High-Speed Text Search Systems and Their Archival Implications

WILLIAM NOLTE

Computing technology has made it possible to create and store ever larger quantities of information. Data bases of every conceivable type are now readily available either to limited, "in house" audiences or to the general public through online services and other shared resources. Retrieving information from these data bases—making it useful—may prove to be more difficult than amassing it in the first place. Much of the archival literature on the subject points, for example, to the problem of maintaining intellectual control over masses of electronic data.

The possibility exists, however, that technology might provide the solution to the problem technology has created. In this instance, the development of high-speed text search (HSTS) systems may represent a breakthrough with significant applications for archivists and others who need to bring electronic retrieval capability into balance with storage capacity.

HSTS systems represent a new development in what has long been a desirable computer application: the ability to search an entire data base, not just indexed terms derived from the data base. In the 1960s and 1970s, several researchers suggested that full-text search would become feasible contingent upon expected developments in computer technology.[1] Later research suggested that full-text search would prove deficient in either recall (the system's ability to retrieve all documents pertinent to a subject) or precision (the system's ability to retrieve only those materials desired). In particular, the results of one experiment in evaluating full-text search questioned the ability of then (1985) available search systems to handle variations in syntax or vocabulary. For example, a query to recover information on a given automobile "accident" would not recover records in which the word "incident" was consistently used instead of accident.[2] The authors of the most pessimistic study of full-text search argued that their results rebutted not just a particular search system, "but the principles on which . . . full-text document retrieval systems are based."[3]

The most common of these systems have involved software indexing, the creation of an "inverted file" of virtually every word appearing in a text. Inverted files work, but only by creating index files "at least as large as the text data base itself." Inverted files are expensive and waste storage, and their performance deteriorates as queries become more complex.[4] An alternative has been hardware scanning, a character-by-character search through a data base looking for character streams matching those created in the query. This can be a successful approach, but one which can expend hours or days of processing time, even with the use of state-of-the-art mainframes. One would not recommend its adoption,

[1]Don G. Swanson, "Searching Natural Language Text by Computer," *Science* 132 (October 1960): 1099–104.

[2]David C. Blair and M. E. Mason, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the Association for Computing Machinery* 28 (March 1985): 289–99.

[3]Ibid., 289.

[4]Kwang-I Yu, Shi-Ping Hsu, Robert E. Heiss Jr., and Lee Z. Hasiuk, "Pipelined for Speed: The Fast Data Finder System," *Quest* [TRW Electronics and Defense Sector] (Winter 1986/1987): 5–19.

William Nolte, formerly Chief of Electronic Records and Archives at the Department of Defense, Ft. Meade, Maryland, is currently on assignment as an area studies analyst for the department.

for example, to search a future data base consisting of the digitized holdings of the National Archives.

Despite these limitations, the allure of full-text search has led to the development of several commercially available systems, ranging from microcomputer-based products[5] to those based on larger equipment, such as IBM's STAIRS. Before one rushes to the assumption that the giant of the industry has solved the problems associated with full-text retrieval, it should be noted that the gloomy assessment cited above resulted from a test using STAIRS and not the micro-based or small vendor systems.

The development of HSTS systems represents a new attack on the problem, an attack made possible in large measure by the appearance of parallel processing, "the technology of fifth-generation computers."[6] As its name implies, parallel processing represents not a significant leap in processing speed, but rather an increase in the number of processors handling a given requirement. This architecture permits a system to have multiple processors looking at a given data source at the same time, reducing response time proportionately.

HSTS equipment does more than simply increase the number of pieces of equipment searching for a specified part of a data base. It also reduces the costs associated with text searches, both by searching more quickly and by permitting the use of less expensive processing equipment. Because searches can be loaded from a mainframe onto a smaller host machine served by the HSTS device, processing costs will be cut significantly, as will capital expenditures.

Commercially available HSTS systems generally fall into the hardware-based versus software search categories for searching devices, though at least one developer, as will be noted, claims to have produced something of a hybrid. Software devices include BRS/SEARCH and BASIS. Though faster than earlier products, these systems continue to encounter the traditional software search problem: the amount of space used to store the inverted index. BRS/SEARCH attempts to deal with this problem by storing its index in a compressed format. In claiming a solution, the developers of hardware-based systems present search speeds that are little short of astonishing. General Electric's GESCAN, for example, is said to search at 250,000 characters of data (the equivalent of fifty to seventy pages of a book) per second. Recent enhancements to the GESCAN have at least doubled the speed claimed for the system.

