情 報 学 基 礎 25－2
（1992. 5. 12）

# 日本語文書用高速全文検索の一手法

菊池 忠一
(株) テレマティーク国際研究所

本文では、日本語文書における高速全文検索手法と広辞苑を用いた検索実験結果を報告する。一般的に、欧文文字と比較すると、日本語文書における文字や文字列の出現頻度は低い。本手法は、日本語文書のこれらの特徴を用いたもので、以下の３要素で構成される。

(1) 文書から抽出した文字あるいは文字列から作成する照合単位の文字位置情報別のグループ化
(2) 照合単位内における文字位置情報の昇順配列
(3) 低出現照合単位からの照合単位間における文字位置情報の照合

# A Fast Full-Text Search Method for Japanese Text Database

Chuichi KIKUCHI
Telematique International Research Lab.
Column Minami Aoyama Bldg. 7th Floor, 7-1-5 Minami-Aoyama Minato-ku, Tokyo 107 Japan

This paper describes a fast full-text search method and the result of retrieval experiments using the Japanese-language dictionary KOJIEN. Generally, compared of European languages, there are fewer occurences of both single characters and character strings in Japanese text databases. This method uses these characteristics of Japanese text databases and is comprised of the following three parts.

(1) grouping character-position datas according to collation units formed from one character or a character string taken from the text
(2) arraying character-position data in each collation unit in ascending order
(3) collation, in ascending order, of character-position data from two collation units

## 1．まえがき

情報化社会の進展に伴い、電子メディアの普及が著しく、作成された大量の文書情報を電子的にファイリングし、必要に応じて容易かつ高速に検索できる文書検索技術の確立が求められている。

これまでの文書検索技術では、インデックス方式と全文検索方式が主流である。インデックス方式は、登録に先立って全文から検索単位に分割した文書に付加したキーワードを用いて、該当する文書を探索する方式である。キーワードをインデックステーブルに登録することにより高速に検索できるが、キーワードの抽出に専門的な知識や多くの労力と時間を要する上、インデクサによって付加するキーワードが異なる問題がある。これらの問題を解決する技術として、全文検索方式の進歩が期待されてきた。全文検索方式は、検索対象である文書に出現するすべての文字を、最初から最後まで検索指定文字列と照合し、検索者が指定する文字列と同じ文字列を有する文書を選出するものである。大規模な文書の全文検索では、検索時間が主要な課題になるため、検索処理の高速化が試みられ、１個の文字列を対象とするＫＭＰ法[1]やＢＭ法[2]，複数の文字列を１度の走査で探索するＡＣ法[3]，ＢＭ法を拡張して複数の文字列検索をできるようにしたＥＢＭ法[4]などの高速パターン照合アルゴリズムが考案されている。

しかし、全文検索方式では、多くの場合、ソフトウェアで文字列照合を行うため、小規模な文書の検索には便利であるが、大規模な文書には不向きであった。これを解決するためハードウェアによる高速化が試みられ、連想メモリ法[5]，セルラーアレイ法[6]，ＦＳＡ法[7]やＤＰ法[8]などが提案された。国内では、データベースプロセッサ[9]，専用プロセサ[10][11]，文字列検索用ＶＬＳＩ[12]を用いたもの，および階層型プリサーチ方式を採用した専用マシーン[13]などが開発されているしかし、専用プロセッサや文字列検索用ＶＬＳＩを用いた方式では、これらのホストから検索対象である文書を転送するのに時間がかかり、これらの高速性が十分に活かしきれていない[11][14]。また、データベースプロセッサや専用マシーンでは、ハードウェアが高価になり、使用する機器が限定されることが多い。このため、大規模な文書でも高速検索ができ、廉価で計算機種に依存しない全文検索方式が求められていた。

筆者は、これらの問題を解決するために、日本語文書における同一文字および同一文字列の出現頻度が低い特徴に着目し、任意文字列を入力とする検索の高速化に取り組み、小規模から大規模な文書まで適用できる、ソフトウェアによる廉価な高速全文検索を実現した。先に、本テーマに関連して、図書名を対象とする部分一致検索の高速化手法[15]，漢字列を検索キーとする全文検索の高速化手法[16]および仮名文字列を検索キーとする全文検索の高速化手法[17]を報告した。

以下、本文では、これまでの報告をまとめ、日本語文書の例として広辞苑を対象に、見出し語の説明文に出現する文字や文字列の出現度数の調査結果および全文検索の高速化手法と実験結果を示す。

