Lost in Space? An Intelligent Retrieval Solution

In the early part of this century, the poet, dramatist, and critic T. S. Eliot noted how fast new books were being published. He complained that works of literature were printed faster than their merit could be evaluated. In today's world of nonfiction, a similar phenomenon is occurring: So much information is available that its usefulness is doubtful.

In technical and scientific fields, the information explosion has magnified the need for an effective means by which multigigabyte text bases can be searched. Laboratory notes, research documentation, and technical reports represent a critical intellectual resource of possibly incalculable value—a value that is dependent on whether someone who could make use of the resource can find what they are looking for when it could solve a problem, answer a question, provide a new insight, or analyze a pattern of events.

An information management project at Chemical Warfare/Chemical Biological Defense Information Analysis Center (CBIAC) at Aberdeen Proving Ground, Md., provides a case study of a means to handle mission-specific documentation. The CBIAC is a DOD-chartered IAC operated by Batelle Memorial Institute (Edgewood, Md.) under the auspices of the Defense Technical Information Center. A key objective of the center is to use state-of-the-art technology to help its clients overcome information overload, and increase document handling efficiency, while keeping costs competitive.

One of CBIAC's missions is to operate an "information refinery" for federal government personnel, contractors, and university researchers who constitute the chemical and biological (CB) community. CBIAC's task is to provide timely and accurate answers to inquiries from that user community. Typical requests are for reference information or materials and design recommendations that would ensure equipment survivability in potentially contaminated environments.

These requests are handled by a technical staff that has highly developed expertise in CB-related technology. The problem CBIAC faced was that it had to manage more than 14,000 documents with more than 23,000 citations in an on-line database. CBIAC personnel were annually processing hundreds of inquiries, conducting in-depth internal studies, and manually indexing thousands of documents. Therefore, CBIAC reasoned, a technology that could enhance the productivity of these highly skilled people would greatly reduce operating costs and improve responsiveness to its users.

Because the mission of the CBIAC has always been information-intensive, its technical personnel was very familiar with searching on-line document collections. Through this experience, the staff had identified shortcomings inherent in using traditional key-word search techniques (Boolean and string-matching). Particularly, these search technologies could not respond to queries about an idea or subject; namely, those for which the user could not specify a particular word or words, author's name, or date of publication. Thus, employees found that traditional systems require an unrealistic *a priori* knowledge of the exact words used by authors in writing about a particular concept or subject. Furthermore, they realized that Boolean-based systems typically miss more than half of the relevant information in the document base, a result that is increasingly supported by research in text retrieval.

In response to the requirement, Battelle, which serves industry and government by developing, commercializing, and managing technology, created an internal research

Trying to locate information in large databases can be as confusing as being lost in space. One company has found their solution to the information retrieval problem: the Intelligent Text Management System

By Donald McGonigle and LeRoy Golly, Jr.

and development (IR&D) project unit to develop a system for information handling that would allow rapid access to relevant text. The primary task of this unit centered on how to retrieve conceptual information. The IR&D objectives were based on the premise that an intelligent search capability for text databases would greatly contribute to the total mission of the CBIAC.

The unit was given the task of identifying an advanced-retrieval system that maximized research effectiveness. Many of the other needs identified in support of the CBIAC mission were typical of any organization involved in the information-processing field. As a technical information center, the collection, processing, storage, manipulation, and distribution of information were a large portion of its work. The final set of system requirements included:

- Improving access to information and enhancing technical research quality
- Providing the capability to process large full-text information sources
- Integrating an imaging system (document conversion) capability
- Hosting new tools within existing retrieval applications
- Providing automated document indexing.

In the first step of the IR&D project, Battelle identified commercially available products that came the closest to meeting these requirements. (These products could be modified to incorporate additional features that were required by the project.) In its search, Battelle identified only one product that had the technology that was needed (see sidebar)—the Intelligent Text Management System (ITMS), a product of Information Access Systems Inc. (IAS) of Boulder, Colo. ITMS is based on IAS's J-SPACE technology, which is a human judgment simulation technique.

The system intelligence, referred to as a Judgment Base, is built before document input by experts in a particular field. A Judgment Base consists of a series of subject matter relevance ratings between a set of subdomain areas and system vocabulary terms. On input, then, the ITMS classifies (rather than indexes) the text segments by subject, calculates a numerical value that represents the subject matter of the text, and places that numerical value in an *N*-dimensional space, each dimension of which is a subdomain of the Judgment Base (Figure 1).

Retrieval is based on an overall "understanding" of the subject matter of the document base it is managing. In response to a query, the ITMS can then select documents for retrieval that are conceptually similar and list them in order of similarity.

