the approach used. The following sequence of steps has proven most effective:

1. Author known
   Go directly to author folder.
2. Subject known
   a. Visually examine subject deck, retrieve documents from author folders.
   b. Consider alternate subjects if material retrieval in (a) is insufficient.
3. Olin Key List Compound or Project number known
   a. Visually examine decks 2 or 3.
   b. Use the subjects listed in 2 or 3 as a clue to the subject deck.
4. Patent information (memoranda of invention, active files, etc.)
   Visually examine deck 4.
5. Correspondence with Olin subsidiary or affiliated corporations and with outside companies
   a. Material from others to Olin will be in file folder of respective author or company.
   b. Material to outside will be filed in Olin author folders. These are sectioned by the location of the authors as listed on the tab card.
   Visually examine deck 5 in fields 8–22 and 23–37 under the appropriate name (as given on letterhead of organization).
6. A "zero" deck has been created and the cards therein filed alphabetically by author and then by recipient within the author group. By means of an IBM 870 Document Writer or by use of a Xerox machine, it is possible to provide any author with a listing of his correspondence for any given period in a matter of minutes. For those with voluminous correspondence, this service has proven valuable for checking purposes and will probably be continued only for these specific individuals rather than for all authors.

COMMENTS

As the volume of material in our center has grown, the chance of finding desired information in this centralized location has also grown. Although a few individuals were reluctant to participate at the outset, our present ability to retrieve information rapidly has converted them. We now have a total of approximately 8500 items after not quite one year of operation.

Current work load is about 1100 to 1200 items filed per month, with an average of 2.23 punched cards per item. Thus, the load on the system is roughly 50 items per day, requiring the punching of about 125 cards per day. Input is steadily increasing as acceptance grows. However, increasing operating efficiency has so far been able to keep step. We expect the value of TCC to increase as the years go by.

A fringe benefit of TCC has been reduction of space devoted to file cabinets in offices and laboratories. In fact, a surplus of used cabinets is soon going to present a problem in some areas.

Merging of individual file contents has resulted in a marked deproliferation of paper—we save only one copy of each document. Our "burn" box must be emptied frequently and stacks of accumulated and unfiled papers are disappearing from atop files, desks, and tables in the various offices and laboratories. The unofficial motto appears to be "send it to TCC, let them worry about it," and that is exactly the purpose of our service.

Weighted Term Search: A Computer Program for an Inverted Coordinate Index on Magnetic Tape*

F. W. MATTHEWS and L. THOMSON
Canadian Industries Limited, Central Research
Laboratory, McMasterville, Quebec, Canada

Received January 5, 1967

Ten to 15 years ago, much technical literature was accurately indexed according to strict rules of classification. Today, because of the large amount of technical data written, it is no longer practical to apply the same rigid indexing procedures. We should therefore be concerned with new search techniques which will allow us to handle input at the lower level we are forced to accept. This paper describes such a search technique.

The problem of searching a coordinate index is discussed by Fairthorn (*1*). The classical method is to use a logic based on Boolean algebra. Combinations of terms using this logic involve a series of operations which are fully discussed by Becker and Hayes (*2*) and Williams (*3*). The argument is that the possible relevance of a document is worth investigating if a stated combination of terms is found in the document. Since in coordinate indexing

* Presented before the Division of Chemical Literature, 152nd National Meeting of the American Chemical Society, New York, N. Y., September 1966.

there may be little control of terminology, expressing a search concept usually involves a series of terms with their synonyms and near synonyms. Handling these requires either a series of Boolean statements or one statement in which several sets of alternate terms are used.

The first alternative may involve preparing a large number of statements which for *n* terms can become (2^n-1) statements. This is discussed in a recent book by Sharp (*4*). The second alternative gives no indication of the particular set of terms which combined to give an answer. In an attempt to overcome this difficulty Catley (*5*) used weights to tag the terms used to express a concept in the Boolean expression. The onus of preparing the Boolean statements is left to the searcher. In practice it is all too easy to compose statements which produce either no answers or a larger number than is practical to use.

