

分類, 索引, 抄録間の関係: 自動処理の観点からみて
Relations among Abstracting, Indexing and Classification :
From the View-point of Automatic Processing

細 野 公 男
Kimio Hosono

Résumé

Recently the increase of documents or reports in volume is rapid, so it becomes necessary to develop new techniques or methods to process such documents by using computers. This entails the new approaches and thoughts about conventional concepts of information and library science.

First of all this paper describes new thoughts based on a statistical approach about and relations between such terms as abstracting, indexing and classification, all of which are concerned closely with the following concepts, namely, category, information storage and retrieval pattern, and probabilistic and deterministic approaches.

Secondly the paper describes known existing techniques of abstracting, indexing and classification from the above mentioned point of view.

Lastly it describes the lack of consideration about retrieval in today's such methods, especially classification methods.

(School of Library and Information Science)

- I. 序
- II. 蓄積・検索パターン
- III. カテゴリー
- IV. 決定論的考え方と確率論的考え方
- V. 抄録, 分類, 索引間の関係
- VI. 結語

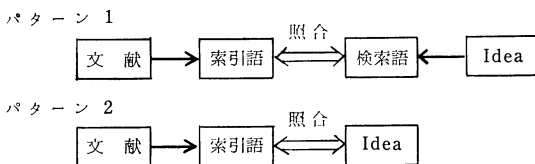
I. 序

近年雑誌, 報告書類の増大に伴い, 収録される論文数は幾何級数的に増大し, その結果個人が自分の対象領域中に属すると思われる論文から, 必要な情報を全て獲得することは困難になった。これら雑誌, 報告書等資料と云われるものの増加は人間社会の, 特に物質的發展に伴う新しい概念, 領域, 手法の出現によるものであるが, このような資料の増加は旧来の資料処理の形態に, 電子計算機による処理の機械化等大きな影響を与え, 図書館, 資料室等の名前で代表される資料を扱う組織体の機能をも, 今までの整理保存中心から検索中心へと変化させた。従って資料処理の方法も検索に便利な方法, すなわち検索を常に念頭においた方法が考案され, しかも迅速, 的確な情報処理が重要な問題となっている。その一つの表われとして自動分類, 自動索引, 自動抄録等の機械処理手法が不完全ながらも開発されている。この論文では文献処理の典型的な形態である抄録, 索引, 分類三者間の関係をそれらの機械処理と結びつけた立場で論述し, 得られた関係から既存の機械処理手法を論じるが, その場合, 抄録, 索引, 分類(いずれも操作を意味する場合と, その結果を意味する場合があるが, ここでは操作に限定する。)はいずれも検索を容易にするための手段を考えるべきであるので, これらの手段間の関係, および手法の説明は検索面から強く光をあてることにする。又処理の対象は文献とする。

II. 蓄積・検索パターン

人間の知識活動の所産である文献の整理, 蓄積操作と, その検索操作間の関係は第1図に示した二種類に分けられる。すなわちパターン1は照合が索引語と検索語間で行なわれるものであるが, これは機械の蓄積検索パター

第 1 図



ンであり, 文献処理の機械化を考える場合はこのパターンを使うことが必要である。パターン2は索引語と Idea が照合されるもので, 人間の蓄積検索パターンを示すものである。

III. カテゴリー(類)

ここで扱うカテゴリーは厳密な規定のもとで使用しているのではなく, “事物を分ける際の器となるもの”, すなわちあるカテゴリー中に含まれる要素は共通の特性を持つという意味で使用。共通の特性はどんな種類のものでもよい。例えば Vickery のカテゴリーではファセットを通して捕えられた特性がそれである。従ってクランプもカテゴリーである。

1. カテゴリーの種類

1.1 顕在カテゴリー

カテゴリーが明確に定義され, 従ってその性質も明らかかなカテゴリーで, 旧来のカテゴリーはいずれもこれに属する。

1.2 潜在カテゴリー

とにかくカテゴリーは存在するが, カテゴリーの性質が明らかでなく, そのカテゴリーの定義が出来ないもの。この種の例は心理学や社会学の分野にみられる。

2. カテゴリーの作成方法

2.1 知識体系中心

既存の知識体系に着目してカテゴリーが作られており, 各種図書館で使われている分類表中のカテゴリーはほとんどこの方式であるが, NDC, DC, UDC 等の様に知識体系そのものを重視する場合と, LC の様に資料に合うように知識体系を修正する場合とがある。

