

データ活用に関する手法(Ⅱ)

オンライン対話形時系列データ解析システム

粕谷 英樹 (宇都宮大学工学部)

1 はじめに

医療データは“特殊”であると言われる。たとえばデータの形式という面を考えてみても、臨床検査から得られる数値データ、心電図等の時系列データ、X線写真などの図形データ、患者の病状や経過のレポートのような言語データ、病名や医薬品のような用語データ、と極めて多様である。これらのデータの、コンピュータによる活用という問題は、2つに分けて考えた方がよい。1つは、データを、統一的にしかも効率的に処理するための、ソフトウェアシステムの設計思想ということであり、他の1つは、個々のデータを目的に適つたように処理するための手法、つまり応用プログラムの問題である。勿論、考えるシステムに関連性の深い応用プログラムもあるが、一般にデータ処理のための応用プログラムは、システムと独立に考えてよい。たとえば臨床研究用に頻りに利用される各種の統計処理の手法は、システムがどんなものであつても変わるものではない。一方、ソフトウェアシステムの面に関しては、データを統合化した共用ファイルとして構成し、データの集収・登録・修正・追加・検索等の処理を統一的に行おうという、いわゆる医療データベースがある。データベース言語としてのMUMPSやGEMISHデータベースシステム等がこの例である。このようなシステムは、病院とか地域医療に応用されて、その有効性が確かめられている。

しかし、たとえば耳鼻咽喉科領域だけとか、あるいは病院のある1つの教室での臨床あるいは基礎研究のためのデータ処理を考えた場合は、データベースシステムよりは、もつと小まわりのきく、しかもデータの処理手法の開発(応用プログラムの開発)に便利なシステムを持ちたいという要求も当然あろう。その場合には、1研究組織に属するそれぞれの研究者が、別々のプロジェクトを持つていても、全体として無駄のない効率的なデータ処理システムを設計するということが大切である。

こゝでは、前回の講習会で報告されたような時系列データ(アナログデータ)を、効率良く統一的に処理するソフトウェアシステムを構築するための問題点について、若干論じてみたい。

手本として、米国カリフォルニア州の音声通信研究所で開発された、ILS(Interactive Laboratory System)を取り上げる。これは、主に音声データの処理に設計されたもので、国内は勿論、欧州そして日本等の音声工学・医学の研究者達によつて、広く使われているものである。

2 対話形システム

ある問題が発生したら、その解法を発見してプログラミングし、それを実行して結果を得、更に解法を修正したり、別の解法を試みたりというように、コンピュータとの対話を繰り返しながら最適解を求めて行く、対話形システムは、バッチ処理などに比べて格段に融通性に富み、研究を能率的に逐行する上で欠かせない。この場合コンピュータは、人間の思考過程の中に一体となつて存在することになるので、コマンドやプログラミング言語が使い易いものであり、端末装置も問題の解決に適したものでなければならない。ILSシステムでは、標準の対話用端末装置として、グラフィックディスプレイを用いる。

3 高級プログラミング言語

対話形オンラインデータ解析システムを作るには、機種によつて異なるアセンブラ言語のような低級言語ではなく、高級言語を用いる方が良い。高級言語を用いれば、プログラミングの時間が少なく、オペレーティングシステムが変更されたり、機種が変つても、それ程大きな影響を受けることはない。高級言語の中でも、対話形に適したBASICやAPLとか、科学技術計算に適したFORTRANが、よく利用される。ILSシステムは、FORTRANをプログラミング言語として用いている。

4 プログラムのモジュール化

システム全体に柔軟性を持たせるためには、プログラムをモジュール化することが重要である。モジュール化することによつて、ある問題を解くのに何回かのコマンドに分けて実行しなければならず、一見不便のようであるが、それを補つて余りあるような利点がある。その第1点はプログラミングの労力を節約できることである。たとえば、A、B、Cという3種類のプログラムモジュールがあつて、それぞれが異つた解法による結果を、中間結果として出すとしよう。次のステップで、それらの中間結果を利用して、更に3種類の処理D、E、Fを行うとしよう。これらのプログラムを一本化して実行する場合には、9つの組み合わせすべて用意しなければならないが、モジュール化することによつて、6種類のプログラムだけで充分である。第2点は、モジュール化することによつて、主記憶の容量やディスクのスペースを大巾に節約できるということである。また、多くのミニコンピュータでは、一人の利用者が利用できる主記憶の容量は限られている。このためにも、プログラムのモジュール化は必要である。

