自然言語処理　54－2
（1986. 3. 28）

多変量解析によるキーワードの自動抽出と文献の自動分類

竹内晴彦　（製品科学研究所）
岩坪秀一　（大学入試センター）
西野博二　（筑波大学）

1．はじめに

大規模な文献の蓄積と検索において、キーワードの自動抽出、及び、文献の自動分類は重要な課題である。キーワードの自動抽出の手法として、文の構造を考慮した解析的手法と、文の構造を考慮せずに単語の出現頻度のみに注目する統計的手法とがある。解析的手法は、対象に応じた細かい処理が可能であるが、大規模なシステムを必要とすることが多い。一方、統計的手法は、局所的な処理を施すことは困難であるが、事前に大規模な辞書を必要とせず比較的小規模なシステムで文献の取扱いが可能であり、文献の自動分類を行うための拡張も容易である。本研究では、統計的手法によるキーワードの抽出、及び文献の自動分類を行なう。特に、単語の局在性を示す指標を新たに導入することにより、識別力の高いキーワードを抽出し、多変量解析の一手法である判別分析を適用することにより合理的な文献の分類を試みる。

統計的手法による文献情報の取扱はLuhn(1957)に，さかのぼる。Luhnは、文献中の単語の出現頻度とその分布を統計的に調べることにより、その文献の内容についての手がかりが得られることを示した。Luhnは、この概念を重要な文の抽出と結びつけ、Science 誌中の１論文の自動要約を試みた。頻度順に単語を配列しなおした時、自動要約のために有効である単語が多く含まれる特定の区間を設定することができるとしたが、キーワードを抽出するには、このような頻度データだけでは無理があり、いくつかの分野についての相対的頻度の情報を利用する必要がある。

Maron (1961)は、90語のキーワードをあらかじめ選び、確率的手法により文献の自動分類を試みた。32グループで正しい分類率は30～50%という結果を得ているが、グループ数が多い割りにキーワード数が少ないため、キーワードを１つも含まない文献が数多くある。因子分析による分類の試み(Borko・Bernik, 1963)もあるが、Maronと同程度の結果を得るのにとどまっている。

西村・岩坪(1970)は、数量化３類を用いて、単語と文を多次元空間の点として位置付けた。そして、これらの単語や文を表す点の位置が、我々が直感的に認識している単語や文の意味の類似や対立とよく合致していることを示した。

Salton・Yu(1975)は、キーワードの数を次元数とするベクトル空間に文献を位置付けた。そして、自動索引を行うために、単語が何文献にわたって出現しているかを示すＤＦ(Document Frequency)に基づいて、有効な単語を選択した。Saltonらは、文献数をｎとする時、ＤＦ値が、n/100 からn/10の間にある単語が分類のために最も有効な単語であると結論づけている。

Hamil・Zamora (1980)は、Chemical Abstractsに含まれている文献を、タイトルのみをつかって、５グループ、及び、80グループへ自動分類することを試みた。正分類率は、５グループでは66.7%，80グループでは45%であった。しかし、データとしてタイトルを用いているため、分類規準を作るために大量のデータを必要とする。又、分類規準を作るために47,283文献を使い、キーワードとして16,366語を登録しても、新たな文献に対してその約10%の文献は、該当するキーワードを含んでいないため分類不可能となっている。

Dillon・Federhart(1982) は、よい検索語を同定することを試みた。Dillonらは、３グループに属する68語の統計的性質を調べることにより、新たな68語をこの３グループに分類することを試みた。この実験は、質問文の中からキーワードとなるものを統計的に抽出しようとする試みであったが、３グループの分類で約60%の分類率を得るのにとどまっている。

本研究は次の２点に重点をおく。第一に、キーワードを抽出するために単語の局在性を示す指標を導入し、さらに変数選択法を用いる。第二に、合理的な分類を行うために、次元の減小を伴う判別関数を用いて文献をユークリッド空間上の点として位置付けることである。

