Library and Information science No. 17 1979

引用傾向の類似性に基づく文献クラスタリングの一手法

A Clustering Method of Scientific Literature Based on Averaged Citation Multiplicity

宮本定明
Sadaaki Miyamoto
中山和彦
Kazuhiko Nakayama

Résumé

A Clustering method is proposed in order to form groups of articles in a specific discipline of science, with an expression of their inter-relationship. The technique is based on the similarity of citation, introducing a topological space which is called here "citation space": Each cited article forms one axis of the citation space and a source article is considered to be a point in this space. Then the measure of clustering is defined to be a scalar product for arbitrary pair of articles. Furthermore, any groups of source articles may be represented by a point in the space, not by a set of points, with each coordinate given by the citation probability of the corresponding axis of article in this group. On the other hand, the hierarchical clustering method imposes a restriction on the amount of data, due to memory requirement and the complexity of the resulting dendrogram. The above argument solves this problem by taking initial clusters for the hierarchical connection to be groups of articles instead of individual papers.

This method is applied to 3505 articles in instrumentation/control engineering extracted from *Science Citation Index* (1977) and 225 initial clusters are made by the source articles citing a specific article (axis). On the resulting dendrogram, groups formed above a specified similarity are summarized and named according to their contents, giving 25 reduced clusters hierarchically connected. The major part of the clusters shows theoretical development of the control engineering including 4 distinct groups of reseaches in USSR.

The essential difference between the Garfield's clustering and the present method is that the former is based on set-theoretical definition of the similarity measure, while the latter uses the notion of a topological space.

宮本定明：筑波大学学術情報処理センター
Sadaaki Miyamoto, Scientific Information Processing Center, University of Tsukuba.
中山和彦：筑波大学電子情報工学系教授
Kazuhiko Nakayama, Professor, Institute of Electronics and Information, University of Tsukuba.

引用傾向の類似性に基づく文献クラスタリングの一手法

I. はじめに

文献調査と統計的分析による研究動向把握の様々な試みのなかで，近年，データに含まれる構造をとり出し，幾何学的表現によって視覚的に明らかにしようとする研究が盛んになってきている。[1),2)] この傾向は，クラスター分析，多次元尺度構成法などに代表されるような，社会科学，行動科学における計量的方法がビブリオメトリックスの分野にも適用されるようになったものである。

特にクラスター分析は，計量的分類の有力な方法として，生物科学，パターン認識等の分野に広く適用され，興味ある結果をもたらしている。[3)]

文献データによる分類の試みとしては，Garfieldらによる *Science Citation Index* のクラスタリング[4)]がよく知られている。彼らの方法は，共引用頻度 (cocitation frequency) の概念に基づいていて，引用文献の任意の対が，それらを引用している文献において同時に引用されている回数によって，それらの間の類似度を定義する。これより類似度の大きいものから順次結合して，グループを形成する。要約すれば，引用文献どうしを，それらを引用している文献を介して，グループ化する方法である。

本稿では，これとは逆に，もとの文献を，引用文献を介してクラスター化する手法を提案する。この方法の特徴は，おのおのの引用文献が1つの軸を構成し，もとの文献がその中の一点として配置されるような空間（本稿ではこれを引用文献空間，あるいは引用空間とよぶ）の設定にある。クラスタリングの基準となるデータ間の類似度はこの空間のスカラー積で定義される。さらに，このスカラー積は任意の2つの文献グループにおいて定義され，2つのグループの間で，平均して何件の引用文献が共通に引用されているかを示すもので，本稿ではこれを平均引用重複度とよぶ。

第II章でこの方法の意義について述べた後，第III章では，*Science Citation Index* によって，実際にクラスタリングを行う。対象は計測制御工学関連文献3505件で，25のグループが抽出される。さらに，得られた結果について考察を行う。

II. クラスタリングの方法

A. クラスタリングにおける問題点

クラスタリングを行うにあたって考えるべき2つの重要な問題は，分類する要素間の近接の程度を表わす測度の定義と，要素どうしを結合しグループ化する方法の選択である。

前者については，扱うデータの性質を最もよく反映するものが選ばれなければならないし，またその選択がグループ化のアルゴリズムに大きく影響することに注意しなければならない。

