Library and Information Science No. 7 1969

文献検索における自動分類手法

Automatic Classification Technique for Document Retrieval

細野公男
Kimio Hosono

Résumé

The concept of classification in library and information Science is being changed to the direction from the storage-oriented to the retrieval-oriented.

Classification is one of men's approaches to analyze or infer something from the objects and determine their decision from it. As classification is connected closely with human knowledge, there are many different kinds of view-points and approaches about it's concept and purpose. Therefore it is necessary to have an insight into the concept, purpose and techniques of automatic classification to clarify the difference of these from those in other fields, to consider the possibility of adaptation of automatic classification techniques developed in other fields, to refer to the problems when they are adapted, then we can determine what kind of classification technique is useful for the library and information science field.

First af all this paper describes the concept of automatic classification and the difference of it's purpose in the library and information science field and other fields.

Secondly the paper describes the factors which affect automatic classification techniques, such as characteristics of objects, processing steps of which classificatory operation is composed, criteria that determine to which category the objects should be distributed and general problems of the techniques.

Lastly it describes what kind of approach or viewpoint is necessary to make classification algorithms in library and information science, and describes a new technique from this point of view. The technique is based on what Bonner has developed. It has the following processing steps:

1) To select a number of characteristics which are representative of a group of documents, and to make a data matrix from the group of documents and characteristics.
2) To make a similarity matrix by calculating similarity measure between documents.
3) To make clumps which are categories and to distribute documents into them from the similarity matrix.

(School of Library and Information Science)

文献検索における自動分類手法

I. 序

情報社会と呼ばれる今日において，多量の情報を蓄積し，検索することは我々にとっては大きな関心事となった。特に蓄積された対象の検索の問題はそれ自身で一つの研究分野を構成する。一般に検索について考える際に二つの側面が考えられる。一つは検索の理論的な面で，これはどの様な方針でどの様な手段を使うかに関連する。他方は実際面であり，対象処理手段のアルゴリズム化が可能かどうかの問題である。

ところで図書館・情報学においては対象処理の一手段として分類があげられる。分類の概念，手法は種々の分野で独立に研究され，その過程の一部がアルゴリズム化された（すなわち自動分類の）手法が開発されている。よって検索と自動分類の概念，手法を結びつけること，すなわち分類を検索面から光をあてて分析することは，有効な検索方法の開発に役立ち，かつ必要である。

それ故，本稿では自動分類の概念，目的，手法等について解析し，次に分類操作を行うにあたっての考え方，アプローチの手段（アルゴリズム）等において他の分野で開発された自動分類手法の援用可能性，およびそれが援用される場合の問題点を図書館・情報学の面から論じ，最後に文献情報の検索に役立つ分類手法の一私案を提示する。

II. 自動分類理論

1. 分　類

分類の概念は広範なものであり，図書館・情報学におけるその概念は図書中心から雑誌等逐次刊行物中心への変化で代表される対象物の質的・量的変化，および電子計算機による情報処理の考えで代表される，処理方法の変化等の影響をうけ変りつつある。一方人間が事物を分析し，また事物から何かを推定し，何らかの判断を行なおうとする時の一つのアプローチとして分類があげられる。従って分類操作は人間と深く結びつき，その応用範囲の広い事と相まって各分野ごとにその概念，目的等にずれが生じている。

それ故，これら種々の変化および違いをも考慮した分類の一般的定義は次の様になる。

抽出した属性を基として対象を一つ，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，その中に対象を分け入れる。

（ii） 対象から特定のカテゴリーをその対象の属性に基づいて自動的に設定し，その中に対象を分け入れる。

の二つが考えられる。又その目的は個々の対象の検索を容易にするため，あるいは個々の対象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り，より高次の分析に役立たせるためである。[1)]

ところで対象認識の能率化には次の二つの面がある。それは

（ｉ） 対象間の結合関係の明確化

a. 単に類似性の強い対象同志をまとめる。

b. 個々の対象間に全て結合関係をつける。

（ii） 対象に影響を与えている因子の抽出および明確化。

であり，（ii）の場合は分類された個々の対象よりもカテゴリーそのものが大切である。（ii）を目的とする方法としては因子分析法がある。図書館・情報学分野においては，分類の主な目的は文献検索，および対象間の結合関係を明確化する形をとる対象認識の能率化の二つであり，能率化は a. と b. が結合した形が望ましい。後者の例として索引語の質的決定，シソーラス作成があげられ，検索効率向上のために役立つ。

他方一般に他の分野においては，分類の目的は対象認識の能率化であり，その結果は検索と結びつかない。例えばある発掘された人間の頭骸骨が，縄文，弥生，古墳のどの時代に属するかを骨の高さ，巾，長さから決定する場合を考える。この場合カテゴリー化された人骨を後に，文献検索の場合のように引き出すことはありえない。従って検索手段として他の分野の手法を援用することはむずかしく，援用出来たとしてもかなりの修正が必要である。対象認識の能率化を目的とする場合は他の分野の手法の援用は可能であるが，分類対象の特性を考えて，援用手法を選びかつ修正することが必要である。

分類の形態としては，分類されたその対象間の結びつきに包含関係（上下関係）を考える場合と，考えない場合があるが，前者を体系分類，後者を非体系分類という。

2. 自動分類手法を構成する各処理過程

自動分類手法は分類を自動的（機械的）に行なうことを意味するが，直接には分類を行なうにあたっての適正なアルゴリズムを開発することである。それは対象中の関連性に基づき要素の集合を作成することであるといえるが，このある種の関連性により一つのグループを形成する要素の集合を，クランプ又はクラスターと呼ぶ。これはカテゴリーの一種である。ところで **II.1** でふれた様に，分類の形態としては体系分類と非体系分類があるが，自動分類における体系分類は，結果として概念の上下関係が表わされる事もあるが，旧来の体系分類のように概念的関係が直接，表に現われるものでなく，近似度による要素間の結びつきの順序を表わしたものであり，この場合もクランプ化の過程が含まれうる。従ってどちらの形態をとるにしてもクランプ化は重要な問題である。

自動分類においては開発されたアルゴリズムをプログラム化することにより，機械処理が可能になるわけであるが，分類手法を構成するどの処理過程がアルゴリズム化されているかは，各種手法によりかなり異なる。

分類手法を構成する各処理過程は

a. 対象を分類するにあたっての基礎データとなる属性の決定，および基礎データの抽出。
b. 対象が分け入れられるカテゴリーの数，および性格の決定。
c. 対象を分け入れる方法およびその基準の決定。

から構成される。分け入れる方法とは分類の尺度として何を選ぶか，体系分類を使うか否か等の考察と実際のその手段からなる。上記のうち

（ｉ） 基礎データの抽出
（ii） カテゴリー数の決定
（iii） 分類の手段および基準

はアルゴリズム化の対象となる。なおカテゴリーの数をあらかじめ決定してしまわないで，分類の結果プロダクトとして得る場合，カテゴリーの性格決定は作成されたカテゴリーに含まれる要素から人間が判断して決めることになる。現在の手法では（iii）のアルゴリズム作成，(iii）と（ii）を同時に含むアルゴリズム作成が主に開発されている。