## 2．日本語文書における文字および文字列の出現頻度の特徴

文書中の文字位置を文字種ごとにグループ化しておき検索文字列と同じ文字種グループ間で、文字位置が連続する組合せを抽出すると、文書中の検索文字列を探索できる。この探索方式は、文書の出現文字種が多いほど、文字種グループ内の文字位置情報が少なくなるので高速になる。また、同一文字列の出現頻度が少ないほど、初めの文字位置連続確認で検索結果を絞り込めるので高速になる。

そこで、日本語文書における文字や文字列の出現頻度を知るために、現代語から古語までを含む広辞苑記載の見出し語の説明文書を対象に、同一文字や同一文字列の出現度数を調査した。広辞苑には約19.8万の見出し語があり、これらの説明文書に出現するＪＩＳ第１水準およびＪＩＳ第２水準漢字表（以降、ＪＩＳ漢字表と略す）に記載された文字の合計は約895 万文字であった。

### 2．1　同一文字の出現度数

一般的に、文書中の特定文字列の出現頻度は、特定文字列をｎ文字（ｎ≧２）で構成する文字セットとすると、ｎが大きくなるほど冗長性が出て出現頻度が低下する。すなわち、ｎを大きくするほど、特定文字列を高速に探索することができる。

いっぽう、ＪＩＳ表記載の文字種は6,879 であるから、特定文字列を最小文字セットの２文字としても、文字セット種は$6,879^2 = 47,320,641$になり、計算機での処理が容易ではない。

そこで、一般的な日本語文書の半数程度を占める仮名文字を２文字の仮名セットとすると、「つ」や「っ」などの仮名の大文字と小文字および「あ」や「ア」などの平仮名と片仮名を同一文字とみなし、更に、特殊記号の「ー」を加えても、仮名セット種は$75^2 = 5625$と少なく、他の漢字などを１文字で扱うとしても、仮名セット種とその他の文字種の合計は12,334にすぎなく、計算機での処理が容易な範囲である。

したがって、本文では、漢字は１文字、仮名文字は２文字を単位として、出現度数を調査した。平均出現度数を表１に示す。

表１　出現度数表

| | 平均出現度数 | 最多出現度数 |
|---|---|---|
| 第１水準漢字 | １１５５ | ６３１８３ |
| 第２水準漢字 | １７ | ７１２ |
| 仮名連続２文字 | ４６５ | ６３４２０ |
| 仮名文字 | ５３２２３ | ３８２００８ |

平均出現度数は、第１水準漢字が1,155 ，第２水準漢字が17，仮名２文字セットが465 であった。なお、仮名１文字の平均出現度数は52,525であった。

### ２．２　同一文字列の出現度数

広辞苑記載の見出し語の説明文書から、ＪＩＳ漢字表記載の３～８文字の漢字列，３～４仮名セット（すなわち５～８文字）の片仮名および平仮名の文字列を各々1000個抽出し、文字列の先頭文字から順に１文字ずつあるいは１仮名セットずつ取り出して作成した文字列の出現度数を調査した。調査結果を表２～表４に示す。

表２　漢字列における平均出現度数

| 照合入力 | n文字の平均出現度数 | | | |
|---|---|---|---|---|
| 文字数 | n＝1 | n＝2 | n＝3 | n＝4 |
| 2 | | | ---- | ---- |
| 3 | 7150 | 326 | 77 | ---- |
| 4 | 7631 | 603 | 122 | 80 |
| 5 | 12122 | 637 | 147 | 55 |
| 6 | 8106 | 519 | 144 | 48 |
| 7 | 14996 | 766 | 75 | 16 |
| 8 | 11267 | 657 | 70 | 34 |

表３　片仮名文字列における平均出現度数

| 照合入力 | n仮名セットの平均出現度数 | | | |
|---|---|---|---|---|
| 文字数 | n＝1 | n＝2 | n＝3 | n＝4 |
| 2 | 1948 | ---- | ---- | ---- |
| 3 | 1230 | 174 | ---- | ---- |
| 4 | 1819 | 232 | ---- | ---- |
| 5 | 1918 | 56 | 53 | ---- |
| 6 | 1781 | 118 | 114 | ---- |
| 7 | 1889 | 32 | 13 | 13 |
| 8 | 1650 | 36 | 4 | 3 |

表４　平仮名文字列における平均出現度数

| 照合入力 | n仮名セットの平均出現度数 | | | |
|---|---|---|---|---|
| 文字数 | n＝1 | n＝2 | n＝3 | n＝4 |
| 2 | 12869 | ---- | ---- | ---- |
| 3 | 8028 | 2386 | ---- | ---- |
| 4 | 5209 | 554 | ---- | ---- |
| 5 | 4053 | 111 | 57 | ---- |
| 6 | 3894 | 87 | 14 | ---- |
| 7 | 4562 | 123 | 7 | 5 |
| 8 | 4069 | 141 | 12 | 7 |

