漢字の出現頻度情報を用いた日本語文献の自動分類

細野公男* 後藤智範* 守屋智* 原田隆史*
諸橋正幸** °梅田茂樹**
（* 慶應大学文学部 ** 日本アイ・ビー・エム株式会社 サイエンス・インスティチュート）

1．はじめに

文献情報を有効に活用するためには、その管理、運用のためにキーワード、分類コード等の二次情報を付与する必要がある。日本語の文献を考える場合、その文中にはひらがな、カタカナ、英数字、漢字等の多種の文字が使用される。そのうち特に漢字は、文字自体に本来の意味があり、造語作用により類似概念をもつ語を生成する働きがある。例えば、「計」という字は、「計測」、「計器」、「計算」等の語を造るが、いずれも「計」という字の意味－何かものの大きさ等を計測するーを表現することに違いはない。

筆者らは、この様な漢字本来の意味に着目し、漢字の頻度情報を基に漢字を含む日本語の科学技術文献を分類する事を試みた。

2．日本語と漢字

日本語の文献中には、ひらがな、カタカナ、英数字等の多種の文字が含まれている。なかでも、漢字は最も種類が多く一字一字が意味をもっている。日本語の文章を分析し、その処理を行うために様々な観点から調査、研究が行なわれている。それらは、以下の様な例がある。

1) 漢字の出現頻度を大規模な形で調査、分析を行う。
2) 漢字一字が熟語の構成要素として果たす役割について分析する。
3) あらゆる漢字を統合し、漢字の体系化を図る。
4) 漢字を種々の観点から分析し分類する。

1)の様な調査は、漢字キーボードの設計、コード体系の作成を目的にしたものが多い。この例として、国立国語研究所の「現代雑誌の９０種の用語用事」「現代新聞の漢字」があげられる。[1] これらの調査は、雑誌・新聞という一般的な分野で使われている漢字の、種類、用法を解明する目的で行われた。この調査では漢字の読み方の調査を行うことにより、音訓読みの傾向を明らかにしている。

2)の例としては、田中による「漢字調査における統計的尺度の問題」がある。[2] ここでは、個々の漢字について、ある漢字が構成した語の数を全体の語の数で除したものをカバー率と定義し、漢字が構成できる語の数と出現頻度との関連を明らかにしている。その中で、出現頻度の高い漢字は、カバー率の高い漢字つまり熟語を多数にわたって構成しやすい漢字であると指摘している点が注目される。

3)の例としては、国文学資料館の田島らにより、全く意味、用語が同じでありながら、字体の異なる異体字を統合・整理しようとする漢字シソーラスの研究がある。[3]

4)の例としては、国立国語研究所の野村らによる「漢字のパターン分類」があげられる。[4] 野村は、先に述べた国研の調査結果をもとに、代表的な漢字についてなりたち、字画数、音訓率、使用率等の11の要素に着目し、林の数量化理論Ⅲ類の手法により分類した。この結果、使用率という頻度情報を基にした要因が、パターン分類するうえで最も有効であったと報告している。

筆者らは、科学技術文献を対象にそこに出現する漢字を、複数の分野ごとの出現確率からなる多変量変数に対応させ、これによって文献で使用された漢字を互いに近似した幾つかのクラスターに分類した。そこで得られた各クラスターは、各々関連性の強い分野と対応付けることができた。[5] この結果、特定の分野に関連の深い漢字を定量的に抽出し、漢字のもつ分類分野に対応する意味が、文献の分野毎の頻度で説明できることを明らかにした。

3．自動分類研究

本稿では、漢字の頻度特性を利用した文献の分類法を提案し、考察することを目的としているが、これについて議論する前に、文献集合を類似したカテゴリーに自動的に分類しようとする自動分類の研究についていくつか例をあげ、それらの特徴を整理する。

3.1 西村、岩坪らの分類法

通産省電気試験所（現在電子総合研究所）の西村、岩坪らは、特定の主題分野を与えられた文献群からその主題特性に基づいて分類体系を自動的に設定してゆく方法を提案した。[6] [7]

