データベース・システム　83－1
（１９９１．５．２４）

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

菊池　忠一
（株）テレマティーク国際研究所

先に報告した「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」には、仮名文字を照合単位とすると、漢字ほどの効果が得られない問題があった。仮名２文字で構成される仮名セットを照合単位とすると、日本語文書に出現する仮名セットおよび仮名セット列の出現頻度が低い特徴があり、本手法はこの特徴を活かしたものである。本文では、仮名文字を照合単位としたソフトウェアによる全文検索の高速化手法を提案するとともに、検索実験結果を報告する。

A Rapid Full-Text Retrieval Method for Kana-String with Character-String Collation

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

There was such a problem in "A Rapid Full-Text Retrieval Method with Coding Character of Full-Text using both an Attribute and Character Location" as this method cannot retriev more rapidly using Kana character than Kanji character as retrieval unit. Because Kana character does not have so many kind of characters as Kanji character. Using Kana-set, which is constructed with 2 kana characters, as retrieval unit, generally, same Kana-set rarely appears and also same Kana-set-string rarely appears in Japanese document. The method herein uses these characteristics of Japanese document. This paper describes a rapid full-text-retrieval method with software using Kana character as retrieval unit and the result of its-experiment.

1　はじめに

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。従来のインデックス方式によるＤＢ化では、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、抽出するキーワードが異なることもあった。また、きめ細かい検索を実現するために、抽出するキーワードが膨大な数になる欠点もあった。これらの解決手段の一つとして、任意の文字列を検索キーとする全文検索が注目されている。全文検索は、全文の中から検索者が指定する文字列と検索条件に合致する文書を選出するもので、全文の最初から最後まで、検索入力文字列と文字列照合する逐次検索が一般的である。しかし、逐次検索は多くの時間を要するため、高速性が求められる。これを実現するものとして、専用プロセサ[1][2]や、文字列検索用ＬＳＩ[3]を用いたものが開発されている。他に、プリサーチ方式を採用した専用マシーン[4]も開発されている。

筆者は、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、ソフトウェアによる任意文字列を入力とする検索の高速化に取り組んできた。先に、図書名を対象とする部分一致検索の高速化手法[5]および漢字列を検索キーとする全文検索の高速化手法[6]を報告した。しかし、これらの手法は、文字種の少ない仮名文字を対象とするとき、文字種の多い漢字ほどの効果が得られない問題があった。

本文では、連続する仮名２文字を仮名セットと呼ぶ照合単位とすると、全文中に出現する仮名文字の出現頻度より仮名セットの出現頻度が低減することを利用した、仮名文字を対象とする全文検索の高速化手法を提案するとともに、広辞苑に記載されている見出し語の説明文を対象とする検索実験結果を報告する。

2　基本概念

全文検索においては、検索対象文書の構成文字列と検索入力文字列との照合回数を低減することが、検索の高速化に有効である。一般的に、日本語文書のように字種の多い文書では、同じ文字や文字列の出現頻度が低い特徴がある。特に、漢字は、第１水準で2,965 字種、第２水準で3,388 字種と字種が多く、登録時に、文書に出現するすべての漢字を文字種別に分類してグループ化しておき、検索時に、これらのグループの中から、検索入力文字と同じ文字グループだけを抽出し、文字グループ間で文字位置の連続性を照合すると照合回数を削減でき、検索の高速化が図れることを先に報告した。[6]

しかし、平仮名が第１水準で83字種、片仮名が同じく86字種と文字種が少ないため、同じ仮名や仮名文字列が文書に出現する頻度は漢字に比して一般的に高い。例えば、「っ」や「ゅ」のような小文字を大文字に、平仮名を片仮名に統合し、これらに特殊記号の「ー」を加えた仮名文字75字種を対象に、広辞苑に記載された見出し語の説明文書を調査すると合計出現度数は3,939,354 、平均出現度数は52,525であった。第１水準漢字の平均出現度数は1,155 である[6]から、仮名文字の平均出現度数は漢字の約45倍になる。このため、仮名の全文検索に漢字に適用した手法[6]を用いると、照合回数が多くなり漢字ほどの高速性が得られない。