2.2 文献中心

内容や主題等文献の特徴を表わすと思われる具体的な因子(例えば重要語)に適当な処理を施すことにより, カテゴリーを作る。従って類はそのカテゴリーが作られた文献群, 又はそれと類似な性質を持つ文献群に対してのみ有効で, 文献の主題と関連する知識体系から作られるカテゴリーとは全く異種のカテゴリーが作成されることがある。これは知識体系から作られるカテゴリーが主題中心であるのに対し, 後者は必ずしもそうでなく, 語の連合を重視する方法もあるからである。その例として, phrase maker, phoneme, Markov process, terminal language¹⁾ を要素とするカテゴリーがあげられる。

カテゴリー作成に計算機を導入する場合は 2.2 の形態をとる。例えば Needham¹⁾ は文献中で共出現する語に着目し, それからクランプを作ることを試みた。その方法ではまず N 個の語に対し, 任意の二つの語が対となって出現する文献数と, その少くとも片方の語が出現する文献数との比を表わす共出現係数を求め, それを要素

とする $N \times N$ の係数行列 A を作る。次に N 個の語からなる集合の部分集合を示すベクトル U を考える。 U は N 個の要素からなり、 N 個の語のうち U に含まれるものに対しては $+1$ を、含まないものには -1 を要素として持つ。そしてクランプは Q を非負の三角行列として、 $AU=QU$ を満足する U と定義されるが、それはこの U が $(1'A1-U'AU)/(1'A1+U'AU)$ を極小にする U であるところから得られる。実際の計算は任意の U を選び、非負ではないが $AU=QU$ を満足する三角行列 Q を求め、 Q 中の負の要素の符号を変え、かつそれに伴い U の値を変えることによりクランプを求める。この外クランプを求める方式としては Dale²⁾ がある。

カテゴリーをその構造面からみると、体系カテゴリーの形式をとるものと、単に並列形式をとるものが考えられるが、2.2の方法から得られるカテゴリーは後者とならざるを得ないと思われる。それは一般に言われる体系カテゴリーが、体系分類のための知識内容の細分化を示しているからで、その結果どうしても意味や概念の問題が入ってくるからである。ただ体系カテゴリーを広義に解釈し、カテゴリー間にも上下関係があるという意味で使用すれば、2.2で与えられたカテゴリー内をさらにカテゴリー化することにより、2.2の方法でも理論的には体系カテゴリーの作成は可能である。

IV. 決定論的考え方と確率論的考え方

文献の処理にあたって文献の形で表わされる対象と、それを処理するための鍵との関係を、どう扱うかに関して二つの考え方がある。一つは旧来の決定論的考えで、それは文献には一義的に定まるその文献固有の鍵があると云う考えに立つもので、例として Coordinate indexing があげられる。この方式においてはディスクリプタ A を持つ文献は必ず A というカテゴリーに参照される。他方は確率論的考え方である。それは一般に対象全部を完全に把握することは不可能であり、あいまいさが当然現われるから、たまたま表われた事象のみに着目するのではなく、その事象を生みだした背景を推測する事が必要で、そのため文献語として現われる事象を客観的に処理する確率論的考えが必要であるとの立場である。この立場は Maron によって導入されたもので、あるディスクリプタを持つ文献は、そのディスクリプタと関連したある定められた確率でいくつかのカテゴリーに参照される。例えば重要語として “Social” という語が文献に表われると、その文献は確率 0.4 で “General psychology” お

よび確率 0.1 で “Social psychology” にそれぞれ参照される。³⁾ 検索時にディスクリプタに重みをつけ検索する場合も、その例と考えられる。多量のデータを処理する情報蓄積検索において fact 的なものが対象である場合を除き、後者の考えがより一般的であり、既存の手法もほとんどこの考えに基づいている。

