Chapter 1

Information Retrieval: An Introduction

0 PREVIEW

This chapter examines the information retrieval problem by considering the social and technological world in which retrieval systems exist. Later chapters will deal with individual system functions and parameters. To render this discussion meaningful, it is necessary to understand the context in which information retrieval systems operate and be aware of the various types of existing information systems.

The chapter closes with an examination of the functional components of information retrieval and a description of a few basic methods for organizing information retrieval files. The second chapter covers retrieval systems whose operations are based on one of these file organization methods, the inverted file.

1 OVERVIEW

Information retrieval (IR) is concerned with the representation, storage, organization, and accessing of information items. In principle no restriction is placed on the type of item handled in information retrieval. In actuality, many of the items found in ordinary retrieval systems are characterized by an em-

phasis on narrative information. Such narrative information must be analyzed to determine the information content and to assess the role each item may play in satisfying the information needs of the system users. The items processed by a retrieval system typically include letters, documents of all kinds, newspaper articles, books, medical summaries, research articles, and so on.

Most people are faced with a need for information at some time or other. Typically one might first turn to friends and acquaintances for help, but if that is to no avail, a more formal search might be initiated in a library or information center. A first search effort might then lead to one or more information items that are selected for detailed examination. In some cases these initially chosen items might suffice in satisfying the existing information needs. If not, additional items might be sought. One possibility for extending a search for information consists in using references to previously available information items to find additional items in related areas. Alternatively, the information need could be redefined. For example, a person interested in information about the effect of tetraethyl lead on the environment and on human beings may conduct separate searches for articles dealing first with the effects of tetraethyl lead on humans, and then with the effects of tetraethyl lead on the environment.

To facilitate the task of the information user in finding items of interest, libraries and information centers provide a variety of auxiliary aids. Each incoming item is analyzed and appropriate descriptions are chosen to reflect the information content of the item. Each item is classified in accordance with the established procedures and incorporated into the collection of existing information items. Procedures are established for formulating requests designed to satisfy an information need and for comparing these requests, or *queries*, with the descriptions of the stored items. These comparisons are the basis for deciding which items are appropriate for the respective queries. Finally, a retrieval and dissemination mechanism is used to deliver the information items of potential interest to the users of the information system. These steps are all carried out in conventional libraries where a card catalog forms the principal auxiliary tool used in an information search. The processes and methodologies needed to carry out those tasks automatically are described in the remainder of this book.

It is often claimed that the usefulness of a collection of information items depends crucially on *currency* and *completeness*. The desire to maintain currency implies that new items must constantly be added to the collections. Completeness implies further that the collection contains a large proportion of the items of potential interest, and that obsolete items are removed only when the obsolescence of an information item can be established without doubt. The U.S. Library of Congress which attempts to maintain both currency and completeness, is adding about 3,500 new items to the collections every day [1].

Currency and completeness are obviously impossible to achieve simultaneously in an age of limited resources. Hence it is necessary to compromise by attempting to incorporate into the collections all the "important" items. But item importance is difficult to evaluate in advance: many information items attract little attention and are never used; others, such as, for example, Vannevar

Bush's "As We May Think," outlast most contemporary items [2]. In practice, somewhat arbitrary decisions are often made to control the acquisitions and the collection maintenance procedures.

The collection development problem is aggravated by the growth in the available information. In early times, the total available knowledge changed relatively slowly. However, by the year 1800, the amount of scientific publication was already doubling every 50 years [3]. More recently with the impressive growth of science and technology, the rate of increase of available knowledge has vastly accelerated. Between 1800 and 1966, the number of scientific journals has increased from 100 to over 100,000. At the present time, no upper limit is apparent in the rate of increase of available information items.

Consider now the problem of actually locating a particular item included in a collection of documents. Various access mechanisms may be provided, related to either the physical or the logical organization of the items. In a library the *physical organization* is generally controlled by the arrangement of call numbers. In the United States common call numbers in use in libraries of academic institutions are those provided by the Library of Congress classification system [4]. Books placed in order according to these call numbers are clustered on the library shelves by topic area. Thus, books about information retrieval may be assembled under common call numbers beginning with Z699. Unfortunately, the same call number (Z699) may also be used for other related subjects such as library automation, cataloging, and general library processing. Furthermore additional information retrieval items can also appear in various other sections of the library, notably in classes identified by call numbers TA and TK in the Library of Congress system.

A person seeking a given information item may then be forced to outguess the library cataloger who made the original decision about the placement of the particular item. To render this guessing task easier, a *logical organization* of the data may be superimposed on the physical organization. Thus, books published on information retrieval can also be identified by looking in a library subject catalog under the term "information retrieval." In some libraries the correct term might be "computer-based information retrieval" or perhaps "information systems retrieval." In any case, once the appropriate term is found, adjacent cards will identify books related to the topic being sought. These books may belong to various call number locations (that is, Z, TA, TK, etc.); all those locations will provide some reference to information retrieval. Given a particular call number, the corresponding item should be found at the designated location on the library shelves. If the item is not at the designated location, one presumes that it is in use or that it may be lost.

When a subject catalog is available, changes can be made to the subject terms without actually reshelving the books themselves. In particular, the items can be logically reorganized by suitably changing the library catalog without altering the physical arrangement. A large number of different logical organizations can be used to characterize the various items. Thus, the items can be placed in order by author, size, date of publication, date of acquisition, title,

subject, and so on. Each logical organization then corresponds to a different set of cards in the catalog.

One problem faced by all users of information systems is the need to reduce to a manageable size the number of items that are to be examined. It is not obvious that the methods currently available for this task are adequate. As early as 1945, the existing methods for information organization were criticized [2]:

> There is a growing mountain of research. . . . The investigator is staggered by findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less remember. The summation of human experience is being expanded at a prodigious rate and the means we use for threading through the consequent maze to the momentarily important item is the same that was used in the days of the square rigged ships.

Similar sentiments have been voiced by many other observers. In Alvin Toffler's "Future Shock"—a book dealing with society's inability to cope with change—Emilio Segre, Nobel prize-winning physicist, is quoted as saying that "on k-mesons alone, to wade through all the papers is an impossibility" [5]. In other words even in specialized, relatively narrow topic areas, one tends to become overloaded with information very rapidly.

