自然言語処理　58 － 5
（１９８６．１１．２２）

漢字クラスターによる日本語文献の重要語抽出

梅田茂樹　諸橋正幸　（日本アイ・ビー・エム株式会社　東京基礎研究所）
細野公男　原田隆史　（慶応大学文学部　図書館・情報学科）
後藤智範　（愛知叔徳大学文学部　図書館・情報学科）

漢字をキーにして、日本語の文章中の重要語を自動的に抽出する方法を提案し、それに基づく実験を試みた。キーにする漢字は、データベースの論文抄録中に出現する漢字の頻度を統計的に解析することにより、生成した。重要語の候補を考えるときの条件として、語の長さ、字種などテキストの表層的なものだけを用いた。抽出した重要語の妥当性の評価は、同一の抄録をその分野の研究者が抽出したものと、比較することにより行った。

その結果、多少ノイズとなるものがあるが、再現率指標ではほぼ妥当な結果が得られた。本稿で提案した方法は、分野のようなカテゴリーがあらかじめ与えられている時は、ある程度有効であると考えられる。

AUTOMATIC KEYWORD EXTRACTION OF JAPANESE TECHNICAL DOCUMENTS BY KANJI USAGE CLUSTERS

Shigeki UMEDA, Masayuki MOROHASHI [1]
Kimio HOSONO, Takashi HARADA [2]
Tomonori GOTOH [3]

1.Tokyo Research Lab. IBM Japan,Ltd.
5-19,Sanbancho,Chiyoda-ku,Tokyo,102
2.School of Library and Information science KEIO University
Mita,Minato-ku,Tokyo,102
3.Scholl of Library and Information science
AICHI SHUKUTOKU University
9,Katahira,Nagakutecho,Aichi-gun,AICHI,480-11

This paper describes a method that extracts the important terms in technical papers automatically by computer. Japanese documents include many kind of characters; Kanji, Hiragana, Katakana, and so on. Among them, Kanji is a kind of ideogram which has an own meaning. We assumed the function of Kanjis' role in Japanese text is almost correspond to words' in English text. By way of Kanji Usage Frequency among several technical fields in Japanese technical documents' database , we made several Kanjis' cluster which has strong connection to technical fields. After experiments of automatic keyword extraction by the clusters, we confirmed the accuracy and recall ratio by the questionnaires. The accuracy is almost same degree in comparison with the word based method , and recall ratio is 70-80 % , which is sufficiently practical.

1．はじめに

文献システムの構築では、現在のところ人手による分類・索引・抄録作成作業に多大な労力と経費をかけている。このため、文献がデータベースに登録されるまでに時間がかかるので、情報の鮮度が落ちるだけでなく、検索機能も低くなることが予想される。Luhnに始まる自動索引の研究はこういった背景から出てきたものである。Luhn[1]の考え方は、文中に出現する語の頻度を手掛りに重要語を抽出するものであるが、これと同様の方式をとるものとしてBaysianモデルによる自動分類を提案したMaron[2]，条件付確率の手法を導入したHammilとZamora[3]，因子分析の手法を導入したBorkoとBernick[4]等があげられる。一方、構文上各語がもつ役割、語と語との関連についての情報を用いるものもある。米国DDCの機械補助索引(MAI) はこの例である。[5][6]

一般に、文章の内容を表現するのに特定の単語をキーワードとして代表させることが多い。この場合、単語を意味の基本単位としてそれらを合成したものが文章全体の意味を代表するという考え方に立っている。Salton[7]らの単語を要素にした文献ベクトルモデルはこの例である。

欧印語と比べて日本語を考えたとき、その特徴として使用文字の多様性があげられる。なかでも、漢字は一字単位に固有の意味をもつ表意文字である。又、日本語の科学技術文献ではその文献の内容を表わす重要語の多くが、漢字もしくはカタカナ表現されていることも事実である。日本語文献中の漢字を、欧印語の単語と対応させて意味表現の基本単位と考えれば、欧印語と同様の手法が活用できる可能性がある。