2．処理手続き

文献の自動分類は、次の２段階よりなる。第１は、すでに正しく分類されている文献データを使ってキーワードを抽出し、判別関数を導出することである。第２は、この判別関数を用い、分類が未知なる文献を実際に分類することである。

2．1　判別関数の導出

すでに分類されている文献からキーワードを抽出し、判別関数を導出するためのデータ処理について述べる。

まず、対象とする文献データから単語を切り出し、出現頻度の高い順に配列する。単語がある文献に含まれているかいないかに着目すると、元の文献は、１－０データとして表現できる。次に、定冠詞、指示代名詞等の、文献分類のためのキーワードとしては不適切な単語を除去する。このために２つの基準を設ける。第１は、単語長が４以下の単語は取り除くという基準である。第２は、式（２）で定義する単語の局在性を示す指標Ｌが0.5 以下の単語は取り除くという基準である。第１の規準は、単語長が４以下の単語はキーワードとして不適切なものが多いことによる。第２の規準は、どのグループにも同じように出現する単語は識別力が低いためキーワードとしては不適切であり、特定のグループに集中して出現する単語は、そのグループを特徴付けるのに有効なキーワードであるという前提に基づいている。

各単語について、各グループのＤＦ(Document Frequency)を求め、各単語ごとに最大ＤＦが１になるように正規化する。１つの単語のグループ別のＤＦをx_i(i=1,…,g）とすると、正規化後の値は式（１）で定まる。ただしｇはグループ数である。

$$y_i = \frac{x_i}{\max_j x_j} \quad (i=1, \cdots, g) \qquad (1)$$

指標Ｌを次式で定義する。

$$L = \frac{1}{g-1} \sum_{i=1}^{g} (1-y_i)^2 \qquad (2)$$

指標Ｌは、０以上１以下の値をとる。Ｌ値が１に近いほど、その単語は、特定の分野に集中して出現しているといえる。

続いて、出現頻度の高い単語を対象として、変数選択法を適用する。変数選択は、元来、ノイズとしての役割しかしていない変数を除去し、小数の変数で効果的な判別を行うことを目的とする。変数選択を行うことにより、統計的に識別力の高いキーワード集合を得ることができる。変数選択の手法としては、Wilks のΛ統計量に基づく方法、マハラノビス汎距離に基づく方法、Hotteling のT^2統計量に基づく方法、RaoのＶ統計量に基づく方法などがある。ここでは、比較的計算量の少ないWilks のΛ統計量に着目した変数選択法を適用し、文献分類に有効なキーワードを選択する。

こうして得たキーワードを変数として判別分析を行ない、判別関数を導出する。判別分析は、各郡内の分散を小さくし、郡間の分散を大きくするように、各変数に重みをつける手法である。第ｋ郡のｉ番目のサンプルの変数x_jに対する値を

$$\{x_{ij}^{(k)}\} \quad (i=1,\cdots,n_k;\ j=1,\cdots,p;\ K=1,\cdots,g) \qquad (3)$$

で表す（図１）。郡間分散共分散行列をＢ、郡内分散共分散行列をＷ、求める重みを

$$u = (u_1, \cdots, u_p)' \qquad (4)$$

とおくと、次の固有方程式が導かれる。

$$(B - \theta W)\, u = 0, \quad u'Wu = 1 \qquad (5)$$

郡	個体	変量 x_1 ···· x_j ···· x_p
G_1	1 … n_1	…
：	：	：
G_k	1 … i … n_k	····· $x_{ij}^{(k)}$
：	：	
G_g	1 … n_g	

