Library and Information Science No. 22 1984

漢文資料の計量的分析
——「黄帝陰符経」の註を対象として——

Statistical Analysis of Chinese Classic Text Structure of Keywords in Commentaries of Huang-ti Yin-fu-ching

松本浩一 *Koichi Matsumoto*
宮本定明 *Sadaaki Miyamoto*
中山和彦 *Kazuhiko Nakayama*
星野聰 *Satoru Hoshino*

Résumé

In a previous paper the authors considered a statistical method for analyzing Chinese classic text by computer and applied them to a commentary of Huang-ti Yin-fu-ching. This method is to make a permutated matrix by a two-way clustering which shows the mutual relationship between important words and commentary units, which are assumed to be units of meaning in the commentary. In spite of its usefulness, this method has following problems to be overcome.

(1) It is difficult to compare different permutated matrices from different commentaries.
(2) The previous method depends upon commentary units which is a particular form of commentary. Therefore its application is severely restricted.

In the present paper new methods are considered and discussed to solve these two problems. For problem (1), we try cluster analysis of commentaries by different authors based on unified re-phrasing of the original canon and the commentary to enable comparison of different permutated matrices. For problem (2), instead of a commentary unit, we consider a sentence and neighborhood of a keyword as a unit of meaning. These methods can be applied Chinese classic texts of any format. In particular the latter method need not any manual phrasing.

松本　浩一，図書館情報大学助手，茨城県筑波郡谷田部町
Koichi Matsumoto, Research Assistant, University of Library and Information Science, Yatabe-machi, Tsukuba-gun, Ibaraki-ken
宮本　定明，筑波大学電子情報工学系講師，茨城県新治郡桜村
Sadaaki Miyamoto, Assistant Professor, Institute of Information Science and Electronics, University of Tsukuba, Sakura-mura, Niihari-gun, Ibaraki-ken
中山　和彦，筑波大学電子情報工学系教授，茨城県新治郡桜村
Kazuhiko Nakayama, Professor, Institute of Information Science and Electronics, University of Tsukuba, Sakura-mura, Niihari-gun, Ibaraki-ken
星野　聰，図書館情報大学教授，茨城県筑波郡谷田部町
Satoru Hoshino, Professor, University of Library and Information Science, Yatabe-machi, Tsukuba-gun, Ibaraki-ken

漢文資料の計量的分析

I. はじめに

近年，人文科学研究におけるコンピュータ利用の波は，中国史や漢文学の分野にも及び，京都大学を中心にこの分野のデータベース開発が進められつつある。この分野におけるデータベースは，(1)漢文テキストからなるデータベース，(2)二次的資料から成るデータベース，(3)書誌情報データベースに分けて考えられる[1]。(2)は原史料になんらかの形で加工をほどこし，整理されたデータとしてからデータベース化したもので，京都大学で既に実用に供されている例として，明代科挙合格者のデータベースがある。また(3)については，同じく京都大学人文科学研究所発行の，「東洋学文献類目」がデータベース化されている[2]。(1)の漢文テキストから成るデータベースとしても，京都大学では既に唐代の詩人李商隠の文集「樊南文集」を，富士通の検索システム FAIRS を用いてデータベース化し，公開している。

(1)の種類のデータベースを実現するには，漢文テキストそのものを扱うため，多くの漢字を処理できる必要があること，またテキストデータを扱うに適したデータベースシステムが未だ開発されていないことなど，困難な点が多い。しかしこの種のものが研究者にとって特に興味深いのは，入力されたテキストの検索が容易に行なえるという点ばかりではなく，更に進んでは，計算機を利用して様々な計量的分析を試みることができることである。

こうした計算機を利用した，漢文テキストの計量的分析の一例として，以前われわれは，道教の主要経典の一つである「黄帝陰符経集註」の分析を試みたことがある[3]。

中国の古典においては，しばしば経に対する註という形で，新しい思想が述べられる場合が多いが，特にこの「黄帝陰符経」の場合，経文自体が短かく，表現も象徴的であるため，様々な思想的立場に立った註がほどこされてきた。しかし註文は，あくまで経文一句一句についての解説，解釈といった形をとるため，註文の主題は筋道に沿って展開されるよりも，経文の句に従って，かなり分散されてしまう場合が多い。以前の方法はこの点に着目して，註文に現われる重要単語相互の関係と，ある重要単語が註文のどの部分に集中的に現われてくるかを，二元クラスタリング[4]手法を用いて表化するものであり，次のような手順から成っていた。

①単語（ここでは一漢字一単語と仮定した）の頻度と，出現の際のコンテキストを勘案して，註文の重要単語を抽出する。コンテキストの検討のためには，漢字一字をキーワードとする KWIC 索引を作成した。

