# 日本語フルテキスト検索プロセッサ

4B－3

金田悟　菊地芳秀　辻澤隆彦

NEC機能エレクトロニクス研究所

## 1. はじめに

フルテキスト検索の高速化技術を大別すると、文字列照合ハードウェアを用いる手法と、テキストデータから自動抽出した二次情報を用いて検索するソフトウェアによる手法がある[1]。二次情報を小さくすると、二次情報だけでは一致が特定できず、テキストデータを改めて文字列検索する必要がある。この高速化には、テキストデータの蓄積媒体からの読み出しと文字列照合の高速化が必要である。

文字列照合速度の向上を図るため、計算機に内蔵可能な回路規模の文字列検索プロセッサをゲートアレイ(7万ゲート、160pinQFP)で開発した。目標として、文字列照合に関連する各種の機能を１チップ化することと、周辺回路を含めたハードウェア規模を小さく抑えることを目指した。

## 2. 機能と仕様

日本語テキストデータベースを検索し、曖昧な検索機能もある程度求められた。このため、

・異なるバイト長の文字コードが混在するテキストを検索できること。

・正規表現で記述される文字列を検索できること。

の２つの機能をハードウェアで実現した。論理条件検索は、多くの場合は小規模な演算ですむので、ソフトウェアで実現することとした。

性能上の仕様は、以下のように設定した。

・同時検索語数 32語(4個並列化で128語まで可能)

・検索速度 30MByte/sec

## 3. 照合方式

従来の文字列照合方式を大別すると、外付けの状態遷移表メモリを用いる方式と、内部にキーワードの文字数分の比較器を持つ方式に別れるが、今回は高速性と周辺回路規模の縮小のため、後者をベースとした。曖昧な文字を含むキーワードの検索も可能なように、文字の集合を検出する回路を新規に設けた。また、後者の欠点である検索語数の少なさを補うため、今回は４チップの並列動作を可能とし、多くの場合に十分と予測される128語の検索語数を得た。

## 4. ブロック構成

多機能化と高速性の両立のため、内部処理はパイプライン化した。内部のブロック構成を図1に示す。このパイプラインに検索対象となるテキストデータをDMA転送などで流し込むと、必要な処理が行われ、一致結果がキーワードの番号とその一致テキストアドレスとして出力され、DMA転送などで引き取られる。あらかじめキーワードなどはI/O書き込みにより内部レジスタセットに設定する。

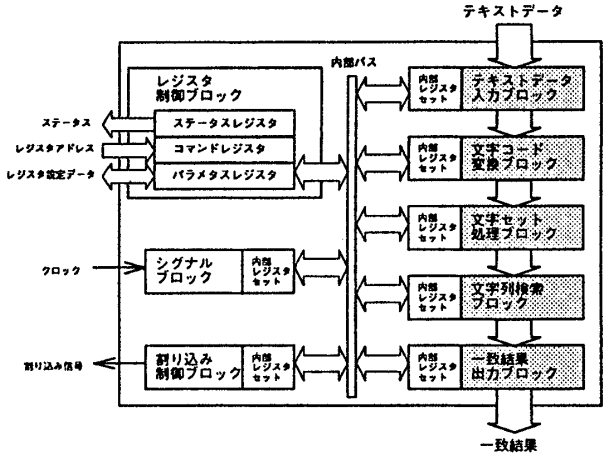


図１文字列検索プロセッサのブロック構成

## 5. パイプラインの構成

パイプライン上では、まず、一定でない文字コード長を内部処理用の２バイトコードに変換する(文字コード変換ブロック)。次に、キーワードの曖昧な文字の集合(文字セット)の検出や、空白文字などの意味を持たない制御文字の検出してテキストデータ

Full text retrieval processor
Satoru KANEDA, Yoshihide KIKUCHI, TakahikoTSUJISAWA
NEC Corporation

にマークする(文字セット処理ブロック)。これとキーワードを比較し、一致結果を出力する(文字列検索ブロック)。

### 5-1 文字コード変換ブロック

日本語規格の EUCコード系テキスト(1から3バイト長の文字コードで構成される)を、内部処理用の2バイトコードに変換する。各バイトの最上位ビットで、4種類あるコードセットの種類を表す。

