Seepによる科学ファクトデータベースシステム

5V－9

佐藤 誉夫　竹田 正幸　松尾 文碩
九州大学工学部

1. まえがき

Seepは，著者らが開発した，ルールベースとデータベースを統合管理するシステムである．Seepは，UNIXとMS-DOSのもとで動作する．九州大学大型計算機センターでは，演繹DBMSであるAdbis[1]を使って，結晶構造データベースシステムXDT[2]，遺伝子情報データベースシステムGENAS[3]のサービスを行なってきたが，これらをSeepで再構築する作業を行なった．Seepを使うことにより，ルールベースに基づく規則推論機構を用いたデータ解析が可能となる．

2. データベースシステムの構成

Seepを用いた科学ファクトデータベースシステムの構成を，XDTを例に説明する．図1にXDTの構成を示す．図のように，Seepはルールベース管理機構

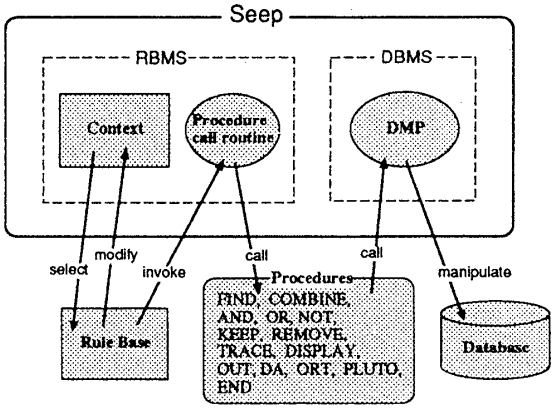

図1 XDTの構成

(RBMS)とデータベース管理機構(DBMS)から成っている．SeepのDBMSの外部へのインターフェースは，DMP (Data Manipulation Primitives) である．これは，find, insert, retrieve, createなどのデータベースの基本操作の集合であり，C言語で記述された手続きのライブラリである．このDMPを応用プログラムなどから呼び出すことにより，データベースを操作することができる．

通常のルールベースシステムと異なり，Seepではルールから手続きを呼び出せるようにしている．SeepのDBMSは固有の問合せ言語をもたない．そこで，FIND, DISPLAYなどの情報検索系のコマンドに対応する手続きを，DMPを用いて作成し，ルールから呼び出すことにした．また，応用プログラムに関しては，ケンブリッジ結晶データ付属のFORTRANで書かれた解析プログラム等をもとにORT, PLUTOなどの手続きを作成した．すなわち，情報検索系のコマンドと応用プログラムを区別しない．図2にXDTの検索例を示す．

```
kyu-cc%xdt
.find triazole
  1:     253 document(s) found
.find amino
  2:    7676 document(s) found
.combine 1*2
  3:      71 document(s) found
.display
  3: combine 1*2
      1/      71
MF = C9 H10 N6 O1 S1
CN = 4-AMINO-3-(BETA-BENZOYLHYDRAZINO)-5-MERCAPTO-1,2,4-TRIAZOLE
BI = R.C.SECCOMBE,C.H.L.KENNARD  : J.CHEM.SOC.,PERKIN T, VOL.    , PAGE
     1973.
+
      2/      71
MF = C24 H30 N6 O2                                          ...OLE-N$1|,N')-
CN = 5-AMINO-1-BUTYL-3-(N-BUTYL-N',N'-DIBENZOY
BI = G.RECK,M.CZUGLER,L.PARKANYI,E.SAUER
     PAGE  565, 1981.
+                                          ...2357, 1990.
      3/      71
MF = C10 H11 CL2 N6 1+ O
CN = 4-AMINO-3-(2,6-
     HYDROCHL
BI = P.-E.W                 AMINO-3-METHYL-1,2,4-TRIAZOLE-N$1|,N$2|)-BIS(MU|2$-CHLORO
+
          ...GOLI,M.LANFRANCHI,M.A.PELLINGHELLI  : J.CHEM.RES., VOL.  214, PAGE
     1712, 1990.
+
      71/      71
MF = C3 H6 N4 S1
CN = 4-AMINO-3-METHYL-1,4-DIHYDRO-1,2,4-TRIAZOLE-5-THIONE
BI = F.BIGOLI,M.LANFRANCHI,M.A.PELLINGHELLI  : J.CHEM.RES., VOL.  214, PAGE
     1712, 1990.
+
```

