フルテキストサーチシステム Bibliotheca/TSの開発(2)
ーサーチアルゴリズムー

2C−4

川下靖司　浅川悟志　坂田　淳　畠山　敦
(株)日立製作所

1　はじめに

近年、企業内における文書情報の量は増大しており、それらの殆どは電子文書化される傾向にある。

そのような電子文書を検索するシステムとして、フルテキストサーチシステム“Bibliotheca/TS”を開発した。ここでは、検索サーバのサーチアルゴリズム、特に、階層プリサーチ方式という検索アルゴリズムで用いる連接文字成分表について報告する。

2　文字成分表の実現方式

文字成分表とは、図1に示すように文字が文書中に存在するか否かを0か1のビットリストで表したものである。これまで我々が検討してきた方式は、1文字を1成分とする単一文字成分表であった。[1][2]　検索時には、検索語を構成する全ての文字に対し、文字成分表で対応するそれぞれのビットリストを取り出し、そのビットリストを用いて検索語の存在の有無を判定し、文書を絞り込むのである。

階層プリサーチ方式での検索速度は、文字成分表でいかに精度よく文書を絞り込めるかにかかっている。なぜなら、文字成分表サーチでの絞り込み率が悪ければ、次の階層で多くの凝縮テキストとテキストを検索することになり、検索時間が増大するからである。このことから、今回は文字成分表の絞り込み率をできるだけ向上するように、検索精度について検討した。

【テキスト】

文書1	あいまいな表現は、東洋圏……
文書2	表現の違いこそあれ内容は……
文書N	表情からは、彼の意図する……

【単一文字成分表】

	あ	…	い	…	表	…	現	…	
文書1	1		1		1		1		
文書2	1		1		1		1		
文書N	0		0		1		0		

【単一文字成分表】　↓

	あい	…	いま	…	違い	…	表現	…	
文書1	1		1		0		1		
文書2	0		0		1		1		
文書N	0		0		0		0		

図1　文字成分表の概要図

2.1　文字成分表における課題

以下に、単一文字成分表の主な課題をあげる。

(1) 高頻度な文字の検索精度の向上

例えば、図1の単一文字成分表を用いて、“あいまい”という語を検索した場合、文書1、文書2が候補文書としてあげられるが、文書2は検索語が含まれておらず、ノイズとして検出されている。これらノイズを削減することが課題となる。

(2) 検索精度の一様化

文字成分表の容量を抑えるために、ハッシュ法を用いて1ビットに複数文字を対応させているので、ノイズが発生する。

例えば、文字“あ”が100件、文字“濱”が1件の場合、文字“あ”と“濱”が同一エントリにハッシングされると両文字の論理和分がヒットする。この場合、文字“あ”で検索すると検索精度は非常に良いが、文字“濱”では“あ”を含む文書が全てヒットするので大量にノイズが含まれることになる。このようなアンバランスを無くすことが課題である。

これらの課題に対して、次のようなアプローチを行った。

2.2　連接文字成分表

絞り込み率向上のために、連接した2文字を文字成分とする連接文字成分表を開発した。これにより、単一文字成分表ではノイズになっていた文書を減らすことが可能となる。

2.3　頻度情報ハッシュ方式

文字出現頻度差によるノイズを抑えるために、文書に使

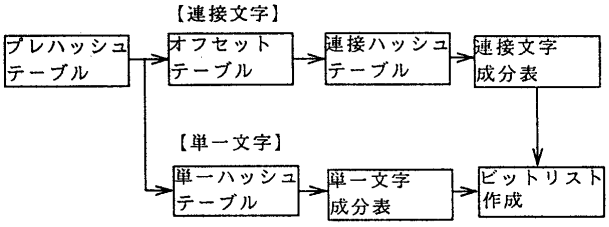

図2　文字成分表アクセスのアルゴリズム

Development of Full Text Search System Bibliotheca/TS (2)　Search Algorithm
Yasushi KAWASHIMO, Satoshi ASAKAWA, Jun SAKATA, Atushi HATAKEYAMA
HITACHI, Ltd.