筆者らは、この点に特に注目し、文献中に出現する漢字を手掛りにして、重要語を自動抽出する方法について検討する。

2．漢字をキーにした重要語の抽出

一般に欧印語は、単語の概念が明確で、意味の基本単位を単語におく考え方が馴染みやすい。一方、日本語の場合は、語と語の区切を意識せずにべた書き表記するのが一般的であり、しかも単語の概念が曖昧で、造語作用が強いために、いくつかの語が結合して容易に新たな語を形成する。

日本語では、重要な語は漢字書きされることが多い。そこで、語よりも細かいレベルではあるが、漢字を、文章の内容を表現する意味の基本単位と考えることもできる。筆者らは、このような観点から漢字と文献の分類カテゴリーとの関係について調査し、この両者の関係を、漢字の使用頻度を用いた指標により説明することを試みた。[8][9][10]

Luhn[1]は語の使用頻度と重要語の関係について、高頻度の単語は一般的な語で主題を識別する力はないとし、又、非常に低頻度のものも、その力は小さいと仮定した。

語の使用頻度を手掛りに重要語を選別する方法の他に、分野間での頻度分布に着目する方法もある。長尾[11]らは、いくつかの分類カテゴリーに属する文献群に出現する語のうち、いずれのカテゴリーにも均等に出現する語を、その文献を識別する力のない不要語とし、その他の語を重要語とした。その均等性を表わすパラメーターとしてχ^2値を用いている。χ^2値はその定義から、サンプリング数（この場合、単語の出現数）と分散値の独立性を仮定しているが、科学技術文献中の単語出現数にこの仮定があてはまる保証はない。又、このχ^2値は適合度検定の検定指標としての意味をもっておらず、これらの点で議論の余地がある。しかしながら、出現頻度傾向が分野間で著しく異なるものが、文献の属する分野の内容を表現するものであろうという考え方は、直感的にみても妥当なものであり、得られた重要語も納得できるものとなっている。

筆者らは、漢字をキーにした重要語抽出を考えているので、単語の場合以上に、絶対頻度と分野間の依存度が大きいと予想される。この点を考慮して、漢字を頻度順に並べ累積頻度を手掛りに、あらかじめいくつかの頻度階層を設定し、各頻度階層毎に分野間の出現傾向を調べることにした。

3．漢字のクラスター分析とキーになる漢字の設定

漢字の分類カテゴリー間での出現傾向をみるために、データベース中の文献抄録を用いて頻度調査を行った。データは JICST理工学文献ファイル電気工学編1983年度版である。同ファイルは、表１に示す12の分類カテゴリーに区分されている。全24巻のうち、第1巻から8巻までのテープから抄録をすべて抜出した。その中に含まれる漢字を、頻度順に並べ、累積頻度 50,75,95,99,99.7,99.9(%) を区切の目安として、1～7の 7つの頻度階層に分けた。(図１) このうち重要語抽出のキーにするという目的から、 頻度階層7は対象外とした。対象とする漢字は1260字であり、出現した漢字1630字のおよそ77%にあたる。

頻度階層ごとに、個々の漢字の分類カテゴリー内出現率と対応した多変量を考える。分類カテゴリーの数(12)の次元をもつ空間で、互いに距離の近いもの同志をひとつのク

表１．　電気工学を構成する１２分野

分類カテゴリー番号	主題内容
1	計測工学　計測機器
2	電磁気学　光学
3	電子物性　磁性　光物性
4	生体工学
5	システム制御工学一般
6	制御工学
7	計算機方式　ハードウェア
8	計算機利用技術
9	電気工学一般
１０	電力工学
１１	電子工学
１２	通信工学