V. 抄録、索引、分類間の関係

A. 定義

1. 抄録

文献内容の凝縮操作であり、その目的は情報要求者が何らかの方法により選択した文献から、情報を得させたり、その文献が真に必要なものかどうかを判定させることである。文献の処理方法として

i) 意味的内容を間接的に表わす要素 (例えば重要語の頻度や分布) に着目する方法と、

ii) 文献構成上の特徴に着目する方法がある。i) の方法としては Luhn³⁾ がある。この方法では重要文の抽出は二段のステップを経て行なわれる。すなわち、まず出現頻度に基づいて重要語を決定し、重要語、非重要語をその間にはさむ重要語の跨がり (cluster と呼ぶ) が短かければ短程、その cluster を含む sentence は重要であるとする。次に significant factor を (cluster 間の重要語数)² / (cluster 内の全ての語数) で表わし、その結果ある定められた値よりも高い値の factor を持つ sentence から抄録が構成される。後者の例としては論文作成者の意図が、文献中における各節の第一および最終 sentence に多く現われる傾向に着目し、第一および最終 sentence を機械的に抽出する方法や、前置詞句の形で重要な内容が表現されることが多いとの仮定のもとに前置詞句に重点を置く Baxendale⁴⁾ の方法がある。これは本来は索引を目的としたものであるが、後述する理由から抄録にも使用出来る。

2. 索引

抽出した文献の特徴を基として、文献相互の認識を可能にするために記号を付することで、その記号としてカテゴリー名も含まれる。個々の文献の検索を容易にするため記号化するのが索引の目的である。文献の特徴としては内容、主題、その他が考えられるが、ここで云う内容は文献そのもの、又はその一部を示し、主題とは内容から抽象される、内容を表わす概念とする。索引結果の形態としては KWIC の様な題名索引、Coordinate indexing や分類索引 (カテゴリー名を索引語とするもの)

の様な主題索引, Baxendale⁴⁾の手法で示される様な内容索引等があり, その表示法としては単なる記号の羅列の場合と, 電報抄録や Doyle⁶⁾の association-map の様に記号間に関係をつけたもの等がある。ところで検索は索引語と検索語間の照合作用で行なわれるのであるが, 索引語や検索語を決定する際に二つの考え方がある。すなわち索引時に索引語間の関係等種々の考慮を施すことにより, 検索語及び検索式を索引時に定められた枠内で考える方法と, それとは逆に索引時には融通性を持たせ, 検索時に工夫をこらし検索する方法である。前者の例としては電報抄録, Doyle⁶⁾があげられ, 後者の例としては後述する Stiles⁷⁾の方法があげられる。これらの方法の適否は recall, relevance の問題と関連させて論じなければならないが, 検索要求は個人によって種々な形態, ニュアンスを持つものであり, 従って検索語および検索式に枠をはめるべきでないという観点の方が一般的であり, その意味で後者が望ましい。Stiles⁷⁾の方法は, (1) ある語との連合の程度が高い順に配列された語のリスト (term profile) を全ての語について作成。(2) 検索語の各々の profile を相互に比較して一定数の profile に共出現する語 (first generation term) を選択。(3) first generation term を検索語として (2) の操作を行ない, second generation term を求める。この操作により同義語を処理する。(4) 索引語又は検索語として表われる語の各々について残りの全ての語との間の association factor を計算した expand list から各語の重みを計算する。(5) 各文献の索引語と検索語の expand list を比較し, 一致する語の数を数え, その重みの和を求める。この和の大小が質問と文献の一致度の程度を表わす。従ってこの方法では索引の段階では何らの規定なく語を使うことが可能であり, 又文献に索引されてない検索語からでも適合文献を検索でき, 不完全ではあるが前述した考えの後者の立場に立つ。

3. 分類

抽出した主題を基として文献を一つ, 又はそれ以上のカテゴリーに分け入れることで, この場合

(i) あらかじめカテゴリーを設定しておき, その中に文献を分け入れる。

(ii) 文献群から特定のカテゴリーをその文献群の特質に基づいて自動的に設定し, その中に文献を分け入れる。

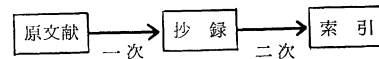
の二つが考えられる。なおカテゴリーの自動設定だけでは(自動)分類にはならない。又その目的は個々の文献

