Patent Office
Research and Development
Reports . . . . . No. 13

Frome, Julius

A SYSTEM OF RETRIEVAL COMPOUNDS, COMPOSITIONS, PROCESSES, AND POLYMERS

Prepared by

Julius Frome, Jacob Leibowitz, and Don D. Andrews
Office of Research and Development
Patent Office

Joseph D. Grandine, Steven T. Polyak, and Karl G. Siedschlag, Jr.
Research Division, Textile Fibre Department
E. I. duPont de Nemours & Co.

Office of Research and Development
Patent Office

November 17, 1958

Robert C. Watson
Commissioner of Patents

Lewis L. Strauss
Secretary of Commerce

For sale by the U. S. Department of Commerce, Washington 25, D. C. Price 25 cents.

Table of Contents

A SYSTEM OF RETRIEVAL COMPOUNDS, COMPOSITIONS, PROCESSES, AND POLYMERS

INTRODUCTION

This paper describes a continuation of the efforts of the U. S. Patent Office to develop and use a mechanized search system for patent searching.

It includes a general discussion of the logic developed by the Patent Office in its approach to the problem and a description of two machine applications of such logic, namely, (a) Use of the "Ilas" (4) punched card machine; (b) Use of the Bendix Computer.

The application to the Bendix Computer is the result of a joint effort between the U. S. Patent Office and I.E. du Pont de Nemours, Textile Fibres Dept.

It has been demonstrated in a limited field, such as the steroid art, a machine search system can be of much practical assistance in patent searching (1).

Following the experiences with the steroid search system, an extension in scope and versatility of the system to encompass various other types of subject matter and relationships appeared to be desirable. A subsequent method, termed "Variable Scope Search System" (2) presents a more universally applicable machine search method insofar as organic chemical compounds are concerned.

A further development has been the extension of machine searching to encompass not only organic compounds, but also inorganic compounds, functions, processes, reaction conditions, and so on.

ART SELECTED

The art involved for experimentation with the new method is the polymer art; selected for the reason that the subject matter is fairly complex and it was felt an adequate mechanized search system in this art would indicate how to handle comprehensive subject matter involving a great deal of variety in compound disclosures, functions and processes.

The art of the ethylenic unsaturated homo and copolymers, classified in Class 260, subclasses 80-94.9 (3) constituted the initial point for this investigation.

CRITERIA AND DESIDERATA

In determining the criteria and desiderata for the system, a sample of the patents and literature was thoroughly analyzed. The important and significant features were noted and tabulated. From the tabulated data, the following types of disclosure entities were selected as significant.

1. Monomers used in polymerization
   - a. Homomonomers, e. g., ethylene, acrylonitrile
   - b. Copolymers, e.g., butadiene and styrene
   - c. Terpolymers
2. Inerts in the reaction, e.g., CO_2, N_2, etc.
3. Solvents in the reaction
4. Catalysts
5. Processes
   - a. Chemical
     - Sulfonation
     - Polymerization
     - Chlorination
   - b. Physical
     - Washing
6. Conditions of the processes
   - a. Temperature
   - b. Time
   - c. Pressure
7. Properties of the product
   - a. Molecular Weight
   - b. Density
   - c. Crystallinity
   - d. Melting Point
   - e. Viscosity
8. Uses

As a result of this analysis, two major subdivisions of disclosure were set forth, namely, "ingredients" and "functions." Ingredients are terms relating to chemical compounds and "functions" include nonstructure terminology identifying processes, properties, conditions of reaction, and so forth.