②経文一句に付せられた註文ひとまとまりを，一つの意味の単位とし（以下註文単位と呼ぶ），註文単位ごとに重要単語の出現頻度を数えたマトリックスを作る。

③ここで重要単語の集合を $W=\{w_1, w_2, \ldots\ldots, w_m\}$，註文単位の集合を $T=\{t_1, t_2, \ldots\ldots, t_n\}$ とし，単語 w_i が註文単位 t_j に現われる頻度を χ_{ij} とする。まず重要単語相互の関係については，次の式によって単語相互の類似度 $S(w_i, w_j)$ を定める。

$$S(w_i, w_j)=\frac{\sum_k \chi_{ik}\ \chi_{jk}}{\sqrt{\left(\sum_k \chi^2{}_{ik}\right)\left(\sum_k \chi^2{}_{jk}\right)}} \tag{1}$$

同様にして註文単位相互の類似度 $S'(t_i, t_j)$ を次のように定める。

$$S'(t_i, t_j)=\frac{\sum_k \chi_{ki}\ \chi_{kj}}{\sqrt{\left(\sum_k \chi^2{}_{ki}\right)\left(\sum_k \chi^2{}_{kj}\right)}} \tag{2}$$

これらの類似度をもとに，重要単語相互の関係を現わす樹形図と，註文単位相互の関係を現わす樹形図

とを作成する。

④二つの樹形図において示された順序に従って，②で作成したマトリックスの重要単語と註文単位の順序を並べかえ，両者の相互関係を反映したマトリックス（以下置換マトリックスと呼ぶ）を新たに作成する。

この方法によって得られた置換マトリックスからは，註文全体におけるキーワードが互いにどのような関係をもち，主として註文のどの部分に現われてくるかが読みとれ，一見混乱した註文の思想を解読する上でよいチャートとなるものであったが，次のような点が満たされていなかった。

①著者を異にする註それぞれについて作成した置換マトリックスを，相互に比較するのが難しい。筑波大学学術情報処理センターですでに入力ずみの，全部で18種類の「黄帝陰符経」の註について，それぞれにこの置換マトリックスを作成してみると，個々の註を読解する上では有益であっても，註によって経文の句切りのしかたが異なるため，二つの置換マトリックスをそのまま比べたのでは，二つの註の特色を経文の解釈という点で比較することはできない。

②分析方法が註釈という形式に依存している。ここでは，註釈という，意味のひとまとまりを機械的に区切りやすい形式に依存して分析を行なったが，ふつうの散文の形式をとる論著等においては，この方法を適用することはできない。註釈以外の形式のテキストの場合では，当然註文単位相互の関係を考える必要はないが，重要単語相互の関係を表わす樹形図を作成するためには，類似度の計算方法を別に考えねばならない。

今回の論文では，主としてこの2つの問題点を解決するために，試みたいくつかの方法について報告し検討することにしたい。

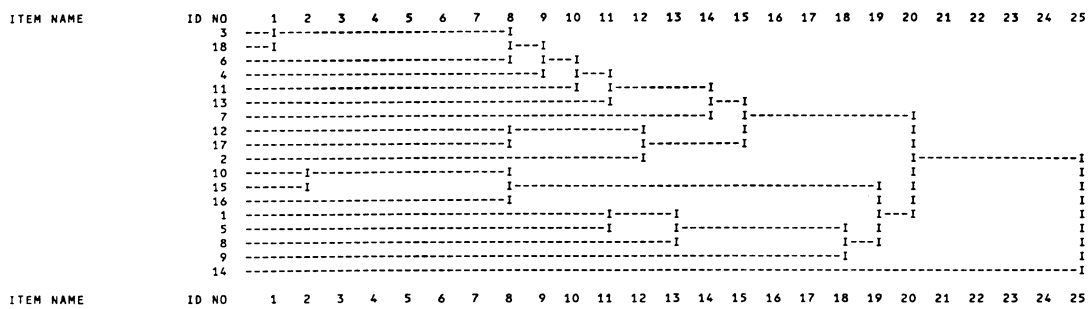
II. 著者を異にする註の相互比較

これには2つの方向が考えられる。1つは個々の註について作成された置換マトリックスを，相互に比較可能なものにすることで，序文に述べた①の問題点についての直接的な解決となるものである。またもう1つの方向は，著者を異にする註相互の類似度を何らかの方法によって定義し，註相互の関係を表わす樹形図を作成することである。