一方，グループを形成するための方法には，大別して，階層的方法と非階層的方法の2つがある。[5)]

階層的方法は，扱うデータ集合中の任意の2つの要素間に定義された実数値をとる類似性の測度によって，類似性の大きい対から順にデータ対の結合を行い，生じたクラスター間の類似度を再定義することで，さらにグループ間の結合を進める。この結果は，データの結合の順序をすべて表わすため，階層的構造をもつ樹形図 (dendrogram) として示される。

この過程で重要なのは，2つのクラスターがそれぞれ複数の要素をもつとき，何によってクラスター間の類似度あるいは距離を定義するかである。もし，個々のデータが，あるユークリッド空間に配置されているならば，2つのクラスターの重心間の距離によって類似性の測度を定義することができる。一方，測度が任意の2要素の対

に対して定義されてはいるが，距離空間の存在を仮定しないならば，重心を考えることはできない。このような場合にも適用できる方法として有力なものは，2つのクラスターからそれぞれ1つずつ任意に要素を選んだとき，それの最短距離によってクラスター間の距離を定義する最短距離法，あるいは最長距離による方法であろう。[6]

階層的手法は樹形図によって，クラスターの関連について豊富な情報を与えるが，一方で，計算実行の際の主記憶量の限界などから，あまり大きなデータ集合に対しては適用が困難であるという欠点がある。これに対して非階層的方法のいくつかは，大量のデータを分類するのに適している。たとえば K-means 法や ISODATA 法に代表される反復法[7]は，あらかじめ生成されるクラスターの個数や大きさを指定しておいて，まずデータを分類するための種子（seed）となるいくつかの点を発生させて，データを最も近接した種子に所属させ，生じたグループの重心を再計算して新たな種子とすることをくり返す。このアルゴリズムによれば，データの量に対する制限ははるかに緩やかとなる。

後の方法は明らかに距離空間を前提としていて，かつ結果は樹形図のようにクラスター間の関連を完全に表わすことはできない。このことに注意すれば，考えるべき問題点は次の2つであろう。

(1) データがユークリッド空間あるいは，より一般的な位相空間に配置されていると考えるか，あるいはこのような空間を前提としないか。

(2) データ量を何らかの方法で制限することによって階層的方法を用いるか，あるいは，非階層的方法によって大量のデータを処理するか。

ここでは，樹形図によるクラスターの表現が，結果を視覚的に把握する上で有効と考え，階層的方法を用いて分析を行う。また，類似度は，以下に示すように，引用文献を軸とする空間上で定義される。

B. 引用空間と平均引用重複度

ある文献Aと別の文献Bが類似の引用傾向をもつというのは，どのようなときであろうか。AとBが同一の文献を1つだけ共通に引用しているならば，両者には何か関連があると考えられるが，その関係がきわめて密接とは判断できないであろう。しかしながら，AとBの共通引用文献数が2あるいは3ならば，それらにかなり強い関連があると考えられる。このように，共通引用文献数が増えるに従い，両者の関連はより密接となるので，引用の重複数（共通引用件数）を類似性の測度とする考えは有力であると思われる。

一方，文献のクラスタリングに際して，先に述べたような量の問題がある。通常の場合，階層的クラスタリングに適するデータ数はせいぜい数百であるが，後の例でもみられるように，特定分野の文献数は少なくとも数千のオーダに達すると考えなければならない。そこで，ここではまず文献を数百のグループに分類し，おのおののグループを1つのデータ単位のようにみなし，これらを初期クラスターとして階層的結合を行う。

このようにグループを対象とするとき，重複度の考えをそのまま利用できるであろうか。たとえば，初期クラスターCが100件の文献を含み，Dが1件を，またEが2件を含むとしょう。CとDの両方に共通に引用されている文献が5件，DとEの間に3件あったとする。このときCとDの共通引用数はDとEの共通引用数より大きいが，おのおのの含む文献数をみれば，明らかにDはCよりもEに近い。

この点を解決するため，平均化を行う。上の例でCの中の任意の文献は，Dと重複した文献を平均して5／100＝0.05件もっていると考える。一方，Eには 3／2＝1.5 件の共通な引用文献がある。したがって 0.05 と 1.5 を平均重複度として類似性の測度にとれば，DとEの近接性が表わされる。

