データベース比較・評価(1)

確率集中楕円による技術情報の解析
—CA SEARCHファイルのたばこ関連情報について—

高木 義和*

たばこ関連情報に付与されたフリーキーワードをもとに，情報を整理し，その結果を用いて，文献に付与されていた分類相互の関係を検討した。たばこ関連情報という特定情報のなかで，一般技術分類であるCA Search の分類の相互関係が明らかとなった。その結果，逆にたばこ関連情報全体が，どのような一般技術によって構成されているかを知ることができ，また数量化理論Ⅲ類によって得られた，たばこ関連情報全体の構造を明らかにすることができた。

1．はじめに

情報の解析は，①収集した情報を解析の目的に応じて整理し，②その結果を総合的に判断する，2つの作業から成る。後者②において正確な判断を得るためには，手数を要するけれども，前者は不可欠な作業である。そこで，統計的手法を前者に適用することにより，効率~~をよくする~~の良い情報解析について検討した。

2次文献情報は，複数のキーワードによって，内容が表現されているので，使用されたキーワードの相互関係を明らかにすることができれば，それに基づいて，元の情報を整理することができるはずである。

筆者は，前回CA Searchファイル中のたばこ関連情報に付与されていたフリーキーワードについて調査し，その結果を報告した[1]。CA92, 93巻(1980)のなかで「TOBACCO」をキーワードとして持つ1,087件に付与された語は，4609種，延べ22,260語あった。使用頻度の高い頻度1％以上の語は，名詞を中心として329種存在し，たばこ関連情報のなかで一定の技術範囲を表現した。

つぎに，この中から適切な語を選び，多変量解析の一手法である数量化理論Ⅲ類を適用した解析結果を報告した[2]。数量化理論Ⅲ類は，マーケッティングや経営活動の分野で発展してきた手法である。情報解析の手法として，その適用の可能性が水谷[3]や松尾[4]らによって指摘されている。前報では，変数として使用するフリーキーワードの選定に留意して解析を進め，たばこ関連情報は，大きく3方向に散布し，情報全体が4種のグループに分かれることを示した。

いっぽう，CA Searchファイルでは80の分類が用いられている。この分類は，一般的な技術分類であるので，CA80分類がたばこ関連情報の中で占める位置を明らかにすることにより，たばこ関連情報が既存のいかなる技術から成立しているか，レビューできる可能性がある。そこで，数量化理論Ⅲ類の結果に，分類による確率集中楕円を組み合わせ，たばこ関連情報におけるCA80分類とフリーキーワードの関係について解析を試みた。

* たかぎ よしかず 日本専売公社技術調査室

2．解析方法

フリーキーワードの選定と数量化理論Ⅲ類による解析については，すでに前報で述べたが，今回の確率集中楕円の基礎データとなるため，重複するが両者についても2.1と2.2とで簡単にふれる。

2.1 フリーキーワードの選定

解析に変数として使用するフリーキーワードが，特殊な要因を持つと解釈が複雑になり，よい解析結果がえられない。CA Searchファイルから D2 レコードとして抽出した，たばこ関連情報に付与されたフリーキーワードの性格を検討したうえで，変数として使用する語を決定した。

フリーキーワードの頻度分布は，一般に Bradford の法則 ~~5,6,7~~ に従うことが知られている。[5] 本法則によれば，横軸にフリーキーワードの頻度順のランク r を対数目盛でとり，縦軸に r 番目までのフリーキーワードの累積語数 S(r)をとると，グラフは近似的に線型になる。[6] Bradford の法則では，2 とおりの分布関数が考えられているが，実際のグラフでは一般に直線となり，r の小さいところと大きなところで，直線からのかたよりが見られる[7]。

CA Search ファイルから抽出した4,609種のフリーキーワードが，本法則で表現できるかどうか，検討を行なったところ，ランク r の小さな部分では，コア領域に相当する直線から上方へのずれが認められた。直線からのずれがやや大きかったため，Bradford の法則による表現型を，その r に関しての微分型である Zipf 型に変形し検討を行なった。

フリーキーワードの頻度を f で表し，頻度 f の場合のフリーキーワードの数を n (f)とすれば，f-n (f)の両対数グラフは，直線になる。

頻度 f が1-1199の全数を用いて両対数の回帰直線を求めたところ，式(1)が得られた。

$$\log n(f) = -1.7561 \log f + 2.8305 \cdots (1)$$

頻度 f の小さな部分で実測値からのずれが大きいので，n (f) が平均1語に満たない部分｛log(f)＜0｝を除き，その結果，残った59点から再度，回帰直線を求め，式(2)を得た。

$$\log n(f) = -1.9771 \log f + 3.493 \cdots (2)$$

式(2)の傾きは－2に近い値を示した。式(2)の算出に用いなかった n(f)＜1の範囲を，Bradford の法則を延長するには無理がある部分と判断し，その範囲に属した42語を領域Aの語とした。領域Aよりも範囲の広い log n(f)＜0.5の部分を，領域Bとした。領域Bの語は，式(2)から，ランク r の上位74語が相当した。ランク r の上位100語を第1表に示す。

2.2 数量化理論Ⅲ類による解析

TOBACCO をキーワードとして持つ1,087件の文献の解析に，数量化理論Ⅲ類を適用した。数量化理論Ⅲ類は外的基準がない場合で，かつ反応パターンから要因カテゴリーを分類するための方法である。要因カテゴリーにフリーキーワードを選び，サンプル（1,087件の文献）が各語を持つか否かのフリーチェック反応（1要因1カテゴリー）で処理を行なった。領域Aに属したランク r の上位42語を，基本の要因カテゴリーとして解析に使用した。

2.2.1 領域Aの語（ランク r の上位42語）による解析

領域Aの42語を用いて数量化理論Ⅲ類による解析を行なった。42語のそれぞれの語の間には，ほとんど共出現関係が認められた。解析の結果得られた固有値のうち，大きい方から順に5固有値を示す。（ ）の中は，その平方根で相関係数に相当する。

1λ	0.4351	(0.6596)
2λ	0.2870	(0.5358)
3λ	0.2466	(0.4966)
4λ	0.2407	(0.4906)
5λ	0.2137	(0.4623)

λ が0.5程度以上の高い相関を示す固有値は得られなかった。第1軸は＋の方向に「原料としてのたばこ」に関する情報が整理され，－の方向に「製品としてのたばこ」に関する情報が整理された。また，第2軸には＋方向に，たばこ植物の成長調整に関する情報が整理され，－方向に生化学関連の情報が整理された。