In the search method described in this paper, the searcher selects terms to express each concept involved. He is asked to rank the concepts and the terms used to express each one. The Boolean logic is handled by the program in that all possible combinations of "and," "or," and "but not" are considered, and within limits set by the searcher presented in an ordered list which takes into consideration the preferences stated in the ranking of terms. This method of searching a coordinate file was proposed at a conference sponsored by Information for Industry (IfI) in June 1965 (*6*).

WEIGHTED SEARCH

A weighted search is based on selecting a group of search terms which express each concept of the inquiry, assigning to each a weight, indicative of its relative importance and computing a "score" for answers resulting from combinations of these terms. Answers with a high "score" indicate the cooccurrence of preferred terms for each concept and hence should have a high relevance to the inquiry. The search is not restricted to any one expression of terms, but considers all possible combinations. Those search statements likely to give answers with a low degree of relevance are eliminated by a screening process which takes into account a minimum score and a minimum number of concepts to be coordinated.

Answer sets corresponding to each acceptable combination of search terms are printed in descending sequence of total score. The user is able to follow the particular combination of term-weights associated with each answer set since they are displayed in the form of a scan-column index. This feature assists in the evaluation of answers and in finding combinations of terms which on analysis prove to be particularly relevant to the search. Limiting conditions may be specified to suppress the printing of a long list of answers, in which case the total number of answers within each set is displayed.

Scores computed for answer sets are based on the cooccurrence of preferred terms, which express each concept or subject area to be considered in the search. Having derived a term combination which includes groups of several alternate synonyms and generic terms, the program selects the preferred term from each basic concept and computes a score based only on the preferred terms. Nonpreferred terms are included in the coordination but their assigned weights do not contribute to the total "score." For this reason, the presence of several alternate synonyms does not affect the order of presentation of answer sets. This strategy provides an effective means of evaluating the relevance of answers, since relevance has been determined by the number and importance of the subject areas which are present in each answer. It is not simply based on the cooccurrence of search terms but on the cooccurrence of the concepts, taking into account the relative importance attached to words used to express each.

This weighted search system has been evaluated in its application to the retrieval of patents from the Information for Industry Patent File (*7*). This file contains all patents in the chemical section of the *U.S. Patent Gazette* for the years 1950–1964 plus selected patents from the electrical and mechanical sections which have significant chemical relevance. The number each year is shown in Table I. The total number for the 15 years 1950–1965 is 158,876. Each has been indexed by an average of 28 terms. While there is a variation in rate of growth of the file, the general trend is an increase in growth from 600 patents per year, to a present growth of about 2000 patents per year. Table I also shows the portion of the file selected from the mechanical and electrical sections. On the average this number is about 10% of the total, but can be as high as 26%. In 1956 the chemical section of the *U.S. Patent Gazette* consisted of 8233 patents; the IfI Patent File included 11,108 patents, of which 2875 were selected from other sections.

In indexing the patents, terms have been selected from the text without use of a control list. This has led to some rather interesting terms which might surprise a user who has not read the entire IfI vocabulary. An example of this is shown in the following list (the letter underlined shows the location of the term in the alphabetized list):

All Skin Rayon
Non Soap Detergent
B-Stage Resin
Thin Boiling Starch
Super Polyester
Unleaded White Gasoline
Semidrying Oil
T-joint
Ready-to-Cook

Such a list emphasizes the need for structuring this vocabulary of terms in a logical manner. A move in this direction has been made by the Dow Chemical Co. who developed a permuted term index (KWIC) which lists each term and significant syllable. We also have compiled a KWIC index for terms, in which we have separated chemical and nonchemical terms into two lists. A page from the nonchemical section is shown in Table II.