Since the technology used by ITMS is radically different from other retrieval tools, it was difficult to compare the product's features with traditional methods. Therefore, the staff chose to use the IAS Application Prototyping System (APS) so that they could become familiar with ITMS before it was used in a production environment. The APS includes a five-day seminar in judgment base development and a license to an ITMS version that is limited to a 1.5MB document base. The seminar teaches the conceptual foundations of the technology and includes exercises in selecting terms and subdomains as well as the building and quality review of the prototype Judgment Base.

The unit saw several advantages that the ITMS solution could bring to CBIAC. First, the system could preserve the knowledge of Battelle experts. The five people chosen for the project collectively represented more than 200 years of experience in defense systems. Battelle thought that preserving such knowledge alone would be valuable, particu-

larly later when it would not have direct access to these highly skilled personnel.

A second benefit, which was closer to the task at hand, was that ITMS supports retrieval based on subject matter instead of relying on a collection of key words. The advantage of this system was that it offered direct support for the user, unlike traditional systems that are supported by librarians or other information intermediaries. The user's task is to formulate the subject matter or conceptual specification as completely as possible using the ITMS conversational language query capability. (A traditional keyword search system requires the user to guess which words to enter with the hope of having the greatest number of useful "hits" while eliminating useless documents.)

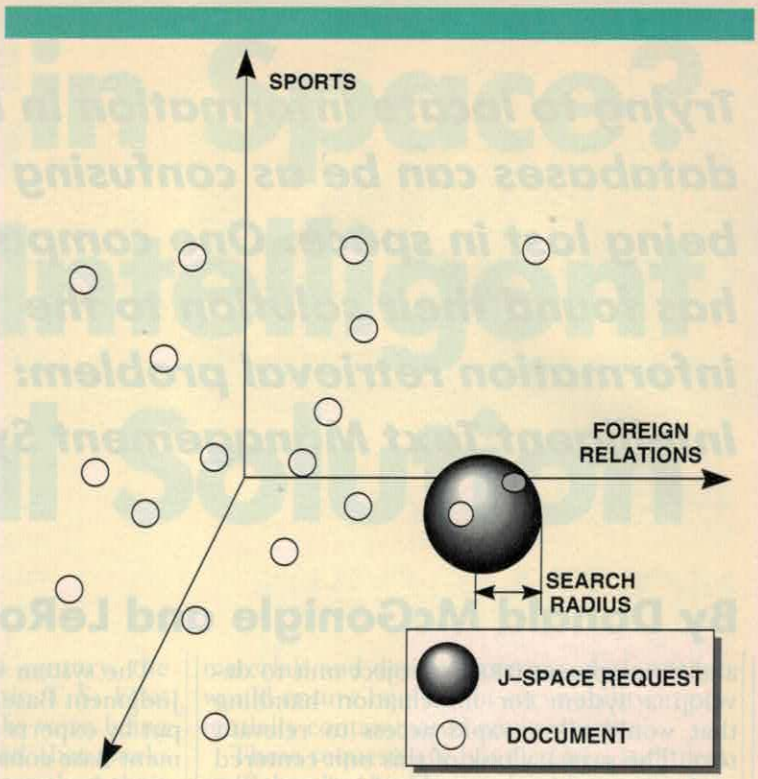
Since ease of use was a major requirement, the unit considered the conversational language query support to be a big advantage. Furthermore, almost no training would be required because users did not have to learn a query language, which meant that the system could be available to many users and that training costs could be kept to a minimum.

Next, the unit foresaw that future systems might need to manage document bases in several subject areas (domains) or at different complexities of Judgment Base development (granularity). The unit considered it a strong plus that other production ITMS installations manage multiple Judgment Bases.

Finally, ITMS addressed the "integratability" requirement, since the J-SPACE functionality could easily be incorporated into existing systems to preserve existing data structures or provide complementary capability. An API is available to incorporate J-SPACE technology into other database products. Users can have both key-word retrieval and intelligent research support.

However, because the J-SPACE is a new technology, the unit had some reservations. It seemed that the experts required to build the Judgment Base would need to put in a lot of initial labor, although the benefit of the time expenditure was not obvious. The salient question was whether the improved retrieval capability provided by the ITMS was significant enough to justify the level of effort. Would the "front-end loading" pay off for the users?

When the staff carried out rapid prototyping to develop a Judgment Base with its five experts, they used their existing document conversion system to scan more than 275 documents (approximately 5,000 pages) as input to the ITMS. Some 30 sample documents were selected and entered in the Vocabulary Development Tool, a module of the ITMS running on a DEC VAX 3100. After the initial Judgment Base instruction was presented by Jalyn Busch, manager of Judg-



Intelligent Text Management System

The Intelligent Text Management System (ITMS), a full text indexing and retrieval system, automates the input, classification, storage, and retrieval of unstructured documents. ITMS integrates three retrieval strategies (formatted field, key word, and "intelligent" retrieval) and distributes, in real-time, both structured and unstructured documents according to individual interests. The Intelligent Text Distribution System (ITDS) module compares incoming text messages to a series of "interest profiles." It then directs the text to users according to the subject matter similarity between the document and the interest profile. ITDS integrates three functions: the monitoring of incoming electronic information, intelligent distribution of documents via a network, and notification of users about relevant messages.