第1表　ランクｒの上位100語

ランクr	フリーキーワード	頻度	ランクr	フリーキーワード	頻度	ランクr	フリーキーワード	頻度	ランクr	フリーキーワード	頻度
1	TOBACCO	1087	26	CONTROL	83	51	LUNG	49	76	DEOXYRIBONUCLEIC	32
2	OF	808	27	NICOTINE	83	52	VIRAL	47	77	GAS	32
3	AND	750	28	NICOTIANA	80	53	N	46	78	INFECTION	32
4	IN	465	29	LEAF	75	54	AS	44	79	METABOLISM	32
5	THE	365	30	ACID	73	55	DURING	43	80	INHIBITION	31
6	SMOKING	348	31	CONTG	73	56	CELLS	42	81	PROTOPLAST	31
7	SMOKE	347	32	LEAVES	73	57	DETERMINATION	42	82	TOXICITY	31
8	PLANT	318	33	BLOOD	72	58	DISEASE	42	83	AFTER	30
9	ON	232	34	EFFECTS	71	59	PREPN	41	84	FILTER	30
10	CIGARET	215	35	DETN	69	60	CONTENT	38	85	HORNWORM	30
11	ACIDS	143	36	REVIEW	68	61	REACTION	38	86	LEVELS	30
12	A	137	37	RIBONUCLEIC	68	62	REGULATORS	37	87	MONOXIDE	30
13	VIRUS	132	38	METAB	63	63	HORMONES	36	88	PRODUCT	30
14	BY	116	39	RNA	63	64	MANDUCA	36	89	STUDY	30
15	FOR	111	40	COMPN	60	65	SEXTA	36	90	TOMATO	30
16	FROM	108	41	ACTIVITY	59	66	BIOLOGICAL	35	91	CALLUS	29
17	FORMATION	99	42	ANALYSIS	59	67	PROTEINS	35	92	PROPERTIES	29
18	TO	99	43	WITH	58	68	SOIL	35	93	SUGAR	29
19	MOSAIC	98	44	TABACUM	57	69	CHARACTERIZATION	34	94	INSECT	28
20	CULTURE	93	45	STUDIES	56	70	INFECTED	34	95	SPHEROPLAST	28
21	PRODUCTS	91	46	GROWTH	54	71	SOME	34	96	DNA	27
22	EFFECT	89	47	NITROGEN	54	72	CHROMATOG	33	97	HUMAN	27
23	CELL	87	48	DEVELOPMENT	53	73	MATERIALS	33	98	SYSTEM	27
24	TISSUE	86	49	CHEMICAL	52	74	METHOD	33	99	WATER	27
25	PROTEIN	84	50	CARBON	50	75	COMPOSITION	32	100	AMINO	26

第2表　1,087文献の主分類の内訳

CA分類	件数	CA分類	件数	CA分類	件数	CA分類	件数
001	16	011	436	026	1	042	2
002	6	012	28	027	4	043	4
003	14	013	10	028	1	048	1
004	207	014	10	030	5	059	29
005	112	015	3	031	1	062	2
006	41	017	6	033	1	063	6
007	17	018	2	035	1	067	2
008	8	019	45	036	1	071	1
009	17	023	1	037	1	075	1
010	33	024	3	039	1	079	3
						080	2

2.2.2　領域Bの語（ランクｒの上位74語）による解析

変数として使用するフリーキーワードの数を増せば，よりくわしく情報を解析できる可能性があるので，領域Bに属した，頻度ｒの上位74語に範囲を広げ，2.2.1と同様の処理を行なった。

得られた固有値を大きい方から順に5個示す。

（　）の中は相関係数を示す。

1λ　0.4454　(0.6674)

2λ　0.3279　(0.5726)

3λ　0.2938　(0.5420)

4λ　0.2804　(0.5295)

5λ　0.2527　(0.5027)

領域Aの42語で行なった場合に比べ，固有値はいずれも高くなっている。しかし，領域Aの場合と同様にλが0.5以上の高い相関を持つ軸は得られなかった。1λ，2λに対応する要因カテゴリーの数量化の結果を，第1，2図に示す。大勢としては，領域Aの42語による場合と差はなかった。

