007：614.8

情報検索における出力と有効性の研究

— 防災情報の検索の諸方法に関する研究（その1）—

高　橋　　博

国立防災科学技術センター第2研究部地震防災研究室

A Study on Output and Validity in Information Retrieval

— Preliminary Report of Study on Methods of Information Retrieval for Disaster Prevention Research(I) —

By

Hiroshi Takahashi

National Research Center for Disaster Prevention, Tokyo

Abstract

As a researcher in the field of disaster prevention has to find out documents, which contain useful information for his study, from among a tremendous number of literatures, observational data and other kinds of materials, it is desirable for him that information retrieval by machine would be developed more and more. Validity of information retrieval is checked by recall factor (number of relevant documents in output/total number of relevant documents in question) and by pertinency factor (number of relevant documents in output/number of output). These factors are independent of each other, but they can be associated with each other by introducing the output index (number of output/total number of relevant documents in question), and the result of retrieval may be unitedly considered.

Output index = recall factor/pertinency factor.

When output index is very much smaller than unity, if output is all relevant, that is the best condition where pertinency factor equals unity, then the recall factor is, the same as the output index, very smaller than unity. In such a case it is needed to renew the instruction for retrieval to the machine and to enlarge the output remarkably. When output index is slightly smaller than unity, and if output is all relevant, then pertinency factor is unity and recall factor is nearly unity, and validity of retrieval is well. If output index is unity and pertinency factor is also unity. recall factor is unity and validity of retrieval is the best. When output index is a little over unity, if all of relevant documents in question are perfectly recalled, recall factor is unity, and pertinency factor is nearly unity, and validity of retrieval is well. When output index is far over unity, and

if all of relevant documents are perfectly recalled, that is the best condition of this case, then the pertinency factor, which is the reciprocal of output index, is very smaller than unity, and it is needed to repeat the retrieval of output.

Validity of subsequent retrieval is also restricted by the recall factor or pertinency factor and output index of the first or former retrieval. On a validity diagram, where the recall or pertinency factor is represented as ordinate, and the output index as abscissa, both in the logarithmic scale, the above-mentioned relations can be graphically expressed.

The total number of relevant documents in question in the file of documents is usually unknown. This number can be statistically estimated by the inspection of a selected part from the file documents. But this method often requires a great deal of work and time. By the application of Leslie's A-method which is for estimation of the total number of foxes in the mountain by using the number of repeatedly captured foxes, the author has tried to estimate the total number of relevant documents in question by using the number of relevant documents in each output. And examination was made on repetition of more retrievals in order to catch the omitted documents by renewing the instruction to be given to the machine, all the way doing the check of validity of each retrieval by using the total number of relevant documents estimated from each output.

1. 序　論

今日，防災に関連した諸業務が国の行政機関はもとより，地方自治体・各種企業ならびに大学・研究所などにおいて，ますます比重をまし，それに関連して防災に関係をもつ各種の資料が日夜ますます多量に生みだされている．一方，災害現象は，自然に関係したものの場合，そのくり返し周期に著しく長いものがあり，履歴性をもち，さまざまの環境的素因と多くの内部的要因の複雑な組合せに支配されているため，たがいに類似性を示しながらも個々の災害は複雑な様相を呈している．したがって，防災のためには，災害時およびその前後のできるだけ長い期間についての資料を幅広く集め，解析することが問題の根本的検討にはもちろん，応急的見解をたてる場合にも必要であるという考えが，次第に高まってきた．このような傾向はひとり防災分野に限らず，単におびただしいデータの処理というより，むしろ，既往の多種多様な資料を解析して信頼性の高い見解に到達する必要のある分野，たとえば経済・経営・外交・地学・化学・医学などの分野で最近ひとしく著しくなった傾向である．この最近の情勢に対する深い見とおしと資料の活用に関する今日的感覚から，当国立防災科学技術センターの設置法に，その主要業務の一つとして資料の収集・整理・解析および提供を行なうことが特に明記されたものである．ところで，収集された多種多様な資料を効果的に活用するためには，それぞれの人の知りたいことがらや調査目的にかなった有効な情報の検索方法が確立されなければならない．もし，有効な検索方法が確立されず，多くの人が知っているような古い型の図書館[注1)]のように沢山の資料がただ静かに保管されているだけならば，多数の資料も内外の利用者にとって，種類と量が多ければ多いほどかえって利用しにくい集積となってしまい，必要な資料を完全にえて自分に課せられた仕事を果

たしたいという利用者の期待に答えられず，その人たちをいつまでもというよりますます限られたその人の知識と個人的な情報源に依存するようにさせてしまう．一方，多種多量な情報源から質問者の真に欲している内容の情報を，もれなく正確にひきだす技術（information retrieval，以下ＩＲと略記する．[注2)]）は強力な電子データ処理システム（ＥＤＰＳ）技術の発達によって現実性をもつに至り，今日各方面で盛んに研究され，その実用化のために粘り強い努力が重ねられている．当センターの上述した課題も，強力なＥＤＰＳの設置とＩＲ技術の開発適用によってのみ解決できるものである．筆者は，数年来，ＥＤＰＳとドキュメンテーションの結合によるこの新しい分野について深い関心を払ってきたが，たまたま自然言語による技術文献の機械検索[注3)]に関する研究に接する機会を得た．その際，当センターにおける情報の機械検索の方法に関する研究の準備の一つとして，検索出力と有効性[注4)]について若干の検討を行なったのでここに報告し，今後に役立てたいと考える．

なお，この報告は前に口頭で発表したものを[2)]主として筆者が進展させたものであるが，用いた実験データは電気試験所電子計算機部佐々木久子技官が行なったもので，この報告に使用を許されたことに対し同技官に感謝します．また，統計的方法について助力を得た統計数理研究所赤池弘次氏に謝意を表し，本報告について検討を賜わった第２研究部長丸山文行氏に感謝します．

注１）

「従来，ともすると日本では，図書館の仕事といえば，日の当たらない日陰の仕事のように考えられてきた．」「ブライアント氏（ハーバード大学図書副館長）は，『大学図書館の図書は読むためにあるものであって，書庫のたなの中に保存しておくためにあるものではない．……大学中の図書は学内のすべての教授・学生・研究者に利用されねばならない．ことに学問の研究分野は日増しにその幅を広げている．今日では，図書の有機的な相互利用なしに十分な研究や教育をすることは不可能である．……

図書のダイナミックな利用をはかることが図書館の役割である．……そうした働きを強化するためには，……近代的技術も思いきってとり入れなければならない．しかし，それにもまして教授や学生に対する積極的なサービスこそ図書館人の使命である．』と強調している．」（岸本英夫東大図書館長：図書館の近代化について．朝日新聞，昭．38. 7. 5）

注２）

ＩＲの具体的内容は次のとおりである．

(1) 文献の材料やテキストを集めること．
(2) 集めた素材を分類したり呼出し語をつけること．
(3) それらをたくわえておくこと．
(4) 要求どおりの情報を取り出すこと．
(5) 取り出した情報をディスプレイすること．
(6) 取り出した情報のハード・コピーをとること．
(7) 情報を必要な人に配付すること．

ＩＲシステムの終局の型は『どんな情報をえたいかがわかっている人』のためにではなく，『その人の現在おかれている状態を察知して』その人のために最適の情報を提供することである．しかし，現状では一歩手前の『要求どおりの情報』を迅速確実に提供することにある．

このような仕事を行なうためには，人間の知能と電子的ならびに光学的装置を結合し，十分の強度と信頼度をもつ機械を作りあげ，この人間機械システムを完備することなしにはできない．（以上参考文献(3)参照）

注３）

情報を機械で検索する場合，各情報記事（注４参照）に情報内容を示す目印をつけておく．機械はその目印によって検索を行なう．目印のつけ方にさまざまあって，昔から有名なものにＵＤＣのコードがある．このような階層的な分類法は体系的ではあるが，もともと多面的性質をもっているものに１点の地位しか与えない矛盾のほかに，そのような欠点を除くと一応考えられる他の分類法でも同様であるが，それまでの知識をもとにして作りあげた人為的分類とは関係なく発展する部門が必ず生じるために，新しいものを合理的にその体系ちゅうに組みこめられない欠点をもつ．（たとえばＵＤＣでは電子計算機（681.142）は精密機械器具（681）として取り扱われ，電気通信工学（621.39）とは別の体系に属す．）また，今回の実験においても生じたが，分類体系が高度であり，精細であればあるほど，特定の限られた専門家以外に正確なコードをつけられない欠点もある．

また，分類体系によらず適当に見出し語をつける方法もあるが（分類法による場合も同様であるが），観点をかえた検索をしようとする場合には，全くお手あげになる場合が生じうる．（例：災害の型でつけられていたとき，富士山でおこる災害は？ときかれた場合．）さらに，情報内容の一部を摘出してあらわすため，情報内容の損失を生じる．特に創造的な仕事にたずさわっている研究者や技術者にとっては，情報の記述者や要目の摘出者が，一見して，情報内容，意味内容がとぼしくても，新しい観点から見なおすときには，非常に重大な情報価値をもっているものがある．このようなことから見出し語を詳細につけた極限として，情報全部を見出し語とみ，全文を検索する方法がある．これが自然（人間の）言語による情報検索で，検索の有効性も高かったという報告もある．（以上参考文献(4)参照）

また，今回の実験に用いた機械は，自然言語の文章をそのままの形で取り扱え，また，検索指令は必要条件を単的に示し，複雑なプログラムを要せず，単語の組合せで表現できるもので，電子計算機についての専門知識をもたない人でも使えるように特に考案された検索専門機[5]であった．当センターにおいても分野が著しく広く，また，質問内容が多様をきわめるであろうことと，電子データ処理システムに質問者の多くがなじんではいないこと，当センターのEDPSおよび資料部門の人員が飛躍的に増加することも考えにくいことなどから，本格的な研究や検索は，多目的の通常の電子データ処理システムによるべきことは疑いを入れないが，自然言語を用いた検索専用機の活用についても考えるべきであろう．