5 操作性

対話形システムでは、システムの操作性が重要である。特別に、コンピュータの知識を持つていなくとも、目的とするデータの解析を滞りなく進めることができるためには、操作性を重視したシステム設計がなされていなければならない。ILSシステムでは、ミニコンピュータのオペレーティングシステムのコマンドインタープリタを上手に利用することによつて操作性を上げている。たとえば、グラフィックディスプレイに、収集した時系列データ（たとえば音声波形）を表示するにはD (Display)を、聴いてみるにはL (Listen)を、またそのデータの周波数スペクトルを計算して、ディスプレイに表示するにはFD (Frequency Display)という文字を、ディスプレイのキーボードから入力するだけで良い。この場合、どのデータファイルの何番目のデータを処理するかの指定は、処理に先立つて、あらかじめ行わなければならない。

6 プログラムやデータファイルの標準化

多くの利用者が、1つのデータ解析システムを共用する場合の、一番重要な問題が、データファイルの標準化とプログラムの規約に関することである。データベースシステムという考え方からすれば、データファイルは(応用)プログラムから独立であるのが望ましい。しかし、前にも述べたように ILS システムは、データベースシステムを指向したものではないために、データの一部はプログラム(の COMMON)が受け持っている。たとえば、使用しているコンピュータのハードウェアとしての定数(1語長、ディスクのブロック長等)とか、時系列解析に用いるいろいろなパラメータ(サンプリング周波数、周波数スペクトルを求めるときの波形の切り出しに関する窓の条件等)である。ILS システムでは、4種類の標準的なデータファイルが用意されている。

- (1) 時系列データファイル(“生”のデータファイル)
- (2) 解析ファイル(基本的な時系列解析プログラムの処理結果用ファイル)
- (3) 記録ファイル(いろいろな処理の中間結果などの記録用ファイル)
- (4) 標記ファイル(時系列データに標記するための ASCII ファイル)

これらの標準化されたファイルを用いて、データの解析処理が進められる。それぞれのファイルは、ファイルに包摂的な情報を含むファイルヘッダを持ち、ファイルヘッダのフォーマットも標準化されている。たとえば、ファイルヘッダの決められた場所に、そのファイルのタイプを示す符号が割り当てられており、プログラムが実行される度に、最初にチェックし、そのプログラムが実際にアクセスしても良いかどうか調べる。こうすることによつて、貴重なデータファイルを間違つて消してしまうことのないようにしている。

ILS システムは、時系列データの中でも音声データの処理を目的にして設計されているため、4種類のファイルは、更に音声データの処理に便利のように標準化されている。したがつて、耳鼻咽喉科領域における音声検査への応用ということであれば、このまゝのファイル構造でも充分利用価値があるが、もしもつと広範囲の時系列データを処理することを目的としたシステムを作ろうとするのであれば、ファイルヘッダの仕様を再検討する必要があるだろう。

7 ILS プログラムの種類

ILS システムには、すでに 100 個を超えるプログラムモジュールが組み込まれている。それ等を分類すると、

- | | |
|---------------|---------------|
| (1) 図形・グラフの表示 | (5) データの変換 |
| (2) 統計処理 | (6) 信号処理 |
| (3) 統計的パターン認識 | (7) 音声の分析(合成) |
| (4) データ検索 | |

等である。上の(1)から(6)のプログラムは、音声データに限らずいろいろな時系列データの処理に利用できるのもので、耳鼻咽喉科領域への応用範囲はかなり広い。

8 おわりに

紙幅の関係で、中途半端な報告になつてしまつたが、講習会では具体的な音声検査への応用例を示しながら、オンライン対話形データ解析システムの特徴について述べてみたい。