図１　判別分析のためのデータ

式（5）を解き、

$$m = \min(g-1,p) \qquad (6)$$

と置くと、m個の0でない固有値が得られ、各々の固有値に対応する固有ベクトル（$u^{(k)}$）（k=1,･･･,m）が、判別関数となる。

2．2 新しい文献の分類

前節で得た判別関数と、判別空間上での各グループの重心の位置を、新しい文献の自動分類に用いる。まず自動分類したい文献が、前項で抽出したキーワードを含んでいるかどうかを調べる。この文献を1－0データで表して

$$x=(x_1,x_2,\cdots,x_p)' \qquad (7)$$

と置く。次にm個の判別関数より成る行列U

$$U=(u^{(1)},u^{(2)},\cdots,u^{(m)})' \qquad (8)$$

を用いて、この文献を判別空間上の点として位置づけると、文献のm次元空間上の位置yは次式で表される。

$$y=Ux \qquad (9)$$

この位置と、各グループの重心との距離を比較することにより、この文献の所属グループを判定することができる。

3．実験

3．1 データについて

データとして、データベースCOMPENDEX に含まれているIEEEの文献データを用いる。本研究では、IEEEの1つのトランザクションを1つのグループと考え、3グループの場合、及び、5グループの場合について、キーワードの自動抽出、判別空間の構成、及び文献の自動分類を行なう。本研究で用いるグループを表1、及び、表4に示す。3グループの場合は、判別関数を構成するために合計450文献、検証のために合計150文献用いる（表1）。5グループの場合は、判別関数を構成するために合計750 文献、検証のために合計250 文献用いる（表4）。

各文献の処理対象として、アブストラクトとタイトルがよく用いられる。統計的手法によりタイトルのみを用いた処理を施すには、極めて大きなキーワードリストを作る必要がある。本研究では複数のキーワードの組合せにより効果的な分類を行うのでアブストラクトを対象とする。

3．2 3グループの場合

＜判別関数の導出＞

すでに所属のわかっている450 文献（表1）を使って、キーワードを選択し、判別関数を導出する。この450 サンプルの総単語数は41992 語で、1つのアブストラクトは、平均93語からなる。

前処理として単語長が4以下、及びL値が0.5 以下の単語を除去し、頻度順に上位150 語の単語を選ぶ。この150 語をデータとして、Wilks の変数選択法を用いて90語の単語を選択し、文献分類のためのキーワードとする。選ばれたキーワードを表2に示す。

この90語を変数として、判別分析を行なう。3グループの判別なので独立な判別関数は2個得られる。第1固有値の寄与率は60%、第2固有値の寄与率は40%であった。判別関数を導出するのに用いたサンプルを判別空間上で分類すると、450 サンプル中21サンプルが誤分類された。見かけ上の誤分類率は、4.7 %ということになる。

表1 グループ名とサンプル数（3グループの場合）

No.	グループ名	サンプル数	
		判別関数の導出	新しい文献
1	Acoustics, speech and signal processing	150	50
2	Magnetics	150	50
3	Plasma sciences	150	50
	合計	450	150

表２　自動抽出されたキーワード

1	MAGNETIC	31	APPROACH	61	CHANGES
2	PLASMA	32	ADAPTIVE	62	MODES
3	CURRENT	33	APPLICATIONS	63	PERMEABILITY
4	ALGORITHM	34	ESTIMATION	64	STATE
5	TEMPERATURE	35	WIDTH	65	STELLARATOR
6	DIGITAL	36	DOMAIN	66	SYNTHESIS
7	FILTER	37	DYNAMIC	67	THICKNESS
8	PARTICLES	38	FILMS	68	DETERMINE
9	SPEECH	39	SIGNALS	69	PERPENDICULAR
10	NOISE	40	DEVICE	70	CLASS
11	MAGNETIZATION	41	TRANSFORMS	71	ELECTRIC
12	TRANSFORM	42	LAYER	72	ALLOYS
13	SPECTRAL	43	CIRCUIT	73	ANALOG
14	FILTERS	44	FLUIDS	74	BOUNDARY
15	BUBBLE	45	INCREASES	75	CALCULATIONS
16	FLUID	46	REQUIRED	76	DEFINED
17	LINEAR	47	SHAPE	77	MATERIALS
18	RECORDING	48	SPATIAL	78	MODULAR
19	MINUS	49	SWITCH	79	SAMPLING
20	DEGREE	50	VOLTAGE	80	AXIAL
21	ELECTRON	51	FILTERING	81	EMISSION
22	ENERGY	52	IMAGE	82	HEADS
23	NUMBER	53	ANODE	83	LASER
24	PROCESSING	54	DISCRETE	84	ANALYTICAL
25	COEFFICIENTS	55	INCLUDED	85	AUTOCORRELATION
26	PRESSURE	56	WAVES	86	CONVOLUTION
27	MATERIAL	57	HARDWARE	87	DEPENDENCE
28	FOURIER	58	TERMS	88	ESTABLISHED
29	RECOGNITION	59	VACUUM	89	FERROFLUIDS
30	PLASMAS	60	ARITHMETIC	90	MAGNET