図1．漢字の累積頻度と頻度階層

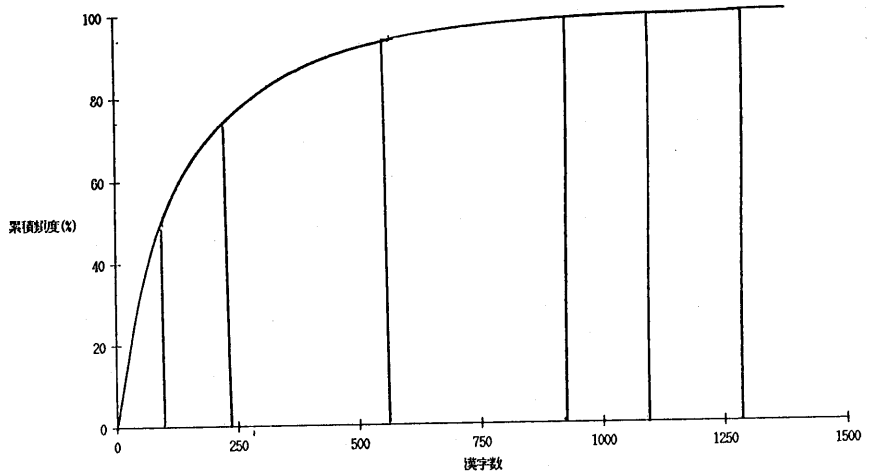


表2．　クラスター分類の結果

頻度階層	頻度範囲	クラスター数	各クラスター内の漢字数										
			1	2	3	4	5	6	7	8	9	10	計
1	0-50%	5	1	1	2	93	2						99
2	50-75%	5	119	5	4	3	3						134
3	75-95%	10	1	16	6	272	7	2	10	9	3	3	329
4	95-99%	8	290	11	4	1	3	5	13	2			329
5	99-99.7%	6	156	11	3	16	19	4					209
6	99.7-99.9%	7	136	1	4	1	7	23	7				179
計													1279

ラスターとする。この動作を順に繰返していくことで、順次クラスターを生成していく。クラスター数はクラスター内の距離の数とクラスターの最大直径より小さい値をもつ距離の数との比から決定した。各頻度階層において生成したクラスターに含まれる漢字数を表2に示す。表中のクラスター数は、頻度階層ごとにいくつのクラスターができたかを示している。又、各クラスターに属する漢字を表3に示す。表の一部は、クラスターに属する漢字が多数に及ぶため省略してある。

クラスター分析の結果、どの頻度階層でも、最も漢字数の多いクラスターは、12の分類カテゴリーに平均的に出現する性質をもつことがわかった。その例を図2に示す。それ以外のクラスターは特定のカテゴリー以外にはほとんど出現することがない性質をもつ。その例を図3に示す。

生成された各頻度階層のクラスターのうち、分類カテゴリー間で平均的に出現する傾向のあるクラスターの漢字は、文献を識別する力がないものと考える。これらの漢字を除いたものを、キーとなる漢字とする。分類カテゴリー間で平均的に出現する漢字は、高頻度のものほどキーにならない漢字と考えられるが、その閾値を判定するのが困難である。そこで、対象とする全漢字から頻度階層1から頻度階層6で平均的に出現するクラスターに属する漢字を順次累計して除いていき、各々それらの漢字群をキーとして重要語を抽出した。

表3．　各クラスター中の漢字

頻度階層	クラスター	漢字
1	1	電
	2	用
	3	性 定
	4	的 法 化 計 方 分. . .
	5	制 御
2	1	可 利 値 造 技 状. . .
	2	報 手 文 列 処
	3	界 膜 熱 温
	4	系 間 次
	5	音 画 像
3	1	金
	2	半 起 密 移 晶 酸 障 着 薄 依 属 壁 太 陽 池
	3	材 絶 縁 抵 抗 石
	4	誤 誘 運 案 無 布. . .
	5	網 話 符 局 星 衛
	6	元 堆
	7	領 境 曲 類 情 知 図 語 字
	8	保 付 規 紹 産 械 駆
	9	心 人 血
	10	声 声 職
4	1	合 寸 足 世 衡
	2	湿 校 粘 稲 架 土 網 希 永 久
	3	覆 捕 獲
	4	抽
	5	辺 読 令
	6	覚 弁 臓 肺 筋
	7	識 臨 皮 激 頭 左 床 胞 脳 症 刺 神 患
	8	骨 脈
5	1	敬 阻 迅 携 描. . .
	2	盛 鋳 暖 富 株 掘 噴 抗 旋 迅 擦
	3	韻 筆
	4	講 刻 頂 凹 族 随 服 該 矛 盾 括 裕 倒 律 又 枝
	5	剛 胸 健 薬 糖 死 尿 犬 腹 謝 肉 炎 呈 毛 乳 疾 週 虚 首
	6	耳 肝 児 拍
6	1	絞 匹 老 艦 康. . .
	2	殻
	3	菱 崎 泥 培
	4	絞
	5	芸 師 隠 絵 忠 辞 淡
	6	老 菌 免 疫 牛 糸 卵 迷 痛 女 胃 襲 腸 麻 酔 胎 胆 鼓 才 剖
	7	冠 髄 膝 襲 眉 鼻