調査結果から、先頭から２文字目あるいは２仮名セット目で、先頭の漢字あるいは仮名セットと同じ漢字あるいは仮名セットを有する文字列の大部分が対象外となり、絞り込まれることがわかった。例えば、表２の照合入力文字数＝８の場合、２文字目で約88.9%に、３文字目で約98.0%に絞り込まれた。同様に、表３の照合入力文字数＝８の場合、２仮名セット目で約97.9%に、３仮名セット目で約99.7%に絞り込まれた。また表４の照合入力文字数＝８の場合、２仮名セット目で約96.5%に、３仮名セット目で約99.7%に絞り込まれた。（絞り込み率：式１参照）

絞り込み率＝ａ÷（ｂ－ｃ）×１００［%］　（１）

ａ：照合各段階の対象外漢字（仮名セット）の累計

ｂ：先頭の漢字（仮名セット）の合致数

ｃ：末尾の漢字（仮名セット）の合致数

これらの結果から、広辞苑記載の見出し語の説明文書に出現する任意の文字列の先頭文字から順に説明文書を文字位置照合するとき、先頭から２文字目の漢字あるいは２仮名セット目の仮名文字で、検索対象がほとんど絞り込まれることがわかった。

## ３．全文検索の高速化手法

本手法の３要素である文字位置照合単位種別のグループ化，文字位置番号の昇順配列および低出現照合単位からの文字位置照合について述べる。

### ３．１　文字位置照合単位種別のグループ化

任意の検索入力と同じ文字列の出現場所を検索対象から探索する場合、検索対象の先頭から逐次に文字列照合するより、検索入力の構成文字と同じ文字だけを検索対象から取り出し、文字位置をキーとして文字位置照合するほうが照合回数が少ない。

広辞苑見出し語の説明文書を対象とする場合、２．１で、ＪＩＳ第１水準漢字の平均出現度数が1155回であることから、検索入力がｎ文字の第１水準漢字列の場合、広辞苑見出し語の説明文書に出現する検索入力と同じ第１水準漢字は、平均1155×ｎ個になる。すなわち、平均1155×ｎ個の漢字の中から文字位置をキーとして、検索入力と同じ漢字列を構成できる漢字列を選出するだけで全文検索を行うことができる。例えば、広辞苑見出し語の説明文書で最も出現度数の高い漢字「一」と、次に出現度数の高い漢字「人」で構成される漢字列「一人」で検索する場合、この二文字の出現度数の合計が99,795であったから、99,795文字の中から「一人」と同じ文字列をさがすことになる。同様に、仮名文字の場合、最も出現度数の高い仮名セット「こと」と、次に出現度数の高い仮名セット「など」で構成される仮名文字列「ことなど」で検索する場合でも、この二個の仮名セットの出現度数の合計が97,286であったから、97,286文字の中から「ことなど」と同じ文字列をさがすことになる。一般的に、検索入力は数10文字以下であるから、検索時の照合対象文字数は全文構成文字数に比してはるかに少なくなる。

以上から、全文中の構成文字を文字種あるいは仮名セット種の文字位置照合単位種別にグループ化して登録しておき、検索時に、これらのグループの中から検索入力を構成する漢字あるいは仮名セットと同じグループだけを取り出し、文字位置をキーとして、検索入力と同じ文字

列を構成できる文字列を選出することにより、全文検索の高速化が図れる。

### 3．2　文字位置番号の昇順配列

数値を昇順に配列した2グループ間で、任意の差を有する数値の組合せを選出するとき、該当しない2数値間では小さい数値が、該当する2数値間では2個の数値が照合対象外になるので、2グループのすべての数値との照合が不要になる。例えば、昇順に数値を配列した2個のグループΣAおよびΣBを

$\Sigma A = A_1, A_2, A_3, A_4, \cdots, A_m$

$\Sigma B = B_1, B_2, B_3, B_4, \cdots, B_n$

およびrを自然数とすると、これらのグループから、$A_i$ と $A_i$ よりr大きい $B_j$ の組み合わせを選出するとき、以下の条件を満足する数値の項が照合対象外になる。

