画像とテキストの高速検索について

国立民族学博物館　杉　田　繁　治

1. ハードウエアでできることはハードウエアで

コンピュータが出現しだした当初は，まだその能力が低かった為に，あることがらを実現するのに専用のハードウエアで論理回路を組あげていたものである．ところがコンピュータの記憶容量が大きくなり，またそのスピードが速くなってくると，かなりの部分をコンピュータのソフトウエアで実現しようということになってきた．ハードウエアによる場合は変更することができないので融通のきかないシステムになってしまうからである．

ところが，コンピュータを計算や制御に使う場合はかなり威力を発揮するが，パターン認識や大量のデータを検索する場合には現在のコンピュータはまだ満足できるものではない．そこである特定の機能を高速で実現するハードウエアを集積回路技術を活用して実現しようという傾向が最近でてきている．これはすべての機能をハードで実現するのではなく，汎用コンピュータで制御しながら特定の機能だけを専用のハードで実現しようというものである．ステップバイステップの論理の適用ではなく，力ずくでやってしまおうという発想であるが，現在のマイクロエレクトロニクス技術の発達を考えれば，以前とは異なりうまくいくのではないかと思われる．

2. 構造化データベースから単純データ集合へ

コンピュータの発達につれて，その応用としてデータベースが各分野で作成されるようになった．そして検索をす速くする為に様々な工夫がこらされるようになった．対象とすべきデータが多くなるにつれて単純な検索では時間がかかり過ぎるからである．いくつものインデックスをあらかじめ作成しておき，それによって検索時間を短くする工夫がなされている．またそのインデックスを検索する方法自体もプログラミングのテクニックを工夫して高速化してきた．

ところが，もし生のデータ集合に対してある文字列の存在を高速に検索することができれば，そのような工夫は必要がなくなるのである．実際それを実現するハードウエアが存在し，今回共同研究の設備として導入し，実験している．これは数メガバイトのデータ全部を検索するのに１秒以下である．また同時にいくつかの文字列を AND, OR で結合した条件で本の頁やドキュメントを検索することができる．また完全一致のみならず，ドントケヤを含む文字列でも検索できる．例えば「民？学」と入力すれば「民族学」や「民俗学」なども一致したとして検索できる．「変？？法」とすれば「変形文法」や「変形方法」などが検索できる．

？の数を可変にしたような検索もコンピュータとの連動で可能である．とにかく検索が速いからデータにあらかじめ細工をせずにいれておいてもよいのである．コンピュータが広く利用されるようになると，コンピュータやプログラムの知識がなくても使えるようなシステムが必要になるが，このテキスト検索ハードウエアは今後のデータベースの方向に大きな変化を与えるものとして期待されるものである．

3. 広い領域を表示

従来のコンピュータは文字を表示することが中心であった為に，表示面積もあまり大きくな

くても役に立っていた．またマルチウインドウのような工夫によって，いくつかの情報を同時に表示することができた．しかし画像を扱うようになると大型のスクリーンが必要になることがすぐに分る．国立民族学博物館では2000×2000ドットのディスプレイを導入して，標本資料の画像を複数個同時に表示するシステムを構築している．しかし，例えばランドサットやスポットなどのデータは1シーン当り縦横6000から9000ドットあり，大型のディスプレイでも一度には表示できない．

今回導入した装置は8層のフレームを持ち，各層は64メガドットの容量があるのでスポットのデータもスッポリ入ってしまう．そしてスムーズな高速スクロールによって任意の領域を表示させることができる．これは全体が一覧できるというわけではないが，それに匹敵する表示能力を有していると考えられる．とくにこの装置では，拡大，縮小が非常にスムーズに行なわれるので全体から局部まで自由に見ることができる．これも専用ハードウエアとコンピュータとの連動による新しい機能の実現である．

4. イメージデータに対する高速パターン認識の実現へ

文字の自動読取りや図形の認識は，コンピュータの歴史と共に研究の対象となってきた問題であるが，まだ解決されていない部分が多い．しかし文字列の高速検索や画像の高速スクロールや拡大縮小が行なえるようになると，文字パターンの検索や図柄の検索も不可能でないように思われる．文字や図形の特徴を分析するのではなく，ドットパターンの一致の度合を高速にチェックする機能を活用すれば，実現できそうである．電子ファイルにイメージとして入力されているドキュメントに対して，その中に記述されている文字を認識することが可能かもしれない．そうすればフルテキストデータベースの作成も非常に容易になる．既に出版されている書籍などは電子ファイルに蓄積しておけば，やがてそれをコードに変換することができるであろう．それを期待している．

国文学研究とパーソナルデータベース

国文学研究資料館 北村 啓子・安永 尚志

国文学者の中でも身近なパソコンを研究の道具として使う人が増えてきており，パーソナル環境での研究支援の要望が多い．また，パソコン上で様々なメディアを比較的容易に扱えるようになってきており，メディアを越えて相互にリファレンスしながらの研究環境が期待されている．

そこで，従来の“大型計算機環境”と新しい“パーソナル環境”それぞれでの研究支援方法，および両者の有機的な利用形態の構想を試みる．ここではそれぞれの環境に加え，両者をコミュニケートする環境として，オンラインで結ぶ“ネットワーク環境”，ニューメディアを介して(オフラインで)結ぶための“プロバイダ”(環境というよりシステム)を考え，これら4者が担うべき機能の分担を行い，それぞれを有機的に利用する研究支援環境の提案を行った．またこの環境において，利用者がパーソナルデータベースをいかにして獲得，利用していけるかを，利用者の立場から見た具体的イメージで説明した．

パーソナル環境での研究支援として，最初に着手したのが目録型データベースをCD-ROMで提供することである．今回開発したマイクロ資料目録CD-ROMについて，データの内容，データ構造上の特徴，ならびに検索システムの特徴と主な機能を紹介した．特に検索システム