DEVICES (LOGICAL)

Identification of Compounds

In considering how the chemical compounds should be identified various factors entered into the determination of the method to be used. It was desired to use the system if successful, on a production basis on a relatively large body of art. The total number of patents in the selected resin patents is about 12,500 originals and cross-references. It was estimated that about 5,000 different compounds existed in the resin art. It was desirable that the method of identification of the compound be compatible with the method used in the "VS_3" system (2). Keeping these features in mind the following method was used to identify the chemical compounds. Each compound was given a unique number. The hexadecimal system of numbering was used. Four hexadecimal characters were used to constitute a subject matter field. Since there are 16 variations in each hexadecimal character, four hexadecimal characters in combination allow for 65,536 variations. In some instances, the individual bits of the characters are used to obtain searches on an "inclusion" basis.

a. In addition to unique identification of each specific compound its genus-species relationship is also provided for. It is quite common for a document to state "olefins such as ethylene, butylene and propylene can be homopolymerized." The specific subject matter codes for ethylene, propylene and butylene are set forth, i.e.,

0 4 D 1	(ethylene)
0 4 D 6	(propylene)
0 4 C E	(butylene)

However, only one set of generic codes, which is equally applicable to all three constituents, is used, i.e.,

0 1 8 5 (Mono Olefin Hydrocarbon)

This generic description of the species disclosed is done as an inherent descriptor of the species even if not explicitly stated in the disclosure.

In the identification of inorganic compounds, the cations and anions are assigned unique codes and are grouped together to indicate their association as being in the same compound. Thus sodium nitrate would be represented as:

0 2 3 7	(sodium)
0 4 4 D	(nitrate)
0 2 3 5	(Group IA metal)

The three codes would be treated as a group by means of a "grouping" signal which associates the three descriptors as belonging to the same entity.

Identification of Functions (Modulant)

A survey of the functions necessary for this body of art indicated that provision for 256 different functions would be adequate. By using two hexadecimal characters in combination it is possible to obtain 256 unique codes. The following codes are examples:

03	Solvent
71	Homomonomer
72	Comonomer
73	Termonomer
3D	Catalyst

The codes for the functions are put in the "modulant" field of the word, since the modulant is a modifier of the subject matter code. In a disclosure of "ethylene as a homomonomer," the codes would be as follows:

Modulant	Subject Matter
71	0 4 D 1

If it is desired to express ethylene as a comonomer, the code would be:

Modulant	Subject Matter
72	0 4 D 1

If one wanted to express the fact that a particular compound functions as a solvent, the code would be 03 in the modulant field. Although the modulant field and subject matter field modify each other, they can be searched independently of each other. For example, if disclosures of all compounds acting as solvents are desired, 03 would be the code, in the modulant field, and the subject matter codes would be ignored. Although only a few functions have been illustrated, it will be apparent that many variations are possible. Thus, if the code in the subject matter field for polymerization is 0 4 A 5, by using the modulant field the following combinations can be obtained.

	Modulant	Subject matter	
	0 0	0 4 A 5	Polymerization
Solvent	0 3	0 4 A 5	Solvent polymerization
Catalytic	3 D	0 4 A 5	Catalytic polymerization

Relationships—Groupings

There are two devices used by this system to show grouping relationships or associations, namely signals and interfix.

a. Grouping (signals)

The signals are devices used to bracket or group component parts of an entity together. For example in describing sodium nitrate a grouping signal is used as follows:

0 2 3 7	(sodium)
0 4 4 D	(nitrate)
0 2 3 5	(Group IA metal)
S_5	(Signal)

The signal serves to make a phrase or group out of the three codes by machine recognition of the signal as the terminus of the preceding codes. The grouping signal prevents the various codes from one group being confused with that of another group.

In the same manner the various codes pertaining to a process entity are grouped together by another signal S_7.

Finally, all the codes are grouped together as pertaining to the same document by signal S_6 which indicates the end of a document.

b. Interfix

The interfix shows interrelationship among the various groups. For example the group for specific compounds and specific process can be interfixed or connected together to show that they are in the same process.

Modulant	Subject matter	Interfix
Homo	Ethylene	
Signal 5		X
Solvent	Benzene	
Signal 5		X
Catalyst	Aluminum	
Catalyst	Chloride	
Signal 5		X
Catalyst	Polymerization	
Time		
Temperature		
Pressure		
	Starting material	X
	End material	X
Signal 7		X X
Molecular weight	Polymer product	
Melting point		
Signal 5		X
Signal 6		
Patent or document #		

It can be seen that the various groups are connected by the interfix device. The interfix device also permits the showing of sequence. Thus, those groups which are interfixed to the term starting material are starting materials and the terms interfixed to the product are products. It should be noted that a product of one reaction could be a starting material of another reaction.