表３　誤分類率（３グループの場合）

	見かけ上の値	真の値
サンプル数	450	150
誤分類されたサンプル数	21	24
誤分類率	4.7%	16%
キーワードを１以上含むサンプル数	441	145
誤分類されたサンプル数	17	21
誤分類率	3.9%	14.5%

＜新しい文献の分類＞

新しい150 文献を自動分類し、正しく分類できるかどうかを調べる。まず、対象とする文献から表２に示したキーワードを抽出し、前項で求めた２つの判別関数を使い、文献を判別空間上の点として位置づけ、所属グループを判定する。この結果、150 文献中24文献が誤分類された。したがって、真の誤分類率は16%となる。この結果を表３に示す。ここで、キーワードを１つ以上含む文献に限れば、真の誤分類率は14.5%となる。

3．3　5グループの場合

＜判別関数の導出＞

すでに所属のわかっている750文献（表4）を使って、キーワードを選択し、判別関数を導出する。まず、出現頻度が高い順に単語を並べかえる。上位40語の単語を表5に示す。前処理として単語長が4以下及びL値が 0.5 以下の単語を除去し、頻度順に上位250語の単語を選ぶ。表5に示した単語については、＊印をつけた単語が選択

表4　グループ名とサンプル数（5グループの場合）

No.	グループ名	サンプル数	
		判別関数の導出	新しい文献
1	Acoustics, speech and signal processing	150	50
2	Magnetics	150	50
3	Plasma sciences	150	50
4	Computer	150	50
5	Software engineering	150	50
	合計	750	250

表5　単語リストの一部（頻度順）

		DF OF GROUP						
NO.	WORD	1	2	3	4	5	FREQUENCY	L
1	THE	147	148	148	140	144	5041	0.0009
2	OF	142	149	144	144	143	3412	0.0015
3	A	138	134	128	140	139	2191	0.0024
4	AND	126	136	137	134	133	1788	0.0020
5	IS	135	108	120	133	139	1686	0.0178
6	TO	109	121	111	115	124	1375	0.0079
7	IN	106	123	117	118	111	1261	0.0082
8	FOR	116	100	90	118	117	1061	0.0200
9	ARE	104	95	88	103	105	927	0.0089
10	AB	150	150	150	150	149	749	0.0000
11	REFS	147	145	148	148	146	735	0.0002
12	WITH	72	96	78	68	66	588	0.0701
13	BY	78	76	64	64	67	547	0.0212
14	THAT	72	70	59	67	70	504	0.0097
15	AN	60	56	59	71	73	449	0.0309
16	WHICH	57	51	52	71	63	402	0.0506
17	BE	59	53	36	56	69	398	0.0848
18	ON	56	63	57	58	56	392	0.0100
19	AS	52	60	46	48	56	389	0.0292
20	THIS	55	38	36	61	69	335	0.1213
21	IT	49	40	37	57	51	307	0.0607
22	＊MAGNETIC	0	95	30	5	0	260	0.8414
23	PRESENTED	63	23	37	44	49	252	0.1784
24	CAN	35	31	28	36	50	238	0.1266
25	SYSTEM	22	26	12	32	48	224	0.2943
26	DATA	28	19	10	21	47	211	0.3610
27	FROM	39	44	40	19	30	210	0.1113
28	AT	25	37	48	15	23	206	0.2565
29	BEEN	23	52	46	19	14	206	0.3153
30	RESULTS	43	36	45	27	20	203	0.1277
31	HAS	35	38	32	27	25	193	0.0580
32	＊FIELD	5	59	42	2	1	188	0.7051
33	USED	37	29	23	29	37	186	0.0592
34	THESE	25	36	28	27	36	185	0.0513
35	MODEL	13	25	29	18	28	172	0.1171
36	OR	27	33	17	27	29	170	0.0790
37	DOUBLE	13	16	5	16	13	164	0.1357
38	METHOD	35	14	21	23	24	161	0.1841
39	SHOWN	44	19	14	35	24	158	0.2590
40	USING	40	28	23	21	19	158	0.1930

