What's happening in the field of information retrieval? This article provides an up-to-date review.

INFORMATION RETRIEVAL IN MEDICINE: STATE OF THE ART

WILLIAM R. HERSH, M.D., AND
ROBERT A. GREENES, M.D., PH.D.

The field of information retrieval is concerned with the representation, storage, and retrieval of heterogeneous textual information. Information retrieval systems are distinct from database management systems, which process structured data, and expert systems, which make inferences about highly organized data. While information retrieval systems can be integrated with these other types of systems, the chief goal of research in this field is to develop methods to represent and query diverse collections of textual resources, such as journal articles, textbooks, and more recently, hypertext.

Because of the heterogeneity of the data stored in information retrieval systems, they are not designed to provide the knowledge sought in a user's query directly. As Lancaster has stated, an information retrieval system does not inform the user on the subject of an inquiry; it merely indicates the existence (or nonexistence) and whereabouts of documents relating to this request [1]. With the decreasing prices of powerful hardware and the increased storage capacity offered by CD-ROM drives, documents can often be located on the same machine as the information retrieval system.

ABSTRACT

Conventional information retrieval systems usually involve searching by terms from controlled vocabularies or by individual words in the text. These systems have been commercially successful but are limited by several problems, including cumbersome interfaces and inconsistency with human indexing. Research on methods that automate indexing and retrieval has been performed to address these problems. The three major types of automated systems are vector-based, probabilistic, and linguistic. This article describes these systems and provides an overview of the field of information retrieval in medicine.

[KEYWORDS: *information retrieval, bibliographic retrieval, MeSH, Medline, full-text retrieval, unified medical language system, vector-based information retrieval, probabilistic information retrieval, automatic indexing, natural-language processing*]

THE PROCESS OF INFORMATION RETRIEVAL

In his book on the field of information retrieval, Salton [2] provides a functional model of an information retrieval system (Fig. 1), composed of a set of items of information (DOCS), a set of requests (REQS), an indexing language (LANG), and a function (SIMILAR) to determine which, if any, of the items of information match the requests.

Indexing is the process by which descriptors of the items of information are chosen. These descriptors are part of the indexing language, which contains all the allowable words or terms that a user can specify in requests for information. The types most commonly used in commercial systems are a controlled vocabulary of terms and a collection of all words that occur in the items of information. Once the items of information have been represented by the in-

dexing language, the user can retrieve them by formulating a search in that language. The search request is usually stated with terms and Boolean operators. In addition, some systems allow specification of a word's position in relation to other words. Once the search has been formulated, the SIMILAR function provides a list of items of information (DOCS) that match the items that the user has specified (REQS) in the indexing language (LANG).

EVALUATION OF RETRIEVAL

There have been studies devoted to evaluating various methods of indexing and retrieval. A theoretical framework for evaluation was developed by Fidel and Soergel [3], who grouped the variables in searching into seven categories: the user, the request, the database, the search system, the search intermediary, the search process, and the search outcome. While most of these variables are related to system users and available content, the performance of different indexing and retrieval strategies is usually measured in terms of recall and precision. Recall measures the proportion of relevant items in the document collection that are retrieved by a search, and precision measures the proportion of retrieved items that are relevant. These proportions are calculated by the following equations [4]:

$$\textit{Recall} = \frac{\textit{number of relevant items retrieved}}{\textit{number of relevant items in collection}}$$

$$\textit{Precision} = \frac{\textit{number of relevant items retrieved}}{\textit{total number of items retrieved}}$$

These measures are analogous to sensitivity and specificity, the variables used to describe the utility of medical testing. (Precision is actually equivalent to positive predictive value.) Just as a diagnostic test with high sensitivity will identify almost all patients who have a disease, even at the expense of identifying some who do not, an information retrieval system aiming for high recall will retrieve almost all documents that are relevant, even at the expense of retrieving some that are not. Likewise, just as a diagnostic test with high specificity will exclude most patients who do not have a disease, even at the expense of missing a few who do, an information retrieval system aiming for high precision will retrieve as few irrelevant documents as possible, even at the expense of missing a few that are relevant. Furthermore, just as setting the normal range of a medical test requires a tradeoff between sensitivity and specificity, the nature of information retrieval systems requires the searching process to be a tradeoff between precision and recall. If a system is designed to retrieve as much as possible on a given search term (higher recall), then irrelevant documents will be retrieved (lower precision). This type of system would be most useful for a researcher who was seeking all the relevant information on a topic. On the other hand, someone who was simply seeking a general review of a subject would be better served by a system designed to retrieve as few irrelevant documents as possible (higher precision), with some relevant documents being missed (lower recall).