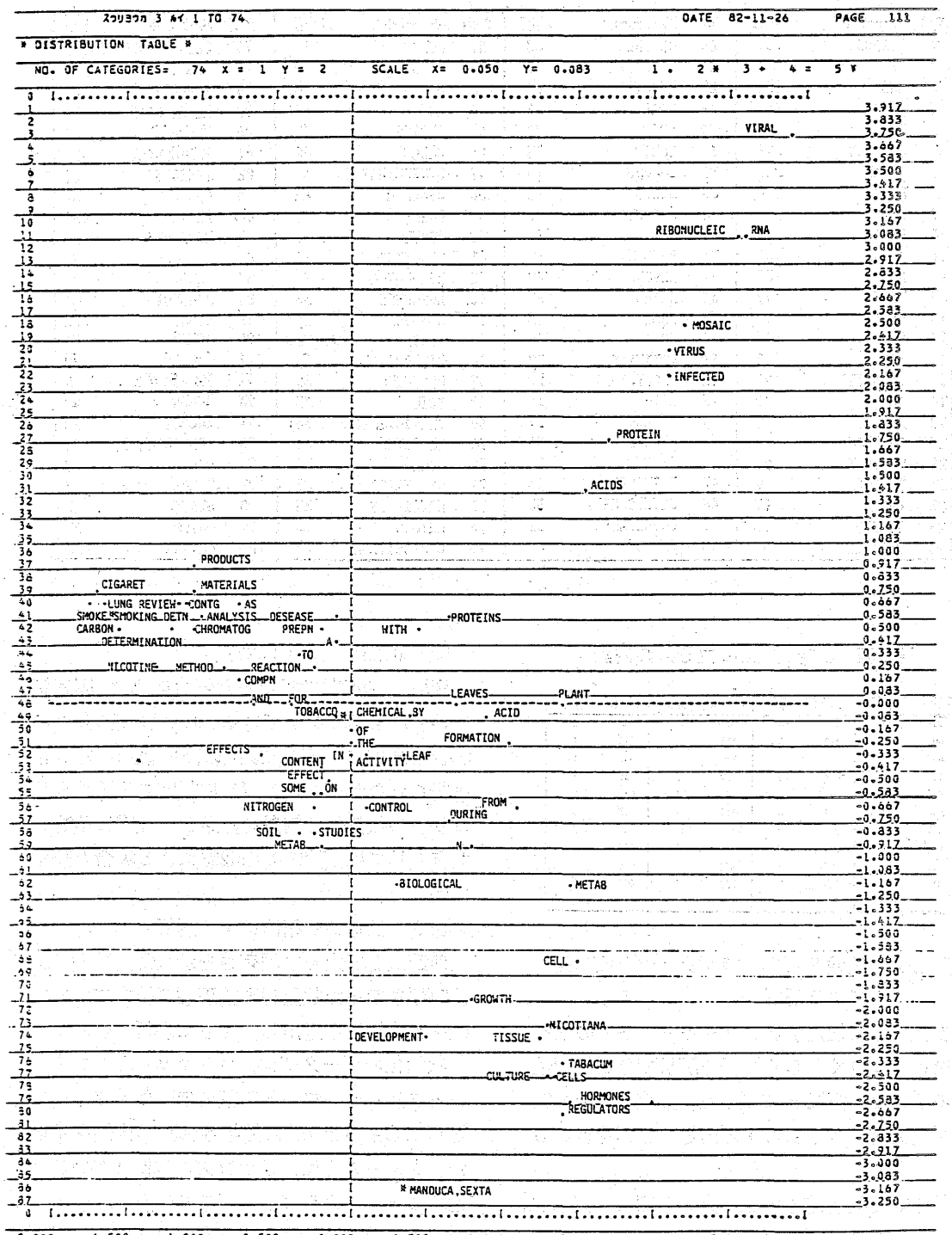
2.3　確率集中楕円による解析

前節2.2で大きな固有値を示した，頻度の上位74語を使った数量化理論III類の結果について，確率集中楕円を適用し，さらに解析を進めた。

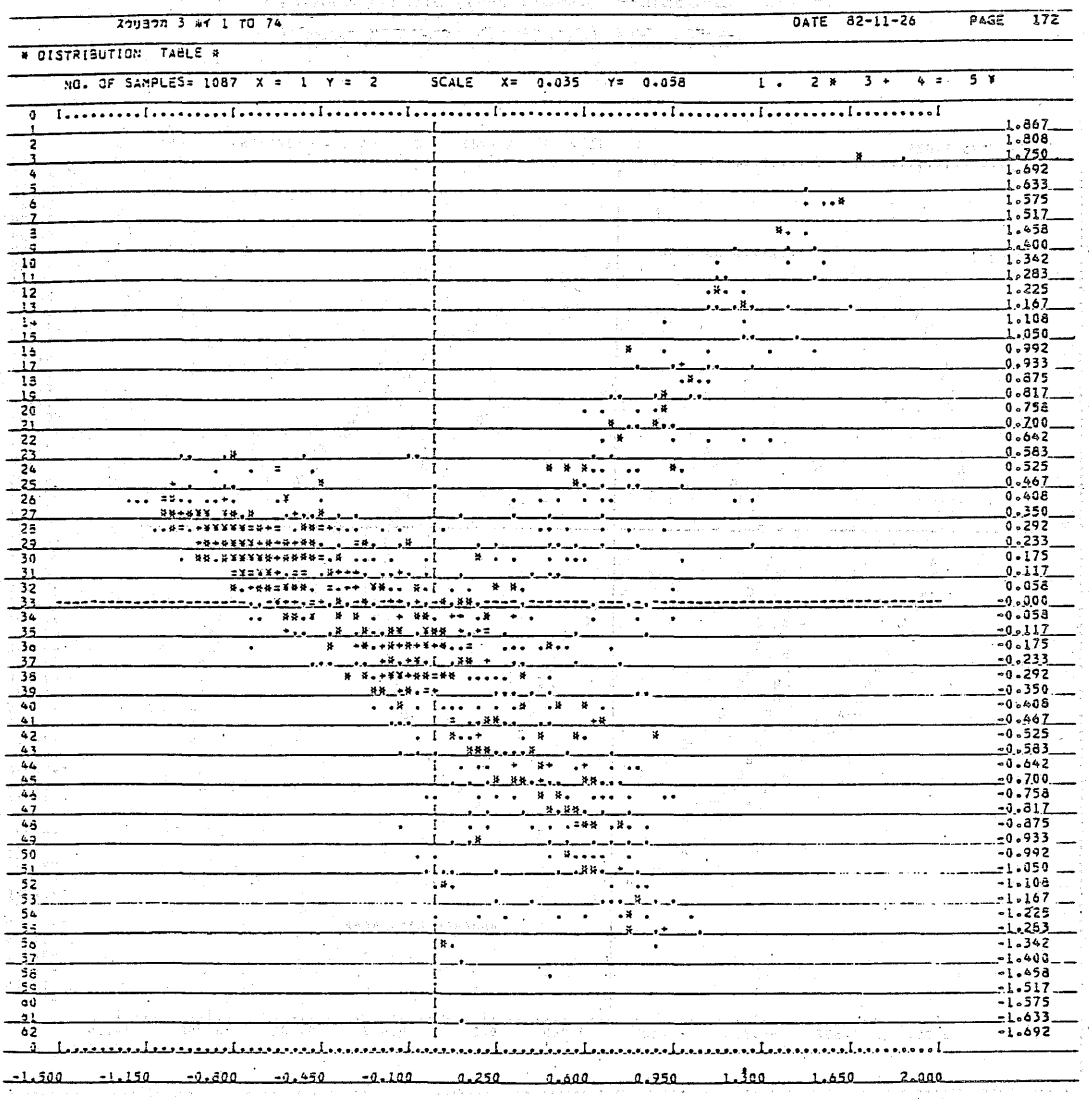
数量化のデータには，カテゴリーとしてもちいたフリーキーワード74種のほかに，フェースとしてCA Searchに付与された80分類を用いた。フェースとは，各サンプルが持っている特性のことで，カテゴリー（フリーキーワード）に対する反応とは違った観点で分析したい場合や，カテゴリーに対する反応との関係を明らかにしたい場合，有効に作用する。

