2部グラフの分割理論を利用した概念構造決定法

1. はじめに

言葉によって表現される意味や概念を，計算機を使って自動的に抽出するなどということは，無謀な試みに思えるかも知れない．しかし，人間の行なう代表的な意味抽出作業である辞書づくりについて，言語学者[3]はつぎのように言っている．「辞書をつくるということは語の真の意味についての権威的な叙述を打立てる仕事ではなくて，種種の語が遠いまたは近い過去に著者達にとって，どんな意味であったかを記録する仕事である．」すなわち，意味の抽出とは言葉の“使われ方”のデータを客観的に整理することである．また，客観的データ処理こそ計算機の得意とする仕事である．したがって，計算機による意味抽出が無謀にみえるとしたら，それは計算機を使うこと自身が無謀なためではなく，意味のように量よりも質のほうが本質的と考えられるデータを扱う手法が，まだ充分開発されていないためであると思われる．以下に紹介するのは，言葉の使われ方のデータから意味あるいは概念の構造と見なし得るものを抽出するための新しい数理的一手法である．

意味の構造を数学的に抽出しようという試みはすでに多くなされている．それらは，因子分析法を利用して意味を多次元空間の点として把えようというもの[10]，類似度解析法を利用して意味と意味の近さを計量化しようというもの[7]，[8]，[9]等，多変量解析という名で総称される手法のどれかを言語データに適用したものがほとんどである．そこでは，言葉の使われ方のデータが，なんらかの形で量的データに変換されて処理されている．意味の間に得られる“構造”も“多次元空間の配置”，“相互距離”，あるいはそれからクラスタ分析的手法により得られる“無向グラフ構造”のようなものである．ここで述べる方法は，それらとは対照的に，組合せ論的な処理によって質的なデータを（計量化することなく）質的なまま扱い，その結果，概念の半順序構造（有向グラフ構造）を決定するというものである．

2. 記号・対象・概念

図1に示されるようなものを名指すために，人はいろいろな言葉を使う．「犬」，「ペット」，「生き物」などと名指すこともあるだろうし，時には「ワンチャン」，「漫画」などと名指すかも知れない．図1のようなものを対象とよび，「犬」，「ペット」，「生き物」等の言葉を記号とよぶことにする．

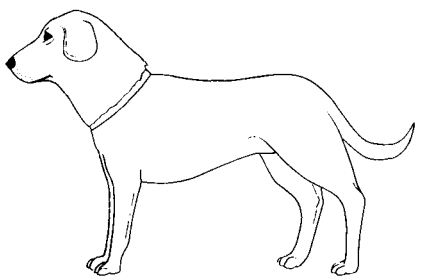

図1 対象の例

さて，ある対象をある記号で名指すという形の言語行動データが与えられたとしよう．同一の対象を名指すのにいろいろな記号が使われ（対象の多面性），同一の記号がまたいろいろな対象を名

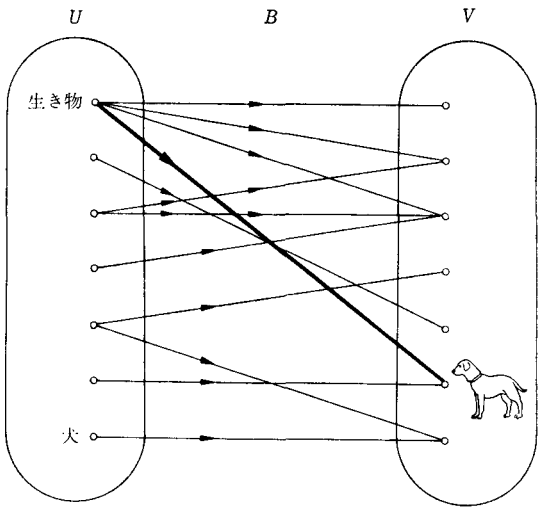

図 2 記号と対象の2部グラフ

指すために使われる（記号の多義性）ため，記号と対象の対応は一般に多対多となり，**図 2** のようなデータが得られる．図中，Uは記号の集合，Vは対象の集合，Bは記号とそれで名指すことのできる対象とをつなぐ枝の集合である．2個の点集合U, Vと両者をつなぐ枝集合Bで定義されるグラフ$(U, V; B)$は，2部グラフとよばれる．記号と対象を点とする図2のような2部グラフは，言語行動データのもっとも単純な一表現形式である．