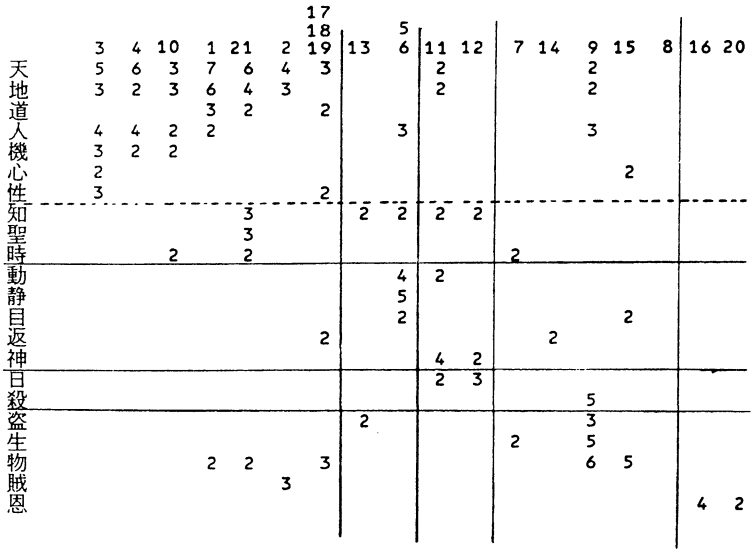
A. 証相互の関係を表わす樹形図の作成

まず後者の方法から述べる。この場合最も簡単な方法は，すべての註に現われる重要単語を抽出し，各々の註におけるそれらの単語の頻度をもとに註相互の類似度を計算することで，前回の方法における註文単位を，1つの註全体に置きかえたものである。ここでは18種類の註から，全体における頻度と個々の註における重要性とを勘案して，91個の重要単語を選んだ。ここで単語の集合を $W=\{w_1, w_2, \ldots\ldots, w_m\}$，註の集合を $T=\{t_1, t_2, \ldots\ldots, t_n\}$ とし，単語 w_i が註 t_j に現われる頻度を χ_{ij} とすれば，註相互の類似度 $S(t_i, t_j)$ は，前回同様式(2)で表わされる。この類似度をもとにPAB[5]のプログラムによって作成した樹形図が図1である。

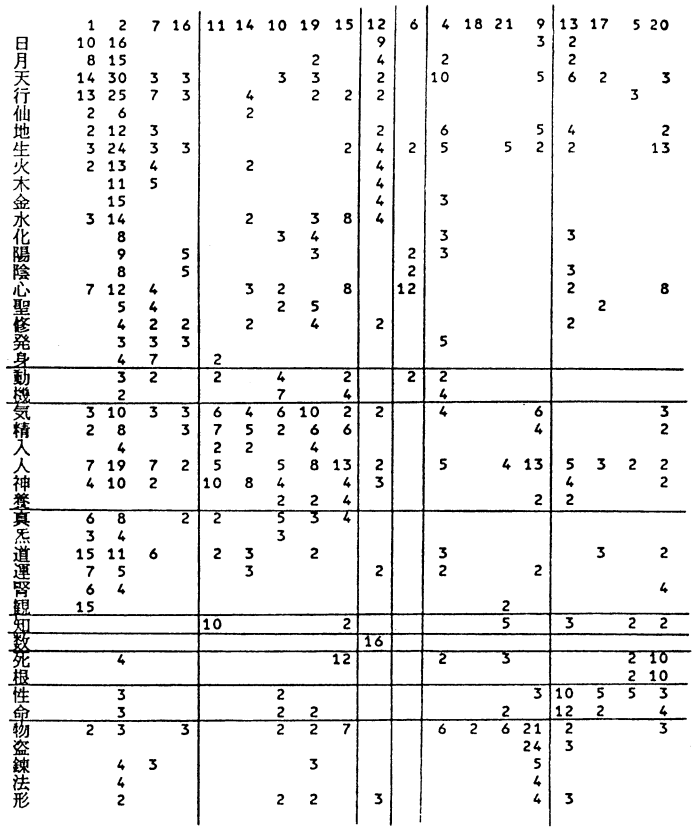
この図を見ると，まず3（李筌疏）と18（袁淑真註）がかなり密接な関係をもっていることが読みとれるが，それはこの2つの註がほとんど同じものであるためである。李筌は唐の玄宗のころの人で，この他に1の「黄帝陰符経集註」の著者の一人でもあるが，3の疏はごく短い註のあとに疏が続く形式になっている。ところが北宋