この考えに従い，グループ間の平均引用重複度を計算するため，次のような引用文献空間を考える。

まず，対象とする文献の集合全体が引用しているすべての引用文献をとり出し，引用回数が一定値（たとえば2）以上のものを保存して，残りを捨てる。引用回数が1回限りの文献は重複度を考える上で必要でないから捨てられる。そこで，とり出された引用文献を 1,2,3,……, n としよう。

一方，文献グループA, B, C,…… を考え，それらに含まれる文献数を $\bar{a}$, $\bar{b}$, $\bar{c}$, ……, 文献 1,2,……, n を引用している回数をそれぞれ $(\bar{a}_1, \bar{a}_2, \cdots\cdots, \bar{a}_n)$, $(\bar{b}_1, \bar{b}_2, \cdots\cdots, \bar{b}_n)$, $(\bar{c}_1, \bar{c}_2, \cdots\cdots, \bar{c}_n)$, ……とする。これより，文献 1,2,……, n の平均引用件数を表わすベクトルを，グループA, B,……に対して

$$a=(a_1, a_2, \cdots\cdots, a_n), \quad a_i=\frac{\bar{a}_i}{\bar{a}}$$

$$b=(b_1, b_2, \cdots\cdots, bn), \quad b_i=\frac{\bar{b}_i}{\bar{b}}$$

$$\cdots\cdots\cdots\cdots$$

$$1 \leqq i \leqq n$$

によって定義する。このように平均引用件数を表わすベクトルは，各引用文献を1つの軸とする空間（引用文献空間）の一点とみなされる。

上の例でAが文献1を引用している平均件数は a_1，Bでは b_1，であり，一般に $0 \leqq a_i \leqq 1$，$0 \leqq b_i \leqq 1$ であるから，これらを文献の引用確率とみなす。このとき $a_1 b_1$ はA，Bが同時に文献1を引用している確率とみることができるから，[8] これが文献1に関するAとBの類似度を表わすと考える。

そこで，同様のことを文献 2,3,……,n について考えそれぞれの文献における類似度を加えあわせて，引用文献空間における平均引用重複度 $m(\mathrm{A},\mathrm{B})$ を定義する。

$$m(\mathrm{A},\mathrm{B}) = a_1 b_1 + a_2 b_2 + \cdots\cdots + a_n b_n$$

$m(\mathrm{A},\mathrm{B})$ は引用文献空間に自乗ユークリッド距離を与えたときの内積に他ならない。

このことを具体例を用いて説明しよう。

文献グループAには3件の文献が属し，グループBには2件が含まれているとする。引用文献が ①，②，③，④，⑤ の5種類に限られると仮定する。Aが①を3回，②を2回，③を0回，④を1回，⑤を1回引用していることは，上の議論によって

$$\bar{a} = (3,\ 2,\ 0,\ 1,\ 1)$$

とかかれる。

Bについては，

$$\bar{b} = (2,\ 0,\ 0,\ 2,\ 1)$$

であるとする。

文献①については，Aの3件の文献が3回引用しているので，Aの個々の文献は平均して1回引用しているとみられる。一方，①はBについても平均2／2＝1回ずつ引用されている。同様に④ついてはAの文献は平均1／3回引用し，Bは1回引用している。

このことから①について，AとBにおける平均化された文献の1対には $1\times1=1$ 回の共通引用件数があると考えられる。同様に④については $(1/3)\times1=1/3$ 回と考える。

引用空間のすべての文献に関する平均共通引用件数（重複度）は，個々の文献に関する平均共通引用数を加えて定義するのが自然であるから，この場合

$$m(\mathrm{A},\mathrm{B}) = 1\times1 + \frac{1}{3}\times1 + \frac{1}{3}\times\frac{1}{2} = \frac{3}{2}$$

すなわち，正規化されたベクトル

$$a = \left(1, \frac{2}{3}, 0, \frac{1}{3}, \frac{1}{3}\right)$$

$$b = \left(1, 0, 0, 1, \frac{1}{2}\right)$$

の内積で与えられる。

III. 計測制御工学関連文献のクラスタリング

分析の対象は，Institute for Scientific Information (ISI) の提供している *Science Citation Index* (*SCI*) 1977年のファイルである。