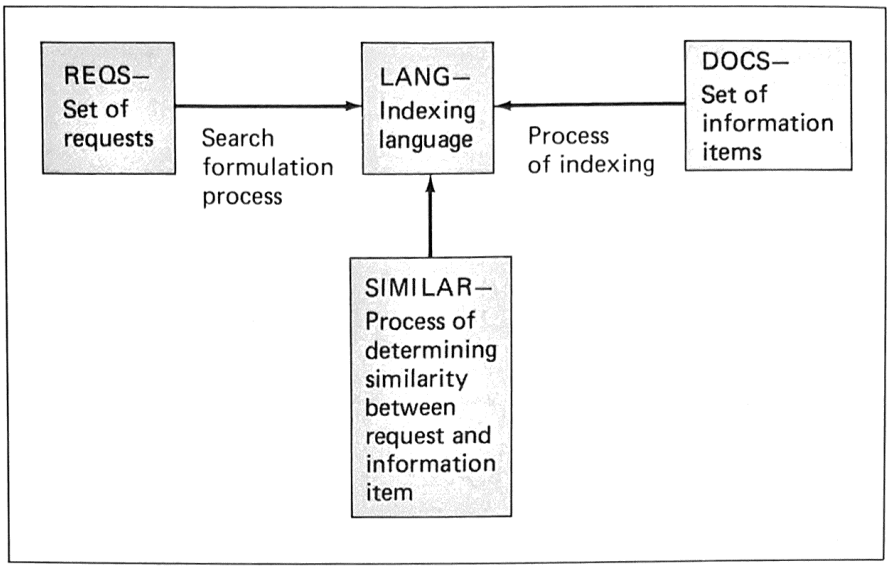

Figure 1. Functional model of information retrieval (adapted from Salton [2]).

CONVENTIONAL SEARCHING METHODS

The original and still most widely used format for organizing online databases is the format of bibliographic citations. Each reference is organized into "fields," such as author, title, keywords, and source. This is similar to the organization of a database management system, although the fields are less highly structured and the individual words in them can be searched. In a database management system such as dBASE III (Ashton-Tate Corporation), only the entire item in a particular field can be searched. Recently, database vendors have begun to offer full-text searching, which is the ability to search freely over the entire text of an article, report, or book.

Bibliographic Citation

Searching a database of bibliographic citations is usually done with one or more fields of each record. The most important field generally consists of the article's keywords, which contain descriptors of the content in the record. These keywords have usually been designated by an expert human indexer. Keyword-based information retrieval systems use a vocabulary consisting of words and phrases that serves as the indexing language to represent the content of the records. The indexing language is usually a controlled vocabulary, in which each term is represented in a fixed way. Some controlled vocabularies contain synonyms that lead to preferred

terms. In the MeSH (medical subject headings) vocabulary, for example, both "hypertension" and "high blood pressure" occur, with "hypertension" being the preferred term. Some vocabularies also have a hierarchical organization.

One of the most famous bibliographic databases is Medline [5]. Created and maintained by the National Library of Medicine (NLM), Medline contains bibliographic references from 3300 journals dating back to 1966. There are now 6 million references in the Medline database, with about 350,000 new entries each year.

The main method of searching Medline is by the use of MeSH terms, which make up the indexing language for the database. MeSH is a controlled vocabulary consisting of 16,000 terms, called "headings," and 12,000 synonyms for the headings, called "entry terms." MeSH has a hierarchical organization based on 15 "tree structures," such as "Anatomy," "Chemicals and Drugs," and "Diseases."

MeSH has features that enhance recall and precision. One such feature is the "explode" function, which allows the user to specify a term and have all the terms underneath it in the tree added to the search. This function has the effect of improving recall because more possibly relevant articles will be retrieved. In addition, the MeSH vocabulary contains 76 "subheadings" that can be added to MeSH terms to increase their specificity. For example, one can add the subheading "diagnosis" to the term "acquired immunodeficiency syndrome" and thus retrieve articles on the diagnosis of AIDS (and not other aspects of the disease, such as its treatment). This has the effect of enhancing precision because fewer irrelevant articles will be retrieved.

The indexing process for Medline is entirely manual. Human indexers use a detailed procedure to designate MeSH terms and subheadings [6]. The indexer spends an average of 15 minutes per article, at an average cost to the NLM of $4.17 per article (Wright N: personal communication). Indexing the 350,000 references added to Medline annually requires about 44 full-time-equivalent indexers at a cost of $1.4 million.