第1図 「黄帝陰符経」の18種類・註の相互関係を表わす樹形図



第2図　註10（蔡氏註）についての置換マトリックス



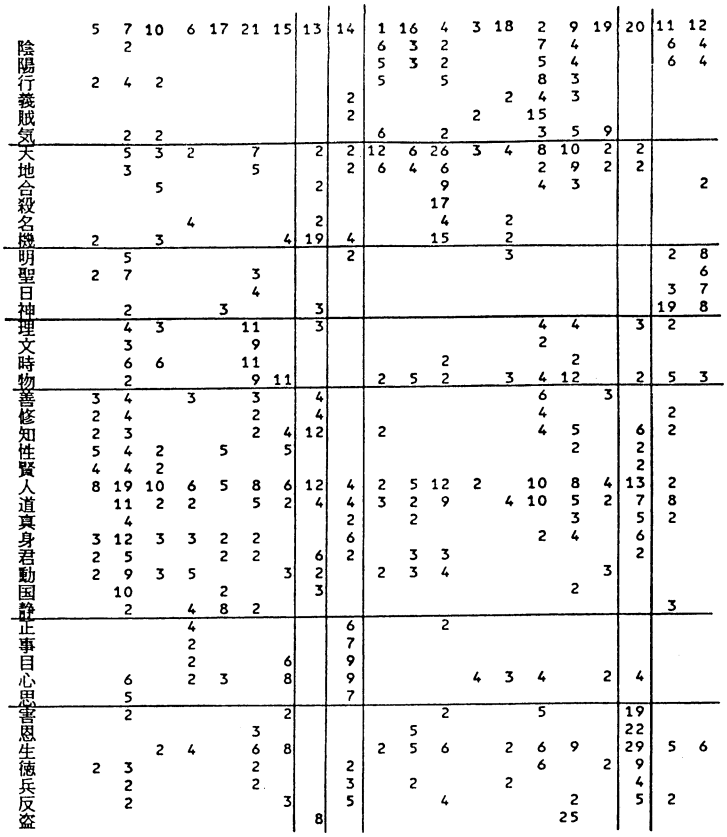
第3図　註12（唐淳註）についての置換マトリックス

の人と考えられる袁淑真の註文は，経文のどの句につけられたものをとっても，その部分に対応する3の註文と疏をあわせたものとほとんど同一である（おそらくは李筌疏とされているもののほうが偽作と思われる)[6]。また同じように深い関係が図に示されている，10の蔡氏註と15の朱子註についていえば，後者の註文は，前者の註文と全く同じものか，あるいは前者の註文に，「朱子語類」等にみられる朱子のその経文に対する評言や，編者らの考証をつけ加えたものからなっている。

またこの図において，1つのグループを形成していると考えるものには，次のようなものがある。まず図の中ほどにある12（唐淳註），17（王道淵註），2（夏元鼎註）からなるグループの註は，いずれも金丹道，即ち人の身中にある精，気，神は錬ることによって金丹に化し，不老不死を得て仙人になるという教えの立場から書かれたものである。次に10，15，16（兪琰註）からなるグループは，15の朱子註に象徴されるように，儒教的色彩が強い註のグループといえる。また1と5（張果註）はこれら18の註の中で最も古い成立と考えられ，二者の関係も深い。それで以上2つのグループよりまとまりは弱いが，図の下方の1，5を中心としたグループに属する8（黄居真註），9（沈亜夫註）等も，これら二者の説を引いていることが予想される。このほか先述の3，8あるいは6（蹇昌辰解）は，いずれも北宋の成立であり，これらを核とした図上方のグループも共通の特色をもつと考えられる。更にそれがいかなるものかを考察するには，それぞれの註を個々に考察していく必要があるが，この図はそのような註どうしの系譜関係を探るための，第一の手がかりとなるものである。

B. 置換マトリックスの相互比較

次に個々の註についての置換マトリックスを相互に比較可能なものにする方法であるが，先述のようにこの相



第4図　註18（袁淑真註）についての置換マトリックス

互比較が難しいのは，個々の註によって経文の句切りの仕方が異なることが主たる原因となっている。そこでこの解決策として考えられるのは，分析を行なう前に予め個々の註における経文の分割の仕方を統一し，註文もその句切り方に従ってそれぞれ統合・分割することである。まず経文について，18個の註各々において句切れとなっている箇所を書き入れていく。そして10個以上の註において，句切れとなっているところを標準的な句切りとする。これで経文は全体で21の句に分かれることになる。それぞれの句には通し番号をつける。次に個々の註における経文の句切りを標準的な句切りにあわせ，各々の句に付けられた註文もそれにあわせて統合，分割していく。即ちある註において経文が標準的句切りよりも更に細かく分けられている部分では，いくつかの句をそれにあわせて1つの句とし，註文も経文にあわせていくつかの註文単位を1つの註文単位に統合する。しかし逆に，ある註における経文の句切りが標準的句切りより粗く，一句が標準句のいくつかを含む場合は，句を再分割し，註文もそれに従って意味の上から，いくつかの註文単位に分割すべきであるが，註文においては必ずしも経文の語句を上から順に解説しているわけではないため，実際には再分割は難しい。そこでここでは統合すべきところだけを統合し，分割すべき部分は通し番号によるラベルによって識別することとした(図2)。そして註文単位の通し番号を，経文の通し番号に一致させるために，経題，章題等につけられた註は一切省略する。