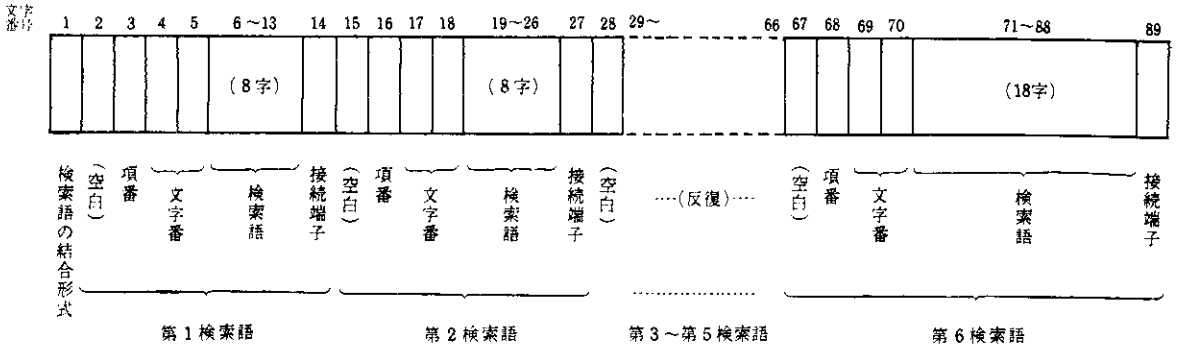
なお，ここに今回の実験を簡単に紹介する．

検索実験は，Proc. IREに掲載された1年分

始符号　¶通し番号　¶ UDC　標数
¶ 標題
¶ 著者名
¶ 雑誌名　¶ 巻数　¶ ページ　¶ 刊行年月
¶ 抄録（約 300字）終符号

実例　（ただし始と終の符号は略）
¶ 0042　¶ 681.142
¶ Pattern Detection and Recognition
¶ Unger, S. H.
¶ Proc. IRE　¶ 47　¶ 1737-1552　¶ 1959. 10.
¶ Both processes have been carried out on an IBM 704 computer which was programmed to simulate a spatial computer. The programs tested included the recognition process for reading hand-lettered sans-serif alphanumeric characters.

図－1　記事の形式



図－2　検索指令の形式

の抄録4455件を自然言語のまま磁気テープに転写されたものについて行なった．抄録1件を記事と呼び各記事は書誌的に著者名・雑誌名・標題・抄録等9項に分けられ，1記事550字以内に収められている（図－1）．検索指令（図－2）は，自然言語の論理的組合せ（and，or，not）で記事中の項をそれぞれ指定して行なう．今回は，標題（3項）と抄録（9項）それぞれについて，見出し語と相互参照（例：Absorption，see also Ionosphere；Resonance；Wave Propagation）によって別々に検索した．

注4）

検索結果が質問内容にどれだけ答えているかを調べるのを検索の有効性の検討という．以下に有効性の検討に関係ある用語の定義を行なう．

情報検索において，機械が検索する対象となる情報の蓄積体を資料ファイルと呼ぶ．そして資料ファイルを構成している細胞に当る単位の情報―文献のファイルであれば個々の文献 ― を記事と呼ぶ．記事は通常，さらに標題，著者名等図－1で見たように多くの項よりなる．検索に際しては，情報をえたい人の質問内容に応じて検索指令が作られ，みかけ上合致する記事が質問に対する「答」として出てくる．（この「答」を以下出力と呼ぶ．）使用した資料ファイルによってはもちろんのこと，資料ファイルが一定の場合（以下原則としてこの場合）でも検索指令の作り方いかんによって検索される記事 ―「答」― はかわる．こうして得た「答」のすべてが真の答であるとはかぎらず，また，出力が必要とする情報のすべてを含んでいるともかぎらない．ここに検索の有効性の問題が生じ，通常，次の四つの指数によって検討される．

I：必要情報記事数（真の答の数），

O：検索された記事数（「答」の数―出力），

F：出力中の適合記事数（出力中の真の答の数）

とすると，

$$\left.\begin{aligned} &\text{呼出率（recall factor）} = \frac{F}{I}, \\ &\text{除外率（omission factor）} = \frac{I-F}{I} = 1-\frac{F}{I}, \\ &\text{適合率（pertinency factor）} = \frac{F}{O}, \\ &\text{雑音率（noise factor）} = \frac{O-F}{O} = 1-\frac{F}{O}. \end{aligned}\right\} \quad (1)$$

資料ファイルがきまればそのときの質問内容とそのときの検索指令によって，I，F，Oはきまりそれらは互いに独立である．また，(1)の四つの指数のうち，互いに独立なものは二つであるので，以下必要ないかぎり呼出率と適合率によって検討する．

2. 出力指数について

検索結果が良好であるためには，Iがほぼ等しい方がよいともいうが，単純一様な性格をもつ情報源でもないかぎりそのようなファイルはありえず，大体Iの大きさそのものが質問内容のわずかな変化によっても変化するものであり，それにもともとIにあわせて情報源を作るわけにはいかないものなのであるから，情報検索においてIの分布範囲は広いものと考えておく必要がある．そこで，筆者ら（高橋・佐々木）は出力の測度としては，出力そのものより出力と必要情報記事数との比であるO/Iを用いる方がよいと考えたが，[2]ここにこれを出力指数と名付け，λであらわすことにする．なお，λに似たものとしてresolution factor$=O/S$（S：資料ファイルの記事総数）

がある．これは機械の「目」の「みえる」度合の良否を示す指標ではあるが，各Iがいずれも等しい場合以外には意味がない．

今，呼出率をr，適合率をpとするとλに対しrとpのそれぞれとりうる最大値は次のようになる．

$\lambda \leqslant 1$の場合　$F_{max} = O$，

$\lambda \geqslant 1$の場合　$F_{max} = I$

であるから

$$\left.\begin{aligned} r_{max} &= \frac{F}{I} = \frac{O}{I} = \lambda, \quad (\lambda \leqslant 1), \\ &= \frac{I}{I} = 1, \quad (\lambda \geqslant 1), \\ p_{max} &= \frac{F}{O} = \frac{O}{O} = 1, \quad (\lambda \leqslant 1), \\ &= \frac{I}{O} = 1/\lambda, \quad (\lambda \geqslant 1) \end{aligned}\right\} \quad (2)$$

となる．

したがって，もし$\lambda \ll 1$の場合は，Iの大きさに十分見合った大きさのOを得て，λを1に近づけなければ必要情報を完全に得ることはできない．反対にIに比べて非常に大きな出力を得ると$\lambda \gg 1$となり，よしんば$r = 1$となってもpが著しく低くなって出力を再検索しなければならないハメにおちいる．したがって$\lambda \fallingdotseq 1$となるような出力を得なければ有効性の高い検索（$r \fallingdotseq 1$，$p \fallingdotseq 1$）は行なえないことが定性的に知られる（表－1）．

次に，rとpは互いに独立であるが，λを介して互いに次の関係にある．

$$\left.\begin{aligned} r &= \frac{F}{I} = \frac{F/O}{I/O} = p \times \lambda, \\ r/\lambda &= p. \end{aligned}\right\} \quad (3)$$

表－1　出力指数と最大呼出率および最大適合率との関係

λ	r_{max}	p_{max}	検索の有効性
< 1	$= \lambda < 1$	1	非常にわるい．出力をうるよう検索方法の再検討が必要．
$\lesssim 1$	$= \lambda \lesssim 1$	1	おおむね良好．
1	1	1	良　好
$\gtrsim 1$	1	$= 1/\lambda \lesssim 1$	おおむね良好．
> 1	1	$= 1/\lambda < 1$	非常にわるい．出力の再検索が必要．

したがって，rとλを独立変数として取り扱うこともできる．従来，呼出率と適合率を統括的にはあくしたり，表示することに成功していないので，[注5)](3)の関係で，すでにみたようにrとpの値はλによって限界が与えられるという条件をも加えて検索の有効性を一つに表示する図を筆者らは作った．[2)]（例：図－3　以下これを有効性図とよぶ．）この図でr_{max}とp_{max}をそれぞれ太い実線と破線で示し（以下L_r，L_pとそれぞれを表わす．），rを中黒，pを中白の記号で表わすと，rとpの存在可能域はそれぞれL_rとL_pより下方となる．[注6)]

注5）

たとえば呼出率と適合率にウェートを与えて一次結合を作り，両者を統一的にはあくしてみようという試み[6)]などがあるが，合理的な方法はいまだ得られていない．

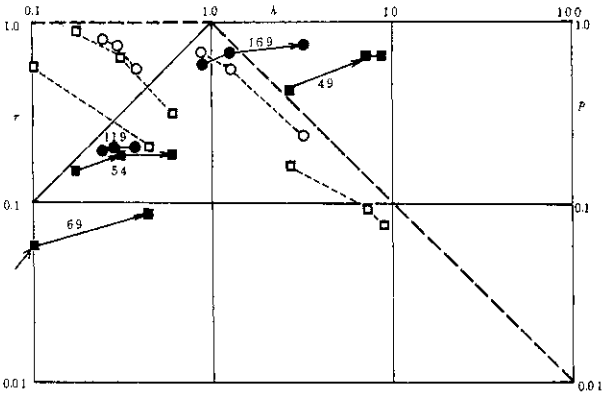
注6）

対数目盛の3角ダイヤグラムによって，図上の1点で$r = p \times \lambda$の関係をあらわすことができるが，そのようなダイヤグラムは一般に市販されていないので，両対数目盛の図を用いて表示することにした．

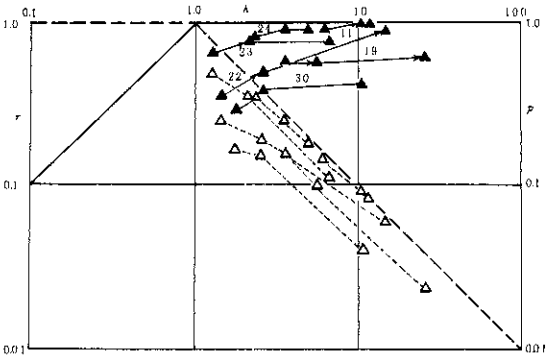
今回の実験の出力を

検索語／項	見出語	相互参照	
標　題	$O_1^{③}$	$O_2^{③}$	$O_1 \equiv O_1^{③} \cup O_1^{②}$
抄　録	$O_1^{⑨}$	$O_2^{⑨}$	$O_2 \equiv O_2^{③} \cup O_2^{⑨}$
			$O^* \equiv O_1 \cup O_2$