When a coordinate index is available on magnetic tape, statistical information concerning the use of indexing terms may be obtained at low costs, through the use of the computer. A computer program was written which prints a frequency table of terms giving the total posting for each term as well as the frequency by year. An example from this table is shown in Table III. In the column at the extreme right, a calculation of the proportion of the references which occurred in the last two years is given. This is a measure of the recent activity of that

Table 1. U. S. Chemical and Chemically Related Patents

Year	Chemical Patents Issued[a]	All Patents IfI File	Chemically Related Patents in File	% Added
1950		6,777		
1951		7,273		
1952		6,848		
1953	5,802	6,361	559	8.8
1954	5,588	6,134	546	8.9
1955	4,691	6,065	1,374	20.6
1956	8,233	11,108	2,875	25.8
1957	7,620	8,833	1,213	13.8
1958	9,551	10,633	1,082	10.2
1959	10,382	11,532	1,150	10.0
1960	8,866	9,801	935	9.6
1961	10,127	10,998	871	7.9
1962	11,753	13,521	1,768	13.1
1963	10,041	12,269	2,228	18.0
1964	11,333	14,675	3,342	22.8
1965	13,193	16,048	2,855	17.6
		142,828		

[a] U. S. Patent Office (2).

term. Since computer time for a given search is largely dependent on the number of patent references associated with search terms, the frequency table often forms the basis for selecting or not selecting terms of questionable importance to a given search.

In addition to a frequency table in term sequence, a table was prepared which lists terms in descending sequence of total frequency count. These data were used to prepare a graph of a percentage frequency count for terms (Figure 1). The graph shows that 50% of all terms have a total posting of 100. About 30% of the terms have a frequency count of less than 10, while 10% of the terms have a frequency count of over 1000. About 1% have a posting of over 10,000, while about 2% of the terms have a frequency count of 1.

For additional assistance in phrasing an inquiry, the user should also consult a technical thesaurus, such as the Engineers Joint Council Thesaurus (*8*), for possible alternate synonyms, generic, and related terms which may be suitable for search.

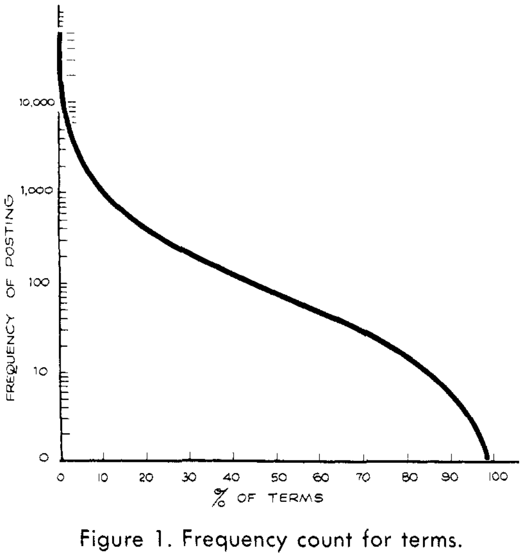

Figure 1. Frequency count for terms.

As an illustration of the steps involved in preparing for a computer search, consider the following example: Find patents concerning acrylic ester polymers which are soluble in natural hydrocarbons. As prepared for search it could be expressed by the following terms selected from the IfI vocabulary as shown in Table IV.

The inquiry has been broken down into three main concepts—*i.e.*, (A) acrylic polymer, (B) solubility, (C) natural hydrocarbon. Each concept is expressed by a number of terms indicated by the alphabetic code. Synonyms and related terms falling within the same concept have thus been identified. The concepts have been ordered according to their importance in the inquiry and within the concepts by importance of terms. The weighting affects only the order of presentation, not the selection; hence, it cannot lead to the rejection of pertinent items. The cooccurrence of a number of concepts may be specified in this case; at least two of the three concepts is required for an acceptable answer. Coordinations within a single concept have a low degree of relevance. A minimum score is used to eliminate irrelevant answers. In this case the minimum score is that associated with term 14. This assures that the A concept is included in all answers.

The integration of all three concepts which have been expressed by the preferred terms would give the highest scoring answers. The presence of alternate synonyms within each concept would then determine the order of presentation of answers within each set having the same score. The answer to the inquiry would consist of a series of sets of patent references, each set having a lower score, and hence a poorer match with the question.

The first 10 answer sets obtained from this inquiry are shown in Table V.

The left side of Table V gives the references expressed as IfI accession numbers, along with the conversion to U.S. patent numbers. The term-weights associated with each answer set are shown on the right in the form of a type of scan-column index (*9*).

