Chapter 10

Future Directions in Information Retrieval

0 PREVIEW

This concluding chapter considers new developments which may be expected to affect the information retrieval world in the not too distant future. In a few years the text entry problem will have been solved either by using character recognition equipment to recognize textual materials or through the widespread use of word processing. It will also be possible to store full documents digitally or in microform, and to use automated graphics equipment to handle illustrations and pictures. Various advanced technologies that may be helpful for the processing of full documents are briefly introduced. This is followed by a summary of theoretical approaches for the representation and analysis of document content, including fuzzy set theory, term dependency analysis, and composite document representations. A number of sophisticated automatic document processing systems are also described that will be capable of processing full documents and of servicing large user populations.

Eventually paperless electronic systems may be created which will offer a wide variety of individually tailored information services to the users. Such systems could include many facilities that are not yet currently available, in-

cluding natural language recognition, graphics processing, speech recognition, and inexpensive point-to-point communications.

1 INTRODUCTION

The material contained in the previous chapters covers the existing theory and practice in information retrieval, as well as various extensions that could in principle be implemented at once in the proper environment. This last chapter deals with new ideas and technologies that may be just beyond the current state of the art. Technological innovations are mentioned which should significantly alter the search and retrieval process as it exists today, and theoretical developments are discussed that may provide new insights into the information search and retrieval functions.

One of the characteristics of the existing operational retrieval systems is the large investment in manpower and resources necessary to provide even relatively simple retrieval services. The design and implementation of retrieval programs constitute major tasks in themselves; in addition, substantial resources must be devoted to the generation or acquisition of the data bases to be manipulated and searched. In these circumstances, even minor adjustments in procedures require careful consideration; more far-reaching innovations are often out of the question because of the large resources that are required. This may explain the tendency among many observers to think of the ideal retrieval facility as a simple extension of the currently existing systems and procedures. Thus a good deal of attention is devoted to the implementation of sophisticated user-system interfaces permitting users and search intermediaries to carry out the retrieval operations without some of the limitations that hamper the existing search efforts. From time to time new search protocols are proposed for conducting iterative searches in such a way that earlier search results are utilized to formulate improved query statements usable in subsequent search operations. The relevance feedback process described earlier is an example of such a system. Efforts are also made to extend the retrieval operations to several different data bases while merging the respective search output.

The future directions in information retrieval may be considered by reviewing theories that are currently under active consideration and in studying technologies that are likely to be prevalent in the foreseeable future. The theories of most interest deal with natural language processing systems using extended representations of information content. The new technologies include special processing "chips," microprocessors, optical character readers, optical memories, and micrographic devices. One may expect that new theoretical developments could in time be coupled to the new technologies, leading to the implementation of flexible, new user support systems capable of controlling many different file processing activities. Sooner or later, the conventional information processing systems may be replaced by "paperless" systems in which machine-readable entities are processed instead of hard-copy products, and all information flow operations are carried out electronically.

It is premature to submit a definitive design of the information system of the future. However, it is not too early to study the developments that may be expected to form the basis for the design of the information handling systems of the future.

2 TECHNOLOGICAL DEVELOPMENTS

A Automatic Document Input

The existing information retrieval systems have enjoyed increasing popularity in the last few years. As more and more items are added to the files, one may expect that an ever larger proportion of the population will become interested in using the automatic search and retrieval facilities. Unfortunately, many of the items that need to be added to the files are not currently available in a form which allows incorporation into existing data bases. In principle, it is possible to charge a typist, keypuncher, or data entry clerk with the task of converting paper documents into machine-readable form. An exceedingly proficient typist might produce errorless copy at the rate of 100 words per minute. This would generate 192 double-spaced pages of machine-readable text in an 8-hour working day, assuming no coffee breaks, lunch hours, errors, or corrections. A more reasonable rate for experienced typists may be 50 typed pages per day. As a mechanism for information input into an automated retrieval system, such a process is inconvenient and labor-intensive, particularly if one considers that the data input operation constitutes a second typing operation, following initial document preparation in many cases.

Fortunately, the data input problem may be on the way to a solution. A typing operation may still be initially required to create the original documents. This typed copy is now increasingly produced with equipment capable of capturing the information in machine-readable form. *Word processing machines* are used in many places to produce edited, machine-readable text that can be converted into a final printed product. In the foreseeable future, one may expect that increasing quantities of text will become available in a format acceptable for input to automated information retrieval systems [1–3].

The word processing concept was initially introduced as a means of simplyifying certain secretarial tasks—notably the typing of multiple copies of letters. A standard typewriter would be used supplemented by a paper or magnetic tape storage unit. The information being typed by the human operator could then be punched out or recorded automatically on the auxiliary storage equipment. When additional copies of previously typed materials were needed, the recorded tape could be used as input without additional keyboarding.

At the present time, far more sophisticated word processing stations are used, including the following components:

1 A keyboarding unit
2 A storage unit, often consisting of a floppy (soft) disk capable of storing, inexpensively, between 50 and 1,000 pages of text

3 A video display unit capable of displaying any page of stored or typed text

4 A print unit that can print out a final ouput product

5 A connection to a computer or to communications lines capable of joining several word processing stations.

The enhanced word processing equipment can now be used not only for the basic typing and further reproduction of text but also for changing, revising, and editing text. The video display unit is particularly helpful in this respect because appropriate editing commands can be supplied from the keyboard. Final copies of text can also be obtained, in a format ready for printing. Furthermore, the materials generated by word processing can be disseminated to the recipients electronically using the available communications lines. Word processors could then function as originating and receiving stations in an electronic mail system. Since a large proportion of the correspondence in a business organization is internal to the company itself, it is not hard to see that electronic word processing systems can in principle take over the vast majority of the routine communications in an office. The effect of word processing on the publishing and information retrieval world is equally far-reaching: the multiple typing operations normally needed to produce final versions of books and documents may soon be a thing of the past; instead the word processor output can be used to drive automatic typsetting equipment, and the material stored in word processors can be taken over directly by an information storage and retrieval system.

Word processing machines do not of course help in converting materials that may already be available in standard, printed form, nor do they solve the input problem for users who may not have access to machine-readable input that may already exist elsewhere. In that case it may be useful to consider *character reading* equipment using optical reading methods to convert printed information directly to machine-readable form. Optical character readers are now available that convert text printed in a variety of type fonts to machine-readable form at a rate of about 80,000 characters per hour, producing over 500 double-spaced pages in a standard 8-hour day [4].

Character readers do not require coffee breaks; however, the existing devices are error-prone, the quality of the output being dependent on the characteristics of the printed input. For example, the input character "t" represented in Fig. 10-1 may be interpreted as a "c" by a standard optical character reader when the upper portion of the character is misformed or exhibits less contrast than the lower portion of the character. The most sophisticated available character reading equipment is now able to recognize correctly about 90 percent of the characters, and about 70 percent of the full words, contained in arbitrary printed input texts. When a machine-aided, human posteditor is added to "clean up" the output of the character recognition system, close to 100 percent correct output may be produced.

The postediting process is typically carried out as a two-step process:

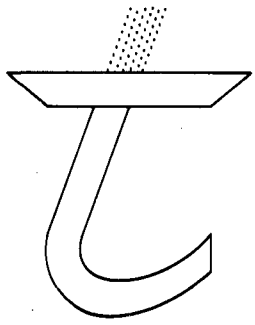

Figure 10-1 Typical input to character reading equipment.

1 An automatic process is used to highlight each character falling below a given threshold of recognition acceptability.

2 A manual character replacement phase then allows a human operator to correct the output of the optical recognition system.

The automatic highlighting process determines a level of certainty for each character as a function of the degree of agreement between features recognized in the input and the patterns stored by the recognition system. Thus, for the character "M" represented in Fig. 10-2 the recognition characteristics may include certain identified vertices as well as line segments (vectors) with appropriate directions.

Given an input character such as the one used as an example in Fig. 10-3, it may be seen that only the vector information matches. Such a character may then be recognized as an "M" with a certainty level of $^4/_7$, or 0.571. Assuming that the threshold for acceptable characters is 0.75, the corresponding input would be submitted to a human operator for appropriate action during the postediting phase. Obviously, the postediting task increases the cost and decreases the efficiency of the character recognition process.

