フルテキストサーチ用フィルタリング型高速文字列照合方式

1C－9

多田勝己*1 川口久光*1 畠山 敦*1 加藤寛次*1 浅川悟志*2

*1(株)日立製作所 中央研究所 *2(株)日立製作所 ソフトウェア開発本部

1．はじめに

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一般の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，我々は階層型プリサーチ方式を用いたソフトウェアによる全文検索システムBibliotheca/TSを開発した[1]。このシステムでは，文字列照合処理を全てサーチエンジンエミュレータ(ＳＥＥ)と呼ぶソフトウェアで実現しているためシステム検索速度としては約10MB/sであり，オフィスにおいてグループで共有される数百MBの文書容量を対象とした場合，性能的に不十分であった。

この問題に対し，ＳＥＥの照合速度をさらに高速化するフィルタリング型ＳＥＥ方式を開発したので，この方式について報告する。

2．サーチエンジンエミュレータ高速化の課題

2.1 サーチエンジンエミュレータの概要

サーチエンジンエミュレータＳＥＥでは，検索対象文書から読み込んだ文字列(入力テキストと呼ぶ)中に，指定した検索タームが含まれるか否かを判定する文字列照合処理を行う。ＳＥＥは図1に示すように文字コード変換部と照合処理部で構成される。

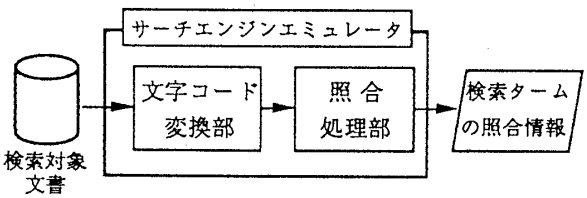

図1．サーチエンジンエミュレータ

(1) 文字コード変換部

文字コード変換部では，照合処理部で用いるオートマトンの状態遷移テーブルのサイズを縮小するために，1バイトコード文字と2バイトコード文字が混在する入力テキストを，文字コード間の隙間を取り除いた内部文字コード体系に圧縮変換する。

(2) 照合処理部

照合処理部では，コンカレントステート型オートマトン照合方式(ＣＳＡ方式)と呼ぶ文字列照合方法を採用している[2]。このＣＳＡ方式では図2に示すように，入力文字コードに対して逐一，状態番号0からの遷移(初期遷移)の有無を調べ，遷移が存在するときにはその遷移先状態に遷移情報を格納する。そして，次の文字コードが読み込まれたときに，その文字コードに対して初期遷移の有無を調べると同時に，前ステップで保存した遷移情報をもとにその状態からの遷移の有無を調べることにより同義語や異表記語を含む複数の検索タームを一回のテキスト走査で照合する。

検索ターム1:「データ圧縮」　検索ターム2:「圧縮比」

¬{デ,圧}

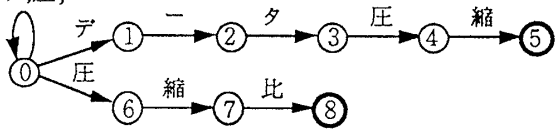

図2．ＣＳＡ方式

2.2 サーチエンジンエミュレータ高速化の課題

ワークステーション3050/R(公称57MIPS)を用いて約10MBのテキストをＳＥＥで照合した時の処理時間の内訳を図3に示す。この結果から，ＳＥＥ処理時間の大半を照合処理部が占めていることが分かる。

図3．ＳＥＥの処理時間内訳

照合処理部では，図2の例において「デ」または「圧」が入力されて初めて，初期遷移が発生する。すなわち，「デ」または「圧」が入力されるまでは遷移が発生する可能性がないにもかかわらず，初期状態0からの状態遷移の有無を判定するという無駄な処理をしている。したがって，ＳＥＥの照合速度を高めるためには，この無駄な処理を省き照合処理部の負荷を軽減することが課題となる。

A Fast String Searching Argorithm with Filtering Function for Full Text Search

Katsumi TADA*1, Hisamitsu KAWAGUCHI*1, Atsushi HATAKEYAMA*1, Kanji KATO*1, Satoshi ASAKAWA*2

*1 Central Resarch Laboratory,Hitachi,Ltd. *2 Software Development Center,Hitachi,Ltd.