このような2部グラフは単に記号と対象の対応関係を列挙したものにすぎない．

そこで，記号集合Uと対象集合Vの間に新しい集合Wを挿入し，**図 3** のような形に整理することによって，対応関係を構造化してみようというのが，ここで述べるデータ処理法の要点である．すなわち，

1) 新しい集合Wの元の間にはある半順序関係が存在する（この半順序関係を，元と元を結ぶ有向枝で示す．このときWは無サイクルな有向グラフとなる），

2) Uの各元はただ1個のWの元に対応する（この対応を有向枝の集合B^+で示す），

3) Vの各元はただ1個のWの元に対応する（この対応を有向枝の集合B^-で示す），

4) 2部グラフの各枝は，B^+の枝とW内での有向道とB^-の枝とをつなぎ合わせることによって再現できる（たとえば，図2の太線の枝は図3の太線の有向道で再現される），

の4条件を満たすように，半順序集合Wと対応B^+, B^-をみつけようというものである．このようなW, B^+, B^-の決め方は次節で論ずることにして，ここでは，図3のような形にデータを整理することの意味について考えてみよう．

記号を使って対象を名指すとき，記号によって表現されているのは，対象そのものではなく，対象のある側面である．たとえば，2個の記号「犬」と「生き物」が図1の対象を名指すために使われるのは，その2個の記号が対象のもつ異なる性質（これを概念とよぶことにしよう）をあらわすためだと考えられる．すなわち，人が対象を記号で名指すのは，まず対象からある概念を抽出し，つぎにその概念をあらわすために記号を選ぶという2段階の行為の結果であるとみなすことができる．したがって，記号と対象の関係を，記号と概念の

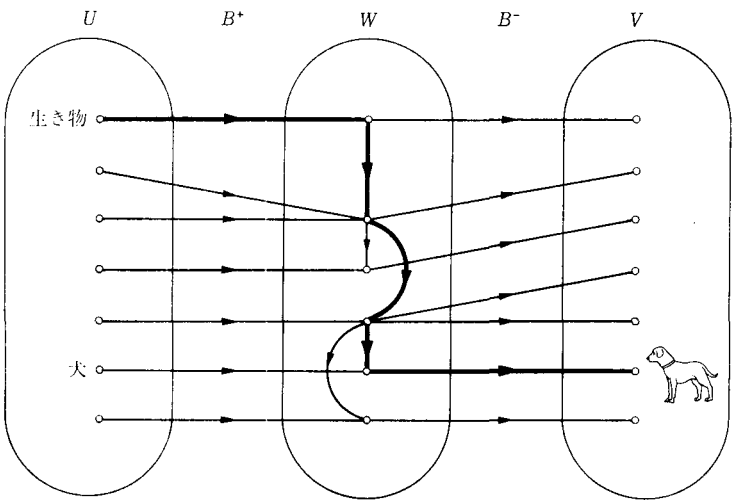

図 3 記号と対象の間に介在する概念

関係および概念と対象の関係に分解することが，意味をもつことになろう．これが，新集合 W を U と V の間に介在させる意味である．

概念の間には，半順序関係を考えることができる．たとえば，「生き物」という記号であらわされる概念は，「犬」という記号であらわされる概念を包含するより広い概念であると考えられる．このように，ある概念がもう一つの概念を包含するとき，前者を後者の**上位概念**，後者を前者の**下位概念**とよぶことにする．（このような上位・下位関係を整理する作業は，古くからなされている[11].）各概念を点とみなし，上位・下位関係にある2個の概念を上位概念を始点とし下位概念を終点とする枝で結んでグラフをつくることができる．上位・下位関係に3すくみ(一般には n すくみ)が生じては都合が悪いから，そのようなグラフは無サイクルでなければならない．このことは，上位・下位関係が半順序関係であるという要請と等価である．これが，集合 W に半順序構造を仮定する意味である．