The recognition process described earlier may furnish a serviceable solution to the text input problem. Alternatively, a training system could be used where the characteristics of some sample input are used by the system to set appropriate recognition procedures capable of handling new text whose features are similar to that of the training sample. Unfortunately, some aspects of the text input problem are still not treated satisfactorily. A few text characters may be left unrecognized even when a human editor is available, because the

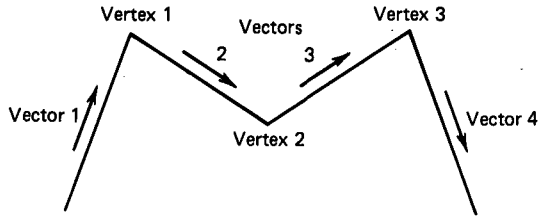

Figure 10-2 Recognition characteristics for letter "M."

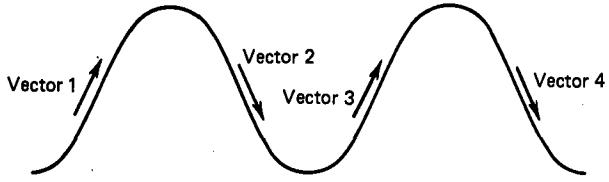

Figure 10-3 Character recognized as "M" with certainty level of 0.571.

recognition threshold may be erroneously met for misinterpreted characters. Thus, the input of Fig. 10-1 may be recognized as a "c" without ever being shown to the editor. Second, special portions of the text, such as author names, document titles, and descriptive indexing information, may require separate treatment from the remainder of the text, and the character recognition equipment may not necessarily be adapted to the identification of these special text elements. Finally, the equipment does not treat pictorial and graphical portions of documents that may also need to be processed. Special graphics equipment is needed to handle illustrations and pictures.

An automatic character recognition system might also be supplemented by an automatic *voice input* system. With such a system, information could be dictated, or read, directly to an automated typewriter or display input equipment, thereby eliminating the distinction between readability and machine readability. Some voice input systems already exist, for example, children's games are sold which properly recognize spoken input words such as "stop," "right," "left," or "go," provided the words are properly uttered in isolation. Certain computer systems in fact accept up to 1,000 individually specified words.

Unfortunately, a substantial distinction must be made between the recognition of distinct spoken words in a given language uttered by speakers to which the equipment has been properly adapted, and the recognition of normal running speech in arbitrary dialects by unknown speakers [5]. A solution to the latter problem appears to be as difficult as the natural language analysis problems previously discussed in Chapter 7. Whether the remaining character recognition problems can be handled effectively in the foreseeable future to solve the information input problem without any keyboarding remains to be seen.

B Optical Storage

In addition to the input problem, the storage problem is a primary concern in information retrieval. In situations where only limited data classes are processed—for example, in systems where the retrieval activity is based on document citations and keywords only—the conventional magnetic disk technology normally proves satisfactory. Magnetic devices are erasable and lend themselves easily to most file updating and maintenance requirements. Furthermore auxiliary index files can then be constructed and maintained to guarantee reasonably rapid access to individual information items.

In many cases, it may, however, be desirable to access or display for the user's attention copies of the full contents of articles. In that case it becomes

necessary to handle different type sizes, pictorial information, signatures and graphical data. A photographic or other optical storage medium may be needed that is capable of storing digital as well as video information. Videodiscs, holograms, and micrographic storage devices are of main concern in this connection.

Videodiscs A videodisc is a picture storage device that can be connected to the home television set. If it is used in the home, the videodisc can provide a color TV movie for the $15 purchase price of the disk. The player for the videodisc was priced at approximately $500 in 1981, but this price may decrease in time. The importance of the videodisc for information retrieval is not necessarily to keep the viewer entertained but rather to store large quantities of information at little cost [6,7].

Consider first the manner in which information is stored on a videodisc. Figure 10-4 shows graphically the method of "burning" information representations onto the disk. As information is received as input, it is translated through a series of electronic devices to control parameters which guide a laser beam. The beam in turn creates pits in the surface of a rotating master disk. These pits may then be "read" to reproduce the information.

Videodiscs are not currently capable of rewriting. That is, once the information is put onto the master disk, it cannot be erased and rewritten. However, duplicates of the master disk may be created. This process is similar to creating

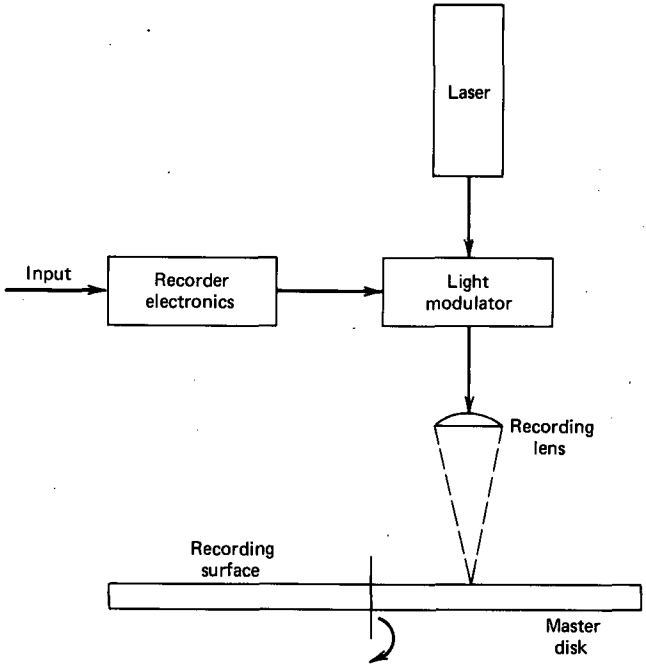

Figure 10-4 Videodisc recording.

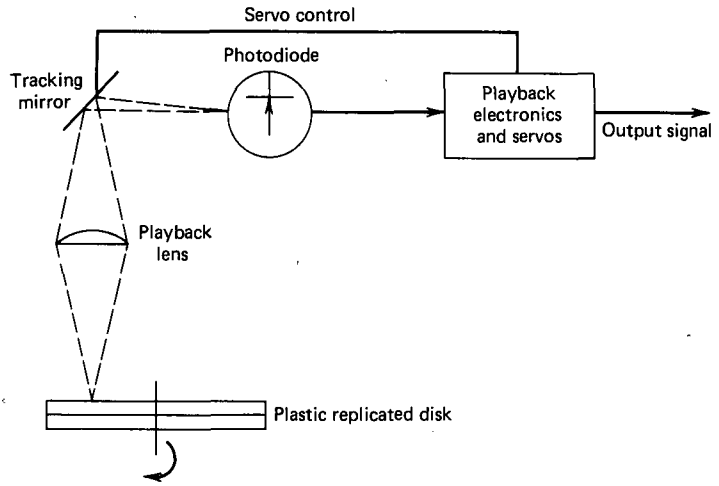

Figure 10-5 Videodisc playback.

duplicate copies of phonograph records. One rotation of the videodisc (one track) produces one visual image (one television picture). There are 54,000 concentric tracks on most current videodiscs. In the more sophisticated videodisc readers, these tracks may be accessed in any order. Figure 10-5 represents the system used to recover the information from the videodisc. Note that the playback system uses reflected laser light to encode the information.

Initial work with videodiscs indicates that 10 billion bits or approximately 1 billion characters of information may be stored on a single disk. Reproductions of a videodisc in quantities of about 1,000 will cost less than $20 per disk. This includes the cost of the disk and of the reproduction process but not of the information contained on the disk. Thus, given a data base of information items each averaging 1,000 characters, 1 million of these records can fit on a videodisc. In addition, work is currently underway to increase the capacity of a single videodisc to 10 billion characters.

