the researcher's facial expressions and body language.

While this electronic reference guide is not yet fully integrated, many of the components are under development and were discussed at the symposium. Some day an electronic reference guide will help researchers navigate their ways through text, graphics, art, musical recordings, still and motion pictures in search of information relevant to the problem at hand. Undiscussed were questions about the implications of this technology for researchers, research methodology and the reference room in the year 2000. Perhaps now would be a good time to begin thinking about these implications.

[1] Article based on notes taken at The Symposium on Advanced Information Processing & Analysis, held in Reston, Virginia, March 24 - 26, 1992.