3. 自動分類手法に影響する諸因子

旧来の分類手法が対象を質的に把握したのに対し，自動分類では量的に把握するので，分類手法に影響する諸因子を明確にする必要がある。

ところで自動分類手法は分類方針により種々な形態をとるから，手法に影響する因子は分類方針の対象となるものであり，**II.2** で述べた各処理過程を構成する個々の要素である。分類方針として次の事が考えられる。

A．カテゴリーの形態をどうするか。
B．属性として何を選ぶか。
C．分類方法をどうするか。
　1．体系分類か否か。
　2．分類尺度（関連性尺度）として何を選ぶか。
　3．どのような分類手段を使うか。

従ってこれら個々についての考察が必要でそれを以下に順次示す。

A. カテゴリーの形態

すでに述べたようにあらかじめカテゴリーを作る場合と，要素の分け入れと同時に作る場合との二種類がある。あらかじめカテゴリーが作られている場合には分類手法は，分け入れのための判定基準を作ることが目的となり，分類手段においてクランプ化の過程は経ない。カテゴリーをも同時に求めようとする場合はクランプ化の過程が必要であるが，そのクランプ化の意味が体系分類をとる場合と非体系分類を行なう場合では異なる。

B. 属性および基礎データの特性

分類対象を充分に代表しうると思われる属性を選択することは，その属性からデータが得られ，このデータに対して分類操作が行なわれるので，分類手法を開発するにあたって最も大切である。すなわち自動分類では対象を分類するにあたって，対象の特性を代表する属性群に関するデーターに処理を施して対象を分類するからである。従って属性として何を選ぶか，属性間に相関はあるのか等を考えねばならない。次に属性を表わすデータはどのような形態をとりうるであろうかを考えると以下の二つがあげられる。

（ｉ） 属性を質的に把握。すなわちデータは0か1の二進的形態をとる。
（ii） 属性を量的に把握。すなわちデータは十進的形態をとる。

（ｉ）と（ii）のどちらを選ぶかは分類の目的，対象の特性，分類操作と密接に関連する。理論的には適正な数の属性を選んだとしてもそれが必ずしも分類システムに合ったものであるとは云えない。分類手法を考える場合，例えば処理に用する時間も又大きな因子となるからであり，理論と実際（処理操作）は同じ重みを持つことが必

要である。このため対象の属性を表わす情報の減少をなるべく少くして属性の数を減らすことが分類に関する研究の一つとなる。

基礎データの性格(これは分類対象の性格でもあるが)としては以下の二つがあげられる。

(i) 基礎データは母集団そのもののデータである。すなわち対象それ自身で母集団を構成する。

(ii) 基礎データは母集団からランダムに抽出された標本のデータである。すなわち対象はある母集団からランダムにとられた標本である。

(ii) の場合分類は対象が抽出された母集団に対して行なわれることになる。それ故対象は母集団を適正に代表するものであるという仮定が暗黙のうちにあり，このように基礎データを考える場合，分類操作は数学的仮説に基づいた理論的根拠からそのアルゴリズムが作られる。

C. 分類方法

(i) 関連性の尺度（分類尺度）

対象間の関連性の尺度として次の二種類が考えられる。

(イ) 類似性

(ロ) 非類似性

類似性を示す尺度の例としては，属性の共出現度数，相関係数があげられ，非類似性の例としては，各種の距離（すなわち距離の遠い要素同志は非類似性が高い）があげられる。類似性と非類似性は表と裏の関係にあるので(イ)と(ロ)の差は本質的なものでなく，どちらの側に光をあてるかの差だけである。しかし一般に類似性の尺度はグループ内の均一性を表わし，非体系分類に使い，非類似性の尺度はグループ間の異質性を表わし，体系分類に使うとみられるが，考え方によっては必ずしもそうでない。何故ならば均一性を表わす尺度は異質性を表わす尺度に変換されうるからである。例えば類似度を示す値として 0 から 1 までの実数を考え，その類似度を S_i として，これから考えられる新たな尺度として $-\log S_i$ を選べば，この値は小さい S_i に対して大きな値をとる。すなわち類似性が小さければ大きな値をとるのであるから，非類似性を表わす尺度であり，ある種の距離とみなすことが出来る。[2]

関連性の尺度として何を選ぶかは属性の特徴，分類方針等によって定められる。例えば属性が母数的であれば尺度として相関係数が使える。

(ii) 分類形態

分類結果をどのように配置するかも分類方針と結びついて分類手法に影響するが，その配置法として次の二つがある。

(a) 体系分類に基づく配置

(b) 非体系分類に基づく配置

分類を実際に行なうにあたって，基礎データから類似関係にある要素を互いに結合させてクランプを作っていく融合的 (agglomerative) 方法と，基礎データを全体と考え，それを分割していくことによりクランプを作成する分割的(devisive)方法があるが，この区別が明瞭でない手法もある。[3] 非体系分類についてはこの種の考察はなされていないが，これは非体系分類においては要素又はクランプ間のつながりが無視されているからである。又要素を一つのクランプのみに属させるか，二つ以上のクランプに属させうるかの選択も分類形態に付随した問題である。

体系分類では要素間，およびクランプ又はカテゴリー間の結びつきの順序が明らかにされるのがその特徴で，あらかじめ定められている体系カテゴリーに分け入れる方法ではない。従ってカテゴリーを最終的に定める道筋を最適化するのが目標である。[4]

ところで体系分類におけるクランプおよびカテゴリーは非体系分類におけるものよりも明確でないので説明が必要である。すなわち要素間の類似度を表わす数値レベルを有する体系は樹木図 (dendrogram) と呼ばれるが，その樹木図においてある数値，又はそれ以下の数値レベルのもとで結合状態にある対象の集合をクランプと云い，ある判断基準に合致する数値又は，それ以下の数値レベルのもとで結合状態にある対象の集合をカテゴリーという。本稿ではカテゴリーは最終的に得られる要素の集合のみを意味するのに対し，クランプは分類中に生じる要素の集合をも含む。又個々の要素そのものがカテゴリーになることもありうる。

体系分類では要素間又はクランプ間の関係をつけるのが主目的のため，クランプ間の異質性に重点をおき，従ってクランプ内の均一性は，ある程度犠牲にされる。要素は当然一つのクランプのみに属する。

この方法の欠点は融合的方法，分割的方法のいずれをとるにせよ，前段階の結果が後々にまで影響を与え，誤った融合および分割は累積して，後には無視出来ない誤りとなる怖れがある。なおこの点に関して Macnaughton-Smith は融合的方法よりも分割的方法がより安全であると述べている。[5]

非体系分類ではクランプ内の均一性を最適化するのが

目標であり，ここに重点がおかれる。その主目的は要素のグループ化，すなわちあらかじめ定められた類似性の尺度に従って，要素をカテゴリーに分け入れることである。カテゴリーはあらかじめ定めてある場合と，ない場合があり，従ってクランプ化の過程をとるものとそうでないものがありうる。要素は一つのクランプのみに属する場合と，二つ以上に属する場合がある。この方法の欠点はクランプ化の過程が，一般に試行錯誤的であり，これは処理時間を長くするだけでなく処理時間の予測をつけにくくする。体系分類よりも処理時間が長いのが通例である。利点はあるクランプに割り当てられた要素の再割当が，後に可能であることで，これはクランプの修正が可能なことを示す。