確率集中楕円は，各集団の分布の状態を超楕円面として，定性的に相違点がわかるように楕

第１図　Ⅰ，Ⅱ軸に対する頻度の上位74語の散布



第 2 図　I, II 軸に対する1,087文献の散布（74語を使用した場合）

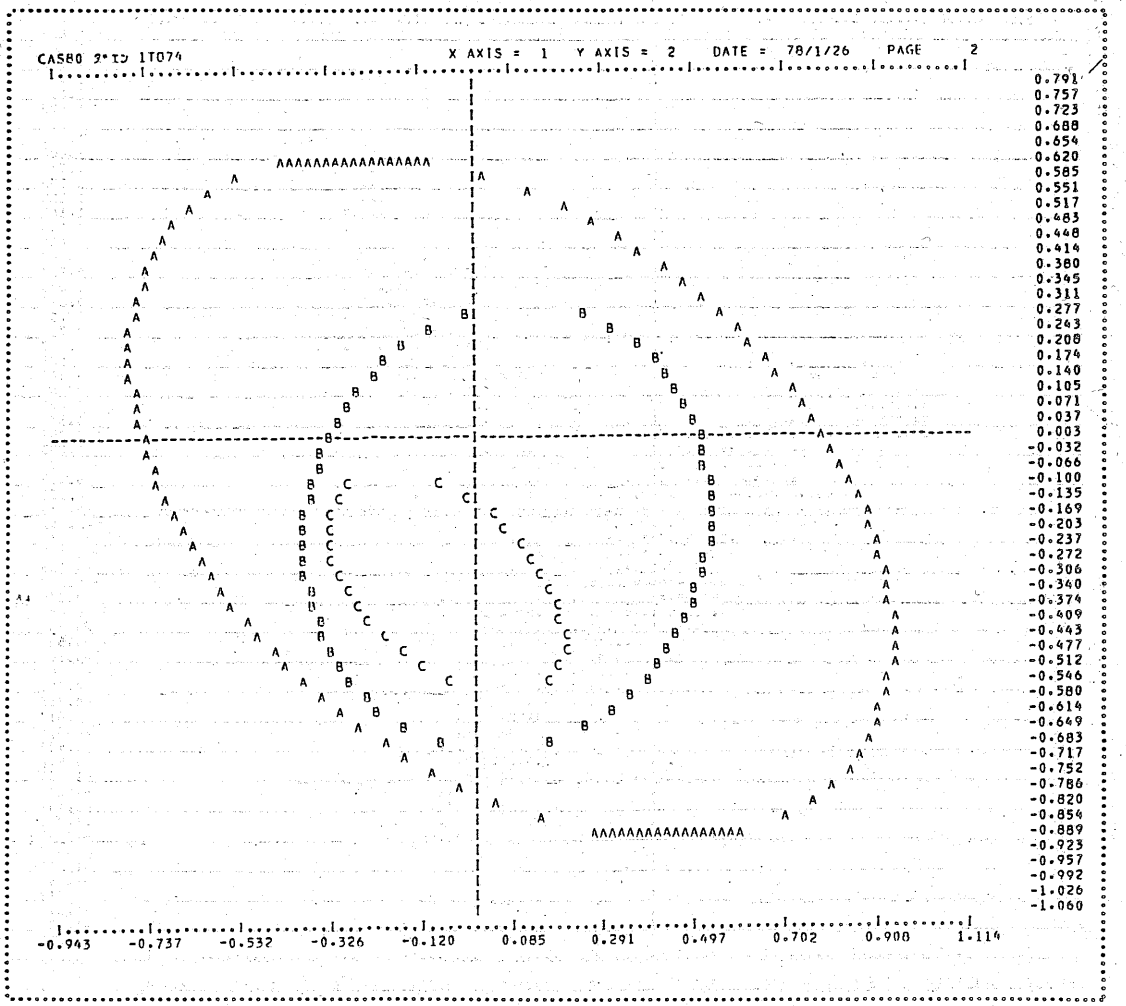


円面を作成し，2次変数の統計量として分布が捉えられている集団を，各集団が1つの楕円となるように，出力したものである。確率集中楕円の各軸を$\sqrt{2}$倍したものが50%の確率集中楕円となる。

分類は，副分類を除いた主分類だけを解析の対象とした。主分類の内訳は，第2表に示すとおりであった。確率集中楕円を描くには，最低3サンプル必要であるが，サンプルの件数が少ないと，正確な結果は得られず，不都合である。この報告では，サンプルの文献数が10件以上あったCA001，003，004，005，006，007，009，010，011，012，013，014，019，059の14分類を対象として解析を行なった。たばこ関連情報の1,087文献に付与された分類は全部で41分類あったが，分析の対象とした14分類は1,015文献に付与されており，全文献の93%強に相当するため，たばこ関連情報の全容を把握するのに十分な量と考えた。

数量化理論III類によって得られた，第I，II軸に関する各サンプルのスコアについて，フェース（分類）ごとの平均・標準偏差・サンプル数を求め，その数値を用いてフェースごとの確率集中楕円を求めた。得られた確率集中楕円から，たばこ情報の中における各フェース（分類）の関係を検討した。第III軸からは，解析の

第 3 図　植物生化学（CA011；A），農薬（CA005；B）
土壌・肥料（CA019；C）の確率集中楕円



対象に使用しなかった。

3．たばこ関連情報の確率集中楕円による解析結果

1,087文献に付与された主分類のうち，もっとも数が多くたばこ関連情報の主体をなすと考えられた分類は，植物生化学（CA011；現在のCA111に相当）で，この分類を持つ文献は436件存在した。CA011分類をもつ436件の確率集中楕円を，第 3 図に記号 A で示す（第 3 図の中で最大の確率集中楕円）。CA011セクションは，全情報の中心に位置する，たばこ関連情報の中核となる情報であることが確認された。

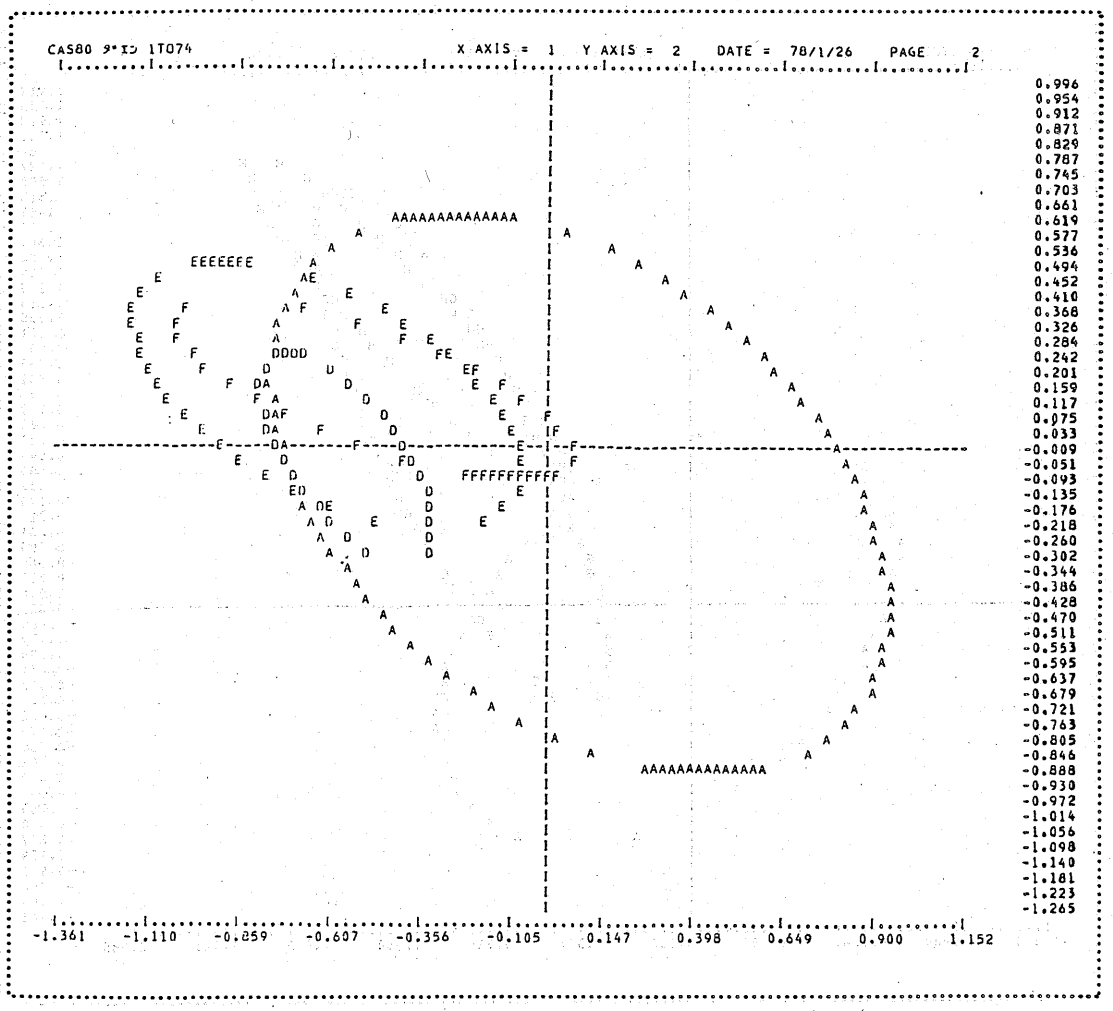
3.1　CA011セクションと重なりの大きなセクションの確率集中楕円

3.1.1　植物生化学（CA011；A），農薬（CA005；B），土壌・肥料（CA019；C）セクションの関係

CA011セクションが，たばこ関連情報の中核をなすので，各セクションの確率集中楕円と比較することによって，たばこ関連情報の中での，CA 分類の相互の関係について検討した。