一方、連続する仮名２文字を１セットとすると、各セットの出現頻度は構成する仮名に比して大幅に低減させることができる。この結果、漢字と同様に仮名セットを文字列照合を行う照合単位とすると、仮名セットや仮名セット列の出現頻度を低減できるので、照合回数が減って、全文検索を高速化できる以下に、高速化を実現する基本５要素を説明する。

2．1　仮名セットによる高速化

ここでは、文字列照合を行う照合入力の数を減らすことによる検索の高速化について述べる。

仮名セットは、（仮名＋仮名）で構成される文字セットである。文字セットには他に、（非仮名＋仮名）で構成される混合セットＡ、（仮名＋非仮名）

で構成される混合セットBがある。登録文書を任意の文字列で検索できるように、登録文書の先頭から１文字ずつ、下記の①～④の規則で文字セットと非仮名に分解する。

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、登録文字列「遠くのものを指していう語」は、「遠，遠く，くの，のも，もの，のを，を指，指，指し，して，てい，いう，う語，語」に分解でき、仮名セットは「くの，のも，もの，のを，して，てい，いう」、混合文字セットAは「遠く，指し」、混合文字セットBは「を指，う語」、非仮名は「遠，指，語」になる。すなわち、この文字列は、14種／14個の照合単位に分解できる。これらの照合単位を照合単位種別にグループ化して照合対象とすると、検索時には、検索入力文字列の先頭から２文字単位に作成する仮名セットで、検索入力と同じ文字列を再生できる仮名セットの組合せを抽出できる。　従って、ｎを自然数とすると、検索入力文字数が２ｎあるいは（２ｎ－１）の仮名文字列の場合、仮名セットはｎ個になるから、文字列照合を行う照合入力の数は１／２あるいはｎ／（２ｎ－１）になり、照合回数が減少し検索の高速化が図れる。

２．２　文字セット種別グループ化による高速化

文字列照合を行う文字を検索対象から特定することによる検索の高速化について述べる。

２．２．１　日本語文書における仮名セットの出現度数

広辞苑に記載された見出し語の説明文書には、75字種の仮名から作成できる仮名セット5,625 種(=75x75)のうち5,068 種が出現し、仮名セット種の合計出現度数は2,358,677 、平均出現度数は465 であった。このうち、出現度数が１万回以上２万回未満の仮名セットは10種、出現度数が２万回以上の仮名セットは９種あり、これらは「こと，など，るこ，して，また，もの，した，つて，ある」で、最多は「こと」の63,420回、次に多い「など」は33,866回であった。なお、広辞苑には約19.8万の見出し語があり、説明文書のＪＩＳ表記載文字は、約885 万文字であった。

２．２．２　照合対象を特定することによる高速化

仮名セットの平均出現度数が465 であるから、ｎを自然数とするとき、検索入力が２ｎあるいは（２ｎ－１）の仮名文字列の場合、平均465 個の仮名セットを有するｎ個の仮名セットグループから、検索入力を再生できる仮名セットの組合せを抽出することができる。例えば、最も出現度数の高い「こと」と次に高い「など」で構成される文字列の「ことなど」で検索する場合でも、文字列「ことなど」を再生できる組み合わせを、63,420個の「こと」と33,866個の「など」から抽出することができる。一方、文字列「ことなど」を、「こ」「と」「な」「ど」の文字で検索すると、97,521個の「こ」，179,850 個の「と」，119,221 個の「な」，53,377個の「ど」から、文字列「ことなど」を再生できる組合せを抽出することになり、仮名セットより照合対象が多くなる。このように、検索入力から作成する仮名セットに照合対象を特定すると、検索入力の構成文字に照合対象を特定するより照合対象を少なくでき、さらに、説明文書より照合対象をはるかに少なくできる。