The construction of an effective system of information organization which permits efficient use of the information items is difficult for at least two reasons. First, the volume of information expands unevenly for different topics. Some areas such as computer science, for example, are growing at a very fast rate, while other subjects such as certain foreign language studies may not be growing at all. Future growth patterns of information are difficult to predict and any predictions are subject to large error rates. To take care of future growth, one may want to provide for some expansion in each and every topic area. Ultimately these expansion mechanisms will be overtaxed in some areas while not being used at all for other topics [6].

A second difficulty in creating effective information organizations is the desire to keep related items relatively close together. For example, books on algebra, matrix theory, graph theory, and topology should appear close to one another in the collection [7]. At first glance this may appear to be easy enough, especially when these topics all clearly fit under the more general topic of mathematics. Special problems do, however, arise for interdisciplinary topics such as systems analysis. This particular subject is related to several major topics including computer science, operations research, engineering, management science, education, and information systems, as shown in the scheme of Fig. 1-1. An organizational arrangement which would allow items on systems analysis to appear close to other items in all related topic classes cannot be achieved by placing the items in order on a bookshelf (an organization based on only one dimension). Rather the organization must be multidimensional.

A two-dimensional organization could, for example, take into account shelf locations above and below a given area rather than only those situated

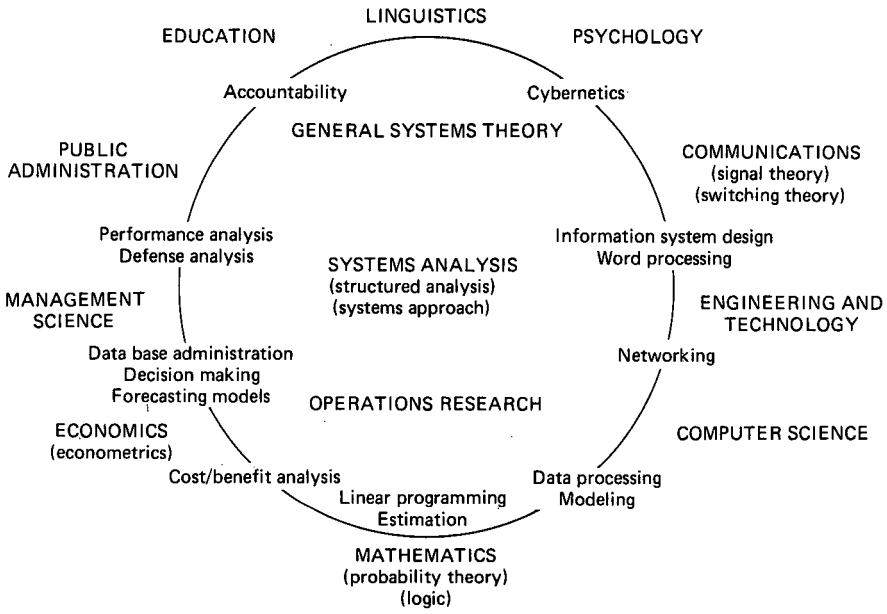

Figure 1-1 Systems analysis: related disciplines, tools, and techniques.

adjacently on the same shelf. Figure 1-2a depicts a one-dimensional organization, while Fig. 1-2b and c represents two- and three-dimensional organizations, respectively. The actual number of dimensions necessary to keep related subjects close to each other must be determined from the number of required relationships. Unfortunately, the physical organization of information stored in computer systems is limited to the few dimensions of the existing hardware systems. Hardware devices and software methodologies are available that may effectively increase the number of available dimensions. These are introduced in Chapter 8.

2 CHANGING TECHNOLOGY

Up to this point a fairly desperate picture has been presented. Barring substantial logical and technological advances, a situation may soon arise where it may not be possible to isolate the useful items from the mass of available information. People may be forced to use whatever turns out to be most accessible and ignore the remainder. Fortunately, computer technology is providing increased capabilities for the manipulation of all types of information. The trend over the past 15 to 20 years has been for computational capabilities of every sort to double every 3 to 4 years. That is, twice as many numbers can be added together in the same time and twice as many items can be stored in the same space as 3 or 4 years ago. The exception is the cost of computing which has decreased. Thus for equivalent dollars one can buy two or three times as much computational capability as was available 3 or 4 years ago [8–10].

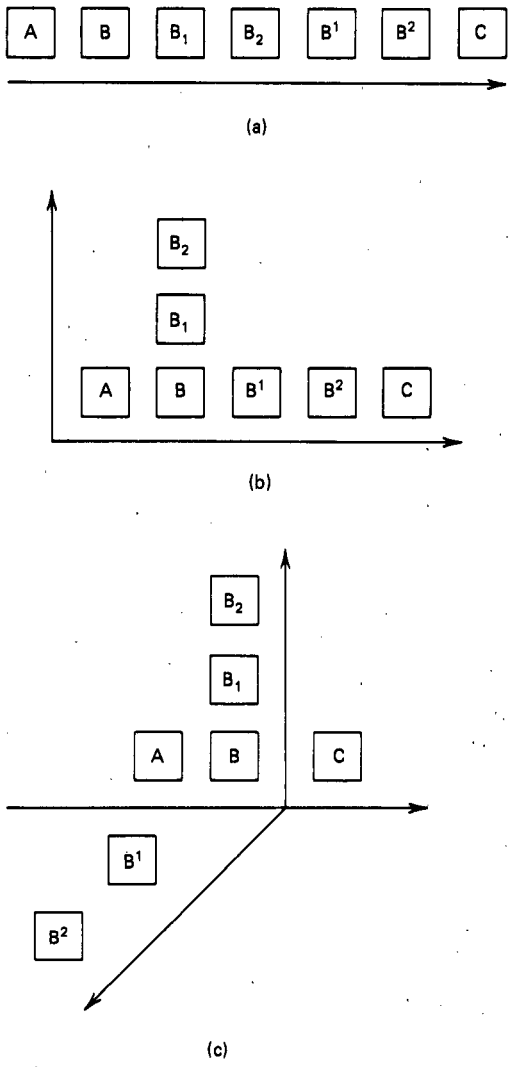

Figure 1-2 Data organizations. (a) One-dimensional data organization. (b) Two-dimensional data organization. (c) Three-dimensional data organization.