ISI の，「収録誌マスターファイル」によれば，全収録誌は52の分野に分類される。本研究で対象としたのは，このうちの計測および制御工学 (Instrumentation and Control Engineering) であり，*SCI* には 34 誌が収録されている。

第1表は収録誌名およびそれぞれの収録件数を示している。抽出された文献の総数は3505件であり，それらが引用している文献の総数は26096件であった。

これより，前章の方法に従ってクラスタリングを行うため，3回以上引用されている引用文献1384件をとり出した。これらのおのおのが引用空間の1つの軸を構成していると考える。

階層的手法への入力となる初期クラスターの数としては，主に計算の際の記憶容量の制限から，200～300程度にとどめるのが望ましい。ここでは5回以上引用されている引用文献 225 件をとり，それぞれを引用している文献を個別に抽出して，225のグループとした。これより，初期クラスターは，それぞれのグループを代表するように選ばれる。すなわち，これらのグループの重心（平均引用件数を表わすベクトル）を計算し，225の初期クラスターとする。

任意の2つの初期クラスター間の内積をとれば，類似度行列が得られる。これから，最短距離法によって階層的クラスタリングを行った。

一般に，階層的クラスタリングの結果は初期クラスターが統合されていく順序を示す樹形図に表わされるが，この樹形図に含まれているすべての情報，すなわち，結合の各段階で形成された文献グループの属性，著者，表題，引用文献などをすべて表示するのは，実際上不可能であるし，結果の解釈も困難である。

そこで，ここでは，Garfield らが行った[9] のと同様に，比較的緊密にまとまったグループに対し名前をつけて，その内容を明らかにする。すなわち，（類似度）≧2.0 の段階で2つ以上の初期クラスターが結合してできたグループにつき，文献の表題を抽出して，適当と判