分類の目的から考えると体系分類は対象認識の能率化に，非体系分類は検索のために有効である様に思われる。

(iii) 分類手段

分類手段は分類方針により種々な形をとるが大きくわけ，数学的に理論付けられた数学的方法と，試行錯誤又は直観により得られた比較的簡単な経験的方法がある。対象の属性が明確につかめてない時に数学的仮説に基づいた方法を使うのは危険であるが，一方誤って分類される確率等，この方法は分類手段の有効性を測る公式が誘導出来る可能性がある。すなわちあらかじめその分類手段の効率を推定しうる可能性を持つ。一般に分類対象を母集団そのものとみる場合，分類手段は経験的方法をとり，対象を母集団からとられた標本とみる場合は数学的方法をとる。分類手段のアルゴリズムは簡単な方が良いのであるが，後者の場合は標本を通して母集団を分類する以上，数学的仮説をたてざるを得ない。その場合分類の誤りの可能性は前者に比べ当然高まる。

4. 分類操作の手順

分類操作はある定められた分類方針のもとでなされ，データの作成，関連性尺度の抽出又は計算，および分類手段の適用の三つの段階から成り立つ。

(i) データの作成

データはすでに述べた様に属性を質的に表わす二進的データと量的に表わす十進的データがある。このデータから関連性の尺度の抽出，又は計算が行なわれる。データはその後の処理のため，データ行列の形をとるのが一般的である。データ行列の例を以下に示す。なお関連性の尺度も場合により又データとみなしうる。

第1図 データ行列

		属性 a	b	c	d	e	f	g
対象	1	1	0	0	1	1	0	1
	2	0	1	0	1	0	1	0
	3	1	1	1	0	0	1	0
	4	1	0	0	1	0	0	1
	5	1	0	1	0	1	0	0
	6	0	1	1	0	0	1	1

(二進的データの場合)

(ii) 関連性の尺度抽出又は尺度計算

関連性の尺度値は基礎データから種々な変換方法により得られ，分類手段の適用される対象である。例えば共出現度数の個々の出現度数に対する割合を考えれば類似性の尺度が得られ，非共出現度数を求めればそれは一種の距離とみなすことが出来，非類似性の尺度が得られる。データの場合と同様に関連性の尺度値を表わす行列を作る場合がありそれを以下に示す。その行列が類似性に基づく場合は類似行列，非類似性に基づく場合は非類似行列と云う。なお数値形態は二進的，あるいは十進的である。

第2図 非類似行列

		対象 1	2	3	4	5	6
対象	1	0	1	4	5	1	4
	2	1	0	5	4	2	5
	3	4	5	0	3	3	2
	4	5	4	3	0	6	3
	5	1	2	3	6	0	3
	6	4	5	2	3	3	0

(十進的データの場合)

(iii) 分類手段の適用

前段階で得られた関連性の尺度値に分類手段を適用することにより，すでに述べた様に

a. 分類対象を既存カテゴリーに分け入れる。

b. 分類対象をクランプ化する。

ことが行なわれる。ところでクランプ化が非体系分類においては最終目的であるのに対し，体系分類では樹木図を作る際の過程にすぎない。従ってクランプ作成の方法も当然異なる。非体系分類におけるクランプ化には次の4つの過程が含まれる。

1. クランプを作り始める方法
2. 要素をクランプに割りあてる方法。およびクランプ同志を融合させる方法。
3. 要素の割り当ての中止を決定する方法（従ってあ

る要素はそれ自身でクランプを構成することもありうる。)

4. クランプに既に割りあてられた要素の再割り当ての方法，すなわちクランプを修正する方法。[6]

しかし既存の手法がこの4つの過程を全て含んでいるわけでなく，3．4．を欠く手法もかなりある。

分類手段の例として，経験的方法をとる場合には凝集力法 (Needham) および最大法又は最小法 (Johnson) 等があり，数学的方法をとる場合は条件付確率，判別函数等がある。なお基礎データに直接分類手段が施される場合もある。

5. 自動分類手法の問題点

自動分類手法を考える場合，考慮せねばならない問題点は大きくわけて次の三つに要約される。

(i) 処理作業が部分的にせよ機械化されているので，電子計算機の能力に適合する手法でなくてはならない。従って入出力データの形態をどうするか，又処理時間をいかに短縮するかが問題となる。

(ii) 誤って分類される（誤分類）の確率を最小にする，すなわち誤分類の影響を最小にする問題。

(iii) 結果がその後の種々な問題の解析の際，有効に使用出来るか，すなわち目的に対し有効であるかどうかの問題。

これらの問題に対し種々な考察，例えば Wallance および Boulton[7], Goodall[8], Williams および Lance[9], Sneath[10] 等がなされており，一つの研究対象となっている。

我々は個々の対象を直接カテゴリーに結びつけて同定することは出来ず，対象に関して得られたデータから間接に同定するわけであるから，(ii) に影響する因子としてデータの特性，関連性尺度，分類手順を考える必要がある。分類手法を評価するにあたって，誤分類の確率が推定出来ることは望ましいことであるが，そのためにはこの三種の因子個々についての考察が必要である。

誤分類の確率を最小にする問題を考えるにあたって，もし分類手段として数学的仮説に基づく方法を使うならば，その確率をあらかじめ理論的に求めうる場合がある。[11] Jardine および Sibson[12] は分類手段が有効であるために要求される種々の条件について述べ，既存の手法がそれらの条件を満たしているかどうかにふれているが，その様なアプローチも誤分類を最小にするのに役立つであろう。

分類にあたってクランプ化の過程を経る手法では，一般に体系分類を使う方が非体系分類を使うよりも，その手法が複雑かつ精巧であり，その数も多い。従って適当な理論的モデルを作ることにより手法の有効性をチェック出来やすい。

6. 各種分類手法

現在分類手法として種々な形のものが開発されているが，代表的なものとしては以下のものがあげられる。

非体系分類としては，Needham 等を中心とする C.L.R.U. グループ[13,14,15,16,17], Dale, A.G. および Dale, N.[18], Bonner[19], William[20], 山川[21], Maron[22], Borko および Bernick[23], Baker[24] がある。

体系分類としては，Constantinescu[25], Edwards および Cavalli-Sforza[26], Johnson[27], Lance および Williams[28] 等がある。

III. 図書館・情報学における自動分類手法

1. 一般的特徴

分類の概念およびその手法は情報システムの改良のため充分活用される必要があるが，その場合単に存在する資料を，蓄積のために分類するだけでないのはもちろんである。分類の概念，手法が生かされる研究対象として

(i) 索引語の質的決定，シソーラスの作成

(ii) 文献検索

があげられ，(i) は対象間の結合関係の明確化を，(ii) はカテゴリー内の均一性を目標とする。よってこれらを方針とする手法を使うことが必要である。

図書館・情報学においては，現在対象として文献，語（ことば），言語が主であり，属性としては語および文献が使われている。

次に分類方針ごとに前章で列記した各種手法をとりあげ，この分野にとって望ましい分類手法の形をのべる。

(1) 対象間の結合関係の明確化

この分野においては語間の結合関係を調べることが必要になる場合はしばしばである。しかもその関係は旧来の概念関係でなく，文献群を媒介とした関係である。従ってこの場合考察の対象は語，その属性としては文献が使われる。索引語としてどの語を選ぶべきかを考えるとき，語と語の間の関係がわかっていることは非常に有効である。又シソーラス中には語の結合関係が含まれており，(1) を方針とする手法はシソーラス作成に役立つ。

