



Searching for Meaning in Databases: Linguistics, Mining, and Context

Is document information ever truly unstructured? Of course not. If it were, documents would be useless. Is most document information highly structured and disciplined? No, again. Most document information is structured, albeit subtly structured. Whether the weak and unimposing structure of HTML, the highly disciplined rubric of SGML, or in the lesser-known middle ground of XML and others, the structure of a document often derives from linguistic conventions or subliminal idioms, specific to culture and language.

If document information critical to your electronic publishing efforts is locked up in databases, you have two problems: first, how to gain ready access to that information; second, how to find information when it is, at best, only subtly structured. Virtually all modern text-retrieval systems can build expensive bridges to database systems, but bridges usually exact tolls, including the following:

- speed because an external retrieval system must interact with, instead of being part of, the database
- space because information is likely to be duplicated locally in the external retrieval system
- support, because a text retrieval system bolted onto your structured data repository requires separate support and learning a new query language.

Every IT organization has SQL experts; how many have full-text retrieval experts? How many, for that matter, have staff linguists?

ConText 2.0, Oracle's new linguistics-based search system, offers one solution to this two-faceted problem. An option initially with Oracle7 Universal Server

and now available as Oracle8 ConText Cartridges, ConText helps users gain easy access to document information within an Oracle database or external sources, including the Web. It also enables users to leverage SQL expertise that their IT organizations already have. With ConText 2.0, Oracle seems well on the way to acquiring the necessary skills for finding meaning within subtly structured document information.

LINGUISTIC SMARTS AND MINING ARTS

Typical content-based retrieval systems assume relevancy by counting words that match queries. If these words are found close together, they are considered even more relevant. However, consider a query to find "Presidential policy regarding sales of Boeing jets to Australia." The following sentence would be considered as highly relevant because of several matches: "On his flight to *Australia*, while strolling down the aisles of the *Boeing 747*, President Clinton discussed his Vatican policy." In fact, you could scramble the words in this sentence and render it meaningless, and a retrieval system without linguistic awareness would consider it equally relevant.

ConText 2.0's linguistic smarts alone make it an impressive and useful product, but in fact the tool offers much more. If building your electronic information product requires you to do some information mining, looking for recurrent summary topics, Oracle can help with that as well. ConText's "Themes" option, rooted in natural language technology, automatically identifies and delivers key themes and theme summaries from searches in large-scale text databases. In essence, ConText builds themes which are the most relevant sentences in a document.

ConText searches are not limited to a single database or even a collection of them; searches can extend to Web sites too. Web-enablement makes the information available from anywhere to anywhere at all times.

Storing document information into databases can be challenging. Documents come in all sizes—from little snippets that fit nicely into text fields of databases to chunks in the thousands of virtual pages. And network performance sometimes demands that database-databases be replicated for fast economical access. To this end, Oracle8 defined a large object type called (appropriately) an LOB. And ConText, as you might expect, supports LOBs in Oracle8.

The new LOB data type extends the potential size of a document to 8GB. Further, LOBs can be stored in-line within a table or out-of-line on separate storage. Keeping track of changes in replicated database-database hybrids is also cheaper and faster. Oracle8 can replicate just the changes and not the entire LOB data.

CONTEXT IN CONTEXT: REAL-WORLD APPLICATIONS

How do ConText customers really use this technology? COM.sortium, a three-year-old developer of Web-based applications, has been an all-Oracle shop for the last year and a half. COM.sortium founder Larry Footer says he picked Oracle originally because his company needed a Web server and browser that could talk to a database management system. Before making COM.sortium "100 percent Oracle," Footer says he had looked at other content-based retrieval systems from Verity and Open Text. His clients needed a way to generate document themes

dynamically, and Footer says he found that in ConText. "We actually built a Java applet to produce an output chart showing relevancy rankings. We are convinced that 90 to 95 percent of the time, these rankings summarize the document meaning perfectly."

COM.sortium clients cite several benefits of using the company's ConText-analyzed databases. The management ranks of pharmaceutical company Pfizer, for example, required faster access to complex information generated by its sales force. Working with COM.sortium and ConText, they developed a Web-enabled Sales Force Automation application that could receive data dynamically and enabled senior managers to use Web browsers to evaluate the status of Pfizer pharmaceutical products sold throughout 1,000 HMOs. Another COM.sortium client, MCI, needed to speed up and improve the quality of its global sales proposal management. Customizing COM.sortium's DocuMine enabled Web-based searching of MCI's heterogeneous document database by theme, whether natively stored

as WordPerfect, Word, Acrobat PDF, HTML, or other formats.

DocuMine's agent feature also stores personal profiles and automatically updates users via email each time an applicable proposal is entered into the repository. Lucent Technologies needed better returns on the \$20 million it spent worldwide on recruiting, so it used COM.sortium's WebRecruiter to manage complex databases of jobs and resumes. This system matches applicants' skill sets with posted job openings. Linguistic searching allows better and faster evaluation of applicant skills and openings in Lucent. Intelligent agents notify recruiters and applicants via email as each step in the recruitment process is completed. According to Reginald Bearfield, Director of Global Recruitment Strategies at Lucent, "WebRecruiter will pay for itself within 45 days."

THE WORD FROM THE ORACLE

Where is Oracle taking ConText in the rest of 1997 and beyond? While Oracle's software schedule commitments remain

unavailable, several initiatives are clearly coming. SGML or XML-tagged text searching is high on the list. Beefing up ConText's international linguistic strength is also very likely. Lastly, it is probably a safe bet that Oracle will open its linguistic framework for third-party developers. Such value-added products could take the form of specialized linguistic products like technical thesauruses.

Applying database technologies to documents may lead to competitive advantage. You may even double the advantage if the system you choose infuses traditional database systems with leading-edge linguistic tools.

Robert J. Boeri and Martin Hensel are columnists for INFORMATION INSIDER. Boeri is Advanced Systems Specialist in the Information Services Division of Factory Mutual Engineering of Norwood, Massachusetts. Hensel is founder of Martin Hensel Corporation, a Newton, Massachusetts-based consulting firm that does SGML-based editorial and production systems for publishers, corporations, interactive services, and composers.

Comments? Email us at: letters@on.orc.com, or check the masthead for other ways to contact us. ■