Startling advances have been registered also in the capabilities of the physical devices which store information. The amount and type of information that can be retained in storage has changed significantly. The first electronic computers were capable of storing only a few numbers. At the present time, devices are being designed that are capable of storing a trillion characters of information. The growth in capacity of storage technology is such that one should soon be able to store library-sized quantities of information in a cost-effective manner. For the stored information to be useful, the information must be delivered to the potential users. Communication technologies such as the telephone play

an important role in this delivery process. Unfortunately, communication devices are still relatively slow and costly at the present time. However, numerous changes are occurring such as, for example, the introduction of satellite communications, the use of optical fibers as communications channels, and the design of sophisticated communication networks. If this technology continues to improve, large quantities of data should be transmittable efficiently in the foreseeable future.

In summary, the existing information problems appear difficult to master without the use of sophisticated methodologies for storing, processing, and transmitting information. The availability of modern information retrieval systems has greatly improved the access to many stored information collections. However, even the current systems are unable to deal effectively with the increasing growth of information. In spite of many efforts on the part of researchers, current operational capabilities remain at a relatively elementary stage [11]. In part, this may be due to the unwillingness or inability of individuals in the information fields to try out and utilize devices and methods which would render the traditional methods obsolete [12]. The present text describes current capabilities as well as promising areas of work and the main research efforts dealing with the information retrieval problem. Before proceeding with a detailed description, it is necessary to place information retrieval in the context of other information processing work. This is done in the next section of this chapter.

3 INFORMATION SYSTEM TYPES

Information retrieval exhibits similarities to many other areas of information processing. The most important computer-based information systems today are the management information systems (MIS), data base management systems (DBMS), decision support systems (DSS), question-answering systems (QA), as well as information retrieval systems (IR).

A Information Retrieval Systems

Information retrieval is best understood if one remembers that the information being processed consists of documents. In that context, information retrieval deals with the representation, storage, and access to documents or representatives of documents (document surrogates) [13]. The input information is likely to include the natural language text of the documents or of document excerpts and abstracts. The output of an information retrieval system in response to a search request consists of sets of references. These references are intended to provide the system users with information about items of potential interest. The users of information retrieval systems have a wide variety of different information needs. They include research scientists seeking articles relating to particular experiments, engineers trying to determine whether a patent covering some new idea has previously been obtained, buyers of vacuum cleaners trying to obtain new product information, as well as attorneys searching for

legal precedents. In other words, information retrieval system users exhibit many different backgrounds, and many different reasons may lead them to use the retrieval facilities.

B Data Base Management Systems

Any automated information system is based on a collection of stored items (a *data base*) that needs to be accessed. Thus data base management systems might simply be systems designed to manipulate and maintain control of any data base. In actual practice, data base management systems are concerned with the storage, maintenance, and retrieval of data facts available in the system in explicit form. That is, the information does not appear as natural language text but is available instead in the form of specific data elements stored in tables [14]. In a data base environment each item, or *record*, is thus separated into several *fields*, and each field contains the value for a specific characteristic or attribute identifying the corresponding record. The characteristics used to identify a set of personnel records might be the names of the individuals involved, the addresses of the various people, as well as the social security numbers, and the job classifications. Specific values of these characteristics, that is, specific names or specific social security numbers, are used as identifiers for the individual records.

The processes of concern in data base management include the storage and retrieval of data, the updating or deletion of data, the protection of data from unintentional or deliberate damage or misuse, and perhaps even the transmission of data to remote users or other data management systems. The output of the systems may consist of individual records, portions of records, tables, or other arrangements of the data from the data base.

In order to be suitable for data base processing, each search request must state the specific values of certain record identifiers for the records of interest. A user may, for example, wish to retrieve all the personnel records corresponding to people of a certain age, height, sex, and so on. The retrieved information will consist of all records which match the stated search request exactly. In information retrieval, as opposed to data base management, it is often difficult to formulate precise information requests, and the retrieved information may include items that may or may not match the information requests exactly. The information contained in data base management systems includes a good deal of numerical data, and statistical or computational facilities may be provided to manipulate the numbers.

C Management Information Systems

A management information system is a data base management system tailored to the needs of managers. The functions performed by a manager in a given corporation depend on the availability of many kinds of data. Of particular interest may be information leading to the choice of possible alternatives by the manager presented in terms of ranges of values of particular attributes. A management information system therefore fits the general data base management

framework. However, in order to be useful to the manager, the information may be subjected to special processing not normally available in data base management systems. For example, a data base of geothermal dispersion data may be used by a manager to determine the effect of differing equipment configurations in a plant. In such a case, data relating to particular equipment configurations and to dispersion characteristics would be connected to a modeling capability to support the manager's activities. Such special-purpose systems useful for management are known as management information systems.

D Decision Support Systems

The systems described so far perform specific operations on homogeneous classes of information items. Normally, information retrieval systems do not perform management information functions, and vice versa. However, it is in principle possible to conceive of information systems in which a variety of different components are assembled into a single cooperating structure that includes information retrieval systems, data base management systems, computer graphics systems, and other technical capabilities which collectively provide powerful tools in support of the decision making process. If, for example, the manager interested in the dispersion characteristics of plant equipment were additionally given the ability to match the dispersion effects on a television screen and to use the graphics equipment also to review written materials in this area, the complete integrated system might be characterized as a decision support system. Decision support systems exist on a limited basis for narrow ranges of users employing data bases in restricted subject areas [15].

E Question-Answering Systems

Question-answering systems provide access to factual information in a natural language setting. The stored data base often consists of large numbers of facts relating to special areas of discourse, together with general world knowledge covering the context within which conversations between persons usually take place. User questions may be received in natural language form, and system responses may also be furnished as natural language formulations. The task of the question-answering system consists in analyzing the user query, comparing the analyzed query with the stored knowledge, and assembling a suitable response from the apparently relevant facts.