以上から、全文中の仮名文字を仮名セット種別にグループ化して登録し、検索入力から作成する仮名セットに対応するグループから、検索入力文字列を再生できる仮名セットの組み合わせを抽出することにより、全文検索の高速化が図れる。

２．３　文字位置番号の照合による高速化

２個の照合単位を作成する３～４文字の仮名文字列の検索の高速化手法について述べる。

2．3．1　昇順配列された数値列間の照合

昇順に数値を配列した2個のグループ間で、任意の差を有する数値の組合せを抽出するとき、該当しない組合せの場合には小さい数値を、該当する組合せの場合には2個の数値を照合の対象外にできるので、すべての数値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　ΣA＝A1, A2, A3, A4, ・・・, Am

ΣB＝B1, B2, B3, B4, ・・・, Bn

およびrを自然数とすると、これらのグループから、AiとAiよりr大きいBjの組み合わせを抽出するとき、以下の条件を満足する数値の項を照合対象外にできる。

Ai ＋n＞Bj ならBj を対象外

Ai ＋n＜Bj ならAi を対象外

Ai ＋n＝Bj なら合致でAi とBj を対象外

従って、Am ＞Bn の場合、照合回数Nは式1で表せる。

N＝k＋n－1－q・・・式1

kは、Ak ＋r＞Bn になるグループΣA内の数値Ak 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、

Bj ＝B7 ＝24＜25＋2 ＝A8 ＋2 ＝Ak ＋2 からk＝8，n＝7，Ai ＋2＝Bj を満足するのは、

A2 ＝9とB3 ＝11，A7 ＝22とB6 ＝24

従って、合致数はq＝2、照合回数はN＝12

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個の数値を照合対象外にできるので、照合回数を削減できる。

2．3．2　文字位置番号の昇順配列による高速化

文字位置番号を、登録文字列から2．1の①～④の規則で作成する文字セットあるいは非仮名文字の、登録文字列における配列順位とする。文字セットは2文字構成であるから、文字セットの文字位置番号は、文字セットの先頭文字の登録文字列における配列順位を示すものとする。これらの文字位置番号を照合単位種別にグループ化し、グループ内で文字位置番号を登録順に配列すると、各グループに同じ照合単位種の文字位置番号を昇順に配列できる。従って、検索入力から作成する仮名セットあるいは非仮名文字の照合単位に対応する照合単位種別グループから、検索入力を再生できる文字位置番号の組み合わせを抽出して検索を行うと、2．3．1で述べた照合回数の低減が図れ、検索を高速化できる。この高速化は、特に、照合単位種グループにある文字位置番号が多いほど効果が大きく、照合単位が2個の場合、あるいは照合単位が3個以上における初めの照合単位種グループ間の照合に有効である。

2．4　文字列照合による高速化

3個以上の照合単位を作成できる5文字以上の仮名文字入力による検索の高速化手法について述べる。

2．4．1　文字列照合時の不連続な文字の削除

一般的に、文字列の文字数が増えるごとに、同じ文字列の出現頻度は減少する。すなわち、検索入力文字列を文字セットあるいは非仮名文字の照合単位に分解し、照合単位を増やしながら、検索入力を再生できる照合単位の組合せを抽出していくと、検索入力から作成した照合単位を増やすごとに、検索入力を再生できる照合単位の組合せが減少し、検索対象が絞り込まれていく。そこで、仮名文字を検索入力として、仮名文字列の先頭にある仮名セットから順に照合単位を増やしながら、広辞苑に記載された見出し語の説明文書を対象に、検索入力を再生できる照合単位の組合せが減少し、検索対象が絞りこまれていく状況を調査した。