用されている文字の頻度を調べ、この頻度情報とにハッシュ方式を定義する頻度情報ハッシュ方式を開発した。これは、頻度の大きい文字については別々のエントリに割り付け、頻度の少ない文字に対しては同一エントリに複数個の文字が入るように調節する方式である。つまり、同一エントリに複数文字を割り付ける頻度基準を決定し、この基準より低い文字を低頻度側に積み上げるわけである。これにより、平均的に安定した精度が得られるようになる。

2.4 プレハッシュ方式

Bibliotheca/TSでは、使用可能な文字コードを１２，１５６文字に設定した。これに対する連接文字の種類は、約１億４千通りになってしまい、扱う連接ハッシュテーブルの大きさが膨大になってしまうため、実用的でない。そこで、連接ハッシュテーブルの大きさを抑えるために、連接文字成分表を参照する前に１度ハッシングを行うことにした。これにより、テーブルの大きさを必要最小限に抑えることができた。

3 連接文字成分表のアルゴリズム

連接文字成分表は、ひらがな等、出現頻度の高い文字について、効率的な絞り込みを行うために使用する。メモリ削減の観点から、第２水準等の出現頻度の低い文字については従来通り、単一文字成分表を使用する。

図２は、連接文字、単一文字の文字成分表へのアクセス方法を示す図である。まず、検索語を構成する文字を用いてプレハッシュテーブルを参照する。プレハッシュテーブルには連接文字の使用の有無を記載しており、隣接した２文字をプレハッシュした時点で、連接文字成分があるか否かの判定ができるようにした。もしあれば、オフセットテーブルを参照して連接ハッシュテーブルへのアクセス場所を算出し、得られたエントリ番号で連接文字成分表を参照する。また、連接文字成分がない、または、検索語が１文字からなる場合には、単一ハッシュテーブルにより、単一文字成分表を参照する。以上の処理により、連接文字、単一文字に対応する文字成分のビットリストを得る。

この処理を、全検索文字に対して行い、得られたビットリストの論理積から文書を絞り込むことになる。

4 連接文字成分表の性能評価

１件平均１ＫＢの新聞記事を１９，９１１件登録したＤＢを用いた場合の、単一文字成分表と連接文字成分表の検索結果の比較を表１に示す。

連接文字成分表は単一文字成分表に比べ、実件数の多い場合、ノイズは殆ど無くなっている。また、実件数の少ない場合でも単一文字成分表に比べ、精度が向上されていることが分かる。

文字成分表サーチの処理時間はＤＢの登録文書件数と検索語の文字数に依存する。そこで、検索語が２文字と５文字の場合について、登録文書件数を横軸に文字成分表サーチ時間を縦軸にとったグラフを図３に示す。

図３の直線は、回帰直線であるが、この直線の傾きが文字成分表サーチ速度と見なすことができる。検索語数が２文字のときは143万件／ｓ、125万件／ｓである。

5．まとめ

今回、Bibliotheca/TSで高速検索を実現するため、連接文字成分表を用いた階層プリサーチ方式を実現した。これにより、文字成分表の検索精度をあげ、ノイズの削減を図ることができた。

また、文字成分表の大きさを抑えることで文字成分表をメモリにのせ、アクセス速度を向上することができた。

《参考文献》

１．加藤、他４、「大規模文書システム用テキストサーチマシンの研究」情報処理学基礎 14-6 1989.7

２．加藤、他６、「大規模文書システム用テキストサーチマシンの開発」1991年情報学シンポジュウム予稿集

３．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」情報処理学基礎 25-4,1992.5

表１　文字成分表性能比較(19,911件/ＤＢ)

検索語	実件数	単一文字成分表 ヒット件数	単一文字成分表 精度 [%]	連接文字成分表 ヒット件数	連接文字成分表 精度 [%]
スト	2,798	8,427	33.2	2,799	99.9
スリ	206	6,132	3.4	562	36.7
もち	473	5,895	8.0	1,243	38.1
のど	134	13,273	1.0	1,395	9.6
うめ	13	10,460	0.1	1,128	1.2
日本	7,169	11,089	64.6	7,170	99.9
一日	2,361	13,416	17.6	2,364	99.9
国内	1,964	5,215	37.7	1,974	99.5

(精度＝実件数／ヒット件数)

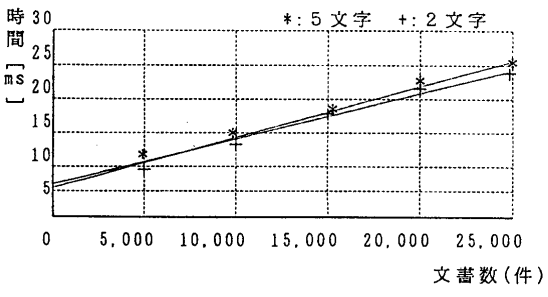

図３　文字成分表の処理時間