Question-answering systems currently exist only as experimental devices. The extraction of meaning from natural language and the determination of general rules of intelligent behavior seem to be major barriers to creating effective question-answering systems for general use. An overview of question-answering systems is provided in references 16 and 17.

Figure 1-3 sketches the relations between the various types of information systems. Although the management information and data base management systems have a great deal in common, this is not apparent for the other system types. Data base and management information systems process structured data, often in the form of tables of numeric information. Document retrieval

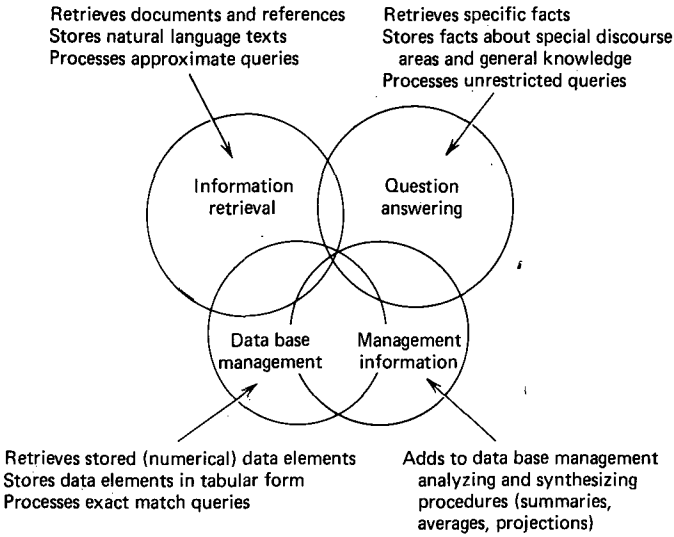

Figure 1-3 Overlap among types of information systems.

and question-answering systems, on the other hand, are concerned with natural language data, the former to retrieve documents and the latter to retrieve specific facts in answer to incoming queries. The users of such systems may have many different interests and backgrounds, and will thus require a variety of services and end products from the systems [18].

The analysis of natural language information is examined in Chapter 7, and the design of question-answering systems is described in more detail in Chapters 7 and 9.

4 FUNCTIONAL APPROACH TO INFORMATION RETRIEVAL

Many different information retrieval systems currently exist. To put these systems in reasonable perspective and to understand the advantages and disadvantages, it is necessary to understand the key functions performed by these systems.

Every information retrieval system can be described as consisting of a set of information items (DOCS), a set of requests (REQS), and some mechanism (SIMILAR), for determining which, if any, of the information items meets the requirements of the requests [19]. Figure 1-4 shows the relationship of these components. SIMILAR represents a relationship operator mapping specific queries to particular items included in the stored document set. In theory, the relationships between queries and documents can be obtained by direct comparison, as suggested in the model of Fig. 1-4. In practice, the relevance of specific information items to particular requests is not determined directly. Rather the documents or information items are first converted to a special form using a classification or indexing language referred to here as the indexing language,

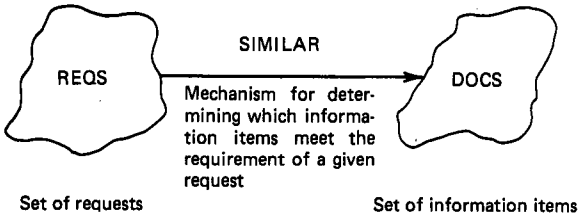

Figure 1-4 Information system environment.

LANG. The requests are also converted into a representation consisting of elements from LANG. Figure 1-5 exhibits the processing of the information items and the requests into LANG.

The mapping of the information items to the indexing language may be carried out manually, automatically, or by a combination of the two processes. This mapping is known as the indexing process. The mapping operation for the requests represents the query negotiation process. The procedures for determining which information items should be retrieved in response to a query are based on representations of the requests and information items consisting of elements from the indexing language. SIMILAR is a relation operator determining the similarity of the various information items to a given request. The similarity measuring process produces identifications of information items that are potentially relevant to the request. SIMILAR may also be considered to be the retrieval function because it identifies the specific items that are to be retrieved. In some instances, the system places the retrieved items in order of probable relevance to the request; in general, however, the retrieved information items will appear in the order in which they are located in the files.

The indexing language is either prespecified (controlled) or taken freely from the text of the information items and information requests (uncontrolled). Sometimes a combination of controlled and uncontrolled elements may be included in the indexing language. Whichever type of language is used, an infor-

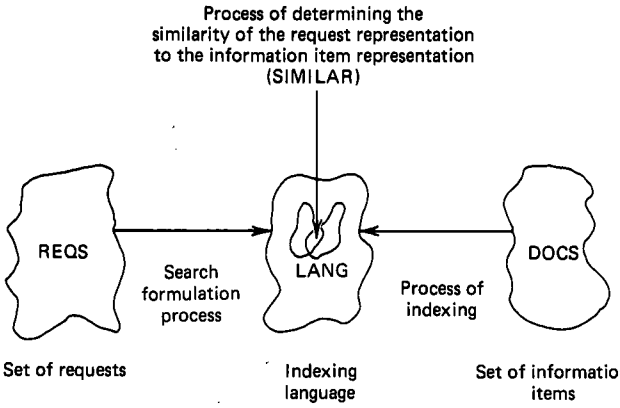

Figure 1-5 Functional overview of information retrieval.

mation item is usually assumed to be representable by a list of elements from the indexing language. For example, a particular information item entitled "Environmental noise assessment study, Part I" could be represented by using the controlled index terms "noise," "environmental influence," "utilization review," and "resident environment" [20]. To form a unified representation of this information item, a four-element *vector* can be used. For instance, the representation of this information item may be a structure such as

$$\langle 1 \quad 1 \quad 1 \quad 1 \rangle$$

where the first 1 is understood to stand for the indexing language term "noise," the second is "environmental influence," the third is "utilization review," and the fourth is "resident environment." A second document that includes the same topics except for the term "noise" would be represented as

$$\langle 0 \quad 1 \quad 1 \quad 1 \rangle$$

The particular method used to represent the individual information items is less important than the actual choice of the elements of the indexing language made during the indexing and text analysis processes. These questions are discussed in Chapter 3.