3．フィルタリング型サーチエンジンエミュレータ方式

上記課題に対し，照合処理の前処理として，入力された文字が検索タームに含まれるか否かを判定し，検索タームに含まれない場合にはこれを切り落し，照合処理の対象から外してしまうフィルタリング処理を文字コード変換部で行うことにより，照合処理部の負荷を軽減し，全体の照合処理スループットを引き上げるフィルタリング型ＳＥＥ方式を開発した。

この方式では，図４に示すように検索タームの先頭に指定された文字(先頭文字と呼ぶ)が現われるまで，全ての入力文字を削除し照合処理の対象から外す。そして先頭文字が現われて初めて検索タームに含まれる文字(検索指定文字と呼ぶ)を連続的に照合処理部に出力する。次に，検索タームに含まれない文字(非検索指定文字と呼ぶ)が現われると，その文字に対しては照合処理を行うが，その次の入力文字からは再度先頭文字が現われるまで全ての文字を削除し照合処理の対象から外す。このようなフィルタリング処理を行うことにより，文字コード変換部からの出力を「デ」,「ー」,「タ」,「圧」,「縮」および「を」だけに削減でき，照合処理部の負荷を軽減することができる。

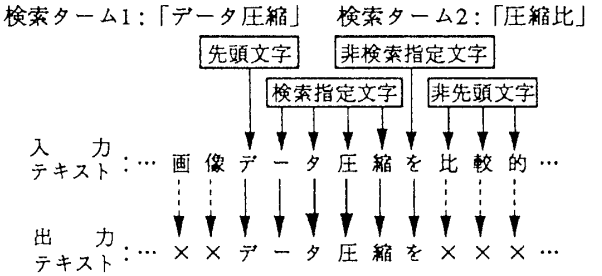

図４．フィルタリング処理の例

4．性能評価

4.1 性能測定条件

上記方式に基づくフィルタリング型ＳＥＥのプロトタイプを作成し，その性能を評価した。フィルタリング型ＳＥＥの性能はフィルタリングを通過する文字数に大きく依存するため，パラメータとして下に示すフィルタリング率を定義した。

$$\text{フィルタリング率}=\frac{\text{フィルタリング出力文字数}}{\text{フィルタリング入力文字数}}(\%)$$

特許明細書テキストを対象として，フィルタリングの効果を試算した結果を表１に示す。特許明細書テキスト中で出現頻度が高い「出力」や「装置」が検索タームに指定された場合でもフィルタリング率は1%を切る値になるため，照合処理部の負荷を1/100以下に軽減できると推測される。

表１ フィルタリング率

検索ターム	認識	音声	消滅	画像	圧縮	出力	装置
フィルタリング率(%)	0.02	0.03	0.04	0.13	0.37	0.66	0.83

4.2 測定結果

測定用ワークステーションとして3050/R(公称57MIPS)を使用し，フィルタリング型ＳＥＥと従来方型ＳＥＥの単体照合速度を測定した。この結果を図５に示す。このように，従来方式の約３倍に相当する最高5.4MB/sの照合速度を実現することができた。

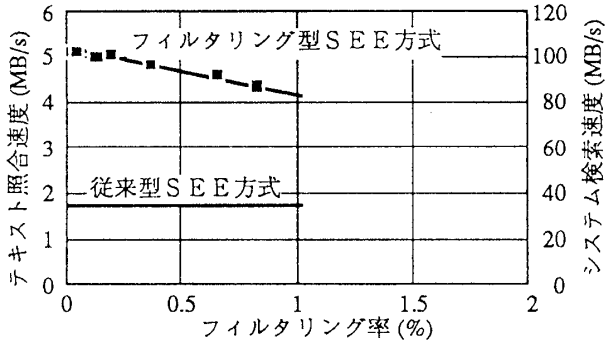

図５．特許明細書検索時のSEE照合速度

5．おわりに

文字列照合処理の前段階として検索タームに含まれない文字を全て削除し検索対象から外すことにより等価的に照合速度を向上させるフィルタリング型サーチエンジンエミュレータ方式を開発した。これにより，従来のＳＥＥ方式に比べ照合速度を約３倍高速化することができ，ワークステーション3050/R(公称57MIPS)上で最高5.4MB/sの照合速度を実現することができた。その結果，システム全体では，従来のワークステーション3050(公称20MIPS)上のBibliotheca/TSに比べ約10倍の300MBの文書容量を，３～４秒で検索することが可能になった。

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)
[2] 川口,他,「自由語検索のための高速文字列検索方式」,情報処理学会第39回全国大会,pp.1078～1079,(1989.10)