

# COMPUTERWORLD

THE VOICE OF THE I.S. COMMUNITY

JUNE 16, 2000 VOLUME 16 NUMBER 12 ■ AN LTI PUBLICATION

## Finding a needle in a haystack

*dtSearch helps companies find information in text-based documents quickly and efficiently*

BY CHRIS CONRATH

As we enter the new millennium, there is a lot of talk about how information is the key to success. But with gigabytes becoming terabytes, finding specific data in a specific document is becoming a daunting task best left to machines.

The problem is machines are not as intuitive as humans.

dtSearch Corp., a Bethesda Md.-based software company, has released beta code for version 6.0 of its developer components in its dtSearch product

*With gigabytes becoming terabytes, finding specific data in a specific document is becoming a daunting task*

line. The dtSearch Web and the dtSearch Text Retrieval Engine are designed to allow developers to incorporate text retrieval technology into their own products for the PC, LAN or Internet, and to allow users to access information in a variety of manners, from natural lan-

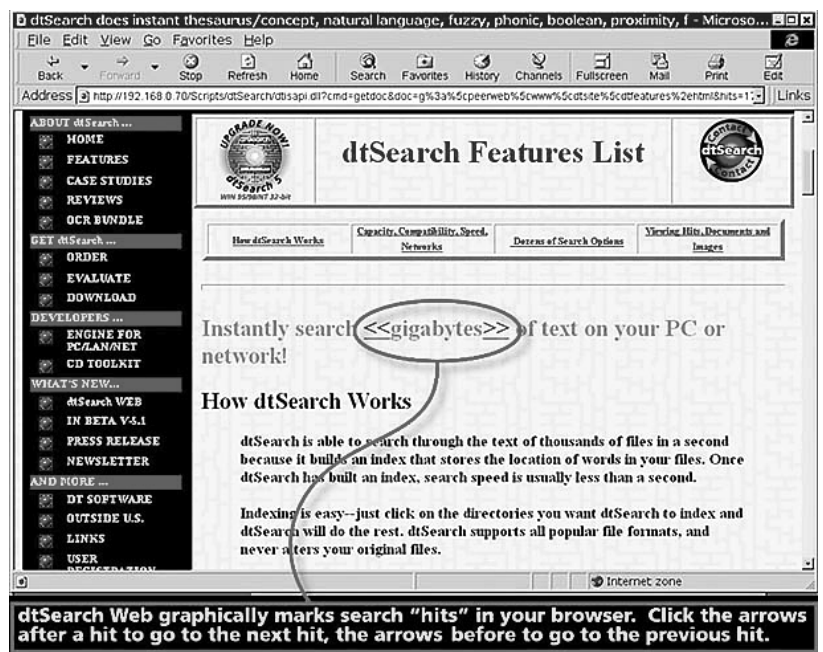
guage to advanced Boolean logic.

The principles behind a good search engine are quite simple, though the technology is not.

“Essentially it goes through every single document and every word and it builds a database that stores the locations of the words in the document collection,” explained David Thede, president of dtSearch.

So if the word “university” occurred seven times in a document, the database would identify its location by its numerical placement in the document. Every character is classified as a letter, space or ignore, the latter neither indexed nor searchable with dtSearch. An example of ignore might be an exclamation point at the end of a sentence, since it is seldom the basis for a text search. dtSearch does, however, allow the addition of punctuation or signs if they are required. Many recruiters at large companies need the plus sign to identify those people with knowledge in C++.

The key to creating an



dtSearch Web has a frames-based user interface for entering queries and navigating search results and retrieving documents.

*The key to creating an efficient text-based search engine is designing specific algorithms for each search technique. Some ... are surprisingly simple ... while others ... are “fiendishly difficult.”*

efficient text-based search engine is designing specific algorithms for each search technique.

Some, like stemming (-ed, -ing) are surprisingly simple, according to Thede, while others, such as query searches, are “fiendishly difficult,” he said.

A query search could be a combination of Boolean and proximity, like searching for A and B but not C (Boolean) where A is the word apple located within five words of pear (proximity).

## FUZZY ALGORITHMS

"Developing algorithms is tricky. If you go down the wrong road you can quickly develop an inefficient way to do it, or you can spend five years struggling with it and not figure out the one way that produces correct results," Theede said.

The fuzzy search algorithm was one of the more difficult to develop, he said. It is designed to find words even if they are misspelled, an option particularly useful to the layperson searching

***The addition of comprehensive XML technology into the engine has put it in position to tap the growing business-to-business e-commerce market.***

for technical information. With more and more medical information on-line, patients will no longer have to know how to spell some of those 20 letter Greek and Latin disease names. Getting close will be good enough. Fuzziness can be set from

one to 10, with 10 casting a wider net than one.

Getting everything to work is a matter of continually testing and refining the algorithms until they work perfectly. If done correctly the net result will appear to be child's play.

"It is not complex and it is easy to learn and to use," said Danielle Raymond, Ottawa-based assistant director of economics and market research, and responsible for IS, for the Dairy Farmers of Canada. She is using dtSearch to index thousands of Excel files that employees can use to search for statistical information like the number of cows in Quebec or the level of cheese production in the Ottawa Valley. "We publish a book only once a year but internally (we) would like to have the information on a regular basis," she said.

## XML ENGINE

The addition of comprehensive XML technology into the engine has put it in position to tap the growing business-to-business e-commerce market.

Jordan Worth, a telecom and internet analyst with IDC Canada in Toronto, agrees with the need to increase XML functionality in search engines. "One of the problems is that

everybody wants different things (when they search)," he said. "I think XML answers a lot of these questions ... [XML] gives you the ability, ultimately, to more clearly define what you want and what you don't want."

Max Rottersman, president of New York-based Reports Automation, is excited about Version 6.0

***For Adam Yahre, CEO of Tampa Bay Systems, a Web-based document management company, the selling point was the search engine's ability to cover the gamut of features customers would want when they search.***

because of these XML capabilities. The company, which automates business reports, has clients who want to do on-line full text searching of their maintenance reports. Without XML, searches would be too broad.

"It allows you to use quantified database-type structures

on unstructured documents."

For Adam Yahre, CEO of Tampa Bay Systems, a Web-based document management company, the selling point was the search engine's ability to cover the gamut of features customers would want when they search. "We needed something to search across the board, in as many general ways as possible so our clients would have something that is flexible," he said.

---

## CONTACT

More information is at  
[www.dtsearch.com](http://www.dtsearch.com)  
or at 1-800-483-4637.

---

Theede says the market for the engine is quite varied, from government agencies with terabytes of data to hospitals and law firms who need to access all information pertinent to a case, quickly and accurately.

The product is accessible to developers using ActiveX, C/C++, Visual Basic, VBScript, Active Server Pages (ASP) and Delphi, and includes sample source code in all supported languages.

Version 6.0 will be available for the Linux platform as well as Windows NT, for US\$999 for use on a single LAN server.