Zeroes are used as values in the file to denote that the compound failed the test. Blanks are left in the test record to indicate the compound was not screened by that test. The program permits bypassing blank data while searching.

II. Inquiry Routine Using "AND" Logic. In this routine, the computer accepts any tests that meet criteria set by the inquiry. A typical question could be: "Are there any compounds on file which have a value of 6 or greater as a Soil Fungicide and a value of 4 or greater as a Nematocide?" All specified parameters must be present in the record for a hit to be printed. Up to 50 test parameters may be set in conjunction with each other in a single inquiry. As in (I), blank data can be deleted if requested. The basic pattern of search and print options is the same as in (I).

III. Inquiry Routine Using "OR" Logic. In this routine, the computer accepts only those records that meet inquiry criteria which specify a number (up to 50) of disjunctive parameters—e.g., 1 or 2 or 3 or. . .*n*. A typical question might be: "Are there any compounds with values of 7 or greater as Aquatic Herbicides or Contact Insecticides or Nematocides?" Again, blank data can be deleted if desired, and the basic pattern of search and print options is the same as in (I).

IV. Inquiry Routine for Compound Number. This routine searches for Compound Numbers, to answer questions dealing with the existence of a specific compound or compounds in the file. The computer searches against the file for up to 50 numbers at one pass. It compares the file against Compound Number and prints out all data for each Number retrieved until all Numbers have been retrieved or the entire file has been read.

NEXT STEPS

When hits indicate promising compounds, additional data about these compounds may be obtained through other parts of the Olin Research Information System. This information may be retrieved through the Olin Compound Registry Number, since this Number is also used in the indexing of all internal technical reports.[1,2]

A permuted index of Wiswesser Line Notations (WLN) of Olin compounds permits rapid assembly of a list of all compounds related to those which have shown promise in the screening tests. As described by Granito *et al.*,[3] other WLN files may be merged with the internal file of Wiswesser Line Notations.

OTHER ACTIVITIES

The results from advanced stages of screening (secondary screening and field testing), can be readily incorporated into the system. Related information from research on analytical procedures, residue procedures, toxicology studies, and market data provide a system searchable for almost any combination of useful data in the development of new agricultural compounds.

Finally, we should mention that involvement of laboratory personnel in retrieval system development not only helped them to re-evaluate their work, but also to become familiar with the fundamentals of the retrieval system. In this way, they became more efficient users and contributors of useful suggestions for improvement.

ACKNOWLEDGMENT

Contributions to the development of this system by F. A. Dowell (formerly, Olin, New Haven), D. W. Ryker, and D. A. Shermer are gratefully acknowledged.

LITERATURE CITED

(1) Ackermann, H. J., J. B. Haglind, H. G. Lindwall, and R. E. Maizell, J. CHEM DOC. 8, 14–19 (1968).

(2) Schlessinger, B. S., and R. E. Maizell, "A New Approach to Indexing Technical Reports in an Industrial Information Center" (paper in preparation).

(3) Granito, C. E., J. E. Schultz, G. W. Gibson, A. Gelberg, R. J. Williams, and E. A. Metcalf, J. CHEM. DOC. **5**, 229–33 (1965).

Heuristic Retrieval: Variable Search Strategies for Identification

EUGENE S. SCHWARTZ
IIT Research Institute, 10 West 35th St., Chicago, Ill. 60616

Received September 12, 1968

Information retrieval is an empirically derived technique for identifying, locating, and retrieving specified information in a data file. The data file is a set of records each of which contains a description of an item in the form of numerical values or alphabetic descriptors. A set of values or descriptors constitute an information vector. A complex item can be described by a set of information vectors, each vector defining a different characteristic.

Given a data file with its sets of information vectors structured in coordinate and subordinate relationships, the problem of information retrieval is to isolate all items that match a specified description. The match is equivalent to satisfying the conditions of a Boolean expression.

Two types of strategy are generally employed in a search operation:

1. Sieve: selection by sorting on designated parameters in a described sequence.
   a. Positive sort: search for items in the file that have specific parameters.
   b. Negative sort: search for items in the file that omit specific parameters.
2. Interactive: open loop feedback between system and user.
   a. Parameter redefinition: results of a previous search are used to change search parameters.
   b. Relevance feedback: user "homes" in through question-and-answer procedure by adjusting the search requests to correspond to relevant items.

EUGENE S. SCHWARTZ

A heuristic retrieval program for identification is described and illustrated by a search experiment with a chemical data file. An adaptive search process in which the next step is dependent upon assessment of future alternatives in the light of past experience is developed from the interaction of three stages of analysis: policy, search, and effectiveness. Data for the analyses are provided by a monitor that maintains data file and search statistics. Heuristic retrieval can provide rapid selection and greater accuracy at lower cost when file items are described by more than one information vector, when both unary and multiple vectors are present or the number of elements in the multiple vectors vary, and the discrimination powers of the vectors also vary.