An unknown factor is the error rate associated with the stored information. For the home entertainment application that is not important; however, for the storage of digital information it may be critical. That is, if one watches a TV picture and a single blue dot suddenly turns red for $^1/_{30}$ second, there is little concern. However, if that mistaken blue dot is the difference between CANCEROUS and NOT CANCEROUS in a medical data base, then it could amount to catastrophic error. Mechanisms for assuring the integrity of digitally encoded information on videodiscs remain to be worked out as of this writing. When this problem is resolved, the videodisc technology may actually produce significant changes in the processing of large data bases with periodic updates. Using videodiscs may, for instance, reduce the need for large communication networks to remote data bases. Instead, institutions such as the National Library of Medicine may simply mail periodic updates of their data bases to geographically dispersed sites. These sites will own inexpensive videodisc readers

attached to inexpensive computers for immediate and reliable access to the information.

Additional applications may also become economically attractive. For example, if it were feasible to store digital information on the videodiscs in addition to pictorial data, the combination of the two might be applicable to the storage of museum information, photography retrieval, etc. In addition, one may also opt to intersperse audio information among the recorded video and digital information. Time and thought will be required to ensure the success of the videodisc technology in the information retrieval environment.

Holographic Devices Another approach for information storage consists in using holographic devices. A hologram is a picture showing a three-dimensional object. The hologram is created by illuminating an object with laser-generated light from two angles. As the light reflected from the object hits one of the laser beams (the reference beam) an interference pattern is created. This pattern is recorded on film as the hologram. The recording operation is represented in Fig. 10-6. When a laser beam is passed through the hologram, it is deflected so that a viewer perceives the original object as shown in Fig. 10-7.

An interesting feature of holographic recording is that the holographic image is literally dispersed across the entire hologram. That is, if a hologram is cut in half and illuminated, the entire object may still be perceived. The resolution of the object is reduced, however.

If the object photographed by the holographic device is replaced by a digital pattern representing characters or other digitally encoded data, then the digital data will also be redundantly stored in the hologram. Reproduction of the data requires that the display of the hologram be focused on a photodetector device to convert the picture back to digital signals.

Holographic storage of information is "safe." Because of the natural re-

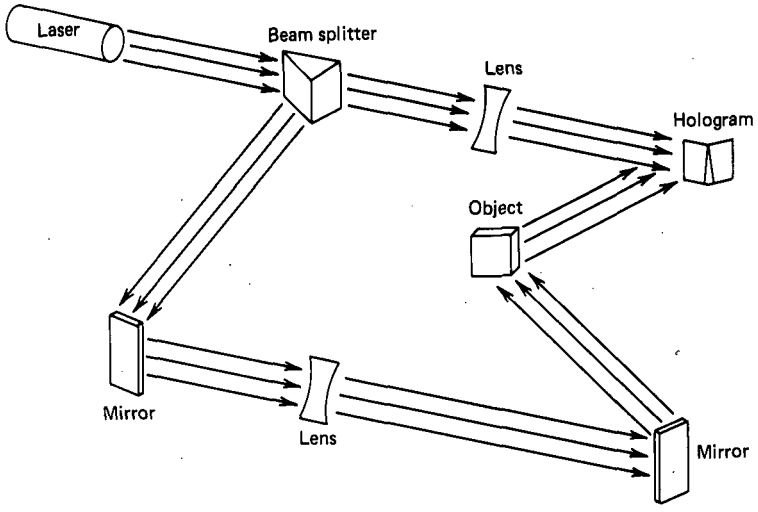

Figure 10-6 Holographic recording.

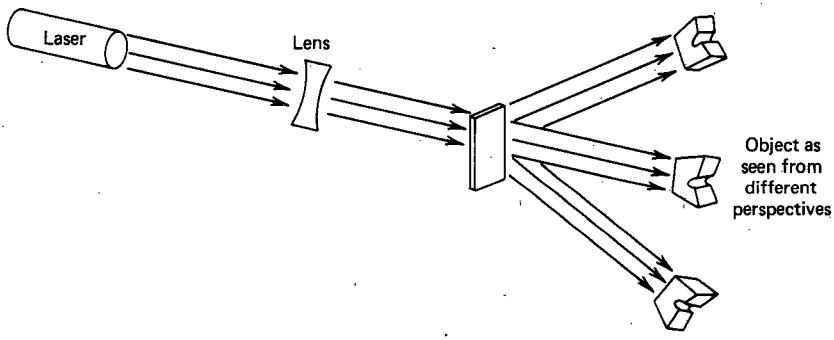

Figure 10-7 Holographic viewing.

dundancy, a substantial portion of the hologram may be destroyed without actually losing the information. The hologram is also capable of huge quantities of storage. It has been stated that a single 4 by 6 inch piece of film can hold up to 20,000 individual holograms or 20 million characters of information. A device for reading this information may cost as little as $5,000 [8].

The holographic storage devices, like the videodiscs, are not capable of rewriting. However, reproduction costs of $1 or even less for a 4 by 6 inch film (fiche) is possible, and as for videodiscs, holograms may be interspersed with video images, making new applications a real possibility.

Micrographic Storage Micrographic storage devices, such as microfilm or microfiche, provide another solution to the storage problem of bulk materials in retrieval. Micrographic devices have been available for many years and have proved to be cost effective in many applications [9]. Micrographic storage, like video equipment, is used to store vast quantities of information, the storage efficiency being achieved by using a substantial size reduction (typically by a factor of 24) for the recorded information. A standard 4 by 6 inch microfiche thus normally stores 98 pages of digital or pictorial information.

Unfortunately, the microform technology does not easily permit erasing or modification of the recorded information. Its application in information retrieval is normally found in situations where the existing file remains largely unchanged, and new film or fiche is used when additional documents are to be incorporated into the file. An additional disadvantage of this equipment is the substantial discomfort felt by many users when confronted by this technology. The size reduction renders necessary the use of special reading equipment which magnifies the recorded information before viewing by the user. Most people would prefer hard-copy document output that can be carried around and directly manipulated to the somewhat remote microforms.

For storage and retrieval purposes special fiche storage devices are used from which a particular fiche can be extracted on demand. It is customary to use the fiche number as the main identification, although in principle, the contents of a particular fiche can be specified by using conventional keywords and index terms. However, since a single fiche may contain many pages of disparate material in addition to pictures and graphs, the full content of a fiche is not

easily captured by standard methods. A particular fiche extracted from the storage device may be automatically sent to a microfiche reader for viewing purposes. Alternatively a microfilm copy can be made for the user's personal attention, or paper copies can be produced of certain documents using a microform-to-hard-copy conversion device.

A typical document processing system using micrographics equipment is shown in Fig. 10-8. Optical character recognition equipment is provided in the illustration for input purposes using the semiautomatic process previously described in which a human operator resolves input ambiguities using a special editing terminal. A combined magnetic and optical storage system is then used for storage and retrieval. Specifically, a standard search is conducted using the document content descriptions stored on conventional disk equipment. Assuming that the standard search produces the location identifications (fiche numbers) of the corresponding full document texts, the corresponding fiches can then be extracted from the fiche storage device for further processing.

In the combined magnetic and optical storage system of Fig. 10-8, the microfilm processing system is completely mechanized. Furthermore, all the graphics equipment is currently in existence and can be bought off the shelf [10]. However, the physical movement of the microforms from the storage area to the viewing position requires a mechanical transport device which is cumbersome and expensive. More generally, the marriage between the magnetic and graphic storage technologies represented in the system of Fig. 10-8 is somewhat forced. Different accessing keys are used in the two systems and the pictorial portion of the information stored on the microforms is not directly used in retrieval. For this reason, systems such as the one represented in Fig. 10-4 will not provide long-range solutions to the existing information retrieval problem.

Efforts have recently been made to permit a digitization of micrographic images following a normal optical scan. That is, as each image is scanned, the images are converted into digital form; the converted image can then be displayed on a normal cathode-ray tube display screen, or entered into computer storage. The perfection of these techniques may solve the interface problem between the micrographic and digital storage equipment and produce a harmonious system involving both technologies [11].

Following a description of certain novel retrieval theories, various extended information systems are considered whose design is based in part on the microprocessing, character reading, and graphics equipment previously mentioned. Such extended systems may then provide a solution to the existing information problems.

