電子計算機を用いた文献情報の蓄積，検索*

加　藤　　緑**　　伏　見　和　郎***

1. はしがき

論文，特許などの文献情報の量は近年急速に増大しつつあるので，これらに関する蓄積や検索の手段の開発は各方面で関心をもたれてきている．これに関してはハンドソート・パンチカードを用いる方法や IBM カードを用いる PCS 方式などが以前から考えられ現在も実用になっているが，最近電子計算機の発達にともなって，これを用いた情報処理方式の開発が盛になってきた．

わが国においてはすでに電気試験所[1,3]，特許庁などでこれについての研究開発が行なわれているが，電気通信研究所においても 1963 年以来，技術情報管理法の一つとして電子計算機による文献情報処理方式の開発を独自に検討してきた．

この報告はまず，われわれの開発した処理方式の基本的な考え方について述べ，つぎに蓄積検索の具体的なプログラムについて説明し，最後に今までに行なった検索の実施例について簡単に述べるものである．使用計算機は NEAC-2206 である．

2. 方式設計の基本的な考え方

論文，特許などの文献情報はその傾向や内容がたえず変化してゆくものである．たとえばレーザという言葉は数年前までは存在していなかったが，今では多くの論文がこれに関係している．このように新しい概念の発生やその比重の消長にともなってその取扱い方も異ってくる．一方情報の蓄積はかなり長期にわたることが必要である．したがって，全体としての体系は変えないまま，部分的には必要に応じて容易に変えうるような融通性のある方式が望ましい．

また機械化ということは，大量の情報を取扱うことを前提としており，電子計算機による処理がいくら早くとも，その準備段階で非常に人手がかかるようでは意味がない．このような考え方にもとづいて，具体的には次のような方法を用いることにした．

（1） 各文献について蓄積すべき内容はその文献の固有番号，分類項目，著者名，著者の所属機関名，論文標題，出典，所属機関の国別，論文の種別とし，さらに必要に応じて論文の内容を補足する言葉をつけ加える．またキーワード欄，備考欄，UDC 欄を別に設けて適当に使い分けができるようにする．これらの項目は記憶装置上で一定の番地に記憶されるように固定語長として取扱う．一文献当りの記憶語数は 6 字を 1 語として 100 語を割当てる．

（2） 項目分類は当研究所の研究分野に即した 43 の大まかな分類項目に従って行なう．そして各文献に対して三つまでの分類項目を重複して指定できるようにする．この分類項目の指定は内容のわかる技術者が行なうのがよい．また UDC による分類記号も必要に応じて用いる．

（3） 分類項目，所属機関，雑誌名などは蓄積の際判読できる程度に簡略化したコード* に書き直す．これによって，たとえば分類項目のように年月とともに変化するものについては修正や前のものとの併用を容易にすることができる．また論文の標題はなるべく原論文のままの形で蓄積する．このため長期にわたってその内容を利用することができるし，また蓄積の場合の労力を軽減することもできる．

（4） 検索方法としては記憶装置上の固定の番地に蓄積されるコード化された情報による検索と自然語による検索との両方を用いる．すなわち，著者，所属機関，機関の属する国名，雑誌名，発行年月日，論文の種別（オリジナル，展望，投書，訂正などの別），分類項目，UDC などはコード化された情報として検索し，一方標題中の技術用語や標題を補足する言葉などは自然語のままで検索する．この両者は場合に応じて適当に組合わせて最も能率の良い検索が行なえるようにする．たとえば，ある主題に属する論文を集めたいという場合，まずそれに関連のある分類項目でしぼ

---

* Information Storage and Retrieval by the Electronic Computer, by Midori Kato and Kazuo Fushimi (The Electrical Communication Laboratory)

** 電気通信研究所

*** 電気通信研究所（現在タケダ理研工業株式会社）

* “コード”というのは“登録された簡略語”の意味に用いる．これ以降も同様．

り，つぎにその主題を最も的確に表わしているいくつかの技術用語を用いて自然語による検索を行なうというような方法をとる．

（5） 情報蓄積の場合の電子計算機への入力には紙テープを用い，その内容を磁気テープ・ファイルの形で蓄積する．1巻の磁気テープ（長さ約 760 m）には約 11,500 件の文献を蓄積することができる．

