M. M. Astrahan

The Role of Large Memories in Scientific Communications

Abstract: Large memories provide automatic reference to millions of words of machine-readable coded information or to millions of images of document pages. Higher densities of storage will make possible low-cost memories of billions of words with access to any part in a few seconds or complete searches in minutes. These memories will serve as indexes to the deluge of technical literature when the problems of input and of the automatic generation of classification information are solved. Document files will make the indexed literature rapidly available to the searcher. However, memory capacity is currently well ahead of our ability to use it, and much work remains in this area. Machine translation of languages and recognition of spoken information are two other areas which will require fast, large memories.

Large memory

The words *large memory* imply, for me, a store of at least a million alphanumeric-coded machine-readable characters, or a million reduced picture images. Automatic access is implied for the coded characters, although manual access may be suitable for the picture images. The access may be directly to a specific body of information, called a record, with little or no scanning, or it may involve scanning the entire file. Records may be identified or addressed by their physical location or by their content.

Coded memories

Access speed is an important characteristic of coded memories. Random-access time is the time required to reach any desired area of the memory and read out a record. Serial scan time is that needed to read all the records.

Present commercially available, large coded memories range from 15-million-bit (binary digit) drums with 1/5 second average random-access time to 40-million-bit disk arrays with 2/3 second average random-access time, and on to reels or strips of magnetic tape which require many seconds for random access to hundreds of millions of bits.

High-speed tapes are available which can be read at a rate of 3/4 million bits per second on 31 parallel channels. If used for continuous serial scan, a 2700-foot reel containing 1/4 billion bits could be searched in 5 1/2 minutes.

The current state of the art of magnetic and photographic recording, as well as the descriptions of laboratory models of large memories which have been presented, make it evident that it is now technically feasible to produce memories of 1 to 10 billion bits with random-access times of less than a second. Economic feasibility is, of course, another question and one which I do not propose to discuss.

High density of recording is an important key to achieving high capacity, high speed, and low cost. Let us examine the densities that seem achievable. It appears that the density limitations in both magnetic and photographic recording will be due for some time to the reading and writing apparatus rather than to the resolution of the medium. For either technique 100,000 bits per square inch is achievable now in the laboratory and a goal of a million bits per square inch is regarded as achievable for a practical machine. Theoretically attractive techniques for achieving a hundred times higher densities have been suggested, but not yet critically evaluated. I will come back to all these numbers later in describing applications.

Scanning speed is another important property, since a complete scan of a file is a straightforward way of achieving an associative access, that is, one based on the content of a record. A complete scan is certainly not aesthetically neat compared to direct associative access, but it definitely is technically attractive if the scanning speed can be made sufficiently high. For magnetic stores, we can expect to achieve scanning speeds of 10 million bits per second for a single reading head. This same rate can be achieved optically, using flying-spot scanners. One attractive possibility for use in searching a photographic file for exact matches is the use of mask comparison for the searching data. This technique leads to severe registration problems at the densities I am bandying about, and it also leads to format restrictions, but it can greatly increase the search-

ing speed with a single searching element. All these speeds can, of course, be greatly increased by parallel reading heads, at greater cost.

I have deliberately not discussed writing rates, since these are less important in the applications I will consider. The rate of duplicating a whole file for distribution is of interest, and here the possibility of high-resolution contact prints makes optical recording attractive.

Picture storage

For storing picture images, there is microfilm, which can put several thousand frames on a 100-foot reel. There are also systems available for putting several page images on a film card. At present, one-hundred-to-one reductions are possible for sheets containing typed text. This results in a storage density of a hundred pictures per square inch. One thousand pictures per square inch is a reasonable goal for a practical machine. Devices which have been demonstrated include Minicards, with a 60:1 reduction and 12 frames plus coding to a card about ⅝ in. x 1¼ in., and the National Bureau of Standards system which provides direct access to any of ten thousand frames on a piece of film 10 inches square. In the latter, each frame represented an individual character, and so did not involve a high reduction.

Indexing

The most evident application of large memory to scientific communication involves communication by means of written information. In the category of written information, I include photographs and drawings. Large memories will help answer the following questions:

1. What is written?
2. Where is it?

Large memories will also help to carry out the following orders:

1. Get it.
2. Translate it.

The other category of scientific communication involves spoken information, and here too, large memory can play a role.

Probably the most pressing application is getting the answers to the two questions, "What is written?" and "Where is it?" Variations of this problem are called Information Retrieval, Literature Search, Indexing, and other similar names. The problems described with these names cover a wide range of complexity. An example of the simplest kind of problem would be the following. Given a set of research reports, each defined by a unique number, find the location of the report corresponding to a specific given number. An example of the most complicated kind of question would be the following. "What has ever been written concerning the photoconductive properties of materials in a specified frequency range of incident radiation?"