Numerical Descriptor—Ranges

The numerically stated descriptors such as temperature, time and pressure, pose a special problem. Consider the following situations.

Given the disclosures of temperature, as follows:

A. 40-49°C   B. 1-45°C   C. 60-65°C

(a) A search for a disclosure of 42°C specifically should result in retrieval of disclosures (A) and (B) but not (C). The specific temperature must therefore be included within the broad ranges.

(b) A search for a disclosure of 40-59°C should result in retrieval of (A) and (B) since the range in the question is partially included within (A) and partially within (B).

(c) A search for a disclosure 1-65°C should result in retrieval of (A) (B) (C) (even though the expressed question range is broader than each of these disclosure ranges)--since the question is partially included within each of (A) (B) and (C).

Hypothetical Dictionary

	20° Range		*10° Range*	*Sample code*
		1-19 (9)	1-10	(1)
	1-39 (12)		11-19	(2)
		20-39 (10)	20-29	(3)
			30-39	(4)
1-79 (14)			30-49	(5)
		40-59 (11)	50-59	(6)
	40-79 (13)		60-69	(7)
		60-79 (15)	70-79	(8)

The three disclosures would be coded as follows:

A for 40-49°C	codes 5, 11, 13, 14
B for 1-45°C	codes 1, 2, 3, 4, 5, 9, 10, 11, 12, 13, 14
C for 60-65°C	codes 7, 15, 13, 14

In searching, the code asked for is the nearest range to that desired but the selected range must be all inclusive in range desired. Thus in searching for a temperature of 42°C, the code selected would be code 5 (40-49°C). In applying the search code to the disclosure it is seen that the disclosures 40-49°C and 1-45° answer the question. If it is desired to search for a range of 40-59°C the proper code selected would be code 11. In searching the disclosure for code 11 it is seen that both (A) and (B) answer the question.

Weighting

It has long been a desideratum for the patent examiner to be able to find not only complete answers to the search question but the closest art, in the absence of the finding of a complete answer. It would also be desirable to be able in effect to ask several questions at the same time. One of the most frequently asked questions is for any of several species or for any member of a Markush group. In cooperation between the U. S. Patent Office and a research group from E. I. du Pont de Nemours Co., Textile Fibres Division, the following procedure called weighting was developed in conjunction with the program to be used on a digital computer.

Each subject group in a question set is assigned a relative numerical value, called a weight in accordance with its considered importance.

During the search, the machine operates in the following manner. When the subject is found for its first appearance in the disclosure the weight assigned to the subject is recorded. The weight is not recorded for any additional finding of the same subject. As each additional subject is found its weight is added to the previously recorded weights of the other subjects. The total weight at the end of the search is compared with a minimum weight which has been assigned concomitantly with the questions. Assume, for example, subject A is assigned a weight of 1, B a weight of 2 and C a weight of 3. When the machine finds subject A, it records weight 1 and when it finds C it adds 3 to one and if it finds B adds 2 to give a total weight of 6. A complete answer to the question is A+B+C. If, in addition, answer B+C is also acceptable, the minimum weight assigned would be, 5, i. e. (B=2, C=3, B+C= 3+2=5). The machine would then find all disclosures containing (A+B+C) and also those containing (B+C).

By requiring the search to have a certain minimum weight, multiple questions can be sought simultaneously. Multiple search can be made for alternatives, such as occur with respect to Markush situations.

The following examples illustrate the procedure.

a. Assume a search for "ethylene homopolymerized in presence of a peroxide catalyst and benzene solvent." In the absence of a complete answer, the next best answer may be considered to be a disclosure of homopolymerized ethylene with a peroxide catalyst, but no solvent being necessarily present.