このような処理を施したのち，註10，12，18（それぞれ蔡氏註，唐淳註，袁淑頁註）についての二元クラスタリングを行なったのが，図2，図3，図4である。この3つの置換マトリックスにおいて，註文単位の番号は，同じ数字はすべて経文の同じ句に対応している。従って各々の註において，経文のどの部分とどの部分が密接に結びつけられているか，また同じ経文がどのような語によって解説されているかを比較するのは極めて容易となる。但し個々の註の性格については，以上の処理を施していない置換マトリックスのほうがよりよく反映しており，また読みとりやすくなっていることは否めない。

III. 註釈の形式に依存しない分析方法

以上の方法は，註文単位という註釈に形式に特有なものを意味のひとまとまりと考え，そこでの単語の頻度を基本データとするという点で，註釈という形式に依存した方法であった。そこで次に，同じく「黄帝陰符経」の註文を分析の対象としながら，この形式に依存しない，従って普通の散文形式にも応用できる方法を考えてみたい。ここでも重要単語の共出現関係によって単語相互の類似度を定め，その関係をクラスター分析によって樹形図の形で表示する方法をとるが，共出現の単位としては，1つの方法では文を採用し，もう1つは単語の近傍というものを考えていく。

A. 文を共出現の単位とする方法

周知の如く，中国の古典においては，もともと句読点は施されておらず，文という概念も明確ではないから，文という単位の決定は研究者自身が行なわなければならない。それでこの単位の決定には，「也」や「矣」などいわゆる文末の助字の出現による以外は，研究者の漢文における文についての考え方の相違によって，かなりのちがいがでてくることが考えられる。また漢文における単語の共出現の単位として考える場合には，いわゆる文という概念によった意味の単位よりも，もう少し拡大された意味のひとまとまりを考えたほうがより適切なのかもしれない。これは別に議論されねばならない問題であろう。ここでは筆者の句読に従って文の単位を決めた。

ここである註の註文における重要単語の集合を $W=\{w_1, w_2, \ldots\ldots, w_m\}$，文の集合を $T=\{t_1, t_2, \ldots\ldots, t_n\}$ と表わし，単語 w_i が文 t_j に現われて頻度を χ_{ij} とする。2つの単語の類似度 $S(w_i, w_j)$ は，2つの式によって計算される。

$$S(w_i, w_j)=\frac{\sum_k \min(\chi_{ik}, \chi_{jk})}{\sum_k \max(\chi_{ik}, \chi_{jk})} \tag{3}$$

$$S(w_i, w_j)=\frac{\sum_k \chi_{ik}\,\chi_{jk}}{\sqrt{\left(\sum_k \chi^2_{ik}\right)\left(\sum_k \chi^2_{jk}\right)}} \tag{4}$$

(4)の式は(2)と同様の内積による方法で類似度を求めるものであり，(3)はミニマックス法[7]によったものである。この2つの方法によって註12の重要単語間相互の類似度を求め，群平均法によって樹形図を作成したのが図5，図6である。2つの樹形図には細部でのちがいはかなり認められるが，いずれの図においても，「修」，「錬」，「神」，「仙」，「聖」，といった単語と，「身」，「気」，「精」，「化」という単語が結びついてでてくる点に，身中の三宝といわれる神，気，精を錬る修行によって神仙，仙人に変化する道を説く金丹道の色彩を強くもった，この註の性格をよく表現しているといえる。

B. 単語の近傍による方法

漢文の文章は漢字のひと続きと考えることができる

が，その中である単語（漢字一字）を中心に置いた場合，その近傍即ち前後何字かの範囲内にもう1つの単語が現われてくる頻度によって，2つの単語の類似度を決める方法を考えることができる。それにはまず，ある単語の前後何字づつを，その単語の近傍と考えるかが問題となる。漢文の場合，漢字四字で一句をなすことが多いことは周知のことであるが，そのことから最も小さな近傍の範囲としては，ある単語を中心に前後3字づつというのが考えられる。しかしこれではあまりに狭すぎてしまうため，次にこの四字句が対句をなすことが多いことを考えて，前後7字づつを近傍とすることが考えられる。ここではこれを近傍の範囲として採用する。つまり単語 w_i, w_j の共出現頻度を考える場合，w_i を中心においた時，その前後7字の範囲に w_j が出現する回数を，w_i と w_j の共出現頻度とするのである。