3 INFORMATION THEORIES AND MODELS

A Natural Language Processing

In attempting to predict the design of future information systems, it appears best not to concentrate on the existing retrieval practice but to identify features that are generally considered desirable but are currently beyond the state of the

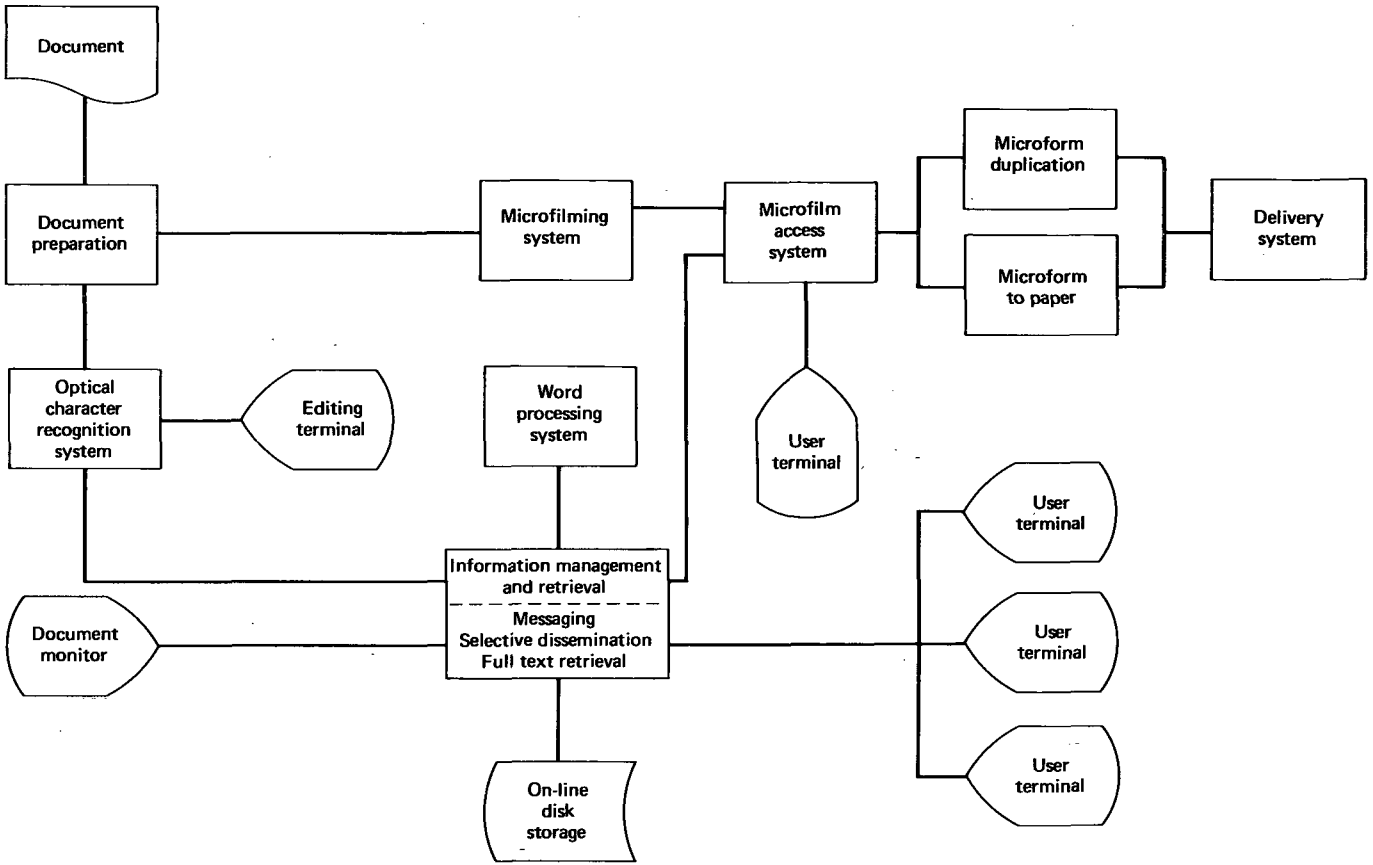

Figure 10-8 Document management system using micrographics.

art. Without a doubt, the most immediately important problem of this kind is the understanding and processing of the written and spoken natural language [12]. If user queries could be submitted using a conventional, natural language formulation, and texts could be automatically analyzed, abstracted, translated where necessary, and properly classified according to the document content, then the main difficulties which currently hamper the information system user would automatically disappear. No longer would it be necessary to worry about the design of controlled indexing languages; the need for well-trained, consistent indexers no longer arises; and the burden currently placed on users and search intermediaries in the formulation of useful queries is lifted.

Unfortunately, the discussion in Chapter 7 on natural language processing indicates that the free manipulation of unrestricted natural language data is not a likely prospect for the foreseeable future. In particular, no agreement exists about the best way for formalizing document content, about the world knowledge (above and beyond the specialized knowledge in a given subject area) that may be needed to understand texts and interpret natural language statements, and about reasoning strategies, inferences, and deductions that may be needed in order actually to respond to user inquiries [13–14].

The approach in information science has been to use specialized techniques in an attempt to provide at least some linguistic input to the standard document analysis and retrieval process [15–16]. Methods have notably been developed for recognizing and assigning noun phrases as document and query identifiers, and syntactic techniques have been used for dealing with individual document sentences instead of complete documents [17–20].

The primary difficulty in natural language processing is due to the flexibility and by extension the ambiguity of most languages. Many ways exist for expressing the same thought; hence it is easy in principle to generate documents and to formulate information requests. At the same time, a given utterance may receive many different interpretations depending on the context in which it appears; hence it is difficult to generate definite or unique interpretations of natural language materials. The approach mentioned earlier in the discussion on character recognition can be followed by considering the use of semiautomatic language analysis methods. That is, the machine might automatically identify potential ambiguities, and a human operator could then attempt to resolve them [21–22].

Various other techniques have been used in information retrieval in attempting to cope with the language ambiguity problem. In the vector processing model, special weights are attached to the terms that identify documents and search requests, reflecting the degree of importance of each term or the degree of certainty with which the term describes the document content. Global similarity computations between weighted term vectors are then used to indicate the relationship between pairs of documents or between queries and documents [23–24]. Thus, the degree of affinity between two items can vary depending on the certainty or interpretation of the respective content identifiers.

Somewhat similar characteristics are exhibited by the fuzzy set and probabilistic approaches to retrieval that are briefly introduced in the next few paragraphs.

**B Fuzzy Set Theory

The basic idea in fuzzy set theory is that elements or entities can be assigned to sets to varying degrees. That is, instead of either including an element in a given set or excluding it from the set, a *membership function* is used to express the degree to which the element is a member of the set. This concept is of interest in language processing, and by extension in information retrieval, because the assignment of individual words to meaning categories is a fuzzy process; so is the assignment of documents to concept categories, and hence also the retrieval of documents in answer to certain queries [25].

In language processing, various attempts have been made to use the fuzzy set approach to model linguistic vagueness, ambiguity, and ambivalence. For example, given a set of well-defined meaning categories, the total meaning of a given word might be expressed as a weighted combination of the membership functions of that word in various meaning classes. Linguistic quantifiers ("most," "some," "a few") and hedges ("sort of") might also be described by using fuzzy measurements of some kind [26–28].

In information retrieval, the fuzzy set approach can be used to classify the documents into fuzzy affinity classes and also to control the actual retrieval process. Consider first a document DOC and a particular term A. If **A** denotes the "concept class" of all items dealing with the subject denoted by A, then the membership function of document DOC in set **A** may be denoted as $f_A(DOC)$. In the usual terminology $f_A(DOC)$ represents the weight of term A in document DOC. Given a number of concept classes **A**, **B**, . . . ,**Z** representing various subject areas, it is now possible to identify each document by giving its membership function with respect to each of the concept classes, that is,

$$D = (f_A(DOC), f_B(DOC), \ldots, f_Z(DOC)) \tag{1}$$

Expression (1) thus takes the place of the normal term vector used in the vector processing model of retrieval.

Given document representations of the kind shown in expression (1), all the usual vector processing operations are now expressible as fuzzy set operations. In particular, the distance (or similarity) between two documents or between a document and a query may be obtained as a function of the differences in the membership functions of the two items in corresponding concept classes. Specifically, given T different concept classes, the *fuzzy distance* between documents DOC′ and DOC″ might be computed as

$$d(DOC',DOC'') = \sum_{x \in T} (f_x(DOC') - f_x(DOC''))$$

or alternatively

$$d(DOC',DOC'') = \sqrt{\sum_{x \epsilon T} (f_x(DOC') - f_x(DOC''))^2}$$

Ranked retrieval is achieved by retrieving the documents in order of increasing fuzzy distance from the query.