Medline also allows searching with other fields, such as author, source, and year of publication. Other systems offer the ability to search titles and abstracts using partial string searching. Some database vendors, such as BRS (BRS Information Technologies) and Dialog (Dialog Information Services), enhance Medline (and other databases) by allowing searching of titles and abstracts using partial string searching as well as positional operators, which enable the user to specify the position of words in a search of a field. BRS features the ADJ operator, which requires that words on either side of it occur adjacently. For example, if the query "asymptomatic ADJ AIDS" is applied to the abstract field, only articles with the phrase "asymptomatic AIDS" in the abstract will match. Dialog has a similar feature but allows specification of the number of words between terms. For example, the query "asymptomatic (5W) AIDS" would match references in which "asymptomatic" and "AIDS" occurred within five words of each other.

Medline is not the only medical database that uses MeSH. Another is the Health Planning and Administration database (American Hospital Association). In addition, MeSH is not the only vocabulary used for indexing databases. For example, the Biosis database (Biological Abstracts, Inc.), which includes references to government documents, conference proceedings, and monographs as well as articles, uses an indexing vocabulary that is markedly different in organization and style from MeSH.

Full-Text Retrieval

The other major method available for searching commercial databases is full-text retrieval. In systems using this approach, there is no controlled vocabulary. Rather, every word that appears in the text is designated as an indexing term, and thus all the words that occur in the database make up the indexing language. The indexing process is essentially the process of creating an inverted file of all words in the database, with each word containing a pointer to each item of information in which it occurs. Searches are usually formulated with words, Boolean operators, and positional operators. An increasing number of medical textbooks and journals are offered as full-text databases by vendors such as BRS and Dialog.

Another well-known full-text retrieval system is IBM's Storage and Information Retrieval System (STAIRS) [2]. This system does not come with preexisting databases; the databases must be supplied by the user. Text is indexed into standard inverted files. Items of information are retrieved by queries that, in addition to Boolean operators, have the operators ADJ for requiring terms to be adjacent, WITH for requiring terms to be in the same sentence, and SAME for requiring terms to be in the same paragraph. In addition, a SYN operator allows the user to designate word pairs or groups as synonyms. The program also offers the ability to rank retrieved documents, which is done by calculating weights for terms in the query. Retrieved documents are then ranked according to the sum of the weights of all terms that were in the query. There are various weighting algorithms; the formula for one is as follows: Value of term t = frequency of term t in document d × frequency of term t in retrieved set/number of documents retrieved with term t.

Limitations of Conventional Systems

Conventional information retrieval systems are helpful to end-users, but have a number of limitations. A common complaint about both keyword and full-text systems is that they are difficult to use. The Elhill program that the NLM provides to search Medline is cryptic and unforgiving. Fortunately, over the past decade, more user-friendly systems have been developed. The first of these was PaperChase [7], which has a user-friendly menu-driven interface. PaperChase includes a function that finds MeSH headings on the basis of words entered. It permits

permutations of authors' names and MeSH headings, it provides automatic word completion and algorithms to handle variations in spelling, and it continuously monitors the user's searching strategy and makes suggestions for improvement. Another system, Grateful Med [8], provides a front end to the NLM's program for searching Medline and other databases. Grateful Med also offers the ability to compose and evaluate a search offline.

Another area that causes problems is formulating Boolean searches. Borgman [9] has documented user errors such as confusion of the logical OR with the logical AND and performing a logical AND with empty lists. Sewell and Teitelbaum [10] note that the majority of end-users tend to use only the logical AND in their searching.

An additional limitation with conventional systems is that the matching documents from a search are not ranked in any way. Often a search will generate tens or hundreds of matching documents, and the user will be required to inspect each one to determine whether it contains useful information. PaperChase ranks documents according to the quality of the journal in which they are published [11]. The Knowledge Finder CD-ROM system (Aries) ranks the output according to how well items match the query (e.g., have the most matching search terms). Rada and Bicknell [12] have looked at using distances in the MeSH tree to rank documents retrieved, aiming to put the documents deemed most relevant to the user's query in the most prominent position. Many of the research systems described below employ ranking strategies.

Other limitations occur in systems based on keywords. Salton [13] argues that human indexing is expensive and leads to inconsistent results. Funk and Reid [14] studied consistency in Medline indexing, looking at the correlations between the indexing terms chosen for 740 articles that were, for a variety of reasons, indexed more than once. They found that central-concept main headings (MeSH terms that are followed by an asterisk in Medline, indicating a highly important term for the document) were consistently chosen 61.1% of the time, whereas main-heading and subheading combinations were consistently chosen only 33.8% of the time. Crain [15] evaluated NLM indexers with a think-aloud protocol analysis, noting that in many instances they did not follow the cognitive processes specified in the indexing protocols. Crain also found that the MeSH terms selected varied with the indexer.