The first strategy is employed ordinarily in retrieving facts from a data file. The second strategy has been used increasingly in on-line document and text retrieval. In the sieve and interactive searches, the search strategy is the same at all stages. The search routine either traces a path through a tree or other hierarchically organized structure or successively partitions data into relevant and nonrelevant subsets. Although the direction of search can be changed, the strategy remains constant.

A third type of strategy, heuristic retrieval, is described in this paper. In heuristic retrieval, the next search step is dependent upon assessment of future alternatives in the light of steps already traversed. The search process is adaptive in that the steps taken during a search are not known *a priori* but are determined at critical points by decisions based on past experience and future possibilities. The method of retrieval can change together with the direction of search. Heuristic retrieval can call upon a number of specialized routines in response to a search state and the decision criteria applicable to the state. Heuristic retrieval is especially suited for identification where the search objective is the selection of one item that satisfies given specifications.

Heuristic programs have been developed in many applications such as line balancing, machine shop and multiprogram scheduling, chess and checkers, pattern recognition, and theorem proving where combinatorial problems have precluded exact preprogrammed solutions. Identification is not primarily a combinatorial problem but is amenable to a heuristic program because of the possibility to optimize search and to incorporate retrieval experience in selecting a unique item.

The SMART system[1] is a retrieval system that allows a user to process a request under several conditions. Although the system offers a choice of search strategies, each strategy proceeds to its preprogrammed conclusion after which the user can request reprocessing under another strategy. The system is, therefore, interactive rather than heuristic. The SOLID system,[2] a large file of chemical structure data, has provision for a "mobile strategy" although this feature has not been implemented. The strategy may contain some features of heuristic retrieval.

A general discussion of heuristic retrieval for identification within the framework of various search environments appears in the next section. The following section describes an experiment in identifying organic chemical compounds by matching input data against a set of chemical information vectors. The final section discusses a heuristic program to aid in the identificaton process.

HEURISTIC RETRIEVAL

Search Actions and Measures. An item is described by a set of information vectors in accordance with a given classification scheme. The vectors can be unary or multiple. A unary vector has one element with a discrete value. A multiple vector contains more than one element each with a discrete value. Identification is the selection from a file of one item whose information vectors satisfy designated match criteria.

Search actions are dependent, in part, upon the completeness of a data file because data may not be available for all vectors in all items. Accordingly, the following search actions are defined.

A match occurs when data stored for a given vector of a known item are equal to or in the tolerance range of input data of an unknown item. A default occurs when input data for a given vector of an unknown are available and a stored item does not have data for that vector. The nonavailability of stored data precludes obtaining a match but does not rule out the possibility that the item with no stored data is a candidate. A skip occurs when input data for a given vector of an unknown item are not available. In this case, a search of the stored data in the vector file is not possible.

AVAILABILITY AND DEFAULT RATIOS. The availability ratio of a vector is the ratio of the number of items having data available to the total number of items in the data base. The default ratio of a vector is the ratio of the number of items not having data to the total number of items in the data base. These ratios can be expressed mathematically:

$$\alpha = S/C, \beta = D/C = 1 - \alpha$$

where

C is the number of items in the data base
S is the number of items with available data in a given vector
D is the number of items with no data in a given vector
α is the availability ratio
β is the default ratio

DISCRIMINATION. Discrimination is the capability to separate sets of items on the basis of well-defined information vectors and its measure is the discrimination factor. The discrimination factor of a vector is a function of the number of matches and the default ratio inasmuch as the number of possible successful matches is reduced by the nonavailability of data. The defaults can be likened to a sea of uncertainty upon which successful matches float. The discrimination factor, δ, is expressed as:

$$\delta = M \times \beta$$

where M is the number of matches obtained in a vector search.

SELECTIVITY. Selectivity measures the screening capability of a sequential search and deals with the intersections of the vector matches. Selectivity is defined as the ratio of the number of surviving candidates at the end of a stage to the total number of items in the data base. Inasmuch as items with defaults can be present in the candidate list, no adjustment for defaults is necessary. Expressed mathematically,

$$\text{Selectivity}_k = \eta_k = (M_k/C)(k = 1, 2, \ldots, n)$$

where the subscript k designates a search stage, and n is the maximum number of stages.

ACCURACY. The most critical of the measures that describe the results of a sequential search is accuracy. Accuracy is the correctness of identification and is measured by the number of candidates selected in identifying an item. Perfect accuracy would result in the unique identification of an unknown item. Poor accuracy results when more than one candidate is retrieved or a legitimate candidate is eliminated erroneously.

