文字成分表を用いた大規模全文検索方式の開発 ―ハッシュレス文字成分表の高精度化方式―

7E－6

多田　勝己　畠山　敦　川口　久光　水谷　奈津子　加藤　寛次
(株) 日立製作所　情報・通信開発本部

1．概　要

近年，電子化文書の急速な増加にともない大量の文書情報をエンドユーザが簡単に蓄積，検索できる文書検索システムに対する要求が高まりつつある。

こうした要求に応えるため，報告者等は文字成分表を用いて検索対象とする文書を絞り込むことにより検索を等価的に高速化する階層型プリサーチ方式を開発してきた[1]。

今回，文字成分表だけで検索結果を得ることのできる大規模文書DB用全文検索方式について検討した。その結果，各文字成分に対し文字成分表の１エントリを割り当てるハッシュレス文字成分表方式とともに，ハッシュレス文字成分表の検索精度をさらに向上させる方式として一文字おきに隣接する二文字を成分とするスキップ連接文字成分表方式とカタカナなどの表記のゆらぎを許容する検索（異表記検索）が指定された場合に表記のバリエーションを部分的に展開してから検索を行う部分展開異表記検索方式を開発することができた。本稿では，この方式の概要と実データを用いた評価について報告する。

2．ハッシュレス文字成分表方式の課題

ハッシュレス文字成分表方式では，これまで文字成分表の容量を削減するために行っていた文字成分表の畳み込み(ハッシング)を行わず，全ての連接文字成分について１対１で文字成分表を割り当てる。このため，一文字または二文字の検索タームでは検索ノイズが発生せず三文字以上の検索タームについても従来の文字成分表に比べ検索ノイズを大幅に低減することができる。しかし，以下に示す二つの要因により検索ノイズが発生する可能性がある。

まず，検索タームから抽出された連接文字成分の組合せによっては検索ノイズが発生することがある。例えば，“動画像”などの“動画”と“画像”の二つの単語の組合せで構成される文字列が検索タームに指定された場合，ここから抽出される連接文字である“動画”と“画像”という単語は含まれるが，検索タームである“動画像”は含まれない文書が検索ノイズとしてヒットしてしまう恐れがある。

第二に，検索タームと文書中の言葉の間の表記上の食違いを許容する検索（異表記検索）を指定した場合に検索精度が低下する可能性がある。すなわち，異表記検索では図１に示すように表記しうるバリエーションを部分的に展開し，各部分文字列毎に連接文字成分を抽出することにより文字成分表サーチを行う。このため，“ンタ”や“タフ”などの部分文字列間にまたがった連接文字成分を文字成分表サーチに利用できないことになり，検索の精度が低下してしまう。

検索ターム：
“インタフェース” ⇨ イン(タ(ー)／タ／タァ)(フェ(ー)／フェイ)ス

〈下線部が検索に用いられる文字成分〉

図１．異表記検索の概要

3．ハッシュレス文字成分表の高精度検索方式

(1) スキップ連接文字成分表方式

“動画像”などのように単語の組合せで構成される文字列が検索タームに指定された場合に生じる検索ノイズを削減する方式として，一文字おきに連接文字成分を抽出するスキップ連接文字成分表方式を検討した。この方式では，従来通りテキストから隣り合う二文字を連接文字成分（逐次連接文字成分）として抽出するとともに，さらに一文字おきに二文字の文字列をスキップ連接文字成分として抽出する。そして，三文字以上の検索タームが指定された場合には，逐次連接文字成分表とスキップ連接文字成分表の両方を参照することにより文字成分表サーチを実行する。例えば，図２に示すように“動画像”という文字列を含む文書が登録された場合には，“動

Development of a Large Full- Text Retrieval System using Character Occurrence Bitmap
--High Precisoin Search Method of Hashless Character Occurrence Bitmap--
Katsumi TADA , Atsushi HATAKEYAMA , Hisamitsu KAWAGUCHI , Natsuko MIZUTANI , Kanji KATO
Infomation Systems R&D Division,Hitachi,Ltd.