3. プログラム

この方式による文献情報の蓄積，検索のプログラムとしては，次のようなものを用意する．

（1） 文献情報を磁気テープに蓄積するプログラム

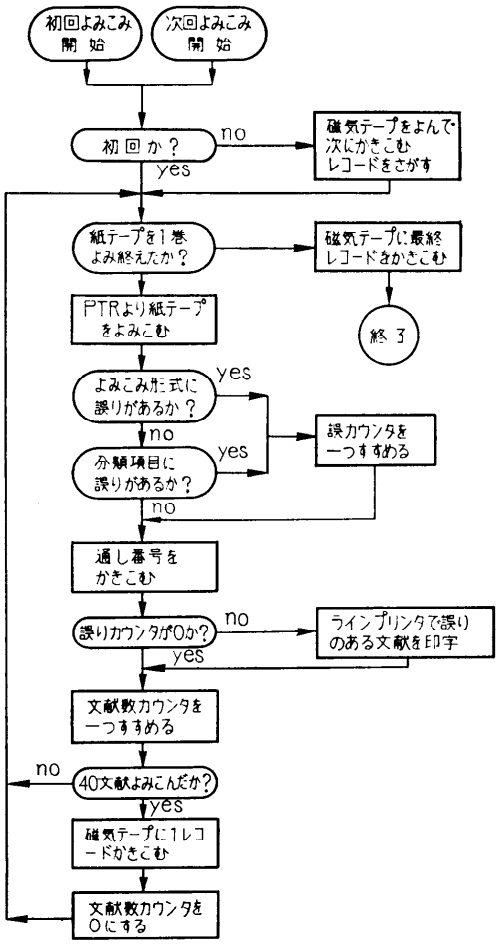
（2） コード化された情報の検索プログラム

（3） 自然語による検索のプログラム

この三つが基本となるものであるが，これらを適当に組合わせて文献のいろいろな統計的調査——たとえば機関別の論文数やその内容分布の分析，技術用語の集積など——のためのプログラムを作ることもできる．つぎに (1) および (3) の内容をフローチャートに従ってやや詳しく述べる．

3.1 文献情報を磁気テープに蓄積するプログラム

これはさん孔紙テープの内容を読みこんだ順に40文献ずつまとめて蓄積するものであるが，読みこみの際には各文献に自動的に通し番号をつけ，これをその文献の固有番号とする．また読みこみ形式の誤りや登録外の分類項目の混入などを検出し，誤りのあった文献はラインプリンタでそのつど打出す．紙テープから磁気テープへの情報転送は 200 ストローク/秒の読みこみ速度で1分間に約 40 文献であるから，1,000 件の文献は約 25 分で蓄積できる．この蓄積プログラムのフロー・チャートを**第1図**に示す．



第1図　文献情報を磁気テープに蓄積するプログラムのフローチャート

3.2 自然語による検索のプログラム

論文の標題中にあらわれる自然語をそのまま検索の対象とするものである．標題を補足する意図で入れるキーワード（情報に近づく手がかりとなる言葉）も同様に取扱うことができる．現在できているプログラムでは6個の検索語*を用いているが，論理和，論理積，否定を含む任意の論理式で表わされる論理的関係で検索が可能である．

検索を行なう場合には，まず問題に応じて検索語を選択し，その間の論理的関係を表わす論理式を設定する．検索語は 17 字以内の自然語を用いるが，第1から第6までの検索語にそれぞれ記号 A, B, ……, F を対応させ，論理式はこれらの記号を用いて表わす．たとえば情報検索に関する文献を検索したいとき，INFORMATION→A, RETRIEVAL→B, STORAGE→C, SEARCH→D, KEYWORD→E, THESAURUS→F のように対応させ P※(A·B)+C+D+E+F@ の形の論理式を用いる．ここで@はエンド・マークの意味である．

これらの検索語および論理式はあらかじめ紙テープにさん孔して検索プログラムとともに磁心記憶に格納しておく．プログラムが実行に移るとまず，各文献に

* 6個という数にはっきりした根拠はない．プログラム作成の容易さ（検索に用いるインデックス・レジスタの数）から自然にきまったようなものであるが，一応妥当な数と思っている．6個以上を用いる場合は2回に分けて検索する．

ついてそれがこれらの検索語を含むか否かを調べ，その結果をはじめにきめた論理式にあてはめて論理演算を行ない，条件に合致したものを選び出すという順序で進む．すなわち，このプログラムは次のような五つの部分に分れている．