片仮名を入力文字列とするときの結果を表1と表2に、平仮名を入力文字列とするときの結果を表3と表4に示す。使用した文字列は、広辞苑に記載さ

れた見出し語の説明文書から、ＪＩＳ表記載の２～８文字の片仮名文字列および平仮名文字列を各々1000個抽出したものである。削除率は、式２で算出した。

削除率=(照合各段階の削除仮名セット累計)
÷(№1仮名セット合致数−最終仮名セット合致数)×100[%] ・・式２

表１　片仮名文字列照合時の削除率

照合入力	n文字セット照合時の削除率[%]			
文字数	n＝1	n＝2	n＝3	n＝4
2	100	----	----	----
3	0	100	----	----
4	0	100	----	----
5	0	99.7	100	----
6	0	99.0	100	----
7	0	96.9	99.8	100
8	0	97.9	99.7	100

表２　片仮名文字列照合時の合致数

照合入力	n文字セット照合時の合致数			
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３　平仮名文字列照合時の削除率

照合入力	n文字セット照合時の削除率[%]			
文字数	n＝1	n＝2	n＝3	n＝4
2	100	----	----	----
3	0	100	----	----
4	0	100	----	----
5	0	99.0	100	----
6	0	97.1	100	----
7	0	97.0	99.9	100
8	0	96.5	99.7	100

表４　平仮名文字列照合時の合致数

照合入力	n文字セット照合時の合致数			
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

２．４．２　不連続な文字の削除による高速化

入力文字数＝５以上における削除率は、№２仮名セット照合結果が、片仮名が入力の場合で96.9%以上、平仮名が入力の場合で96.5%以上である。従って、№１仮名セット～№２仮名セットの連続性照合で、同じ仮名セットを先頭に有する仮名文字列の中から、片仮名が入力の場合に96.9%以上、平仮名が入力の場合に96.5%以上が不連続として照合対象外になり、合致して残るのが各々3.1 %以下および3.5 %以下であることがわかった。

このように、同じ仮名文字列の出現頻度が少ないことに着目すると、５文字以上の検索入力では、№１仮名セット～№２仮名セットの照合で96%以上も絞りこまれ、５文字目（№３仮名セット）以降では、絞り込まれて残った照合結果との照合になるから、照合回数が減少し、検索の高速化が図れる。

なお、入力文字数＝２では、入力文字列から作成された仮名セットに対応する照合単位種グループの文字位置番号がすべて検索結果であるから照合処理は不要で、削除率は100 %になる。また、入力文字数＝３と４では、入力文字列から作成された仮名セットに対応する２個の照合単位種グループから検索結果が得られるから、№１仮名セットの削除率は０%，№２仮名セットの削除率は100 %になる。

２．５　出現頻度の低い仮名ユニットからの照合による高速化

3個以上の照合単位を作成できる5文字以上の仮名文字入力による検索を、さらに高速化する手法について述べる。

2．5．1　文字列照合時の不連続な文字の削除

3個以上の照合単位が作成される文字列が入力のときには、2ステップ以上の文字列照合を行って検索結果を抽出する。第1ステップは、2個の照合単位に対応する照合単位種グループ間の照合であるので照合回数が多いのに対し、第2ステップ以降は、これまでの照合結果と照合単位種グループ間の照合になるから、照合回数は少ない。表2と表4に示すように、第1ステップの照合結果は絞り込まれ少なくなるから、全ステップの照合回数を減らすには、第1ステップに対応する照合単位種グループを文字位置番号の少ないものから選び、照合回数を低減させるとともに照合結果も減らすことが有効である。そこで、5～8文字の仮名文字列を入力として、先頭にある仮名セットから順に文字列照合を行う場合と、出現度数の少ない仮名セット順に文字列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8に示す。入力文字列は、上記2．4で使用したものと同じである。

2．5．2　不連続な文字の削除による高速化

出現度数の少ない仮名セット順に文字列照合を行う場合の照合回数と、先頭にある仮名セットから順に文字列照合を行う場合の照合回数を比較すると、片仮名が入力の場合で前者は後者の約30%～51%であり、平仮名が入力の場合で前者は後者の36%～56%に減少している。また、先頭にある仮名セットから順に文字列照合を行う場合の照合回数は、入力文字数の増加にともない増えるが、出現度数の少ない仮名セット順に文字列照合を行う場合の照合回数は、入力文字数が多くなるほど減少している。これは、入力文字数が多くなるほど入力文字列から作成される仮名セットの中に、出現度数の少ない仮名セットが含まれる確立が高くなることによる。以上のことから、入力文字列の中の出現度数の低い仮名セットから順に、照合単位種グループ間の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ことがわかる。