第1表 *SCI* (1977) より抽出された計測制御工学関連雑誌

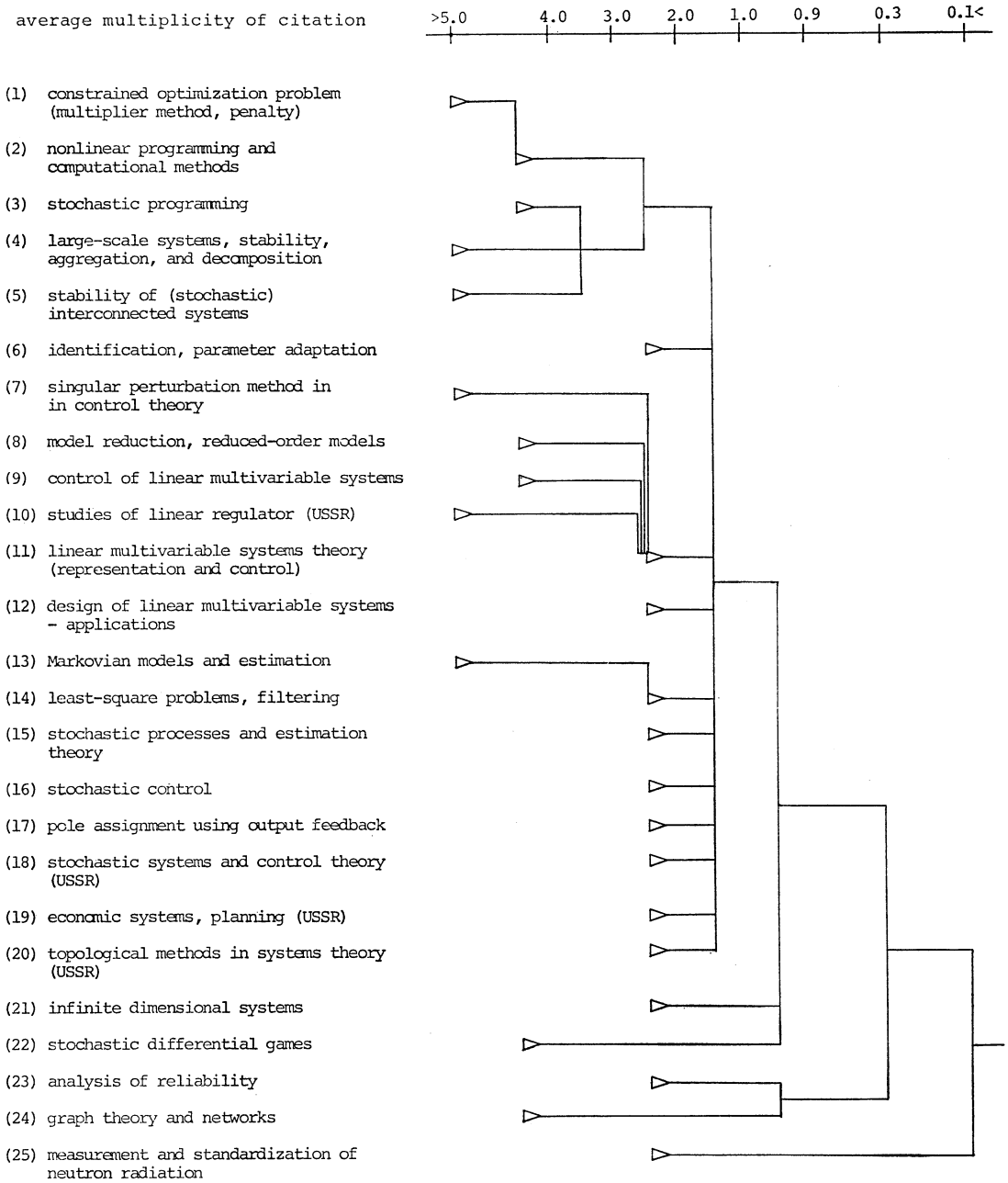
雑誌名	論文数
Applied Spectroscopy Reviews	9
Applied Spectroscopy	127
ATM Messtechnische Praxis	22
Australian Journal of Instrumentation & Control	14
Automation and Remote Control USSR	273
Automatica	43
Automatisme	47
Biomedical Engineering	64
Chemical Instrumentation	18
Control and Instrumentation	74
Control Engineering	98
F&M-Feinwerktechnik & Messtechnik	79
IEEE Transactions on Automatic Control	290
IEEE Transactions on Biomedical Engineering	85
IEEE Transactions on Industrial Electronics and Control Instrumentation	103
IEEE Transactions on Instrumentation and Measurement	61
IEEE Transactions on Reliability	124
IEEE Transactions on Vehicular Technology	27
Instruments & Control Systems	76
Instruments and Experimental Techniques	189
Instrumentation and Control	62
International Journal of Control	159
ISA Transactions	56
Measurement and Control	59
Measurement Techniques USSR	565
Medical & Biomedical Engineering	133
Metrologia	26
Microtecnic	17
NDT International	14
Non-Destructive Testing	13
Proceeding Annual Reliability and Maintainability Symposium	89
Review of Scientific Instruments	378
SIAM Journal of Control	71
Technisches Messen ATM	40

断される名前をつけ，樹形図を簡単化する。

第1図はこのようにして得られた樹形図であり，25のグループおよびそれらの結合の状態を表わしている。また，それぞれのグループの核となった引用文献の著者とグループ内のメンバー数を第2表に示す。表中で，著者の後の括弧中の数字は引用文献の発行年度であるが，発行年の記述が不完全なデータが若干存在するので，一部表示を省略している。第2表のクラスター番号は第1図に対応していて，著者のグループによる研究とクラスターの表題との一致性を確認することが可能である。

第1図において，あるクラスターが別のクラスターに接続されているのは（第2図 *a* 参照） いくつかの初期クラスターがある類似度で結合しているところへ，さらにいくつかの初期クラスターが集まって，より低い類似度で結合し，前者を含む別のクラスターを形成していること，すなわち第2図 *a* のようにクラスターAがクラスターBの部分クラスターとなる場合である。第1図のクラスター（1，2），（7，8，9，10，11）および（13，14）はこれにあたる。第3図 *a* に，実際に計算機により出力された樹形図の一部を示す。

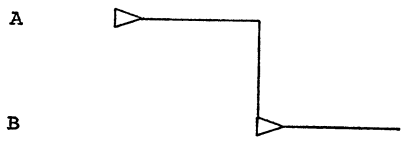
これに対して，第2図 *b* のような場合は，単にクラスターCとクラスターDの結合を示している。第3図 *b* は