の検索を容易にするため, あるいは個々の文献をカテゴリー化することにより対象認識の能率化をはかるためである。⁸⁾又分類手法を使っても検索は可能である。

B. 抄録と索引との関係

内容の凝縮操作を行なう点において抄録は索引と本質的に同じであるが, その凝縮度に差があり, 抄録の方が凝縮度が第2図の様に小さく, 検索操作は第1図のパターン2しかとりえないところに索引との違いがある。もしパターン1の形式をとる場合は fact そのものの検索となり, 文献検索ではない。

第2図



従って抄録は索引機能が低く, 検索の観点からみた場合二次的なものであり, 一次的な索引を補う立場にあり, 索引のための記号化は必ずしも必要ない。凝縮度の観点からみた場合索引手法は全て抄録に使えるが, 索引は文献の特徴を記号化するものであり, その記号化の対象は内容に限らず, 例えば主題, 著者でもよいから索引手法は必ずしも抄録に援用出来ない。なお逆は常に成立する。

C. 分類と索引との関係

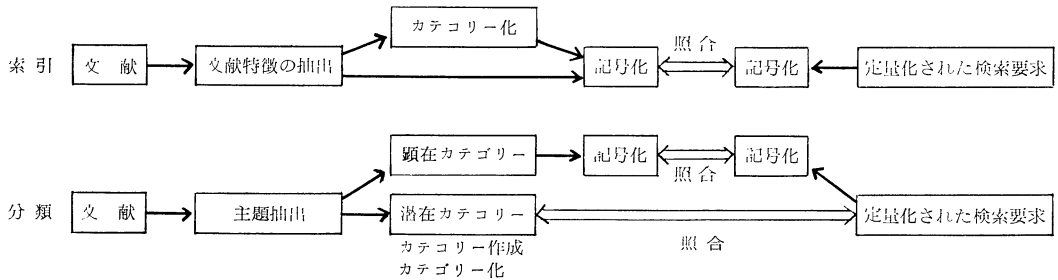
分類は内容ではなく主題を処理対象にするので, 抄録とは処理上何らの共通点を持たないが, 索引のうち主題を処理対象とするものとは共通点を有する。しかし索引が記号化を最終の目標としているのに対し, 分類ではカテゴリー化が最終の目標であるところに大きな違いがある。従ってカテゴリー化されたものを記号化すれば, それは本質的に索引と同じである。ところでカテゴリーの項で述べた様に, カテゴリー化された結果が必ずしも索引としての記号化が出来るとは限らない。すなわち潜在カテゴリーはカテゴリーの存在はわかっているとしてもその性質がわからないため, 索引としての記号化は不可能である。一見潜在カテゴリーは分類の目的に即さない様にみえるが, 実際はそうでない。確率論的考えで文献処理を行なう場合, 文献内のデータから帰納的にカテゴリー, および文献を分け入れる基準が得られるので, 旧来の決定論的考えではカテゴリーそのものが重要であるのに対し, 確率論的考えではデータからカテゴリーや判断基準を生み出す過程そのものが本質的に重要である。従ってこの過程が把握されていれば, カテゴリーは必然的に得られ, かつ適切なカテゴリーに文献が分け入れられる。

このため検索操作を考える場合、丁度カテゴリーに分け入れる際と同じ操作を文献の主題を表わす因子に対応する検索語に施せば、同様の処理過程でカテゴリーに行きつくことが出来、そのカテゴリー中に含まれる文献を抽出出来る。しかし今の場合検索語はカテゴリーや判断基準を作り出した際のデータと同様の形態をしていることが必要である。後述する様に機械処理の手法においては、文献の特徴を質的につかむのではなく量的につかむ方式が多い。例えば旧来の方法が文献の特徴を表わす因子について、それがあつかうかが問題であったのに対し、機械処理による方式では、その各々の因子が、因子相互間でどれ程の割合を示しているかを重視するので、デー

タはあるかないかの二進的でなく、例えば重要語の出現頻度の様に十進的である場合が多い。従って検索要求を検索語に重みをつけることにより、(例えばディスクリプタ A は重み 5, B は 6, C は 3, 等) 数量化することが必要となる場合がある。