Necessary Conditions. The goal of a heuristic procedure for identification is rapid screening of a data file with perfect accuracy. Heuristic retrieval is feasible when the data file is amenable to variations in search order and the search program is amenable to variations in identification procedure. Specifically, heuristic retrieval is feasible when a data file exhibits the following features:

An item description contains more than one vector.

The vector set includes both unary and multiple vectors or the number of elements in the multiple vectors vary.

The discrimination powers of the vectors vary.

Groups of items are susceptible to different identification procedures.

The first condition implies that a sequential search can be made according to an ordering scheme. The second condition provides a guide to the ordering scheme inasmuch as the time to search the vectors will vary. The variable discrimination powers of the vectors provide a differential screening capability and influence search effectiveness while the differences between groups of items permit variable identification procedures.

Retrieval Tools. Heuristic retrieval is achieved by the interaction of three stages of analysis based upon information supplied by a monitor. The stages are: policy analysis, search analysis, and effectiveness analysis.

Policy analysis consists of evaluating input data and establishing guidelines for the search analysis. The availability of input data, data characteristics, and the search stage will be noted and evaluated. If the recognized data characteristics permit, a preliminary, coarse identification can be made.

Search analysis establishes the direction and method of search appropriate to the information developed in the policy analysis. The goal of the analysis is to facilitate the search and enhance the accuracy. Search order, search keys (vectors), values to be assigned variables, and match criteria will be determined. The search analysis routine will operate before the first search stage and after the final search stage. In addition, the routine can be called upon in any search stage in response to information supplied by the monitor concerning the current status of a search.

The effectiveness analysis specifies the procedures that will optimize search by determining the smallest number of searches than can be performed in a minimum time. The program will first estimate the number of candidates that are expected to survive each search stage and then will order the search sequence to minimize the number of matches required to form the intersection of the stages.

The monitor will collect and make available to the analysis statistical information. One set of statistics will describe the contents of the data file; a second set will contain the results of previous searches. Data file statistics will be updated as a part of file maintenance. Search statistics will be updated by means of counters associated with the search program. The monitor will also maintain pertinent information on the progress of a current search.

A CHEMICAL SEARCH EXPERIMENT

Search Problem. Heuristic retrieval is illustrated here in application to a chemical data file that was used to identify unknown organic compounds by matching their spectra (signatures) with those of known compounds.[3] Data for 500 representative pure organic compounds from 23 chemical groups were collected and a series of experimental searches were conducted. One hundred compounds having data for five signatures were treated as unknowns and screened against the data file. The experiment was designed to test the hypothesis that an unknown compound can be identified by sequentially searching a data file consisting of a minimum number of significant data points in each designated signature and obtaining the logical intersection of the data matches.

A composite chemical signature of a compound consists of a representation of the significant parameters of its component signatures. The composite signature for compound C was

$$C_i = f(M_i, R_i, N_i, G_i, U_i,), i, = (1, 2, \ldots, n)$$

where

M = mass spectrometry signature (MS)
R = infrared spectrometry signature (IR)
N = nuclear magnetic resonance signature (NMR)
G = gas chromatography signature (GC)
U = ultraviolet spectrophotometry signature (UV)

Each signature, in turn, was represented by the parameters selected for the experiment, the data values being the vector elements of a signature vector.

State and elements searches, S_i and E_i, preceded the composite signature search in the experiment.

The match criteria for the seven steps of the sequential search were:

1. State: solid/liquid/gas.
2. Elements: match on combinations of significant elements as coded in column 32 of the ASTM punched cards for infrared spectra.
3. Infrared spectrometry: match on any two of the first three highest absorption bands with transmittance equal to or greater than 10% of the most intense band ($\pm 0.1\ \mu$).

4. Nuclear magnetic resonance: match on the highest peak (±0.1 p.p.m.). Solvents were coded but were disregarded.
5. Mass spectrometry: match on any two of the four highest peaks; matches were made on the mass numbers of ranked peaks but not on the relative amplitudes.
6. Gas chromatography: match on relative retention time of either one of two columns, Silicone SE-30 or Carbowax-20M. A variable tolerance was used.
7. Ultraviolet spectrophotometry: match on the highest peak (±2 mμ).

The objective of a search aimed at identifying an unknown compound was to yield a minimal set of compounds that satisfied the Boolean expression

$$C_j = \{C_{S_i}\} \cap \{C_{E_i}\} \cap \{C_{M_i}\} \cap \{C_{R_i}\} \cap \{C_{N_i}\} \cap \{C_{G_i}\} \cap \{C_{U_i}\}$$

where ∩ = logical AND, and the i and j subscripts designate the stored known and input unknown compounds, respectively.