Keyword indexing is also expensive, in both time and dollars, as suggested by the costs associated with indexing Medline. Because the medical literature contains only a small portion of the available and useful medical information, this problem will be exacerbated as more information resources are added to computer-based information systems. CD-ROM disks that contain much information beyond bibliographic references are already available. Compact Library: AIDS (Medical Publishing Group of the Massachusetts Medical Society), for example, contains an electronic textbook and a statistical database in addition to bibliographic references. Most of these CD-ROM systems use full-text retrieval methods.

Another problem with keyword systems is that so many medical vocabularies are used by the various commercially available databases. Most of these vocabularies require reference material as large as the Boston telephone book. Not only do they contain different terms, but their indexers apply terms as keywords in different ways. In addition, there are even more medical vocabularies used for patient records (e.g., SNOMED [16]), patient billing (e.g., ICD-9 [17]), and expert systems (e.g., Quick Medical Reference [18]). Table 1 shows how synonymous terms can be expressed in different vocabularies.

In an effort to overcome this problem, the NLM initiated the unified medical language system project [19]. The goal of this project is to create a "metathesaurus" that will allow translation of terms between different vocabularies. The first version, Meta-1, is now available for research use. For information retrieval systems, vendors will be encouraged to map the

SYNONYMOUS TERMS FROM MeSH AND SNOMED

Table 1

MeSH Term	SNOMED Term
Acute yellow atrophy	Atrophy, acute yellow
Arenaviridae	LCM group virus
Baritosis	Barium lung disease
Facial hemiatrophy	Romberg's syndrome
Facial nerve	Seventh cranial nerve
Homosexuality	Homosexual state
Islands of Langerhans	Islets of Langerhans
Leukemia, myelocytic	Myeloid leukemias
Ross River virus	Epidemic Australian polyarthritis
Round window	Cochlear window
Scleroderma, systemic	Generalized scleroderma
Subacute sclerosing panencephalitis	Dawson's encephalitis
Sturge-Weber syndrome	Encephalotrigeminal angiomatosis

INVERSE DOCUMENT FREQUENCY FOR TERM WEIGHTING

Inverse document frequency of term k =

$$\log 2 \frac{\text{Number of documents in collection}}{\text{Number of documents in which term k occurs}} + 1$$

Term weight of term k in document i =

Frequency of term k in document i × Inverse document frequency of term k

Figure 2. Inverse document frequency for term weighting.

terms in their vocabularies to Meta-1, thus allowing the creation of more generalized searching interfaces based on its preferred terms.

While full-text retrieval does not have the problems associated with keyword systems, it has problems of its own. Full-text retrieval systems are highly dependent on the way the user enters terms. Blair and Maron found that these systems generally have high precision but low recall [20]. They tend to require the user to anticipate all the synonyms of a word that might be used. For example, to ensure that all articles on the treatment of a certain type of pneumonia are found, the user may have to enter the names of all the antibiotics that are possibly used for it, in case the article does not contain general terms such as "antibiotic." These systems are also limited by the fact that many of the words that appear in documents do not accurately represent the major points under discussion. For example, an article containing the word "lead" used as a verb would be retrieved when a user was searching for topics on the chemical element lead.

RESEARCH SYSTEMS

The limitations of keyword systems, particularly the expense and inconsistency of human indexing, have led to research on information retrieval systems that perform automatic indexing. The ground-breaking work in this area was done by Luhn [21], who suggested in the late 1950s that computers could be used to analyze the content of text. Luhn observed that the words in a document could be classified into three categories based on whether they occurred with high, medium, or low frequency. Terms of high frequency, such as "and" and "the," impart no resolving power—that is, no ability to identify relevant documents and distinguish them from irrelevant documents. Terms of low frequency also have little resolving power, whereas terms of medium frequency, Luhn discovered, have the most.

These observations led to research on the use of automated methods to select indexing terms. Although it was initially thought that automatic indexing would require a great deal of understanding of the syntactic and semantic information in documents, most of the success achieved has been with systems based on analysis of word frequencies. However, recent advances in knowledge-based systems and natural-language processing have improved the ability to handle syntactic and semantic information in processing text, and workers in this area hope that these advances can be translated into improved indexing and retrieval.

Other early workers noted that term weighting was useful in measuring the value of words as indexing terms. Sparck Jones [22] defined the "inverse document frequency," which calculates the inverse proportion of documents in which a term occurs. With his approach the weight assigned to each term in a document is based on the product of the term's frequency in that document and the inverse document frequency, as shown by the formula in Figure 2. This gives the highest value to terms that occur infrequently in the collection but frequently in the individual document.