表5　片仮名文字列照合時の照合回数
（先頭にある仮名セットから順に照合）

照合入力	n文字セット照合時の照合回数				
文字数	n＝1	n＝2	n＝3	n＝4	合　計
5	0	4498	332	----	4830
6	0	4481	784	----	5265
7	0	4540	189	168	4897
8	0	4037	189	133	4359

表6　平仮名文字列照合時の照合回数
（先頭にある仮名セットから順に照合）

照合入力	n文字セット照合時の照合回数				
文字数	n＝1	n＝2	n＝3	n＝4	合　計
5	0	11304	1281	----	12585
6	0	10379	1246	----	11625
7	0	11977	911	499	12888
8	0	10643	947	596	12186

表7　片仮名文字列照合時の照合回数
（出現度数の少ない仮名セット順に照合）

照合入力	n文字セット照合時の照合回数				
文字数	n＝1	n＝2	n＝3	n＝4	合　計
5	0	1499	488	----	1987
6	0	1844	856	----	2700
7	0	1167	183	245	1595
8	0	971	101	226	1298

表8　平仮名文字列照合時の照合回数
（出現度数の少ない仮名セット順に照合）

照合入力	n文字セット照合時の照合回数				
文字数	n＝1	n＝2	n＝3	n＝4	合　計
5	0	5291	1814	----	7105
6	0	4947	1401	----	6348
7	0	3704	689	678	5071
8	0	3129	613	711	4453

3　検索方式

一般的な書籍を対象として、登録処理および検索処理を述べる。登録処理では、文字位置情報と検索ファイルの作成を述べる。

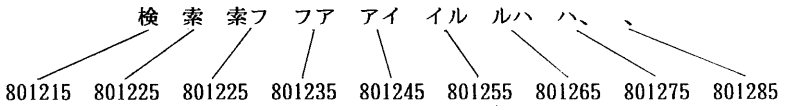

図1　文字位置情報への変換例

3．1　文字位置情報の作成

文字位置情報は、書籍に記載された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される。

(1)　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。

(2)　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。

(3)　文字位置番号

検索単位を構成する文字列から、2．1の①～④の規則で作成する文字セットあるいは非仮名文字の、文字列における配列順位を文字位置番号とした。文字セットは2文字構成であるから、文字セットの先頭文字の文字列における配列順位を、文字セットの文字位置番号とした。

(4)　文字位置情報

式3を用いて、検索単位から分解した照合単位を自然数のコードに変換し、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)　・・式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される。

このとき、文字位置情報を4バイトでコード化すると、最大10000 文字長の検索単位を、

$2^{64} \div (n \times a)$ ＝約4万個取り扱うことができる。

3．2　検索ファイルの作成

検索ファイルは、図2に示すように、第1および第2水準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、各検索単位から分解されたすべての照合単位に付与された文字位置情報が、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「通信ファイルは、」を登録した例を図3に示す。

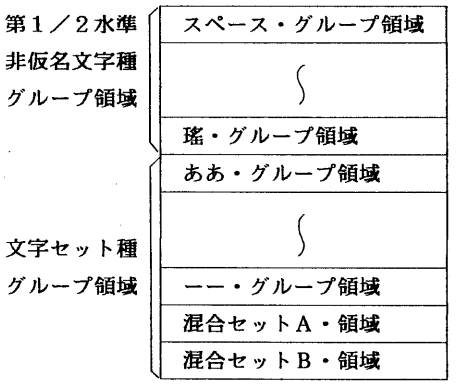

図2　検索ファイルの構成

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×Σ（検索単位の照合単位数）になる。なお、追加登録は、追加する検索単位の各照合