図2．クラスターの出現傾向（1）

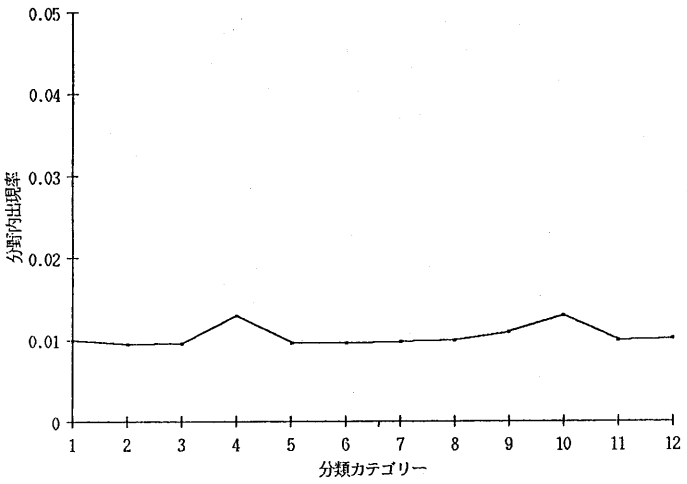
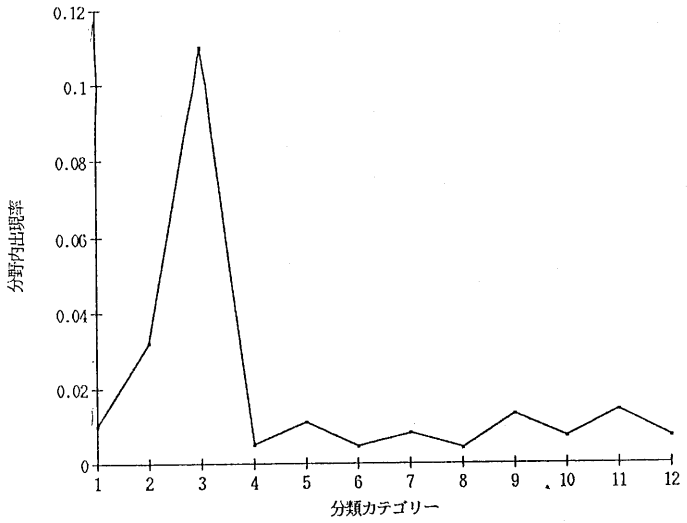


図3．クラスターの出現傾向（2）



4．重要語抽出

科学技術文献中の重要語は、文献の性質上、その多くは、いわゆる専門用語である。その品詞は、一般的に名詞でしかも複合語と考えるのが妥当である。そこで、テキスト中から名詞の並びを取り出し、その中で、次の3つのいずれかの条件をみたす語を重要語とする。

(1) キーとなる漢字を2つ以上含み、かつ文字数が3文字以上の名詞列

(2) キーとなる漢字を1つ含み、かつ文字数が4文字以上の名詞列

(3) キーとなる漢字を1つ以上含み、かつカタカナ列を含む名詞列

テキスト中から名詞の並びを取り出すにはテキストの文章が単語切りされ、品詞が識別されなければならない。このための前処理として、「漢字カナ変換プログラム」を使用して重要語抽出に用いる抄録文の単語切りを行った。このプログラムの精度は92~93%であるので、解析できないものについては人手で修正を加えた。