Consider Set Number 1. It consists of one patent reference, Access No. 2726-1951, U.S. Patent No. 2552775. This reference had the highest score and hence represents the best match between the inquiry and all of the patents in the file. The term-weights associated with it are 9, 13, and 18. These weights had been assigned to the terms "Polyacrylate," "Solubility," and "Hydrocarbon Oil." This combination of terms is the best since it uses the preferred terms found for expressing all concepts. The terms "Polyacrylate" and "Solubility" represent the preferred terms for concepts A and B, respectively; the term "Hydrocarbon Oil" with a weight of 9 is the second best term of the C concept which will coordinate with "Polyacrylate" and "Solubility" to give an answer. The preferred terms of the C concept, which is (C8–C24) Hydrocarbon with a weight of 10, had been coordinated with the terms "Polyacrylate" and "Solubility," but none of the patents in the file had this combination of terms.

The order of presentation of answers is not markedly affected by the presence of several alternate synonyms. This point is illustrated in Set Number 5. It consists of four patent numbers, retrieved from different combinations of search terms. The term-weight combination of 18, 13, and 1 is basic to each of the four search statements. These term-weights are the highest within each of the

Table II. Permuted Term Index

*SEMI-DRYING	OIL	062000
*SESAME	OIL	062260
*SHALE	OIL	062430
*RICE BRAN	OIL	060090
*RICE	OIL	060100
*SYNTHETIC	OIL	068370
*SPRAY	OIL	065755
*SOY BEAN	OIL	065250
*SPICE	OIL	065493
*SPINDLE	OIL	065530
*DISTILLATE, DISTILLATE	OIL	023720
*RESIDUAL	OIL	059650
*RE CYCLE	OIL	058950
*DIESEL FUEL, DIESEL	OIL	021990
*GEAR,-GEAR	OIL	030300
*FOOTS,-FOOTS	OIL	029070
*FLUSHING,-FLUSHING	OIL	028900
*FUEL	OIL	029660
*ESSENTIAL	OIL	026160
*OITICICA	OIL	048190
*OLIVE	OIL	048330
	*OIL	048030
*NON DRYING	OIL	046560
*PETROLEUM	OIL	051520
*PINE	OIL	053230
*BABASSU	OIL	006795
*ANIMAL	OIL	004280
*ANTHRACENE, ANTHRACENE	OIL	004470
*COAL TAR	OIL	015250
*COCO NUT	OIL	015530
*BUNKER	OIL	010450
*CALSOLENE	OIL	011820
*CASTOR	OIL	012840
*INDUSTRIAL	OIL	036145
*ACID	OIL	000550
*LINSEED	OIL	039640
*CRUDE	OIL	017960
*CYCLE	OIL	018630
*CUTTING	OIL	018440
*CRACKED NAPHTHA, CRACKED	OIL	017528
*CRANK CASE	OIL	017600
*TALL OIL FATTY ACID, TALL	OIL ACID	068580
*HYDROGENATED FISH	OIL ACID	034170
*CASTOR	OIL ACIDS SA-RICINOLEIC-ACID	012850
*SOY BEAN	OIL ACIDS, SOYBEAN OIL FATTY ACIDS	065260
*DRYING	OIL AND ACID	024480
*SUNFLOWER	OIL AND FATTY ACID	067770
*PALM	OIL AND FATTY ACID	049840
*DEHYDRATED CASTOR	OIL AND FATTY ACID	019890
*COTTON SEED	OIL AND FATTY ACID	017350
*RAPESEED	OIL AND-FATTY-ACID	058530
	*OIL BURNER	04800
	*OIL CLOTH	048055
*WATER-IN-	OIL EMULSION	075800
*COCO NUT	OIL FATTY ACID	015540

three concepts, and therefore contribute to the score for this set. Alternate synonyms—for example, 12 for the B concept, and 14 and 15 for the A concept—were included in the coordination and determine the order within the set.

An effective screen is provided to eliminate answers with a low degree of relevance by the use of both the minimum number of concepts and the minimum score as limiting conditions. The elimination of answers which fall below these limits during the coordination or logic phase results in a considerable reduction of computer time in the sort and print phases which follow.