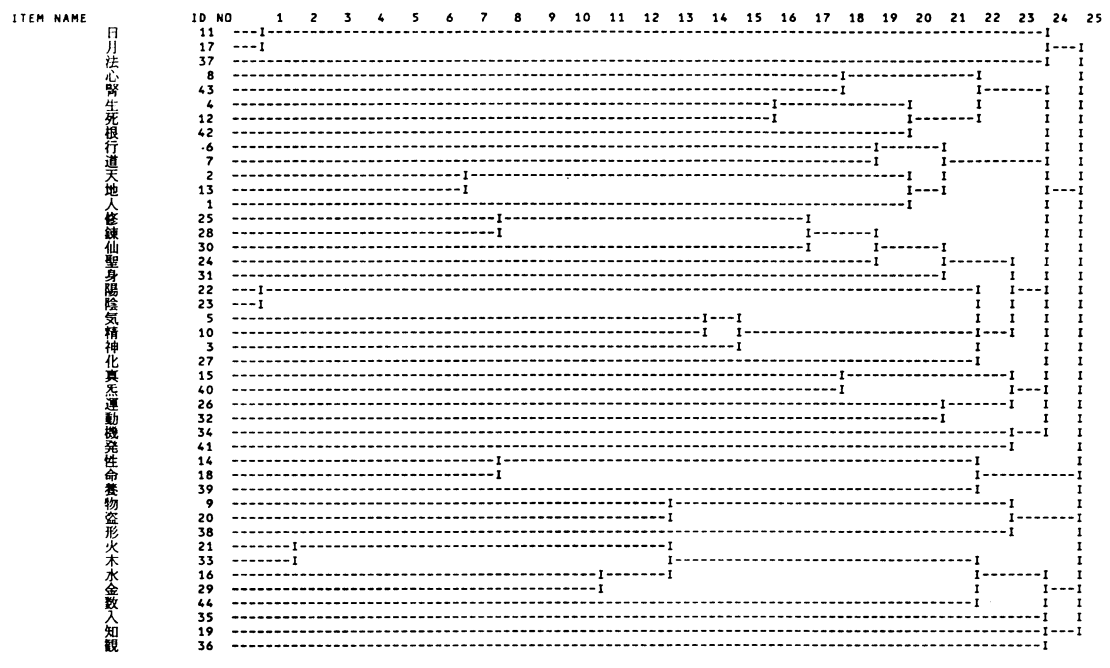
ここで，共出現頻度の数え方には次の3つの方法が考えられる。

①漢文の文章中において，w_i は何回か出現するが，1つの w_i につき，その前後7字の間に w_j が出現すれば，その範囲内における w_j の出現回数にかかわらず，それを1回と数える。この方法によって得た w_i, w_j の共出現頻度を $C_1(w_i, w_j)$ とおく。

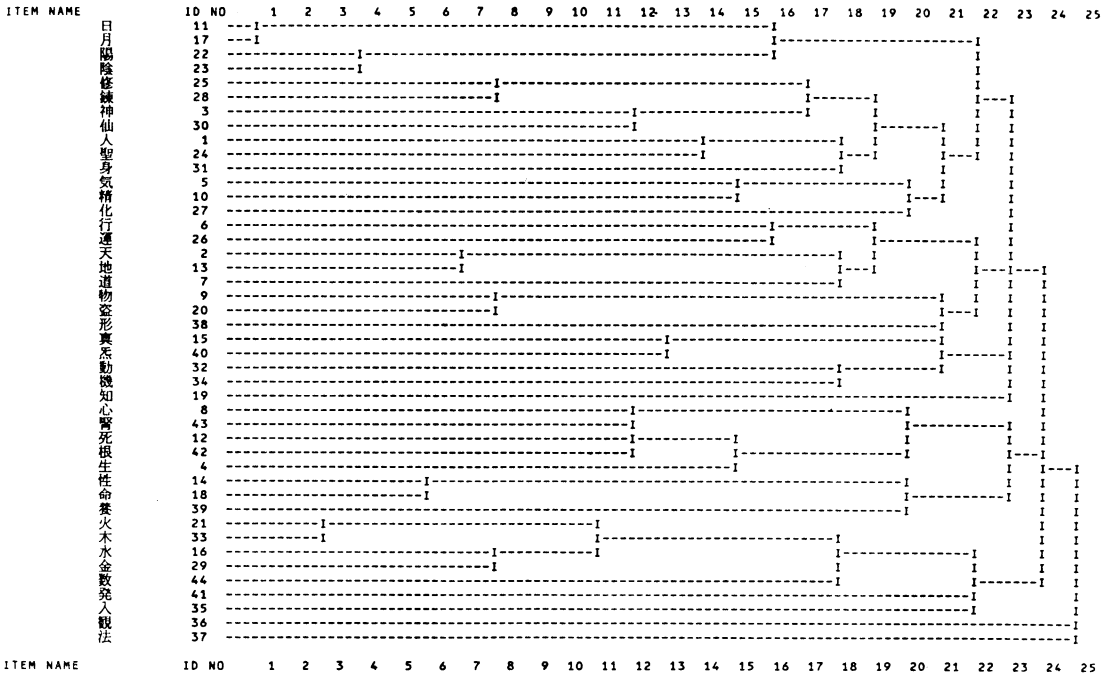
②1つの w_i につき，その前後7字の間に w_j が出現すれば，出現した回数すべてを共出現頻度として加算する。ここではこの方法は採用していない。