された。こうして得た 250語の一部（50語）を表６に示す。この 250語をデータとして、Wilks の変数選択法を用いて 150語の単語を選択し、文献分類のためのキーワードとする。選ばれたキーワードを表７に示す。

このようにして得たキーワード 150語を変数として、判別分析を行なう。５グループの判別なので独立な判別関数は４個得られる。第１固有値の寄与率は40%、第２固有値の寄与率は25%、第３固有値の寄与率は23%、第４固有値の寄与率は12%であった。判別関数を導出するのに用いたサンプルを判別空間上で分類すると、750 サンプル中65サンプルが誤分類された。見かけ上の誤分類率は、8.6 %ということになる。

表６　単語リストの一部（前処理後）

NO.	WORD	DF OF GROUP 1	2	3	4	5	DF	L
1	MAGNETIC	0	95	30	5	0	130	0.8414
2	FIELD	5	59	42	2	1	109	0.7051
3	ALGORITHM	37	2	0	23	20	82	0.5623
4	PROGRAM	7	6	2	6	46	67	0.7865
5	PLASMA	0	0	59	0	0	59	1.0000
6	SIGNAL	35	10	1	8	1	55	0.7482
7	CURRENT	1	14	38	3	5	61	0.7374
8	ALGORITHMS	15	0	0	25	15	55	0.5800
9	NETWORK	1	0	1	23	16	41	0.7306
10	DIGITAL	37	4	0	13	0	54	0.8041
11	FREQUENCY	25	11	15	2	2	55	0.5416
12	PARALLEL	5	1	7	18	8	39	0.5239
13	PROCESSING	22	5	0	14	7	48	0.5486
14	IMPLEMENTATION	20	0	0	12	20	52	0.5400
15	TEMPERATURE	0	22	16	0	0	38	0.7686
16	TRANSFORM	25	1	4	7	0	37	0.7864
17	FILTER	36	2	1	2	0	41	0.9323
18	FAULT	0	0	2	25	4	31	0.8880
19	SOFTWARE	0	0	0	3	30	33	0.9525
20	TESTING	4	3	1	14	15	37	0.5133
21	DENSITY	5	17	23	1	0	46	0.6489
22	ERROR	19	6	0	4	4	33	0.6787
23	PROGRAMS	2	2	0	6	27	37	0.8299
24	NOISE	23	4	1	1	0	29	0.8781
25	PARTICLES	0	23	3	0	0	26	0.9390
26	SPEECH	29	0	0	2	1	32	0.9498
27	FAULTS	0	0	0	23	3	26	0.9390
28	INPUT	17	0	2	18	6	43	0.5594
29	LINEAR	27	7	4	3	4	45	0.6975
30	MULTIPLE	4	4	3	21	8	40	0.6071
31	NETWORKS	0	1	0	21	7	29	0.8379
32	BUBBLE	0	20	0	5	0	25	0.8906
33	EFFECTS	6	12	25	1	4	48	0.6188
34	LOGIC	0	0	0	27	3	30	0.9475
35	MAGNETIZATION	0	34	0	0	0	34	1.0000
36	SINGLE	2	9	3	21	10	45	0.5385
37	SPECTRAL	23	1	2	0	0	26	0.9371
38	CIRCUITS	2	2	2	24	0	30	0.8802
39	FILTERS	20	0	0	0	0	20	1.0000
40	LANGUAGE	0	2	0	6	30	38	0.8778
41	DEGREE	3	18	1	3	3	28	0.7438
42	SCHEME	7	0	1	12	8	28	0.5312
43	FLUID	0	22	3	0	0	25	0.9365
44	MEASUREMENTS	3	19	15	1	0	38	0.6627
45	CIRCUIT	2	5	12	16	0	35	0.5752
46	MINUS	5	17	4	3	0	29	0.6903
47	DETECTION	7	1	0	12	4	24	0.6146
48	FIELDS	1	18	13	2	1	35	0.6628
49	OBSERVED	1	16	18	2	0	37	0.6736
50	PROGRAMMING	3	1	0	4	24	32	0.8446