彼等は、文献を機械的に抽出可能な「要素」の集合と考え、その要素をキーにして文献の内容を識別できると考えた。彼等はこの「要素」として基本的な標本単語を設定し、二つの文献が似たようなキーワードの集まりから成っていたとき、これらの文献は似た内容を示す、すなわち同一カテゴリーに属しているものとしている。

何らかの方法で文献から単語を抽出し、この単語が各々の文献に含まれているか否かの単語と文献の相伴表を作り各単語と文献の反応を数量化する。彼らはこの数量化の方法として林の数量化理論Ⅲ類を適用した。

すなわち相伴表において似ている反応の文献同士を近づけ異なった反応の文献同士を遠ざける様に、単語と文献を並べかえる。これは、数学的には、単語と文献の相関が最大となる様に並べかえることを意味する。

標本単語としては、多くの分野に出現する単語を英文の例では100語、邦文の例では200語を選択した。

実験は、英文ではＪＡＬＭを、邦文では岩波の「科学」をデータとして用いて行なわれた。この方法の問題点として英語の語尾処理や同形異義語の扱い、標本単語の選択基準のあいまいさ等を指摘している。

3.2 Maronの研究 [8)]

各文献にキーワードが決められていて、充分な数のサンプリングの結果、あるカテゴリーに含まれる文献の語数のうちキーワード群の語数の比率と、全分野の文献数とカテゴリーに含まれる文献数の比率が、わかっているとする。

この時、ある語W_iを含む文献がある主題分野C_jに属する条件付き確率$P(C_j/W_i)$は、ベイズの定理により、

$$P(C_j/W_i) = \frac{P(C_j)\cdot P(W_i/C_j)}{P(W_i)}$$

で表すことができる。

Maronは、分類のキーとなりうる語をあらかじめ選定しておき、$P(W_i/C_j)$を調べたうえで先の式により、$P(C_j/W_i)$を全てのｊについて求める、その最大値を与えるC_jに文献をアサインする分類方法を提案した。

260の文献をトレーニングデータとして各$P(C_j/W_i)$を求め、これを用いて 145の文献の分類を行ったところ、51.8%の文献の分類が人手による分類結果と一致したと報告されている。

3.3 Hamil, Zamoraの研究 [9)]

Maronらがベイジアン・モデルを導入したのに対してZamoraらは、あるキーワードK_jを含む文献がある主題分野C_iに含まれる確率を条件付き確率の形で表現し、これを用いて文献の分類を試みている。

文献の付与されたキーワードの数をＮ個とした時、主題分野C_iに対してあるキーワード群が持つ分野識別力Y_iを、$Y_i = 1/N \times P(C_i/K_j)$ で定義する。与えられた文献をどの分野にわりあてるかは、全分野ｉについてY_iを計算した後に、Y_iが最大値を示した分野にわりあてることにする。

彼らは実験データとして、主題分野数80の Chemical Abstractの論文タイトルを用いた。　条件付き確率$P(C_i/K_j)$については、

$$P(C_i/K_j) = \frac{f_{ij}}{\Sigma f_{ij}}$$

の形で定義し、低頻度語、高頻度語をカットした後、全ての語についてこれを計算した。

実験の結果、サンプル 16089タイトルがいずれかの分野に割当てられ、このうち正答率、すなわちＣ．Ａ．Search の分類カテゴリーと一致したものは45%であった。

4．漢字をキーにする文献分類手法

4.1 漢字の分類カテゴリー別頻度解析

本稿で提案する分類手法は、日本語科学技術文献の論文タイトル中に出現する漢字の出現頻度特性を利用して、文献に分類カテゴリーを割当てるものである。そのためには、漢字の頻度情報を文献のカテゴリー志向性と対応づける必要がある。そこで、あらかじめ充分な量の文献データより漢字の分類カテゴリー別出現頻度表を作成した。これは、文献ファイル中の論文タイトル中に出現する全ての漢字について、どの分類カテゴリーに属する文献から抽出されたものかをカウントして表形式にしたものである。