しかし人間が作るシソーラスとはかなり異った物となると思われる。機械翻訳の辞書作成のためにもシソーラスと同じ意味で適用可能であろう。

この方針をとる分類手法は分類形態として体系分類をとる。Edwards の方法では結果が二分された形しかとらないが，図書館・情報学においては必ずしもそうでなく，二つ以上の語が同時に一つの語と結びつくことが当然ありうるし，語間の結びつきが tree 的でなく Needham が示したように lattice 形態をとる事が多い。又例えばシソーラスを考えた場合，語と語の間の関係をつけると同時に語のクランプ化もなされており，前述した体系分類を使う場合，体系中のあるレベルでのクランプ化が必要となるが，前記手法ではいずれもその分割中止の判断基準はない。これらの手法がいずれも各要素間の関係をつけることだけに主眼がおかれているからである。従っていずれの方法をとるにしてもその適用対象はかなり限定されると思われる。

（2） カテゴリー内の均一性

カテゴリー内の均一性を目標とする approach は文献検索，および対象認識の能率化を目的として単に類似性の強い対象同志をカテゴリー化する場合にとられる。後者の例として言語学における語の意味的な分類[29,30]があげられるが，ここでは文献検索の面についてのみ言及する。カテゴリー内の均一性をめざす時は，非体系分類の形をとる。

文献検索を目的として分類手法を考える場合には分類方法と関連していくつかの考慮すべき因子がある。その一つとして検索方法をどうするかがあげられるが，それには検索と分類とを同じ操作で行う場合と，全く分離した場合が考えられ，既存の手法は後者に属する。前者の場合は検索質問も基礎データとみなし文献データと共に分類操作の対象とし，その検索質問が分け入れられたカテゴリーに属する文献を適合文献とする方法が考えられる。既述手法の中には文献の分類を行なっていながら検索方法について何らの考慮がなされていない場合が多いが，それでは単に対象認識の能率化を行なっただけとなる。次にデータが母集団そのものか，又は母集団からとられた標本であるのか，があげられる。現在では文献群が構成する母集団の性質は何ら解明されていない。従ってデータとして標本を考えるのは，それが母集団を的確に代表する保証が何ら得られていないのであるから好ましくなく，この意味でデータ等に数学的仮説をおく数学的方法は危険である。この観点にたつと分類対象は，理論的にきめたある領域全体の文献ではなく，実際に分類したい（現存する）文献群である。それ故作られたカテゴリーは新しい文献が入るたびに影響をうけ，カテゴリーの改定がしばしば必要になるので，この点からも考慮する必要がある。その三として何を対象とし何をその属性とするかがある。文献検索が目的でも対象が必ずしも文献となるわけではない。一般に文献を対象，重要語あるいは索引語を属性とする場合，および索引語を対象として文献を属性とする場合がある。後者の方針をとると分類方法と検索方法は当然ことなり，カテゴリー化された重要語のリストを使って検索が行なわれることになる。なおカテゴリーは相互排他的である必要はない。

次に以上の点から各種手法の適用可能性を論じる。Williams, 山川, Maron, Borko 等, Baker は文献を対象，重要語又は索引語を属性として分類手法を開発しているが，Williams, 山川, Maron は分類にあたっての判断基準を標本データから求めているので，この分野では判断基準の有効性が保証出来ない。又 Borko 等は因子分析法を使っているが，この方法は本来，対象に影響を与えている因子の抽出，および明確化をねらう方法であり，カテゴリー内の均一性をめざす文献検索が目的の場合にはその意味で好ましくない。検索面から考えると Baker を除いた他の手法では検索に対して何らの考慮もなされていない。前述したように検索の際，検索質問を基礎データと同じように考え，カテゴリー化を行なう方法が検索方法として考えられるが，Williams や Borko の場合にはデータが十進的であり，従って検索質問を表わす重要語，又は索引語に重みをつけて十進的データに変換する必要が生じる。ところが重みのつけ方に明確な根拠がなく，この種の検索方法は使えない。山川の手法は分類の基準値を作る際に使用率により重要語をグループ化したが，そのグループは多くて3個なので，検索語間に 1〜3 の重みをつけることを行なえば，この種の検索方法を使うことが可能となる。

Maron, 山川等が文献を分類 (document classification) したのに対し，Needham, Dale 等は語の分類 (term classification) を行なった。文献分類の場合では Needham 等と同種のクランプ化を経て文献をカテゴリー化したものは上記手法中にはないが，文献分類と語分類ではどちらが文献検索により有効であろうか。前述したように文献検索を目的とした時にはデータそれ自身で母集団を形成するのであるから，どちらの方式をとる場合でも新しい文献が加わると今までのカテゴリーが有効

でなくなり，作りなおすことが必要である。語分類を使う場合は，カテゴリー化された語の集合から各々の語が含まれるカテゴリーのリストを作り，それから各文献に対してカテゴリーリストを作る操作を行なうが，それを全てやりなおさなければならない。一方文献分類では文献の分け入れられるカテゴリーだけが変化するので，改定の影響は語分類に比べはるかに少い。よってクランプ化の過程を経て，しかも検索面も含む文献分類法が文献検索に必要である。

2. 文献検索を目的とする新分類手法

前節から文献検索を目的とする場合，分類手法は，データがそれ自身母集団を構成すること，分類対象としては文献を使うこと，検索方法も含んでいることが三つの条件となる。

この条件を備えた手法として以下に述べる方法が有効であると思われる。それは Bonner の手法を基本としたもので，検索は検索質問を基礎データとみなし，分類と同じ操作で該当するカテゴリーに行きつき，同時にそのカテゴリーを形成する文献群を適合文献とする。分類は文献と検索質問両方を含めた集合に対して行なわれる。従って通常は文献を表わすコードはカテゴリー化されてはおらずデータ行列のままであり，検索質問が出されるとその質問をデータ行列の中に組込み，初めて類似行列の作成，クランプ化の手段がとられ，データがカテゴリー化される。この時検索質問とカテゴリーを形成する文献が適合文献となる。この方法をとるとカテゴリーの改定の問題は生じない。従ってこの場合分類と検索は同一操作である。

A. データ行列の作成

索引語を英字，文献および検索質問を数字で表わす。文献又は検索質問が索引語 a を含んでいれば 1，含んでいなければ 0 とする。又文献，検索質問が索引語を論理和の形で含んでいる時は分割し，それぞれを一つの文献又は検索質問とみなす。従って同一の番号を持つ行が生じることがある。データ行列の例を第 3 図に示す。

第3図　データ行列

	a	b	c	d	e	f	g
1	1	0	1	1	0	0	1
2	0	1	1	0	0	1	0
⋮	⋮			⋮			⋮
5	1	0	1	0	0	0	1
51	0	1	1	1	0	0	1
51	0	1	0	1	1	0	1
⋮	⋮			⋮			⋮
80	1	0	1	1	0	0	0