$A_i + r > B_j$ なら $B_j$ が対象外

$A_i + r < B_j$ なら $A_i$ が対象外

$A_i + r = B_j$ なら合致で $A_i$ と $B_j$ が対象外

従って $A_m > B_n$ の場合、照合回数Nは式2で表せる。

$$N = k + n - 1 - q \qquad (2)$$

kは、$A_k + r > B_n$ になるグループΣA内の数値 $A_k$ の配列順位、qはΣA～ΣB間の合致数である。例えば、

ΣAが（5, 9, 13, 15, 17, 19, 22, 25, 30, 34 ）

ΣBが（3, 8, 11, 12, 18, 20, 24）

r＝2の場合、

$B_n = B_7 = 24 < 25 + 2 = A_8 + 2 = A_k + 2$ から $k = 8$, $n = 7$ になる。また、$A_i + 2 = B_j$ を満足するのは、$A_2 = 9$ と $B_3 = 11$, $A_7 = 22$ と $B_7 = 24$ であるから、合致数はq＝2、照合回数はN＝12になる。一方、ΣAとΣBのすべての項を総当たりに照合すると、照合回数は10×7＝70回になる。このように、照合するグループ内の数値を昇順に配列すると、2個の数値を照合するごとに、1個あるいは2個の数値が照合対象外になるので、照合回数を削減できる。

以上から、3．1で文字位置をキーとして検索入力と同じ文字列を構成する文字列を選出できるようにするため、全文を構成する漢字および仮名セットを含むすべての文字位置照合単位に、全文における配列順位を表す文字位置番号を付与し、各照合単位種グループごとに文字位置番号を昇順に配列すると、文字の連続確認時の照合回数が低減し、検索を高速化できる。

### 3．3　低出現照合単位からの文字位置照合

表2～表4において、n＝2における平均出現度数は文字列の先頭および2番目の各々の照合単位と同じ照合単位種グループにある文字位置番号を照合した結果である。また、n＝3における平均出現度数は、n＝2における照合結果と、文字列の先頭から3番目の照合単位と同じ照合単位種グループの文字位置番号を照合した結果であるから、n＝2における照合回数が最も多く、次にn＝2の照合結果を用いるn＝3における照合回数が多いことがわかる。従って、文字列のすべての照合回数を減らすには、文字列を構成する照合単位の中で、低出現度数の照合単位からの照合が有効であることがわかる。そこで、3～8文字の漢字列と3～4仮名セットの片仮名および平仮名文字列（すなわち5～8文字列）を入力として、照合単位の先頭から順に文字位置照合を行う場合と、低出現照合単位から順に文字位置照合を行う場合について、照合回数を調査した。調査結果を表5～表7に示す。入力文字列は、2．2で使用したものと同じである。低出現照合単位から順に照合する場合と、先頭にある照合単位から順に照合する場合の照合回数を比較すると、低出現照合単位順の照合回数は、先頭にある照合単位から順の場合の、漢字が入力の

表5　漢字列照合時の照合回数

| 照合入力文字数 | n文字照合時の照合回数 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 先頭の漢字からの照合 | | | | | 低出現の漢字からの照合 | | | | |
| | n＝1 | n＝2 | n＝3 | n＝4 | 合　計 | n＝1 | n＝2 | n＝3 | n＝4 | 合　計 |
| 3 | 0 | 19017 | 2214 | ---- | 21231 | 0 | 8604 | 2120 | ---- | 18092 |
| 4 | 0 | 19844 | 3241 | 1220 | 24305 | 0 | 6950 | 2164 | 1471 | 10585 |
| 5 | 0 | 29430 | 3735 | 1301 | 35282 | 0 | 7731 | 1462 | 1077 | 11615 |
| 6 | 0 | 20952 | 2798 | 1116 | 26060 | 0 | 4671 | 877 | 589 | 7574 |
| 7 | 0 | 34850 | 4389 | 1022 | 41248 | 0 | 4789 | 797 | 442 | 7368 |
| 8 | 0 | 27357 | 3379 | 887 | 32930 | 0 | 3494 | 501 | 268 | 5746 |

表6　片仮名文字列照合時の照合回数

| 照合入力文字数 | n仮名セット照合時の照合回数 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 先頭の仮名セットからの照合 | | | | | 低出現の仮名セットからの照合 | | | | |
| | n＝1 | n＝2 | n＝3 | n＝4 | 合　計 | n＝1 | n＝2 | n＝3 | n＝4 | 合　計 |
| 5 | 0 | 4498 | 332 | ---- | 4830 | 0 | 1499 | 488 | ---- | 1987 |
| 6 | 0 | 4481 | 784 | ---- | 5265 | 0 | 1844 | 856 | ---- | 2700 |
| 7 | 0 | 4540 | 189 | 168 | 4897 | 0 | 1167 | 183 | 245 | 1595 |
| 8 | 0 | 4037 | 189 | 133 | 4359 | 0 | 971 | 101 | 226 | 1298 |