＜照合単位＞	
	・・・・・・・・・・・・・・・・
（、）	・・801285・・・・・・・・・・
	・・・・・・・・・・・・・・・・
（検）	・・・・・・801215・・・・・・
	・・・・・・・・・・・・・・・・
（索）	・・・・・・・・・801225・・・
	・・・・・・・・・・・・・・・・
（アイ）	・・801245・・・・・・・・・・
	・・・・・・・・・・・・・・・・
（イル）	・・・・・・801255・・・・・・
	・・・・・・・・・・・・・・・・
（ハ、）	・・・・・・・・・801275・・・
	・・・・・・・・・・・・・・・・
（フア）	・・・・801235・・・・・・・・
	・・・・・・・・・・・・・・・・
（ルハ）	・・・・801265・・・・・・・・
	・・・・・・・・・・・・・・・・
（索フ）	・・・・801225・・・・・・・・
	・・・・・・・・・・・・・・・・

図３　検索ファイルの例

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×Σ（検索単位の照合単位数）になる。なお、追加登録は、追加する検索単位の各照合単位に対応する照合単位種グループの末尾に、新たに文字位置情報を追加して行う。また、削除は、削除する検索単位あるいは文字列の各照合単位に対応する照合単位種グループにある該当文字位置情報を特殊記号に変更して行う。これにより、追加登録と削除を短時間で行うことができる。

３．３　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ている文字位置情報の連続性を照合し、検索結果として、検索入力文字列と同じ文字列を再生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、出現頻度の低い照合単位から順に並べ変える。

(2)　文字列照合

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５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。

{(検索入力文字位置番号iの照合単位のグループ格納文字位置情報)－
(検索入力文字位置番号jの照合単位のグループ格納文字位置情報)}=(i－j)×a
・・・式５

a：最大属性種数

(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。

例えば、本文に記述された「検索ファイル」を検索する場合、照合単位は「検，索，フア，イル」になる。これらの照合単位の出現頻度を、検＝1,008，索＝200，フア＝395，イル＝13,873とすると、並べ変えた照合単位列は「索フア検イル」になる。文字列照合では、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５を使用して文字列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。

４　検索実験

４．１　実験条件

本手法の性能を評価するために、ＰＦＵ社製Σ２

３０モデル２２上に、広辞苑に記載された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、説明文書を合計すると約895 万文字になる。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６を用いて、説明文書を構成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)x3108+(文字位置番号) ・・・式６

実験では、前記２．４で使用した片仮名文字列と平仮名文字列を検索入力とし、主記憶に置かれ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測定した。なお、Σ２３０モデル２２は、32ビット・５ＭＩＰＳのＵＮＩＸミニコンである。

４．２　実験結果

初期登録，追加登録および検索の実験結果を以下に示す。

(1)　初期登録

広辞苑の先頭から抽出する説明文書を、100 万文字から800 万文字まで100 万文字ごとに増やし、初期登録の登録時間と登録した照合単位数を表９に示す。

表９　初期登録時間

文書数	文字数	照合単位数（B）	登録時間（m）
21,117	1,000,046	1,182,241	28
43,575	2,000,006	2,361,456	58
66,812	3,000,006	3,536,570	82
88,849	4,000,006	4,711,585	121
112,679	5,000,023	5,883,354	150
134,276	6,000,026	7,060,529	184
155,634	7,000,011	8,236,937	212
177,607	8,000,024	9,411,394	247

表１０　追加登録時間

被追加文書			追加文書			登録時間
文書数	文字数	照合単位数	文書数	文字数	照合単位数	（m）
0	0	0	21,117	1,000,046	1,182,241	28
21,117	1,000,046	1,182,241	22,458	999,980	1,179,215	30
43,575	2,000,006	2,361,456	23,237	999,994	1,175,114	32
66,812	3,000,000	3,536,570	22,037	1,000,027	1,175,015	34
88,849	4,000,027	4,711,585	23,830	999,996	1,171,769	36
112,679	5,000,023	5,883,354	21,597	1,000,003	1,177,175	37
134,276	6,000,026	7,060,529	21,358	999,985	1,176,408	38
155,634	7,000,011	8,236,937	21,973	1,000,013	1,174,457	40