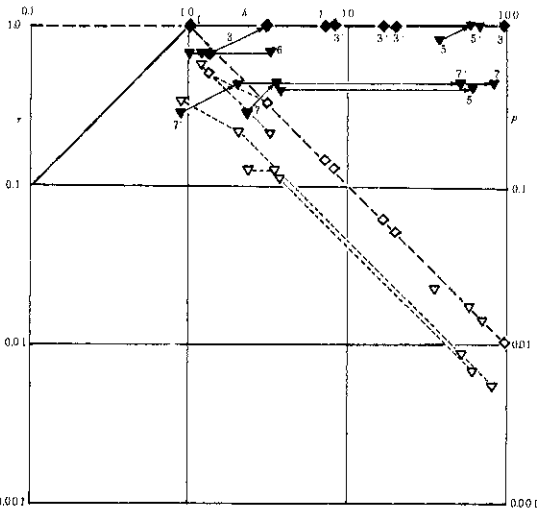
のように表わし，検索もれを拾うために新しい検索語や分野（項）をえらんで検索範囲の拡大を$O_1^{③} \rightarrow O_1 \rightarrow O^*$のように行なった場合の有効性の変化を図－3に示した．図－3において検索の有効性が全体的によく表現されている．たとえば，$I = 54$の場合，$O_1^{③}$では$\lambda = .17$であるから$r = .15$と低いのは当然である．しかし，L_rに非常に近い値をえているから適合率は$p = .89$と高い．次の検索で$\lambda \rightarrow 1$となるような出力$O_1^{⑨}$をえ，そのほとんどが適合するような検索指令を作れば，きわめて有効性の高い検索が行なえるであろうこ



図－３－１　$I > 32$　（記号は円：$I \geqslant 100$，四角：$100 > I > 32$）



図－３－２　$32 > I \geqslant 10$　（記号は三角）



図－３－３　$10 > I \geqslant 1$（記号は逆三角：$10 > I \geqslant 4$，ひし形：$4 > I \geqslant 1$）

図－３　検索範囲の拡大による有効性の変化

注）　両対数目盛．記号は呼出率：中黒，適合率：中白

とが読める．（実際はそうならなかったが．）ま　た，$I = 7$ の場合，$O_1^{⑧} \rightarrow O_1$ にかけてrとrは

共に1.5倍化したので，pは同じ値を保ったが，O^*においてh＝80にもなり，一方rは増加しなかったため，適合率が$p=.0054$と著しく低いものとなった．

以上のように，hを介してrとpを検討すると，検索語や，検索範囲の拡大過程における有効性の意味や変化が読みとれ，次の検索の指針がえられる。

3. 有効性の変化

検索範囲の拡大過程における検索の有効性の変化を有効性図（以下略して図という）について検討する．

(2)から

$h \leqslant 1$では　$L_r = h$,　$L_p = 1$,

$h \geqslant 1$では　$L_r = 1$,　$L_p = 1/h$

であり，図は両対数目盛であるから

$$\overline{L_r - L_p} = \log L_r - \log L_p = \log h. \qquad (4)$$

（ただし，$\overline{L_r - L_p}$は図上の線分の長さ，以下同じ．）

また，(3)から

$$\log r = \log p + \log h. \qquad (5)$$

ゆえに(4)と(5)から

$$\log L_r - \log r = \log L_p - \log p,$$

$$\overline{L_r - r} = \overline{L_p - p}. \qquad (6)$$

ゆえに，検索結果をhとrから図上にプロットすれば，pの位置は(6)からあたかもrの影のごとく従属的に定まる．

また図から

$h < 1$では，つねに　$r < p$,

$= 1$では，つねに　$r = p$,

> 1では，つねに　$r > p$

であることが容易に知られる．

今i回検索を重ねた結果，それまでに得られた出力をO_i,その適合率をF_iとし，さらに$i+1$回目の検索が行なわれた際の出力をO_{i+1}，その適合数をF_{i+1}，そのうち真に新しくえられたものをそれぞれO_{i+1}^+，F_{i+1}^+とすれば，

$O_i \cup O_{i+1} \equiv O_{i+1}$,

$O_i \cup O_{i+1}^+ \equiv O_{i+1}$,

$O_i \cap O_{i+1}^+ \equiv \phi$．ただし$\phi$は空集合．

$F_i \cup F_{i+1} \equiv F_{i+1}$,

$F_i \cup F_{i+1}^+ \equiv F_{i+1}$,

$F_i \cap F_{i+1}^+ \equiv \phi$．

次に，有効性の図上におけるおもな移動形態を検討する．

$O_i \to O_{i+1}$の過程で

ⅰ）出力が増しても適合記事数がふえない場合

すなわち，

$O_{i+1} = mO_i$,　（$m > 1$）

$F_{i+1} \subseteq F_i$

の場合

$h_{i+1} = mh_i$,

$r_{i+1} = r_i$,

$p_{i+1} = \frac{1}{m} p_i$

となる．

$\therefore$　$\log h_{i+1} - \log h_i = \log m$,

$\log p_{i+1} - \log p_i = -\log m$.

したがって，図上でrはhの軸に平行に移動し，pはL_p（$h>1$）に平行に（右45°下方）下がる．

ⅱ）出力と適合記事数の増加率が等しい場合

すなわち，

$O_{i+1} = mO_i$,　（$m > 1$）

$F_{i+1} = mF_i$

ならば

$h_{i+1} = mh_i$,

$r_{i+1} = mr_i$,

$p_{i+1} = p_i$.

$\therefore$　$\log h_{i+1} - \log h_i = \log m$,

$\log r_{i+1} - \log r_i = \log m$.

したがって，図でpはh軸に平行に移動し，rはL_r（$h<1$）の線に平行に上昇する．（右45°上方）

ⅲ）最大値・最小値

$O_{i+1} = O_i + O_{i+1}^+$

$F_{i+1} = F_i + F_{i+1}^+$

において

$O_{i+1}^+ \not< 0$,

$F_{i+1}^+ \not< 0$

であるから，必ず

$r_{i+1} \geqslant r_i$,

$h_{i+1} \geqslant h_i$

となる．

また，

$\phi \subseteq F_{i+1}^+ \subseteq O_{i+1}^+$

であるから

$$\left.\begin{aligned} r_{i+1\ \max} &= (F_i + O_{i+1}^+)/I = r_i + O_{i+1}^+/I \\ \text{or} \quad &= (O_{i+1} - N_i)/I = h_{i+1} - N_i/I \end{aligned}\right\} \qquad (7)$$

ただし，$N_i \equiv O_i \cap F_i^c$　（雑音）

ゆえにもしも$N_i \equiv \phi$　ならば

$$r_{i+1\ max} = h_{i+1} \equiv L_r$$

ただし，この場合 $N_i \equiv \phi$ は $h_{i+1} \leqslant 1$ の領域でしかありえない．また，

$$r_{i+1\ min} = (F_i + 0) / I = r_i \qquad (8)$$

で(i)に相当する．

なお，(7)において

$$F_i + O^+_{i+1} \equiv I$$

となった場合，

$$h_{i+1} = N_i / I + 1 \geqslant 1$$

となり，これは（$i+1$）回目にもっとも効果的な検索を行ないえた場合に相当するが，$N_i \neq \phi$ でないかぎり，$h_{i+1} > 1$ であることを示している．

次に適合率においては

$$p_{i+1} = (F_i + O^+_{i+1}) / (O_i + O^+_{i+1}).$$

その最大値は，$F^+_{i+1} \equiv O^+_{i+1}$ の場合で

$$\left.\begin{aligned} p_{i+1\ max} &= (F_i + O^+_{i+1}) / (O_i + O^+_{i+1}) \\ \text{or} \quad &= 1 - N_i / O_{i+1}. \end{aligned}\right\} (9)$$

もし，$N_i \equiv \phi$ ならば $p_{i+1} = 1$ となる．ただし，この場合 $O_{i+1} \leqslant I$ でなければ $N_i \equiv 0$ はありえないから，$p_{i+1} = 1$ は，$h_{i+1} \leqslant 1$ の領域でなければありえない．

また，(9)において $F_i + O^+_{i+1} \equiv I$ ならば

$$p_{i+1\ max} = 1/h_{i+1} \equiv L_p = I/(N_i + I)$$

で，これは $h_{i+1} \geqslant 1$ においてもっとも効率的に検索もれをなくしえた場合に相当し，また $N_i \equiv \phi$ でないかぎり $p_{i+1\ max} < 1$ となる．

最小の場合は $F^+_{i+1} \equiv \phi$ の場合で

$$p_{i+1\ min} = F_i / m O_i = p_i / m \qquad (10)$$

で(i)の場合に相当する．

したがって $h_{i+1} \gg 1$ or m が大きくなれば，$N_i / O_{i+1} \to 1$，$1/m \to 0$ にそれぞれ近づくから，$p_{i+1\ max}$ でも $p_{i+1\ min}$ でも，共にいくらでも0に近づきうる．この点，呼出率が $r_{i+1} \geqslant r_i$ で1に近づくのと対照的である．

r と p の存在可能域についてはすでに述べたが，(iii)の検討から実際の存在可能域は，最初の検索結果によってまず限定され，そのあと検索の回を重ねるごとによりせまく限定されてゆくことが明らかとなった．そこで，(7)～(10)によってえられる r と p の上限線を，それぞれ uL_r，uL_p，下限線を ℓL_r，ℓL_p とすると変域は次のようになる．

i) $N_i \equiv \phi$ の場合（$h_i \leqslant 1$ においてのみ存在）

$$uL_r \equiv L_r, \quad uL_p \equiv L_p$$

となる．（図－4－2）

i)′ $h_1 \geqslant 1$ （$\therefore N_1 \neq \phi$），$r_1 = 1$ の場合

なんらかの原因で検索をさらに加えたとすると，$r_i = 1$，p_i は L_p の線上を移動する．

ii) $h_1 \ll 1$，$r_1 \ll 1$ の場合

図－4－1からわかるように，h が10倍化するあたりで uL_r は L_r，uL_p は1に近づき，$h \fallingdotseq 1$（1よりわずか多い値）で $uL_r = 1$ になり，$uL_p \fallingdotseq 1$（1よりわずか低い）となる．

iii) $h_1 \lesssim 1$，$r_1 \ll 1$ の場合

図－4－3にみるごとく，uL_r と uL_p は $h = 1$ で大きく交わる．uL_r の $r = 1$ になる点は1.5前後で，p の最高値もやや低下している．しかし，この付近で変化するものは，全体としてみても検索の有効性はもっとも高いといえよう．

iv) $h_i = 1$，$r_i < 1$ の場合（図－4－4）

$r_1 \lesssim 1$ の場合，$r_1 = p_1$ から uL_r は比較的急激に，しかし uL_p はあまり向上せずに最高値に達する．そのとき $h \lesssim 2$ で，$p_{max} \gtrsim 0.5$ である．