One attractive feature of the fuzzy set approach is the possibility of extending the definition of the membership function from single terms to combinations of terms. Thus, given the membership functions of document DOC with respect to terms A and B, the following rules apply for Boolean combinations of terms [25]:

$$\begin{aligned} f_{(A \text{ AND } B)}(DOC) &= \min(f_A(DOC), f_B(DOC)) \\ f_{(A \text{ OR } B)}(DOC) &= \max(f_A(DOC), f_B(DOC)) \\ f_{(\text{NOT } A)}(DOC) &= 1 - f_A(DOC) \end{aligned} \tag{3}$$

It is easy to verify that the fuzzy set rules of expression (3) satisfy the normal rules of Boolean algebra when $f_A(DOC)$ and $f_B(DOC)$ are restricted to the values 0 and 1. The fuzzy set retrieval model thus represents an extension of the normal Boolean retrieval system to the case where the assignment of weighted identifiers is possible for the documents but not for the queries.

Since the vector space model effectively supplies term weights for both the documents and the queries, the fuzzy set model is in a sense intermediate between a conventional Boolean query system where no term weights are allowed and a vector processing system. The attractions of the fuzzy set model are its compatibility with the standard Boolean query processing system and the interpretation of the fuzzy weights as linguistic indicators of term ambiguity or ambivalence. A number of attempts have been made to use fuzzy set models in retrieval; however, the resulting systems have never been evaluated, and the linguistic relationship has not so far been exploited [29–34].

**C Term Dependency Models

The flexibility and ambiguity characteristics of the natural language are reflected in the fuzzy set retrieval model where terms and other content identifiers apply to some degree in particular cases. A different approach consists in using as a measure of term importance values expressing the likelihood that a given term occurs in a particular environment, or the likelihood that the term should be assigned to a given document. Such *probability measures* can be used as term weights, and the documents can then be retrieved in ranked order according to the probability of relevance to the given queries.

Given a particular document identified by a binary term vector $x = (x_1, x_2, \ldots, x_t)$ the optimal retrieval rule in the probabilistic model consists in first evaluating $P(x)$, the probability of occurrence of x in a relevant document,

and $Q(x)$, the probability of occurrence of x in a nonrelevant document, and then in retrieving the document represented by x whenever $P(x) > Q(x)$ [35–36]. The actual values of the probabilities $P(x)$ and $Q(x)$ depend on the occurrence probabilities of the individual terms $x_1, x_2, \ldots, x_t$ in the relevant and nonrelevant documents of a collection, respectively.

It was mentioned earlier that under the assumption that the terms are assigned independently to the documents of a collection, the probabilistic retrieval model produces an optimal term weighting function TERMREL, known as the "term relevance," representing the logarithm of the proportion of relevant documents in which a term occurs, divided by the proportion of nonrelevant items in which the term occurs [37–38]. The normal linguistic intuition makes it clear, however, that words or concepts do not occur in the language independently of each other. Hence one may expect that retrieval models that are based on the term independence assumption will not reflect reality very accurately. A more realistic treatment would take into account various similarities and dependencies between the terms.

In the fuzzy set model of retrieval, the notion of term dependence is directly built in, in that many terms may be assigned to common concept classes and somewhat different membership functions may be used to reflect differences between them. In probabilistic retrieval the same effect is obtainable by computing *term dependence* probabilities as a function of the probability of co-occurrence of several terms in the documents of a collection, compared with the occurrence probabilities of the individual terms alone.

In fact, exact formulations for the occurrence probabilities $P(x)$ and $Q(x)$ of a given document vector x in the relevant and nonrelevant documents of a collection are provided by the Bahadur-Lazarsfeld expansion (BLE). This expansion takes into account the occurrence probabilities of the individual terms x_i, as well as all dependencies between term pairs, triplets, and all higher-order dependencies [39,40]. The correct retrieval strategy then consists in using the BLE expansion to compute the probabilities $P(x)$ and $Q(x)$, and in retrieving a document represented by x whenever the expression $\log[P(x)/Q(x)]$ is sufficiently large. Unfortunately this method is not immediately usable in practice because an exponential number of pairwise and other higher order dependencies exist between the terms, and because no obvious way is provided for estimating all the required probabilities.

Various techniques suggest themselves for constructing approximations to the exact process. One possibility consists in using a *tree dependence* model where each term is allowed to be dependent on only one other term [37,41,42]. Another approach consists in using the BLE expression in a suitably truncated form. A possible measure of pairwise term dependence is given for two sample terms x_i and x_j as

$$I(x_i, x_j) = \sum_{\substack{x_i=0,1 \\ x_j=0,1}} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{4}$$

where $P(x_i,x_j)$ is the joint probability of occurrence of both terms x_i and x_j in the documents of a collection, $P(x_i)$ and $P(x_j)$ are the individual occurrence probabilities of the two terms, and the summation is taken over the four combinations of values obtained by assuming values of 0 or 1 for both x_i and x_j [39]. Since the probability of occurrence of a given document vector x [that is, $P(x)$ and $Q(x)$] must be computed separately for the relevant and nonrelevant documents of a collection, the probability values in the term dependence expression (4) must also be obtained separately for the relevant and the nonrelevant documents of a collection. That is, expression (4) may be assumed to be valid for the relevant documents, and a similar expression with Qs replacing all the Ps would apply to the nonrelevant items.

A realistic method for estimating these probabilities consists in using the *relevance feedback* strategy introduced in Chapter 6. In relevance feedback, tentative search requests are first submitted to a retrieval system; the relevance or nonrelevance of initially retrieved items is then used to generate improved query formulations that are more similar to the documents identified earlier as relevant than the original queries, and less similar to the documents identified as nonrelevant. The reformulated queries are then processed in a second search operation in the expectation that additional useful items may be retrieved. The query reformulation process may be repeated until the user is satisfied with the retrieval results [43–45]. If the relevance feedback process is used for a given document collection over some period of time, relevance information should eventually be obtainable for a substantial number of documents. This implies that the occurrence characteristics of a large number of terms in the relevant and nonrelevant retrieved documents should become known.

It remains then to estimate the occurrence probabilities of the terms in the relevant and the nonrelevant documents of a collection by looking at the corresponding occurrence probabilities in the *relevant retrieved* and in the *nonrelevant retrieved* items, respectively. If $P(x_i)$ and $Q(x_i)$ denotes the occurrence probability of term x_i in the relevant and nonrelevant items of a collection, the estimated values could be obtained as

$$P(x_i) \text{ estimated} = \frac{\text{number of relevant and retrieved items containing term } x_i}{\text{number of relevant and retrieved items}}$$

$$Q(x_i) \text{ estimated} = \frac{\text{number of nonrelevant and retrieved items containing term } x_i}{\text{number of nonrelevant and retrieved items}} \tag{5}$$

The probability estimation methods of expression (5) have been used experimentally with good success [41,45]. However, conclusive evidence about the usefulness of including term dependencies in retrieval is still lacking. In particular, practical methods remain to be worked out for choosing a small number of the most important term dependencies, and for actually generating the corresponding term dependence factors used in the probability formulas.

*D Composite Document Representations

The fuzzy set and probabilistic retrieval models constitute enhancements of the conventional retrieval environment in the sense that term ambiguities and term dependencies are taken into account. Another, even more fundamental question, concerns the proper choice of the initial, basic concepts to be used for the representation of document content. In principle, the vocabulary included in query and document formulations must be used as a basis for the choice of content identifiers. However, it is not clear a priori how many basic concepts ought to be used, and how these concepts are to be defined.

One possible approach consists in taking an indexed document collection in which term vectors are used to represent the various documents, and in performing a *factor analysis* to derive the few independent concepts from the larger class of initially available terms. Unfortunately, a factor analysis process is expensive to carry out, and becomes impractical when more than a few hundred documents and queries are involved.