CA005，CA019セクションの確率集中楕円を

第4図　植物生化学（CA011；A），薬理学（CA001；D），
毒物学（CA004；E），大気汚染（CA059；F）の確率集中楕円

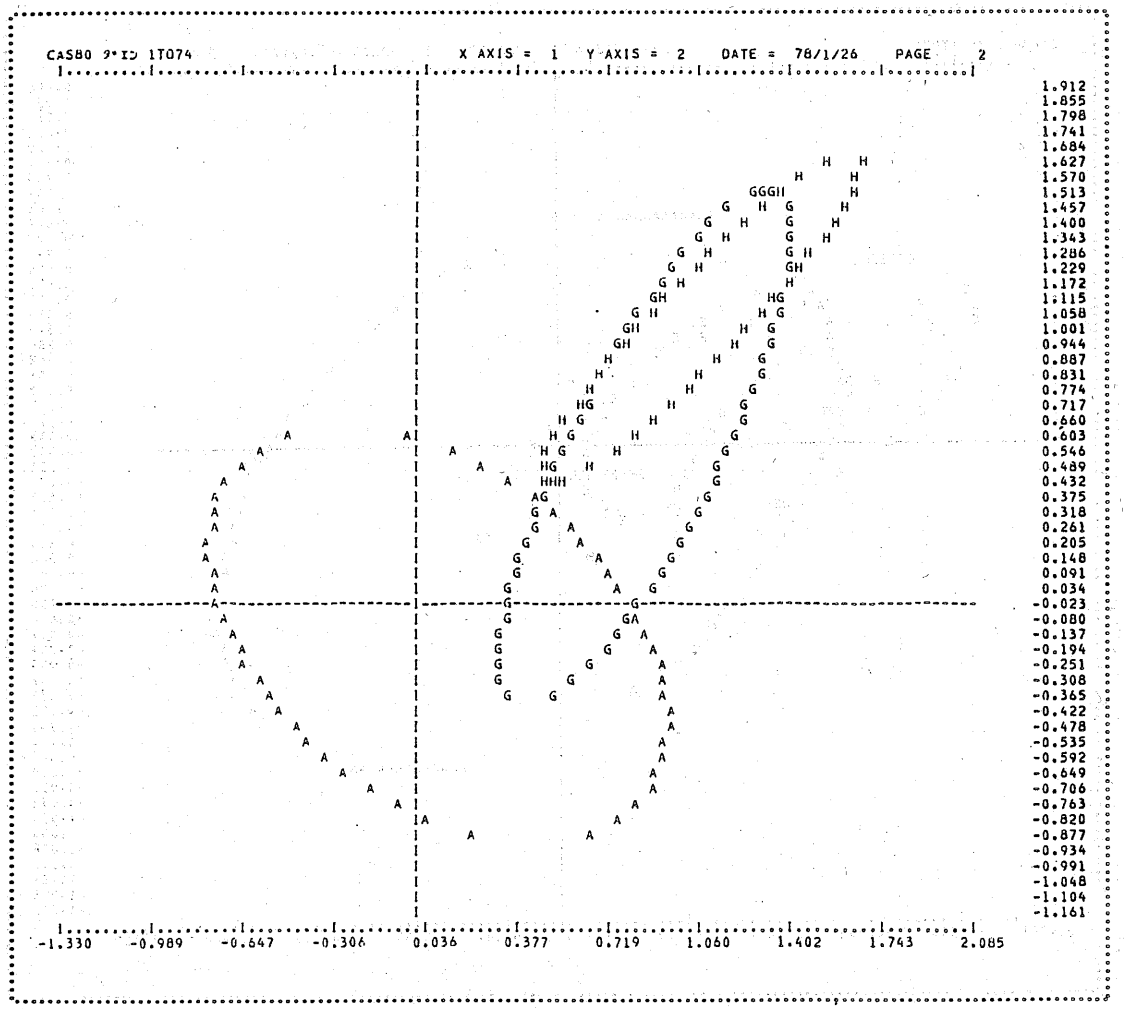


第3図に，記号ＢとＣで示す。植物生化学（CA011）＞農薬(CA005)＞土壌･肥料(CA019)の順で，重なることなく完全に三重の楕円が描かれた。これは，フリーキーワードによって，たばこ関連情報を整理すると，植物生化学（CA011）の情報に農薬（CA005）の情報が完全に含まれ，さらに農薬（CA005）の情報には，土壌・肥料（CA019）の情報が完全に含まれることを示している。このことから，たばこ関連情報の中核をなすのは，農薬と土壌・肥料との情報を含んだ植物生化学の情報であることが示された。

3.1.2　植物生化学（CA011；A）と薬理学（CA001；D），毒物学（CA004；E），大気汚染（CA059；F）セクションとの関係

CA011，CA001，CA004，CA059セクションの確率集中楕円を第4図に，記号A，D，E，Fで示す。薬理学（CA001；D）の確率集中楕円は，植物生化学(CA011；A)の確率集中楕円の左端の中にほぼ含まれた。いっぽう，毒物学（CA004；E）および大気汚染（CA059；F）の確率集中楕円は，CA011の確率集中楕円の左上の方向に，面積比で約1/3ずれているのが認められた。左上の方向は第1図から喫煙に関連のある情報が整理されていた方向である。よって，たばこ関連情報の中で一般技術分類は，薬理学（CA001)≦植物生化学（CA011）→大気汚染