本実験で用いた JICSTデータベースは、シソーラス中のキーワード(重要語)を付与する統制語方式なので、データベース中にあるキーワードと、実験で抽出された重要語とを直接比較することはできない。そこで、電気工学分野の研究者に抄録中の重要語を抽出してもらって、比較の材料とした。抽出する語の長さ、語数などは制限なしに、任意の語を選択できる方式をとった。1つの抄録につき研究者5人を割当て、又1つの分類カテゴリーについては10抄録ずつとした。研究者5人が同一の抄録から重要語を識別することになるが、1つの語について3人以上が重要語候補と判定したものを重要語とする場合と、2人以上が重要語候補と判定したものを重要語とする場合との、2つのケースについて評価を行った。抽出された重要語の妥当性を評価するために、以下に示す指標P，Qを計算した。

指標P＝（C／A）＊ 100（%）
指標Q＝（C／B）＊ 100（%）
ここで
　漢字をキーに抽出した重要語の数　：　A
　研究者が抽出した重要語の数　：　B
　両者が共通して選択した重要語の数　：　C
とする。

Aが文中に出てくるあらゆる語を含むとすれば、Qは100%になるがPは極小になる。これは、文献を検索する立場からみれば、もれはなくなるがノイズが非常に高くなることを意味する。Cが小さくならないように慎重にAを選べば、Qはそれほど下降せずにPをある程度高い値に保つことができる。これらの指標は、各々情報検索でいう再現率と精度に対応している。

抽出結果は、文献の属する分類カテゴリーの影響をうけると考えられる。そこで、分類カテゴリー別に指標P，Qを計算した。

図4．頻度階層と指標P，Q
（3人以上）

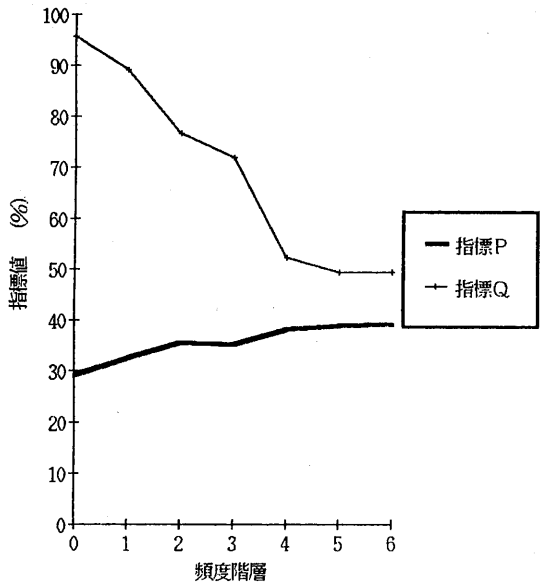
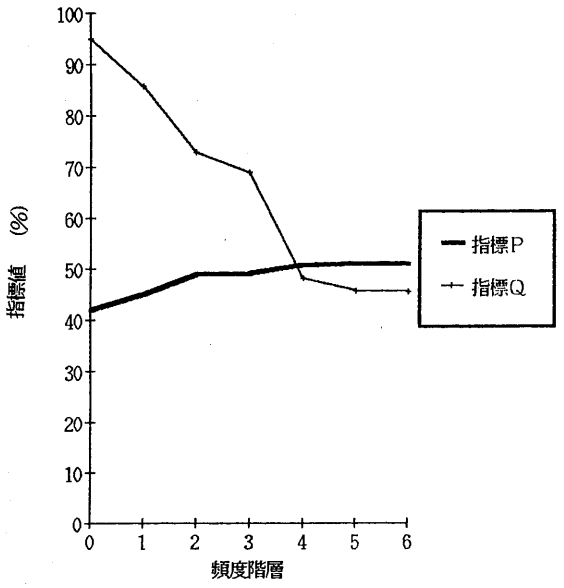


