7E－5

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構成方式－

畠山 敦　多田 勝己　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，特許公報の CD-ROM によるテキストとイメージの配布に伴い，大規模な全文検索に対するニーズが高まっている。筆者等は，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，１エントリに複数文字成分を割り当てていた従来の文字成分表を，１対１に対応させるハッシュレス文字成分表方式を開発することができた。本方式は，文字成分表の容量の増大を防ぐために，登録するデータ中に存在する連接文字だけを文字成分表に登録するハッシュレス文字成分表方式と，ビットリスト形式と文書 ID リスト形式を併用する混在型格納方式で構成される。本稿では，その基本方式と実データを用いた評価結果について報告する。

2. ハッシュレス文字成分表方式

従来の文字成分表方式では，図１に示すように，複数個の連接文字成分を一つのエントリへ割り当てるハッシング型の連接文字成分表を用いている。この連接文字成分表では，２文字の連接文字により文字成分表での絞り込み率を高め，また複数の連接文字を一つのエントリへ畳み込むことにより文字成分表容量の低減を図っている。

しかし，この方式では文字成分表の同一エントリに割り当てられた別の連接文字を含む文書がノイズ文書として検索されるため，文字成分表の段階での検索結果をシステムの検索結果として出力するには検索精度上不十分である。

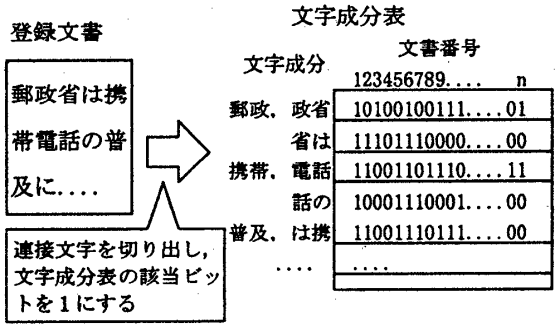

図１　従来の連接文字成分表

そこで，連接文字成分の１エントリへのハッシングを行なわず，連接文字成分と文字成分表のエントリを１対１対に対応付ける。これにより，検索精度の向上が図れると予想されるが，文字成分表の容量が巨大化する恐れがある。

しかし，連接文字の出現状況を実際の特許公開公報 10 万件のテキストデータを対象に調査してみると，出現する連接文字数は，全ての連接文字の組み合わせ数(144M 個)の５％にも満たないことが分かる。

そこで，ハッシュレス文字成分表では，登録文書内に出現する連接文字成分だけを管理するとともに，出現頻度の高い連接文字成分については文書番号をビット位置で記述し，出現頻度の低い連接文字成分については出現文書番号をそのまま格納するビットリスト・ID リスト混在型格納方式を採用した。これにより，文字成分表容量が大幅に削減できることになった。

Development of a Large Full-Text Retrieval System using Character Occurrence Bitmap
-- Structure of Hashless Character Occurrence Bitmap --
Atsushi Hatakeyama, Katsumi Tada, Hisamitsu Kawaguchi, Natsuko Mizutani, Kanji Kato
Information Systems R&D Division, Hitachi,Ltd.

3. 実装方式

図2にハッシュレス文字成分表の実装方式の概略を示す。ビットリストを格納したファイルとID リストを格納したファイルとに文字成分表を区別し，これらのファイルへアクセスするためのディレクトリ情報を第一文字テーブルと第二文字テーブルに格納する。第一文字テーブルは，連接文字の第一文字目を添字とする配列で，第二文字テーブルへのオフセット値を格納する。第二文字テーブルには，連接文字の第二文字を格納し，各連接文字成分に対応するビットリストあるいはID リストを格納するファイル識別子 SFID と各ファイルのオフセット値 offset を格納する。これらのテーブルを用いて，検索タームを構成するすべての連接文字成分のビットリスト又は ID リストを取得し，それらの積集合を取ることにより該当する連接文字を全て含む文書を文字成分表検索結果として得ることができる。

4. 評 価

(1)文字成分表の容量

10 万件の特許公報データをハッシュレス文字成分表に登録し，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，文字成分表は約 20%の 340MB になった。

(2)検索精度

検索に良く使われる検索タームを数語選び，10 万件の特許公報 DB を用いて，検索精度を評価した。その結果を表1に示す。

表 1 検索精度

項番	検索ターム	正解件数	文字成分表ヒット件数
1	文書	1,530	1,530(100%)
2	半導体	14,218	14,282(100%)
3	画像圧縮	185	187 (99%)
4	磁気記録	2,397	2,488 (98%)
5	フルテキストサーチ	2	3 (67%)

注1：( )内は正解件数に対する正解率を表わす

連接文字成分を使っているので，2文字以下の検索タームについては 100%の検索精度が得られる。更なる検索精度の向上に関する検討は，次の発表 7E-6 で説明する。

参考文献

[1] 畠山，他2，「ソフトウェアによるテキストサーチマシンの実現」，情処自然言語処理研報，Vol.92, No.32, 25-4, pp.19-25(1992.5)

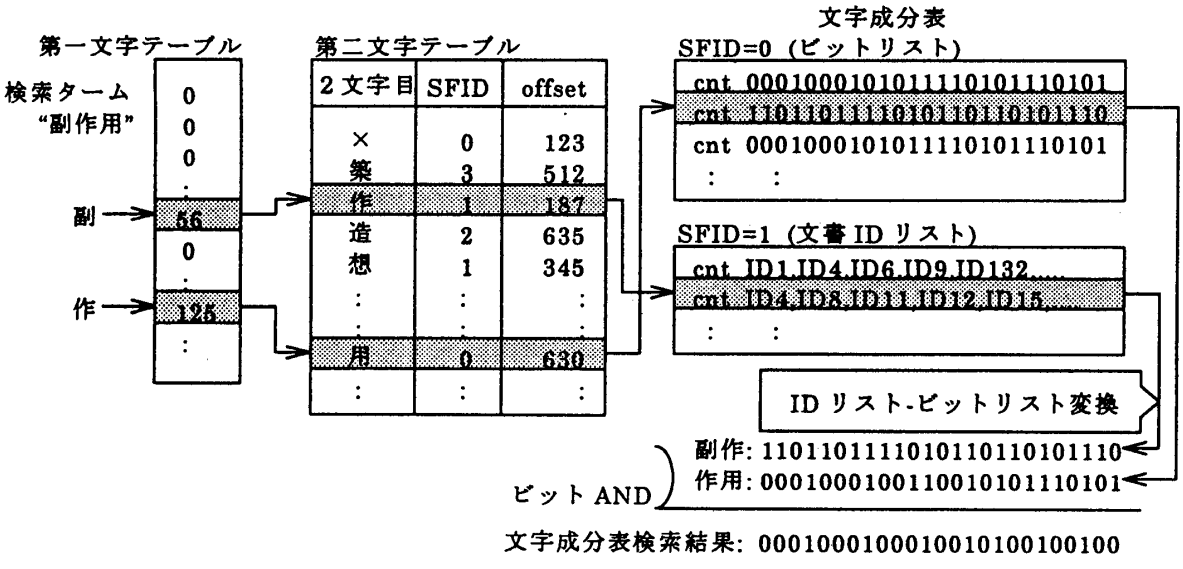

図2 ハッシュレス文字成分表の実装方式