APPLIED INTELLIGENCE

Text Indexing Finally Organizing Corporate Paper Trail

JAMES MARTIN

This is the fifth article in a series on text management and its influence upon the corporate environment.

The efficient storage, indexing and categorizing of text—a process aimed at enhancing search and retrieval—needs to become a priority at corporations. The end result will be to give employees better access to the information required to complete a project.

Typically, text-management applications handle text in one of two ways. An approach prevalent among micro- and LAN-based products is to manipulate text as it resides in its native environment, such as word-processing and spreadsheet formats or as ASCII characters.

The other common text-management style requires all text to be reformatted into a proprietary structure that is dictated by either micro or host software.

For example, a database-management system (DBMS) relies on key fields for access to data in the database. A DBMS automatically generates an index based on the key fields.

With textual information, however, there are no key fields. Therefore, indexing techniques are frequently applied to speed up text retrieval; generally text can be indexed either by keywords or by an inverted list.

With keyword indexing, each document is assigned frequently used words that characterize its content. This technique, which is similar to the subject-heading indexing used by library card catalog to cross-index information, is labor-intensive and requires input from an expert in the domain. Its advantage is that text can be searched easily, but the disadvantage is the appropriate keywords need to be assigned to the text.

Another approach to indexing is inverted indexing, also known as full-text indexing, which employs discrete words to locate data. An inverted index is an alphabetic list of all the words contained in the text collection with the exception of "noise" words such as "of," "and" and "the." The inverted index maintains pointers to every text item that contains a particular word; the word may appear once, twice or in a title or footnote; unfortunately, the index does not recognize the differentiation.

Full-text indexing also has another downfall. It requires users to be fluent in the vocabulary of the textual information being searched, and to understand Boolean logic (AND, OR, NOT and operators). Users also need to determine the relevancy of retrieved text. For example, the following search phrase "(("artificial" OR "machine") AND "intelligence")" would locate documents pertaining to "artificial intelligence" as well as those that might include a reference to "artificial satellites that collect intelligence data for the CIA."

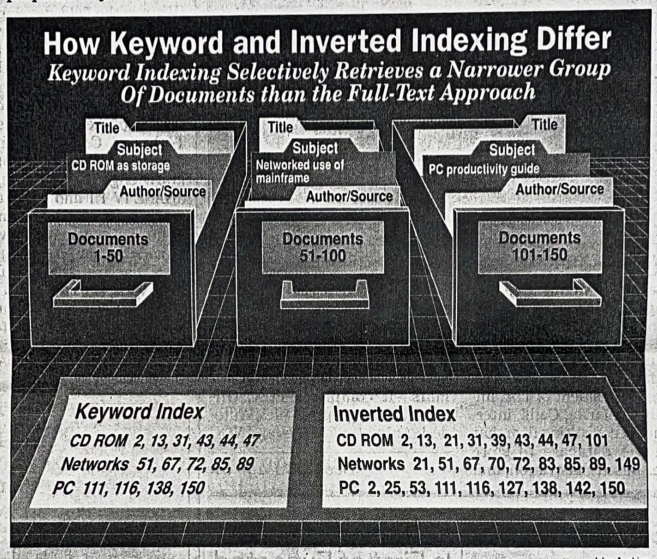
Another obstacle to inverted indexing is the massive storage capacity needed for the text collection, which could be 50 percent to 200 percent more than what was needed for the original text. Vendors of text-management software, however, are employing compression techniques to minimize this overhead.

Inverted indexing does have its benefits, however. Because the index is automatically generated, there is no need for an expert to review each of the items in the text collection.

Regardless of the text-management medium, users need to examine the success factor of a particular search. While speed is an important issue, the relevancy of the returned text items for the user is even more critical. Quickly returning 25 documents that must be reviewed for relevance is not as efficient as returning six documents that all have a high probability of being relevant.

This forces users to word their queries carefully, because the unpredictability of retrieval—facing thousands of documents to review or none at all—can be a source of frustration.