図１にこれを示す。図１で、ｊ番目に総頻度順位ｊ番目の漢字K_jが、分類カテゴリーｉに属する文献中から抽出されたカウント数がｊ行目ｉ項目に記されている。　調査の対象としたデータは、ＪICST理工学文献ファイル（1983年版）のうち、電気工学編のテープ８巻分を用いた。このうち論文タイトルだけを抽出しその中に含まれる漢字を頻度調査の対象とした。これらの漢字は ＪICST漢字表に属する漢字で、当用漢字表から喜怒哀楽を表現するものを除き、さらに人名地名を加えた1884字であり、総出現数は251661字である。

4.2 文献分類法

ある漢字K_jが任意の分類カテゴリーｉに対してカテゴリー指向性S_{ij}をもっていると考える。すると漢字一字に対して、分類カテゴリーの数ｎだけの値が対応していることになる。S_{ij}をカテゴリー番号順に並べ、ベクトル表現にしたものを漢字一字ずつに対応させたものを漢字ベクトルK_jとする。

$$K_j = (S_{1i}, S_{2i}, \cdots, S_{ij}, \cdots, S_{nj})$$

（ｎは分類カテゴリー数）

具体的にS_{ij}を決定するにあたっては、漢字K_jが分類カテゴリー番号ｉに属する文献から抽出される確率によるものとする。

分類 / 漢字	1	2	3	・・・・	j	・・・・・・	t
1	f_{11}	f_{12}	f_{13}		f_{1j}		f_{1t}
2	f_{21}				f_{2j}		f_{2t}
・・・							
i	f_{i1}				f_{ij}		f_{it}
・・・							
t_i	f_{t1}				f_{tj}		f_{tt}

f_{ij} ：漢字 i のカテゴリー j に於ける出現頻度数

f_{tj} ：カテゴリー j の総延べ漢字出現数

f_{it} ：漢字 i の総出現頻度数

f_{tt} ：総延べ漢字出現数

図１．　漢字の分類カテゴリー別頻度表

今、漢字の分類カテゴリー別頻度表において、カテゴリー間の文献数の補正をしたのち、j番目の漢字K_jの出現総数がN_jであったとする。また、各カテゴリー毎の出現数m_{ij}であったとする。すなわち、漢字K_jは分類カテゴリーiの文献からはm_{ij}回カウントされ、全てのカテゴリー毎の総数がN_jであったことになる。これは次の式で表現される。

$$m_{ij} = N_j \quad (n\text{は分野カテゴリー数}) \qquad (1)$$

今、全てのjに対してm_{ij}をN_jで除した値を考える。これをP_{ij}で表せば、

$$P_{ij} = \frac{m_{ij}}{N_{ij}} \qquad (2)$$

$(i=1,2,\cdots,I,\ i:\text{カテゴリー数})$

となる。

(2)式の右辺の意味を考えれば、これはある漢字K_jが、ある文献から抽出されたものである確率を示している。この様な確率は１つの漢字について、分類カテゴリーの数だけ対応していることになる。

すなわち(1),(2)で定義したP_{ij}をS_{ij}と等価と考えれば、カテゴリー指向性S_{ij}を、カテゴリー別出現確率で定義することになる。同様に (2)で与えた値の補正値をS_{ij}と考えても良い。いずれにしても、S_{ij}の算出法としては、漢字K_jが主題カテゴリーiに対して指向性が強いと考えられるものほど、S_{ij}の値を大きくしてやるのが妥当と思われる。また、漢字の総出現頻度の大小に応じて適当に重み付けしてもよい。

文献タイトル中に含まれるj番目の漢字K_jがカテゴリーiに対して示す志向度をS_{ij}できめる時、文献の示すカテゴリーiに対して示す志向度R_iは、各漢字K_jに対応するS_{ij}を出現する漢字の分だけ加算したものと考える。すなわち、文献中に漢字がm ケ含まれているとしたとき、