For purposes of considering large memory requirements, we can describe a general problem which covers the whole range of problems. For purposes of this talk, I will call the general problem "the indexing problem" and avoid the need for naming the variations of it. Any indexing problem involves a finite set of documents, D in number. In the most general case, D might be the total number of documents ever written. In a very restricted case, it might be a set of numbered reports issued by a laboratory. I define the word document as being a unit of written information which I am willing to search through if there is a high probability that the information I am looking for is somewhere within it. This means to me the length of the usual article in a technical journal or a chapter in a book.

In any specific problem there is an average number of characters C per document, including all the text tables, titles of pictures, mathematical symbols, chemical formulas, texts on graphs, etc. I take a character to be 7 to 8 bits, including a check bit.

All of the indexing techniques I have heard used or proposed involve choosing one or more index terms for each document from a set T of all possible index terms. These terms may be merely an identification number, or they might perhaps be chosen from a catalog of suitable index concepts for the subject being considered, or they might be derived from the words in the article using more complicated logical analyses. $(DC/10)$, which is 10% of the total number of characters in all the documents, is probably a good upper bound to the number of characters that might be needed for complete indexing. This is my guess based on work done by Phyllis Baxendale* of this laboratory on determining index terms automatically from original text by use of rules based on language structure. I believe that these techniques offer the only hope of achieving high-speed searching equivalent to the questioner's reading all the searched documents.

Any inquiry will be expressed by means of the index terms, either directly, or perhaps through a logical analysis of an inquiry statement similar to the analysis used in indexing the documents. The indexing process then involves searching the file of index terms for matches with the inquiry terms. The output is the identification and location of documents, and possibly abstracts. The matches could be exact or they might involve other logical rules, such as those which would result in selecting the closest matches.

There are two general ways that a large memory can be organized for indexing. If the memory has direct access, we can organize the memory into a unit record for each index term and store in each of these records the identification of the documents which are characterized by that index term. We then search only the records for the inquiry terms. With a serial scanning memory, we organize the memory into a record for each document, where each record contains the index terms for that document, and inquiry then requires a search through all the records. If the document numbers, which are repeated

*P. B. Baxendale, "Machine-Made Index for Technical Literature—An Experiment," page 354 of this issue.

for every appropriate index term in the first case, average the same number of characters in length as the index terms, which are repeated for every appropriate document number in the second case, then the same file capacity is required for indexing in both systems. Systems exist today using 5 million character magnetic tapes or Ramac Disk Files which can accommodate tens of thousands of documents with several manually assigned concepts or index terms for each document.

I will mention a few of these that have been reported.

At the Naval Ordnance Test Station at China Lake, California, over 14,000 reports on magnetic tape have been searched. Chemical structure searches on tape were tried at Dow and Monsanto. Our experimental system using a Ramac now contains 5,000 documents. I tried it in preparation for this paper and got a long list of references. Of course, I got most of my information by the proven method of asking people who know.

FOSDIC 2 should also be mentioned here. It was developed by the Bureau of Standards and provides for selective reproduction of punched card images stored at a density of 13,000 card images to a 100-foot reel of 16 mm microfilm. It searches at 4,000 cards per minute. The experimental model is being delivered to the Air Weather Service at Asheville, N. C. for evaluation. They have a file of 300 million cards to search.

Two other jobs are worth mentioning as examples of indexing. These are the use of magnetic tape systems in the preparation of a Bible Concordance and in the indexing of the Dead Sea Scrolls.

I will now indulge in speculative mathematics about memory sizes that might be needed in the future for indexing. I have found a second-hand reference* which states that there are approximately 60 million pages of technical literature currently published every year throughout the world. From other references, I have gotten figures of almost 2 million magazine articles and more than 60 thousand technical and scientific books published each year. Together, these probably would account for about 25 million pages, so I will accept the 60 million figure. Let us assume that we wish to search this whole mess and ignore everything but the memory requirements. I will assume that there are an average of 500 words per page, so my 10% rule calls for 50 index terms per page. This comes to a total of about 3 billion index terms, or about 300 per document if we assume 10 million documents. The index store will thus require about 150 billion bits, assuming 50 bits per term. This could be halved with better coding, but more tables and less redundancy in the search matching would result.

At a million bits per square inch, we could put 150 billion bits on a recording surface one inch wide and 12 thousand feet long. A series scan of the whole memory at 10 million bits per second would take 4 hours. If we could afford to be in a hurry, we might provide a hundred parallel channels and a hundred comparing circuits and run the tape at a thousand inches per second. The search would then take about 2 minutes. If the inquiry only takes about 10 words, however, it would be profitable to make a memory with more direct access. For example, the one-inch-wide surface mentioned previously could be divided into 120 hundred-foot strips. Assuming 10 seconds to select a strip and 10 seconds to run an average of 25 feet out and 25 feet back to the center of the strip, each access would take 20 seconds, and the whole search would take about 3 minutes. Using 12,000 one-foot strips we might be able to cut this down to 1 second per access, with a total of 10 seconds for the search. Other arrangements on disks and drums could just as well have been considered.