す。検索ファイル容量は、文字位置情報が４バイトであるから、４バイト×（照合単位数）になる。

(2)　追加登録

広辞苑の先頭から順に説明文書を100 万文字単位に抽出し、初めは100 万文字の初期登録を行い、次からは登録済み文書に順次追加登録したときの登録時間と追加登録した照合単位数を表１０に示す。

表１１　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１２　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

(3) 検索

検索ファイルを主記憶上に展開したときの、片仮名を検索入力とする検索時間を表１１に、平仮名を検索入力とする検索時間を表１２に示す。

４．３ 考察

(1) 検索の高速性について

検索時間は、例えば、片仮名の５文字入力で平均4.8 ms、８文字入力で平均3.3 ms、平仮名の５字入力で平均16.7ms、８文字入力で平均10.6msと短く、本方式が高速であることがわかる。なお、検索入力が２文字の場合、検索時間は平均0.2 msと小さいが、これは検索入力そのものが仮名セットになるため、仮名セットに対応する照合単位種グループ内の文字位置情報がすべて、検索結果になることによる。

(2) 検索入力文字数と検索特性

検索入力が３文字から８文字まで、すなわち、２仮名セットから４仮名セットまで、検索入力文字数の増加にともない検索時間が減少している。これは、検索入力から作成する仮名セットの中で、出現度数の少ない仮名セットから順に、照合単位種グループ間の文字位置照合を行うことによる高速化効果の現われである。すなわち、検索入力の文字数が多くなるほど、検索入力から作成される仮名セットの中に、出現度数の少ないものが含まれる確率が高くなるので、照合回数の低減が図れ、検索を高速化できる。

(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、片仮名入力の検索時間の３～６倍多い。これは、片仮名文字列は単語を抽出したのに対し、平仮名では単語を含む文字列を抽出したことにより、同じ助詞や助動詞などで構成される仮名セットや仮名セット列を多く抽出したことに起因する。

５．おわりに

本方式は、漢字より字種の少ない仮名文字による全文検索の高速化を図るため、連続する仮名２文字で仮名セットを構成するとき、全文中の同一仮名セットおよび同一仮名セット列の出現頻度が低いという日本語文書の特徴を活かしたものである。登録時に、全文に記載された照合単位を、照合単位が出現する検索単位の属性と検索単位番号および文字位置番号を用いて文字位置情報に変換し、照合単位に対応する照合単位種グループに格納して検索ファイルを作成する。検索時に、検索入力から作成する照合単位に対応する照合単位種グループを検索ファイルから取り出し、検索単位属性と検索単位番号および文字位置番号を用いて、検索入力と同じ文字列を再生できる文字位置情報の組合せを抽出すると、全文検索を高速化できることを述べた。

なお、本方式は、検索入力された文字列と同じ文字列を全文から抽出することであるため、実用化にはさらに、曖昧検索入力への対応、同義語や異表記などへの対応、等の課題がある。

６．謝辞

本テーマの取り組みにあたって支援をいただいた前当社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外月主席研究員および藍沢主席研究員、日頃から有益な御意見をいただいている山田室長に感謝いたします。

参考文献

[1] 早川 他 「ストリームデータプロセサＳＤＰ(1)」第37回情処全大．3N-9（昭６３後）

[2] 田村 他 「テキストデータベース検索システム「検蔵君」(2)」第41回情処全大．1F－9 （平２後）

[3] 菊地 他 「ＩＳＳＰを用いたテキスト検索システムの試作」第35回情処全大．5S-9（昭６２後）

[4] 加藤 他 「全文検索用テキストサーチマシーンの開発」信学技報DE-89-38

[5] 菊池 他 「キーワードのコード化による一検索方式」第39回情処全大．2N-5（平１後）

[6] 菊池 他 「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DE-90-24