Vector-Based Systems

Salton et al. [23] noted that resolving power as described by Luhn [21] was a function only of how frequently a word occurred, leading them to introduce the concept of "term discrimination value," which measures the degree to which terms help to discriminate documents from each other. Salton found that for any two documents in a collection, such as Di and Dj, one could calculate a function, SIMILAR(Di,Dj), that represented their similarity on the basis of words that occurred in them. One could then calculate the AVERAGE-SIMILARITY function, which would give a measure of the density of the document space, or how much the documents were bunched up in the document space. The term's discrimination value could then be measured by removing the term from the documents and subtracting the AVERAGE-SIMILARITY for the collection with the term present from the AVERAGE-SIMILARITY for the collection with the term removed.

If a term occurs in many of the documents in a collection, its removal will decrease the average document-pair similarity. High-frequency terms such as this have a negative discrimination value, which has the effect of increasing average similarity when the term is added. They are thus poor discriminators and are not used as indexing terms. When terms are of medium frequency, on the other hand, their removal will increase the average document-pair similarity. These terms have a positive discrimination value, which has the effect of decreasing average similarity when the term is added. Terms of low frequency usually tend to have little effect on average similarity and have discrimination values close to zero. The

decision to include these words as indexing terms is based on whether the system will emphasize precision, which is reduced by their removal. For the terms that are kept as indexing terms, weights are calculated as the product of the term's frequency in a particular document and its discrimination value.

Before the term-weighting process begins, steps can be taken to make automatic indexing more efficient. The first step is to eliminate high-frequency noise words, such as "and," "the," and "also," which are kept in a stop list. According to vanRijsbergen, 250 words should be included in this list [24]. Another step is to use a "stemming algorithm," which removes common suffixes, such as "tion," "ing," and letters making a word plural. Reducing words to their stem forms in this way enhances recall. Lovins developed one of the first frequently used stemming algorithms [25], and Porter [26] developed a simpler one that performs just as well. Efficiency is also increased by indexing document abstracts only. Salton et al. [23] found that recall and precision were identical whether the full text of a document or only its abstract was used.

In a typical automatic indexing scheme, the steps described above are followed by calculation of the term discrimination values. Terms with negative values are then eliminated, so that only terms that discriminate between documents are left in the indexing language. For each document, an array consisting of the number of words remaining as indexing terms is created, with each element consisting of the term weight. This array can also be thought of as a vector multidimensional space, with the length in each dimension equal to the term weight. In addition to the document vectors, term vectors can be created, in which each dimension represents a term's weight for a given document.

Because this approach uses a completely uncontrolled vocabulary, there is no way to specify synonymous words. Salton does, however, describe a method of constructing a thesaurus automatically, on the basis of similarity measures between term vectors [23]. Low-frequency terms are grouped together, which increases the discrimination values of each term. This leads to improved recall because more documents containing the same words will be retrieved.

VECTOR COSINE FORMULA

$$\text{COSINE of document i for query j} = \frac{\Sigma \text{ Weight of term k in document i} \times \text{Weight of term k in query j}}{\sqrt{\Sigma \text{ (Weight of term k in document i)}^2 \times \Sigma \text{ (Weight of term k in query j)}^2}}$$

Figure 3. Vector cosine formula.

Another way to enhance performance is to combine high-frequency terms into indexing phrases. Using statistical, nonsyntactic methods, commonly occurring words such as "blood" and "test" are combined into a phrase such as "blood test," which is much more specific and is likely to be a better discriminator than either word alone. This process enhances precision because only documents indexed by the more specific phrase will be retrieved.

In Salton's system querying is done by entering free text. Each word in the query is run against the stop list and stemming algorithm. A query vector is created for all terms that are found as indexing terms. This vector has the same dimensions as the document vector, and documents are matched to the query on the basis of the cosine measure of the vectors in multidimensional space. As with the trigonometric cosine, angles between vectors that are more similar have a higher cosine. The formula for calculation of the vector cosine is shown in Figure 3. Documents retrieved by this approach will reflect the terms in the query and will be ranked in order of their cosine. Thus, the documents that match the query most closely, which are presumably closest to what the user has sought in the query, will be ranked highest.

Once the initial querying has been completed, performance can be further improved by the use of "relevance feedback," in which the user indicates whether or not the documents retrieved are relevant, leading the system to reformulate the query. Typically, the system will add weight to words that occur mostly in relevant documents and remove weight from words that occur in irrelevant documents. Wu and Salton observed a 30 to 50% improvement in recall and precision with this method [27].

At first glance Salton's methods, which are based on word frequencies without the benefit of a controlled vocabulary and synonyms, seem nonspecific. But studies of recall and precision have consistently shown that these methods perform as well as conventional systems, if not better. Salton showed that the automated indexing process with ranked output and an automatically generated thesaurus performed as well as standard Boolean searches from Medline [28]. When relevance feedback was added, the automated approach showed a 15 to 30% improvement.