Ratings were used to evaluate the probability that an unknown was a compound on the candidate list in accordance with weights assigned to the signatures and with the closeness of the match. A skip was rated 0 and a default was rated 1.

In both the MS and IR searches, all combinations of peaks were searched because numerous permutations were found in the rankings of the four major peaks among different data sources. In MS, for example, input peak 1 was matched against stored data peaks 1, 2, 3, and 4 and similarly with the other input peaks. Although any compound that met the minimum match criterion for any search stage became a candidate for identification, use of the rating scale made it necessary to consider also maximum and intermediate matches. Thus, in MS, a compound that matched on all four peaks was a better candidate than one that matched only on the minimum of two peaks. Candidate rating scores are listed in Table I.

Table I. Candidate Ratings

Search	Parameter	Rating	Search	Parameter	Rating
State		2	MS	1-1	16
Elements		4		1-2	12
NMR	± Tolerance	20		1-3	8
	Exact peak	24		1-4	4
GC	Column A:			2-1	9
	± Tolerance	6		2-2	12
	Exact	9		2-3	9
	Column B:			2-4	6
	± Tolerance	6		3-1	4
	Exact	9		3-2	6
	Maximum	18		3-3	8
IR	1 to 1	18		3-4	6
	1 to ±1	15		4-1	1
	1 to 2	12		4-2	2
	1 to ±2	9		4-3	3
	1 to 3	6		4-4	4
	1 to ±3	3		Maximum	40
	2 to 1	8	UV	± Tolerance	4
	2 to ±1	6		Exact peak	6
	2 to 2	12	Any	Skip	0
	2 to ±2	9	Any	Default	1
	2 to 3	8			
	2 to ±3	6		Maximum	130
	3 to 1	2			
	3 to ±1	1			
	3 to 2	4			
	3 to ±2	3			
	3 to 3	6			
	3 to ±3	4			
	Maximum	36			

Monitor Information. FILE STATISTICS. The availability and default ratios of each signature in the 500-compound data file are listed in Table II. The default figures indicate the number of compounds that are automatically candidates in each stage of search, and range from 22 in MS to 267 for UV.

The distribution of stored IR peaks in the data file is shown in Table III. Similar tables have been prepared for the other signatures but are not shown.

SEARCH STATISTICS. Signature discrimination derived from the search for 100 test compounds is listed in Table IV. The range of matches, the average number of matches, the average number of candidates, and the discrimination factor are tabulated for each signature. The match range column indicates the number of matches (less defaults) that occurred in each signature. The average number of candidates per test compound ranged from 60.39 for IR to 303.77 for UV.

Table II. Signature Availability and Default Ratios

Signature	Number of Compounds (C)	Number of Signatures Available (S)	Number of Defaults (D)	Availability Ratio (α)	Default Ratio (β)
MS	500	478	22	0.956	0.044
IR	500	470	30	0.940	0.060
NMR	500	357	143	0.714	0.286
GC	500	308	192	0.616	0.384
UV	500	233	267	0.466	0.534
Totals	2500	1846	654	0.738	0.262

Table III. Distribution of Stored Peaks Infrared Spectrometry

	Number of Compounds				Number of Compounds		
Band	Peak 1	Peak 2	Peak 3	Band	Peak 1	Peak 2	Peak 3
2.9	5	1	6	7.6			
3.0	31	13	16	7.7	1	4	8
3.1	4		2	7.8	14	9	4
				7.9	8	10	6
5.6	4			8.0	13	9	8
5.7	7	7	1	8.1	12	4	1
5.8	30	27	8	8.2	10	5	7
5.9	27	3	3	8.3	8	9	4
6.0	14	9	2				
6.1	2	4	4	8.8	8	5	6
6.2	12	4	7	8.9	8	4	11
6.3	9	7	5	9.0	6	10	9
6.4	5			9.1	2	3	4
6.5		1	2	9.2	4	10	3
6.6	1	9	1	9.3	1	6	4
6.7	2	5	10	9.4	4	4	2
6.8	33	24	38	9.5	5	12	5
6.9	21	26	19	9.6	2	6	10
7.0	4	9	8	9.7	2	3	6
7.1	1	7	13	9.8	6	4	5
7.2	2	16	15				
7.3	3	26	22	13.3	2		3
7.4	6	15	9	13.4	6	2	8
7.5		1		13.5	5	2	2

The average number of IR matches on peak combinations obtained from a sample of 25 input test compounds is listed in Table V. The results of the input peak 1 on stored peak 1 and input peak 2 on stored peak 2 matches were as expected. The number of 3 on 2 matches was slightly greater than the 3 on 3 matches. The last column of the table lists for each input peak the percentage of candidates in each peak combination.