表7　平仮名文字列照合時の照合回数

| 照合入力文字数 | n仮名セット照合時の照合回数 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 先頭の仮名セットからの照合 | | | | | 低出現の仮名セットからの照合 | | | | |
| | n＝1 | n＝2 | n＝3 | n＝4 | 合　計 | n＝1 | n＝2 | n＝3 | n＝4 | 合　計 |
| 5 | 0 | 11304 | 1281 | ---- | 12585 | 0 | 5291 | 1814 | ---- | 7105 |
| 6 | 0 | 10379 | 1246 | ---- | 11625 | 0 | 4947 | 1401 | ---- | 6348 |
| 7 | 0 | 11977 | 911 | 499 | 12888 | 0 | 3704 | 689 | 678 | 5071 |
| 8 | 0 | 10643 | 947 | 596 | 12186 | 0 | 3129 | 613 | 711 | 4453 |

場合で約17%～51%、片仮名が入力の場合で約30%～51%、平仮名が入力の場合で約36%～56%であった。

以上から、「計算機」あるいは「できること」のような３個以上の照合単位で構成される文字列が検索入力の場合、出現度数の低い照合単位から順に文字位置照合を行うと、照合回数を低減でき、検索の高速化が図れる。

## ４．検索方式

ここでは、文字位置照合単位種別のグループ化，照合単位種グループ内の文字位置情報の昇順配列，低出現照合単位からの文字位置照合を取り入れた全文検索方式を、一般的な書籍を対象として述べる。

### ４．１　文字位置情報の作成

文字位置情報は、照合単位である文字セットあるいは非仮名文字が出現する場所を表すもので、以下に示す属性番号，検索単位番号および文字位置番号を用いて作成できる。

本文では、２．１に示した背景から、仮名の連続２文字を文字セットとした。更に、非仮名文字と仮名文字の接続を考慮し、非仮名文字と仮名文字の連続２文字も文字セットとした。しかし、これらの文字セットの構成は限定されたものではなく、多数の文字種を高速に効率よく処理できる場合には、全文字種で文字セットを構成できる。また、英語，数字あるいは仮名文字のように文字種が少ない場合には、３文字以上の文字セットも可能である。

(1)　文字セット

文字セットには、（仮名＋仮名）で構成する仮名セット，（非仮名＋仮名）で構成する混合セットＡ，（仮名＋非仮名）で構成する混合セットＢがある。文書を任意の文字列で検索できるように、文書の先頭から１文字ずつ取り出し、その次に続く文字との合計２文字から、下記の①～④の規則で文字セットと非仮名文字を作成する。

①　仮名＋仮名　　⇒仮名セットの作成
②　非仮名＋仮名　⇒非仮名＋混合セットＡの作成
③　仮名＋非仮名　⇒混合セットＢの作成
④　非仮名＋非仮名⇒非仮名のまま
　（最初の非仮名を使用、２個目の非仮名は除外）

例えば、文字列「作成された大量の文字情報」は、「作，成，成さ，され，れた，た大，大，量，量の，の文，文，字，情，報」に分解でき、このうち仮名セットは「され，れた」、混合セットＡは「成さ，量の」、混合セットＢは「た大，の文」、非仮名は「作，成，大，量，文，字，情，報」になる。

(2)　属性番号

一般的な書籍は、目次，序文，章／節タイトル，本文，図／表タイトル，参考文献などで構成される。書籍の記載事項を全文検索するとき、検索結果をこれらの構成部分の名称で得られると理解しやすく、便利である。また、これらの構成部分の名称も検索入力として使えると、利用しやすい検索になる。ここでは、これらの構成部分を各々検索単位とし、各検索単位に属性番号を付与した。属性番号は、目次＝１，序文＝２章／節等のタイトル＝３，図／表のタイトル＝４，本文＝５，参考文献＝６とした。

(3)　検索単位番号

属性とは無関係に、すべての検索単位に、１，２，３・・・と出現順に番号を付与し、検索単位番号とした。

(4)　文字位置番号

検索単位を構成する文字列から文字セットあるいは非仮名文字である照合単位を作成し、文字列の先頭からの配列順位を文字位置番号とした。なお、文字セットの文字位置番号は、文字列における文字セットの先頭文字の配列順位とした。

(5)　文字位置情報

式３を用いて、検索単位から作成した照合単位を自然数に変換し、文字位置情報とした。

文字位置情報
＝｛（検索単位番号）×ｎ＋
　（文字位置番号）｝×ａ＋（属性番号）　（３）
ｎ：最大検索単位文字長
ａ：最大属性種数

例えば、ｎ＝10000，ａ＝10の場合、８番目の検索単位である本文（属性番号＝５）の先頭から121 ～128 文字目に出現する「検索ファイルは、」は、図１のような文字位置情報に変換できる。このとき、文字位置情報を４バイトで表すと、最大10000 文字長の検索単位を、
$2^{32} \div (n \times a)$ ＝約４万個取り扱うことができる。