表7　自動抽出されたキーワード

1	MAGNETIC	51	CLASS	101	SEQUENTIAL
2	FIELD	52	HARDWARE	102	SERVICE
3	ALGORITHM	53	STRUCTURES	103	SHAPE
4	PROGRAM	54	DOMAIN	104	MACHINE
5	PLASMA	55	MATERIAL	105	MACHINES
6	CURRENT	56	RECOGNITION	106	MOTION
7	ALGORITHMS	57	COEFFICIENTS	107	REALIZATION
8	NETWORK	58	FOURIER	108	SPECIFICATION
9	DIGITAL	59	NONLINEAR	109	CHANGES
10	FREQUENCY	60	RANGE	110	CORRECTNESS
11	PARALLEL	61	SWITCHING	111	MODES
12	PROCESSING	62	PRESSURE	112	SCHEMES
13	TEMPERATURE	63	REGION	113	SEQUENCES
14	FAULT	64	MULTIPLIED	114	TWO-DIMENSIONAL
15	SOFTWARE	65	COMMUNICATION	115	VACUUM
16	TESTING	66	TYPES	116	VECTOR
17	DENSITY	67	VELOCITY	117	ANALYTICAL
18	ERROR	68	GENERATION	118	CONCURRENT
19	PROGRAMS	69	TRANSFER	119	CORRECT
20	NOISE	70	ESTIMATION	120	EQUILIBRIUM
21	PARTICLES	71	OPTIMAL	121	MEASUREMENT
22	SPEECH	72	PLASMAS	122	METHODOLOGY
23	FAULTS	73	SURFACE	123	PERMEABILITY
24	INPUT	74	ADAPTIVE	124	RECOVERY
25	LINEAR	75	EXECUTION	125	STELLARATOR
26	BUBBLE	76	INCREASES	126	ACCURACY
27	EFFECTS	77	SIGNALS	127	CODES
28	LOGIC	78	TRANSFORMS	128	COMPUTING
29	MAGNETIZATION	79	MECHANISM	129	CONFIGURATIONS
30	SINGLE	80	THERMAL	130	DIRECTION
31	SPECTRAL	81	WIDTH	131	GRAPH
32	CIRCUITS	82	ACCESS	132	INSTABILITY
33	FILTERS	83	FILMS	133	LANGUAGES
34	LANGUAGE	84	SPACE	134	LOGICAL
35	DEGREE	85	VOLTAGE	135	PATTERN
36	FLUID	86	APPROXIMATELY	136	PERPENDICULAR
37	MEASUREMENTS	87	IMAGE	137	RECENTLY
38	CIRCUIT	88	INTEGRATED	138	CALCULATIONS
39	MINUS	89	PULSE	139	COLUMN
40	FIELDS	90	SEQUENCE	140	DIAGNOSTIC
41	OBSERVED	91	SYNTHESIS	141	ELECTRIC
42	PROGRAMMING	92	HIGHER	142	INTENSITY
43	PROPAGATION	93	LAYER	143	MEMORIES
44	RESPONSE	94	SPATIAL	144	OVERHEAD
45	RECORDING	95	SUGGESTED	145	PASCAL
46	PROCESSORS	96	SWITCH	146	SAMPLING
47	ERRORS	97	ARRAY	147	TABLE
48	DATABASE	98	BINARY	148	VIRTUAL
49	ELECTRON	99	BOUNDS	149	WALLS
50	ENERGY	100	COMPUTATIONAL	150	ALLOYS