Table VI lists the average number of candidates obtained with varying match criteria, exclusive of defaults. The term exclusive in Table VI refers to matches on one peak only, two peaks only, and three peaks only. The term inclusive refers to every instance of one, two, or three matches; for example, the number of matches on one peak includes the number of matches on two and three peaks.

The average number of candidates obtained with varying match criteria in the MS search is listed in Table VII. The number of matches obtained on MS peak combinations has been calculated but is not shown. The number of candidate compounds obtained with different GC match criteria is listed in Table VIII. No statistics for NMR and UV are tabulated because their information vectors are unary.

The test results showed that selectivity varied according to the search order and the group of structurally related compounds. The accuracy also varied according to the group with the alcohols and hydrocarbons being the most difficult to identify.

HEURISTIC SEARCH PROGRAM

The chemical information system described in the previous section meets all the conditions specified for a heuristic retrieval program, and the program outlined here has been developed from the evaluation of 100 test searches in the 500-compound file. It is anticipated that the file will consist eventually of tens or hundreds of thousands of compounds and candidate lists in the early stages of search may include many hundreds or thousands of compounds.

The rapid elimination of noncandidate compounds before multiple-vector searches can result in large cost savings. A steep selectivity curve will also reduce the memory required to store candidate lists and ratings and will minimize the intersections of the signature searches.

Policy Analysis. The policy analysis program will include but not be restricted to the followng steps:

1. Search stage

At the start of the first search stage, steps 2, 3, and 4 will be followed. At the end of the last search stage, step 5 will be followed.

2. Availability of input data

The input information vectors will be scanned. Lack of input data for any signature will necessitate a skip in the search of that signature. If the skip is in a highly discriminating signature, a reordering of the search sequence may be desirable to achieve more rapid selection of candidates.

3. Data characteristics

Vector values will be noted in the input data scan. The presence of infrequent values will call for the use of the average number of expected candidates in effectiveness calculations; common values will call for estimating procedures. Values that afford little discrimination—e.g., 85 compounds that were transparent to UV—will be noted and the information will

Table IV. Signature Discrimination for 100 Known Compounds

Signature	Match Range (Less Defaults) Low	Match Range (Less Defaults) High	Average Number of Matches	Number of Defaults	Average Number of Candidates	Discrimination Factor
MS	1	153	49.83	22	71.83	2.19
IR	1	84	30.39	30	60.39	1.82
NMR	1	59	31.22	143	174.22	8.93
GC	1	39	11.74	192	203.74	4.51
UV •	1	85	36.77	267	303.77	19.64

Table V. IR Matches on Peak Combinations (25 Input Compounds)

Peaks Input	Peaks Stored	Average Number of Matches	% Matches in Combination
1	1	41.72	51
	2	22.12	27
	3	18.44	22
2	1	24.36	31
	2	27.28	35
	3	26.20	34
3	1	26.40	28
	2	34.56	36
	3	33.92	36

Table VI. Number of Candidate Compounds Obtained with Different IR Match Criteria (100 Input Compounds)

Number of Peaks Matched	Match Range (Inclusive) Low	Match Range (Inclusive) High	Average Number of Matches Exclusive	Average Number of Matches Inclusive
1	38	282	161.28	191.67
2[a]	2	84	27.05	30.39
3	1	12	3.34	3.34

[a] Criterion used for the experiment: matches on two or more peaks.

Table VII. Number of Candidate Compounds Obtained with Different MS Match Criteria (100 Known Input Compounds)

Number of Peaks Matched	Match Range (Inclusive) Low	Match Range (Inclusive) High	Average Number of Matches Exclusive	Average Number of Matches Inclusive
1	3	313	103.54	153.37
2[a]	1	153	40.23	49.83
3	2	38	7.85	9.60
4	1	6	1.75	1.75

[a] Criterion used for the experiment: matches on two or more peaks.

Table VIII. Number of Candidate Compounds Obtained with Different GC Match Criteria (100 Known Input Compounds)

Match Criterion	Input Data	Number of Matches, Range Low	Number of Matches, Range High	Average No. of Matches
Column A	96	1	26	7.62
Column B	83	1	20	6.32
A and B	79	1	7	1.44
A or B[a]	100	1	39	11.75

[a] Criterion used for the experiment.

be provided to the search analysis for sequence ordering. The presence of particular values will be used for preliminary identification as in step 4.

4. Preliminary screening

A preliminary screening can be made using significant data values. A screen based upon UV transparency or a peak <272 mμ and the presence of an IR band at 3.0 (±0.1) μ, for example, narrowed the chemical group to diol or alcohol. The presence of an IR band at 6.8 (±0.1) μ and values of all MS peaks $\leqq 130$ indicated a possible hydrocarbon. The alcohol and hydrocarbon groups identifications in both cases were not definitive inasmuch as a few other compounds passed the screen. However, all the test alcohols and hydrocarbons were selected.

