数理解析研究所講究録
第461巻 1982年 226-235

構造を有するデータの入力システムCODEの開発

京都大学　工学部　　徳田成穂
上林弥彦
矢島脩三

1. まえがき

最近の科学技術の急速な進歩により、技術移転の効率化が大きな問題となってきている。高品質のデータベース作成はこの問題に対処する一つの方法である。このためには、大きな組織が機械的に収集したもののほかに、特定の専門分野に対してその方面の研究者がこまやく収集したものも重要である。しかし、後者のようなデータベース作成が効率良く行なえることが重要で、本稿ではそのようなシステムを開発した結果について述べている。

さきに我々は、データベース文献を約3900件収集して刊行した[KAMB81]。この作業結果を分析反省し、あらたに効率の良いデータ入力システムを設計した。このシステムは、TSSシステムの便利さをそのまま実現した会話型データベース入力システムCODEと、データ内の構成要素の分

離作業をデータ入力者が行なわなくてすむ、非専門家向きのSIFTという2つの入力システムより構成されている。本稿では、すでに開発を終了したCODEについて述べているが、システム全体の構成は[KAMBL8111]を、SIFTについては[KAMBL820]を参照されたい。

2. データベース入力システムへの要請

さきに行なった文献集[KAMB81]作成作業を分析した結果、文献データのためのデータベース入力システムに対する要請を次のようにまとめてみた。

R1：システムは非常に長いデータや非常に多くの構成要素よりなる集合を扱う必要がある。

たとえば、非常に長い文献標題や、著者が50人もある論文、参考文献が100件を越す論文がある。

R2：論文を表現するための種々の表現形式を扱える必要がある。さらに、本、プロシーディングス、それらの中の論文、レポート、マニュアル等の区別や、それら独自の表現形式を扱える必要がある。ある論文では、著者、その論文の含まれている論文集の編者、さらにその論文集がシリーズのときシリーズの編者等といった人名集合が対応することになる。データ入力中に、予想していなかった属性等を臨時に追加できる機能も重要である。

Y3：よく出現するデータに対する入力作業量を減少させる。会議名，雑誌名，所属等に対して重複入力を避けたい。

Y4：種々の名称に対する標準的記法や標準的略記法を決めて検査できること．1つの会議名が種々の形で記されることが多いので，文献集としては統一をとる必要がある。

Y5：暗黙値の指定のできること．暗黙値には，ほとんどのデータに対して指定されるもの（言語＝英語）と，一時的なものとがある。

Y6：データの等価性が指定できること．会議名(FJCC, SJCC→NCC），著者名（結婚による），大学名等も変化するので，索引作成上必要である。

Y7：集合や順序集合の扱えること。参考文献はその中の値（年や著者名）によって順序が決められるので非順序集合として扱ってもよいが，著者名は順序集合である。とくに集合の要素間に対応付けのある場合順序集合が必要である。

Y8：正規化されていない集合も扱えること．データの正規化は，データ操作を容易にする上で重要であるが，正規化しないで一体として扱う方が良いものもある。コメント集合等は正規化の必要はない。

Y9：集合間の対応のとれること。著者集合と所属集合の

同等集合間で対応関係をとる必要がある。直積や1対1対応といった表現法がある。

Y10：不完全情報の扱えること。たとえば、現在不明なデータをあとで補うための空値や、データの信頼性を示す情報も必要である。

Y11：値の検査。書式や最大値，最小値以外のものとして、pp、i-j における i<j がある。また、標題等の属性には空値は許されていない。

Y12：データの修正。KWICに使うため、ーや／に対して使われ方で区別する。SDD-1やPL/Iは1語として扱うが、A RELATIONAL SYSTEM-INGRESや、CAD / CAM は空白を入れて分ける。また、KWIC のストップ語となってはいるが別の意味になっている語を表現できなくてはならない。

Y13：データの自動作成。文献IDを著者名や年より作る等のデータの作成。また、データの作成者や日時等も記憶しておく必要がある。

Y14：データの修正が容易である。

Y15：意味制約による誤り検出機能を持たせうること。

Y16：データをクラスター化して出力できること

3. CODEの機能

CODEの主な特性は次のとおりである。

(1) 属性名はシステムの初期化時に与える。各属性に対しATOMIC, SET, ORDERED SET, 書式, 属性の定義域等の指定ができる。現在のシステムでは属性数は99個までである。

(2) システムは、グループによる共同作業に適している。マスターファイルは複数の利用者より利用できる。相互排除や利用者間の通信もサポートされている。同じ利用者IDが同時に複数の端末より使えないようになっている。複雑なパスワード機能が用意され、利用者、日時は自動的に登録される。

(3) システムのコマンドはTSSのコマンドとほぼ同じである。端末の性格の違いもある程度扱える。

(4) データ入力操作を減らすため、暗黙値指定を動的に変えたり、割込み的な作業を許すためのスタックがある。

(5) TSSエディタと同等のエディタが用意されている。

(6) システムの拡張が容易なように、プログラム(PL/I)中のすべての変数はEXTERNALとなっている。

現在のCODEでは、2節の各要請は次のように実現されている。

V1：各属性の長さは可変長になっている。システムの起

動時に最大値を指定することにより、長さ0から最大値まで、自由な長さのデータを受け付ける。集合の要素は空白2つを介して連結されるため、要素の数の自由度は属性長に換言できる。