第 5 図　植物生化学（CA011；A），微生物生化学（CA010；H），
生化学一般（CA006；G）の確率集中楕円



（CA059）→毒物学（CA004）といった位置づけにあることが明らかとなった。

3.2　CA011セクションとの重なりの少ないセクションの確率集中楕円

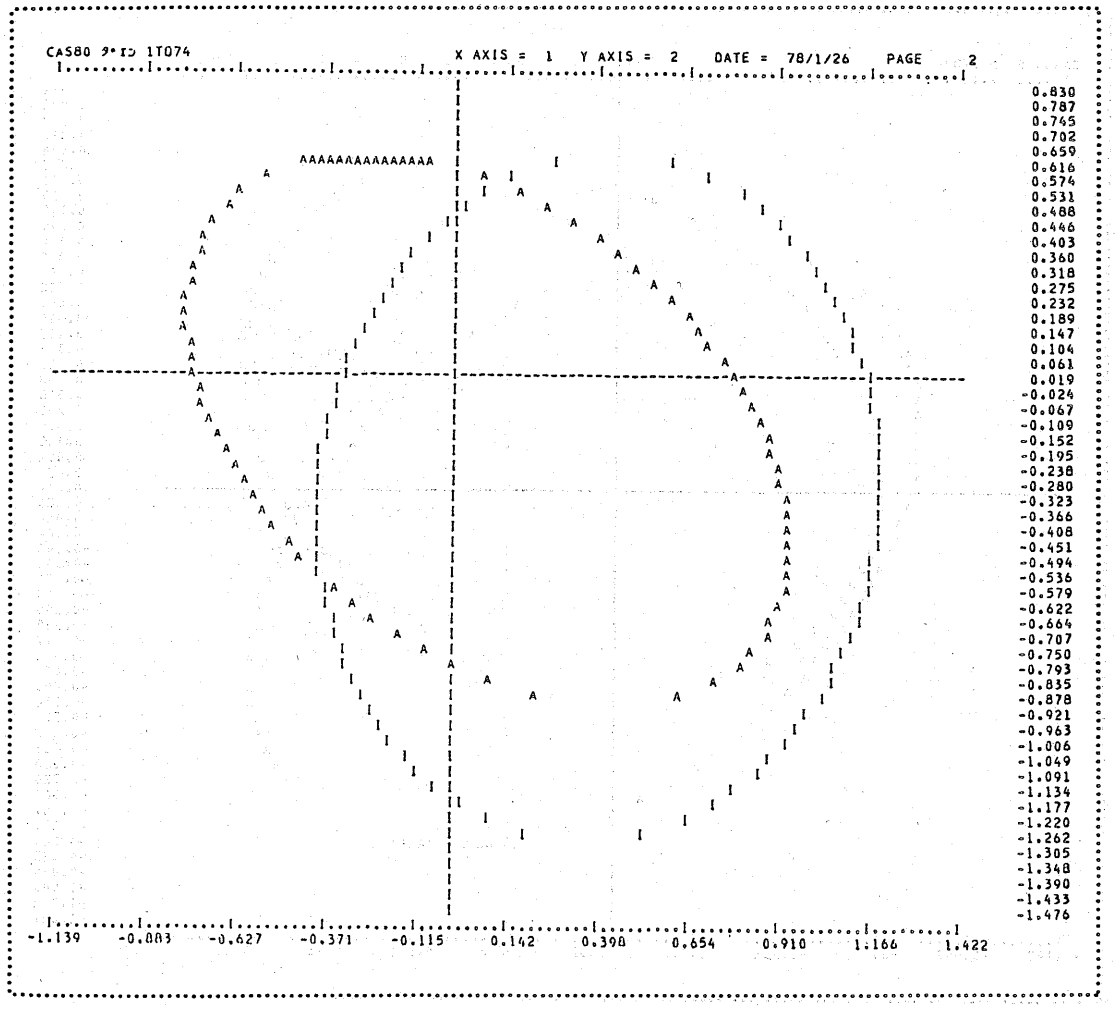
3.2.1　植物生化学（CA011；A）と微生物生化学（CA010；H），生化学一般（CA006；G）との関係

CA011と CA006, CA010の確率集中楕円を第5図に，記号A, G, Hで示す。植物生化学（CA011；A）と微生物生化学（CA010；H）の確率集中楕円は完全に分離した。このことは，両分類がたばこ関連情報のフリーキーワードから見て，完全に異なる技術分類であることを示している。いっぽう，生化学一般（CA006；G）の確率集中楕円は，CA011と CA010の確率集中楕円の中間に位置した。以上の結果から，一般技術分類はたばこ関連情報のなかで，植物生化学（CA011）→生化学一般（CA006）→微生物生化学（CA010）といった関係にあることが示された。

3.2.2　植物生化学（CA 011；A）および酵素（CA007；I），生化学実験法（CA 009;J），動物生化学（CA012;K）との関連

CA011と CA007, CA011と CA009, CA011と CA012の確率集中楕円を，それぞれ第6，7，

第 6 図　植物生化学（CA011；A），酵素（CA007；I）の確率集中楕円



8 図に記号 A，I；A，J；A，K で示す。

第 6 図のように，酵素（CA007；I）は，植物生化学（CA011；A）よりも右下方向にずれている。この方向には，成長調整に関連する情報が，整理されていたことから（第 2 図），植物生化学（CA011；A）よりも生物そのものに関する情報を多く含んでいた。

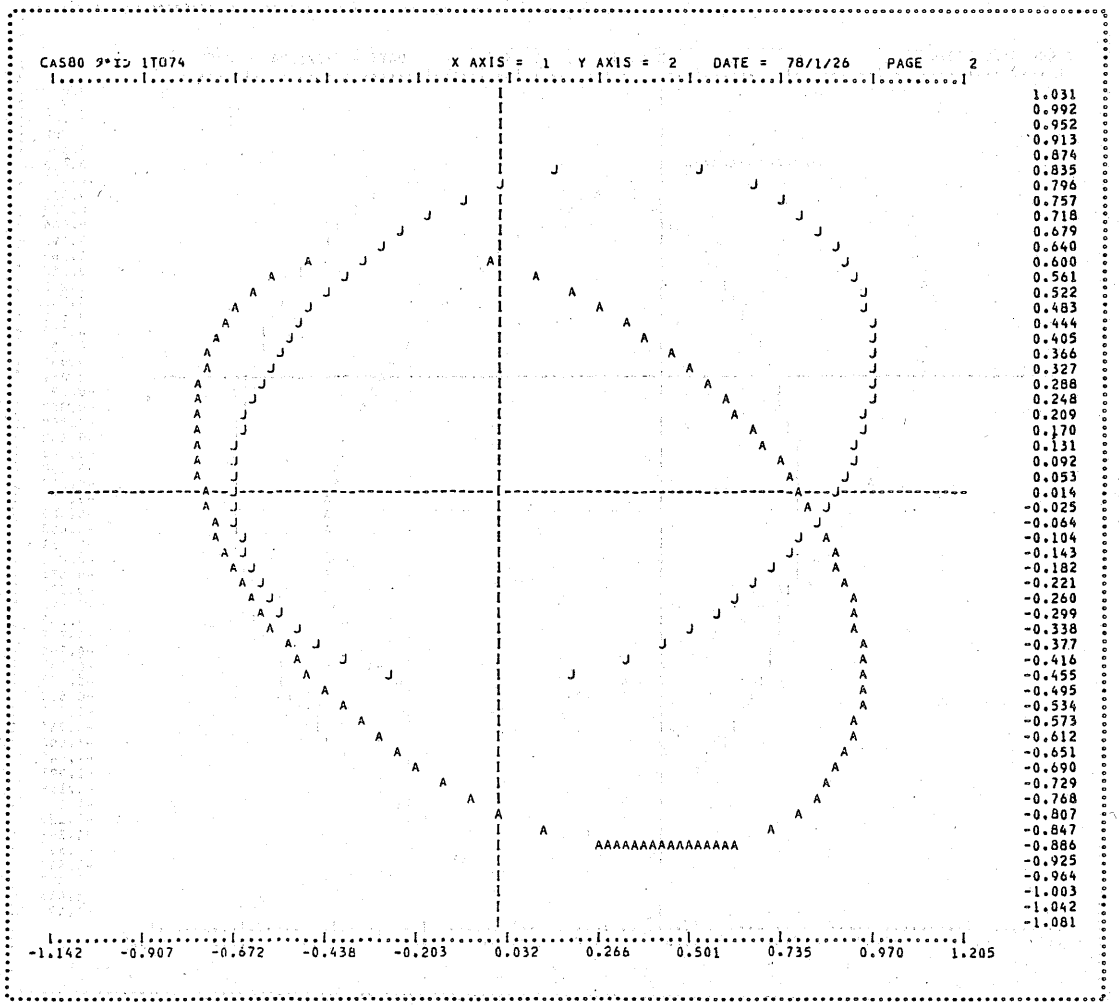
第 7 図の生化学実験法（CA009；J）の分類は，3.2.1 で述べた生化学一般（CA006；H）と同様の傾向を示した。その確率集中楕円は，植物生化学（CA011；A）と生化学一般（CA006；H）の中間に位置したところから，3.2.1 の結果と合わせると，植物生化学（CA011）→生化学実験法（CA009）→一般生化学（CA006）→微生物生化学（CA010）の順となり，たばこ関連情報における一般技術情報の位置づけが明らかとなった。