1〜50 までは検索質問を表わし，51以後は文献に使う。文献51は索引語 c と e が論理和の形であるので左の様に分離されている。

B. 検索質問の定式化

検索質問は次のように処理される。検索質問は検索語の論理式で例えば (A+C)・F・(−G) のように表わされる。この方法では文献の場合と同じように，( ) はばらされ二つの検索質問になおされる。すなわちデータ行列は論理積（否定も論理積である）で表わされる。次に最終的に得られた検索語群はそれぞれ文献を表わすものと仮定し，1 から通し番号をつけてデータ行列に加える。もし検索語が否定の形で使われておれば，カテゴリー作成後検索質問と同じカテゴリーを形成する文献中で，否定の形で検索式中に現われる語を持つ文献は除外される。（クランプ化の過程では除外されない。）

C. 類似行列の作成

類似性の尺度としては以下のように定義される Tanimoto の方法を使い尺度値 R_{ij} を求め，その値があらかじめ定めた値 T よりも大ならば 1，小ならば 0 とする。これは Bonner が使った方法である。

$$R_{ij}=\frac{\text{文献 } i \text{ および } j \text{ の両者に共通な索引語数}}{\text{文献 } i \text{ 又は } j \text{ 中に含まれる索引語数}}$$

第 3 図のデータ行列から得られる類似行列は第 4 図を経て第 5 図のように表わされる。なお $T=0.45$ とした。

第4図　類似行列 B_1

		文献 1	2	…	5	51	51	…	80
文献	1	1	1/6・3/4		3/5	1/3・3/4			
	2	1/6	1		1/5	2/5	1/6・1/5		
	⋮	⋮	⋮		⋮	⋮	⋮		
	5	3/4	1/5・1		2/5	1/6・1/2			
	51	3/5	3/5・3/5		1	1/2・3/5			
	51	1/3	1/6・1/6		1/2	1	1・1/7		
	⋮	⋮	⋮		⋮	⋮	⋮		
	80	3/4	1/5・1/2		2/5	1/7・1			

第5図　類似行列 B_2

		文献 1	2	…	5	51	51	…	80
文献	1	1	0・1		1	1・1			
	2	0	1・0		0	0・0			
	⋮	⋮	⋮		⋮	⋮	⋮		
	5	1	0・1		0	0・1			
	51	1	0・0		1	1・0			
	51	0	0・0		1	1・0			
	⋮	⋮	⋮		⋮	⋮	⋮		
	80	1	0・1		0	0・1			

D. クランプ化

クランプ化は二つの手段でなされる。それぞれをクランプ化 A，クランプ化 B と名付ける。クランプ化 B はクランプ化 A で出来たクランプの修正を行なうものである。これら手段は Bonner の tight cluster 作成，core cluster 作成，およびクランプ修正と本質的には一致するが，方法的にはやや異なり，tight cluster 作成と core cluster 作成の区別がなされていない。文献検索ではカテゴリーが相互排他的である必要はないから，特に core cluster を作る必要性がないのがその理由である。クランプ化 A は互いに類似している文献同志の集合を作る，すなわち類似行列中で共通の場所に 1 を有する文献をグループ化する手段である。どこにも属さない要素はそれ自身だけでクランプを構成する。クランプ化 B はクランプ化 A で出来たクランプ相互間で，クランプを構成する個々の要素を媒介として類似度を調べ，その類似度があらかじめ定めた値よりも大であるクランプの組の中で類似度が一番高い組をまとめて一つのクランプとする手段である。この操作はクランプ間の類似度が定められた値よりも大きいものがなくなるまで続けられる。第 6 図の類似行列にクランプ化 A を適用するとクランプとして (1, 2, 6), (3, 5), (4, 7), (8) が出来，この結果にクランプ化 B を適用すると (1, 2, 6), (3, 5, 8), (4, 7) がクランプとして得られる。

第 6 図　類　似　行　列

	1	2	3	4	5	6	7	8
1	1	1	0	0	0	1	0	1
2	1	1	0	0	0	1	0	0
3	0	0	1	0	1	0	0	1
4	0	0	0	1	0	0	1	0
5	0	0	1	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	0	0	0	1	0	0	1	0
8	1	0	1	0	0	0	0	1

クランプ化 A およびクランプ化 B のフローチャートを第 7 図以下に示す。

クランプ化 B の方がクランプ化 A よりも条件がゆるいので多くの文献が検索される可能性がある。従ってまずクランプ化 A で検索を行なって，検索された文献量が少い場合にクランプ化 B を適用する方針が考えられる。

IV. 結　　語

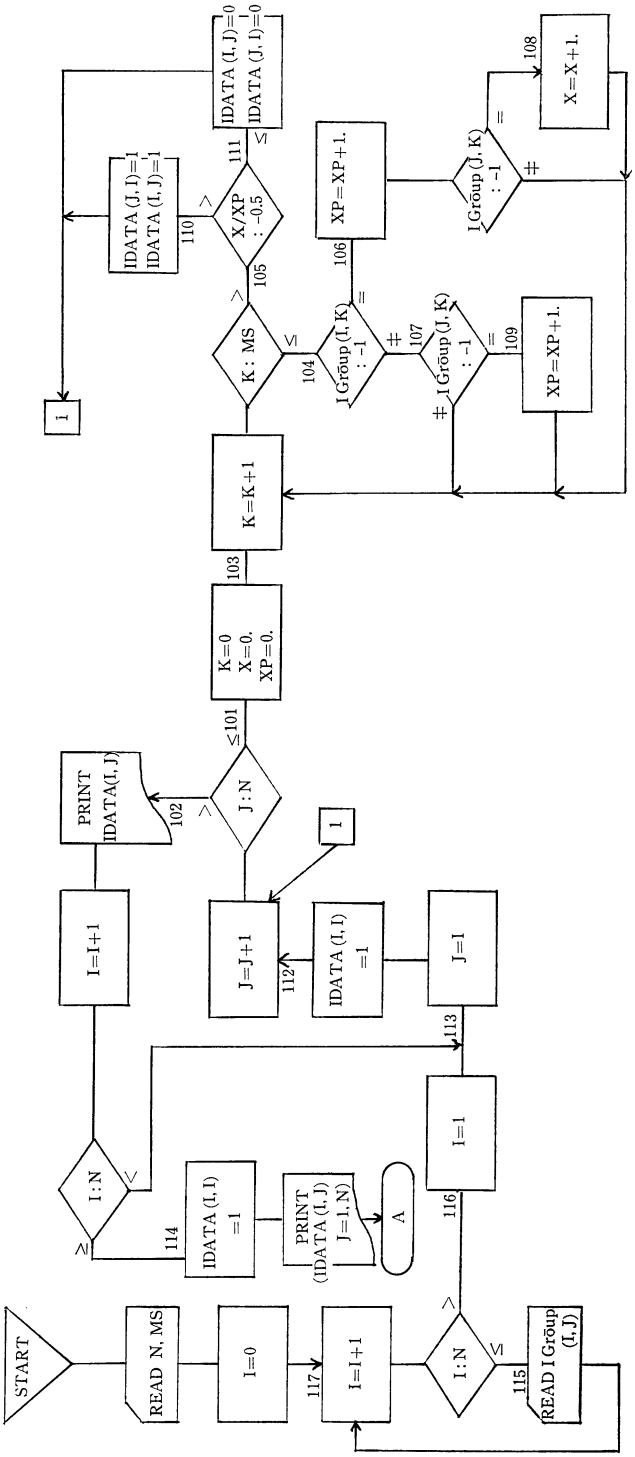
本稿では自動分類手法の特徴，目的，必須条件等について論じ，その条件を満足する手法の一つを提示した。しかしこの手法の大きな問題点は索引語の選択方法および評価と評価方法について考慮していないことである。