1) 検索語を1字ずつに分解し，字数をカウントするプログラム
2) 検索語に合致するものがあるか否かを調べるプログラム
3) 論理式を逆ポーランド記号による表現に変換するプログラム
4) 2) および 3) の結果を用いて論理演算を行なうプログラム
5) 条件に適合した文献を印字するプログラム

論文名やキーワード中に検索語と合致するものがあるか否かを判定する方法はいろいろ考えられるが，著者らは木沢氏の情報検索機[1]にヒントを得て，検索語と検索すべき語との1字ずつを比較してゆく新しい方法を開発した．この方法を用いれば接頭語や接尾語の取扱いや複数の処理も比較的簡単になる．たとえば標題中に SUBMILLIMETER という語を含む論文を MILLIMETER という検索語によって検索することもできるし，ERROR-CORRECTING CODE という言葉を含む文献および ERROR-CORRECTION という言葉を含む文献を ERROR-CORRECT という一つの検索語で検索することも可能である．複数についても VISCOSITY および VISCOSITIES を VISCOSIT で検索するなどの方法をとることもできる．

このため検索に先立って，はじめに紙テープで読みこんでおいた検索語を1字ずつに分解して，おのおのの文字を記憶装置の1語に格納し，同時に字数もカウントしておく．検索の手順は，たとえば標題中の SUBMILLIMETER という語を MILLIMETER という検索語で検索する場合について述べると次のようになる．

まず検索語の第1文字 M を標題中の文字 S, U, B, ……と一致するものが見つかるまで順に1字ずつ比較してゆく．M で一致するものが見つかると検索語 MILLIMETER の字数カウンタを1ステップする．これによって次の段階では検索語の2番目の文字 I が SUBMILLIMETER の次の文字 I と比較されることになる．さらにこの字数カウンタの内容を前に数えておいた検索語の字数と比較する．いまの段階では一致しないからつぎに進む．この過程をくり返してゆくと R のところで字数カウンタが検索語 MILLIMETER の字数と一致するので，SUBMILLIMETER はこの検索語を含むと判断し，検索結果格納番地にその結果を書きこむ．検索結果格納番地としては論理式中の検索語 A, B, …に対応するものと，これらの否定 −A, −B, …… に対応するものとを用意しておく（これらの番地の内容（1または0）がのちに論理演算を行なう際に論理式の変数にそれぞれおきかえられる）．結果を書きこむと字数カウンタをクリアして論文標題の続きを調べ，論文標題やキーワードの文字を全部調べ終るまでこれを繰返えす．以上の過程は六つの検索語について平行して行なう．

つぎにこの結果を用いて論理演算を行なうのであるが，まず演算に先立って論理式を分解し易くするため逆ポーランド記号による表現に変換する[2]．これによるとたとえば，論理式 P※(A+B+C)･(D+E+F)@ は PAB+C+DE+F+･※@ のような括弧を除いた形に変換されるので，さきの検索結果を変数 A, B, C, D, E, F に代入すれば容易に演算を行なうことができる．そして演算の結果，その文献が条件に適合することがわかれば，そのつどそれを印字しつぎの文献の検索に移る．

以上自然語による検索プログラムの大要を述べたがこのプログラムの特徴は

1) 検索語の文字と論文標題中の文字とを1字ずつ比較しているため検索語の選び方に融通性がある．
2) 六つの検索語の論理和，論理積，否定を含む任意の論理的関係で検索できる．
3) 検索動作が六つの検索語に対し，平行して行なわれるため検索時間が節約できる．