Y2：どの様な形式も文字列として記憶するために、この要請は満足されている。

Y3：繰り返し出現するデータは最初の入力時に値を入力し、その属性を暗黙指定すれば、以後の入力ではプロンプトは省略され、値は最初のものが保たれる。

Y4：現在は実現されていない。等しいものを対応させる辞書を付加させることを検討中である。

Y5：暗黙値を設定した後その属性を暗黙指定する。例外的な値が出現した場合はそのデータを一度スタックに積み、例外値を入力した後スタックから取り出し処理をする。

Y6：現在は実現されていない。Y4と同様、辞書を付加することを検討中。

Y7：システム起動時に集合入力を行なう属性を指定すると順序集合を扱う属性になる。要素間のデリミタは2つ以上の空白である。

Y8；Y9：文字列に特別の意味を持たせることで実現している。例えば@（単位記号）は正規化しない区切記号で

あり、／（スラント）は正規化する区切り記号である。

Y10：以下の様に定義されている。

?i　i個値が存在する。

?+　1個以上値が存在する。

?*　値は存在するが個数は不明。

?φ　値が存在しない。

?－　存在し得ない。

?　全く不明

Y11：検査用モジュールを付加することで可能である。

Y12：修正用モジュールを付加することで可能である。

Y13：文献IDを自動生成し検査する。現在、チェックルーチンを作成中である。

Y14：TSSエディタに匹敵するエディタを備えている。

Y15：Y11に同じ。

Y16：現在、バッチバージョンが用意されている。オンライン化が望まれる。

4. 使用例

プログラムは、富士通PL/Iで、READYモード576ステップ，EDITモード569ステップである。使用例は図1に示されている。図2はコマンドの一覧表であり、図3は、EXTERNAL変数の一覧表である。

文献

KAMB81 Y.Kambayashi (Ed.Ass. Konishi, Tanaka, Le Viet), "Database-A Bibliogrophy,"vol.1, Computer Science Press, 1981.

KAMBL8111 Y.Kambayashi, C.Le Viet, S.Tokuda and S.Yajima, "A Database Preparation System," 5-th IEEE COMPSAC, Nov. 1981.

KAMBL8201 Y.Kambayashi, C.Le Viet, S.Tokuba and S.Yajima, "SIFT-A Simple Form Translater for Bibliographic Data," 15-th Hawaii International Conference, vol.1, pp.112-121, Jan. 1982.

```
≠ edit vldb new
INPUT
 1 AUTHOR
* j.a.bubenko,jr
*
 2 TITLE       : data base design tools
 3 PUB. ABBR.  : vldb-4
 4 TYPE        : proceedings
 5 PUB. NAME   : the fourth international conference on very large data bases
 6 VOLUME      :
 7 NUMBER      :
 8 DATE        : sept. 13-15, 1978
 9 PAGE        : 2
10 COMMENT     :
E≠ append 1 /  s.b.yao/
E≠ change 1 /jr/jr./
E≠ list 1
 1 AUTHOR      : J.A.BUBENKO,JR.  S.B.YAO
E≠ set 3 4 5 6 7 8 10
E≠ save vldb78-1
SAVED IN ENTRY VLDB78-1
E≠ input
INPUT
 1 AUTHOR
* b.sundren
*
 2 TITLE       : data base design in theory and practice toward an integrated -
methodology
 9 PAGE        : 3-16
E≠ save vldb78-2
SAVED IN ENTRY VLDB78-2
E≠ input
INPUT
 1 AUTHOR
* j.m.smith
*
 2 TITLE       : comments on the paper "data base design in theory and -
practice" by bo sundren
 9 PAGE        : 17-18
E≠ save vldb78-3
SAVED IN ENTRY VLDB78-3
E≠ end
≠ listcat
IN CATALOG:MASTER
VLDB78-1    VLDB78-2    VLDB78-3
≠ list vldb78-2
 1 AUTHOR      : B.SUNDREN
 2 TITLE       : DATA BASE DESIGN IN THEORY AND PRACTICE TOWARD AN INTEGRATED METHODOLOGY
 3 PUB. ABBR.  : VLDB-4
 4 TYPE        : PROCEEDINGS
 5 PUB. NAME   : THE FOURTH INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES
 6 VOLUME      :
 7 NUMBER      :
 8 DATE        : SEPT. 13-15, 1978
 9 PAGE        : 3-16
10 COMMENT     :
≠
```

図1 使用例

(≠ and E≠ are prompting symbols)

1. CHGCS
2. DELETE
3. EDIT
   3. 1. APPEND
   3. 2. BOTTOM
   3. 3. CHANGE
   3. 4. DOWN
   3. 5. END
   3. 6. FIND
   3. 7. HELP
   3. 8. INPUT
   3. 9. LINEDIT
   3.10. LIST
   3.11. LOCATE
   3.12. POP
   3.13. PUSH
   3.14. RESET
   3.15. SAVE
   3.16. SET
   3.17. TOP
   3.18. UP
   3.19. VERIFY
4. HELP
5. LIST
6. LISTALC
7. LISTBC
8. LISTCAT
9. LISTD
10. LOGOFF
11. LOGON
12. SEND
13. STATUS
14. TIME

図2　コマンド一覧

SYSIN	システム入力ファイル
SYSPRINT	システム出力ファイル
MASTER	データ入出力ファイル
USERCAT	ユーザ登録ファイル
MESSAGE	メッセージ交換ファイル
HELPMSG	HELPメッセージファイル
COND	システムステイタス
USERID	ユーザID
ENTRYID	エントリ名
X	作業用文字列
XX	〃
EXTMSG	拡張メッセージ
COMBUF	コマンド列
OPEBUF	オペランド列
IN	入力行数
OUT	出力行数
LINE	属性行番号
ATTR	属性値
LABEL	属性名
DEFAULT	暗黙指定
INHIBIT	入力禁止指定
OPTCALL	外部呼出し指定
SETBIT	集合入力指定
ENDTIME	終了時刻

図3　　EXTERNAL変数