図5．頻度階層と指標P，Q
（2人以上）



5．重要語抽出結果及び考察

図4，5に、全漢字から、各頻度階層で平均的に出現する漢字を順次累計して除いたものを、キーとなる漢字にした時の評価指標の変化を示す。ここで横軸は、その頻度階層までの平均的に出現する漢字を除くことを意味する。図4，5からわかるように、指標Pは漢字の減少に伴いなだらかに上昇するが、頻度階層1，2で平均的に出現する漢字群を除いたときに、その増加が著しい。このことは、頻度階層3より低頻度で平均的に出現する漢字群は指標Pを上昇させる大きな要因とはならない。指標Qについては漢字の減少とともに、その値は減少する。指標Qの値がそれほど下がらない範囲で、指標Pの値をできるだけ高くするのが理想的である。ここでは、指標Pの変化に比べて、Qの方が大きいので指標Qの値にしたがってきめる。図4，5より、頻度階層1，2で平均的に出現する漢字群をのぞいた場合が、最適と考えられる。そこで、全漢字から頻度階層1，2で平均的に出現した漢字を除いたものを、キーとなる漢字群とする。

3人以上の研究者が選んだ語を重要語とする場合は、指標Pは30~40%であるが、2人以上の場合には、その候補の数が増加するので、指標Pは、 50%前後になる。指標Qについては、両者とも大きな相違はない。

重要語抽出を試みた抄録の例を図6，7に示す。不適切な重要語を抽出したものの特徴を整理して以下に示す。抽出の際のノイズとして、

1．位相制御、温度測定、使用実績、電気抵抗、時間間隔、測定原理、測定技術といった漢字の2文字熟語の組み合わせの形のもの。これらは、位相の制御、温度の測定、使用の実績など、助詞「の」を介して結ばれるべき2文字熟語が1語となっているためである。
2．耐熱性、適用性といった、「2文字熟語＋性」の形のもの。
3．電気抵抗温度計、誘電体、レーダ用材料、周波数範囲といった語。これらの語については、5人中2人が重要語として、選択しており、重要語かどうかの判定が困難である。

一方、抽出の際のもれとして、

1．制御、粒界のように、2文字熟語で重要語となっているもの。
2．結線方法、導波管、共振管、開放形共振器といった語。これらは漢字を多く含んでいるにもかかわらず、その漢字が全て、頻度階層1，2で平均的に出現する漢字である。
3．パルス反響、距離センサ、ディジタル信号発生機械といった、漢字とカタカナを含んでいるにもかかわらずその漢字が全て、頻度階層1，2で平均的に出現する漢字であるもの。

次に分類カテゴリーごとの評価指標を図8に示す。これより、本稿で提案した重要語抽出法が有効な分類カテゴリー

炉心内の温度及び中性子素の測定用検出器について調査した。温度測定に関しては、熱電対、熱雑音及び電気抵抗法について述べた。熱電対では、クロメル、アルメル、鉄、ニッケル、白金、ロジウム、タングステンなどの、中性子照射安定性と結線方法を示した。熱雑音を利用した方法については、原子炉での使用実績と特性を示した。電気抵抗温度計はアルミナなどセラミックの電気抵抗が温度に依存することを利用したもので、試作試験の例を示した。中性子検出器に関しては、耐熱性の分裂計数管とセルフパワード中性子検出器（ＳＰＮＤ）の研究開発の現状を報告した。