動物生化学（CA012；K）は，第 8 図のとおり植物生化学（CA011；A）の下方に狭い範囲の確率集中楕円を示したことから，特定の技術範囲の存在が示唆された。事実，動物生化学（CA012）の分類を持つ文献を調べたところ，これはたばこの害虫である Manduca sexta の生態などに関する情報で，たばこ関連情報の中では，他の情報から独立した情報の群を形成していた。

3.3　会社相互の関係

CA 分類の場合と同様，会社ごとに確率集中楕円を描き，相互の関連を考察した。主なたば

第 7 図　植物生化学（CA011；A），生化学実験法（CA009；J）の確率集中楕円



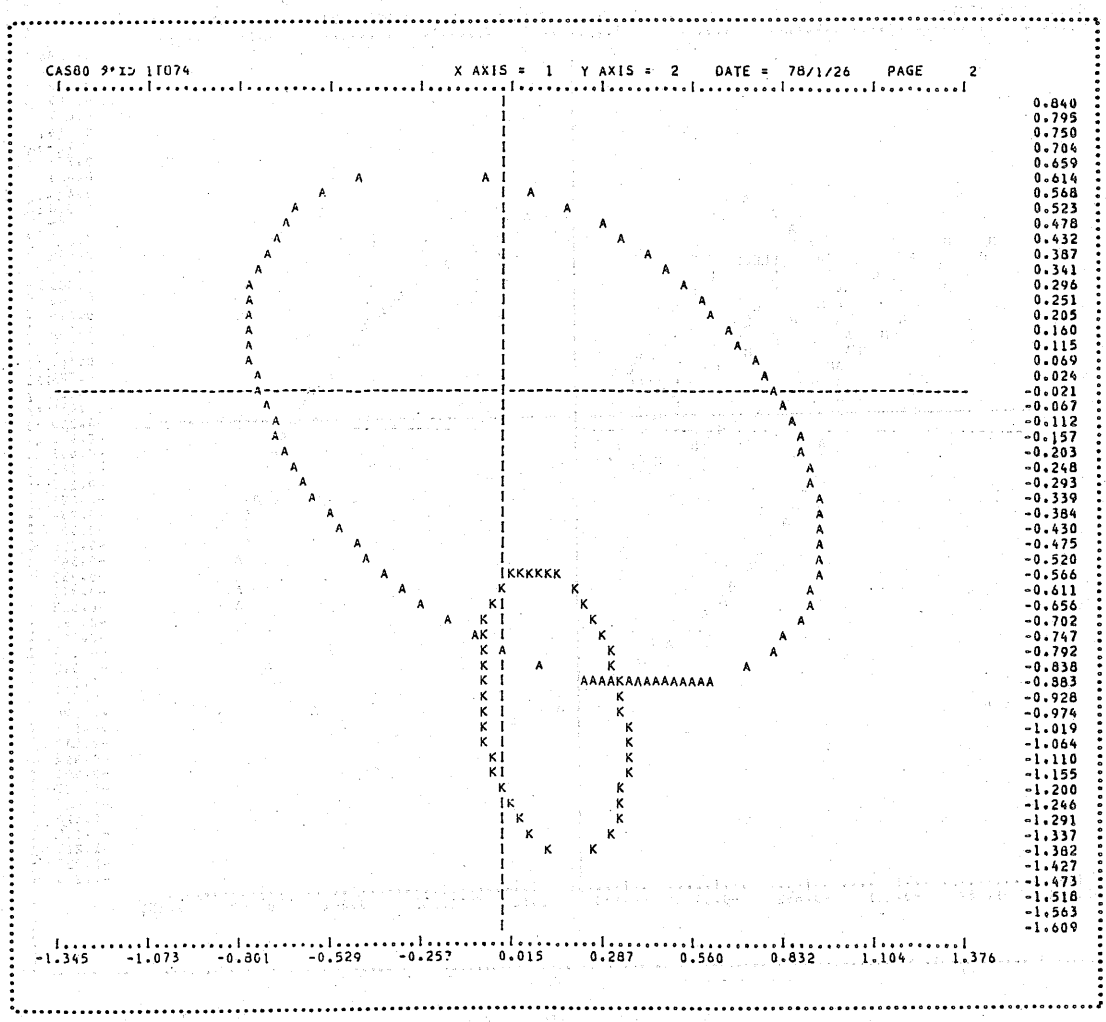
こ関連企業のうち，頻度が 5 以上のものは 5 社存在した。そのなかで，大きな分離を示す 2 社を，第 9 図に記号 L，M で示す。両者の文献の頻度は 5 件，8 件と多くないが，研究に対する興味が異なる会社であることがわかる。分類の場合と同様，基準とする会社を決めれば，相違点が明確になるであろう。