$r_i \ll 1$ の場合，パターンはv)になる．

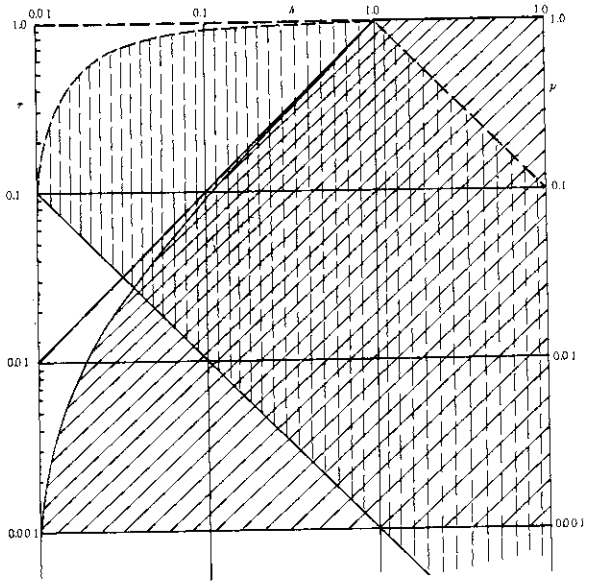
v) $h_1 > 1$，$r_1 < 1$ の場合

uL_r，uL_p ともに r_1 h_1 にほとんど関係ないくらいにほぼ垂直的に上昇する．（図－4－5）しかしながら p の達しうる最高値は低く，たとえば $h_1 = 5$ でさえ0.2である．

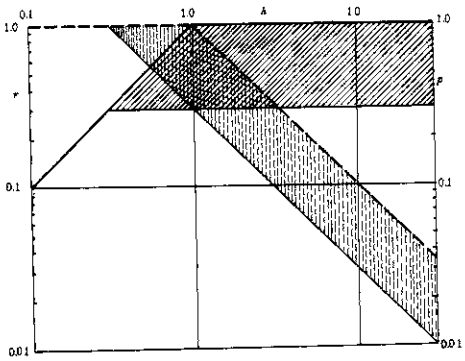
この点は特に注意すべきで，$h < 1$ の領域では $r \ll L_r$ でなければ p の悪いことはないが，$h > 1$ の領域では常に $p < r$ であるため，ただ r の増加をはかるくらいではだめで，p を急激に上昇させることをはからなければ検索の有効性は高くならない．そうして p を上げてももっとも効率的な場合でさえ，たかだか上述のごとき程度にしか向上しない．したがって，検索の有効性の観点からみれば，h が1をこさないように極力努力する必要がある．

4. 必要情報記事数の推定

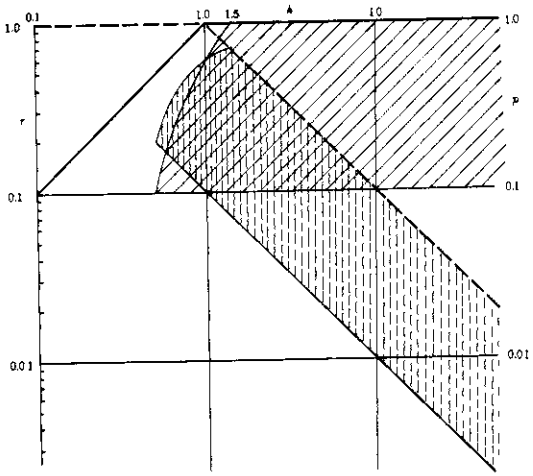
以上のように検索結果の有効性を検討するためには，求めている必要情報の大きさ ― 記事総数 I がわかっている必要がある．すべての記事が単純明確な主題内容をもっており，その内容を完全にあらわす"目印"がつけられている場合を除けば，質問内容のわずかな違いによっても I は変わる．そこで通常 I はあらかじめわかっていないも



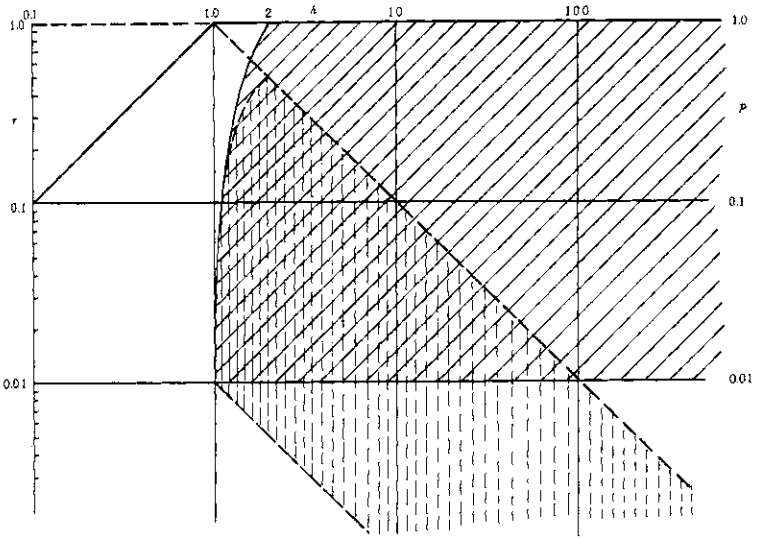
図－4－1　$\lambda_1 \ll 1$，$r_1 \ll 1$　の場合



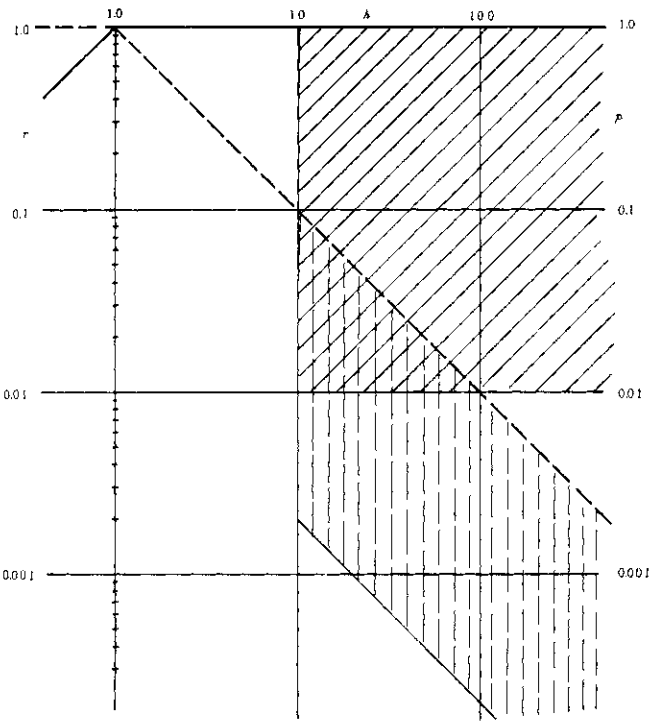
図－4－2　$\lambda_1 < 1$，$r_1 = 1$　の場合



図－4－3　$\lambda_1 \lesssim 1$，$r_1 \ll 1$　の場合



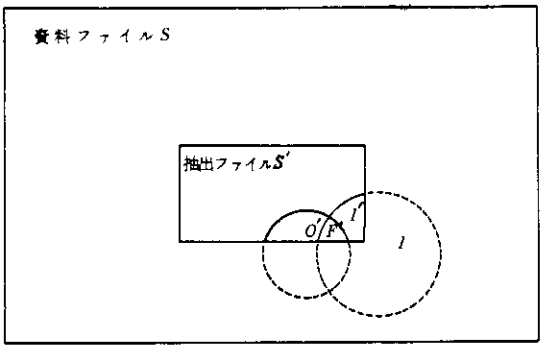
図－4－4　$h_1 = 1$，$r_1 < 1$　の場合



図－4－5　$h_1 > 1$，$r_1 < 1$　の場合

図－4　有効性の変化可能域（両対数目盛）

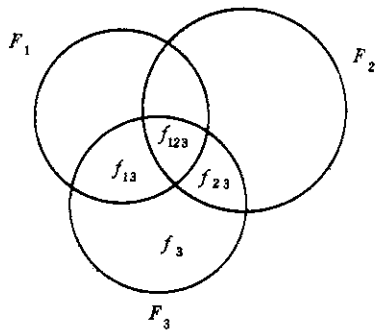
のと考え，Iの大きさを推定する方法について検討した．Iを推定するには，本来は資料ファイルから抽出した小ファイルについて検索し，その結果からもとの資料ファイルのIを推定すべきである．（図－5）ただこの方法は相当手数を要する．



図－5　抽出ファイルによる必要記事数の推定

一般に検索に際して検索落ちがないことを何人も願うが，検索結果の利用目的によっては絶対に検索落ちをゼロならしめなければならない場合と，必ずしもそこまでの完全性を必要としない場合とがある．ゆえに，Iの大きさの推定も厳密さを強く要求される場合と，それほどでもない場合とでは推定のために支払いうる手間にも相当なひらきがある．そこでここではIの推定のために特に調査は行なわず，検索結果だけから推定する方法について述べる．種々のモデルによる方法が考えられるが，ここではマーキング法[注7)]を適用した結果について記す．

適合数について次のように記号を定義する．（図－6）



図－6　マーキング法による適合数の記号の定義（$i=3$の場合）

$$F_1 \equiv f_1,$$
$$F_2 = f_{12} + f_2,$$
$$f_{12} \equiv F_1 \cap F_2,$$
$$f_2 \equiv F_2^{+}.$$
$$F_3 = f_{123} + f_{13} + f_{23} + f_3,$$
$$f_{123} \equiv F_1 \cap F_2 \cap F_3,$$
$$f_{13} \equiv \{F_1 \cap F_3\} \cap f_{123}^{c},$$
$$f_{23} \equiv \{F_2 \cap F_3\} \cap f_{123}^{c},$$
$$f_3 \equiv F_3^{+}.$$

（F_4 以上同様．）

各回の検索が互いに独立に行なわれ，適合記事はランダムに抽出されているとみなせるならば，2回以上の検索結果があればそれらから必要記事数Iを推定することができる．いくつかの方法[7)]があるが，今回はLeslieのA法[8)]によった．