	Weight
homo ethylene............	2
catalyst peroxide.........	2
solvent benzene...........	1
Minimum weight.......	4

Thus both homoethylene and peroxide as a catalyst must be present.

b. A search is desired wherein "either ethylene, propylene or butadiene is homopolymerized with a peroxide catalyst and benzene solvent."

	Weight
homo ethylene..................	1
homo propylene................	1
homo butadiene.................	1
catalyst peroxide..............	4
solvent benzene................	4
A minimum weight of	9

Accordingly those disclosures which have a peroxide catalyst, a benzene solvent and at least one of either ethylene, propylene or butadiene, will answer the question.

OTHER APPLICATIONS

Although the system has been described with respect to chemical subject matter it will be apparent that it can be applied to other subject matter. For example, one could describe a motor as follows.

Modulant	Subject matter
Electric............	Motor
Steam...............	Motor
Gasoline	Motor

The interfix relates the motor with a gear. Thus

	Interfix
Electric motor.............	X
Gear	X

The interfix can show that the gear is connected to the motor. In this manner many relationships can be built up and used.

USAGE

The system described is in operation. This system is currently being used by the Patent Office to search resins in Class 260, subclass 94.9, using the "ILAS" punched card machine (4). Preliminary tests with a computer program have been encouraging.

MACHINES

Use of a Computer

Introduction

It was decided to determine whether the logic of the described system could be adequately programmed and used practicably on a commercially available computer, such as the Bendix (G-15-D). In cooperation with E. I. duPont de Nemours & Company, Textile Fibers Department, a flow chart and program were written for the Bendix Computer and an operable program set up.

Word format in the file

There are two types of words, "subject" word and "interfix" word. The word format for subject word is as follows:

	Modulant		Subject matter			
T	M	M	SM	S-M	S-M	S-M

T = O Hexadecimal characters

Hexadecimal character

The word consists of three divisions, the field "T" Modulant and subject matter. In the file, T is always zero, thereby identifying a subject word. The next two hexadecimal characters contain the modulants and the next four hexadecimal characters are used for subject matter identification. The format of the interfix word is as follows:

S_1	I	I	I	I	I	I

The first hexadecimal character identifies the signal. The next six characters are allotted for the interfixes. It should be noted that the interfix pattern although recorded by hexadecimal characters, is searched on a bit basis.

The "end of patent" word is actually two words as follows:

S_1	C	O	O	O	O	O

N	N	N	N	N	N	N

S_6 Country of patent Patent Number

The first hexadecimal character is a signal S_6. In the second character the country of origin is recorded. The remainder of the 7 digit word contains zeros. The next 7 hexadecimal characters of the succeeding word contain the patent number.

Word format in the question

The word format for a question is as follows:

	Modulant		Subject matter			
T	M	M	S	S	S	S

The question word format is the same as that of the file word except for the first digit "T". In the file word, T is zero, whereas in the question word it varies according to the question. The symbol M=modulant, S. M.=subject matter, D=disclosure, Q=question. T is defined as follows:

When the search requirement is for--

T=V, M and S.M. identical in D and Q.
T=Z, M identical in D & Q, S.M. of Q included in S.M. of D.
T=8, M is disregarded, S.M. of Q to be included in S.M. of D.
T=9 =M is disregarded, S.M. identical in D & Q.

Weighting procedure

By means of the weighting procedure described, the computer can determine if the minimum weights have been met and thus retrieve closely related art in addition to full anticipations, if any.

Print out

The results of the search are printed out in seven columns, according to the following chart.

	Col. 1	*Col. 2*	*Col. 3*	*Col. 4*	*Col. 5*	*Col. 6*
Max. weight..	yes	yes	yes	yes	no	...
Min. weight..	...	...	...	...	yes	yes
S-count OK..	yes	yes	yes	yes	yes	no
I-count OK..	yes	yes	no	no	...	...
Relationships OK..................................	yes	no	yes	no	...	...

Column 7 contains the numbers of the subject groups which were not found.