The "intelligence" of ITMS is provided by a knowledge base, referred to as a Judgment Base Module (JBM). The JBM is developed through a human judgment simulation technique, the results of which are incorporated into the machine decision-making process. Subject fields and rated vocabulary terms constitute a JBM. After the initial JBM construction, no human input is required, except to compose requests and perform minimal maintenance.

Users formulate interest profiles according to a specific format: areas of responsibility, general areas of interest, and specific areas of interest. A user may have one or several interest profiles and may submit additional ones as required. The basic unit for retrieval is a document that consists of formatted fields and text. The number of formatted fields and the amount of text that can be managed is limited only by the supporting hardware.

FIGURE 1.
The ITMS at work.

ment Base development at IAS, the experts identified 16 subdomains covering the document base and 275 distinctive terms. The second task of the experts was to judge the conceptual relevance of the terms to the subject matter areas using a second ITMS tool called the Judgment Capture Tool.

At first, the task of making the required 4,000 ratings appeared monumental. However, the experts discovered that, for a given subdomain, many of the terms were not relevant and did not require deliberation. Furthermore, with the Judgment Capture Tool, the process can be accomplished very rapidly (an expert experienced with the system can make up to 1,200 judgments per hour). The result was a matrix that represented the total judgments. The judgment ratings set became the input to the subject matter classification module of the ITMS. The document base of 5,000 pages was then passed against the ITMS and classified for retrieval. On the final day of the prototyping workshop, the ballistics experts entered requests and critiqued the level of intelligence they had given the ITMS. Once the experts began to use the retrieval system, any initial resistance to the process or the system's value disappeared.

The APS had practical as well as intellectual results. First and most important, the staff achieved the goal of building a conceptual retrieval system that would be integrated into the other document management components. The users clearly saw that the intelligent retrieval capability was easier to use and had better retrieval power than traditional key-word systems. Also, users appreciated the complementary role that a key-word system could play when integrated with the J-SPACE functionality of the ITMS.. Before the workshop, members of the unit found it difficult to understand the idea behind the system intelligence (the Judgment Base) and were therefore uncertain as to the soundness of the technology.

Afterwards, the experts who participated in the training understood the straightforward process by which a Judgment Base is built and why it functions as it does. They became aware of the ways in which their judgments affected the quality of the retrieval when they evaluated the results. Having built the prototype, they realized how to avoid the pitfalls when building a production judgment base.

Once the staff was convinced that J-SPACE afforded a significant advance in management of relevant information within a large document base, they considered steps they wanted to take. One, of course, was to build the Judgment Base for the production CB system. Second, given Battelle's in-house expertise in scientific and technical areas, it could construct "signature" Judgment Base modules for its clients and capitalize on its intellectual resources in yet another way.

Further, they reviewed the second major product based on the technology, the Intelligent Text Distribution System (ITDS), and how it would function in tandem with the ITMS, particularly for open source information management. The ITDS selects text from any specified source (wire feeds, disk, OCR input) and distributes the text based on the intellectual content specified by a user.

The overall assessment of CBIAC's IR&D unit was that the APS was worthwhile from many perspectives. First, building intelligence into a text management system makes it easier to use and results in more retrieval power; second, from an investment point of view, the APS permitted a "try before you buy" advantage; and third, preserving experts' intelligence is a resource that can be put to work for other clients. AI

Key-word and formatted-field queries are created with standard Boolean syntax. Intelligent retrieval requests are typed in conversational language by requesting documents about, or similar to, a particular subject. The results of key-word, formatted-field, and intelligent searches are combined and presented to the user in the order of their relevance to a query. Features of ITMS include:

- Queries to the system using conversational language
- Retrieves documents by subject
- Indexes automatically
- Returns documents in the order of relevance to a query
- Integrates traditional and intelligent search technologies
- Indexes and retrieves subdocuments separately
- Submits a full or partial document as a request
- Has user-definable screens
- Highlights relevant document sections for quick reference
- Supports parallel query processing and distributed database access
- Supports on-line context-sensitive help
- Contains system administration utilities
- Supports DECnet and TCP/IP networks
- Supports integration of third-party DBMS
- Provides immediate query feedback
- Selects only that text that is of interest
- Selects and distributes text based on conceptual content
- Uses conversational English interest profiles
- Constantly updates with incoming messages
- Complements existing e-mail systems
- Compatible with ANPA wire service transmission guidelines
- User definable.

Donald McGonigle is the principle research scientist manager of Battelle's Edgewood facility's new technology applications for information refineries. LeRoy Golly, Jr. is the principle research scientist head of ordnance systems and technology at Batelle's Edgewood facility.