本来，分類はいくつかのより大きな段階から成り立っており，索引語の選択の問題，分類システムの評価の問題は分類手法の開発と同様に，大きな研究分野を構成する。もちろんこの二つの問題は分類手法と密接に結びつくのであるから，分類手法を考える際に無視することは出来ない。しかし分類の全ての段階を総合的に考えるアプローチももちろん必要であるが，相互の影響をなるべく少なくするような方法で各個撃破していくアプローチも必要である。この手法はその意味における一つの私案にすぎないが，この手法を使ってテスト実験をやることにより，上でふれた二つの問題に光をあてることが可能であろう。

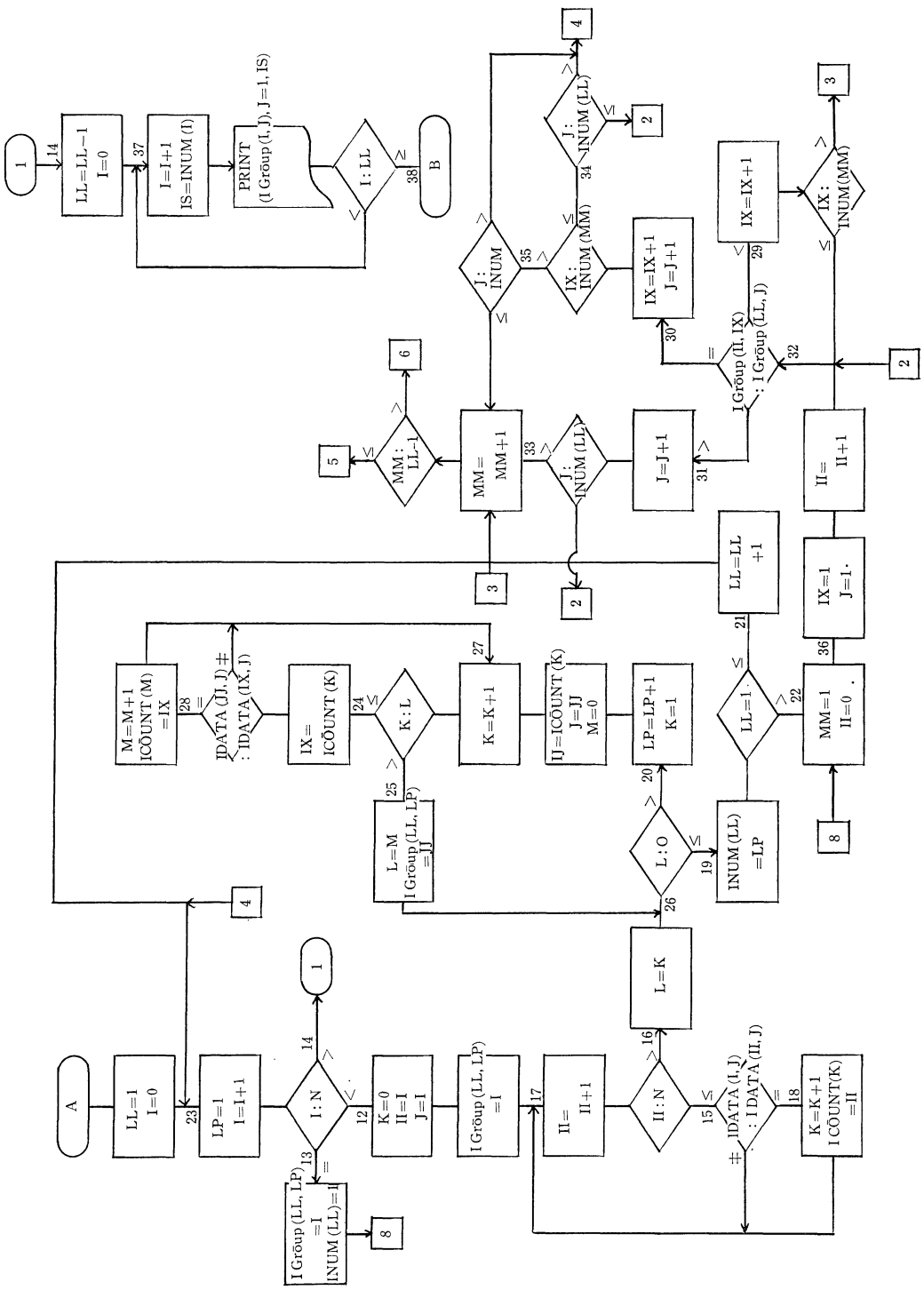
(図書館・情報学科)

1) 細野公男. "分類, 索引, 抄録間の関係: 自動処理の観点からみて," *Library and information science*, no. 6, 1968, p. 115–21.
2) Rogers, D. J. and Tanimoto, T. T. "A computer program for classifying plants," *Science*, vol. 132, Oct. 1960, p. 1115–8.
3) Lance, G. N. and Williams, W. T. "Computer programs for monothetic classification (Association analysis)," *The computer journal*, vol. 8, 1965, p. 246–9.
4) Lance, G. N. and Williams, W. T. "A general theory of classificatory sorting strategies 1. Hierarchical systems," *The computer journal*, vol. 9, no. 4, 1967, p. 373–80.
5) Macnaughton-Smith, P. "Dissimilarity analysis: a new technique of hierarchical sub-division," *Nature*, vol. 202, 1964, p. 426.
6) Lance, G. W. and Williams, W. T. "A general theory of classificatory sorting strategies II. Clustering systems," *The computer journal*, vol. 10, no. 3, 1967, p. 271–7.
7) Wallance, C. S. and Boulton, D. M. "An information measure for classification," *The computer journal*, vol. 11, no. 2, 1968, p. 185–95.
8) Goodall, D. W. "Classification, probability and utility," *Nature*, vol. 211, no. 5044, 1966, p. 53–4.
9) Williams, W. T. and Lance, G. W. "Logic of