前述した理由からカテゴリーの明確な定義は検索に必要でない。従って明確な定義を持たない潜在カテゴリーを文献処理に使う場合には、索引語が存在せず、記号間の照合操作によっては検索が行なわれない。すなわち潜在カテゴリーを扱う分類は索引とはなり得ないのである。ここに索引と分類の違いがみられる。以上から索引と分類とを图示すると第 3 図の様になる。

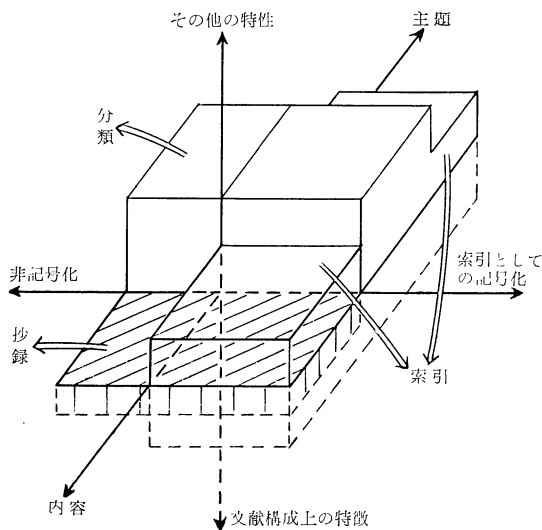
第 3 図



又抄録をも含め、抄録、分類、索引三者間の関係を三次元で表わすと第 4 図が得られる。第 4 図において座標軸に分類、索引、抄録の特性を表わす因子をとったが必

ずしも三者間の関係を示す特性が三次元空間上で表わされ得ないので、今までに記述しなかった特性はその他の特性で代用した。

第 4 図



ところで既知の索引、分類手法は今まで記述して来た視点のどの様な立場に立つものであろうか。

旧来の決定論的索引から確率論的索引の道を開いた Maron の方法⁹⁾から述べていこう。この方法ではあらかじめ定められた 32 個の顕在カテゴリーにサンプル文献を人間により分類し、32 個のカテゴリーのそれぞれで頻度数の高い重要語 90 個を選択した。次に j カテゴリーに分類された文献数を数え、それを全文献数で割ったものを、ある文献が j カテゴリーに分け入れられる確率 $P(c_j)$ とし、又 j カテゴリーに分け入れられた文献に属する語 w_1 の数を、 j カテゴリーに分け入れられた文献に属する全ての重要語の数で割ったものを、ある文献が j カテゴリーに分け入れられる時に語 w_1 を含む条件付確率 $P(w_1|c_j)$ とし、条件付確率の公式から

$$P(c_j|w_1) = \frac{P(c_j) \cdot P(w_1|c_j)}{P(w_1)} = k_1 P(c_j) \cdot P(w_1|c_j) \dots (1)$$

を求めこの値の一番大きいカテゴリーに新しい文献が分

け入れられる。一般的には (1) 式は

$$P(c_j|w_1, w_2, \dots, w_r) = kP(c_j) \prod_{i=1}^r P(w_i|c_j) \dots \dots (2)$$

で与えられる。なお Bayes の定理を使った Birnbaum 等¹⁰⁾ の方法も (2) 式と本質的に同じである。この方法は分類を経て索引が行なわれ、検索はカテゴリー間の照合操作により行なわれる。すなわち検索要求の内容を表わす 90 種の重要語のいくつかから (2) 式により得られるカテゴリーが検索語として照合される。又これら重要語を大きさ 1 の数量データと考え、(2) 式の処理過程で、求めるカテゴリー、従って文献を得ると考えれば、この方式は索引ではない。

Borko 等¹¹⁾ は因子分析法を利用した自動分類手法を考えた。因子分析法が有力な方法として適用出来る領域は、基本的概念が未だ明確でなく、その領域がいかなる要因から構成されているかわからぬ領域や、決定的な実験を行なうことが困難な領域であり、この場合因子分析法によって得られる成果は、未知の領域についてのおおざっぱな輪郭を与えてくれるにすぎない。¹²⁾ 因子分析においては各変数はそれに含まれると考えられる数個の因子の一次結合によって (3) 式の様に表わされると考える。

$$z_{ji} = a_{j1}F_{1i} + a_{j2}F_{2i} + \dots + a_{jm}F_{mi} + u_jU_{ji} \dots (3)$$