Following the choice of a representation for the information items and requests, the item representations must be organized into a *file structure*. Specifically, the item representations must be collected together and organized to ensure the efficiency and/or effectiveness of operations such as file searching or file updating. File structures vary in complexity from those with little or no organization to those structures maintaining various explicit relationships between information items.

5 SIMPLE FILE STRUCTURES

A Linear Lists

The simplest file structure is referred to as a linear list. A linear list is literally an unordered collection of items. Recall that current computer storage devices are largely one-dimensional. Thus, to find a specific element in a linear list kept in a computer store requires that the file items be examined one at a time. The item being sought may fortuitously be the first element examined—or if one is especially unlucky, it may be the last item. On the average, a specific item will appear in the middle of the file. Hence, on the average $(n + 1)/2$ items must be examined to locate a given item, where n represents the number of items in the file.

A linear list of the kind represented in Fig. 1-6 exhibits many advantages for information retrieval. For example, when new items are introduced into the system, they can be added to the file without concern about altering the order of already existing items. Nor need deletions from the file be followed by any

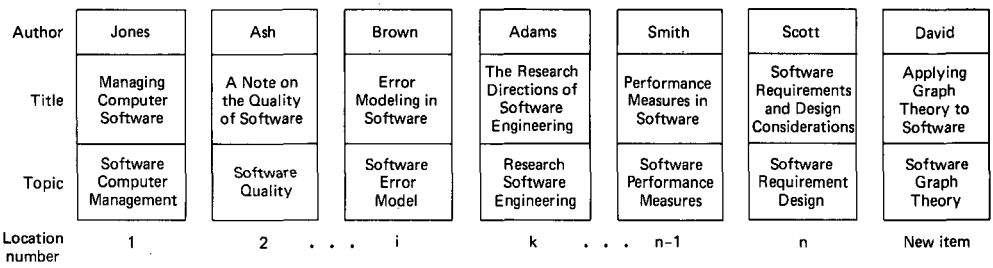

Figure 1-6 Linear list.

file rearrangements. The file maintenance process is therefore largely eliminated.

Since the items relevant to incoming requests are unknown in advance in a linear list arrangement, all items may have to be examined to determine those that must eventually be retrieved. When the file includes only a few items, this is a reasonable process. Unfortunately, for large files the file scanning operation becomes very inefficient. For example, assuming that it takes $^1/_2$ second to decide whether or not to retrieve a specific item, 83.3 minutes will be required to answer a request for a file of 10,000 items. Thus, on the average the search for a specific item will take 41.6 minutes. If, on the other hand, it takes $^1/_{10}$ second to determine the retrieval status of an item, the retrieval time is reduced to 16.6 minutes in all, or to an average of 8.3 minutes when a specific item is wanted. Thus, the usefulness of a linear list is dependent on the size of the file and on the speed of retrieval. Chapter 4 introduces the notion of clustered files that may be used to increase the efficiency of retrieval. In a clustered file, items are grouped into classes in such a way that all items entered into a common class exhibit certain similarities. When a clustered file is used for retrieval, it is no longer necessary to examine every file item; instead the search can be restricted to certain classes of items that appear to be "close" to the request.

B Ordered Sequential Files

In most instances, certain portions (fields) of the records stored in a file prove to be of special importance for retrieval purposes. For instance, the name of the first author of a journal article is often used as the main criterion for finding that article. The fields used to obtain access to the stored records are referred to as *keys*. A given file may be ordered sequentially according to the values of one of the keys, in which case this special order may be used to obtain access to the file. For instance, an ordered file consisting of journal articles may be ordered alphabetically according to the last name of the first author of each article. In contrast, the linear list would store the same journal articles in no specifiable order.

The addition of a new item to a sequential file requires that room be made at the appropriate spot to enter the new item. That is, a new journal article to be placed in the file must appear at a specific location if the order of the file is to be maintained. Other items may have to be moved in order to make room, as

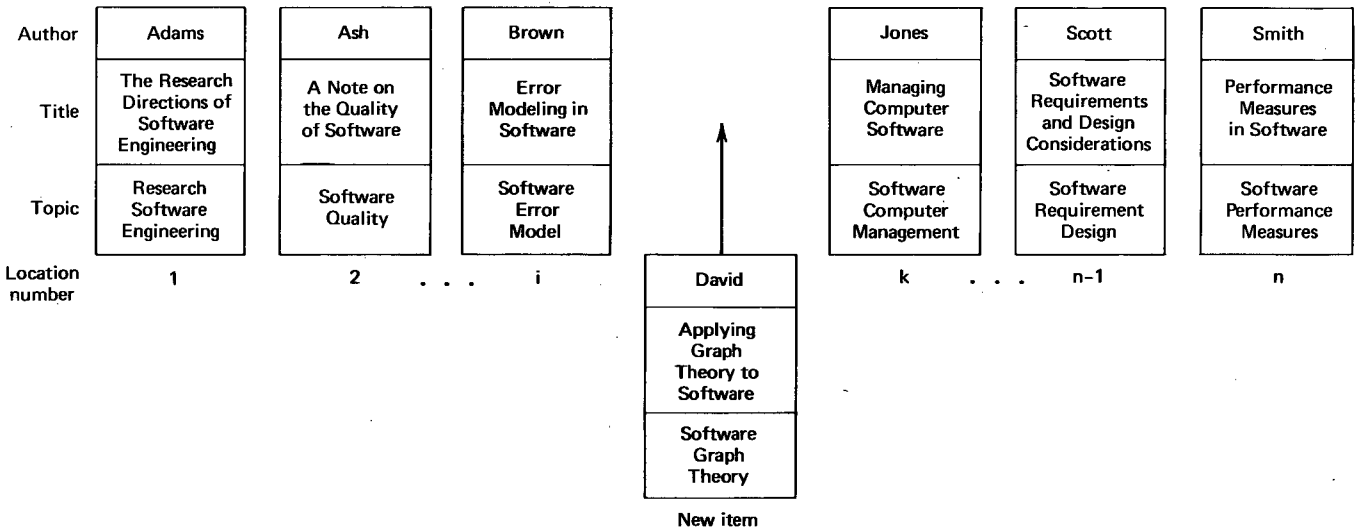

Figure 1-7 Sequential file ordered by author.

shown in the example of Fig. 1-7. On the other hand, in the linear list which is unordered a new item can simply be placed in the next available location.