さて，条件1)，すなわち半順序構造をもった概念集合 W を記号と対象との間におくことを認めれば，他の条件2)，3)，4）はつぎのように解釈できる．すなわち，条件2)，3）はそれぞれ記号と対象を一つの概念と対応させようというものである．条件4）は，ある記号がある対象を名指すのはその記号の属する概念がその対象の属する概念と一致するかあるいは前者が後者の上位概念になっているときに限るという要請である．

3. 概念構造決定法

2部グラフ $(U, V; B)$ の枝の部分集合 M は，M に属す枝の端点が重複しないとき**マッチング**とよばれる．このとき，M に属す枝の端点となっている頂点は，M によって**飽和されている**という．とくに，$|M|=|U|=|V|$ が成り立つとき，M を**完全マッチング**とよぶ．

さて，前節で要請した4個の条件のみでは，集合 W と対応 B^+, B^- は一意には決まらない．しかし，図2のような2部グラフが勝手に与えられたとき，それを前節の4条件は少なくとも満たすように分解でき，しかもその分解は下記の意味で一意的であるということが，組合せ数学の分野で知られている．すなわち，任意の2部グラフ $(U, V; B)$ に対して，U および V の分割 $\{U_1, U_2, \cdots, U_M\}$, $\{V_1, V_2, \cdots, V_M\}$ $(U_i \cap U_j = V_i \cap V_j = \phi\ (i \neq j);\ U = \bigcup_i U_i;\ V = \bigcup_i V_i)$ で，以下の性質を満たすもっとも細かいものが一意に決まる（$B_i = B \cap (U_i \times V_i)$ とおく）：

i） 半順序集合 $W=\{w_1, w_2, \cdots, w_M\}$ が存在する（半順序関係を $\geqq$ とする）；

ii） w_i が W の極大元でも極小元でもなければ $|U_i|=|V_i|$ で，かつ，B_i の任意の枝に対して，2部グラフ $(U_i, V_i; B_i)$ にはその枝を含む完全マッチングが存在する；w_i が W の極大［小］元であれば，$|U_i| \leqq |V_i|$ ［$|U_i| \geqq |V_i|$］で，かつ，B_i の任意の枝に対して，2部グラフ $(U_i, V_i; B_i)$ にはその辺を含み U_i［V_i］の頂点をすべて飽和させるマッチングが存在する；

iii） $(U, V; B)$ において $a \in U_i$, $b \in V_j$, $(a, b) \in B$ であるのは，W において $w_i \geqq w_j$ であるときに限る．また，W において w_i が w_j の直上元であれば，$(a, b) \in B$ であるような $a \in U_i$ と $b \in V_j$ が存在する．

上の定理は，Mendelsohn と Dulmage［1］，［2］によるものであるが，［5］の第6章にそのくわしい初等的な解説がある．また上記の分割を実際に行ない W の半順序構造を定めるための効率のよいアルゴリズム（$|U|$ や $|V|$ の3乗に比例する程度の手間で済むもの）も知られている［5］.

このように，2部グラフが自然にしかも一意的に分割され，そして分割のブロックの間に半順序関係が定められるという事実は，われわれの目的のためにそっくりそのまま利用することができる．そこで，概念とは，記号と対象のつくる2部