Restrictions can also be set to control the number of answers printed in reply to particular inquiry. This is accomplished by means of one or both of the following limiting conditions: (1) a print minimum score limit can

Table III. Frequency Table

Term	Code	Total	Year Totals																% of 65 64
			65	64	63	62	61	60	59	58	57	56	55	54	53	52	51	50	Reference
COAGULATION, CLOTTIN	15200	1778		115	156	145	127	121	122	96	87	107	51	72	87	212	156	124	15.
COAL	15210	1494		79	66	104	61	85	99	107	115	106	75	114	114	251	79	39	9.
COALESCENCE	15220	281		26	50	31	25	30	13	11	9	19	7	7	10	23	9	11	27.
COAL TAR	15230	364		24	24	38	23	27	26	16	14	32	12	16	9	46	33	24	13.
COAL TAR DISTILLATE	15240	27		0	0	1	6	11	3	0	0	0	1	0	2	0	3	0	.
COAL TAR OIL	15250	42		2	2	1	4	10	6	3	2	0	1	0	2	2	4	3	9.
COAL TAR PITCH	15260	111		7	5	6	15	6	12	12	11	10	3	4	6	5	6	3	10.
COARSENESS,-COARSE	15272	347		19	47	38	18	28	29	9	7	20	11	12	24	44	16	25	19.
COATING SA-ENAMEL,-P	15280	13432		1634	1384	1125	884	922	1048	855	663	879	652	525	515	945	783	618	22.
COAXIL	15290	237		81	67	21	10	10	9	2	6	0	10	3	2	7	7	2	62.
COB	15295	22		0	1	2	0	0	0	0	5	7	4	0	0	3	0	0	4.
COBALT	15300	2906		354	180	253	193	187	203	222	216	202	138	125	131	254	124	124	18.
COBALT-56	15305	01		0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	.
COBALT-60	15310	126		18	17	17	14	24	16	8	5	3	3	0	0	0	1	0	27.
COBALT ACETATE	15320	208		26	7	23	15	22	8	16	11	19	6	11	6	18	13	7	15.
COBALT CARBONATE	15330	93		4	7	7	9	8	14	10	8	6	3	6	2	3	5	1	11.
COBALT CARBONYL	15340	216		4	6	8	10	16	13	30	18	27	13	20	11	23	12	5	4.
COBALT CHLORIDE	15350	249		24	19	29	14	27	20	17	17	17	8	11	14	19	7	6	17.
COBALT HALIDE	15357	42		4	1	5	3	1	3	0	1	11	12	0	0	0	0	1	11.
COBALT-COMPOUND, COB	15360	441		10	61	92	22	23	41	41	27	28	0	30	17	31	13	5	16.
COBALT DYE	15370	114		0	2	0	3	21	22	36	12	0	0	5	5	8	0	0	1.
COBALT FLUORIDE	15375	26		3	1	0	0	2	1	0	1	0	1	1	3	0	12	1	15.
COBALT MOLYBDATE	15380	338		24	22	23	29	54	66	26	9	33	8	9	11	13	10	1	13.
COBALT NAPHTHENATE	15390	526		30	17	43	44	38	59	37	36	41	11	25	27	56	38	24	8.
COBALT NITRATE	15400	179		15	6	22	18	17	17	10	8	7	0	9	12	21	9	8	11.
COBALT OCTANOATE	15403	27		0	3	2	11	1	0	3	0	2	0	0	0	0	5	0	11.
COBALT OLEATE	15405	44		4	2	3	5	3	2	6	0	12	0	3	2	0	0	2	13.
COBALTOUS ACETATE	15407	47		4	0	0	8	3	8	5	4	4	0	2	3	3	1	2	8.
COBALTOUS CHLORIDE	15410	100		7	2	6	18	4	6	12	7	5	1	2	5	9	8	8	9.
COBALT OXIDE	15420	712		54	66	78	71	54	71	55	33	55	21	31	25	43	24	28	16.
COBALT SULFATE	15430	212		11	5	17	13	17	18	27	32	21	5	5	11	19	5	6	7.
COBALT SULFIDE	15440	107		15	11	12	11	11	4	0	12	11	5	5	3	0	4	3	24.
CO CATALYST	15450	45		21	3	5	5	10	1	0	0	0	0	0	0	0	0	0	53.
COCCIDIOSIS	15460	108		9	7	17	5	6	10	11	5	8	6	5	2	3	7	7	14.
COCCOMYCES AND-SPECI	15470	02		0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	.
COCKPIT	15480	06		1	0	0	3	0	0	0	0	1	0	1	0	0	0	0	16.
COCKROACH SA-ROACH	15490	124		4	9	9	24	9	9	1	1	1	0	32	3	11	8	3	10.
COCOA	15500	122		6	6	12	16	8	16	8	11	5	2	5	2	5	5	14	9.
COCOA BUTTER	15510	63		3	4	6	13	4	5	4	6	3	1	4	1	3	2	4	11.
COCONDENSATION	15515	12		0	0	0	0	2	0	1	4	0	4	0	1	0	0	0	.
COCO NUT	15520	59		7	2	8	7	0	0	2	8	2	4	5	0	4	1	9	15.
COCO NUT OIL	15530	369		2	23	46	25	19	44	27	28	16	21	18	13	52	16	19	6.
COCO NUT OIL FATTY A	15540	385		44	17	10	41	28	36	27	19	38	5	16	19	47	22	16	15.
COCO NUT SHELL	15545	05		0	0	0	0	3	1	0	0	0	0	0	0	0	1	0	.
CO CRYSTALLIZATION	15550	09		4	1	2	1	0	0	0	1	0	0	0	0	0	0	0	55.
CO CURRENT	15555	18		3	3	12	0	0	0	0	0	0	0	0	0	0	0	0	33.
COD	15560	21		2	3	0	2	4	3	0	0	0	0	4	0	1	2	0	23.
CODE	15570	36		15	10	0	9	1	0	0	0	0	0	1	0	0	0	0	69.
CODEPOSITION	15580	09		0	1	1	1	0	0	5	1	0	0	0	0	0	0	0	11.
CO DIMER	15585	18		1	0	2	0	1	0	1	4	0	0	2	3	0	3	1	3.