注7）

マーキング法とは山にいるキツネの数を推定するために考案された方法で，とらえたキツネにマークをつけて山に返し，次回捕えた際マークのあるものを数え，前回とは別のマークをつけてまた山に返す ── こういうことを何回かくり返して山にいるキツネの総数を知る．A法のほかにB法やJackson法などあり，途中でキツネに増殖死亡のある場合，その増減率をも推定できる．

マーキング法を情報検索に適用する場合，次のような問題があると思われる．キツネの場合は，マークをつけて山にかえしたキツネが，山の中のキツネの中で十分かくはんされ，これをランダムサンプリングすることを前提としている．一方，情報検索の場合は，必要情報以外の多数のノイズを含んだ資料ファイル全体からのサンプリングである．したがって，山で，キツネだけでなく，ウサギもバッタもヘビも含めてサンプリングし，その中のキツネだけをマークによって調べ，その全数を推定しようということになる．したがって，母集団の構造と標本抽出の仕方に相違がある．ただし，情報検索の場合でも，資料ファイルの中からまったくでたらめに抽出するのではなく，求める情報の集合に可能なかぎり焦点をあてて検索をする．すなわち，山のキツネの全数を知りたいのであるから，キツネのみをとらえるように山で作業し，ウサギやヘビ・バッタの類は無視するし，山にいるそのようなものの存在とはおかまいなしに調査する．情報検索の場合も，ノイズはキツネをとらえるときに道でゆきあったヘビや，間違っ

て網にかかったウサギとまったく同じように取り扱い，それらを無視ないし除去する．母集団および標本抽出の仕方の相違を上のように考え，その置き換えができるものとしても，情報検索の場合，マーキングした情報が次の検索の場合，Iの中からランダム・サンプリングされるという保証はない．検索に際し，Iに対し毎回異なった角度から見るように検索指令を作った場合には，独立性が成り立ちうるのではないかと考える．検索の際に，概念的（意味上）には独立性が一応成り立っているようであっても，母集団からの標本抽出として，それが直ちにランダム・サンプリングになるとはいえないこと，それから，実際にはそう検索のたびごとに観点を全く新しく次々にかえられるものかどうかということ，それにも困難があり，あるいはそういうことができたとしても，それで検索結果が本当によくなるかという本質的な目的上の問題がある．

しかし，ここでは実用の立場から，理論的厳密性には欠けるが，試みてその結果を見てみることにした．

次に，マーキング法を適用しても実際上さしつかえがないとした場合にも，方法上次のような問題がある．

今，山にいるキツネの全数をN，そのうちマーキングされたキツネの数をmとすると，この方法はN匹の中からn匹がランダムに抽出された場合，その抽出のされ方は$\binom{N}{n}$通りあるが，その際n匹の中にマークされたものが何匹いるかという確率分布から出発している．$m/N=p$が常に定数であるような抽出がなされていることを前提としている．したがって，$N \gg m$でない場合この前提が成り立たなくなる．ところが，このようなものは超幾何分布を示すが二項分布に接近するので，マーキング法は後者におきかえて式をたて，さらに標本の大きさが大きくない場合のかたよりを少なくするための修正をほどこしてある．今，$N=100$，$m=40$，$n=5$の場合に，5匹の中にi匹マークされたキツネのいる確率をp_iとすると次のようになる．

	超幾何分布	二項分布
p_0	.0725	.0778
p_1	.2591	.2592
p_2	.3545	.3456
p_3	.2323	.2304
p_4	.0728	.0768
p_5	.0087	.0102
$\sum_i p_i$	.9999	1.0000

また，推定されるNの大きさ$E\{\check{N}\}$は，超幾何分布の場合，94.644，二項分布の場合，95.335となり，Nとmがこの程度なら二つの方法とも大差はない．

次に，分散の推定には著しい差が生じる．すなわち，超幾何分布の場合は，上の例で2238，二項分布では2357，マーキング法の修正式によると，nの中のマークされたキツネ数iより，下のような幅広い値をとり，平均をとれば超幾何分布で2914，二項分布で3031となり比較的近い値をとる．

i	$V(\check{N})$
0	24,000
1	3,200
2	800
3	240
4	64
5	0

したがって，分散については，n中のiにより大きくかわるから，精度を高めるためにはnに含まれるiのゆれを少なくする必要がある．そのためには，m/Nの大きさに見あった十分に大きいnを必要とする．

以上から，よい推定値をうるためには$N \gg m$，$n \gg 1$が必要条件となる．

今，$k_{i,e}$をi回目にはじめて検索され，問題としているe回目にも検索された適合記事の数とする．

例　$k_{14}=f_{1234}+f_{124}+f_{134}+f_{14}$．

検索が全部でE回行なわれたとすると，i回目の検索をもとに推定されるIの推定値$\check{I}_i$とその分散$u(I_i)$は次式で求められる．

$$\left.\begin{aligned}\check{I}_i &= F_i\left(\sum_{e=i+1}^{E} F_e+1\right)\Big/\left(\sum_{e=i+1}^{E} k_{i,e}+1\right),\\ u(I_i) &= \check{I}_i\left\{\frac{\sum_{e=i+1}^{E} f_{i,e}}{\left(\sum_{e=i+1}^{E} F_e+1\right)\left(\sum_{e=i+1}^{E} k_{i,e}+2\right)}\right\}^{\frac{1}{2}}\end{aligned}\right\} \quad (11)$$

ただし，

$$\Sigma \ell_{i\varepsilon} = \Sigma F_\varepsilon - \Sigma k_{i\varepsilon}$$

そこで，$O_1^{(8)}$，$O_1^{(9)}$，$O_2^{(8)}$，$O_2^{(9)}$ の順に $i=1,2,3,4$ ととり，各回とも適合記事をえた5例につき I とその信頼区間の推定を試みた．（図－8）

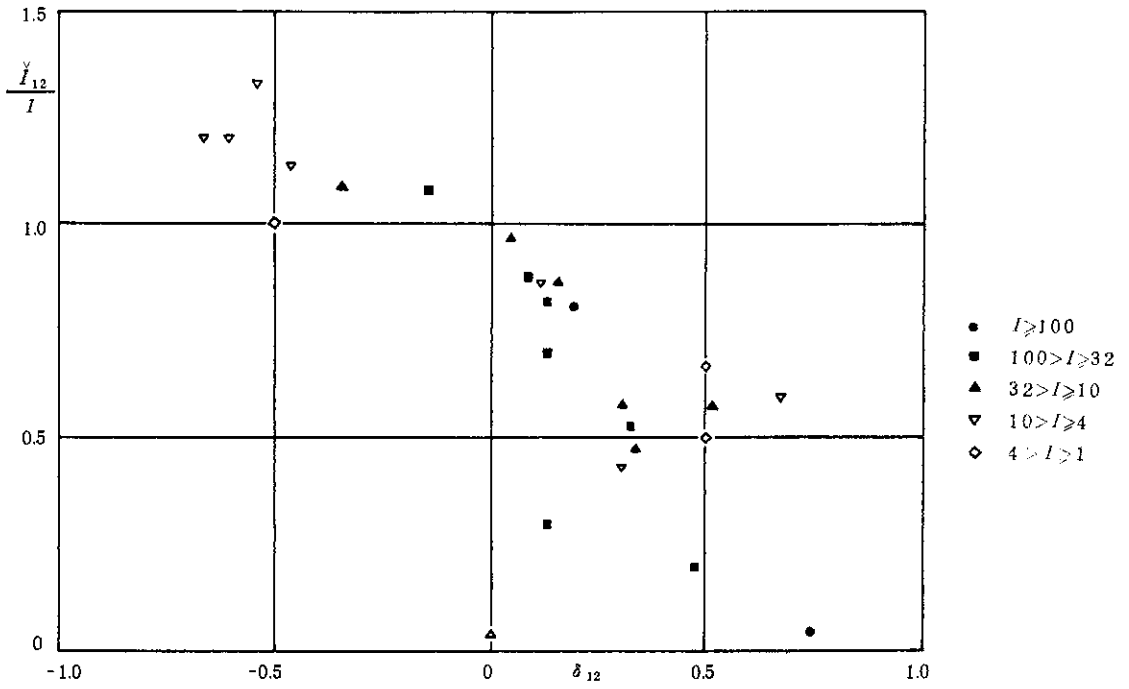
この方法は各検索が互いに独立であることを前提としているが，実際には今回の試行の各回の検索（標題と抄録，見出し語と相互参照）は互いに独立とはいえない．その影響をみるため $O_1^{(8)}$ と $O_1^{(9)}$ について相関係数（δ_{12}）を求め，これと $\check{I}/I$ との関係をみた．（図－7）その結果 $\delta_{12}>.2$ では推定値が著しく低くなっている．

$$\delta_{12}=\frac{\frac{f_{12}}{I}-\frac{F_1F_2}{I^2}}{\sqrt{\frac{F_1}{I}-\left(\frac{F_1}{I}\right)^2}\sqrt{\frac{F_2}{I}-\left(\frac{F_2}{I}\right)^2}}$$