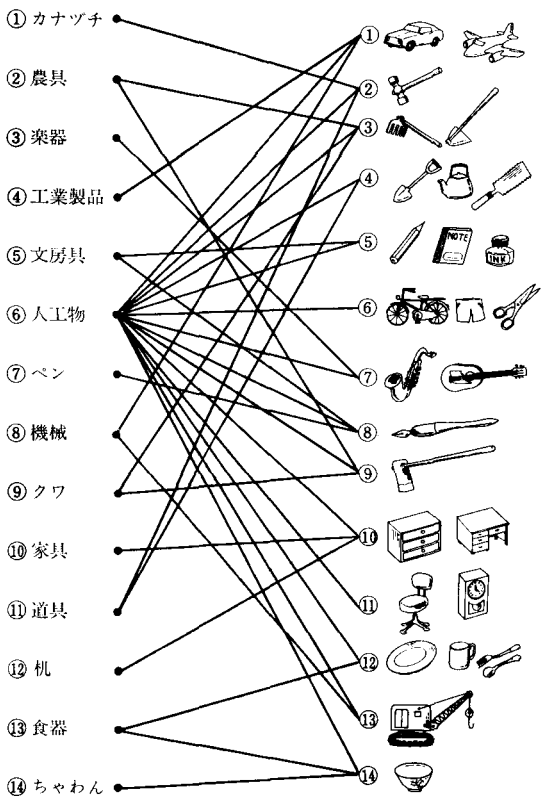

図 4　記号と対象の2部グラフデータ

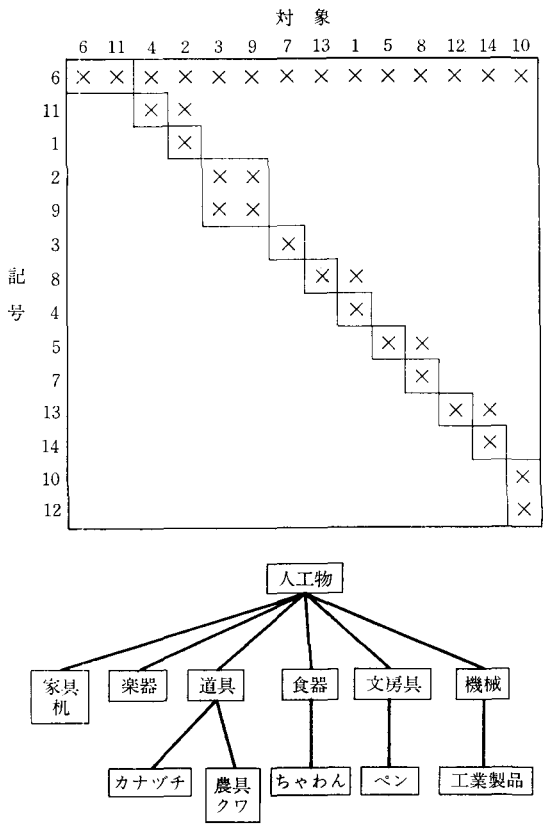

図 5　図4のデータから得られる概念構造

グラフに上の定理を適用して決まる半順序集合 W の元のことであると，あらためて**定義する**ことにしよう．このように定義した概念が，実際のデータではどんな様相を呈するかを次節で見ることにする．

4. いくつかの実験例

4.1 実　験 1

図 4 は記号と対象のつくる2部グラフの例である．これは，左側の記号と右側の対象のみを描いた紙を被験者に渡し，「ある対象をある記号で名指すことができるとき線で結んでください．」と言って線を書き入れてもらった結果である．このグラフに前節の方法を適用した結果が**図 5** 上である．図中，各行は記号を，各列は対象をあらわし，×印は該当する記号と対象を結ぶ枝をあらわす．また，対角線上に並ぶ四角形のブロックは W の元，すなわち概念，をあらわし，それらのブロックに含まれない×印は概念間の半順序構造を決める（半順序構造であるため，図のように，対角ブロックの左下には×印が1個もないように行と列を並べかえることができる）．このようにして抽出された概念構造が図5下である．ただし，ここでは，各概念を対応する記号の集合であらわしてある．

同様の実験で別の被験者の描いた2部グラフを整理した結果を**図 6** 上に，それから得られる概念構造を図6下に示す．図5と図6では構造が異なるが，これは実験時点における2被験者の概念構造の差異によるものであると考えられる．