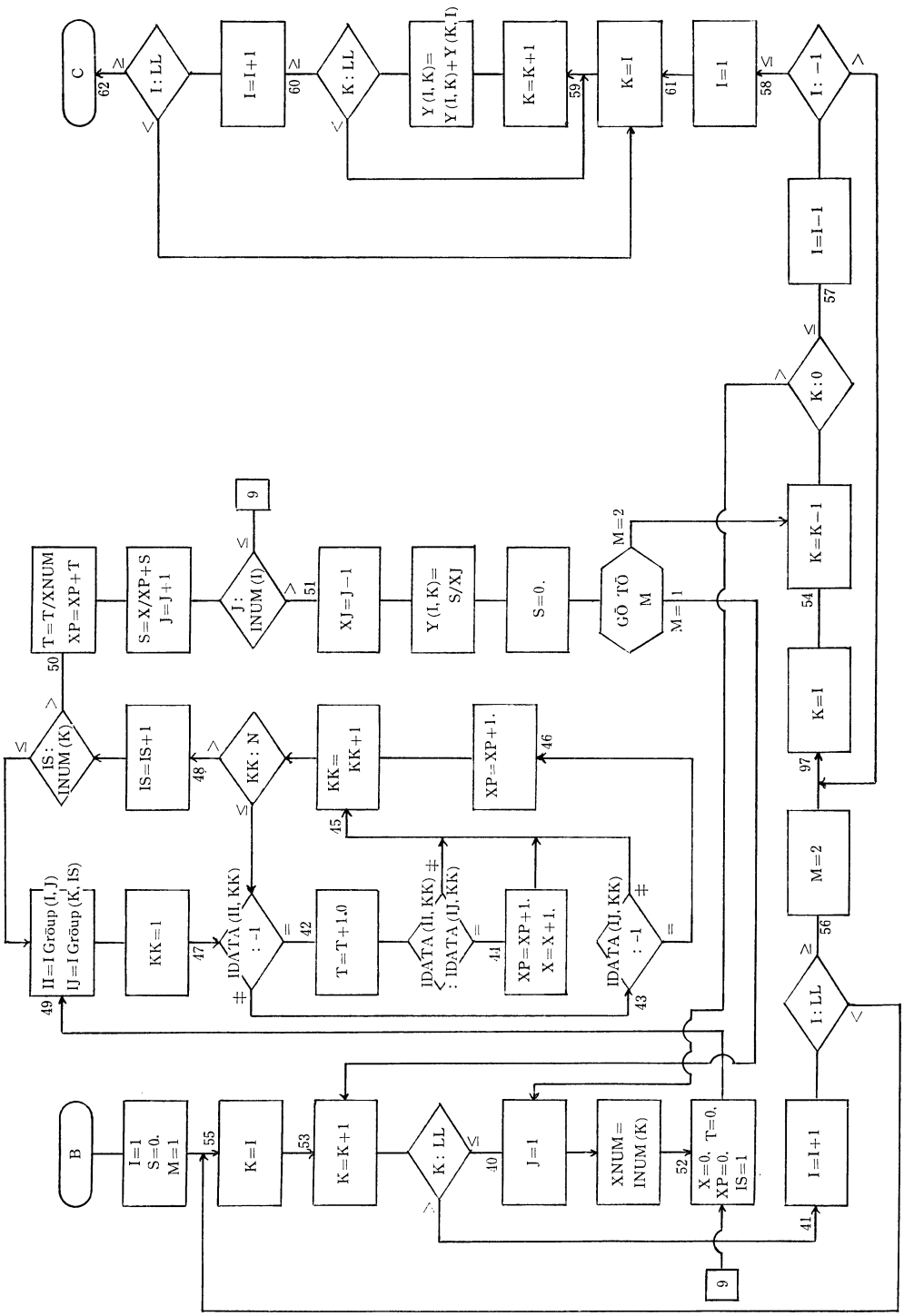
第 7 図

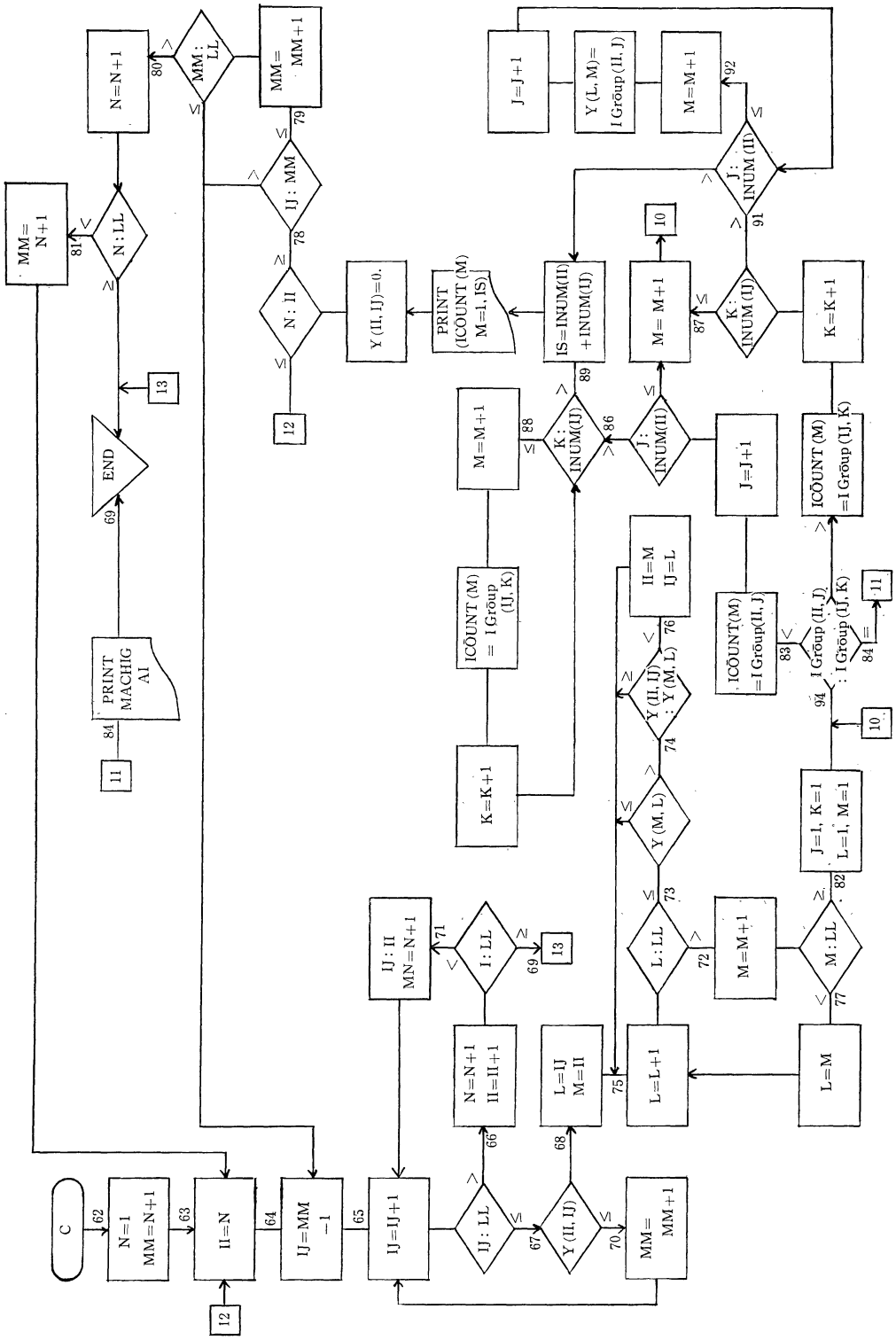

START
READ N, MS
I=0
117
I=I+1
I : N
>
≤
115
READ I Group (I, J)
116
I=1
113
J=1
IDATA (I, I) =1
J=I+1
112
1
PRINT IDATA(I, J)
102
J : N
>
≤101
K=0 X=0, XP=0.
103
K=K+1
K : MS
>
≤
104
I Group (I, K) : -1
=
≠
107
I Group (J, K) : -1
=
109
XP=XP+1.
106
XP=XP+1.
I Group (J, K) : -1
=
108
X=X+1.
≠
105
X/XP : -0.5
>
110
IDATA (J, I)=1 IDATA (I, J)=1
111
IDATA (I, J)=0 IDATA (J, I)=0
≤
1
I=I+1
I : N
≥
<
114
IDATA (I, I) =1
PRINT (IDATA (I, J) J=1, N)
A