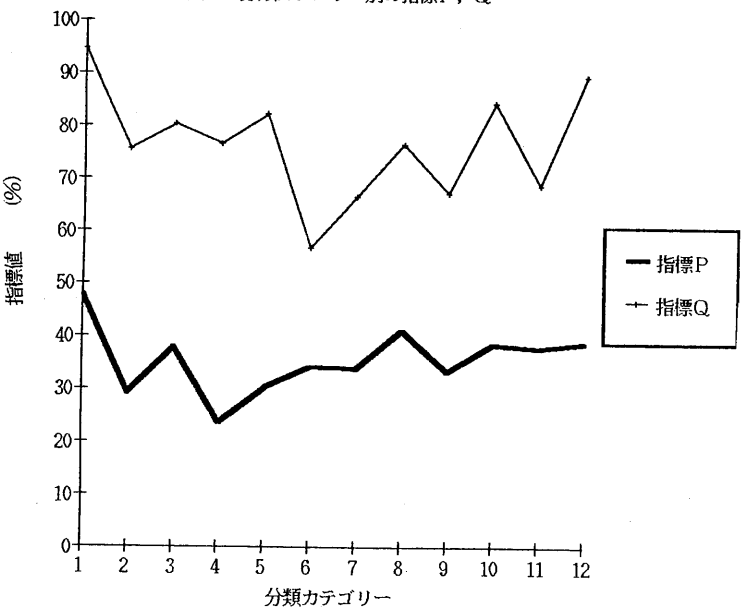
図６．重要語抽出を試みた抄録の例　（１）

波長可変色素レーザ技術は分光測定の分野で急速に発展したが、将来は色素レーザの線幅と同等の精度での波長の測定が必要になると予想される。ここではＮＢＳで進められている干渉技術を用いた色素レーザによる高精度、高分解能な波長測定技術の研究につき概説。標準波長光源を用いたＭｉｃｈｅｌｓｏｎ干渉計法による光路長を変化させない静的あるいは光路長を振動的に変化させる動的波長計、パルスあるいは連続発振レーザの両方に適用できるＦｉｚｅａｕ干渉計による干渉パターンをオンライン計算機で処理して波長を算出する波長計など、いくつかの波長計について性能、適用領域を示した。

図７．重要語抽出を試みた抄録の例　（２）

——— 研究者による抽出
～～～ 漢字による抽出

図8．分類カテゴリー別の指標P，Q



は、計測工学・計測機器、電力工学、通信工学であり、生体工学、制御工学については有効性がやや劣ることがわかる。

6．結論

1．重要語を抽出するキーとならない漢字は、累積頻度50%,75%を目安とする2つの頻度階層に存在し、各分類カテゴリーで平均的に出現する漢字である。

2．漢字をキーにして抽出した重要語は、電気工学分野の研究者による抽出結果と比較して、再現率指標で80%、精度指標で40% 前後であった。検索もれを、最小にしたいという一般的な情報検索の立場からすれば、この結果は充分といえないまでも妥当な値である。又、これらの値は、文献の分類カテゴリーにも依存し、その差はほぼ20% 程度であった。重要語の目安の設定の点で簡略化した点があるが、全体的には、ほぼ満足した結果が得られた。

3．重要語の候補を考えるうえで、キーとなる漢字、および語の長さ、字種など、テキストの表層的な条件だけでしぼり込む簡略的な方法をとったが、全体の評価としては、ほぼ満足した結果が得られた。

謝辞

実験に協力してくださった慶応大学理工学情報センターの落合啓一氏、及び慶応大学、東北大学の研究室の方々に感謝いたします。

参考文献

1) Luhn,H.P.
The Automatic Creation of Literature Abstracts.
IBM Journal of Research and Development.
Vol.2,No.2,p.159-168(1958)

2) Maron,M.E.
Automatic Indexing ; An Experimental Inquiry.
Journal of ACM. Vol.8,No.3,p.404-417(1961)

3) Hamill,K.A. and Zamora,A.
The Use of Titles for Automatic Document Classification. Journal of ASIS.
Vol.31,No.6,p.396-402(1980)

4) Borko,H. and Bernick, M.
Automatic Document Classification.
Journal of ACM.
Vol.10,No.2,p.151-162(1963)

5) 中井浩.
機械補助索引(MAI)について[I].
情報管理. Vol.19,No.4,p.247-259(1976)

6) 中井浩.
機械補助索引(MAI)について[II].
情報管理. Vol.19,No.5,p.350-357(1976)

7) Salton,G., Yang,C.S. and Yu,C.T.
A Theory of Term Importance in Automatic Text Analsys. Journal of ASIS.
Vol.26,No.1,p.33-44(1975)

8) 後藤智範，細野公男 他.
出現頻度に基づく重要漢字と主題分野の関連.
第21回情報科学技術研究集会発表論文集.
p.209-215(1984)

9) 梅田茂樹，守屋智，細野公男 他.
漢字の出現頻度情報を用いた日本語文献の自動分類.
情報処理学会自然言語処理研究会報告. Vol.47,No.7,p.47-54(1985)

10) 細野公男，梅田茂樹，原田隆史 他.
漢字の出現頻度特性を用いた日本語文献の機械処理.
情報管理. Vol.29,No.5,p.410-420(1986)

11) 長尾真,池田浩之,水谷幹男.
日本語文献における重要語の自動抽出.
情報処理. Vol.17,No.2,p.110-117(1976)