4.2 実　験 2

喜怒哀楽に関する概念と味覚に関する概念について同様の実験を行なった結果を**図 7** と**図 8** に示す．この場合も，左側に記号を並べ，右側に対象

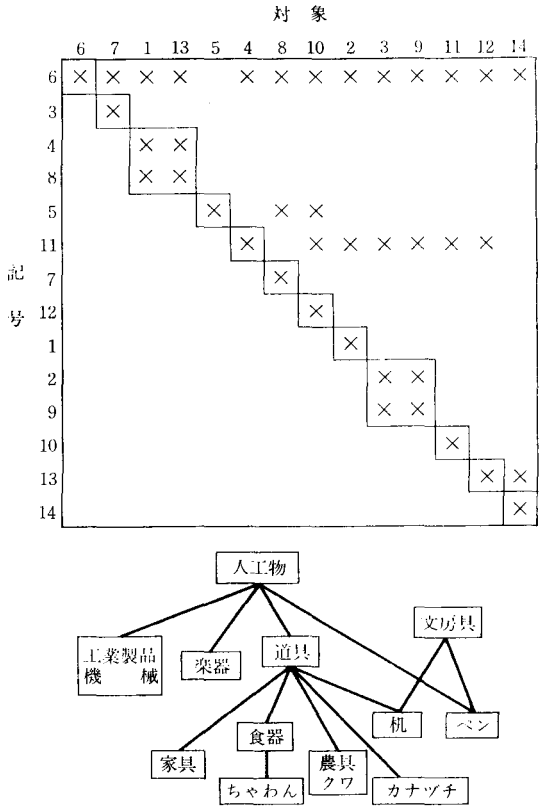

図 6　別の被験者から得られた概念構造

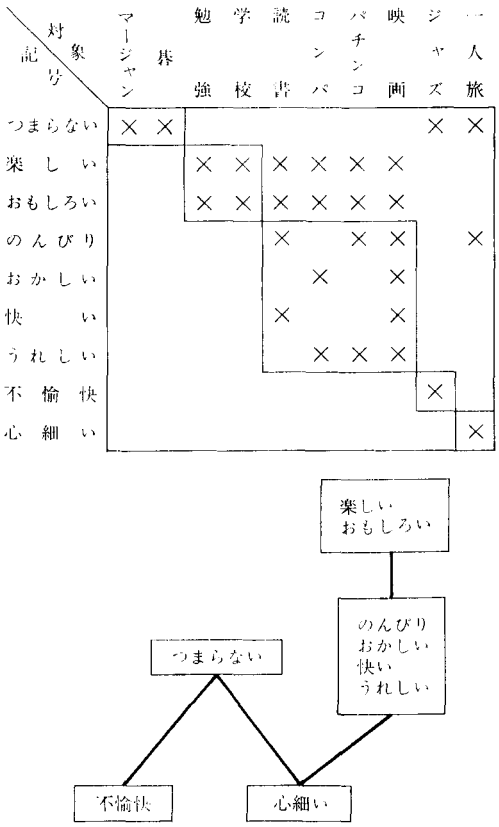

図 7　喜怒哀楽に関する概念抽出例

を並べた紙を被験者に渡し，記号と対象の間を線で結んでもらった．ただし，対象にも「単語」を使った点が実験 1 とは大きな相違である．得られた概念構造はどちらもそれほど意外なものではなく，記号集合と対象集合を“うまく”選ぶとかなり常識に近い構造が得られるということがわかる．

4.3 実　験 3

キーワードを記号とみなし，文献を対象とみなし，あるキーワードがある文献に含まれるという関係を枝とみなしてできる 2 部グラフに，概念構造決定法を適用してみた．情報工学関係の雑誌で，論文の投稿に際してキーワードを添えることが義務づけられているものをいくつか選び，付録 1 に示す方法でキーワードと文献の 2 部グラフをつくったところ，キーワード数34，文献数37の連結 2 部グラフが得られた．このデータから抽出された概念構造を図 9 に示す．