To locate a specific item in an ordered sequential file it may still be necessary to look at $(n + 1)/2$ items on the average. That is, one can start at the beginning of a sequential file and work through the file item by item to find any desired element. However, the efficiency of search in a sequential file can be increased when the file is ordered by the key values used in the search. For example, a *binary search* can reduce the required number of steps to $\log_2 (n + 1)$. In other words, assuming that the file contains 1,023 items, a desired item can be found in 10 steps on the average using a binary search, but the search would take 512 steps on the average with the sequential scanning method described previously.

To use the binary search procedure the file must be ordered according to the key values and the record sought must be specifiable by citing the value of that particular key. For the example used earlier this implies that all records would be retrieved by citing only the particular author name. Instead of initiating the search at the beginning or the end of the file, the record in the middle of the file is examined first. The key value for the middle record is then compared with the key value specified in the search request. If the key values match, the middle record is the record being sought. However, if the values do not match, it is necessary to determine whether the desired key value occurs before or after the key value of the middle element in the sequential order. If the desired key value occurs after the value of the middle element, the beginning half of the file is ignored and the middle element of the remainder of the file is examined. This process continues until the desired record is located. For example, if a file consists of seven elements with key values A-B-C-D-E-F-G and the search is made for the record with a key value of C, then the items examined in order will be D then B and finally C. The middle element of the whole file, D, is examined first. Its key value does not equal the desired key value, C. The desired element logically precedes the D element; so the last half of the file is ignored for the remainder of the search process. The next comparison is made with the middle element of the first part of the file, that is, B. The desired element C comes after B; so the beginning portion of the file is ignored for the rest of the search. At this point the only remaining element is C, the desired element. The binary search process is illustrated in the example of Fig. 1-8 [21,22].

Compared with an item-by-item search the binary search is very efficient. A file of 10,000 elements in which $1/2$ second is needed for each key comparison can be searched in approximately 7 seconds, compared with the 41.6 minutes needed for the sequential search. Unfortunately, in information retrieval many searches are not conducted for specific items of known key values. Instead items must be located that appear meaningful in some sense to particular requests. Furthermore, the unique key values used to identify the records in an ordered sequential file are sometimes difficult to specify in a search request. In general several different keys are therefore specified to identify potentially relevant items, and simple binary searches are no longer useful in these circumstances.

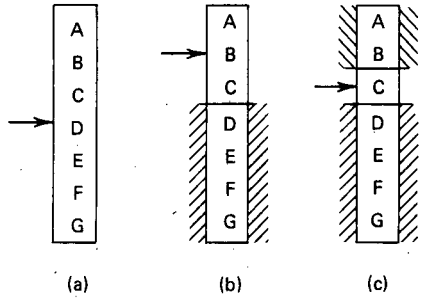

Figure 1-8 Binary search example for search key C. (a) Initial comparison with middle element (D). (b) Next comparison with middle element of upper half (B); ignore lower half. (c) Final comparison with remaining element (C); ignore upper half.

Some simple file accessing methods, including binary searches, are treated in more detail in Chapter 8.

*C Indexed Files

Another way to speed up a search for an information item is to develop an index which provides access to segments of a file. For instance, an index may be constructed using the first letter of the author's last name for the ordered sequential file of Fig. 1-7. The index identifies the location of records corresponding to the first letter of a particular last name. In Fig. 1-7 it can be seen that the authors whose last name starts with the letter J begin to appear at location number k. Figure 1-9 shows an index constructed for the sequential file in Fig. 1-7. Given a particular author name, it is then possible to use the index to find the storage location of records with author names beginning with a given letter. The search for a particular record now requires only a search of the index and a search of that portion of the file specified by the index. The number of steps needed on the average to find a specific item is reduced to the number of steps required to search the index, plus $(n + 1)/2$ additional steps for the sequential subfile where n is the number of records in that specific sequential segment. As an example, consider a sequential file of one million records. A sequential search for a record beginning with a specified letter will require $(n + 1)/2$ or 500,001 steps on the average. If an indexed file arrangement is used

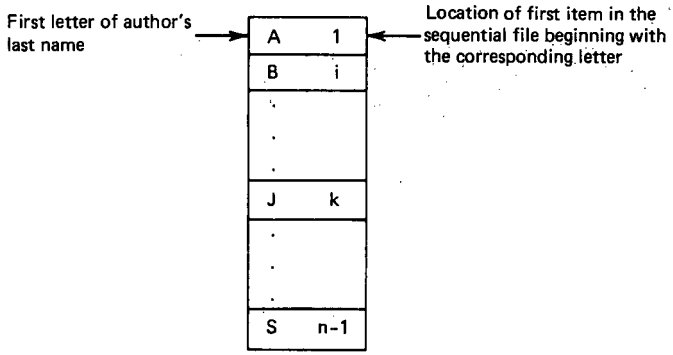

Figure 1-9 Index to ordered sequential file.

instead, the number of records with author name beginning with the specified letter may be equal to 50,000. The index would then normally consist of 26 entries, one for each letter of the alphabet. Thus, the search will require (26 + 1)/2 steps to search the index plus (50,000 + 1)/2 steps to search the records beginning with the given letter, or 25,014 steps. The search has been reduced by a factor of 20.

Indexes are particularly important for certain kinds of computer storage devices such as disks. Disks permit rapid access to consecutive records, but access to particular regions of the disk is slow. An index may then be used to locate the special region of the disk which contains the records of interest for a given query. These records can then be scanned sequentially. By contrast, a binary search requires multiple probes to various parts of the file.

Unfortunately, indexed files carry one big disadvantage. When new records must be added to an indexed file, both the file itself and the index must be changed. This can be an expensive undertaking which may outweigh the gains achieved in the search process. In addition, the previously mentioned disadvantages in updating an ordered file still apply.