③1つの w_i につき，その前後7字の間に w_j が出現した回数はすべて出現頻度として加算するが，この場合一度数えた w_j は，次に出現する w_i の前後7字の間にあっても，それは回数に数えない。この方法による w_i, w_j の共出現頻度を $C_3(w_i, w_j)$ とする。

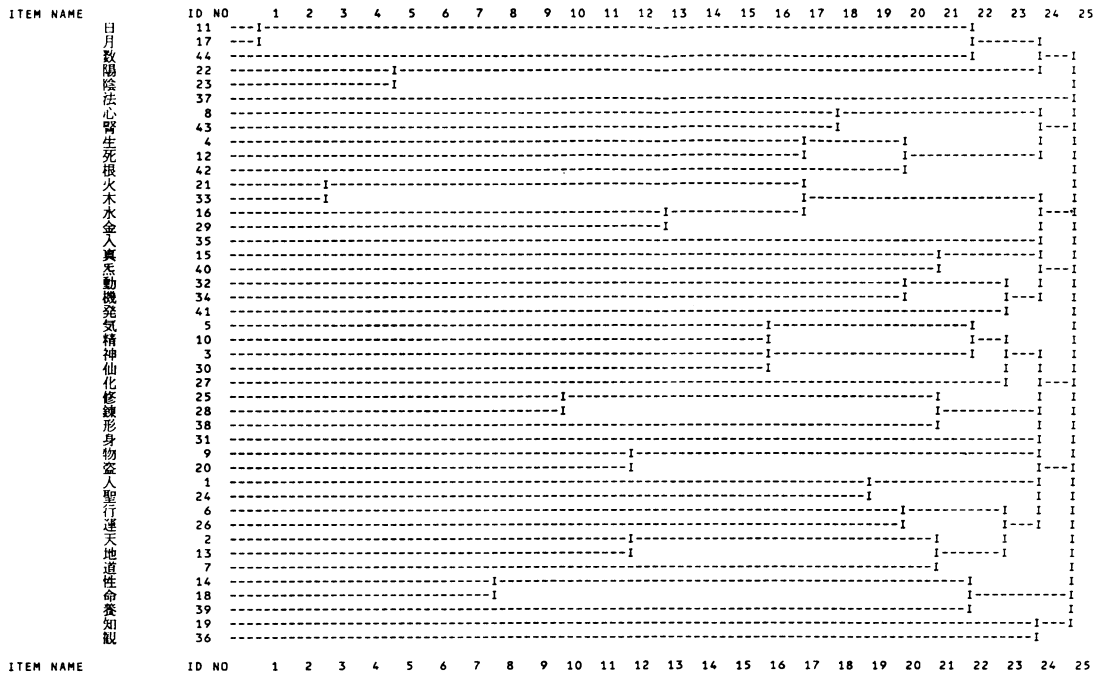
即ち下に示す「地」を索引語とした KWIC 索引の例では，1つめの「地」の近傍には「天」は2回，2つめの「地」の近傍には3回現われている。①の方法によれば，それぞれを1回と数え，この場合「地」と「天」と共出現頻度は2である。②の方法では，それぞれ2回，3回をそのまま加算して，共出現頻度を5とする。そして③の方法では，1つめの「地」の近傍では，「天」の出現回数は2回と数えるが，2つめの「地」については，その近傍にあらわれる3つの「天」のうち，前2つは1つめの「地」の近傍で数えた「天」と同じものであるため，この場合「天」の出現回数は1回となり，共出現頻度は合計3回と数える。



第5図　註12の重要単語相互の関係を示す樹形図（ミニマックス法による）



第6図　註12の重要単語相互の関係を示す樹形図（内積法による）



第7図　註12の重要単語相互の関係を示す樹形図（近傍法による）

天文　地　理也，天時地利也，象
天文地理也，天時　地　利也，象天体制者聖，
察天時人事，若能合天　地　化育，与時設教，乃聖
道分絹為天　地，天地分而為万物，　万
道分而為天地，天　地　分而灰万物，万物之中