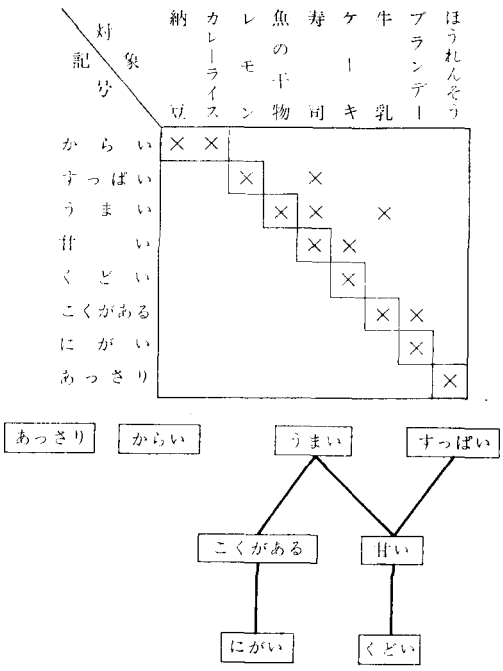

図 8　味覚に関する概念抽出例

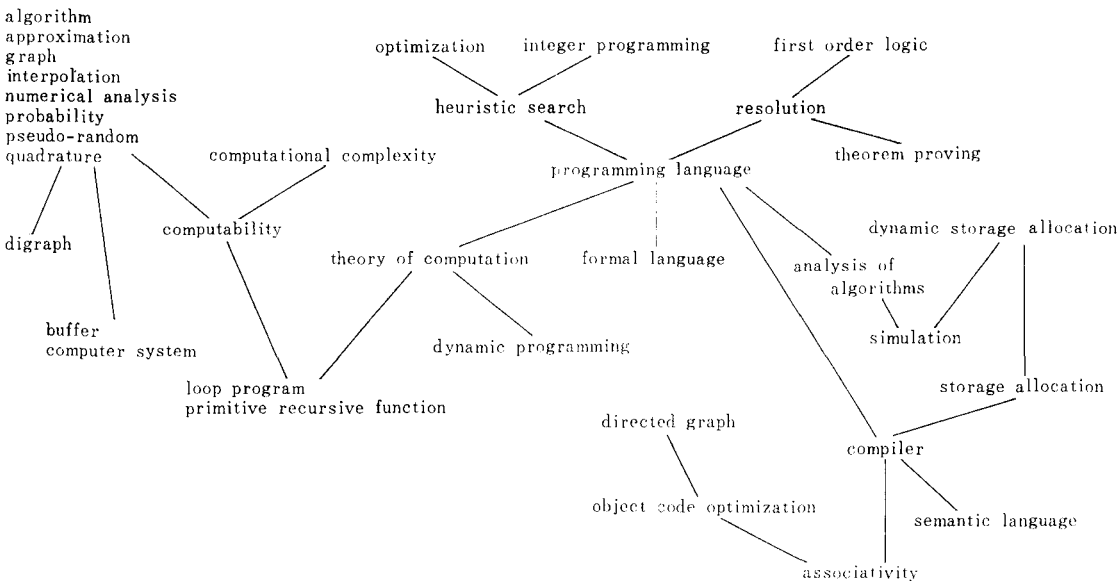

図 9　文献・キーワードのつくる2部グラフから得られた概念構造

5. 応用の可能性と問題点

本稿で述べてきたような問題の定式化と概念構造決定手法の意義について正しい評価を下すためには，まだまだ多くの実験を行なってみなければならない．人間の言語活動に関係したいろいろな場面で本手法を適用し，経験を積み重ねることによって，その効用と限界が明確になるであろう．また，本方法の入力データが記号と対象のつくる2部グラフという単純な形式であること，概念抽出の手順を実現する効率のよいアルゴリズムが知られていること，質的なデータを質的なまま取り扱えること，多変量解析等では得られない半順序構造という魅力的な構造が決定できること等の特徴を考えると，今後多くのデータについて実験してみる価値があると思われる．

その場合，単に得られた構造をながめるだけでなく，たとえば情報処理の効率の向上とか，言語病理学的な診断とかいうような目的に使ってみる（たとえば，実験3の結果を文献検索用のシソーラスとして使ってみる）ことも必要であろう．

本方法の問題点の一つは，記号集合と対象集合の大きさがかけ離れていると良い結果が得られないという点である．そのために，実験3では，付録1に示すように，文献・キーワードのつくる2部のグラフから頂点をランダムに取り除き，左右の頂点数の差を小さくする必要があった．しかし，最近の組合せ論の進歩の結果，第3節で述べた定理はさらに一般化できることが解明され[4]，それを利用すればこの問題点は一部解消すると思われる．