Table IV. Search of Ifl Index to U.S. Chemical Patents

Inquiry No. 03 Requested by No. of Search Terms 18	Hydrocarbon Soluble Acrylic Ester Polymers Dr. Strangelove of the Bond Institute Minimum Score 0000016384	Question No. 32 August 32, 1988. Max. No. Answers Printed 0050

Term Codes			
54170	POLYACRYLATE	18	A
54680	POLYMETHYL ACRYLATE	17	A
54430	POLYETHYL ACRYLATE	16	A
54660	POLYMETHACRYLATE	15	A
54710	POLYMETHYL METHACRYLATE	14	A
65010	SOLUBILITY	13	B
65030	SOLVENT	12	B
65020	SOLUTION	11	B
11185	C-8 C-24 HYDROCARBON	10	C
33960	HYDROCARBON OIL	09	C
65060	SOLVENT NAPHTHA	08	C
44790	NAPHTHA	07	C
37950	KEROSENE	06	C
43590	MINERAL OIL	05	C
15250	COAL TAR OIL	04	C
39470	LIGROIN	03	C
30230	GASOLINE	02	C
33923	HYDROCARBON	01	C

be specified, so that answers whose score is below this limit are not printed; and (2) the maximum number of answers printed within a particular set can be specified. When the second limit is reached the total number of answers in that particular set is given.

As well as providing controls to limit the number of answers printed, provision is made in the print program for documentation of the inquiry. The title of the search, the inquirer, date, and print controls are displayed along with the list of search terms (Table IV).

The weighted search program has been written in "COBOL" specifically for use on IBM 1410 equipment and IBM 360 series and other units for which a "COBOL" compiler is available. The computer system flow chart is shown in Figure 2.

The search routine consists of several phases, which, however, may be run as a series under control of the computer operating system. The first phase is the search of the inverted master tape file for terms entered on inquiry cards, ordered in sequence of terms. This phase is essentially an extract of the list of patent references for the terms specified in the search. The output is a tape containing patent references which are in order of terms and random with respect to inquiry number. The tape records are sorted in sequence of IfI accession number for the logic phase which follows.