①，③の方法では，2つの単語 w_i, w_j の共出現頻度は，w_i を中心において数えた場合と，w_j を中心において数えた場合とでは一致しない。即ち $C_1(w_i, w_j) \neq C_1(w_j, w_i)$ であり，$C_3(w_i, w_j) \neq C_3(w_i, w_j)$ である。しかしこの2つの方法には，$C_1(w_i, w_j)=C_3(w_j, w_i)$ という関係が存在している。即ち w_i を中心において，①の方法によって数えた w_i, w_j の共出現頻度とは，w_j を中心において，③の方法で数えた共出現頻度は等しくなることが証明される（付録の証明を参照）。このことにより，単語 w_i, w_j それぞれの頻度を x_i, x_j としたとき，この2つの単語間の類似度 $S(w_i, w_j)$ を次の式で定めることにする[8]。

$$S(w_i, w_j) = \frac{\min(C_1(w_i, w_j), C_1(w_j, w_i))}{x_i + x_j - \min(C_1(w_i, w_j), C_1(w_j, w_i))}$$

この式に従って，図5，図6と同様，註12の重要単語について相互の類似度を求め，群平均法によって樹形図を作成したのが図7である。先にも述べたような註12の金丹道的性格は，この図にもはっきりあらわれている。この近傍の概念によって，重要単語相互の共出現頻度を求める方法によれば，文を単位としたときのようにテキストに人手を加えることなく，オリジナルな漢文データそのままで計算機処理ができる。更にこの方法はいかなる形の文章にも適用できるため，その応用範囲も広い。ただ近傍の範囲として，前後何字づつを設定するのが適当であるかということは，漢文資料の性格，特性に則して別に検討される必要があろう。

IV. おわりに

しばしば単語の意味はその使われ方によって決定されるといわれるが，ここで得られた樹形図の結果は，個々の漢文資料中における単語の真の意味や，著者の意味世界の特性，更には文献が書かれた時代の思考の枠組などといった，思想史上重要な問題にアプローチする上で，非常に有益な道しるべとなる。特に最後に検討した単語の近傍の概念による方法によれば，計算機に入力されたあらゆる形式の漢文テキストを，予め手を加えることなく，そのままの形で分析することができる点で応用範囲は広いように思われる。

ただここでは，前回からの懸案であった，1つの漢字を1単語とみなすことの再検討を果たすことはできなかった。また今回の研究の過程で，単語の共出現頻度を計算する単位として，文をどのように考えるか，近傍の範囲をどう決定するのかといった問題も新たに浮かび上がってきた。更には単語と単語の関係という問題をより深く考えていこうとすれば，単に共出現という関係のみではなく，データ処理の際に，単語間の意味的な関係というものを反映させることを考えていかなければなるまい。それで今回試みたような方法を更に洗練されたものにしていくためには，次には以上のような，漢文の特性とも深くかかわった問題にとりくんでいく必要があると思われる。

1) 星野　聰，勝村哲也，村尾義和“中国の歴史に関するデータベースの開発”．情報処理学会第23回全国大会講演論文集．p. 523-524 (1980)．
2) 星野　聰，勝村哲也“東洋学文献類目のデータベース化”，情報処理学会論文誌．Vol. 25, no. 2, p. 187-193 (1984)．
3) Matsumoto, K. ; Miyamoto, S. ; Nakayama, K. “Computer application to the research of Chinese history: indexing and statistical analysis of Chinese Classics.” Proceedings of International Computer Symposium 1982. p. 121-128 (1982).
4) 宮本定明“計量書誌学統計処理パッケージの作成”．昭和58年度科学研究費補助金（試験研究）研究成果報告書．1984，175 p.
5) Programs for Analyzing Bibliographic databases, 4) の文献参照
6) 松本浩一“陰符経の諸註についての諸問題”．“アジア諸民族の社会と文化”．東京，国書刊行会，1984，p. 189-215.
7) Miyamoto, S. ; Miyake, T. ; Nakayama, K. “Generation of a pseudothesaurus for information retrieval based on cooccurrences and fuzzy set operations.” IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-14, p. 203-212 (1984).

8) 一般に，クラスター分析の類似度は，総出現頻度 x_i によって影響を受けないように正規化しておくほうが，望ましい結果が得られることが経験上確かめられている．

付録

$C_1(w_i, w_j) = C_3(w_j, w_i)$ の証明について。

$C_3(w_j, w_i)$ を計算するとき，共出現頻度として数えら

れた w_i に対して，1つずつ番号をふっていったと考えよう。明らかに番号をふられた w_i は，文章中の w_j のいずれかの近傍に属する。このことは，番号をふられた w_i のそれぞれの近傍に，少くとも1つの w_j が存在していることを意味する。一方 $C_3(w_j, w_i)$ の計算で，番号をふられなかった w_i は，どの w_j の近傍にも入っていない。このことは，番号をふらなかった w_i の近傍には，w_j が存在しないことを意味している。従って，$C_1(w_i, w_j)$ の計算において，先に番号をふられた w_i については共出現1回と数えられ，番号をふられなかった w_i については数えられない。すなわち，$C_1(w_i, w_j)$ は先にふられた番号の数に等しく，$C_1(w_i, w_j) = C_3(w_j, w_i)$ が成り立つ。