また，外延の包含関係（すなわち，記号 a と枝で結ばれている対象の集合が記号 b と枝で結ばれている対象の集合を含むとき，a は b を含むとみなす包含関係）が，本方法の上位・下位関係として明示的にあらわされない場合があること，本方法で得られる上位・下位関係が常識的なものと反転することがあること等の問題点にも，一応の解釈が与えられることがわかっている[6]．

6. おわりに

概念とその間の上位・下位構造を，「記号・対象・名指す関係」というデータから操作的に構成する一つの手法について紹介した．数理的手法というととかく計量的方法のみが強調されそうな今日，本稿のような定性的手法も，とくに言語活動

のように質的側面の強いデータに対してはいくらかの意味があるものと考えられる．なお，第4節で使用したデータはすべて[12]からとった．

参考文献

[1] Dulmage, A. L., and Mendelsohn, N. S.: Coverings of Bipartite Graphs. *Canadian Journal of Mathematics*, Vol. 10 (1958), 517—534.

[2] Dulmage, A. L., and Mendelsohn, N. S.: A Structure Theory of Bipartite Graphs of Finite Exterior Dimension. *Transactions of the Royal Society of Canada*, Ser. 3, Section III, Vol. 53 (1959), 1—13.

[3] ハヤカワ，S. I.（大久保忠利訳）：思考と行動における言語．岩波書店，1965.

[4] Iri, M.: Principal Partitions of Matroids and Their Applications——A Review of Recent Activities in Japan. *Proceedings of the Second International Conference on Combinatorial Mathematics*, held in New York, April 4～7, 1978, to be published by the New York Academy of Sciences.

[5] 伊理正夫，韓太舜：線形代数——行列とその標準形．教育出版，1977.

[6] 伊理正夫，杉原厚吉：概念構造決定への数理的一接近法．昭和52年度文部省科学研究費特定研究「言語」藤崎班研究資料 No. 52—5，1977.

[7] Jones, K.P.: *Automatic Keyword Classification for Information Retrieval*. Butterworths, London, 1971.

[8] Lawler, E. L.: Cutsets and Partitions of Hypergraphs. *Networks*, Vol. 3 (1973), 275—285.

[9] Москович, В. А.: Статистика и Семантика. Наука, Москва, 1969.

[10] Osgood, C. E., Suci, G. J., and Tannenbaum, P. H.: *The Measurement of Meaning*. The University of Illinois Press, Urbana, 1957.

[11] Roget, P. M.: *Roget's International Thesaurus* (4th edition; revised by Chapman, R.L.). Thomas Y. Crowell Co., New York, 1977.

[12] 杉原厚吉：概念構造を決定するための数理的一手法．東京大学大学院工学系研究科計数工学専門課程修士論文，1973.

付録1　実験3で使用したデータ

下記の雑誌からキーワードを5個以上含む論文を拾い出したところ，論文数は128編であった．

Information Processing Letters, Vol. 1, No. 1 (1971), No. 2 (1971), No. 3 (1972),

NBS Journal-Mathematical Sciences, Vol. 74B, No. 3 (1970), No. 4 (1970); Vol. 75B, Nos. 1 & 2 (1971),

Journal of the Associaton for Computing Machinery, Vol. 17, No. 1～No. 4 (1970); Vol. 18, No. 2～No. 4 (1971); Vol. 19, No. 1 (1972).

これら128編の論文の2編以上に含まれるキーワードのみを残し，論文とキーワードの2部グラフをつくったところ，最大連結成分はキーワード48個と論文64編とによって構成されていた．このグラフから論文16編をランダムに取り去ったところ，残った論文48編とキーワード48個のつくる2部グラフのうち最大連結成分はキーワード数34，文献数37であった．この2部グラフを，実験3で使用した．

すぎはら・こうきち　1948年生
1973年　東京大学大学院計数工学専門課程修士修了
1973年7月　電子技術総合研究所
バイオニクス研究室
いり・まさお　1933年生
1955年　東京大学工学部応用物理学科卒
現在　東京大学工学部計数工学科教授