Definitions

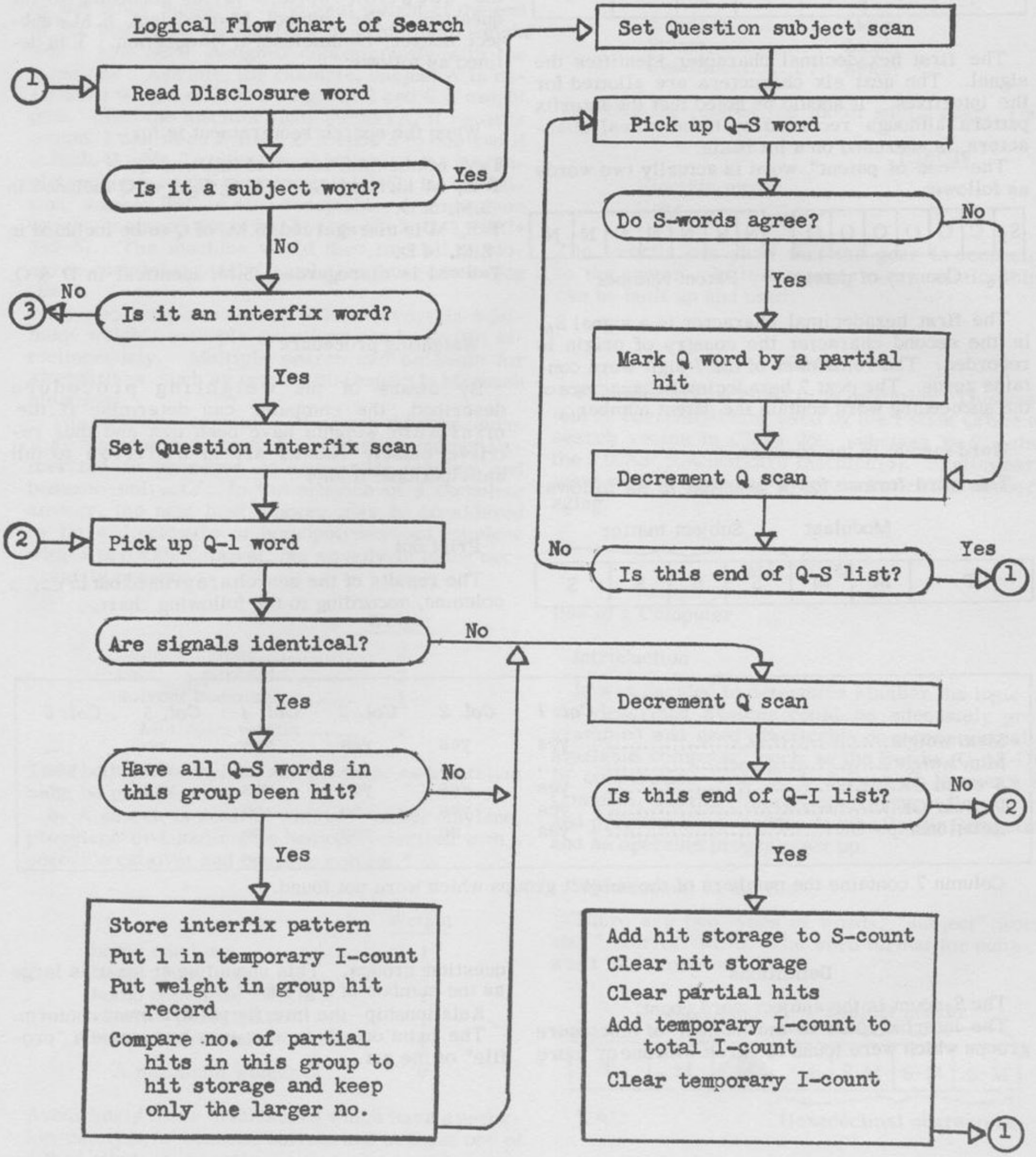
The S-count is the subject word count.