Probabilistic Systems

Probabilistic information retrieval systems represent another type of research approach. Instead of using vectors and query-document similarities, these systems are based on probabilistic measures for term weighting and document ranking. The IRX system of the NLM, for example, is an experimental method with a modular approach, allowing different term weighting measures, stop lists, and stemming algorithms to be

THE ORIGINAL AND STILL MOST WIDELY USED FORMAT FOR ORGANIZING ONLINE DATABASES IS THE FORMAT OF BIBLIOGRAPHIC CITATIONS.

used and evaluated [29]. Indexing is done by calculating term weights for all words not on the stop list. For retrieval, all documents containing one or more words in the query (minus stop words) are given a score, based on the sum of the term weights from the terms that occur in the query. These documents are then ranked and presented to the user, with the documents matching the query best at the top of the list.

Fuhr's probabilistic system [30] uses more factors for term weighting than just the inverse document frequency or term discrimination value. Term weights are calculated by a number of factors, such as whether the term occurs in the title rather than the abstract (the former gets twice the score), whether the term is a stem form, and the maximum frequency of the term in any document in the collection. Croft has devised a probabilistic system that uses a Bayesian approach, based on the probability that a document is or is not relevant for a given word [31].

Linguistic Systems

Although the performance of vector and probabilistic systems is unmatched among research systems, theoretical issues suggest that the use of linguistic methods might improve indexing and retrieval further. Linguistic systems are based on advances in natural-language processing, with the underlying theme being the idea that indexing is based on concepts instead of terms, which are just string-form representations of concepts. That is, the concept "high blood pressure" has several terms that can describe it, such as "high blood pressure," " hypertension," and "elevated blood pressure." Indexing by concepts theoretically allows a much richer variety of words to be used to capture concepts for indexing and match them in retrieval. The indexing language consists of actual concepts (usually represented in a semantic network), which can be expressed in diverse string forms.

Some definitions of linguistic terms may be helpful at this point. The "syntax" of words is their grammatical category, such as noun or verb. Syntax indicates how words can be arranged to form grammatically correct sentences. "Parsing" is the process whereby sentences and phrases are broken down into syntactic categories. The "semantics" of a word or phrase is its meaning, and a "semantic network" is a network of concepts and the allowable relationships between them. For example, the concepts "penicillin" and "pneumococcal pneumonia" have a relationship, "treatment," between them.

Some linguistic systems use syntax to recognize concepts. Fagan [32] developed a syntactic approach based on natural-language processing techniques that used parsed noun phrases as indexing units. However, studies of recall and precision showed that nonsyntactic statistical approaches performed better. In an evaluation of the failure of the syntactic approach in Fagan's study, Salton and Smith [33] noted problems with ambiguous parsing, parsing of meaningless phrases, and an incomplete vocabulary, showing that natural-language processing for information retrieval based on syntax alone was problematic.

Others have attempted to use more semantic information for natural-language processing. Croft and Lewis [34] developed a program based on "case frames," collections of lower-level concepts and relationships between them that signified higher-level concepts. The system was akin to the traditional approach to natural-language processing, which is limited at present to very narrow domains. Other problems included the difficulty of building case frames for large domains and of capturing all domain knowledge in case frames.

In the Linguistic String Project, Sager et al. [35] captured a considerable amount of semantics by limiting the domain and the type of language processed. Their system processed medical records, which could then be stored in a conventional database. Its success was largely due to the fact that medical records are written in a terse and repetitive style, simplifying the tasks of syntactic and semantic analysis. The approach proved difficult with other types of textual information, such as bibliographic citations, in which sentence variation was more pronounced and complex knowledge bases were required for semantic evaluation.

Evans has attempted to overcome the problems of slow performance and limited domains through the use of restricted linguistic methods [36]. Abandoning the goal of complete syntactic and semantic understanding of text, his system uses selected amounts of syntax and semantics to recognize concepts for indexing. The main function of the parser in this system is to recognize noun phrases, which represent the concepts in the text.

Another way of using phrases as index terms has been to map words and phrases that occur in documents into an existing dictionary. Biebricher et al. [37] used a database containing abstracts in physics and a large dictionary of terms. A statistical approach to mapping words in the abstracts and queries into terms in the dictionary was devised on the basis of string-matching.

Another approach to concept-based indexing is to use a "discrimination network" to map phrases into concepts represented by preferred terms. This approach uses a branching pathway through a network to identify the most specific terms in a string. It was first described by Shoval [38] and was subsequently used for in-

dexing and retrieval from a neuropathology database [39]. This method requires the existence of a vocabulary with preferred terms and their synonyms. We have enhanced Shoval's approach by increasing the variety of synonyms it can handle [40]. We added a stemming algorithm to create a concept-matching algorithm on which we based a system that used probabilistic measures applied to concepts instead of terms.