4．おわりに

たばこ関連情報に付与されたフリーキーワードをもとに情報を整理し，その結果を用いて，文献に付与されていた分類相互の関係を検討した。たばこ関連情報という特定情報の中で，一般技術分類である CA Search の分類の相互関係が明らかとなった。その結果，逆にたばこ関連情報全体がどのような一般技術によって構成されているかを知ることができた。また数量化理論Ⅲ類によって得られた，たばこ関連情報全体の構造を明らかにすることができた。なお，技術範囲が広いため，数量化理論Ⅲ類の Ⅰ，Ⅱ 軸に対するたばこ関連情報は“ねじれ”ている。このため，たばこ関連情報の中で等質な集団を明確にしたうえで，解析を進めれば，さらに詳細な結果が得られると考えられた。

最後に，データの作成に協力していただいた化学情報協会の時実象一氏，日本科学技術情報センターの小野寺夏生氏，そして当社コン

第 8 図　植物生化学（CA011；A），動物生化学（CA012；K）の確率集中楕円



ピューターセンターの前浜秀信氏，秋葉伸一氏，特許センターの住谷千秋嬢に感謝いたします。

参　考　文　献

1）高木義和：CA Search ファイルで使用されるキーワード，ドクメンテーション研究，32(10)　1982，p.p.465～471

2）高木義和：CA 情報に付与されたフリーキーワードによるたばこ関連情報の解析，ドクメテーション研究，33(7)，1983，p.p.309～318

3）水谷静夫：数量化理論とその応用，情報管理，21(9)　1978，p.p.692～706

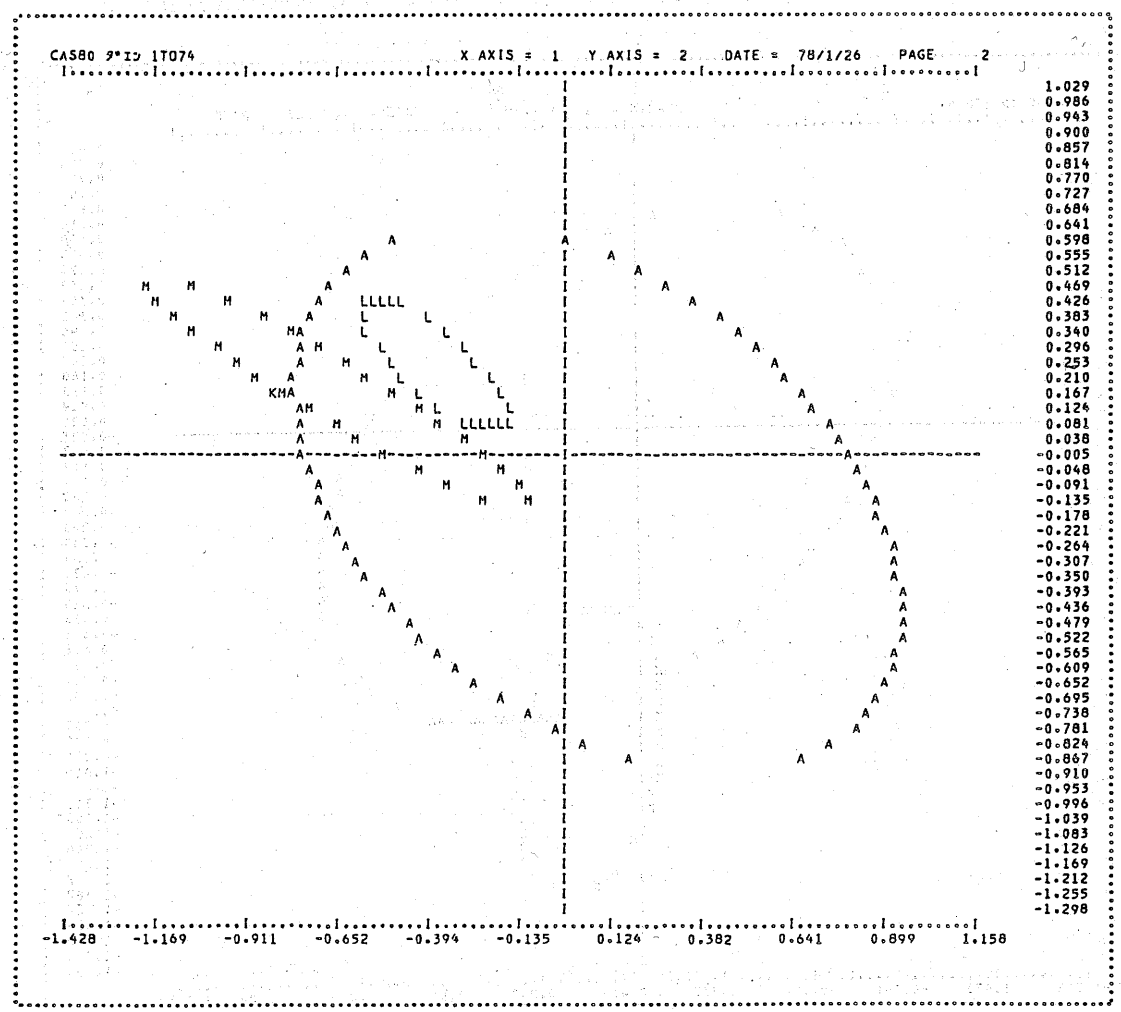
4）松尾雅嗣：テクスト処理プログラム LEX の文献情報処理への応用例，第18回情報科学技術研究集会発表論文集，日本科学技術情報センター，東京，1981, p.p.259～265

5）B. C. Brookes: Theory of the Bradford Law, J. Doc., 33(3)　1977, p.p.180～209

6）小野寺夏生：Bibliostatistics—情報現象の統計的説明—，情報管理，21(10)　1978，p.p.782～802

7）小野寺夏生・中井浩：単純なモデルからの Zipf の法則の導出，第12回情報技術研究集会発表論文集，日本科学技術情報センター，東京，1975，p.p.129～138

第9図　L社（5文献），M社（8文献）の確率集中楕円



質　疑　応　答

質問　平石繁雄（武田薬品工業）

クラスタ理論など，他の解析手法も考えられるが，数量化理論III類を用いた理由は何か。

回答　クラスタ分析を直接行なうと，得られた結果の解釈がむずかしい。まず数量化理論によって，主成分を2つから3つにしぼりたかった。そのごクラスタ分析を適用したいとは考えているが，数量化での問題点がいろいろ出てきて，そこまで至っていない。

質問　村瀬信一（雪印乳業）

CA Search を選んだ理由は何か。CAB の利用は考えなかったか。

回答　ここではフリーキーワードで解析しているが，これがうまくいかなかった場合，統制キーワードが解析の参考になるという考えがあったので，統制語がしっかりしている CA Search を選んだ。現状では，1ファイルで手一杯であるので，他のファイルまでは考えていない。

質問　長塚隆（紀伊国屋書店）

上位のキーワードに，冠詞・前置詞など，意味のないものが含まれているが，解析ミスの原因にならないか。

回答　ストップワードを除いての解析もやってみたが，除かないほうが現実に即した結果が得られた。この理由はよくわからないが，このように一見すると無意味な語も，文脈の中で何らかの役割を果たしているのではないかと思う。