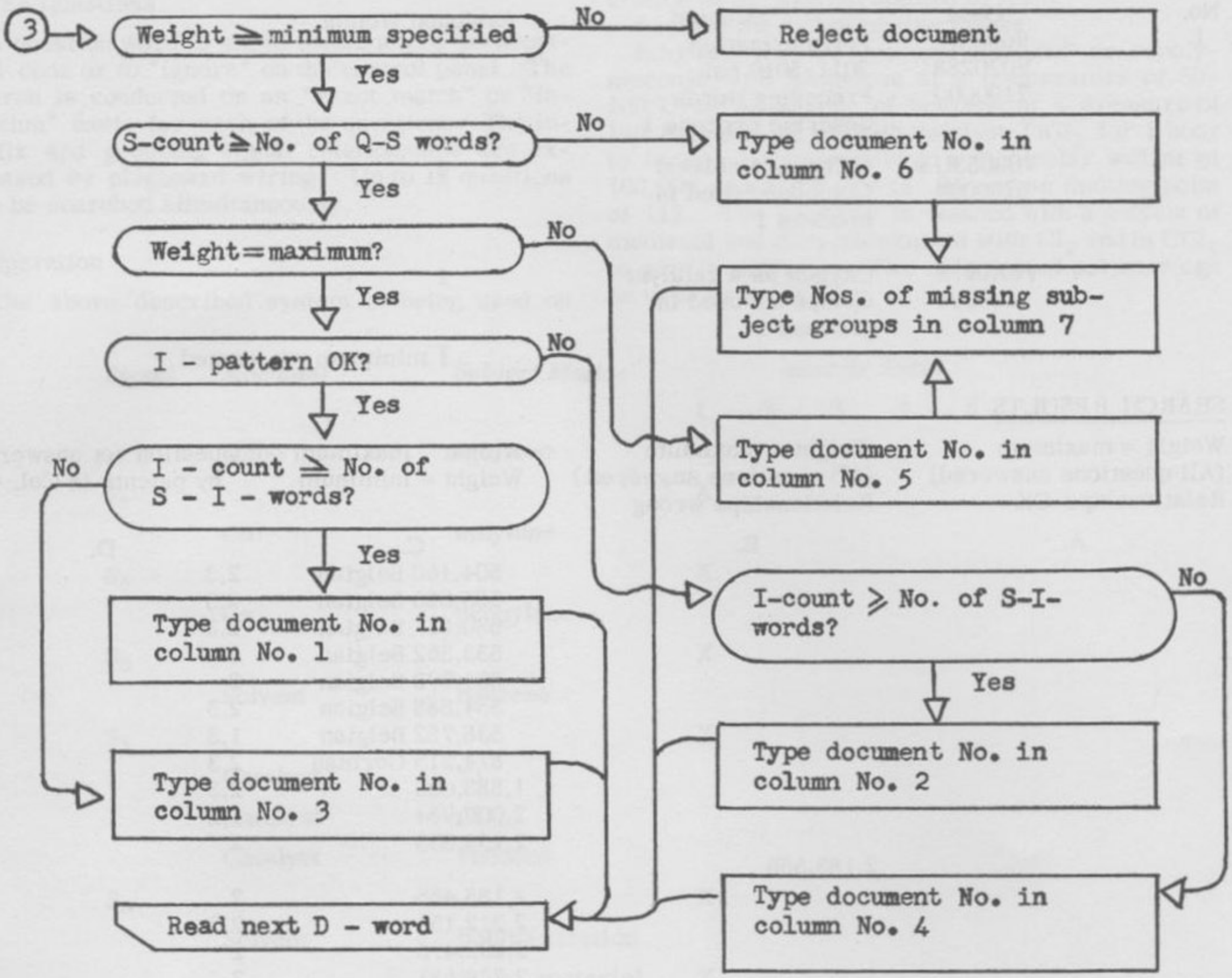
The interfix count is the number of disclosure groups which were found to agree with one or more question groups. This should be at least as large as the number of S groups in Q.