$$R_i = \sum S_{ij} \qquad (3)$$

R_iは、１つの文献タイトルについてカテゴリー数 i だけ計算される。この値の最も大きいものを選んで、それを与える様なカテゴリー番号に文献を分類する。

すなわち、

$$R_{max} = \max(\{R_i\}) \qquad (4)$$
$$(i = 1,2,\cdots\cdots,n)$$

となるiに対応する分類カテゴリーを文献にわりあてる。文献のタイトル中に含まれる漢字の字数が多いほど平均的にみるとR_iは大きくなる。一方、もしタイトル中に一字しか漢字がなければ、その漢字一文字の出現確率で決まる$S_{ij}(S_{i1})$で分類カテゴリーが決定されることになる。　モデルではこの両者の差異は考慮されない。この点を確認するために、分類カテゴリー別に１レコード中に含まれる文字数、及び漢字数を算出し相互に比較してみた。これらの関係を表１に示す。分野７を除いては、各カテゴリー共１タイトル10～12文字の漢字が平均的に含まれている。

本モデルが考えている分類は、科学技術文献ファイル中の１分野に属する文献群をそのサブカテゴリーにわりあてるものである。したがって、分類対象にする文献はもともと似た性質をもっていると考えられるので、タイトルの長さ、含まれる漢字の割合も似ていると推測される。　したがって平均的に見れば、タイトル中の漢字の数の相違は、分類結果に与える影響はたいして大きいものではないと思われる。

表1．分類カテゴリー別の1レコード中の文字数・漢字数と漢字の割合

	1レコードの平均文字数	1レコードの平均漢字数	全文字中の漢字の割合(%)
1．計測工学・計測機器	23.13	10.74	46.48
2．電磁気学・光学	26.39	11.15	42.28
3．電子物性・磁性・光物性	31.25	12.27	39.27
4．生体工学	25.43	12.44	48.92
5．システム制御工学一般	24.90	11.66	46.84
6．制御工学	25.47	10.12	39.74
7．計算機方式ハードウェア	24.95	7.10	28.43
8．計算機利用技術	23.75	9.84	41.43
9．電気工学一般	24.39	11.19	45.87
10．電力工学	25.70	12.60	49.02
11．電子工学	25.60	9.81	38.32
12．通信工学	24.94	10.04	40.21
総　合	25.49	10.64	41.74

Zamora，Maronの分類法と同様、このモデルの特徴の一つは、全てのカテゴリーに対する近似度を相対的に比較することにより、分類カテゴリーを決定する。この時、非常に微妙な差でカテゴリーが決まるものや、明白な差が認められるものがある。値が非常に接近しているものについては、分類を見あわせる、つまりはっきりと分類できないものとして処理する方法も考えられる。

しかしながら、今回の実験は、分類結果を評価することが第一の目的ではなく、漢字の頻度を分類キーに使用することの可能性を追及することが第一義と考えている。したがって、今回は、最大値R_{max}をとるiだけの情報によって分類することにした。

提案するモデルでは、漢字のカテゴリー別の出現確率だけを評価し、漢字の絶対頻度の情報については特に考慮していない。Luhnは語の絶対頻度の情報が文章の意味内容を表現する重要語の抽出に有用であると仮定したが、[10] 筆者らは絶対頻度の情報を特に考慮しないことにした。すなわち、

1) どの程度の頻度をもって中程度の頻度とみなすかがあいまいである。Luhnもこの点については経験的にきめるものとしているが、本稿の様に、漢字単位で集計した頻度データから決定することは、単語レベル以上に難しいというのが第一の理由である。

2) 漢字の場合、単語の場合と違って、高頻度の漢字がカテゴリー間の分布が一様なストップワード的な存在であると一概にはいえない。例えば『電』という字はJICSTテープ6館分のタイトルデータの場合では総計4379回カウントされ頻度順位1位である。そのうち、カテゴリー番号10に属する文献からは、1795回カウントされたが、カテゴリー番号5に属する文献からは、わずかに8回しかカウントされていない。

したがって、Luhnのたてた経験則が、今回解析を行った漢字一文字の場合に関してもそのまま成り立つとは必ずしも言えないと思われる。

5．分類実験

5.1 テストデータのサンプリング

モデルの評価のための分類実験を構成に行うためには、適正なテストデータは無作為に抽出されたものでかつモデルの特性を表現しやすい様に考慮されていることがのぞましい。

又、モデルの評価として分類が成功するカテゴリーと成功しないカテゴリーをはっきりとわかる様にしたい。そして、そこからカテゴリーとしての特徴を考察できるようにしたい。

