DIGITAL LIBRARIES

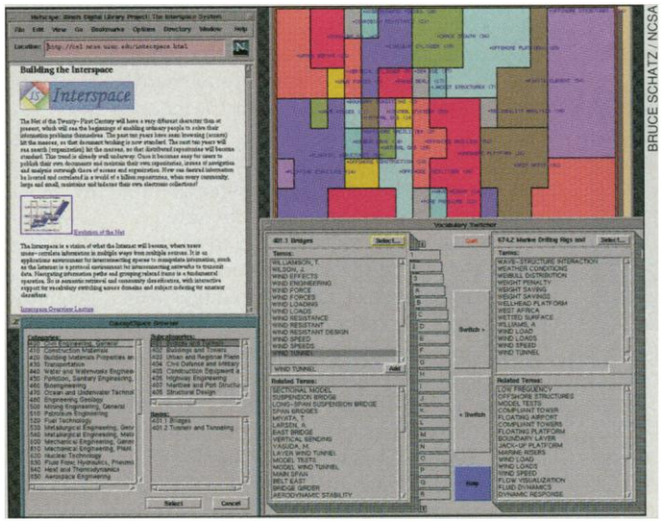
Computation Cracks 'Semantic Barriers' Between Databases

An aeronautical engineer searching a database on ship design for marine counterparts to familiar effects like lift and drag might be stymied. The search would founder on what information scientist Bruce Schatz of the University of Illinois calls the semantic barrier—the differences in vocabulary between fields that make it difficult or impossible to search for a familiar concept in databases from a variety of fields. Now, with little fanfare and no sonic boom, Schatz and Hsinchun Chen of the University of Arizona have opened what they claim is the "first crack in the semantic barrier."

What they've done is lay the groundwork for a system that would provide a user with key words needed to search for information across fields. Their technique is a long way from being practical; for one thing, it only suggests key words for searching new fields, leaving the user to determine which are best. What's more, a test run required one of the largest supercomputer calculations of all time: 10 days on the HP Convex Exemplar supercomputer at the University of Illinois's National Center for Supercomputing Applications (NCSA) earlier this spring. But the computational effort—a statistical analysis of millions of science and engineering abstracts—was worth it, says NCSA director Larry Smarr. The demonstration, reported in the May issue of IEEE's *Computer* magazine, has "the potential to affect virtually every academic field," Smarr says.

In a report last year, the managers of the federal digital libraries project, which aims to develop information repositories that could be accessed and searched via the Internet, identified "semantic interoperability" as one of the "grand challenges" on their agenda. The goal is to develop a system that automatically translates terminology from one field to the next, allowing the user to search by meaning—in effect, to match concepts rather than just matching words. "It's like going to a human librarian who, after hearing what you're interested in, can point you to the best sources," Schatz explains. "But to find something on the Net, in a future world of a billion knowledge repositories, you'd need more reference librarians than there are people, which is why we have to automate the process."

Schatz and Chen aren't the first to try. Even before the burgeoning of on-line databases, researchers in the field of natural language processing tried to create software that could break down sentences following grammatical tenets to discern their meaning, enabling a user to search a database for concepts rather than key words. But these strategies have progressed slowly. Inadequate computers, meanwhile, have held back a simpler approach: inferring meaning from a statistical analysis of word frequency rather than from the structure of sentences. "Now that computers are literally a million times faster than they were 30 years ago," Schatz says, "some of these techniques look feasible for large databases."



Matchmaking. A prototype system for navigating unfamiliar databases allows a user to compare lists of related terms in different fields.

He and Chen decided to put a statistical technique to a test by sifting through 10 million abstracts in 1000 science and engineering fields, sentence by sentence, to see how often any two terms appear together. The computation, carried out over several weekends in March and April, ultimately produced 1000 giant matrices ("concept spaces"), one for each field, representing the "co-occurrence frequency" for every pair of terms in a set of 100,000. Terms found hand in hand presumably are linked to the same underlying concept, so a set of co-occurring terms from one database provides a set of candidate key words for searching an unfamiliar database. "If one term does not pull up the desired documents, you have others you can try," Schatz explains.

Although that may sound modest, other information scientists regard the computation as a significant step forward. "Schatz and Chen have taken theoretical approaches established in the 1960s and '70s and tried them out for the first time on serious, production-sized databases," says Clifford Lynch, director of library automation for the University of California. Yet Schatz concedes that there's still a long way to go before the semantic barrier is truly broken. For one thing, he says, their technique still takes too much computational muscle to really be practical: "We need to get it down to about a day on a big machine like the one Dialog uses, rather than on a supercomputer."

More important, he notes that while the method is useful for "term suggestion," it falls short of actual vocabulary switching: "What we'd like to be able to do is take a term-frequency pattern in one subject, compare that to a frequency pattern in another subject, and then figure out the best terms to search." Doing so will require better "intersection techniques" for finding the commonality between the giant matrices.

Meanwhile, Chen and Schatz have linked with other digital-library researchers at the University of California, Santa Barbara, to run the same type of computation for a database of images rather than words, inspecting thousands of topographic maps and aerial photos to determine the co-occurrence of "textural features" within a given region of a map. The goal is a system that could search databases of images and maps for meaningful features, says Schatz: "Systems available today might be able to show you all the blue things, but what we'd really like is something that could show you all the lakes." The approach they're taking, which is analogous to term suggestion, involves identifying a set of features common to lakes, for example, so that if one of the features fails to locate all the lakes in a given map, another might succeed.

As this work progresses, Rutgers University information scientist Paul Kantor says he keeps hoping that the brute force the technique has required so far may turn out to be unnecessary. "What we hope to find, after the fact, is that some large part of these calculations was not necessary," he says. Schatz agrees. Once simpler algorithms and faster computers make vocabulary switching a routine operation, he says, "scientists will be able to solve problems outside their own narrow specialties, and the walls will come tumbling down."

–Steve Nadis

Steve Nadis is a writer in Cambridge, Massachusetts.