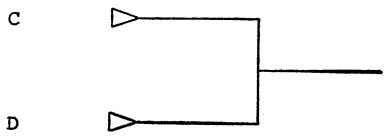
第1図　階層的クラスタリングによる25のグループ

クラスター番号	クラスターの核となった著者名	クラスター内の文献数
1.	Bertsekas D.P. (1973,1975) Fiacco A,V. (1968) Hestenes M.R. (1969) Kort B.W. (1973) Powell M.J.D. (1969) Rockafellar R.T.	15
2.	Luenberger D.G. (1969,1973) Polak E. (1971) Polyak B.T. (1972) Rockafellar R.T. (1970) + 1.	45
3.	Danzig G.B. (1963,1955) Walkup D.W. (1967)	7
4.	Siljak D.D.(1972,1974,1975,1976)	7
5.	Fiedler M. (1962) Kushner H. (1971) Araki M. (1972) Bailey F.N. (1966) Michel A.N. (1972) Porter D.W. (1974)	14
6.	Carrol R.L. (1973) Hang C.C. (1973) Landau I.D. (1974) Kudva P. (1973) Narendra K.S. (1973) Young P.C. (1970)	15
7.	Chang K.W. (1972) Collins W.D. (1973) Freeman M.I. Haddad A.H. (1971) Omalley R.E. (1972,1974) Kokotovic P.V. (1972) Wilde R.R. (1973)	14
8.	Bosley M.J. (1972) Chen C.P. (1968,1970,1974) Danati F. (1970) Hutton M.F. (1975) Shamash Y. (1975)	19
9.	Davison E.J. (1972,1973) Grasselli O.M. (1973) Pearson J.B. (1969)	15
10.	Krasovskii N.N.(1969) Letov A.M. (1960)	6
11.	Brasch F.M. (1970) Brunovsky P. (1970) Coddington E.A. (1955) Desoer C.A. (1075) Davison E.J. (1966,1974) Johnson C.D. (1971) Morse A.S. (1973) Popov V.M. (1970) Wang S.H. (1971) Wonham W.M. (1972) +7.+8.+9.+10	112
12.	Belletrutti J.J. (1971) Bode H.W. (1945) Rosenbrock H.H. (1969)	13
13.	Casti J. (1974) Kailath T. (1973) Reid W.T. (1972) Lainiotis D.G. (1975)	8
14.	Kalman R.E. (1961) Kleinmann D.L.(1968) Ljung L. (1976) + 13.	24
15.	Bucy R.S. (1968) Caines P.E. Doob J.L. (1953) Ljung L. (1974) Wonham W.M.(1968)	18
16.	Astrom K.J. (1970) Meier L. (1971) Pearson J.D. (1971)	11
17.	Davison E.J. (1970,1971) Fallside F. (1971) Luenberger D.G. (1967) Seraji H. (1973,1975)	21
18.	Butkovskii A.G.(1965) Kazakov V.S. (1962) Pugachev V.S. (1962,1974)	13
19.	Dubovskii S.V. (1972,1973) Dyukalov A.N. (1973,1974)	14
20.	Nikaido H. (1972) Opoitsev V.I. (1974)	8
21.	Bensoussan A. (1971) Curtain R.F. (1975) Hille E. (1957) Lions J.L. (1971)	12
22.	Benes V.E. (1971) Davis M.H.A. (1973)	5
23.	Barlow R.E. (1965,1975) Mann N.R. (1974) Proschan F. (1974)	19
24.	Fratta L. (1973) Jensen P.A. (1969)	6
25.	Okeirimmarkus I.B. (1974) Yudin M.F. (1972)	7

第2表 クラスター（第1図）の核となった著者およびクラスター内の文献数



第2図 a　部分クラスターの表示



第2図 b　クラスターの結合

```
┌┌LAINIOTIS DG 1975    24  ---I
│└REID WT 1972        198  ---I
13 KAILATH T 1973      175  ---I------------------I
└─CASTI J 1974         146  ---I                  I-------I
14 KALMAN RE 1961       31  ----------------------I       I---I
  KLEINMANN DL 1968     46  ------------------------------I   I-----
└─LJUNG L 1976         180  ----------------------------------I
```