第 8 図


A
LL=1
I=0
LP=1
I=I+1
I : N
I Group (LL, LP)
=I
INUM (LL)=1
K=0
II=I
J=I
I Group (LL, LP)
=I
II=
II+1
II : N
L=K
IDATA (I, J)
: IDATA (II, J)
K=K+1
I COUNT(K)
=II
L=M
I Group (LL, LP)
=JJ
L : O
INUM (LL)
=LP
LP=LP+1
K=1
IJ=ICOUNT (K)
J=JJ
M=0
K : L
IX=
ICOUNT (K)
IDATA (JJ, J) ≠
: IDATA (IX, J)
M=M+1
ICOUNT (M)
=IX
LL=LL
+1
LL=1
MM=1
II=0
IX=1
J=1
II=
II+1
MM:
LL-1
MM=
MM+1
J:
INUM (LL)
J=J+1
J:
INUM
IX:
INUM (MM)
IX=IX+1
J=J+1
I Group (II, IX)
: I Group (LL, J)
IX=IX+1
IX:
INUM(MM)
LL=LL-1
I=0
I=I+1
IS=INUM (I)
PRINT
(I Group (I, J), J=1, IS)
I : LL
B
1
2
3
4
5
6
8

第 9 図


B
I=1
S=0.
M=1
55
K=1
53
K=K+1
K : LL
40
J=1
XNUM=
INUM (K)
52
X=0.
T=0.
XP=0.
IS=1
9
41
I=I+1
I : LL
56
M=2
97
K=1
54
K=K-1
K : 0
57
I=I-1
I : -1
58
I=1
61
K=I
59
K=K+1
Y (I, K)=
Y (I, K)+Y (K, I)
K : LL
60
I=I+1
I : LL
62
C
GÖ TÖ
M
M=1
M=2
Y (I, K)=
S/XJ
S=0.
XJ=J-1
51
J :
INUM (I)
9
S=X/XP+S
J=J+1
T=T/XNUM
XP=XP+T
50
IS :
INUM (K)
IS=IS+1
48
KK : N
KK=
KK+1
XP=XP+1.
46
45
II=I Group (I, J)
IJ=I Group (K, IS)
49
KK=1
47
IDATA (II, KK)
: -1
42
T=T+1.0
IDATA (II, KK)
: IDATA (IJ, KK)
44
XP=XP+1.
X=X+1.
IDATA (IJ, KK)
: -1
43


C
62
N=1
MM=N+1
63
II=N
64
IJ=MM
−1
65
IJ=IJ+1
IJ : LL
>
≤
66
N=N+1
II=II+1
67
Y (II, IJ)
≤
68
L=IJ
M=II
70
MM=
MM+1
IJ : II
MN=N+1
71
I : LL
≥
69
13
12
L=L+1
75
L : LL
>
72
M=M+1
≤
73
Y (M, L)
>
≤
74
Y (II, IJ)
: Y (M, L)
≥
76
II=M
IJ=L
K=K+1
ICŌUNT (M)
= I Grōup
(IJ, K)
M=M+1
88
K :
INUM(IJ)
>
≤
89
IS=INUM(II)
+INUM(IJ)
PRINT
(ICŌUNT (M)
M=1, IS)
Y (II, IJ)=0.
N : II
≥
≤
78
IJ : MM
>
≤
79
MM=
MM+1
MM :
LL
>
≤
80
N=N+1
N : LL
<
≥
81
MM=
N+1
END
69
PRINT
MACHIG
AI
84
11
13
J : INUM (II)
>
≤
91
92
J=J+1
Y (L, M)=
I Grōup (II, J)
M=M+1
K :
INUM (II)
≤
87
M= M+1
10
K=K+1
J :
INUM(II)
>
≤
86
J=J+1
ICŌUNT(M)
=I Grōup(II, J)
83
<
I Grōup (II, J)
: I Grōup (IJ, K)
=
84
11
ICŌUNT (M)
=I Grōup (IJ, K)
94
10
J=1, K=1
L=1, M=1
82
M : LL
≥
<
77
L=M

第 10 図

computer-based intrinsic classifications," *Nature*, vol. 207, no. 4993, 1965, p. 159–61.
10) Sneath, P. H. A. "Some statistical problems in numerical taxonomy," *The statistician*, vol. 17, no. 1, 1967, p. 1–12.
11) Anderson, T. W. *Introduction to multivariate statistical analysis*. John Wiley, 1958, p. 126–9.
12) Jardine, N. and Sibson, R. "The construction of hierarchic and nonhierarchic classifications," *The computer journal*, vol. 11, no. 2, 1968, p. 177–84.
13) Needham, R. M. "A method for using computers in information classification, <International federation for information processing. *Information processing, 1962*> p. 284–7.
14) Needham, R. M. and Jones, K. S. "Keywords and clumps," *Journal of documentation*, vol. 20, no. 1, 1964, p. 5–15.
15) Needham, R. M. "Applications of the theory of clumps," *Mechanical translations*, vol. 8, 1965, p. 113–27.
16) Jones, K. S. and Jackson, D. "Current approaches to classification and clump-finding at the Cambridge Language Research Unit," *The computer journal*, vol. 10, no. 1, 1967, p. 29–37.
17) Jones, K. S. and Needham, R. M. "Automatic term classification and retrieval," *Information storage and retrieval*, vol. 4, no. 2, 1968, p. 91–100.
18) Dale, A. G. and Dale, N. "Some clumping experiments for associative document retrieval," *American documentation*," vol. 16, no. 1, 1965, p. 5–9.
19) Bonner, R.E. "On some clustering techniques," *IBM journal of research and development*, vol. 8, no. 1, 1964, p. 22–32.
20) Williams, J. H. "A discriminant method for automatically classifying documents," *Proceeding of the fall joint computer conference*, 1961, p. 161–6.
21) 山川邦雄. "文献分類の数量化の考察," **数理科学**, 3巻, 5号, 1965, p. 58–62.
22) Maron, M. E. "Automatic indexing: an experimental inquiry," *Journal of Association for Computing Machinery*, vol. 8, 1961, p. 407–17.
23) Borko, H. and Bernick, M. "Automatic document classification," *Journal of Association for Computing Machinery*, vol. 10, 1963, p. 151–62.
24) Baker, F. B. "Information retrieval based upon latent class analysis," *Journal of Association for Computing Machinery*, vol. 9, 1962, p. 512–21.
25) Constantinescu, P. "The classification of a set of elements with respect to a set of properties," *The computer journal*, vol. 8, no. 4, 1966, p. 352–7.
26) Edwards, A. W. F. and Cavalli-Sforza, L. L. "A method for cluster analysis," *Biometrics*, vol. 21, 1965, p. 362–75.
27) Johnson, S. C. "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, 1967, p. 241–54.
28) Lance, G. W. and Williams, W. T. "Ageneral theory of classificatory sorting strategies 1. Hierarchical systems," *op. cit.*
29) Jones, K. S. "Experiments in semantic classification," *Mechanical translation*, vol. 8, 1965, p. 97–112.
30) Jones, K. S. "Mechanized semantic classification," <1961 International conference on machine translation of languages and applied language analysis. *Proceedings*, vol. 1> p. 418–35.