そこで、テストデータは、漢字の頻度解析のために使用した8巻分のJICST理工学文献ファイルとは別のテープから、各カテゴリーについて同数ずつ、ランダムにサンプリングした。

データ数は各カテゴリー毎45文献ずつ、12のカテゴリー毎に採取し、計 540文献とした。

表2．　サンプリング文献の分類実験結果

カテゴリー	再現率	カテゴリー番号													TOTAL
		1	2	3	4	5	6	7	8	9	10	11	12	NO	
1　計測工学　計測機器	33.3%	15	3	6	4	3	2	0	0	1	5	0	6	0	45
2　電磁気学　工学	48.8%	2	22	6	1	5	0	0	1	1	5	1	1	0	45
3　電子物性　磁性　光物性	66.6%	1	6	30	0	2	0	0	0	3	3	0	0	0	45
4　生体工学	68.8%	1	4	2	31	1	2	0	2	0	1	1	0	0	45
5　システム制御工学一般	80.0%	2	0	0	3	36	3	0	0	0	1	0	0	0	45
6　制御工学	46.6%	2	1	0	2	10	21	0	0	1	8	0	0	0	45
7　計算機方式　ハードウェア	0.0%	5	5	2	1	5	4	0	8	3	5	1	5	1	45
8　計算機利用技術	37.7%	1	1	0	1	9	1	0	17	4	7	1	3	0	45
9　電気工学一般	33.3%	2	3	3	1	7	0	0	0	15	11	0	3	0	45
10　電力工学	93.3%	0	0	0	0	2	0	0	0	1	42	0	0	0	45
11　電子工学	0.0%	2	5	22	0	6	1	0	0	3	4	0	2	0	45
12　通信工学	20.0%	4	8	1	2	13	1	0	4	0	3	0	9	0	45
計	44.2%	37	53	72	46	99	35	0	32	32	95	4	29	1	540

5．2　分類結果と再現性の評価

文献にわりつけられていた分類カテゴリーに属する文献が、実験の結果、どの分類カテゴリーにわりふられたかを示した結果を表2にあげる。分類カテゴリー番号 i の文献でカテゴリー j に assignされたものの文献数が i 行 j 列に示されている。　もし実験結果が100% の再現率を示せば、この表で　右下りの対角線上に45（1カテゴリーあたりの文献数）が並び、他の要素は0となる。　カテゴリー毎の再現率を比較すると、カテゴリー10が最も高く、ついで、5,4,3などが70% 前後の高い再現率を示している。　又、逆にカテゴリー番号 7,11 のものは再現率　0　%で、元来属しているカテゴリーに分類されたものは1件もなかった。

タイトル中の漢字の含有率と再現率の関連、カテゴリーによって再現率の違いが著しいが、カテゴリー別にタイトル中の漢字の量と分類の再現率との関係を調べるために、各カテゴリー毎に、タイトル中に漢字の含まれる割合と再現率を対比して散布図にしたものを図2に示す。

図2に示したプロットが全体的に右上りの直線傾向があることより、1レコード中に漢字が含まれる率が高いほど分類モデルの再現率が向上することがわかる。

カテゴリー7，11の文献については、1件も正しい分類カテゴリーが再現されなかった。

本稿で提案する分類モデルは日本語の文献の特徴についてある前提を立てている。すなわち、

1) キーとなる概念、又はその文献が属するカテゴリーを表現する語が漢字で表現されていることが多い。
2) その様な語は、各カテゴリー毎にユニークに使用されていることが多い。

今カテゴリー7に属する文献群について考えてみる。この表よりタイトル中に含まれる平均文字数も、その中に含まれる漢字の割合も、全カテゴリー平均に比べてかなり低い値を示していることがわかる。カテゴリー7に属する文献についてタイトル中に含まれている平均漢字数をみると、7以外のカテゴリーに比べて半分の値しか示していない。