Other workers have attempted to use semantic relationships to represent the contents of documents in a more knowledge-based fashion. Miller et al. proposed to enhance Medline searching by identifying semantic relationships between MeSH terms, aiming to extend the relationships implicit in MeSH subheadings [41]. They initially identified relationships that could apply to MeSH terms (such as disease X causes or predisposes to disease Y, and treatment X treats disease Y) and then used these relationships in revising the indexing of the literature. Like the use of MeSH subheadings, the use of semantic relationships is an attempt to improve precision by making the criteria for retrieval more specific.

Despite the theoretical potential for improved retrieval with this approach, there are practical issues that need to be addressed, such as whether a comprehensive set of relationships can be developed and whether indexers can identify them in an accurate and consistent fashion [42]. Some perspective on this latter point can be gained by recalling Funk and Reid's finding [14] that consistency between indexers dealing with MeSH main concepts and their subheadings was only 33.8%. In addition, it is not known how well users would understand and employ this type of approach.

RESEARCH WITH MANUALLY INDEXED SYSTEMS

Manually indexed systems continue to attract research interest, with various workers attempting to improve the indexing and retrieval processes, mostly by means of artificial intelligence. Humphrey [43] has developed an interactive expert system designed to help indexers choose MeSH terms. Rules are encoded as "procedural attachments" (or "demons") to concepts in a semantic network in an attempt to mimic the rules in the Medlars indexing manual. The rules are activated when a concept is designated as an indexing concept. The system asks the user for additional information suggesting MeSH headings and subheadings and possibly activating further demons.

Other research on manual systems has concentrated on the MeSH vocabulary. As mentioned earlier, Rada and Bicknell [12] used the hierarchical relations in MeSH to devise a scheme for ranking references on the basis of distances between terms in MeSH trees. They have used other medical thesauri to help suggest revisions in MeSH, based on inappropriate hierarchical relations in the MeSH trees.

FUTURE DIRECTIONS

With the many different lines of research on information retrieval being explored, it seems safe to assume that better information retrieval systems will continue to evolve. While conventional manually indexed and full-text systems will continue to dominate commercial systems, some of the automated indexing schemes from research efforts will no doubt make their way into commercial products. Research with manually indexed systems should also add to the power of commercial systems.

As computer hardware continues to become less expensive and more powerful, better performance and functionality are certain to ensue. Furthermore, as mass storage technologies such as CD-ROM and laser disks improve, increasing amounts of information will be available through the computer. Improved access to information should lead to better medical care.

[From the Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115.

We are indebted to Ms. Cindy Schatz, Reference Librarian, Countway Medical Library, Harvard Medical School.]

THERE ARE NOW 6 MILLION REFERENCES IN THE MEDLINE DATABASE, WITH ABOUT 350,000 NEW ENTRIES EACH YEAR.