第3図 a　部分クラスターの例

```
┌─DANZIG GB 1955       153  ----------------------I---I
3 WALKUP DW 1967       218  ----------------------I   I---I
└─DANZIG GB 1963        94  --------------------------I   I
┌─SILJAK DD 1975         76  ---I                        I
4 SILJAK DD 1974       213  ---I-------I                 I
│ SILJAK DD 1976       124  ---I       I-----------------I-----
└─SILJAK DD 1972       215  -----------I                 I
┌─BAILEY FN 1966       132  ---------------I             I
│ MICHEL AN 1972       191  ---------------I-------I     I
5 ARAKI M  1972        135  ---------------I       I     I
│ KUSHNER H 1971       183  ----------------------I     I
│ FIEDLER M 1962        65  ----------------------I-------I
└─PORTER DW 1974        53  ----------------------I
```

第3図 b　クラスター(3), (4), (5)

注.　著者名，年度の後の数字は初期クラスターの番号を表わす.

第1図のクラスター 3，4，5 である。第3図 a との相違に注意しよう。

IV. 考　　察

A. クラスタリング結果に関して

第1図，第2表にはソ連邦における研究が4つのグループとして現われていることがまず認められる。これらのグループの核のうちあきらかにソ連邦以外の著者と認められるのは1人にすぎない。

わが国における計測制御工学研究の現状をこの結果と比較するため，1978年第21回自動制御連合講演会における講演件数との比較をしてみる。まず大きな分類として次のものをあげる。[10] 項目の後の括弧内の数字は上記講演会での発表件数であり，ダッシュの後の数字は第1図のクラスター番号を表わす。

(1) 理論（103件）——1，2，3，4，5，6，7，8，9，10，11，12，13，14，15，16，17，18，20，21，22，23，24

(2) 制御要素，機器（8件）——なし

(3) 応用（107件）——19（経済計画）

(4) 計測（27件）——25（中性子線の計測）

このように，理論に関するものが大半を占めている。さらに，理論に関して細分をすれば[11]

(a) システム理論（20件）
- 8（モデルの簡単化）
- 9（線形多変数制御系）
- 12（線形多変数システム理論）
- 17（出力フィードバックによる極の移動）
- 20（システム理論と位相的方法）

(b) レギュレータ問題（4件）
- 10（線形レギュレータ）

(c) 制御系設計（5件）
- 12（線形多変数系の設計）

(d) 安定論（4件）
- 5（確率安定性その他）

(e) 大規模系（6件）

4 （安定性，分解，統合）

(f) 最適制御（10件）

16 （確率制御）

18 （ソ連邦における制御理論，確率系）

(g) 数理計画（6件）

1 （制約付き最適問題）

2 （非線形計画と計算法）

3 （確率計画法）

(h) 推定（6件）

13 （マルコフモデルと推定）

14 （最小自乗法）

15 （確率過程と推定）

(i) 同定（8件）

6 （パラメータ同定）

(j) 分布系（6件）

21 （関数空間論による分布定数系）

(k) その他

7 （特異摂動）

22 （確率微分ゲーム）

23 （信頼性理論）

24 （グラフ理論，ネットワーク）

と対応する。ここで（23）信頼性理論，（24）ネットワークは，一般にシステム理論，制御理論と呼ばれているものとはいささか異なり，クラスター上でも他と離れて表わされている。

さらに，わが国での研究が比較的盛んであるのに対し，クラスターにはあらわれていないものとして，理論では，

(l) むだ時間系（4件）

(m) 学習理論（3件）

(n) ファジィ理論（3件）

(o) 適応制御（8件）

(p) オブザーバ（10件）

があげられる。特に，適応制御，オブザーバについては学会での発表件数も多いが，第1図では，（6）パラメータ同定，と関連があるだけで，あまり顕著でない。このように理論部門においては，わが国での研究傾向と欧米のそれに差が認められる。

これに対して，応用については，一般に論文の発表される雑誌も様々で，計測制御工学以外の分野にも広がっていると思われるので，第1図から応用に関する研究がこの分野において盛んでないとは必ずしも結論できないように思われる。ただ，わが国においては，応用部門が全体の半数近くを占め，その範囲が，環境，生体，社会システム，輸送，計算機応用など多岐にわたっていることから，[12] その学際的傾向を指摘することができよう。