Let us speculate further and consider the Library of Congress. Dr. King once estimated that it contains 10^{14} bits. Using the same indexing factor of 10% that I used before, we come out with about 100 times as much storage required as in the preceding example. Thus, it would not take too much further improvement in resolution and speed to make possible a practical single index for all of this information.

Of course, if we had these memories today, we would not be able to make use of them. We do not really know how to use the large memories available today. There are many successful punched card, special purpose indexes in use, but hardly any magnetic tape systems. The tape systems existing are mostly experimental. One reason is that there are many other techniques that need to be developed. These involve input methods and searching logic.

Manual assignment of index terms and preparation of the machine-readable index entry for a document is costly and of limited value. We must have means for automatically entering text information and for automatically generating the index terms from it which are independent of the opinions of a human cataloguer. Character sensing readers which will read all type styles might be achievable at a cost which would make them practical, if used in conjunction with people to point out the things to be read and the proper order. I think that machines to do this whole job, so that you could hand them a magazine or a book and come out with indexed documents, are further off than the index memories themselves. Thus, for some time our systems will be limited largely by the input problems. We can at least hope that in the next few years it will be possible to begin having all published technical information produced in a machine-readable form in addition to the human-readable form. A lot of work is going into the automatic indexing problem and it seems reasonable to hope that solutions to this problem will be available when needed. With automatic entry and automatic indexing, it will become practical for central agencies to maintain up-to-date indexes in special areas. These might contain a hundred thousand documents, at 1000 index terms for each document. A basic billion-character memory would be needed, and it would be revised regularly, either by the addition of supplementary units or by complete replacement of the storage element.

*Vagtborg, H., *Tomorrow Through Research*, 7, No. 1, p. 1, February 1955, Southwest Research Institute.

Document store

You will notice that I have kept the job of indexing separate from the job of retrieval of the document, which is the answer to "get it." We can imagine picture-storage memories in which rapid mechanical access to any of a large number of pieces of film is available. Each piece of film would have pictures stored on it at a density of between 100 and 1000 pictures per square inch. The coded index files would provide the information concerning the specific location of any desired picture. The first justification for such systems will be in terms of access time. With improved print-out and display techniques, these systems will compete on a cost basis with manual access. We might look forward to a system whereby libraries have mainly an indexing search machine and an automatic document file. We might further hope for remote display and print-out at many locations, or we might provide each individual with a complete set of hundreds of thousands of picture images with manual or even automatic access, along with a viewer and copier. In either case, a searcher would call the library to find where to look, and would do the actual reference work at his own desk.

Translation

Language translation is another area which will be helped by large memory. Here again, current memory techniques are ahead of our knowledge of the logic of the problem. Most present systems use dictionaries which translate single words into their alternative translations, leaving the reader to make sense out of the resulting mish-mash. Automatic production of a clean translation, such as a good human translator could produce, will require more storage to hold the language structure and idioms. Even so, memories of one to ten million characters with random access in one-tenth second should be able to handle real time translation, when the language structure and data input problems are solved.

Received May 27, 1958

Voice recognition

Speaking of real time translation brings us to one other means of communication which should be considered, namely, voice communication. Voice recognition for purposes of translation, control, or recording for printing will eventually be handled with the aid of large memories in a manner similar to the language translation problem. Schemes for building up words by recognizing individual phonemes will be replaced by systems wherein whole words at a time are recognized by an indexing process. The index terms will be characteristics such as energy distribution at various frequencies for a succession of phonemes. Simple language-structure rules will make it possible to distinguish between words which sound the same but are spelled differently. The indexing rules would probably have to call for nearest matches, rather than for exact matches. A vocabulary of ten thousand words would probably be quite satisfactory for a voice-operated letter writer. A million-character memory completely scanned several times a second would handle this job and is certainly achievable within the density and speeds mentioned earlier.

Conclusion

Large memories provide automatic reference to millions of words of machine-readable coded information or to millions of images of document pages. Higher densities of storage will make possible low-cost memories of billions of words with access to any part in a few seconds or complete searches in minutes. These memories will serve as indexes to the deluge of technical literature when the problems of input and of the automatic generation of classification information are solved. Document files will make the indexed literature rapidly available to the searcher. However, memory capacity is currently well ahead of our ability to use it, and much work remains in this area. Machine translation of languages and recognition of spoken information are two other areas which will require fast, large memories.