### ４．２　検索ファイルの作成

検索ファイルは、図２に示すように、ＪＩＳ漢字表の非仮名文字種グループ領域と文字セット種グループ領域で構成され、各領域にはそれぞれの文字種あるいは文字セット種に対応する照合単位種グループが配列される。照合単位種グループには、各検索単位から作成されるすべての照合単位に付与される文字位置情報が、登録順に先詰め形式で格納される。従って、各照合単位種グループには、文字位置情報が昇順に格納される。図１の文字列「検索ファイルは、」を登録した例を図３に示す。文字位置情報は４バイトであるから、ファイル容量は、４バイト×（照合単位数）になる。

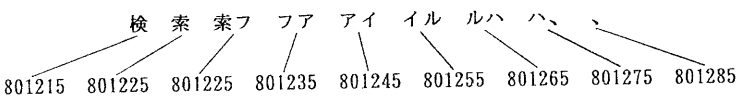

図１　文字位置情報への変換例

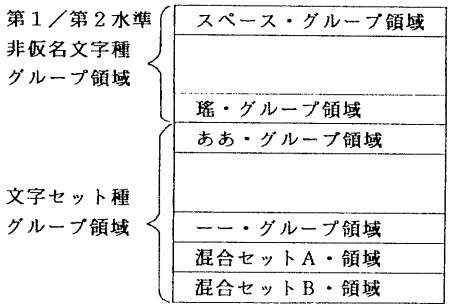


図２　検索ファイルの構造

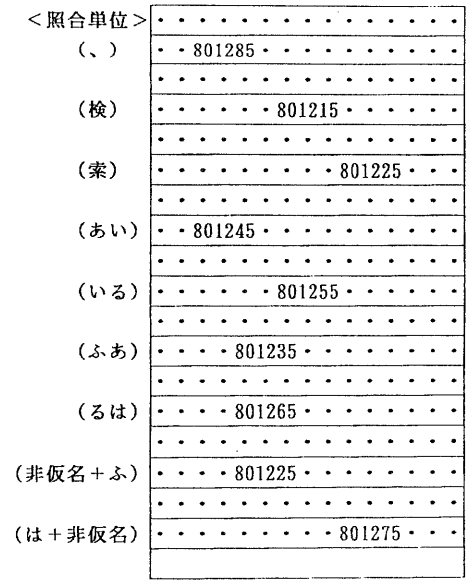


図３　検索ファイルの例

### ４．３　検索方法

検索処理は、以下に示す(1)～(4)の処理で構成され、検索入力文字列から作成する照合単位に対応する照合単位種グループに格納された文字位置情報の連続性を照合し、検索入力文字列と同じ文字列を作成できる文字位置情報の組合せを抽出する。

(1)　検索入力文字列の並べ変え

検索入力文字列の先頭から文字セットあるいは非仮名文字の照合単位を作成し、出現度数の低い照合単位から順に並べ変える。

(2)　文字位置照合

並べ変えた照合単位列の先頭から順に、照合単位に対応する照合単位種グループを取り出し、式４に示すように、検索単位と属性が等しく、文字位置番号の差が検索入力文字列における文字位置差に等しい文字位置情報の組み合わせを抽出する。

$$\Sigma G_i - \Sigma G_j = (i - j) \times a \qquad (4)$$

ここで、$\Sigma G_i$ と $\Sigma G_j$ は各々、検索入力の先頭からｉ番目およびｊ番目の文字を先頭文字とする照合単位に対応する照合単位種グループ内の文字位置情報、ａは最大属性種数である。

(3)　属性照合

抽出した文字位置情報の組み合わせの中から、検索入力と同じ属性を有する文字位置情報を取り出す。

(4)　検索結果の抽出

取り出した文字位置情報から、検索単位番号と文字位置番号を取り出し、検索結果とする。

次に、(1)～(4)の具体例を示す。例えば、本文から「検索ファイル」を検索する場合、照合単位は「検．索．ファ．イル」になる。これらの照合単位の出現度数を、検＝1,008 索＝200 ．ふあ＝395 ，いる＝13,873とすると、並べ変えた照合単位列は「索ファ検イル」になる。文字位置照合では、図３の検索ファイルから、これら４個の照合単位に対応する照合単位種グループを取り出し、式４を使用した文字位置照合を行い文字位置情報を抽出する。属性照合では、検索条件が本文であるから、属性番号を５として照合すると、この中に「索」の文字位置情報801225も含まれる。「検索ファイル」では、「索」が検索入力文字列の２文字目で、文字長が６文字であることから、検索入力文字列と同じ文字列が、検索単位番号８の本文の、先頭文字から121 ～126 の文字位置にあることがわかる。