Next week I will discuss how vendors are minimizing user frustrations with text retrieval. ■



The concepts in this article are described in a new volume, Text Management, of The James Martin Report Series. For more information on this volume, call (800) 242-1240 or (617) 639-1958. For information on seminars, contact Technology Transfer Institute, 741 10th St., Santa Monica, Calif. 90402, (213) 394-8305 (in the United States and Canada). In Europe, contact Savant, 2 New St., Carnforth, Lancs., LA5 9BX United Kingdom, (0524) 734 505.

dBASE IV 1.1 \ *Upgrades Targeted*

Continued from Page 65

demonstrated in its pre-beta form. "The Professional Compiler looks promising. What I see and hear is encouraging," said Charles Desjardin, a systems analyst for the University of Arkansas in Little Rock. "Developers have been waiting for this for a long time."

Many corporate evaluation efforts for dBASE IV 1.1 have been short and sweet, with some companies taking only two weeks to a month to finalize their plans and kick off their upgrades.

"Between 40 and 60 percent of our dBASE III Plus users are switching to dBASE IV," said William Dreizler, a systems analyst for Shell Oil Co. in Martinez, Calif.

Also helping to rekindle user interest in Ashton-Tate is the company's plan to offer dBASE IV versions on many platforms, attendees said.

"I think Ashton-Tate will recover its share of the market, especially with the cross-platform strategy," said Bill Waterman, manager of programs and cost estimating at Honeywell Inc., an electronics firm in Plymouth, Minn.

Ashton-Tate has demonstrated dBASE IV versions running on Sun Microsystems Inc.'s workstations, the Macintosh and Digital Equipment Corp.'s VAX/VMS minicomputers. It has promised dBASE IV versions for Xenix, The Santa Cruz Operations Inc.'s Unix, 386 Unix and eventually, Windows and OS/2 Presentation Manager.

Other users were hesitant about the dBASE IV 1.1 upgrade, deterred by Ashton-Tate's record of product delays and development snafus. They are taking a wait-and-see attitude before they commit to any new dBASE platforms.

"We do support 1.1, but we're not recommending that anyone use it yet," said Steve Cook, vice president of Welcom Software Corp., a project-management software developer in Houston. "It's too early, and there's a bit of a bad taste after the 1.0 debacle." ■

DPMI \ *Solving Compatibility Issue*

Continued from Page 65

plications tap the three modes of Windows. Windows 3.0 incorporates support for the preliminary 0.9 version of the specification, and Microsoft has committed to building DPMI 1.0 into the next versions of OS/2 and Windows.

Once the DPMI 1.0 specification is made public this fall, DPMI-compatible DOS-extended applications will start to make their way to market. Along with their ability to fully exploit Windows, these applications will be highly portable and will also run under DOS, OS/2 2.0 and even in the DOS compatibility box of some Unix variants.

"The DPMI standard provides long-term life for the DOS extender and DOS extender companies," said Terence Colligan, president of Rational Systems of Natick, Mass.

Lotus, of Cambridge, Mass., was able to get a jump on its competitors and work with the DPMI 0.9 version of Rational's DOS extender because it owns a significant share of the DOS extender company, Colligan said. He declined to set a date for the public delivery of Rational's DPMI 1.0-compliant product.

While Lotus 1-2-3 3.1 adheres to the 0.9 specification, it should not have compatibility problems with DPMI 1.0 because that version is a superset of 0.9, according to the DPMI committee's plans.

Phar Lap, the other principal player in the DOS extender market, is aiming to ship its DPMI 1.0 DOS extender during the first part of next year, according to Robert Moote, vice president of the Cambridge firm. As an interim solution, the company will ship a DOS extender that is compatible with Microsoft's HIGHMEM.SYS memory driver, which will allow it to run under the 286 standard mode of Windows, he said.

Ergo Computing Inc., a third DOS extender company, intends to release its DPMI 1.0 offering by the end of the year, according to an official at the Peabody, Mass., firm. ■