The logic phase coordinates patent numbers, attempting all possible combinations of terms within an inquiry, and accumulates a score. The program selects concepts, disregarding synonyms associated with each concept in computing scores. The minimum number of concepts required for a logical statement and the cutoff score for each inquiry are specified by the use of control cards. Answers which do not statisfy these limits are eliminated. An additional number which is the sum of all weights converted to a power-of-two is also developed to provide the mechanism for the playback of terms in the print phase. The output of this phase is a tape containing answers, described by an IfI accession number, its conversion to U.S. patent number, and a score and playback number. Answers are then sorted in descending sequence of score for the print phase to follow. The sorted tape serves as the "save-tape" for the computer search, and may be rerun through the print phase if the limits were set too high.

Several inquiries are usually processed together in any one computer search, providing for more economical use of the computer. The computer time for a given search is largely dependent on the number of patent references associated with the search terms. It is dependent to a lesser extent on the total number of search terms used. A typical question with 10 terms having an average frequency of posting would require 12 minutes on an IBM 1410 (40K) computer, for which current charges are $15.00. This time assumes that the inquiry is one of about five such inquiries which have been processed together. The sort efficiency is the major factor in determining the over-all computer time. As a rule, over 80% of the references listed in the highest scoring answer sets are relevant; if the inquirer scans the first few answer sets, he will have the majority of the pertinent answers.

The user's selection and preference of terms is determined by his interpretation of the inquiry and his knowledge of the vocabulary of the index. The particular group of terms which were selected from the text of a patent at the time of indexing is unknown to him; he can only speculate on the probability of finding relevant patents using his selected group of terms. If the user wants all of the relevant answers that had been retrieved from the inquiry, he should scan the remaining answer sets, using the scan-column index of terms, as a guide. Often, he will find answers which satisfy the inquiry, even though they had not been indexed by the preferred terms of the search.

Frequently the user fails to consider all of the synonyms and generic and related terms which have been used in the index to express one of the concepts of the inquiry.

Table V. First Ten Answer Sets from Inquiry

							C			B		A
INQ	NO	03										
SET NO IF	ACC NO	YR	US PATENT NO							TERMS		
1	002726	51	2552775					09		13		18
2	003453	53	2647101				07			1213		18
3	005503	54	2695270	01		05				13		18
4	001534	53	2634260	0102						1213	15	18
	004355	55	2719833	0102						13		18
5	006436	63	3095388	01						131415		18
	010631	59	2915482	01						1213	15	18
	003337	53	2646416	01						1213		18
	007484	63	3098835	01						13		18
6	001752	60	2928797							131415		18
	010656	59	2915507							13	15	18
	003889	63	3085885									
	000591	52	2583991									
	004167	60	2940952							111213		18
	001593	60	2927906							1213		18
	005562	55	2727015									
	001667	57	2784184									
	001462	58	2825711									
	003335	53	2646414									
	001073	54	2671065						11	13		18
	007403	64	3140204									
	005937	64	3136636									
	009698	63	3107464							13		18
	003873	59	2885271									
	004487	64	3131630									
	009277	60	2964487									
	000556	55	2701391									
	012265	56	2751368									
	007490	57	2812314									
	006110	58	284703									
	001260	57	2782183									
	013047	56	2754240									
7	003298	60	2936300		03	0506		09		12	15	18
	003112	60	293584	01				09		12		18
	002056	53	2637718									
8	005643	57	2803611				07			12	15	18
9	000747	53	2628923			06				12	15	18
10	009008	60	2963388	02						1112	15	18
	002165	60	2930768	0102						12		18
	007664	57	2813083	02						12		18

