APPLIED INTELLIGENCE

1990s Increase Demand for Text-Query Improvements

JAMES MARTIN

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

In the 1990s, users will demand that text-management applications meet certain requirements, including friendlier user interfaces and the ability to generate answers that do not require additional verification. They will assume that the information returned will fit well within the parameters of their search request.

Today, however, most text applications do not live up to these expectations. Many employ an indexing technology that uses inverted lists. This approach maintains an alphabetic listing (referred to as a dictionary file) of all the words that comprise a text collection, omitting "noise" words such as "the," "a" and "and."

To search this data pool, users must apply Boolean logic to phrase queries. However, because the queries are not specific, a single search phrase might retrieve unrelated or unnecessary articles.

To minimize obvious mismatches, most text-retrieval software offers a technique known as proximity searching, which lets the user specify the number of words, sentences or paragraphs that can separate individual search terms.

To support proximity searching, a postings file is created in addition to the dictionary file. This file maintains detailed information pertaining to each word's location in the document.

Also, most full-text retrieval software supports wild-card and prefix/suffix queries to help make the search process more effective. Other commonly used techniques include phonetic searching—where, for example, a search on "Mayer" would also locate files with "Meyer"—and the ability to browse the index itself, letting the user gain familiarity with the words that comprise the text collection.

To minimize user frustration with query results, many vendors offer thesaurus capabilities to improve recall and precision in the retrieval process.

A thesaurus can clarify terminology, provide cross-referencing and supply synonyms. As key words are entered, they can be compared to a thesaurus and switched to a preferred term for consistency.

Topic, a text-retrieval product from Verity Inc., is one product that bridges the methods of inverted-file indexing with controlled vocabulary. Users group terms into symbolic concepts and relate these concepts to each other via a graphical representation that can be browsed and assigned a label. These representations, called "topics," are used to search criteria and can be combined using Boolean operations.

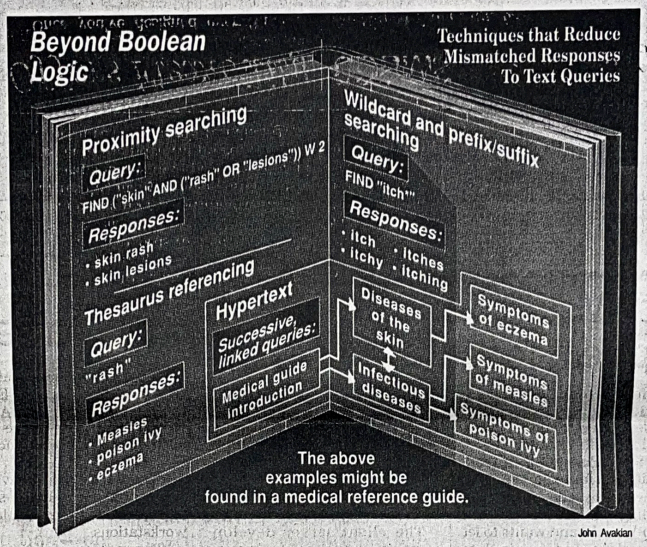
Hypertext is yet another text-management medium. This approach, which links related pieces of text, offers a unique capability—readers can browse through the text collection by following links of interest. Their quest for information can be as specific or as general as they wish.

Hypertext applications require that someone with knowledge of the domain identify the location of the links. In addition, the links must be manually built into the application. They are analogous to GOTO statements in unstructured programming code, with all the attendant maintenance problems.

The assigning of key words, the building of topics and any other labor-intensive front-end work requires a major commitment from an organization. The alternative, inverted-file indexing, can be performed automatically and may be beneficial when a large text collection is involved. However, this approach requires that the user understand the conceptual categories and relationships that exist in the collection.

Ideally, text-management facilities should understand natural language. However, to date, interpreting the meaning of text requires human intervention to supply intuitive and domain knowledge.

Next week I will discuss the role CD ROM plays in text management. ■



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 (617) 639-1958. For information on seminars, contact (in the United States and Canada) Technology Transfer Institute, 741 10th St., Santa Monica, Calif. 90402 (213) 394-8305. In Europe, contact Savant, 2 New St., Carnforth, Lancs., LA5 9BX United Kingdom (0524) 734 505.

SPC \ *InfoAlliance Offers Flexibility*

Continued from Page 63

and menu choices to create different views of the data, regardless of where and in what format it resides on the network, he said.

Coca-Cola plans to tap InfoAlliance to build a nationwide retail survey system that spans multiple LANs in order to help users perform break-down analyses of market data.

With this system, a user at one retail chain, for example, may choose to view sales of a particular Coca-Cola product in one or two regions; InfoAlliance will automatically cull the data from several dissimilar databases.

Later, that same user could review sales of that item in other regions without having to respecify to the InfoAlliance software where the data is located, said Byers.

"People developing [views] won't have to think of everything while they are working on a small part of the application," he said.

"Once you've designed everything to capture data, you can let the end user retrieve data and do the analysis themselves, without needing help from the systems department every time their needs change," according to Bob Holmes, computer technology research analyst for the Southern California Gas Co. in Los Angeles.

This kind of flexibility will boost the efficiency of users, alleviating the need to consult a database programmer whenever new development needs arise, Byers and other InfoAlliance users said.

InfoAlliance, which is the result of SPC's reported $10 million, five-year development effort, will be available for both DOS and Windows by the second quarter of next year.

Pricing for InfoAlliance ranges from $8,500 for a single-server, 10-workstation configuration to $99,500 for a nine-server, 200-workstation version.

SPC, of Mountain View, Calif., can be reached at (415) 335-2082. ■

ODBMS \ *Technology Is Rolling Along*

Continued from Page 63

to its Ontos database along with a new fourth-generation language (4GL) tool.

ODBMS technology, which is currently available from only a handful of players, stores complex data types directly, unlike more traditional relational databases that organize data in tables of rows and columns. This approach, which is well-suited for creating large, multimedia databases of bit-mapped graphics, geometric forms and compound objects, can also serve to boost performance, according to officials of both firms.

Object Design's ObjectStore 1.0 comprises three major components: a database module, the ObjectStore Data Manipulation Language and a mix of development tools, including SchemaDesigner, a graphical database-design tool and code generator, said company officials in Burlington, Mass.

"The reason I need an object-oriented database is that I'm modeling pretty complex objects, and it would be a kludge if it were done in a relational database," said Todd Bashuk, technologist for Southwestern Bell Technology Resources Inc., a division of the telephone company in St. Louis.

Ontologic, also in Burlington, Mass., is tossing its second entry into the ODBMS ring. Its Ontos 2.0 upgrade runs on Sun-3, Sun-4, SPARCstations, Apollo and DECstation workstations, as well as PCs running OS/2.

Ontos is also showing off Ontos Shorthand, a new 4GL option for its database. The 4GL features an interface called Ontos Studio, which provides access to blocks of preprogrammed code for buttons, menus and text boxes. Both options will be available in December.

Pricing for ObjectStore, which is determined by the number of users per site, ranges from $2,000 to $9,000.

Ontologic's Ontos upgrade is free to registered users of the database. Site licenses are available, starting at $1,500 per user.

Object Design can be reached at (617) 270-9797; Ontologic is at (617) 272-7110. ■