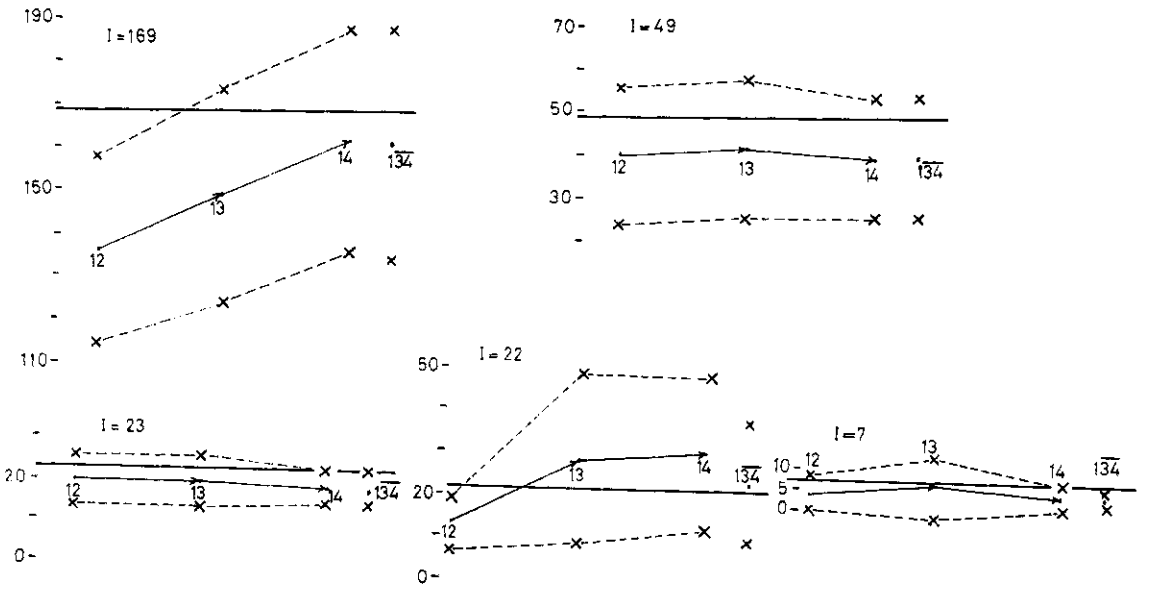
次に図－8で，12, 13, 14 と記したものは，$i=1$ すなわち最初の適合記事数をもととして，$E=2,3,4$，すなわち検索範囲を順次拡大していった場合，推定した $\check{I}$ の変化を示したもので，また $\overline{34}$ としてあるものは $i=3$ で O_2 の適合数をとった場合を示す．図－8をみると $I=22$ を除けばすべて低目に推定しているが，これは各回の検索が互いに完全に独立でないためであろう．推定値 $\check{I}$ は正確には正規分布をとらないと思われるが，目安として信頼限界を正規分布の場合の95％（$\check{I}\pm 2u(\check{I})$）をとったが，結果は I はおおむねその範囲にはいり，信頼区間の幅も検索を重ねるごとにせばまり，精度の向上がみられた．なお，$u/\check{I}$ は $I=169$ で0.08，$I=7$ で0.41と I の大きさに逆比例して精度が劣っている．（図－9）なお，一般に14と $1.\overline{34}$ の値はほとんど同じかあるいは後者の方が精度がよい（例　$I=22$, 7）．これらは相互参照による出力が通常非常に小さく，$O_2^{(8)}$，$O_2^{(9)}$ を区別して取り扱っても精度の向上に役立たなかったためと考えられる．

上述の5例につき，上記のほかに考えられる種々の組合せ（例　$1\rightarrow 23\rightarrow 24$，$\overline{12}\rightarrow\overline{12}\cdot\overline{34}$ 等々）について同じ方法により I の推定を行なったが，$1\rightarrow 13\rightarrow 1\cdot\overline{24}$ すなわち，標題について最初見出し語，次に相互参照，最後に抄録について見出し語と相互参照（$O_1^{(8)}\rightarrow O_1^{(8)}\cup O_2^{(9)}\rightarrow O^*$）という検索の場合のみ上記の場合と同程度の推定精度をえたが，他の場合はいずれも著しく悪い推定値をえている．

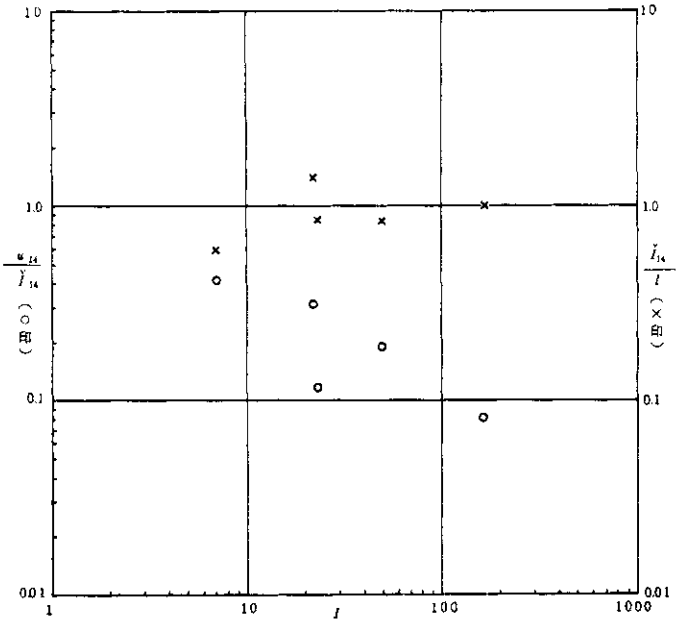
今回の試行において，相互参照によってえられ



図－7　必要記事数の推定に与える検索の独立性の影響
（$\check{I}$：2回の検索（$F_1^{(8)}$ と $F_1^{(9)}$）による I の推定値，δ_{12}：2回の検索の相関係数）



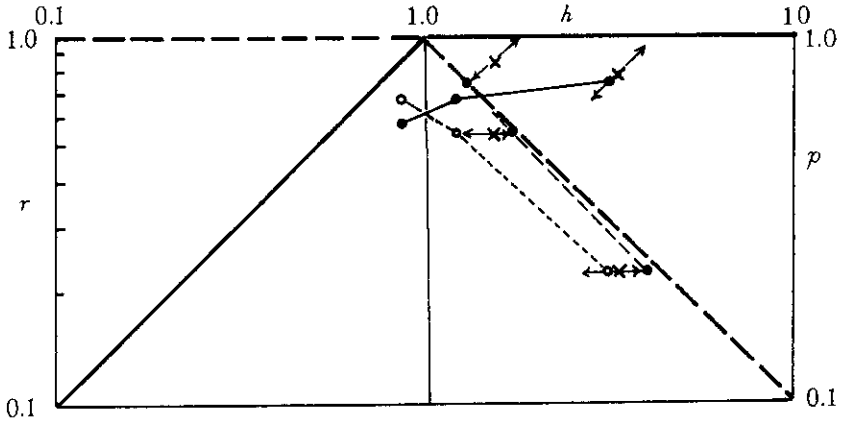
図－8　必要記事数の推定値と信頼限界



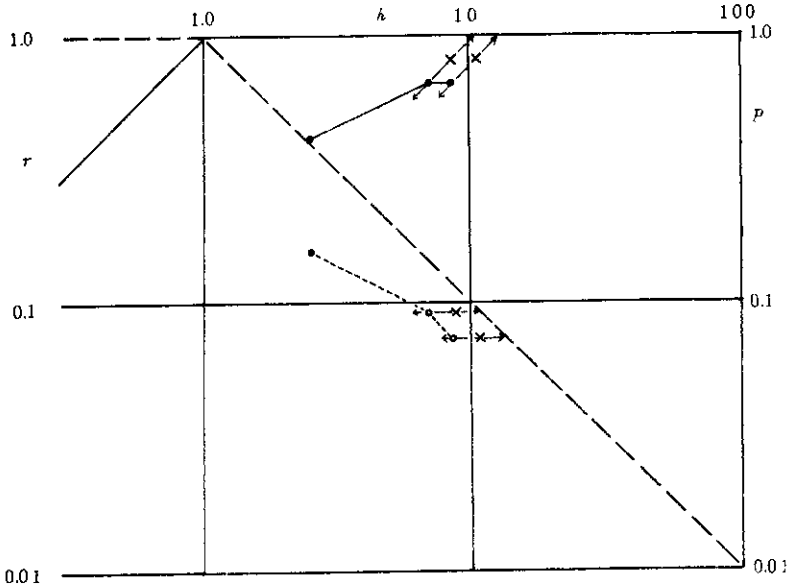
図－9　必要記事数とその推定精度（両対数目盛）

た適合記事数が著しく少ないという問題があったが，Iの推定においても，もっともオーソドックスな検索範囲の拡大方式 — 見出し語によって標題，次に抄録，最後に相互参照によって標題と抄録 — による場合のみよい推定値をえていることは，理由は明らかでないが注目すべきことと思う．

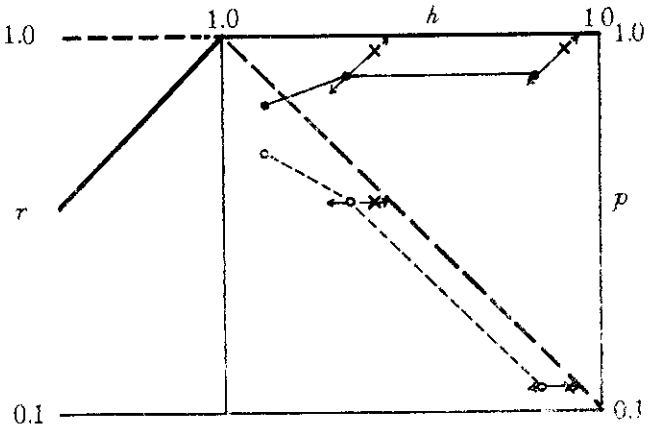
ところで，マーキング法の適用の前提となった各回の検索の独立性と，検索および推定結果についてここで検討してみる．一般に最初の検索は情報を引き出すにもっとも適していると思われる検索方式（たいてい正攻法的）をとる．そしてその