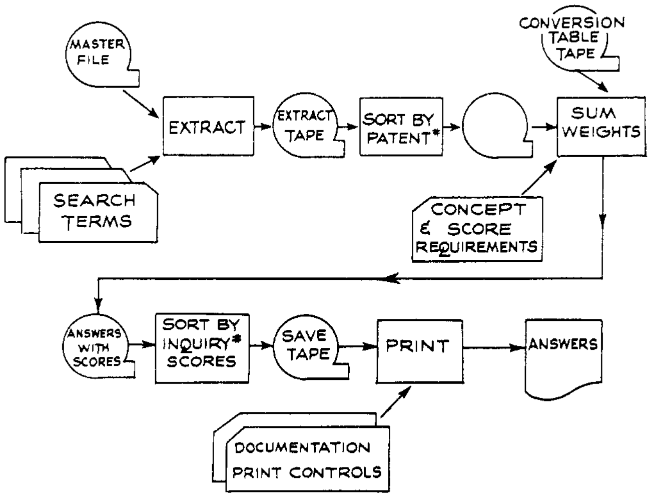

Figure 2. Flow chart.

The flexibility of this technique allows him to retrieve answers resulting from the coordination of terms in which some of the search concepts had not been expressed. These answers would have low scores and would not be printed near the top of the list. The user is, however, able to consider the answers which only partially satisfy his search expression. It is important to note that these answers would not in many cases be available if the Boolean technique of specifying a particular term combination had been applied.

To illustrate this point consider an item from one of the answer sets obtained for the example discussed previously, concerning "Acrylic Polymers Soluble in Natural Hydrocarbons." The term-weights associated with this set were 18 and 12—that is, "Polyacrylate" and "Solubility." The "Natural Hydrocarbon" concept is missing in this term combination. The first claim from one of the patent references in this set, U.S. Patent Number 3050484, is

> **Bituminous Protective Coating and Method of Use** - James Q. Wood. "A bituminous *coating* composition comprising a normally liquid polymerized conjugated diene having from 4 to 8 carbon atoms per Molecule, a *polymer of acrylic acid* selected from the group consisting of polyacrylic acid and a copolymer of acrylic acid and a conjugated diene having from 4 to 8 carbon atoms per molecule, and from 50 to *90 weight percent asphalt.*"

In listing terms which cover the concepts of "Natural Hydrocarbon," the searcher did not consider "asphalt," the final fraction in the distillation of petroleum. If the search technique had required the presence of one of the specified "Hydrocarbon" terms, this pertinent reference would not have been retrieved. This example points out the risk of suppressing relevant answers by restricting the search to specific term combinations.

The search technique described combines efficient search of an inverted file with an ease and flexibility in expression of the search. The output is ranked in probable order of relevance and provides a further aid to retrieval: a scan-column index of terms coordinated in each answer.

LITERATURE CITED

(1) Fairthorn, R. A., "Towards Information Retrieval," p. 61, Butterworths, London, 1961.

(2) Becker, J., and Hayes, R. M., "Information Storage and Retrieval," p. 335, John Wiley, New York, N. Y., 1963.

(3) Williams, W. F., "Principles of Automated Information Retrieval," p. 219, The Business Press, Elmhurst, Ill., 1965.

(4) Sharp, J. R., "Some Fundamentals of Information Retrieval," p. 99, London House and Maxwell, New York, N. Y., 1965.

(5) Catley, J. M., Moore, J. E., Banks, D. G., and O'Leary, P. T., J. CHEM. DOC. **6**, 15 (1966).

(6) Matthews, F. W., Communication to the Information for Industry Users Conference, Washington, D. C., 1965 (not published).

(7) "Uniterm Index to U. S. Chemical Patents," Information for Industry, 1000 Connecticut Ave., Washington, D. C. 20036.

(8) "Thesaurus of Engineering Terms," Engineers Joint Council, 347 East 47th St., New York, N. Y., 1964.

(9) O'Connor, J., *Am. Doc.* **13**, 204 (1962).

CORRECTION

Quality Control and Auditing Procedures in the Chemical Abstracts Service Compound Registry

The Computer-Based Subject Index Support System at Chemical Abstracts Service

In the above articles by D. P. Leiter, Jr., and H. L. Morgan [J. CHEM. DOC. **6**, 226–9 (1966)] and D. J. Whittingham, F. R. Wetsel, and H. L. Morgan [J. CHEM. DOC. **6**, 230–4 (1966)], respectively, Figure 2 has been interchanged. The captions should remain with the respective articles, but the location of the illustrations should be reversed.