## ５　検索実験

本手法の性能を評価するために、広辞苑を用いて実験を行った。

### ５．１　実験条件

ＰＦＵ社製Σ２３０モデル２２上に、広辞苑記載の見出し語の説明文書を登録した実験システムを構築した。広辞苑には約19.8万の見出し語があり、説明文書のＪＩＳ漢字表記載文字を合計すると約895 万文字であった。説明文書の最大文字長は3108、本文だけを対象としたので属性は無く、式３でａ＝１，属性番号＝０とした式５を用いて、説明文書に出現する照合単位を文字位置情報に変換した。

文字位置情報＝（検索単位番号）×３１０８
＋文字位置番号　（５）

検索入力は、２．２で使用した各々１０００個の漢字列片仮名文字列および平仮名文字列を用いた。なお、Σ２３０モデル２２は32ビット・５ＭＩＰＳのワークステーションである。

### ５．２　実験結果

主記憶に置いた検索ファイルから該当する文字位置情報を抽出するまでの時間を検索時間として、実験結果を表８～表１０に示す。

表８　漢字入力の検索時間

| 検索入力文字数 | 平均 (ms) | 最小 (ms) | 最大 (ms) |
|---|---|---|---|
| 1 | 0.1 | 0.1 | 0.4 |
| 2 | 29.7 | 0.2 | 303.4 |
| 3 | 24.8 | 0.5 | 156.5 |
| 4 | 24.7 | 0.4 | 147.1 |
| 5 | 26.9 | 0.5 | 156.5 |
| 6 | 17.7 | 0.9 | 93.3 |
| 7 | 17.3 | 0.7 | 100.5 |
| 8 | 13.6 | 1.0 | 86.0 |

表９　片仮名入力の検索時間

| 検索入力文字数 | 平均 (ms) | 最小 (ms) | 最大 (ms) |
|---|---|---|---|
| 2 | 0.2 | 0.1 | 0.3 |
| 3 | 6.2 | 0.3 | 5.7 |
| 4 | 7.7 | 0.3 | 53.2 |
| 5 | 4.8 | 0.4 | 38.9 |
| 6 | 6.6 | 0.4 | 25.0 |
| 7 | 4.0 | 0.5 | 42.7 |
| 8 | 3.3 | 0.5 | 22.3 |

表１０　平仮名入力の検索時間

| 検索入力文字数 | 平均 (ms) | 最小 (ms) | 最大 (ms) |
|---|---|---|---|
| 2 | 0.2 | 0.1 | 2.9 |
| 3 | 38.1 | 0.4 | 233.7 |
| 4 | 32.9 | 0.6 | 343.8 |
| 5 | 16.7 | 0.4 | 157.8 |
| 6 | 15.0 | 0.4 | 167.2 |
| 7 | 12.0 | 0.5 | 125.1 |
| 8 | 10.6 | 0.6 | 161.5 |

### ５．３　考察

(1)　検索の高速性について

検索時間は、漢字の１～８文字入力で平均19.4ms、片仮名の２～８文字入力で平均4.69ms、平仮名の２～８文字入力で平均17.9msであった（表８～表１０参照）。検索対象が約895 万文字であるから、等価的に、漢字で平均4.6 億文字／秒、片仮名で平均19.1億文字／秒、平仮名で平均5.0 億文字／秒の検索速度が得られ、本方式が高速であることがわかった。検索入力が漢字の１文字の場合、片仮名および平仮名の２文字の場合、平均検索時間はそれぞれ0.1 msと0.2 msと短いが、これは検索入力そのものが照合単位になるため、照合単位に対応する照合単位種グループ内のすべての文字位置情報が検索結果になることによる。なお、検索ファイルを磁気ディスクに置いたときの検索時間は、磁気ディスクから主記憶への照合単位種グループ転送時間を表８～表１０の各検索時間に加算した時間になる。

(2)　検索入力文字数と検索特性

照合単位の増加に伴い、すなわち、漢字では入力文字の増加に伴い、仮名文字では仮名セットの増加に伴い検索時間が減少した。これは、検索入力から作成する照合単位の中で、低出現照合単位から順に、照合単位種グループ間の文字位置を照合することによる。すなわち、検索入力文字が多くなるほど、検索入力から作成される照合単位の中に出現度数の低いものが含まれる割合が高まるので、低出現照合単位から順に照合すると照合回数が減少し、検索時間が短くなる。なお、漢字入力の９文字以上を調査すると、検索入力文字の増加に伴い検索時間が増加していた。これは、９文字以上の検索入力では、低出現照合単位からの照合による検索時間の減少に比して、文字位置照合に要する時間の増加が大きくなるためと考えられる。