Relationship--the interfix pattern must conform.

The print out gives what may be termed a "profile" of the art.

SEARCH FLOW CHART


Logical Flow Chart of Search
1
Read Disclosure word
Is it a subject word?
Yes
No
Set Question subject scan
Pick up Q-S word
Do S-words agree?
No
Yes
Mark Q word by a partial hit
Decrement Q scan
Is this end of Q-S list?
No
Yes
1
Is it an interfix word?
No
3
Yes
Set Question interfix scan
2
Pick up Q-l word
Are signals identical?
No
Yes
Decrement Q scan
Have all Q-S words in this group been hit?
No
Yes
Is this end of Q-I list?
No
2
Yes
Store interfix pattern
Put 1 in temporary I-count
Put weight in group hit record
Compare no. of partial hits in this group to hit storage and keep only the larger no.
Add hit storage to S-count
Clear hit storage
Clear partial hits
Add temporary I-count to total I-count
Clear temporary I-count
1


3
Weight ≥ minimum specified
No
Reject document
Yes
S-count ≥ No. of Q-S words?
No
Type document No. in column No. 6
Yes
Weight = maximum?
No
Type Nos. of missing subject groups in column 7
Yes
I - pattern OK?
No
Yes
Type document No. in column No. 5
No
I - count ≥ No. of S - I - words?
Yes
I-count ≥ No. of S-I-words?
No
Type document No. in column No. 1
Yes
Type document No. in column No. 2
Type document No. in column No. 3
Type document No. in column No. 4
Read next D - word

Bendix Computer Search in Patent Class 260 sub 94.9

QUESTIONS

No.	Code	Meaning	Assigned Weight
1	9000001	Polymerization	1
	V070288	2011-5010 psi	
	7100000	Preceding words describe process 1	
2	V03053U	CCl_4 as a solvent	1
	5100000	Compound used in process 1	
3	V3X0475	Oxygen as a catalyst	1
	5100000	Compound used in process 1	
			$\bar{1}$ minimum wt. wanted

SEARCH RESULTS

Weight = maximum (All questions answered) Relationships OK	Weight = maximum (All questions answered) Relationships wrong	Weight maximum Weight = minimum	Question not answered by patents in Col. C
A.	B.	C.	D.
		504,160 Belgian	2,3
		525,025 Belgian	2,3
		530,617 Belgian	2,3
		533,362 Belgian	2
		534,792 Belgian	2
		534,888 Belgian	2,3
		538,782 Belgian	1,3
		874,215 German	2,3
		1,885,060	2,3
		2,000,964	1,2
		2,153,553	2
	2,183,556		
		2,188,465	2
		2,212,155	2,3
		2,232,475	2
		2,238,681	2
	2,261,757		
		2,325,060	2,3
		2,327,705	1,2
2,334,195			
		2,342,400	2
		2,377,779	2,3
		2,387,755	2
		2,388,138	2

ILAS

ILAS, a punch card machine designed by U. S. Patent Office and built by Bureau of Census has been previously described (4).

Although ILAS has available an 80 bit word, it was decided for experimental reasons to use a 40 bit word.

The logic of the system used has been described above.

Word format in file or disclosure

The 40 bit word was divided as follows:

Signal	Modulant		Subject Matter				Interfix											
S	M	M	S.M	S.M	S.M	S.M												

the 1st four bits are for signals, the next eight bits are for the modulant, the next 16 bits for subject matter, and the next 12 bits for interfix.

The signals

S_5 signifies a grouping signal to indicate the end of an item or compound.

S_7 indicates the end of a process group. S_6 signifies the end of patent or document.

The Questions

A question word is set to the required hexadecimal code or to "ignore" on the control panel. The search is conducted on an "exact match" or "inclusion" basis for each of the questions. The interfix and grouping signal relationships are expressed by plugboard wiring. Up to 12 questions can be searched simultaneously.

Operation

The above described system is being used on ILAS patent searching with respect to Class 260, subclass 94.9.