REFERENCES

1. Lancaster F. Information retrieval systems. New York: Wiley, 1968.
2. Salton G. Introduction to modern information retrieval. New York: McGraw-Hill, 1983.
3. Fidel R, Soergel D. Factors affecting online bibliographic retrieval: a conceptual framework for research. JASIS 1983; 34:163-80.
4. Salton G, Fox EA, Wu H. Extended Boolean information retrieval. Commun Assoc Comput Machinery 1983; 26:1022-36.
5. Feinglos S. MEDLINE: a basic guide to searching. Chicago: American Medical Association, 1985.
6. Charen T. MEDLARS indexing manual. Springfield, Va.: National Technical Information Service, 1983.
7. Horowitz GL, Jackson JD, Bleich HL. PaperChase: self-service bibliographic retrieval. JAMA 1983; 250:2495-9.
8. Haynes RB, McKibbon KA. Grateful Med. MD Comput 1987; 4(5):47-57.
9. Borgman CL. Why are online catalogs hard to use? Lessons learned from information retrieval studies. JASIS 1986; 37:387-400.
10. Sewell W, Teitelbaum S. Observations of end-user online searching behavior over eleven years. JASIS 1986; 37:234-45.
11. Underhill LH, Bleich HL. Bringing the medical literature to physicians: self-service computerized bibliographic retrieval. West J Med 1986; 145:853-8.
12. Rada R, Bicknell E. Ranking documents with a thesaurus. JASIS 1989; 40:304-10.
13. Salton G. Another look at automatic text-retrieval systems. Commun Assoc Comput Machinery 1986; 29:648-56.
14. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Lib Assoc 1983; 71:176-83.
15. Crain CJ. Appendix A: protocol study of indexers at the National Library of Medicine: final report on the Automated Classification Retrieval Project. Washington, D.C.: National Library of Medicine, 1987.
16. Cote RA. Systematic nomenclature of medicine. Skokie, Ill.: College of American Pathologists, 1982.
17. Slee VN. The international classification of diseases: ninth revision, ICD-9. Ann Intern Med 1978; 88:424-6.
18. Miller RA, McNeil MA, Challinor SM, Masarie FE, Myers JD. The Internist-1/Quick Medical Reference project: status report. West J Med 1986; 145:816-22.
19. Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. In: Kingsland LC III, ed. Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care. Washington, D.C.: IEEE Computer Society Press, 1989: 475-80.
20. Blair DC, Maron ME. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Commun Assoc Comput Machinery 1985; 28:289-99.
21. Luhn H. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of R&D 1957; 1:309-17.
22. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Documentation 1972; 28:11-21.
23. Salton G, Yang CS, Wu CT. A theory of term importance in automatic text analysis. JASIS 1975; 26:33-44.
24. vanRijsbergen CJ. Information retrieval. London: Butterworth, 1979.
25. Lovins JB. Development of a stemming algorithm. Mech Translation Comput Linguistics 1968; 11:11-31.
26. Porter MF. An algorithm for suffix stripping. Program 1980; 14: 130-7.
27. Wu H, Salton G. The estimation of term relevance weights using relevance feedback. J Documentation 1981; 37:194-214.
28. Salton G. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). JASIS 1972; 23:75-84.
29. Harman D, Benson D, Fitzpatrick L, Huntzinger R, Goldstein C. IRX: an information retrieval system for experimentation and user applications. SIGIR Forum 1988; 22:2-10.
30. Fuhr N. Models for retrieval with probabilistic indexing. Inf Proc Mgmt 1989; 25:55-72.
31. Croft WB. Document representation in probabilistic models of information retrieval. JASIS 1981; 32:451-7.
32. Fagan JL. Automatic phrase indexing for document retrieval: an examination of syntactic and non-syntactic methods. SIGIR 87 1987; 91-101.
33. Salton G, Smith M. On the application of syntactic methodologies in automatic text analysis. SIGIR 89 1989; 137-50.
34. Croft WB, Lewis DD. An approach to natural language processing for document retrieval. SIGIR 87 1987; 26-32.
35. Sager N, Friedman C, Lyman MS. Medical language processing: computer management of narrative data. Reading, Mass.: Addison-Wesley, 1987.
36. Evans DA. Notes on the CLARIT project. Technical report, Laboratory for Computational Linguistics, Carnegie-Mellon University, 1989.
37. Biebricher P, Fuhr N, Lustig G, Schwanter M. The automatic indexing system AIR/PHYS: from research to application. Technical report, Technische Hochschule Darmstadt, Darmstadt, West Germany, 1989.
38. Shoval P. An expert consultation system for a retrieval database with a semantic network of concepts. Ph.D. thesis, University of Pittsburgh, 1981.
39. Vries JK, Shoval P, Evans DA, Moossy J, Banks G, Latchaw R. An expert system for indexing and retrieving medical information. Technical Report, University of Pittsburgh School of Medicine, 1986.

40. Hersh WR, Greenes RA. SAPHIRE: an information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. Comput Biomed Res (in press).
41. Miller PL, Barwick KW, Morrow JS, Powsner SM, Riely CA. Semantic relationships and medical bibliographic retrieval: a preliminary assessment. Comput Biomed Res 1988; 21:64-77.
42. Miller PL, Morrow JS, Powsner SM, Riely CA. Semantically assisted medical bibliographic retrieval: an experimental computer system. Bull Med Libr Assoc 1988; 76:131-6.
43. Humphrey SM. Interactive knowledge-based indexing: the Medindex System. RIAO 88 1988; 883-98.

WILLIAM R. HERSH, M.D.
Dr. Hersh is a staff scientist at the Biomedical Information Communications Center of Oregon Health Sciences University in Portland, Oregon, where he is also assistant professor of medicine in the Division of General Internal Medicine. Previously, he was a fellow in medical informatics at the Decision Systems Group of Brigham and Women's Hospital and Harvard Medical School in Boston. He received his M.D. at the University of Illinois. His research interests include the design and evaluation of medical information retrieval systems.

ROBERT A. GREENES, M.D., PH.D.
Dr. Greenes is an associate professor of radiology and is director of the Decision Systems Group, Brigham and Women's Hospital, and of the Medical Informatics Research Training Program at Harvard Medical School. His M.D. and Ph.D., in applied mathematics and computer science, are both from Harvard. He has chaired the Symposium on Computer Applications in Medical Care and is a member of numerous editorial boards and committees in the fields of informatics and radiology. His research interests include medical knowledge management, the physician-computer interface, and computer-based education and decision support.