A more efficient method for generating a set of independent basic concepts may be obtained by first choosing a set of *core documents* from a sample document collection, the core being selected in such a way that core items have no common terms. The assumption is then made that the full concept space corresponds to the space initially defined by the core documents. Specifically, the documents outside the core are processed one at a time, and an appropriate place is found in the initial core concept space for each term included in these documents. That is, each of the new terms is defined as a combination of the originally chosen concepts [46,47]. Various methods can be used for defining the mapping of the new terms into the original concept space. So far, no conclusive information is available about the effectiveness of this type of process.

In addition to document content represented by the standard content terms, or keywords, documents may also be characterized by *scope,* or *extent;* survey articles tend to have greater scope than research notes dealing with specific technical subjects. Documents also carry various degrees of *influence;* many articles are effectively "lost" in the sense that they are unknown and are never referred to by other items in the literature; other documents become famous and serve as starting points for additional developments. Individual documents may also be characterized by citing other items in the literature that may exhibit similar or related information content. For example, a document dealing with a legal problem may be explicated by citing related cases covering legal precedents; similarly, a religious article may be better understood by adding references to authorized interpretations of the text.

This suggests that the normal vector representations of document content in which a term represents a particular content identifier be extended to cover these other factors. In particular, a document might be identified by using three types of identifiers:

1 Normal content terms including single words, phrases, and thesaurus categories

2 Factual or objective identifiers covering proper names of authors, publishers, dates, language of publication, and so on

3 Interpretational identifiers representing related documents designating scope, influence, and other interpretations

It has been suggested that the interpretational aspect of document content be represented by citations to other related bibliographic items [48,49].

A particular document might thus be represented as

$$D = (d_1, \ldots, d_k, o_{k+1}, \ldots, o_n, c_{n+1}, \ldots, c_t) \tag{6}$$

where d_f represents the weight of the fth content term, o_m represents the mth objective identifier, and c_p represents the pth citation to a related document.

Instead of using simple bibliographic citations to related documents for content representation, the *cocitation strength* could be used, defined as the degree to which two particular documents are cited in common [50].

The extended vector representations can be utilized in retrieval experiments by introducing *composite matching* coefficients to measure the similarity between a document and a query as a combination of the similarities for the various classes of identifiers [51]. The similarity SIM(D,Q) between a document D and query Q might then be computed as

$$\begin{aligned} \text{SIM}(D,Q) = {} & \alpha[\text{content identifier similarity}] \\ & + \beta[\text{objective term similarity}] \\ & + \gamma[\text{citation similarity}] \end{aligned} \tag{7}$$

for suitable weighting factors α, β, and γ. Much work is still required in the future for the generation and evaluation of useful extended content representations and composite query-document comparison methods.

4 ADVANCED INFORMATION SYSTEMS

A Mixed Information Retrieval Systems

Bibliographic retrieval is concerned with the processing of books and documents, available generally in natural language form. In many situations, even the most advanced bibliographic retrieval system will not satisfy the users' needs. Consider, for example, the problem of determining whether a particular chemical substance is toxic, that is, proves harmful to human beings who come in contact with it. In that case, textual data must be coordinated with numeric information produced by tests conducted with the substance in question. The user will thus require various resources in order adequately to resolve the toxicity question, including bibliographic information related to the chemicals, health effects, production processes, chemical uses, and so on. In addition experimental output and test information are needed about the production and

use of chemicals, and about the known hazards of related substances as they affect populations in certain geographic areas.

The factual information relating to chemical characteristics and hazards differs from normal textual data in the sense that the *values* of specific attributes are important rather than the attributes themselves; furthermore, these values require analysis instead of simple retrieval. Data base management systems are normally designed to offer storage and retrieval capabilities for stored numeric information, as well as data analysis capabilities to perform statistical computations and to produce output summaries.

The chemical toxicity problem may then be attacked by supplying a variety of resources including in particular those contained in data management as well as bibliographic retrieval systems. More generally, an increasing need exists in many situations for sophisticated information systems capable automatically of coordinating related information from many sources. For example, an article dealing with the development of a chemical information system by a specific chemical company might be supplemented by information about the governmental regulations affecting data required from chemical companies, and by articles dealing with the availability of other chemical information systems, or articles written by the same author or citing the original article.

A system capable of pulling together many different but related resources would obviously expand the user's knowledge of a given problem by supplying a particular piece of information, as well as the context for this information. The ultimate information system is thus often conceived as a *network* of different information facilities including bibliographic retrieval, data base management, data analysis, citation indexing, and text processing systems. In order to utilize such a network effectively, the user ought to understand the capabilities of each resource and the interactions between them. System aids could be provided in the form of catalogs of available facilities, sample search protocols, lists of search commands and their effects, tutorial sequences, and so on.

Of particular interest in this connection are common interface systems which would permit the user to formulate in a common language information requests which are to be submitted to many different information systems. It is then up to the interface system to convert the original user statements into the different internal languages required by the various resources [52]. Common interface systems are helpful in that they render it unnecessary for users to concern themselves with the internal details of many different information systems. Of even greater interest might be systems that could themselves decide, given a particular query, which of many resources should be called upon to furnish the answer. In this mode, the system "knows" the capabilities of different resources and decides how best to use the available facilities to answer a given search request.

The design of such intelligent information systems is still somewhat beyond the state of the art. However, considerable advances are being made in developing special-purpose systems that facilitate access to many different resources and relieve the user from the operational details that must normally be mastered before file access can be obtained [53–55].

B Personal Computing and Paperless Information Systems

A special class of information system that has been much in the news in recent times is designed to mechanize the paper handling routines normally taking place in office environments. An extension of these *office automation* systems are the so-called paperless information systems from which the hard-copy paper products are eliminated [56,57].

Until recently the supply of paper was plentiful and the printing operation was inexpensive. As a result the production of large journals and the bulk mailing of heterogeneous materials to large classes of recipients appeared justified. Printing and paper costs have been increasing steadily in recent years. At the present time, it appears much more important to target the disseminated information precisely to recipients who are actually likely to be interested in receiving each item. This results in the production of smaller, more specialized journals, and eventually in the replacement of journals by electronic distribution systems where individual items are distributed only to specially intended receivers.

The following main points are often cited in support of the paperless information systems of the future:

1 The volume of material to be processed and stored is becoming too large to be handled by a full-text, hard-copy paper system.
2 The materials of interest to any one person are becoming increasingly fragmented over many files, books, and journals, and the time available for finding and selecting the interesting portions is limited.
3 The cost of paper publication continues to increase at precipitous rates, in part because the publishing industry is labor-intensive and has not so far fully benefited from automation.
4 The conventional publication system continues to experience increasing delays in publishing research materials in part because the work force that originates the materials increases while at the same time the scope of many journals is constricted by rising costs and decreased subscription income.
5 An increasing proportion of the information materials is initially available in machine-readable form produced as a by-product of word processing or automated typesetting systems.

A situation is then postulated in which users, including office workers, scientists, technical personnel, and so on, have access to personal on-line console terminals. These terminals would be used to receive text, compose letters and documents, search for stored information, seek answers to factual questions, build information files, converse with colleagues by sending and receiving messages, receive mail, and generally conduct many kinds of information processing activities. Many different files could be maintained in such a system, including personal files accessible to the owner alone, mail files for incoming and outgoing messages, central files belonging to a given organization, and external files such as those maintained by outside information banks and libraries. A summary of various files and file operations of interest in such an environment is contained in Table 10-1.

Table 10-1 File Operations and Files for Typical Office Automation System

Typical operations	Typical files involved
Information search	Text files in digital form
Information retrieval	Text files in microform medium
Storage of new information	Personal files including personal data, addresses, comments, personal bibliographies
Refiling of old information	
File maintenance	
Data computations	Mail files including data and status of incoming and outgoing messages
Text and letter composition	
Text processing including correction, editing, hyphenation, justification, and so on	Central business files including information and status of stored correspondence
Communications, that is, sending and receiving of messages	External files such as generally accessible library files
	Master index files giving access to other files

In an automated file processing system a particular information item may be contained in several different files, for example, in certain publicly accessible files as well as in a variety of private files. Users may be allowed to use their own access points (content identifiers, keywords, etc.) to find a particular item, and different access conditions may obtain for the different users. Nevertheless each item need be stored only once assuming that index files are available to translate the different types of content descriptions into pointers referring to a given common storage location.