画”と“画像”を逐次連接文字成分表に登録するとともに一文字おきに二文字の文字列“動像”をスキップ連接文字成分表に登録する。そして検索タームに“動画像”が指定された時には，“動画”と“画像”で逐次連接文字成分表をサーチするとともに“動像”でスキップ連接文字成分表をサーチし，これらの結果間の論理積をとることにより検索結果を得る。こうすることにより，“動画”と“画像”の二単語を別々に含むが“動画像”という単語を含まない文書からはスキップ連接文字成分“動像”が抽出されないため，文字成分表サーチの結果から除外されることになり検索ノイズとして削除することができる。

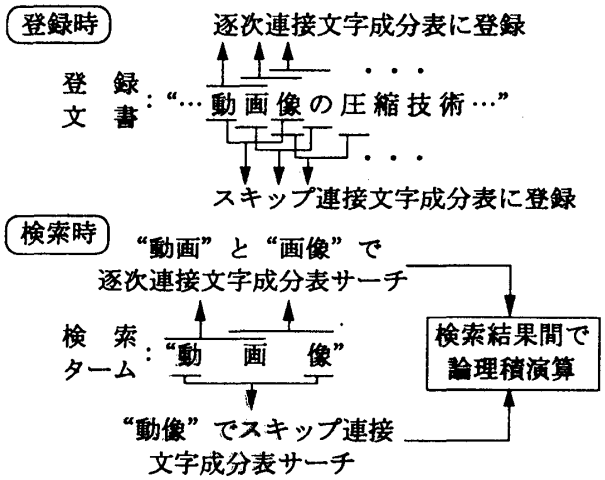

図２．スキップ連接文字成分表方式

(2) 部分展開異表記検索方式

異表記検索が指定された場合に生じる検索ノイズを削減する方式として，異表記のバリエーションを部分的に展開してから連接文字成分を抽出する部分展開異表記検索方式を検討した。

この方式では図３に示すように，まず初めに入れ子構造で表された異表記文字列を入れ子のないレベル（第一レベル）へ展開する。そして，各層の先頭二文字を前の層の文字列の末尾に付加することにより異表記文字を部分的に展開する。そして，展開された各部分文字列毎に連接文字成分を抽出することにより，指定された検索タームに対する文字成分表サーチ結果を得る。以上の処理により，従来抽出できなかった“ンタ”や“ーフ”などの部分文字列間にまたがった連接文字成分を検索に利用することができるため，文字成分表の段階で十分に検索精度の高い検索結果を得ることができるようになる。

“インタフェース”の展開処理例：

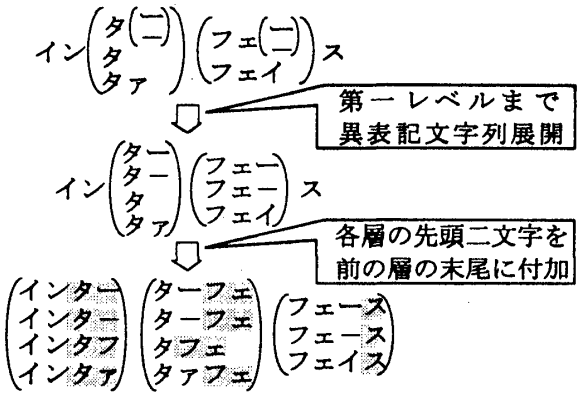

図３．部分展開異表記検索方式

４．評価結果

(1) スキップ連接文字成分表の容量

10万件の特許公報データに対しスキップ連接文字成分表を作成し，ファイル容量を測定した。その結果，テキストデータの総容量1.5GBに対して約40%の600MBになった。すなわち，逐次連接文字成分表とスキップ連接文字成分表を併せて本文の約60%の容量で実現できる見通しが得られた。

(2) 検索精度

いくつかの検索タームについて検索精度を評価した結果を表１に示す。従来方式では，比較的多く検索ノイズを発生していた検索タームについても精度の高い検索結果が得られていることが分かる。

表１．検索精度の評価結果（単位：件）

番号	検索ターム	正解ヒット件数	従来方式	本方式
1	動画像	329	654	329 (100%)
2	色変換	162	243	163 (99%)
3	インタフェース	6,056	15,250	6,059 (100%)
4	データベース	12,378	12,616	12,378 (100%)
5	コンピュータ	1,302	29,294	1,334 (98%)

注1：カタカナ文字列については異表記検索を指定
注2：()内は正解ヒット件数に対する正解率を表す

(3) 検索速度

10万件の特許公報データを10回登録することにより100万件のデータベースを作成し，検索レスポンスを評価したところ3文字の検索タームで100万件のデータベースを約1秒で検索できる見通しを得た。

参考文献

[1] 畠山,他,「ソフトウェアによるテキストサーチマシンの実現」,情報処理学基礎研究会,25-4,(1992.5)