分類モデルは、タイトル中の漢字にのみ着目し、その頻度特性に基づいて分類カテゴリーを算出する。すなわち、カテゴリー7に属する文献は、分類のキーとなる情報が、他と比べて半分の量しか与えられていないことになる。したがって、他のカテゴリーと同様に重要概念が漢字表現されていたとしても、その再現率がそれほど大きな値になると期待できない。

又、漢字を含む割合が低いということは、科学技術文献という題材から考えて、キーとなる語はカタカナで表現されていると推定される。実際サンプリングデータを調べてみると、この中にも極端に漢字が少ない例がかなり検出された。その一部を図3に示す。

図3は、タイトルが分類される過程を表現するためのトレース情報である。例えば、
『カラーファクシミリとカラー複写機』というタイトルの場合、『複写機』という漢字列だけで分類を行うことになる。『写』の出現確率の列のうち、分野8に対しての影響度が大きい。

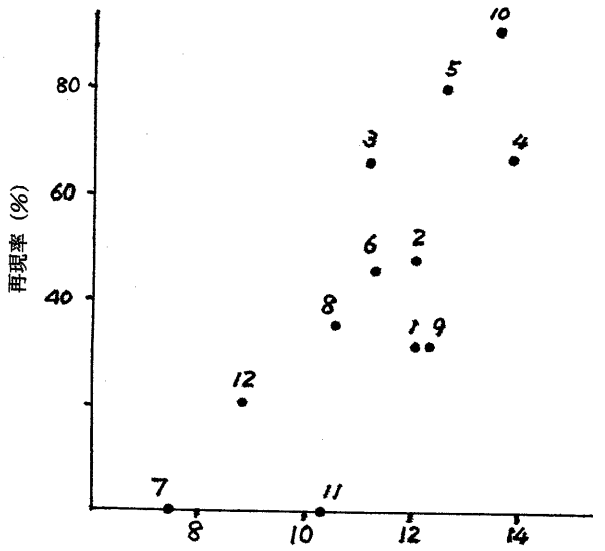

図2．　分類結果の再現率とタイトル中の漢字の割合

分類に用いられる文字は3文字しかないので、結局このカウントが大きく効いて、カテゴリー8にアサインしてしまう。又、より顕著な例として、『LEDアレーを用いた光学プリンター』などがある。このタイトルのキーワードは、『LEDアレー』、『光学プリンター』の2つであり、プリンターが主としてキーになるので、計算機関連のカテゴリーになる。しかしながら本来『プリンター』にかかる修飾語としての『光学』特に『光』の字が物理・光学のカテゴリーから検出される場合がきわめて多いので、逆に分類の際、このカテゴリーにアサインしてしまう。

E83111866JC040050U
カラーファクシミリとカラー複写機

複	0.04	0.11	0.04	0.03	0.10	0.04	0.07	0.08	0.16	0.11	0.05	0.12
写	0.01	0.08	0.00	0.15	0.15	0.00	0.06	0.37	0.00	0.00	0.06	0.08
機	0.07	0.02	0.01	0.08	0.02	0.19	0.14	0.05	0.08	0.16	0.03	0.09
	0.13	0.22	0.05	0.26	0.28	0.24	0.27	0.52	0.24	0.29	0.14	0.30

8(0.5203) 12(0.3025) 10(0.2908) UNKNOWN = 13 MOJI = 16

E83090866JC040050U
LEDアレーを用いた光学プリンタ

用	0.10	0.06	0.03	0.07	0.04	0.13	0.08	0.09	0.08	0.09	0.07	0.08
光	0.06	0.44	0.19	0.02	0.00	0.00	0.01	0.01	0.02	0.07	0.06	0.06
学	0.08	0.19	0.12	0.18	0.08	0.04	0.02	0.08	0.06	0.03	0.04	0.03
	0.25	0.70	0.35	0.29	0.13	0.18	0.13	0.18	0.17	0.20	0.19	0.19