TRW claims that its Fast Data Finder can support a system with twelve disk drives capable of storing 5.4 billion characters—and search "every character in it" in thirteen minutes. One of the more entertaining aspects of this technology is the attempt to find ways to put its capacities into some human perspective. According to TRW, the search described above (5.4 billion characters in thirteen minutes) equates to reading five years of the *Los Angeles Times* in less time than the average reader could browse through a single day's edition.[7] An even more incredible speed is claimed by the manufacturers of the Utah Text Search Engine (Contexture, Inc.), which uses hardware *and* a modified inverted index approach to achieve claimed search rates of 50 million characters per second (or five years of the *Los Angeles Times* in about two minutes).

All of the systems described above have problems, not the least of which is meeting in practice the search speeds described in

[5]For example, TEXTBANK from Group L Corporation, Herndon, Virginia, and MARCON from AIRS, Baltimore, Maryland.

[6]Richard K. Miller, *Parallel Processing: The Technology of Fifth-Generation Computers* (Ft. Lee, N.J.: Technical Insights, 1985).

[7]Kwang-I, et al., "Pipelined for Speed," 15.

marketing brochures. The limiting factor in this area has been data transfer rate. In other words, it makes no difference if a system can search a given number of characters per second unless storage and transfer devices can move data onto the HSTS as fast as the HSTS can devour it.

Archival Implications of HSTS

Archival administration is a reflective profession in that it traditionally accepts or reflects the organizational practices of the institutions creating records. In many respects, this somewhat passive approach will need to be modified if archivists dealing with machine-readable records hope to have their requirements introduced into systems and acquisition efforts. In other respects, however, archivists will remain dependent on decisions made by records managers and creators. One key to the archival application of HSTS technology is whether such systems become common in corporate and governmental institutions. Effective use of HSTS systems will depend on developments in other areas, including expert systems or other forms of artificial intelligence that can effectively limit or define searches. Advances in high capacity storage media make this the least problematic area associated with the application of HSTS, as optical disks featuring greater storage and reduced costs frequently appear.

Assuming that high-speed text search and related technologies mature and become commonplace, how will their development affect archival practice and principle? One of the effects archivists may experience is a possible change in the operating procedures of the organizations and institutions creating records that, in current practice, would at some point be transferred to an appropriate repository for disposition. The decision to retire records is a critical one, and any technological or other shift affecting that decision will have serious implications. The development of mass storage systems and declining costs associated with both their acquisition and operation already threaten the retirement and transfer mechanism. Effective text-search systems would contribute to this trend. If an organization can cheaply store data on its own premises (or at least its own system) and can systematically and accurately retrieve parts of that data, what motivation exists for a transfer at all? This self-service approach to record storage and retrieval would thus seem to eliminate the need to have an archivist provide the intellectual tools to describe records and facilitate their retrieval. Carl Becker's "everyman a historian" concept may have a latter day equivalent: "everyman an archivist."

Before consigning themselves to the scrap heap of lost professions, archivists should at least consider attempting to adapt to this potentially threatening environment, a first step being to understand the nature of the threat. Even if one assumes that the combination of cheap mass storage and effective search systems will tempt organizations to alter their behavior in retiring and transferring materials, serious problems of intellectual control of masses of data will not be solved. Text search systems might make searches of entire data bases possible; they will not render them perfectly efficient. The proper application of these devices should entail automatic segmentation of a data base and the creation of a multitiered retrieval capability.

Segmenting the data base means nothing more than ensuring that records from a company's marketing department go to the electronic equivalent of a designated storage area, with records from other departments being similarly "tagged" for retrieval purposes. In this way an efficient retrieval of records known to be from marketing would search only that segment and not the entire data base, resulting in a saving in processing costs.

Multitiered retrieval strategies offer similar economies. Many requests for records are triggered by the appearance of a spe-

cific citation in memoranda, correspondence, or discussions. When a serial number or other identifier is known, retrieval should be aimed at that identifier. This would require automatic indexing of key record fields (in standard office memoranda, such items as originator, recipient, date, subject, and—where used—serial number), and this in turn would require standardization of record formats within an organization. Archivists should discover that many of their skills are largely adaptable to the new technology, but they need first to understand that technology, and then to be able to define their skills conceptually and functionally, rather than in terms of specific procedures.