＜新しい文献の分類＞

新しい250 文献を自動分類し、正しく分類できるかどうかを調べる。まず、対象とする文献から表7に示したキーワードを抽出し、前項で求めた4つの判別関数を使い、文献を判別空間上の点として位置づけ、所属グループを判定する。この結果、250 文献中74文献が誤分類された。したがって、真の誤分類率は29.6%となる。この結果を表8に示す。ここで、キーワードを1つ以上含む文献に限れば、真の誤分類率は28.3%となる。

表８　誤分類率（５グループの場合）

	見かけ上の値	真の値
サンプル数 誤分類されたサンプル数 誤分類率	750 65 8.6%	250 74 29.6%
キーワードを１以上含むサンプル数 誤分類されたサンプル数 誤分類率	743 61 8.2%	243 69 28.3%

４．おわりに

大容量のＲＡＭやＣＤ－ＲＯＭの実用化に伴って、キーワード抽出、自動分類、自動索引付けなどの文章解析の研究は今後益々重要になるものと思われる。

本論文では、多変量解析によるキーワードの自動抽出と文献の自動分類について述べた。キーワードを自動抽出するために、単語の局在性を示す指標Ｌを新たに導入し、さらに変数選択法を用いた。文献の自動分類に判別分析を適用することにより、３グループの場合約85%、５グループの場合約70%の文献を正しく分類した。今後の課題として、統計的手法と解析的手法との融合したシステムを構成することが挙げられる。

[謝辞]

日頃御討論頂く筑波大学の板橋秀一助教授はじめパターン認識研究室の皆様、及び東海大学の上村龍太郎講師に深く感謝いたします。

参考文献

1) Borko, H. and Bernick, M. : Automatic document classification, JACM, 10, pp.151-162 (1963)
2) Dillon, M. and Federhart, P. : The use of discriminant analysis to select content-baring words, JASIS, 33, pp.243-253 (1982)
3) Hamill, K.A. and Zamora, A.: The use of titles for automatic document classification, JASIS, 31, pp.396-402 (1980)
4) 細野公男・後藤智範・諸橋正幸：パターン・マッチングによる重要語の自動抽出，情報処理学会自然言語処理研究会資料，39-1 (1983)
5) Luhn, H.P. : A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Res. and Dev., 1, pp.309-317 (1957)
6) Maron, M.E. : Automatic indexing - An experimental inquiry, JACM, 8, pp.404-417 (1961)
7) 諸橋正幸：自動索引付けの研究の動向，情報処理，25，pp.918-925 (1984)
8) 西村恕彦・岩坪秀一：計算機による文献の自動分類，第６回情報科学技術研究集会発表論文集 (1969)
9) 西村恕彦・岩坪秀一：計算機意味論の実験，情報処理，11，pp.127-134 (1970)
10) Salton, G, Yang, C.S. and Yu, C.T.: The theory of term importance in automatic text analysis, JASIS, 26, pp.33-44 (1975)