2(0.7041) 3(0.3545) 4(0.2921) UNKNOWN = 13 MOJI = 16

図3．　漢字が少ないタイトルの分類例　（計算機方式，ハードウェア）

E83090957JC040010C
高性能集積回路の設計

高	0.10	0.12	0.09	0.03	0.01	0.05	0.09	0.04	0.10	0.10	0.13	0.06
性	0.06	0.10	0.13	0.09	0.11	0.04	0.05	0.02	0.13	0.07	0.08	0.06
能	0.07	0.04	0.02	0.10	0.07	0.10	0.17	0.07	0.06	0.07	0.07	0.10
集	0.10	0.07	0.02	0.03	0.10	0.12	0.07	0.06	0.13	0.06	0.13	0.04
積	0.09	0.17	0.07	0.08	0.05	0.01	0.06	0.03	0.08	0.03	0.20	0.08
回	0.08	0.06	0.03	0.02	0.03	0.05	0.05	0.04	0.23	0.08	0.17	0.10
路	0.06	0.10	0.03	0.01	0.01	0.05	0.05	0.04	0.25	0.07	0.20	0.08
設	0.04	0.02	0.00	0.02	0.12	0.13	0.10	0.07	0.09	0.18	0.09	0.08
計	0.24	0.05	0.01	0.05	0.09	0.09	0.10	0.07	0.05	0.10	0.05	0.05
	0.87	0.78	0.44	0.48	0.63	0.67	0.79	0.47	1.16	0.79	1.16	0.69

11(-1.1657) 9(1.1618) 1(0.8754) UNKNOWN = 1 MOJI = 10

図4．　漢字を多く含むタイトルの分類例　（計算機方式，ハードウェア）

E83090490NB03090G
太陽光発電用構造物の設計における風荷重の考察

太	0.03	0.00	0.27	0.00	0.00	0.01	0.00	0.00	0.00	0.34	0.31	0.01
陽	0.02	0.01	0.27	0.00	0.00	0.01	0.00	0.00	0.03	0.31	0.30	0.00
光	0.06	0.44	0.19	0.02	0.00	0.00	0.01	0.01	0.02	0.07	0.06	0.06
発	0.05	0.07	0.03	0.09	0.02	0.08	0.03	0.04	0.07	0.34	0.07	0.04
電	0.05	0.06	0.13	0.04	0.00	0.04	0.01	0.01	0.15	0.28	0.10	0.06
用	0.10	0.06	0.03	0.07	0.04	0.13	0.08	0.09	0.08	0.09	0.07	0.08
構	0.03	0.07	0.19	0.02	0.11	0.08	0.10	0.11	0.06	0.03	0.08	0.06
造	0.04	0.09	0.24	0.02	0.09	0.08	0.05	0.06	0.09	0.05	0.11	0.02
物	0.10	0.08	0.23	0.10	0.04	0.04	0.02	0.10	0.08	0.05	0.09	0.02
設	0.04	0.02	0.00	0.02	0.12	0.13	0.10	0.07	0.09	0.18	0.09	0.08
計	0.24	0.05	0.01	0.05	0.09	0.09	0.10	0.07	0.05	0.10	0.05	0.05
風	0.05	0.02	0.00	0.02	0.02	0.05	0.00	0.00	0.05	0.66	0.03	0.05
荷	0.06	0.06	0.17	0.05	0.01	0.12	0.02	0.01	0.07	0.24	0.10	0.02
重	0.10	0.15	0.05	0.04	0.05	0.05	0.13	0.05	0.06	0.05	0.07	0.16
考	0.06	0.02	0.00	0.06	0.21	0.04	0.09	0.10	0.07	0.14	0.04	0.10
察	0.11	0.02	0.08	0.08	0.11	0.04	0.09	0.09	0.09	0.08	0.04	0.09
	1.22	1.31	1.95	0.75	0.99	1.07	0.92	0.87	1.11	3.10	1.69	0.96