図2 XDTの検索例

3. Seepのデータベース

Seepのデータベースは，関係の集合である．関係の特別な形として関数を区別した．例えば，関係$ATMP$

$$ATMP \subseteq ID \times ATOM \times APN$$

は，三つの領域$ID, ATOM, APN$の直積の部分集合を意味する．これらの領域は関係データベースにおい

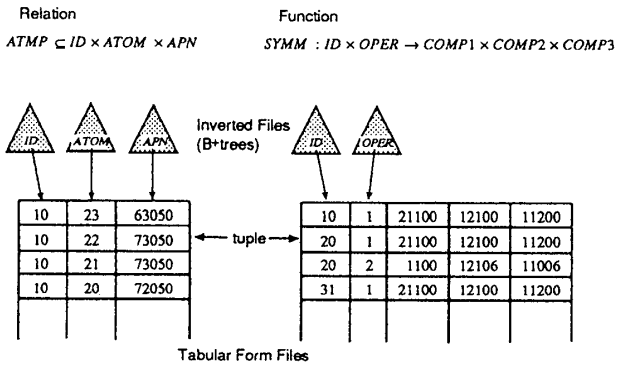

図 3 関係・関数のデータファイル構成

て属性と呼ばれる．一方，関数 $SYMM$

$$SYMM : ID \times OPER \rightarrow COMP1 \times COMP2 \times COMP3$$

は，$ID \times OPER$から$COMP1 \times COMP2 \times COMP3$への写像を意味する．関係および関数は，図 3に示すように，表形式ファイルと索引転置ファイルとして実現される．関係ではすべての領域について索引転置形を作成するが，関数では関数値の領域に関して索引転置形を作成しない．

関係データベースのすべての関係がひとつの属性を含むとき，その属性をデータベースキーとよぶ．XDT や GENAS などの科学ファクトデータベースでは観測ごとに付与した参照番号をデータベースキーとすることができる．Seep では，データベースキーをもつデータベースについては，キー集合を単位とした検索や集合演算などの操作を行なうことができる．

キーワード検索のためには，キーワード KWから関連するデータベースキー IDの集合を求める必要がある．これは，関係

$$KY \subseteq KW \times ID$$

によって実現できるが，Seep では，時間・領域量の効率化を図るため，

$$KY : KW \rightarrow 2^{ID}$$

という特別な関数を用いることができる．この関数は，図 3の表形式をもたない．

4. データベースの定義と構築

データベースを作成するためには，データ定義言語 (Data Definition Language; DDL) を用いて，データベースの定義をあたえなければならない．図 4に XDT のデータ定義の一部分を示す．図において，

```
DATABASE(XDT)
ADMINISTRATOR(g70073a)
MASTER_PASSWORD(*****)
DATABASE_KEY(ID)
  .........
DEF(ATMP < ID*ATOM*APN)
    L('ATOM PROPERTIES')
    PATH(/home/center/database/w_vfl/xdt/)
DEF(SYMM : ID*OPER -> COMP1*COMP2*COMP3)
    L('SYMMETRY POSITION')
    PATH(/home/center/database/w_vfl/xdt/)
  .........
DEF(ID)   T(L4) L('ID. CODE')
DEF(ATOM) T(I2) L('NUMBER OF ATOM')
DEF(APN)  T(I4) L('10000*ELEM+1000*NCA+100*NH+NCH+50')
DEF(OPER) T(I2) L('NTH SYMMETRY OPERATOR')
  .........
```

図 4 XDT のデータ定義

```
DATABASE_KEY(ID)
```

は，領域IDをデータベースキーとすることを意味する．また，

```
DEF(ATMP < ID*ATOM*APN)
DEF(SYMM : ID*OPER -> COMP1*COMP2*COMP3)
```

はそれぞれ，関係 $ATMP$, 関数 $SYMM$を定義している．

図 4の形式によるデータ定義のファイルをもとに，DMP の define によってデータベースの定義が行なわれる．各関係・関数は，それぞれの入力形式データファイルから DMP の create によって創成が行なわれる．

5. むすび

Seep を用いた科学ファクトデータベースの構築について述べた．既存のデータ解析プログラム等は，ルールベースにより効果的にデータベースの応用プログラムとして統合することができる．今後は，ルールベースによる規則推論機構を用いた遺伝子情報の解析等を行なうことにより，Seep の規則推論機構の実用環境下での有効性を検証したいと考えている．

参考文献

1) 松尾，二村，高木：データベース統合支援システム Adbis(1)–Adbis の概要と DMP の仕様–, 九州大学大型計算機センター広報 **16**, 6, 172–217 (1981).

2) 河野，高木，松尾，二村，鬼塚：結晶構造データベースシステム XDT の使用法, 九州大学大型計算機センター広報 **16**, 1, 32–53 (1983).

3) 久原，榊，高木，松尾，二村，鬼塚：核酸塩基配列データベースシステム GENAS の使用法 (1), 九州大学大型計算機センター広報 **16**, 5, 497–521 (1983).