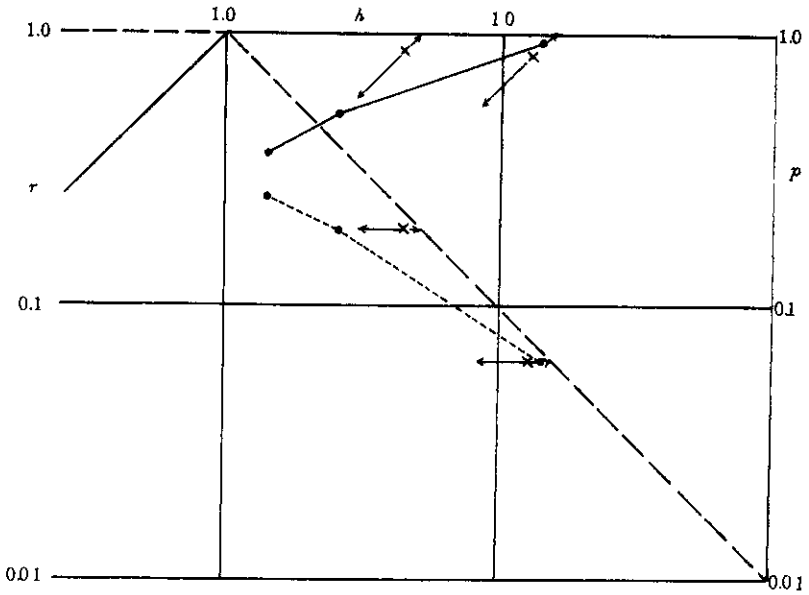
図－10－1　例1（$I=169$）



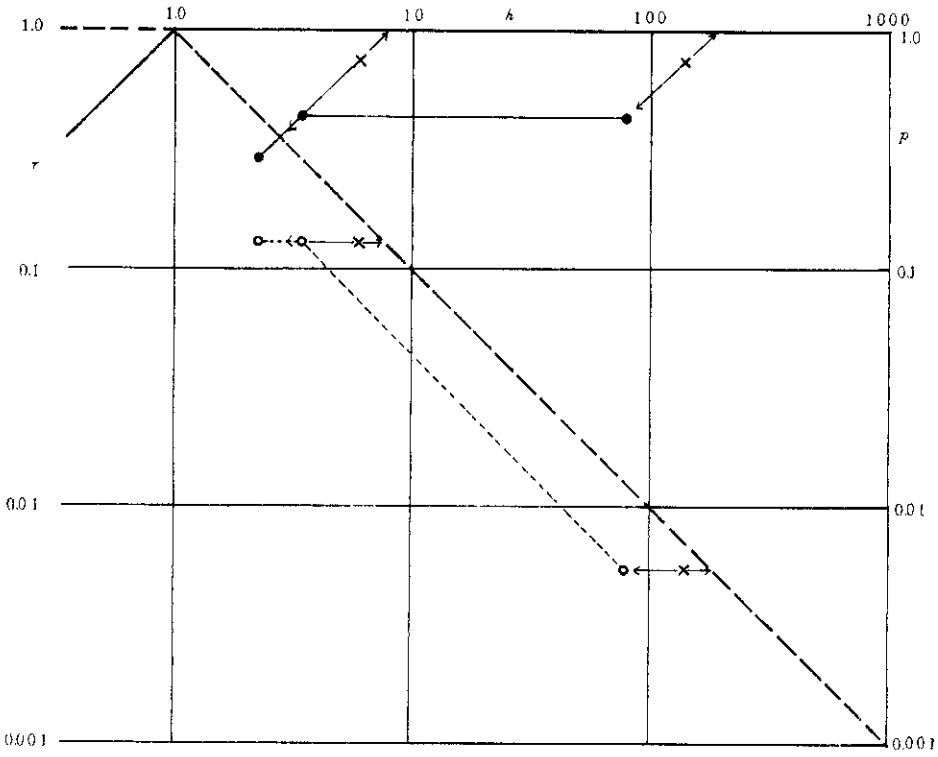
図－10－2　例2（$I=49$）



図－10－3　例3（$I=23$）



図－10－4　例4（ $I=22$ ）



図－10－5　例5（ $I=7$ ）

図－10　必要記事数を推定しつつ行なった検索の有効性の信頼限界（両対数目盛）

呼出し結果に不満な場合，新しい出力をうるため，検索語とその論理関係や検索分野を新しく選ぶ．その際，大別すると同じ範囲中で拡大をはかる（平板的）場合と，異なった観点からの（多側面的）場合に分かれる．

前者は検索方法（考え方）に根本的な問題はなく，適当な修正を行なって検索の有効性を高めようとする場合で，たとえば $I=169$ で（amplifiers，相互参照 circuits or valves）や22（acoustics，相互参照 sound，ultrasonics，diffraction，absorption）などで，概念的に同類か上下の包含的関係にある語で行なわれる．そして検索を重ねるごとに着実に出力をまして有効性をまし，I の推定もより接近する場合（$I=169$）や，しばしば質問に近い内容のものではあるが，不要な記事を多数えて，呼出率をましえても適合率は下がり，I の推定も F_i に応じて変化する場合（$I=22$）などあるが，一般に出力は必ずえられ，（$F_{i+1}=0$ の場合もあるが，例 $I=30$ atmosphere，相互参照 ionosphere or troposphere）概して確実な方法で，I の推定からみても互いに独立性があるとみなしても大体よさそうである．ところで後者の場合は，検索の互いの独立性は保たれ，前回の検索方法ではえられない出力をえられる可能性がある．しかし，最初最適と思われた観点を変えるのであるから，時によると出力をほとんどえられない場合（例 $I=119$ aerials，相互参照 lens）やまったく無関係な記事を多量に検出する危険（例 $I=7$ converters，相互参照 frequency or power supply）が少なくなく，I の推定も変動が多いかのようである．

以上の検討の結果から，検索結果より I を推定し，検索の有効性を検討しつつ検索範囲を拡大してゆく場合を考える．上記の5例の $O_1^{(3)} \rightarrow O_1 \rightarrow O^*$（$1 \rightarrow 12, \rightarrow 1\cdot\overline{34}$）の場合につき，$h$ と r の推定値とその信頼限界を算出し，その時の p とともに有効性図にプロットすると（図－10），5%程度の危険率の推定値は矢印の線分で示される範囲となる．ただし，$I \supseteq F$ であるから $\check{I}_i - 2u < F_i$ の場合は，F_i の値をもって $\check{I}_i$ の下の限界値とした．各検索の間に，通常は正の相関がある程度存在すると考えられるので，$\check{I}$ は少なめとなり，r と h の値は実際より大きめとなるから，呼出率 r の値は信頼限界の下限値付近にあるとみる方が今回の試行からはよいように思われる．

ゆえに有効性の高い検索を行なうためには，I の推定精度が高い必要がある．そのためには，ひとつには I が小さいと推定精度が落ちるから，I があまり少なくないと思われる資料ファイルをはじめから選ぶか，作るか，また，検索結果から I が小さいと思われたら資料ファイルから考えなおす必要がある．この点今回の試行をみても $I<10$ が全体の $2/3$ を占めている．（図－11）防災資料の場合は幅の広さと，多様性のため I の小さいものが多い傾向はもっと著しくあらわれる可能性が大きい．もうひとつには I を推定しつつ検索をすすめてゆく場合には，各検索の互いの独立性が必要で，検索回数も幾回も行なえる方がよい．したがって資料ファイル中の各記事に項数がごく少数ではなく，種々の検索語が選べ，種々の論理関係で選べるようである方がよい．たとえば，UDCのほかに種々の分類，見出し語，記述等を十分多くもりこんでおく必要があろう．

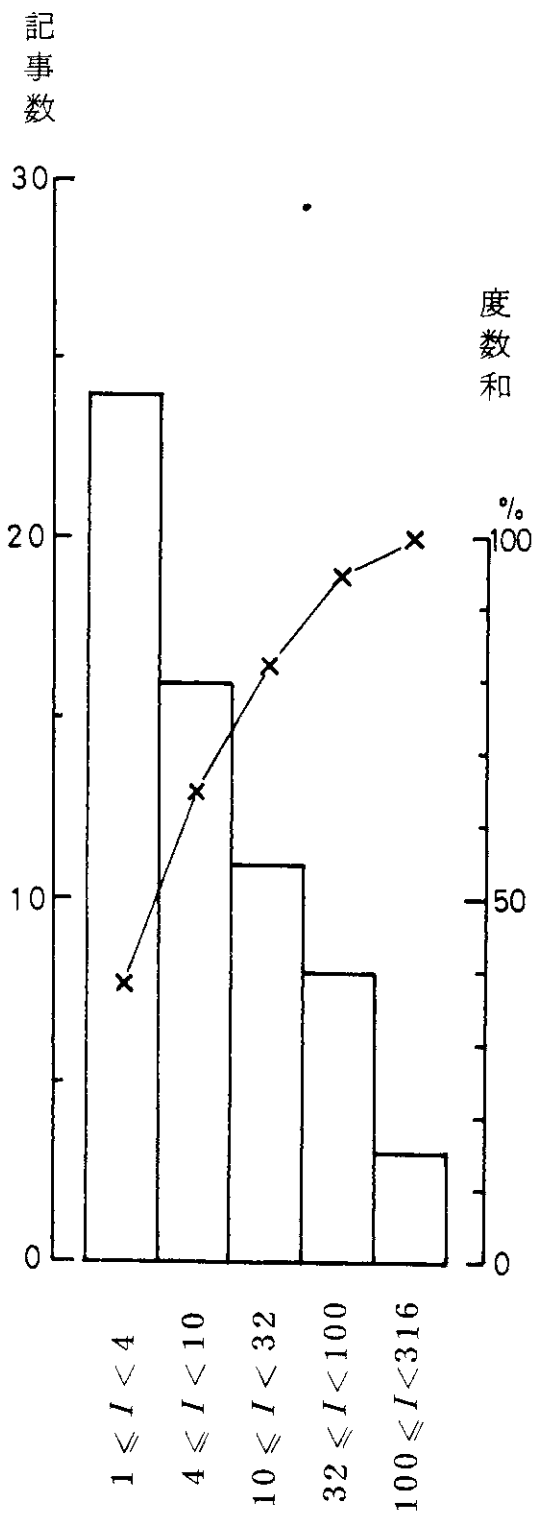
この点防災分野は領域が広く，内容を表示する方法を多数選ぶには一見ことかかないようであるが，起こりうる質問に非常に幅があるから，効果的な表示を多面的にいくつもつけることはたやすいことではない．ただし，検索回数が多くなると，すでにみたように $h \gg 1$ となり，多量の雑音に悩まされることになりがちであるから，各回の出力が大きくならないよう押え目にする必要がある．

5. あとがき

これまで検討してきたことから結論的にいえることは次のとおりである．

(1) 検索の有効性の検討は呼出率と適合率に出力指数を加えることにより，互いに独立な二つの指標を統一的にはあくできる．縦軸に呼出率，横軸に出力指数をそれぞれ対数目盛にとると，適合率は従属的に表示され，図上で呼出率と適合率の変化が効果的に追跡でき，両指数のとりうる可能域も読みとれ，有効性の高い検索を行なう上で指針をうるに役立つ．

(2) 通常，求めている記事の総数はわからないものであるが，出力からマーキング法により推定できる．ただし，一般に各検索は正相関をもつことが多いようで，そのため推定値は低目になり，その結果出力指数と呼出率の推定値は大き目になる．



図－11　必要記事数の度数分布

(3)　検索を行なう場合，まず求めている記事の総数が少なすぎない資料ファイルを用いること，また記事はなるべく項が多く，種々な側面からの検索が行なえるようコード，見出し語，記述文などが多く，かつ多種類であることが望ましい．求めている記事の総数を推定しつつ検索を幾度か重ねてゆく場合，出力指数が1を大きくこえないよう努力すること，各検索が確率論的に互いに独立であるよう努力することが必要である．