などである．

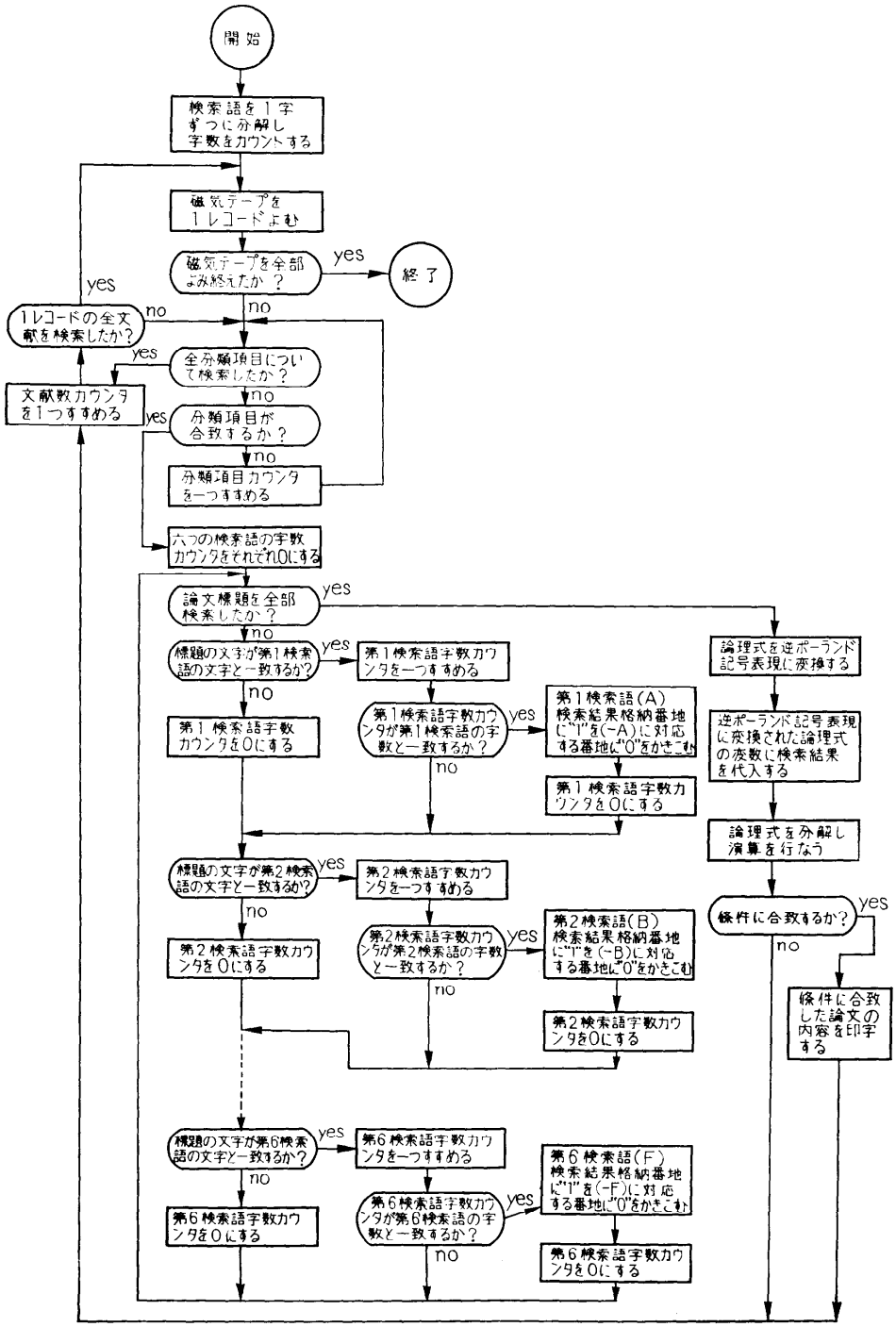
第2図に自然語による検索プログラムのフローチャートを示す．なおこの検索に要する時間は検索のとき指定した分類項目内の文献数によるが大体 100〜150 件/分である．

4. 実 施 例

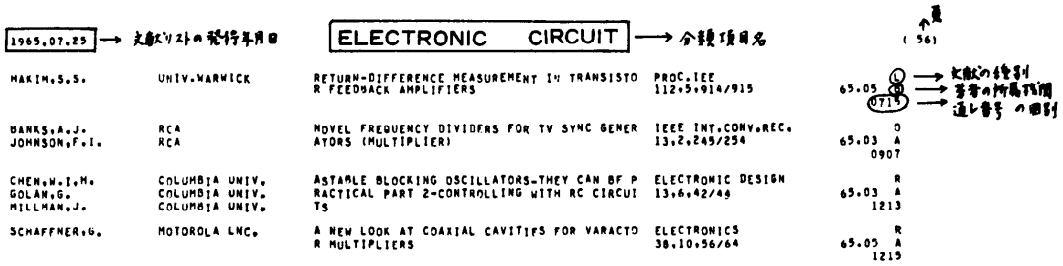
つぎに磁気テープ・ファイルを用いた検索の実施例について述べる．

4.1 海外文献リスト

第3図に示すような海外文献の分類項目別リストを 1964 年 2 月から毎月所内用に発行している．これは1カ月分の磁気テープ・ファイルを用いて分類項目別検索を行ない，所定の形式に印字したものであって，