The obvious advantages of an automated, personal file processing system include the possibility rapidly to access a large variety of information products, including possibly the full text for many items, the convenience of being able to maintain and search private and public files at various levels of specificity and complexity, and the saving in space and paper handling.

The trend toward personal computing is accelerated by the recent startling advances in VLSI (very large scale integration) technology which makes possible the construction of powerful small computers coupled to large inexpensive storage devices (32-bit microprocessors with 128,000 bytes of storage) at costs not much in excess of $1,000 [58]. (See Table 10-2.) This technology makes it possible to expand computer services by simply furnishing small machines to large classes of potential users. This possibility is especially attractive in view of the increasing availability of personalized information services variously known as teletext or viewdata, where prestored information is disseminated on demand to individual recipients [59], and by advances in communications and networking that may in time allow the transmission of large masses of data using novel technologies such as fiber optics transmission lines [60].

The use of small, individual computers avoids some of the resource allocation problems which complicate life when a single large computer is used to service large user classes. Furthermore, users may prefer being "in charge" of their own machines, instead of having to contend with the restrictions normally imposed by computing center rules and regulations.

Table 10-2 Microprocessor Configurations

Configuration	Typical use	Typical cost
Handheld calculator Digital readout Input keyboard Several dozen storage registers Arithmetic unit Automatic sequence control	Small programs and calculations	$50–$200
Minimal processor 16–32 bit address 32,000 to 128,000 characters of data storage 2 to 5 microseconds add time Floppy disks for storage Cathode ray tube display	Word processing capability Text editing Program editing Indexing Cataloging	$1,000–$5,000
Full microprocessor configuration Printer added to minimal processor Hard disk for bulk storage	Information retrieval Inverted file processing Output printing	$10,000–$15,000

On the other hand, the future paperless systems also raise difficult legal and social problems that have not so far been adequately considered: for example, methods must be worked out for safeguarding the interests of a variety of parties in the information chain, including authors, publishers, information product vendors, and so on. At the present time royalty payments are not made when a user withdraws an item from a conventional library. An automated system with point-to-point communication facilities is, however, equivalent to an unrestricted, universal photocopying system, where anyone can easily obtain all stored materials. This possibility beclouds the future of the publishing industry and of the conventional library systems [61]. A good deal of thought must also be given to the role of the existing copyright legislation that protects ownership of information, as well as to the financial arrangements that must be made between the parties involved in the electronic communications system and to the provisions required for safeguarding the information stored in the publicly available files.

One may expect that many of the objections to the institution of paperless information systems will eventually disappear as the automatic systems become more sophisticated and more user-friendly. In time, a large proportion of the manual transactions currently performed with paper systems should be carried out automatically with electronic counterparts. And even if some paper products were to be maintained indefinitely, a more rational and hopefully more convenient information handling system will surely be instituted to serve the growing number of information users.

5 CONCLUSION

After reading some of the foregoing material, it will be obvious that a wealth of information and know-how exists about the theory and practice of information

retrieval. Many insights relating to the information retrieval problem are obtainable by studying elements of related areas such as decision theory, artificial intelligence, software engineering, information theory, combinatorial mathematics, linear algebra, computational linguistics, pattern recognition, scene analysis, and logic. Viable solutions to the information problem will eventually be found by combining results derived from these various disciplines.

Many challenges lie ahead for researchers and practitioners in information retrieval. First, the knowledge and technology already available ought to be incorporated into the existing retrieval system implementations. The brief summary of technological and theoretical developments contained in this chapter indicates how much can already be done to improve the current information handling facilities. Second, there is a continued need for more basic work in various areas where progress has been relatively lacking, such as language processing, voice recognition, graphics storage and display, and the performance analysis of existing or projected systems.

There is little doubt that the pressures of organizations and individuals who demand a more effective utilization of information will continue to increase. The importance of timely and useful information necessary to carry out almost any task continues to grow, and the penalties in time, effort, and money to be paid when information resources remain unused or poorly used are fast increasing. A continued need then exists for high quality work in information retrieval and for the training of individuals knowledgeable in the various aspects of information handling. It is hoped that this text will make a modest contribution toward that aim.

REFERENCES

[1] J. Whitehead, Word Processing: An Introduction and Appraisal, Journal of Documentation, Vol. 36, No. 4, December 1980, pp. 313–341.

[2] R.M. Woelfle, The Impact of Word Processing on Engineering Communications, IEEE Transactions on Professional Communications, Vol. PC-23, No. 4, December 1980, pp. 159–163.

[3] J.C. Lawlor, What Can WP Do for Managers, in Information Choices and Policies, Proceedings of the ASIS Annual Meeting, American Society for Information Science, Vol. 16, Washington, D.C., 1979, p. 355.

[4] M.J.F. Poulsen, Optical Character Readers, Encyclopedia of Computer Science, A. Ralston and C.L. Meek, editors, Petrocelli/Charter, New York, 1976, pp. 1017–1022.

[5] A. Newell, J. Barnett, J.W. Forgie, C. Green, D. Klatt, J.C.R. Licklider, J. Munson, D.R. Reddy, and W.A. Woods, Speech Understanding Systems, North-Holland/ American Elsevier, London, 1973.

[6] K. S. Winslow, Videodisk in Your Future, Educational and Industrial TV, May 1975, pp. 21–22.

[7] P.B. Schipma and D.S. Becker, Text Storage and Display via Videodisk, Proceedings of the ASIS Annual Meeting, Vol. 17, American Society for Information Science, Washington, D.C., 1980, pp. 103–105.

[8] T.H. Maugh II, Holographic Filing: An Industry on the Verge of Birth, Science, Vol. 201, August 1978, pp. 431–432.
[9] National Micrographics Association, 1978-1979 Guide to Micrographic Equipment, Vols. 1 and 2, 7th Edition, New York, 1979.
[10] C.F.J. Overhage, Plans for Project Intrex, Science, Vol. 152, No. 3725, May 20, 1966, pp. 1032–1037.
[11] G. McMurdo, The Interface between Computerized Retrieval Systems and Micrographic Retrieval Systems, Journal of Information Science, Vol. 1, 1980, pp. 345–349.
[12] National Technical Information Service, Natural Language Processing, Special Bibliographies, Vols. 1 and 2, 1964-77 and 1978-79, Springfield, Virginia, 1980.
[13] T.R. Addis, Machine Understanding of Natural Language, International Journal of Man-Machine Studies, Vol. 9, No. 2, March 1977, pp. 207–222.
[14] G. Silva and C.A. Montgomery, Knowledge Representation for Automated Understanding of Natural Language Discourse, Computers and the Humanities, Vol. 11, No. 4, July–August 1977, pp. 223–234.
[15] K. Sparck Jones and M. Kay, Linguistics and Information Science, Academic Press, New York, 1973.
[16] C.A. Montgomery, Linguistics and Information Science, Journal of the ASIS, Vol. 23, No. 3, May–June 1972, pp. 195–219.
[17] R. Grishman, A Survey of Syntactic Analysis Procedures for Natural Language, American Journal of Computational Linguistics, Vol. 13, No. 5, 1976, Microfiche 47.
[18] W.A. Woods, Transition Network Grammars for Natural Language Analysis, Communications of the ACM, Vol. 13, No. 10, October 1970, pp. 591–606.
[19] J. O'Connor, Retrieval of Answer-Sentences and Answer-Figures from Papers by Text Searching, Information Processing and Management, Vol. 11, Nos. 5/7, 1975, pp. 155–164.
[20] J. O'Connor, Data Retrieval by Text Searching, Journal of Chemical Information and Computer Science, Vol. 17, 1977, pp. 181–186.
[21] P.H. Klingbiel, A Technique for Machine-Aided Indexing, Information Storage and Retrieval, Vol. 9, September 1973, pp. 477–494.
[22] P.H. Klingbiel and C.C. Rinker, Evaluation of Machine-Aided Indexing, Information Processing and Management, Vol. 12, No. 6, 1976, pp. 351–366.
[23] G. Salton, The Smart Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1971.
[24] G. Salton, Dynamic Information and Library Processing, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1975.
[25] L.A. Zadeh, Fuzzy Sets, Information and Control, Vol. 8, No. 3, 1965, pp. 338–353.
[26] L.A. Zadeh, Fuzzy Logic and Its Application to Approximate Reasoning, Information Processing 74, North-Holland Publishing Company, Amsterdam, 1974, pp. 591–594.
[27] J.A. Goguen, The Logic of Inexact Concepts, Synthese, Vol. 19, 1968-69, pp. 325–373.
[28] G. Lakoff, Hedges—A Study in Measuring Criteria and the Logic of Fuzzy Concepts, English Regional Meeting of the Chicago Linguistic Society, Chicago, Illinois, 1972.
[29] T. Radecki, Mathematic Model of Information Retrieval Based on the Concept of a

Fuzzy Thesaurus, Information Processing and Management, Vol. 12, No. 5, 1976, pp. 313–318.