5. Post-search identification

Upon completion of all search stages, the results will fall into one of four categories:

a. Zero selection—no compounds
b. Plural selection—more than one compound
c. Conditional selection—one compound isolated with less than "perfect" rating; identification cannot rule out other possibilities
d. Unique selection—positive identification of single compound having a "perfect" rating

Results a and d require no further action. Results b and c may be improved by calling on a fine search. The chemical group of the single conditional candidate or the compounds in b that exceed a threshold rating will be looked up and the information provided to the search analysis program.

Search Analysis. For input items that have a complete set of vectors with infrequent values, a standard search sequence and procedure will be established initially. For other items, the search analysis program will prescribe search procedures based upon the information provided by the policy analysis. As clues to identification are obtained in the course of the sequential search, the search analysis will be alerted by the monitor and the program can call upon any combination of the procedures below to achieve faster selection.

1. Information vector

Information vectors can be modified in two ways. First, the number of elements in a vector can be fixed or variable. Ordinarily, a fixed number of elements will be designated, for example, three absorption bands in IR. It may be necessary to search additional bands to aid in the identification of some chemical groups. Second, the vector element itself may be modified. For a fine search, for example, the relative intensity of designated MS peaks may be searched together with the mass numbers.

The inclusion of additional parameters and variable information vectors in the search necessitates, of course, the incorporation of these data in the data file. However, these data can be restricted to those compounds that search experience indicates may be difficult to identify.

2. Discrimination level

The discrimination of a vector is a function of the default ratio, the tolerance limits, and the match criteria as has been demonstrated in Tables IV, VI, VII, and VIII. By varying the tolerance and/or the match criteria, greater or lesser discrimination can be achieved.

3. Fine search

To aid in identification, especially in cases where candidate ratings are close, it may be desirable to focus a signature search on a narrow band or a specific parameter. Thus, in IR, for example, a narrow band may be searched or the absence of designated peaks may be noted. The fine search will require that appropriate data be included in the data file.

Effectiveness Analysis. SEARCH ENVIRONMENT. The choice of procedures for optimizing the search is dependent upon the monitor statistics and the search environment. The environment is defined by file characteristics, file structure, and computer system.

Significant file characteristics are the number of items in the file, vectors per item, dimensions of the vectors, and the length of a record.

The file can be structured in one of five major types or their variants: serial, sorted serial, inverted, list, and threaded.

The serial file is a single file with all information vectors of an item appearing in one record and the records are arranged according to a designated key. A sorted serial file is ordered so that items indexed by certain keys can be found in a specified portion of the file. Only that portion of the serial file that contains the query values need be searched. In the inverted file all items whose information vectors have a designated value are located in a single record. Search for a specific vector will lead directly to the vector record.

A list is a sequence of items that are connected through a field in each item's record which contains the address of the succeeding item. For items that are described by a set of information vectors, a separate list can link each vector. In the threaded file, each record contains all vectors of an item as in the serial file. Lists are threaded through the records to link vector elements of similar value. One record will be linked in as many lists as there are elements in the information vectors. The lists and threaded lists create the logical equivalent of inverted files.

Computer characteristics that affect search effectiveness include: tape or disk operation, bit transfer rate and access time, instruction execution times, memory cycle time, core storage capacity, and operating system, especially input-output handling.

Other factors being equal, the file structure will have the greatest bearing on search effectiveness. Of the three major processing considerations in selecting a structure, storage, updating, and retrieval, retrieval is the most important because it is repetitive. Updating will be infrequent compared to retrieval and will also involve fewer records. The storage required by each file structure will differ both in regard to the data file and the working or temporary storage during search operations.

A comparison of retrieval times required for searches on serial and inverted files has been made by Curtice.[4] A discussion of inverted and threaded lists including timing estimates is contained in a paper by Lowe.[5] A chemical information system that utilizes threaded lists has been described by Lefkowitz and Powers.[6] Fossum analyzes inverted and list files in a study of retrieval systems and presents the advantages and disadvantages of each structure in regard to search outputs, search algorithms, file maintenance, file storage, and query processing requirements.[7]

The experiment described in the previous section was performed using an inverted file in a Termatrex manual card system. Numbers of compounds having specified data values were drilled in corresponding cards, and intersection was accomplished by optical coincidence. However, because the system was not amenable to complex logical searches a great deal of manual recording and searching was necessary. It is anticipated that an expanded retrieval system will be processed by computer and the data file will be list-structured, each signature forming one list. Each compound in a list will also be linked to the other signatures of the compound. By this organization full advantage can be taken of the difference in information vectors and the difference in discrimination power of each vector.