機械による情報検索を行なえるようにするためには，理論的にも技術的にも解決しなければならないことは沢山ある．しかも解決策は実際的でなければならない．自然言語を用いて文章をそのまま検索しない場合，コーディングや見出し語をつけるためのばく大な作業が生じる．また，一次資料そのままで検索するのもEDPSがいかに強力とはいえ相当に大変であるから二次資料によって能率的な検索をはかろうとすると，二次資料を作るためのばく大な作業がいる．したがって，どうしても機械による自動索引，自動抄録の実用化をはかる必要がある．ところで，自動抄録化の場合にせよ，文の検索（意味的に高度な）を行なう場合にせよ文の構造の問題がある．人の書く文章は，どんな論理的に書かれたものでも，論理学的に完全な記述にはなっていないし，また文の構造はいかなる言語でも，決して固定した完全な構造をもってはいない．このへんにも大きな問題がある．もっとも低い段階のシソーラスにも問題がある．防災のように広い分野では，同じ言葉が作られる分野によって皆異なった意味内容をもっている場合が少なくない．たとえば，土質でいうロームと地質学でいうそれとは異質である．であるから使用状況に応じた弾力性のあるシソーラスを作る必要がある．ひところ，シソーラスを精力的に作ることによって，検索の向上がかなり期待できるように思われたが，新しい用語がどんどん生まれ，使用限界がかわったりなど，雑音が多くでたり，概念の包含関係のゆれや用語としての使用度数など根本的問題のほかにも問題が生じてシソーラスにたよりすぎてはいけないこともわかってきた．したがって，情報検索を行なうには，これらIR技術の基礎的問題についての研究と実用化の努力を，他の機関とともに相当払ってゆかなければならないものである．

さらに，防災分野そのものの検索にふさわしい

ファイルやコードは，何かということをよく調べ研究する必要がある．そうしなければ利用者の本当の用には答えられないであろう．たとえば，コードについては，本当に効果的な見出し語を多数そろえるのにどうしたらよいか？また，防災に関する資料には必ずその資料に対応する地域があるから，それを地理学的名称ででも，測地学的座標系ででも，どちらでも検索可能なコード系を開発してつけなければならないであろう．ところで，防災における資料としては図書，雑誌などのいわゆる文献のほかに，通常学術文献のはんちゅうには属さない種々の調査資料を加える必要があることは大方考えられている．だが，この他に本や冊子のようなていさいはととのえてない，ただのデータや各種の図などが不可欠であろう．なぜなら，地球に関する研究や調査には，論文の記述や式よりも，データそのものを必要とする場合が少なくなく，データの不足に一番の問題がある場合が少なくないからである．データや図面と関連して，たとえばボーリング・コアーのような実物は二度と入手できないか，地理的，時間的，経費的に容易にはえられないのが普通であるから，データや図面の意味を正しく知り，あるいはそのものについてより詳しい情報をうるため，標本の有無とその保管場所を知りたいことも従来はまず困難であきらめられていたことである．ゆえに，標本についての情報も収集する必要があり，その逸散消滅に対しても本当は配慮すべきであると考える．[注8)]

抄録は，人が読んで内容を判断するには日本文では300～400字ぐらいが適当とされているが，[9)] 報知的抄録は検索用二次資料ファイルの記述として，その程度でよいか否かの検討はまだこれからである．一次資料や準一次資料[注9)]で検索ファイルを作ることは，（後者は別としても）今のところ得策とも思えないが，二次ファイル[注10)]としては抄録を伴った通常の図書カード式の内容のものでよいのかという問題，二次資料ファイルも報知的，書誌的なもののほかに，一次資料から抽出されたものも含め，主要な対象別のデータや図面についての二次資料ファイルなども作る方が活用上も検索上もよいのではないか？あるいは，地域によって検索できるような索引的な二次資料ファイルも必要ではないか？　それら種々のファイル全体をみるための全索引的なこと（三次ファイル）は何でいかにさせるか等，資料ファイルにも問題が沢山あると考える．

注8）

たとえば，松代地震センターの資料収集について，ボーリングのコアーも集めるべきことを，杉山隆二教授（信州大学理学部長）が主張し，実際に関係者の賛同をえている．また，第四紀の研究と実務にたずさわっている多くの研究者および技術者も，現在多額の経費を用いてえられている，ボーリング・コアーのほとんどが，工事完了とともに，すてられていることが，特に軟弱地盤の研究と技術の進歩と防災や施工の実務に大きな損失を与えているので，これらを一元的に収集，管理し，整理解析する機関のできることを強く望んでいる．その有効性の実例として，関東大震災のあと，帝都復興院の行なった，横浜から浦和にかけての地盤調査の際のボーリング・コアー（トラック3台分ぐらい）が，地質調査所において，たまたま，保管されていたため，他の資料とあわせて，新幹線の前身，弾丸列車の路線の調査計画の際に，非常に役立ったことなどが知られている．

注9）

一次情報とは，本来人類にとって未知の知識（観測，観察，実験，調査，理論等）を伝えるもので，総説的なもの（review）は準一次情報とよぶ．[10)] 一次情報の存在を知らせるものを二次情報，多数ある二次の情報の存在を知らせるものを三次情報とよぶ．一次資料ファイルは一次情報の蓄積をいうが，一次情報をあらわした論文は上述の定義からは本来一つであるが，場合によってはその内容を関係するむきに応じて修飾した論文として幾つかあらわされることがある．利用者にとっては，そのような論文も利用価値が高いので，ここでは上述のような絶対的一次情報のみでなく，それと同質とみなされる情報をもふくめて一次の情報資料とみなす．また，二次，三次情報をもふくめてここでは二次情報としてあつかう．一次情報からデータ（または図）のみを抽出して作った資料ファイルは，データそのものは一次情報で，必要な場合それから一次情報を知れるような使いかたができる点では二次情報ともなる．このような準一次情報の資料ファイルも複雑さをさけるために二次の資料ファイルとしてここでは一括しておく．

機械による情報検索を軌道にのせるためにはハードウェアの面にも問題があり，今日直面してい

るのは入出力機器の弱体である．特に入力の作成には，現状では多大の人力がいる．最近商品化されつつある光学的文字読取機（OCR）は，ローマ字でさえ手がき文字を読めないことはもちろんのこと，活字により印字されたものでも使用条件が極度に限定されていて，通常の文献の入力作成には役立たない．OCRが，漢字と手がき文字は別として，表音文字なら読めるようになったとしても，一次または二次情報を磁気テープにしまいこんであったのでは，記事内容を一見したい時や，出力のコピーをとるのに必ずしも容易でなく，また経済的でもない．そこで，マイクロフィルム（またはマイクロフイッシュ）を入出力メディヤにできないかと考えられ，この考えの1部をなすと思われる機器もあるようであるが，完全な経済的な装置の出現はまだ先のことである．マイクロフィルムは多量の（活字）情報を高密度に収納するために開発され，発展してきたもので，人がそれを見，複写をとるシステムも一応便利に開発されている．しかし，多量の情報を処理する場合にはなお欠点がある．すなわち，特殊なものを除けば検索に直接使えないほか，ひんぱんなそう査には情報保持の点から寿命がやや短いこと，適合ないしは必要なコマを自動的にえらびだすシステムの二，三の試みはあるが，機器の面でも，方式の面でも実用化していない．防災に関係ある分野には多種多様な図が沢山あり，その情報内容はしばしば非常に高い．したがって，図類の入出力装置を特に考える必要があり，そのうち高度なものは“図形処理”や“図形認識”に関した装置となろう．現在当所のTOSBAC 3400には多チャネルの高速A－D変換器と2チャネルのD－A出力装置が直結しており，データ処理に偉力を発揮している．しかし，カーブ・リーダやデジタル・プロッタ，アナログ・ディスプレイ装置などがない．これらを至急ととのえることが，図形資料の処理検索に進むうえでぜひ必要だと考える．なお，これからは後に手のかかるデータ処理，解析を必要とするような測定，調査，観測機器の場合，その出力を最近の地震探鉱装置のように，EDPSにより直接処理できる出力をえられるようにすべきであろう．

最後に専用機か汎用機かの論議は，汎用機の性能の著しい向上により結着がついた観がある．この問題は使用目的の限定性，使用の激しさ，利用者の程度などと経済性などの具体的条件によって選ばれるもので，しかも大型EDPSが存在する場合はそのオフライン機のような形で考えられるものであろう．[11] その場合，汎用機とまったく独立のシステムになってしまわないようにならねばならない．“図形処理や認識”に関して，部分的にアナログで処理やシミュレートする利点があるとすれば，その装置と汎用機との連結や交信の仕方なども新しい課題となろう．

今後，これら根本的問題を研究しつつ，防災にかかわりある情報検索のシステムを一つ一つ作りあげ，また作りかえてゆくことがダイナミックで，利用者に本当のサービスの行なえる資料センターを作るみちであると考える．

参　考　文　献

(1) 佐々木久子（1965）：自然言語による機械検索の有効性について．電気試験所彙報，29，863－878．

(2) 高橋博・佐々木久子（1965）：自然言語による情報検索の研究．第2回ドキュメンテーション研究会講演集，275－281．

(3) ORA研究会報告書　インフォメーション・リトリーバルに関する研究 ― アパーチャー・カード・システム ―（1964）．日本機械会工業連合会，p. 127（39）

(4) 木沢誠（1961）：言語で表わした情報の機械検索をめぐって．情報処理，5，277－282．

(5) 木沢ら（1961）：磁気テープを用いた情報検索機．電気通信学会誌，44，210－217．

(6) 安倍浩二・富永勲（1965）：金属工学文献の機械検索実験報告V．検索効率におよぼす索引深さの影響の調査．第2回ドキュメンテーション研究会講演集，261－263．

(7) 伊藤嘉昭（1953）：動物生態学入門 ― 個体群生態学編 ―．古今書院（東京）．（p. 202～214参照）

(8) Leslie, P. H.（1952）：The estimation of total numbers. Biometrika, 39, p. 363－388.

(9) 藤川正信（1963）：第二の知識の本．新潮社，p. 340（p. 242参照）

(10) 小林胖・野村悦子（1966）：一次情報．情報管理，9，p. 63－65.

(11)　カニング，R. O. ：経営のためのエレクトロニック・システム．玉井訳．産業図書．p. 363.

（1967年4月6日原稿受理，1968年1月5日改稿受理）