10(3.1017) 3(1.9537) 11(1.6999) UNKNOWN = 6 MOJI = 22

図5． 「電力工学」カテゴリーの分類例

さらに、このカテゴリーに属する文献のもう一つのパターンである全ての語が漢字で表現されているタイトルについて例をあげて考える。

『高性能集積回路の設計』

この場合のトレース情報を図4に示す。

図4で、『集』、『積』、『回』、『路』の4文字が、いずれかのカテゴリー寄与得点が 0.20 ~ 0.25と平均値の2.5~3.0倍の値をしめす他は、いずれも平均値0.08近傍をほぼ上下している。したがって、カテゴリーのアサインに寄与した字はこの4字であろうと推定できる。これらのいずれの字も電気工学一般、電子工学のカテゴリーから抽出された確率が高いので、その効果の影響で、これらのカテゴリーにアサインされたことを示している。

次にカテゴリー番号11に属する文献群の中から同様に例をとって考えてみる。このカタゴリーについてはキーとなる語を作る漢字が他のカテゴリーのキーとなる語も同時に作りやすい性格を持っていると考えられる。特に表2をみるとカテゴリー3と重複する場合が多いと考えられる。カタゴリー3が、「電子物性・磁性・光物性」であることを考えれば、この推測は正しい考えられる。又、同様の傾向がカテゴリー 12,6 にもみられる。

再現率がきわめて高いカテゴリーはカテゴリー番号10の『電力工学』であり、45 件中 42 件的中した。

このカテゴリーの一例を図5にあげる。

『太』、『陽』、『発』、『電』、『風』、『荷』の6文字がカテゴリー10から抽出された確率が高い字であり、このカテゴリーに対してのみ高い寄与率を有漢字であることがわかる。

この傾向は、全てのカテゴリー10に属する文献に現れる漢字の多くは、このカテゴリーに特に集中して出現するものであることを示している。逆にどこにでも現れる性質のものは、このカテゴリーには、それほど多く出現しないことになる。

6．結論

漢字の出現頻度だけを手掛りに文献の自動分類を試みた。その結果、以下の点が明かになった。

1) カタカナ、英字の情報を全て落とした漢字だけの頻度情報だけとする方式で英文の場合に単語をベースとした方式とほぼ同精度の再現率をえた。このことは漢字本来のもつ字としての意味がかなり強いカテゴリー表現力をもっていることを示している。
2) 分類実験の結果、再現性の高いカテゴリーと低いカテゴリーが明確になった。
3) 再現性の高いカテゴリーは、電子物性、生体工学、システム制御工学一般、電力工学であり、これらのカテゴリーに集中して出現する漢字は、そのカテゴリーに対する識別力があると推定できる。
4) 再現性の低いカテゴリーは、計算機方式、電子工学であり、これらのカテゴリーだけに集中して現れる漢字はきわめて少ない。又、このカテゴリーに属する文献の主要概念は漢字以外の文字（カタカナ、英字）等で表現されていると推定できる。

７．おわりに

漢字には自ら造語作用があり他の漢字と結び付いて新たな語を生成すると同時にその字本来の意味をも有している。筆者らは、その意味とカテゴリー毎の頻度との関係に着目し、大量のデータを統計的に処理した漢字の頻度情報だけを基にした時、分類のキーとしての漢字が、英文で単語をキーにした場合とよく類似していることを確認した。

８．参考文献

1) 国立国語研究所編
現代新聞の漢字，1976

2) 田中，漢字調査に於ける統計的尺度の問題，
電子計算機による国語研究,vol18,1975

3) 田島，漢字シソーラスの作成，漢字情報システムの問題点と対策，
第16回情報科学技術研究集会発表論文集,1977

4) 野村，漢字のパターン分類，
電子計算機による国語研究,vol10,1979

5) 梅田,細野,他，
漢字出現頻度に基づいた日本語文献の定量的分析,情報処理学会第28回全国大会発表論文集，3M-1,1983

6) 西村,岩井，
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,1968

7) 西村,岩井，
邦文文献の自動分類，
第7回情報科学技術研究集会発表論文集,1969

8) M.E.Maron,
Automatic Indexing : An Experimental Inquiry, J.A.C.M. Vol.8, No.3, 1961

9) K.A.Hanil,A.Zamora,
The Use of tittles for Automatic Document Classification, Journal of American Society for Information Science,Vol.31, No.6, 1980.

10) H.P.Luhn,
The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development,Vol.2,No.2,1958.