[30] A. Bookstein, Weighted Boolean Retrieval, Proceedings ACM-BCS Conference on Research and Development in Information Retrieval, in Information Retrieval Research, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, Butterworths, London, 1981.

[31] A. Bookstein, Fuzzy Requests: An Approach to Weighted Boolean Searches, Journal of the ASIS, Vol. 31, No. 4, July 1980, pp. 240–247.

[32] D.A. Buell and D.H. Kraft, A Model for a Weighted Retrieval System, Journal of the ASIS, Vol. 32, No. 3, May 1981, pp. 211–216.

[33] W.G. Waller and D.H. Kraft, A Mathematical Model of a Weighted Boolean Retrieval System, Information Processing and Management, Vol. 15, No. 5, 1979, pp. 235–245.

[34] V. Tahani, A Fuzzy Model of Document Retrieval, Information Processing and Management, Vol. 12, 1976, pp. 177–187.

[35] M.E. Maron and J.L. Kuhns, On Relevance, Probabilistic Indexing and Information Retrieval, Journal of the ACM, Vol. 7, No. 3, 1960, pp. 216–244.

[36] A. Bookstein and D.R. Swanson, A Decision-Theoretic Foundation for Indexing, Journal of the ASIS, Vol. 26, No. 1, 1975, pp. 45–50.

[37] C.J. van Rijsbergen, A Theoretical Basis for the Use of Cooccurrence Data in Retrieval, Journal of Documentation, Vol. 33, 1977, pp. 106–119.

[38] S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the ASIS, Vol. 23, No. 3, 1976, pp. 129–146.

[39] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973.

[40] C.T. Yu, W.S. Luk, and M.K. Siu, On Models of Information Retrieval, Information Systems, Vol. 4, No. 3, 1979, pp. 205–218.

[41] D.J. Harper and C.J. van Rijsbergen, An Evaluation of Feedback in Document Retrieval Using Cooccurrence Data, Journal of Documentation, Vol. 34, No. 3, September 1978, pp. 189–206.

[42] S.E. Robertson, C.J. van Rijsbergen, and M.F. Porter, Probabilistic Models of Indexing and Searching, ACM-BCS Conference on Research and Development in Information Retrieval, in Information Retrieval Research, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, Butterworths, London, 1981.

[43] J.J. Rocchio, Jr., Relevance Feedback in Information Retrieval, in The Smart System—Experiments in Automatic Document Processing, G. Salton, editor, Prentice-Hall, Englewood Cliffs, New Jersey, 1971, Chapter 14.

[44] E. Ide and G. Salton, Interactive Search Strategies and Dynamic File Organization in Information Retrieval, in The Smart System—Experiments in Automatic Document Processing, G. Salton, editor, Prentice-Hall, Englewood Cliffs, New Jersey, 1971, Chapter 18.

[45] H. Wu and G. Salton, The Estimation of Term Relevance Weights Using Relevance Feedback, Department of Computer Science, Cornell University, Ithaca, New York, 1980.

[46] M.B. Koll, WEIRD—An Approach to Concept-Based Information Retrieval, SIGIR Forum, Vol. 13, No. 4, Spring 1979, pp. 32–50.

[47] M.B. Koll, Information Retrieval Theory and Design Based on a Model of the User's Concept Relations, ACM-BCS Conference on Research and Development in

Information Retrieval, in Information Retrieval Research, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, Butterworths, London, 1981.
[48] G. Salton, Automatic Indexing Using Bibliographic Citations, Journal of Documentation, Vol. 27, No. 2, June 1971, pp. 98–110.
[49] E. Garfield, Citation Indexing, John Wiley and Sons, New York, 1979.
[50] H.G. Small, A Co-Citation Model of a Scientific Specialty: A Longitudinal Study of Collagen Research, Social Studies of Science, Vol. 7, 1977, pp. 139–166.
[51] R. Knaus, A Similarity Measure on Semantic Network Nodes, Paper presented at the Annual Meeting of the Classification Society, Gainesville, Florida, 1978.
[52] R. Marcus and J. Reintjes, Computer Interfaces for User Access to Heterogeneous Information Retrieval Systems, Massachsuetts Institute of Technology, Electronic Systems Laboratory Report ESL-R-739, Cambridge, Massachusetts, April 1977.
[53] R.A. Winter, T. Lozano-Perez, and B.O. Marks, An Overview of the Chemical Substances Information Network, Computer Corporation of America, Technical Report No. CCA-79-20, Cambridge, Massachusetts, September 1979.
[54] D. Eastlake III, T. Lozano-Perez, and D. Low, Design of Version I Prototype Chemical Substance Information Network, Computer Corporation of America, Technical Report No. CCA-80-6, Cambridge, Massachusetts, June 1980.
[55] M. Bracken, J. Dorigan, and J. Overbey, Chemical Substances Information Network, MITRE Corporation, Technical Report MTS-7558, McLean, Virginia, June 1977.
[56] C.A. Ellis and G.J. Nutt, Office Information Systems and Computer Science, ACM Computing Surveys, Vol. 12, No. 1, March 1980, pp. 27–60.
[57] F.W. Lancaster, Toward Paperless Information Systems, Academic Press Inc., New York, 1978.
[58] D.P. Bhandarkar, The Impact of Semiconductor Technology on Computer Systems, Computer, Vol. 12, No. 9, September 1979, pp. 92–98.
[59] J. Martyn, Prestel and Public Libraries: An LA/Aslib Experiment, Aslib Proceedings, Vol. 31, No. 5, May 1979, pp. 216–236.
[60] L.M. Branscomb, Information: The Ultimate Frontier, Science, Vol. 203, January 12, 1979, pp. 143–147.
[61] G. Salton, Suggestions for Library Network Design, Journal of Library Automation, Vol. 12, No. 1, March 1979, pp. 39–52.

BIBLIOGRAPHIC REMARKS:

Two kinds of materials suggest themselves as additional sources for information on the future of information retrieval. The first consists of reviews outlining the current state of the art and including also short-term projections for the future. Such material can be found in certain conference proceedings and in some of the widely circulated journals in the computer field. There exist yearly proceedings of a Conference on Office Automation sponsored by the American Federation of Information Processing Societies (AFIPS). The Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM) jointly sponsor a yearly conference on Very Large Data Bases. Conference proceedings are also available covering many computer

technologies of interest such as microprocessors and computer graphics. The journals that regularly include reviews and projections into the future of information science include Computerworld, Datamation and the ACM Computing Surveys.

Another type of article takes a more frankly visionary approach. In that area it is however necessary to distinguish materials with a science-fiction approach from articles exhibiting a solid technical foundation. The following articles or books are certainly worth reading:

V. Bush, As We May Think, Atlantic Monthly, Vol. 176, No. 1, 1945, pp. 101–108.

J. G. Kemeny, A Library for 2000 A.D., in Management and the Computer of the Future, M. Greenberger, editor, MIT Press, 1962, pp. 134–178.

F. W. Lancaster, Toward Paperless Information Systems, Academic Press Inc., New York, 1978.

J. C. R. Licklider, Libraries of the Future, MIT Press, Cambridge, Massachusetts, 1965.

G. Salton, Dynamic Information and Library Processing, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1975, Chapter 10.