PROCEDURES. With the information from the policy and search analyses and the file and search statistics provided by the monitor, the effectiveness program will establish an optimum search sequence. The program will calculate the number of compounds that will be expected to survive each search stage and the sequence will be determined using the principle of intersection by minimum overlay.

In intersection by minimum overlay, the smallest set of items is matched against the larger set(s). Given expected candidates of 50 and 20 in two signatures, the signature with 20 candidates would be searched first inasmuch as the candidates obtainable from the intersection can never exceed the minimum number.

The expected number of candidates of a signature search can be estimated in two ways. If the value of a signature is not a frequent one, it need not be included in a table of file statistics and the expected number can be considered to be the average discrimination. For a commonly occurring value that appears in a table, the expected number will be obtained from the table.

For estimates made while the search is in progress, the expected number of candidates will be the product of the average discrimination or table value and the selectivity of the preceding stage. The selectivity will serve as a proportionality factor on the assumption that the expected candidates are equally distributed throughout the data file and hence are directly proportional to the number of surviving candidates.

In MS and IR where multiple vector signatures are present, the number of expected candidates can be calculated by taking the product of a table value or average discrimination and the percentage of matches in a peak combination. Consider the IR peak of a 6.8 μ in Table III, for example. The expected number of candidates for the combinations of input peak 1 on stored peaks 1, 2, and 3 are:

Input Peak 1	Stored Peak	No. Compounds (Table III)	% Matches (Table V)	Expected Candidates
	1	33	0.51	16.83
6.8 μ	2	24	0.27	6.48
	3	38	0.22	8.36
			Total	31.67

Thus an IR input peak 1 of 6.8 μ can be expected to match with 32 compounds from a data base of 500 compounds. The expected number of 1 on 1, 1 on 2, and 1 on 3 matches appear in the column on the right.

The expected number of candidates can be calculated for each signature in the manner described. For signatures where the match criterion requires agreement with more than one value, combinations of single-valued matches must also be matched. The combinations in IR, with a match criterion of two out of three peaks, are matches on 1 and 2 peaks, 1 and 3 peaks, and 2 and 3 peaks, in addition to the triple match of 1, 2, and 3 peaks. In MS, with a match criterion of 2 out of 4 peaks, there are six two-peak combinations, three three-peak combinations and one four-peak combination.

An algorithm for determining a minimum number of records to be searched given an arbitrary logical condition has been described in an Auerbach report.[8] The algorithm converts an expression from parenthesis to prefix format and then replaces each symbolic term by the appropriate statistic. In a right to left scan, when an OR operator is found, the preceding two terms are replaced by the smaller of the sum or the total number of items in the file. When an AND operator is found, the preceding two terms are replaced by the minimum of the numbers.

The gain in effectiveness obtainable from a variable search sequence is illustrated with two compounds in the state and elements search stages. The number of compounds listed in the file statistics for *p*-methoxyphenol and tetraethyl lead are:

Compound	State		Element	
p-Methoxyphenol	Solid:	95	Oxygen:	300
Tetraethyl lead	Liquid:	387	Lead:	1

In the case of *p*-methoxyphenol, the first search should be on state (= solid) and the second on element (= oxygen) to achieve a minimum number of compounds to be searched: 95 plus the intersection of 95 and 300. In the case of tetraethyl lead, the inverse sequence should be followed inasmuch as a search on element will yield one compound and only this compound need be searched in the state stage.

Given the list structure described above, only those compounds that are candidates need be accessed and searched in any stage. The signatures of the candidates in the next successive stage will be readily obtainable from the signature links.

The results of two sequential searches ordered by different criteria are illustrated by an example from the experimental test. The first sequence of search for crotonaldehyde was ordered by ascending dimensionality of the information vectors. The second sequence was ordered by ascending value of the default ratio. The selectivity of the former is shown in Figure 1 and the latter in Figure 2 (state and element stages were omitted in these sequences).

The presence of 143 defaults in the NMR data and 267 in the UV data accounted for 80 candidate compounds that were carried into the third stage of search, as indicated in Figure 1. The 22 and the 30 defaults in the MS and the IR data, respectively, gave rise to no double-default candidates, as indicated in Figure 2. In the sequence ordered by default ratio, 80 signature searches were performed after 69 compounds were selected from 500 in the first stage. Contrasting with this number were the 309 signatures searched in the sequence ordered by the dimensionality of the information vectors. In tests on 10 compounds, the signatures searched in the former sequence totaled 855 while only 170 were searched in the latter sequence.