On the other hand, if the search process appears most important, and the file needs to be updated only at infrequent intervals, then the index provides a valuable tool. If a record carries more than one key, such as an author name as well as individual words from the title that may also be used to gain access to the record, the idea of an index for a single key may be extended to cover all the record keys. A single index structure may then be built that includes every value for each key for the records in the file. That is, the data organization is turned around (inverted) to create an index for all unique key values in all documents as in Fig. 1-10. The illustration of Fig. 1-10 shows an information file arranged in order by item number, each item being further identified by various topic terms. A file in which the items themselves provide the main order of the file is known as a *direct file*. The inverted index, on the other hand, is arranged in order by topic, and each topic includes the corresponding list of item numbers. When an inverted index is available, each topic term is then usable as a key to obtain access to the corresponding items.

An *inverted file* ensures quick access to the information items because the index alone is examined in order to determine the items which satisfy the search request, rather than the actual file of items. Furthermore, the index is sequentially ordered by the key values. For example, to retrieve all the information items dealing with the "systems" topic, a search is made of the index to locate "systems." Items 1, 3, 4, 5 are then identified as candidates for retrieval in the example of Fig. 1-10. One need not examine the individual records to determine their actual key values, because that information is already contained in the index. It should be obvious that the addition of new items to the file is potentially laborious, because not only must a new item be placed in the main file, but the index entries relating to this new item must also be updated. The effect of adding a new item (number 6) to the file structure of Fig. 1-10 is shown in Fig. 1-11. Note, however, that the new item itself is placed at the end

		Related information items				
	Computer	1		3		
	Information	1	2		4	
Topic	Retrieval	1	2		4	5
	Systems	1		3	4	5
	Users		2			5

(a)

Item number	1	2	3	4	5
Author	Ash	Brown	Jones	Reynolds	Smith
Title	Aspects of Computerized Information Retrieval Systems	A Survey of Users of Information Retrieval	The History of Computer Systems	The State of the Art of Information Retrieval Systems	Users of New Retrieval Systems
Topic	Computer Information Retrieval Systems	Information Retrieval Users	Computer Systems	Information Retrieval Systems	Retrieval Systems Users

(b)

Figure 1-10 Sample inverted file organization. (a) Inverted index identifying item numbers corresponding to particular topics. (b) Sample information items.

of the file, or wherever convenient, and that no special order is required in the main record file.

In many cases the inverted index constitutes a file of considerable size. This index file must be searched to find a desired key value and changed to reflect the addition and deletion of key elements. To increase the efficiency of the index search, one can build an index to the index. In fact, hierarchies of indexes are often required to render quick searches possible. Unfortunately, given a hierarchy of indexes, the addition of new information items may require changes to the indexes at all levels of the hierarchy.

Inverted files are used in almost every commercially available information retrieval system. Experimental systems may use inverted files, or alternatively a direct file organization may be used, as in standard linear lists or ordered document files. The relationship between the inverted and the direct file structures is surprisingly simple. The inverted file identifies all information items associated with a given term; that is, for each term in the indexing language a list of information items indexed by that term is carried in the index. On the other hand, the direct file includes for each information item the list of terms which are associated with the item through the indexing process. If the presence of a term or of an item is indicated by a 1, and the absence is indicated by a 0, then the example of Fig. 1-12 shows the relationship between the two file structures. The inverted file is accessed using the individual terms, while the individual information items themselves are used to access the direct file. Thus an access to

		Related information items					
	Computer	1		3			6
	Information	1	2		4		6
Topic	Retrieval	1	2		4	5	
	Systems	1		3	4	5	6
	Users		2			5	

(a)

Item number	1	2	3	4	5	6
Author	Ash	Brown	Jones	Reynolds	Smith	David
Title	Aspects of Computerized Information Retrieval Systems	A Survey of Users of Information Retrieval	The History of Computer Systems	The State of the Art of Information Retrieval Systems	Users of New Retrieval Systems	A Study of Computerized Information Systems
Topic	Computer Information Retrieval Systems	Information Retrieval Users	Computer Systems	Information Retrieval Systems	Retrieval Systems Users	Computer Information System

(b)

Figure 1-11 Inverted file with added item. (a) Inverted index with a new item 6. (b) Sample information items with added item 6.

Information items

Topics		Document 1	Document 2	Document 3
	Term 1	1	0	1
	Term 2	1	1	0
	Term 3	0	1	1
	Term 4	1	1	1

(a)

Topics

Information items		Term 1	Term 2	Term 3	Term 4
	Document 1	1	1	0	1
	Document 2	0	1	1	1
	Document 3	1	0	1	1

(b)

Figure 1-12 Inverted and direct file examples. (a) Inverted file. (b) Direct file.

the inverted file for term 1 identifies information items 1 and 3 for retrieval in the example of Fig. 1-12. On the other hand, accessing item 1 in the direct file structure reveals that the item is indexed by terms 1, 2, and 4. The relationship of inverted and direct files is exactly that of matrix transposition.

The items obtained from a direct file as a result of a search must be processed to determine which, if any, are likely to be of interest to the user of the retrieval system. On the other hand, the items identified through an inverted file access are already known to be indexed by the given search term. Such items are then assumed to be of potential interest. Of course, the searcher may wish to combine different terms with each other to formulate a search request. Such processes are described in Chapter 2.

Differences in the file organizations used for retrieval play a major role in the retrieval process. For example, in the direct file an access to a particular information item reveals information concerning *all* the index terms assigned to that item. Using the inverted file, one obtains no information about other terms that may be assigned to a given information item, unless the item itself is retrieved. As a second example, consider the efficiency of the search process. If one assumes that information items are to be retrieved by citing particular search terms among the item identifiers, then an access to the inverted file immediately identifies all information items to be retrieved. On the other hand, an access to an information item in the direct file simply begins the process necessary to determine whether the particular item should be retrieved. That determination must be made for each and every information item in a direct file.

6 SUMMARY

Some of the principal information retrieval problems have been outlined in this introductory chapter. Several different types of information systems were introduced, and two simple file structures currently in use were examined in some detail. The information retrieval environment is complex. The information to be stored is often available in the form of natural language texts, and the mechanisms used to represent the texts include both controlled and uncontrolled vocabularies which separately or in combination constitute the indexing language. The indexing process transforms the natural language text of the information items into the elements of the indexing language, and the search formulation process converts statements of information need into elements of the same indexing language. The identification of information items to be retrieved then depends on the item representations rather than on the original natural language texts of the items.