### 5-2 文字セット処理ブロック

文字セット処理ブロックは、文字列検索の前に、テキストデータからキーワードに含まれる文字セットを検出したり、制御文字を検出する。これにより、単純な文字列検索回路で、正規表現文字列の検索や、制御文字を無視することが可能になる。

図2に示すように、テキストデータ列中に文字セットがあると、その後に文字セットに対応したコード(文字セットコード)を挿入する。また、テキストデータ中の制御文字には、制御文字であることを示すフラグを付加する。

### 5-3 文字列検索ブロック

文字列検索ブロック(図3)は、4文字×32個の文字照合セルと、その一致順序から文字列一致を判定する順序回路で構成され、 文字照合セルに設定されたキーワードを入力されたテキストデータから文字列検索する。4文字ごとの文字列一致結果を次段に送ることで、任意の長さのキーワードを検索できる。

## 6. システムへの実装上の課題

本プロセッサをUNIX ワークステーション(NEC製EWS-4800/360)の汎用I/Oバス(APバス)を介して利用するボード(図4)を作成し、フルテキスト検索システムを開発した。本体メモリとI/Oバスの間でのデータ転送の高速化が課題であった。これに関して、I/Oバスの問題は、より大きなワード数でのバースト転送にはDMAコントローラをFPGA等で独自に開発しなければならないことである。デバイスドライバでは、データ転送のオーバヘッドを小さくする必要があった。また、アプリケーション開発でも、処理の並列化とプロセス間通信の最適化が必要となる。

## 7. おわりに

フルテキスト検索で必要となる文字列検索処理を行なう専用プロセッサを開発した。複数の候補文字を含む文字列をテキストデータから照合する機能や、異なるバイト長の文字コードを含むテキストデータを検索する機能を、パイプライン状に構成することで、高速性と多機能性を両立させた。

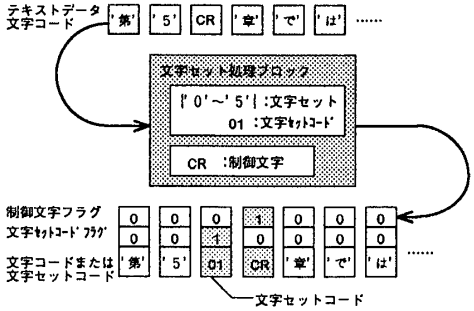


図2 文字セット処理ブロック

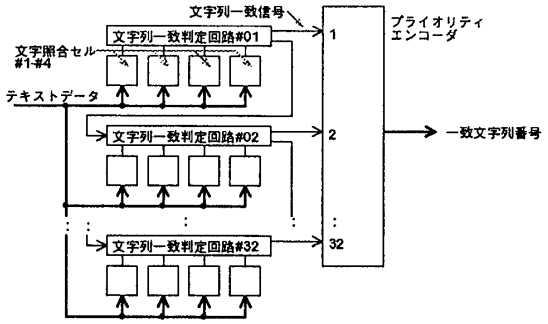


図3 文字列検索ブロック

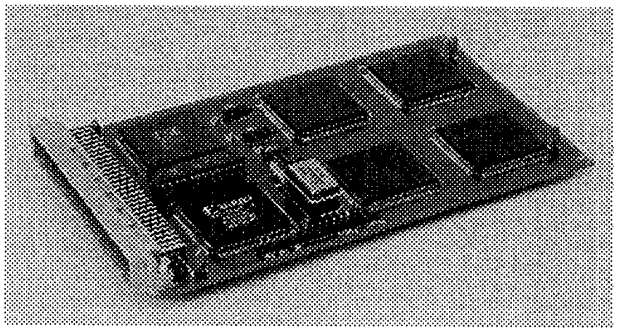

図4 フルテキスト検索ボード(170mm×100mm)

## 参考文献

[1] 菊地他,全文検索の技術動向とｼｽﾃﾑ事例, 情報学基礎 25-1, 1992