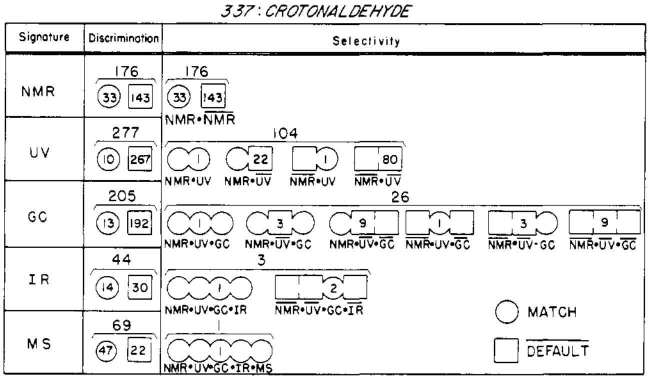

Figure 1. Selectivity with sequence ordered by number of elements in information vectors

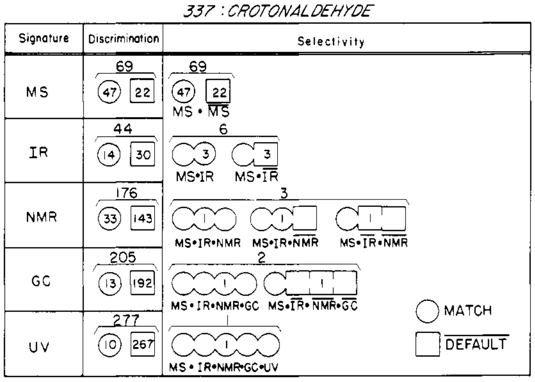

Figure 2. Selectivity with sequence ordered by default ratio

In the case where signature searches involve the same number of elements per vector and number of match criteria, the cost differential of the two sequences is obvious. In the case where the vectors and match criteria differ, the potential cost saving must be balanced against the cost per signature search. To determine the economic feasibility of an heuristic retrieval program, in general, the cost to provide the necessary statistics and perform the necessary calculations must be balanced against the savings in the search program as compared to a standard, nonvariable search.

CONCLUSIONS

A heuristic retrieval procedure especially suited for identification has been described. An adaptive search process in which the next step is dependent upon assessment of future alternatives in the light of past experience is developed from the interaction of three stages of analysis: policy, search, and effectiveness. Data for the analyses are supplied by a monitor that maintains data file and search statistics and also information on the progress of a current search.

Policy analysis evaluates input data and establishes guidelines for the search analysis. Search analysis establishes the direction and strategy of search. Effectiveness analysis specifies the procedures that will optimize search by determining the smallest number of searches in minimum time.

Heuristic retrieval can provide rapid selection and greater accuracy at lower cost when file items are described by more than one information vector, when both unary and multiple vectors vary, and the discrimination powers of the vectors also vary.

The proposed heuristic procedure was formulated in connection with a chemical data file that was used to identify unknown organic compounds by matching their spectra against those of known compounds. Results of the search experiment were used to illustrate the techniques of heuristic retrieval.

ACKNOWLEDGMENT

The study of identification of organic chemical compounds by sequential matching of spectral data on which this paper is based was conducted with the support of the Chemical Information Unit of the National Science Foundation under contract No. NSF-C514.

LITERATURE CITED

(1) Salton, G., and M. E. Lesk, "The SMART Automatic Document Retrieval System—An Illustration," *Comm. ACM* 8 (6), 391, June 1965.

(2) deMaine, P. A., and B. A. Marron, "The SOLID System 1: A Method for Organizing and Searching Files," in "Information Retrieval: A Critical Review," George Shecter, Ed., Thompson Book Co., Washington, D. C., 1967.

(3) Scholz, R. G., E. S. Schwartz, and M. E. Williams, "Feasibility Study of the Development of a Specialized Computer System of Organic Chemical Signatures of Spectral Data," IIT Research Institute, Report No. C6104-4, prepared for the National Science Foundation (PB-178 354), April 29, 1968.

(4) Curtice, Robert M., "Magnetic Tape and Disc File Organizations for Retrieval," Center for the Information Sciences, Lehigh University, July 1966.

(5) Lowe, Thomas C., "Design Principles for an On-Line Information Retrieval System," Moore School of Electrical Engineering, University of Pennsylvania, AD 647 196, December 1966.

(6) Lefkowitz, David, and Ruth V. Powers, "A List-Structured Chemical Information System," in "Information Retrieval: A Critical Review," George Shecter, Ed., Thompson Book Co., Washington, D. C., 1967.

(7) Fossum, Earl G., and G. Kaskey, "Optimization and Standardization of Information Retrieval Languages and Systems," Sperry Rand Corp., AD 630 797, January 1966.

(8) Development of a System Design Specification for the Automatic Data Processing Subsystem of Reliability Central, Auerbach Corp., April 1965.