A similar ability to adapt may be required for archivists to adjust the appraisal process to deal with HSTS systems. Given the importance twentieth century archivists place on appraisal, this may prove difficult. The development of appraisal theory has been a major accomplishment for the archival profession. The discovery that the archivist is no longer a "passive custodian," but has become an "active appraiser," is of enormous significance not only in the practice of the profession but to its self-perception.[8] A former Archivist of the United States has described appraisal as "central" to the archivist's work;[9] it is seen by many as the most intellectually demanding, and therefore professionally defensible, aspect of an occupation that is uneasy about defining itself in purely curatorial terms.

HSTS systems could significantly alter the underlying assumptions concerning appraisal, which developed, after all, as a reluctant concession to the realities of growing masses of documents, the consequent costs of storing them, and the difficulty of retrieving useful information from them.[10] HSTS would not be the only challenge to conventional appraisal practice, being simply one part of a related series of technologies that together alter the operating environment of records creating entities and thus require a change in the archival procedures created in response to an earlier environment.

It is at least possible that records creators could choose not to support appraisal activities when purging and weeding unessential records seems not to be required. If storage is cheap and retrieval—even from an extremely large data base—efficient, why bother to sift and dispose? If the cost of appraising exceeds the cost of simply keeping even useless material, and if the chaff does not inhibit storage or retrieval, one could then argue for keeping chaff.

Will archival appraisal have a role to play as HSTS systems proliferate? In the absence of any experimentation and evaluation in the field, any answer must be speculative. What seems likely, however, is that the answer, when it comes, will neither be neat, global, nor fixed. Some archival theorists have conceded that technology could affect appraisal, making it economical to keep "less information-dense materials than in the past."[11] The governing assumption, however, has remained that technology would not permit the abandonment of the principle that the

[8]Nancy E. Peace, "Deciding What to Save: Fifty Years of Theory and Practice," in *Archival Choices: Managing the Historical Record in an Age of Abundance*, ed. Nancy E. Peace (Lexington, Mass.: D.C. Heath, 1984), 10.

[9]Robert M. Warner, "Foreward," in *Archival Choices,* ed. Nancy E. Peace, vii.

[10]Philip C. Brooks, "Archival Procedures for Planned Records Retirement," *American Archivist* 11 (October 1948): 308–15 and "The Selection of Records for Preservation," *American Archivist* 3 (October 1940): 221–34.

[11]This conclusion is derived from an assessment of the work of Swedish archivist Nils Nilsson. See Nancy E. Peace, "Deciding What to Save: Fifty Years of Theory and Practice," *Archival Choices,* 12. Peace and others argue that technology may eliminate the problems of physical bulk without eliminating the need for selection.

only way to make valuable information available is to dispose of that which is not valuable. This assumes, of course, that one can make that distinction.

Archivists may have to learn to live in situations in which appraisal decisions will be based on a number of new assessments. How rapidly is technology developing in a given field? What are the costs associated with a given technology? Are costs—relative to capability—stable, increasing, or declining? In short, archivists will have to become far more adept at planning, cost assessment, technology management, and other skills for which they have generally not been prepared. Several years ago F. Gerald Ham noted that archivists had failed to develop fiscal tools that would, among other things, permit them "to attach a price tag to their appraisal decisions."[11] This failure may be difficult to overcome, especially as technology adds to the options available and thus complicates the issue.

The archival profession's ability to answer questions about the costs of its activities will become increasingly critical as archivists attempt to involve themselves in defining and acquiring systems for the creation and storage of records. Failure to contribute to the decisions that organizations make about their information systems will largely relegate archivists to the role of curators of those materials created in an earlier environment. Ironically, excessive concern about protecting the specific (activist) procedures resulting from the development of appraisal theory could doom archivists to a return to passivity.

High-speed text search technology is not a development unto itself. Its ultimate implications for the archival profession will depend on concurrent developments in storage systems and artificial intelligence, among other fields. Even more, its impact on the profession will depend on the profession's ability to anticipate technological developments and to project the changes in archival procedure they require. Though the future of full-text search as an economical and efficient technique in archival settings remains to be demonstrated, the appearance of commercially available high-speed text search equipment suggests that such a strategy may have a role in the archival future.

[12]F. Gerald Ham, "Archival Choices: Managing the Historical Record in an Age of Abundance," *Archival Choices*, 137.