B. クラスタリング手法に関して

本稿で提案したクラスタリング手法の特徴として，

(1) 平均引用重複度による引用傾向の類似度の定義。

(2) 初期クラスターとして，個々の文献をとるかわりに，文献グループを用いることにより，数を減少させ，大量のデータ処理を可能にした。

の2つがあげられる。後者について，第Ⅲ章では，引用回数の多い文献を引用しているものを初期クラスターとしたが，このほか，様々なグループ化が可能である。

たとえば，キーワードなど特定の情報をもつものを初期クラスターとして階層的結合を行えば，引用関係からみたキーワード間の関連が明らかになるであろう。また，階層的クラスタリングのための初期クラスターは非階層的クラスタリングを用いて発生させることも可能である。

階層的統合の方法として，ここでは最短距離法（single linkage method）を用いた。この方法はアルゴリズムが簡単で，枝の交叉[13]が生じないなど好ましい性質を備え，最も広く用いられている。他に，最長距離法（complete linkage method）や重心法（centroid method）なども適用できる。重心法を用いる場合，平均引用重複度の再計算をくり返し行う必要があり，かつ枝の交叉の可能性がある。

クラスターの表示については，前章で述べたように，初期クラスターすべてを表示するのではなく，初期クラスターが結合してできたグループに名前をつけて表示した。分析する文献数および初期クラスターの数からみて，このような操作は結果の理解を容易にするため必要であると考える。しかしながら，ここには複数の解釈の余地があり，機械的にはできないので，十分注意する必要があろう。なお，Garfieldを中心として行われたクラスタリング[14]でもこの部分は人手により行われている。

Garfield らの方法と本稿での試みの本質的な相違点として，類似性の測度の性質の違いがあげられる。Garfield らによれば，ある一対の引用文献の類似度は，それらを引用している2つの文献集合の交わりにおける要素の数で定義されるので，いわば集合論的であるが，これに対して，ここでは，引用空間の座標で文献グループの属性を規定し，内積によって測度を与えるので，位相空間に基づいているといえよう。

V. おわりに

本稿では，引用空間の設定により文献集合上に類似性の測度を誘導し，階層的クラスタリングを行った。

文献検索には筑波大学学術情報処理センターで使用されているデータベースマネジメントシステム IDEAS/77 を用い，処理は同センターの大型計算機 ACOS800/II を使用した。なおクラスタリングには，Anderberg によるプログラム[15]を一部に用いた。

本手法については，測度の定義，階層的結合法，クラスターの表示などについて様々な変形が可能で柔軟性があり，かつアルゴリズムも比較的簡単であるので，研究動向把握のための道具として十分利用できると思われる。

最後に，本稿作成にあたって，数々の有益な助言を頂いた筑波大学学術情報処理センター上田修一氏に心から謝意を表する。

1) Garfield, E. "Historiographs, librarianship, and the history of science," *Toward a theory of librarianship:* papers in honor of Jesse Hauk Shera, ed. by Conrad H. Rawski. Metuchen, N. J., Scarecrow Press, 1973, p. 380–402.
2) Narin, F., Pinski, G., and Gee, H. H. "Structure of the biomedical literature," *Journal of the American Society for Information Science*, vol. 27, no. 1, Jan.–Feb. 1976, p. 25–45.
3) "クラスター分析特集号" 数理科学, no. 190, April, 1979.
4) Garfield, E., Malin, M. V., and Small, H. "A system for automatic classification of scientific literature," *Journal of the Indian Institute of Science*, vol. 57, no. 2. 1975, p. 61–74.
5) Anderberg, M. R. *Cluster analysis for applications.* Academic Press, New York, 1973.
6) *Ibid.*, p. 137–9.
7) *Ibid.*, p. 162–73.
8) 厳密には，AとBでの文献の引用が独立であると仮定しなければならない。ここでの議論は発見的である。
9) Garfield, *et al.*, *op. cit.*
10) システムと制御. vol. 22, no. 9, 1978, p. 3.
11) *Ibid.*, p. 4–9.
12) *Ibid.*
13) Anderberg, *op. cit.*, p. 142.
14) Garfieled, *et al.*, *op. cit.*
15) Anderberg, *op. cit.*, p. 278–90.