The following is a further description of the operation of the system applied to ILAS.

Assume the following disclosure:

Ethylene can be homopolymerized or copolymerized with propylene at a temperature of 50-100°C in a solvent of benzene at a pressure of 140 lbs. per sq. in. with catalyst $TiCl_2$ for 1 hour to form a polymer having a molecular weight of 100,000 and a density of .94 and an melting point of 112. The polymer is washed with a solvent of methanol and then chlorinated with Cl_2 and in CCl_4 to give a polymer. The chlorinated polymer can be used as a film.

Signal	*Modulant*	*Subject Matter*	*Interfix Field*					
			1	2	3	4	5	6
	homo	ethylene						
S_5			X					
	Co	ethylene						
S_5			X					
	Co	propylene						
S_5			X					
	Solvent	benzene						
S_5			X					
	Catalyst	Ti						
	Catalyst	Cl						
	Catalyst	reduced						
S_5			X					
	Solvent	polymerization						
		Start. material	X					
		end material		X				
	Temp.	50-100°C						
	Pressure p.s.i.	140						
	Time	1 hr.						
S_7			X	X				
		polymer						
	M.W.	100,000						
	M.Pt.	12						
	Density	.94						
S_5				X	X	X	X	X
		Washing						
		Start. material			X			
		End material				X		
S_7					X	X		
	Solvent	CH_3OH						

Signal	*Modulant*	*Subject Matter*	*Interfix Field* 1	2	3	4	5	6
S_5					X			
	Solvent	$C\ Cl_4$						
S_5					X			
		Cl_2						
S_5					X			
	Solvent	Chlorination						
		Start. Material				X		
		End. material					X	
S_7						X	X	
		Film						
S_5								
S_6	Patent No.							

In the above punched card the column under Signal discloses the--

Signals (a) S_5 for end of compound
(b) S_7 for end of process
S_6 for end of patent or document

In the modulant field is described--

Modulant

(1) function or use of material e.g. solvent, catalyst
(2) conditions, such as temperature, pressure, time
(3) Properties, molecular weight, density, melting point
(4) Use i.e. film
(5) Process as polymerization, washing

In the Interfix column

The following relationships are disclosed--

(a) ethylene and propylene are copolymerized
(b) benzene is a solvent
(c) $TiCl_2$ is a catalyst
a, b and c are starting materials in a polymerization process and give
(d) a polymer
(e) the conditions of this polymerization process are describ?d as to time, temperature and pressure
(f) This polymer product is now a starting material in a washing step and is washed by CH_3OH
(g) this material is then chlorinated to give
(h) a chlorinated polymer

The interfix shows not only the relationship between the various ingredients but to some extent, a reaction sequence.

CONCLUSION

An abbreviated description has been presented of some of the efforts and thinking with respect to developing an effective and practical machine searching system for the Patent Office. Various approaches are being developed and tested, with the hope and expectation that the ultimate goal will be arrived at by a series of successive short goals, each constituting an advance over the previously attained objective.

Another approach which offers a great deal of promise is a method, described elsewhere, which involves the features of (1) parallel access, (2) finding correlations at the time of search rather than by a prearranged correlation in the coded file, and (3) generating the code dictionary from the disclosures as presented in the documents.

BIBLIOGRAPHY

(1) Julius Frome and Jacob Leibowitz
"A Punched Card System for Searching Steroid Compounds" Patent Office Research and Development Report #7, Washington 25, D. C., Department of Commerce, 1957.

(2) J. Leibowitz, J. Frome and D. D. Andrews
"Variable Scope Search System" p. 291-316 Preprints of papers for the International Conference on Scientific Information, 1958

(3) Classification Bulletin of the United States Patent Office, Class 260--Chemistry, Carbon Compounds, Bulletin No. 200, Revision 1

(4) Don D. Andrews
"Interrelated Logic Accumulating Scanner (ILAS) Patent Office Research and Development Report #6, Department of Commerce, Washington, D. C.

43443--U.S.Dept.of Comm--DC--1958