ここで z_{ji} は変数 j の i 番目の標準得点, $a_{j1}, a_{j2}, \dots, a_{jm}$ は変数 j に関する m 個の共通因子負荷量, u_j は変数 j の独自因子負荷量, $F_{1i}, F_{2i}, \dots, F_{mi}, U_{ji}$ は i の各因子得点を示す。 n 個の変数 ($j=1, 2, \dots, n$) に関する N 個 ($i=1, 2, \dots, N$) の得点 z_{ji} を求めると $n \times N$ 個の z_{ji} が得られる。これから n 個より少い共通因子負荷量 a_{j1}, \dots, a_{jm} を求めるのが因子分析の目的である。

Borko 等は z_{ji} として Maron によって与えられた 90 種の重要語がそれぞれ一文献中に出現する回数とし、文献数を 260 とした。よって Borko 等の場合 (3) 式において $j=90, i=260$ となる。(3) 式から求められる相関行列 R は A を因子行列, A' をその転置行列として $R=AA'$ の形で表わされる。 A は共通因子負荷量を要素

| | | | | | | | | |
|----|----|---|---|---|---|---|---|-------------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | |
| A= | 1 | 0 | 0 | ⋮ | ⋮ | * | 0 | * は 0 でない要素 |
| | 2 | 0 | * | ⋮ | ⋮ | 0 | * | |
| | ⋮ | * | 0 | ⋮ | ⋮ | 0 | 0 | |
| | 20 | 0 | * | ⋮ | ⋮ | * | 0 | |
| | 21 | * | * | ⋮ | ⋮ | * | * | |
| | ⋮ | 0 | 0 | ⋮ | ⋮ | 0 | * | |
| 88 | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| 90 | * | * | ⋮ | ⋮ | 0 | 0 | | |

とする行列で、基準軸を回転させることにより A の要素の一部を 0 とする事が出来、0 でない要素からカテゴリーの性質が主観的に決定出来る。例えば左下の A において列は重要語を、行はカテゴリーを表わし、カテゴリー 5 は重要語 1, 20, 21, 88 で表わされ、その値の大小によりカテゴリーの性質、従って名前が主観的に定められる。Borko 等の実験では 21 個のカテゴリーが得られた。カテゴリーへの文献の分け入れは次の (4) 式でなされる。すなわち

L_j を重要語 j の当該文献中における出現頻度
 T_{ij} を i カテゴリーに対する重要語 j の因子負荷量
 P_i を i カテゴリーの重みとし、

$$P_i = \sum_{j=1}^{90} L_j T_{ij} \dots \dots \dots (4)$$

の値が一番大きいカテゴリーに分け入れられる。この方法も分類を経て索引が行なわれ、検索もカテゴリー間の照合作用で行なわれる。もし検索の際、重要語間に重みをつけることにより検索要求を丁度分類を行なった際のデータの形で考えるならば、記号間の照合ではなく分類と同じ操作により適合文献が得られるから、その様に考えればこれは索引手法ではない。

Baker¹³⁾ は社会学で使用されている潜在構造分析を文献処理に応用した。この手法は文献群にいくつかの潜在カテゴリーがあると想定し、個々の文献をカテゴリーに割りあてるものである。この手法はまず k 個の重要語の各々に対し文献がそれを含まるか、含まないかを基準として文献を分類することを考える。 k 個の重要語に対しては 2^k 個の異ったパターンがある。 π_i を重要語 i がある文献に含まれる確率、前もって定められる相互排他的なカテゴリーの数を m , V^α を文献が α 番目のカテゴリーに属する確率、 λ_i^α をカテゴリー α に属する文献が重要語 i を含む確率とすると、 $\pi_i = \sum_{j=1}^m V^\alpha \lambda_i^\alpha$ 等を得る。潜在構造分析はサンプルからのデータ π_i 等を使って V と λ を推定し、これから重要語のパターンを反映する ordering ratio を計算し、文献は最大の ordering ratio を持つカテゴリーに分類される。この方法と Maron の方法の違いは手法そのものの外に、Maron が文献がカテゴリーに分類される確率、すなわち $P(c_j), V^\alpha$ をデータとしているのに対し、Baker は重要語が文献に含まれる確率、すなわち $\pi_i, P(w_i)$ をデータとしているように Approach の違いがある。