第2図　自然語による検索プログラムのフローチャート



第3図 文献リストの形式

THE NUMBER AND PERCENTAGE OF DOCUMENTS CLASSIFIED BY SUBJECTS AND CORPORATE AUTHORS

	BTL		IBM		RCA		PHILCO		SPERRY RAND		UNIV.CALIF.		WESTINGHOUSE	
GENERAL	1	1	2	2	1	1			1	1				
THEORY-1	3	12	2	10	2	4	1	1	1	1	3	8	1	1
THEORY-2	7	19	3	6		3			1	1	5	8		
THEORY-3			1	3	4			2	2	1	2	3	4	1	1
COMPUTER-1	4	7	5	10		1					1	1		

第4図 機関別分類項目別文献数一覧表の一部分

コード化された情報による検索の一つの例である．このプログラムには各分類項目記載の頁を示す目次作成のプログラムも含まれている．プログラム実行に要する時間は 900 ライン/分のラインプリンタを用いて 1,000 件当り約 35 分である．

4.2 機関別分類項目別文献数一覧表

横軸に BTL, IBM, RCA などの機関名をとり，縦軸には分類項目名をとった表であって，機関別の合計および全文献数に対する比率，各分類項目別の合計等が計算して書きこまれる．プログラムの説明は省略するが，とり上げた機関数 70，検索の対象とした文献数 11,356 件の場合，検索時間は約 25 分で，このうち磁気テープを読んで機関別分類項目別に検索するのに約 20 分，合計や比率を計算し，作表するのに約5分かかった．この表の1部を**第4図**に示す．

4.3 自然語による検索

i） Ge, Si および化合物半導体に関する文献の動向調査

これは Germanium や Silicon, Gallium·Arsenide 等の言葉を標題中に含む文献を検索するのであるが，この場合言葉で表わされるものの他に Ge, Si, GaAs などの記号で入っているものもかなり多いので，両方の論理和で検索した．検索の対象とした文献数は 17,018 件で，そのうち半導体関係の分類項目に属するもの 1,680 件について自然語による検索を行なった．

検索時間は磁気テープまきもどしの分も含めて1物質当り平均約 20 分である．このような検索は人手ではかなり面倒であるので，計算機を用いるのに適した問題である．

ii） 導波管および論理数学に関する文献の検索実験

自然語による検索法はある主題に属する文献を集めたいという要求に対して有効である．

この場合得られた検索結果にどのくらい不要なものが含まれているか，あるいは必要なものがどのくらい脱落しているかがその有効さの尺度となるので，これについての簡単な実験を次のような手順で行なった．

1） 検索の対象として二つの母集団 Ⓐ, Ⓑ を考える．ここで Ⓐ は6カ月分の海外文献リストのアンテナ・導波管関係（あるいは論理数学関係）の分類項目に含まれる文献，ⒷはⒶ以前の4カ月分の同じ分類項目に含まれる文献である．

2） Ⓐに含まれる文献のなかから検索条件に適合するものを内容を読んで選び出し，その標題にでてくる

第1表　導波管に関する文献検索結果

検索語	Ⓐ (214 件)					Ⓑ (167 件)				
	r	l	e	p	q	r	l	e	p	q
W-1	97	(2) 13	50	(98.0) 86.6	(34.5) 37.4	65	(13) 25	35	(80.0) 61.5	(40.3) 46.7
W-2	97	(9) 26	20	(90.6) 73.2	(18.5) 22.0	65	(16) 29	20	(75.3) 55.4	(29.0) 35.8

Ⓐ: 最近の6カ月分の文献，　Ⓑ: Ⓐ以前の4カ月分の文献

W-1: waveguide, microwave, wave, power, band, circulator, high, frequency, ferrite, dielectric, junction, filter