The organization of the items in the files has a great deal to do with the efficiency of search and retrieval. Clearly the use of inverted files is more efficient for searching than the use of direct files. On the other hand, the inverted indexes must be stored and updated when changes occur, and these auxiliary operations may substantially affect the efficiency of the retrieval operations [23].

REFERENCES

[1] The Bowker Annual of Library and Book Trade Information, 25th Edition, R.R. Bowker Inc., New York, 1980, p. 97.
[2] V. Bush, As We May Think, Atlantic Monthly, Vol. 176, No. 1, 1945, pp.101–108.
[3] H.S. Heaps, Information Retrieval, Academic Press, New York, 1978, pp. 2–3.
[4] A.C. Foskett, The Subject Approach to Information, 3rd Edition, Chapter 21, The Library of Congress Classification, Linnet Books and Clive Bingley, Hamden, Connecticut, 1977, pp. 359–367.
[5] A. Toffler, Future Shock, Random House, Inc., New York, 1970, p. 159.
[6] F.W. Lancaster, The Measurement and Evaluation of Library Services, Chapter 5, Evaluation of the Collection, Information Resources Press, Washington, D.C., 1977, pp. 165–206.
[7] G. Salton, Dynamic Information and Library Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975, p. 7.
[8] F. Withington, The Changing Profile, Datamation, Vol. 25, No. 6, May 1979, pp. 10–12.
[9] J.M. Gabet, VLSI: The Impact Grows, Datamation, Vol. 25, No. 7, June 1979, pp. 109–112.
[10] S.A. Caswell, Computer Peripherals: A Revolution Is Coming, Datamation, Vol. 25, No. 6, May 1979, pp. 82–89.
[11] M.J. McGill and J. Huitfeldt, Experimental Techniques of Information Retrieval, Annual Review of Information Science and Technology, M.E. Williams, editor,

Vol. 14, Knowledge Industry Publications, Inc., White Plains, New York, 1979, pp. 93–127.
[12] M. Taube, Machine Retrieval of Information, Library Trends, Vol. 5, No. 2, October 1956, pp. 301–308.
[13] J. Minker, Information Storage and Retrieval—A Survey and Functional Description, SIGIR Forum, Association for Computing Machinery, Vol. 12, No. 2, Fall 1977, pp. 1–108.
[14] C. Meadow, Applied Data Management, John Wiley and Sons, Inc., New York, 1976.
[15] E.C. Carlson, editor, Proceedings of a Conference on Decision Support Systems, Data Base, Vol. 8, No. 3, Association for Computing Machinery, New York, Winter 1977.
[16] B. Raphael, The Thinking Computer: Mind inside Matter, W.H. Freeman and Company, San Francisco, California, 1976.
[17] P.H. Winston, Artificial Intelligence, Addison Wesley Publishing Company, Reading, Massachusetts, 1977.
[18] F.W. Lancaster and E.G. Fayen, Information Retrieval On-Line, Melville Publishing Company, Los Angeles, California, 1973.
[19] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill Book Company, New York, 1968, pp. 210–215.
[20] Resource Directory, Health Information Sharing Project, School of Information Studies, Syracuse University, Syracuse, New York, 1978.
[21] E. Horowitz and S. Sahni, Fundamentals of Data Structures, Computer Science Press, Woodland Hills, California, 1976.
[22] C.C. Gotlieb and L.R. Gotlieb, Data Types and Structures, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.
[23] M.J. McGill, Knowledge and Information Spaces: Implications for Retrieval Systems, Journal of the American Society for Information Science, Vol. 27, No. 4, July–August 1976, pp. 205–210.

BIBLIOGRAPHIC REMARKS

This text assumes some knowledge of computers and computer science. If the reader has little or no background in this area, any of the following texts may be helpful:

Marilyn Bohl, Information Processing, 3rd Edition, Science Research Associates, Inc., Chicago, Illinois, 1980.
Richard Dorf, Introduction to Computers and Computer Science, 2nd Edition, Boyd and Fraser Publishing Company, San Francisco, California, 1977.
Gerald A. Silver and Joan B. Silver, Data Processing for Business, 2nd Edition, Harcourt Brace Jovanovich, Inc., New York, 1977.
Fred Gruenberger, Computing: An Introduction, Harcourt Brace and World, Inc., New York, 1969.
D. Wilde, An Introduction to Computing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1973.

For a different look at information retrieval the reader may wish to examine:

F.W. Lancaster and E.G. Fayen, Information Retrieval On-Line, Melville Publishing Company, Los Angeles, California, 1973.

C.J. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, London, England, 1979.

The first text presents an excellent overview of information retrieval systems as of 1973. The second covers research problems in information retrieval and experimental retrieval systems.

EXERCISES

1-1 **a** Assume that a file contains one million information items. If the file is organized as a direct file, how many items would need to be examined on the average in order to locate one specific item from the file?

b How many items if the file is an ordered sequential file using a binary search?

1-2 Using any programming language take an arbitrary string of text and break up the text into individual words. Be sure to consider text items consisting of numbers containing decimal points such as 37.235, abbreviations such as Mr. and Mrs., and all punctuation marks. Hyphenated words such as semi-automatic should be treated as single words.

1-3 What extensions are required to the answer to Exercise 1-2 in order to handle a chemical data base which includes formulas such as H_2O and chemical names such as N-benzoyl-glycine and di-n-propyl-ether?

1-4 Describe in detail the procedure necessary to add a single new document to a system based on an inverted file. Consider both the case in which the new document is indexed by some terms not already found in the inverted file and the case where the new document is completely indexed by terms already in the inverted file. Describe this process either in a flowchart or as a step-by-step narrative.

1-5 Consider the area of American history. This topic subsumes a number of subareas such as vocabulary construction, the American Revolution, and the Civil War. Develop a one-dimensional physical organization for books on American history. How will this change if the organization becomes two-dimensional? three-dimensional? etc.