(3)　片仮名と平仮名の検索時間

平仮名入力の平均検索時間は、片仮名入力の平均検索時間に比して３～６倍多い。これは、平仮名の平均出現度数が片仮名の約１１倍と高いことと、片仮名文字列が単語として出現するのに対し、平仮名は単語以外に助詞や助動詞などを構成し、同じ文字列として出現するのが多いことに起因する。

## ６．むすび

本文では、日本語文書から全文検索の照合単位である非仮名文字と文字セットを取り出し、照合単位種ごとの文字位置情報のグループ化，照合単位種グループ内の文字位置情報の昇順配列，低出現照合単位からの文字位置照合により、全文検索を高速に行えることを述べた。また、文書を複数の検索単位で構成する場合には、照合単位が出現する検索単位番号，属性番号および検索単位における配列順位から文字位置情報を作成できることも述べた。

ここでは、日本語文書のように文字種の多い文書における高速全文検索方式を述べたが、本手法を英語文書などのような文字種の少ない文書へを応用するのが、今後の課題である。

## ７．謝辞

取り組みにあたってご支援をいただいた前テレマティーク国際研究所研究技術員飯島豊氏、助言をいただいた東京電力システム研究所卯月主席研究員および藍沢主席研究員、有益な御意見をいただいた前テレマティーク国際研究所山田室長、広辞苑を使用させていただいた岩波書店辞書部の方々に感謝いたします。

## 文　献


(1) D.E.Knuth, J.H.Morris and V.R.Pratt ;" Fast Pattern Matching in Strings", SIAM J.Comput, 6 ,2, pp.322-350(June 1977)

(2) R.S.Boyer and J.S.Moore."A Fast String Searching Algorithm" ,Commun.ACM, 20 ,10 ,pp.762-772(Oct. 1977).

(3) A.V.Aho and M.J.Corasick, "Efficient String Matching: An aid to bibliographic search", Commun.ACM, 18 , 6, pp.333-340 (June 1975).

(4) G.Kowalski and A.Meltzer,"New Multi Term High Speed Text Search Algorithms", First International Conference on Computers and Applications(CAT. No. 84ch2039-6), pp514-522, XIV+905, 1984.

(5) F.J.Burkowski;"A Hardware Hashing Scheme in the Design of a Multiterm String Comparator", IEEE Trans. Comput, COM-31, 9 ,pp825-834(Sept.1982).

(6) M.J.Foster and H.T.Kung;"The Design of Special-Purpose VLSI Chips", IEEE COMPUTER , pp26-40, (Jan.1980).

(7) R.L.Haskin;"Special Purpose Processors for Text Retrieval", Database Enginnering ,4(1), pp16-29(Sept.1981).

(8) G.Salton;"Automatic Information Retrieval", IEEE COMPUTER, pp41-55(Sept.1980).

(9) 井上潮，速水治夫，福岡秀樹，鈴木健司，松永俊雄："データベースプロセッサＲＩＮＤＡの設計と実現"，情処学論，３１，３，pp. ３７３－３８０（１９９０－０３）．

(10) 伊藤正雄，菅野祐司，田村登，安藤敦史，早川佳宏："フルテキストデータベース検索システム「検蔵君」(1)"，１９９０情処秋季全大，１Ｆ－８

(11) 田村登，菅野祐司，伊藤正雄，安藤敦史，早川佳宏："フルテキストデータベース検索システム「検蔵君」(2)"，１９９０情処秋季全大，１Ｆ－９

(12) 山田八郎，平田雅規，永井肇，高橋恒介："文字列検索ＬＳＩ"，信学技報，ＣＡＳ８７－２５

(13) 加藤寛次，藤澤造道，大山光男，川口久光，畠山敦，兼岡則幸，秋沢充："全文検索用テキストサーチマシンの開発"，信学技報，ＤＥ８９－３８

(14) 高橋恒介，永井肇，山田八郎："テキストデータベースの学習型検索方式"，信学技報，ＤＥ８８－３

(15) 菊池忠一，飯島豊："キーワードのコード化による一検索手法"，第３９回情処全大，２Ｎ－５（平１後）

(16) 菊池忠一，飯島豊："構成文字の属性／文字位置を含むコード化による全文検索の高速化手法"，信学技報，ＤＥ９０－２４

(17) 菊池忠一："文字列照合を用いた全文検索における仮名文字検索の高速化手法"，情処学研報，９１－ＤＢＳ－８３