W-2: waveguide, microwave, band, circulator, junction, filter, ferrite, dielectric

r: 与えられた主題に適合する文献数

l: 与えられた主題に適合するにもかかわらず検索もれになる文献数

e: 与えられた主題に適合しないにもかかわらず混入する文献数

p(呼出率)$=\frac{r-l}{r}$,　q(雑音率)$=\frac{e}{r-l+e}$

(括弧内はキーワードの補足によって検索されたものを含む場合の数)

第2表　論理数学に関する文献検索結果

検索語	Ⓐ (176 件)					Ⓑ (141 件)				
	r	l	e	p	q	r	l	e	p	q
L-1	38	(1) 6	35	(97.4) 84.2	(48.6) 52.3	29	(3) 5	31	(89.6) 82.8	(54.4) 56.4
L-2	38	(1) 7	7	(97.4) 81.5	(15.9) 18.5	29	(5) 7	3	(82.8) 75.9	(11.1) 12.0

Ⓐ: 最近の6カ月分の文献，　Ⓑ: Ⓐ以前の4カ月分の文献

L-1: function, threshold, logic, sequential, circuit, machine, method, network, state, element, using, minimiz

L-2: threshold, logic, sequential, state, Boolean, minimization, switching, asynchronous

r, l, e, p, q に関する定義は第1表の場合と同じ.

(括弧内はキーワードの補足によって検索されたものを含む場合の数)

技術語の出現頻度を調べる.

3) 2) で調べた出現頻度の高い順に12個の語を選び，それを検索語としてⒶについて検索を行なう.

4) 2) で調べた出現頻度の高いものからthreshold function の function, microwave power の power のような意味の広い語を除外して選んだ8個を検索語として再びⒶについて検索を行なう.

5) 3) および 4) の場合と同じ検索語を用いた実験を母集団Ⓑについて行なう.

実験結果は**第1表**および**第2表**に示すとおりである．この結果から次のようなことが考えられる.

1) 同数の検索語を用いた場合検索すべき主題および検索語の選定によって呼出率* にかなりの差がある．第1表は導波管に関する文献を集めるには 8～12 個の検索語では不足であることを示している.

2) function や power のような意味の広い言葉を除外すれば，呼出率にはあまり影響なしに雑音をかなり減らすことができる.

3) 標題が十分に内容を表わしていない論文についても呼出率を良くするには文献蓄積の際に適当な言葉を補足することが考えられる．母集団Ⓐは蓄積の際に組織的にキーワードの補足を行なったものであるが，呼出率を上げるのにかなり役立っている.

5. むすび

電子計算機を用いた文献情報処理方式について情報蓄積方法と自然語による検索法を中心に述べた.

当研究所においては，この方式による文献の蓄積を1964年2月以来継続しており，蓄積された磁気テープファイルは文献リスト（毎月発行）や機関別分類項目別文献数一覧表の作成，発行ですでに利用されている.

* 呼出率とは検索条件に適合する文献数を r，条件に適合するにもかかわらず検索もれになる文献数を l としたとき $(r-l)/r$ で表わされるものである.

また自然語による検索もこの1年間の実験期間中に32件（1件の検索語は1語～25語）程度の依頼をうけた．この結果にもとづいて1966年1月よりサービスを開始することになった．計算機操作は電子計算機室に一般の計算サービスと同じ方法で依頼している．

本文の説明からわかるように，本方式は1）多数の文献の処理に適すること，2）蓄積した文献ファイルからの最新の文献の項目別速報や文献に関する統計表の作成，所要文献の検索など多目的に利用できること，3）汎用の電子計算機を利用できること，などが特徴である．また4）自然語による検索を前提としているため，分類作業が比較的容易であること　なども電子計算機の機能を活用した方法といえよう．

この仕事は小口，伊藤，中村各情報特許部長および草間情報課長の御理解とはげましの下で行なわれた．また遠藤調査役，山宮調査員はじめ情報課内の各員にも随時有益な御意見をいただいた．さらに数値計算課の鈴木社員にはプログラム作成に御協力いただいた．これらの方々にここで深く謝意を表する．

参考文献

1）木沢　誠：情報検索，39年度電気四学会連合大会予稿 S. 5-3
2）計算機言語とプログラミング，電気通信学会誌 Vol. No. p. 103 (1965).
3）木沢　誠：情報の検索とその機械化，日科技連数学計画シンポジウム，報文シリーズ No. 11 情報数学

（昭和40年12月29日受付）