次にこの方式を使った場合の検索であるが、この場合カテゴリーが顕在していないため、検索要求から得られ

たカテゴリーと索引語としてのカテゴリー間での照合操作によって検索を行なうことは出来ない。従って検索操作を Baker は次の様に考えている。検索要求を表わすと思われる重要語から、重要語のパターンを求め、ordering ratio を計算し、その様なパターンを持つ文献が分け入れられるカテゴリーをつきとめる。次に検索要求を表わす重要語パターンから得られる ordering ratio に関係なくあらかじめ定めた ordering ratio よりも大きい値を持つ文献をそのつきとめられたカテゴリーから抽出する。ここで問題となるのは、あらかじめ ordering ratio を定めるにあたって、何らその根拠が与えられていないことであり、検索要求を表わす重要語パターンから得られる ordering ratio と等しい、又はそれ以上の ordering ratio を持つ文献を抽出する方が妥当性があると思われる。

VI. 結 語

文献量の増大に伴い、Coordinate indexing で代表される索引、検索システムは大きな影響をうけ、前述した手法等が生れて来たのであるが、その中にはカテゴリーを作ることにより処理操作を簡素化しようとする方針を取るものが多い。しかしその場合でも蓄積又は索引のみを目的として、検索をも考慮したものはほとんどないといえる。例えば Baker¹³⁾ の例では、重要語として computer だけしか含まない検索要求 (A) ではカテゴリー 2 の文献が得られ、重要語として、computer, automated を含む検索要求 (B) ではカテゴリー 1 の文献が得られる。Baker の例はカテゴリーが 2 種しかない特殊な例であるにしても、検索要求 (A) で得られる文献中には検索要求 (B) で得られる文献を含んでいなければならないはずであるが、その様な論義は全然なされていないのである。Maron, Borko 等の手法においても、ある検索要求により得られる文献中に、それよりも specific な検索要求で得られる文献を含んでいる何んらの保証もない。単に文献を蓄積するだけでは適正な情報処理が行なわれたことにはならない。従ってこれからの文献処理手法は検索の面により多くの考慮がなされることが必要である。

(図書館・情報学科)

- 1) Needham, R. M. A method for using computers in information classification, <International federation for information processing. *Information processing, 1962*> p. 284-7.
- 2) Dale, A. G. and Dale, N. "Some clumping experiments for associative document retrieval, *American documentation*, vol. 16, Jan. 1965, p. 5-9.
- 3) Luhn, H. P. "The automatic creation of literature abstracts," *IBM journal of research and development*, vol. 2, no. 2, Apl. 1958, p. 354-61.
- 4) Baxendale, P. B. "Machine-made index for technical literature—an experiment," *IBM Journal of research and development*, vol. 2, no. 4, Oct. 1958, p. 354-61.
- 5) Borko, H. "The construction of an empirically based mathematically derived classification system." <AFIPS. Spring joint computer conference, 1961. *Proceedings*> p. 279-89.
- 6) Doyle, L. B. "Semantic road maps for literature searching," *Journal of Association for Computing Machinery*, vol. 8, 1961, p. 553-78.
- 7) Stiles, H. E. "The association factor in information retrieval," *Journal of Association for Computing Machinery*, vol. 8, 1961, p. 271-9.
- 8) 外務省大臣官房電子計算機室. 自動抄録法および自動索引法の研究. Mar. 1966.
- 9) Maron, M. E. "Automatic indexing; an experimental inquiry," *Journal of Association for Computing Machinery*, vol. 8, 1961, p. 407-17.
- 10) Birnbaum, A. and Maxwell, A. E. "Classification procedures based on Baye's formula," *Applied statistics*, vol. 9, 1960, p. 152-69.
- 11) Borko, H. and Bernick, M. "Automatic document classification," *Journal of Association for Computing Machinery*, vol. 10, 1963, p. 151-62.
- 12) 清水利信, 齊藤耕二. 因子分析法. 日本文化科学社. 1966. p. 1.
- 13) Baker, F. B. "Information retrieval based upon latent class analysis," *Journal of Association for Computing Machinery*, vol. 9, 1962, p. 512-21.