

DBMS上のスキーマ診断システム GSC

黒瀬博靖, 安次富伸浩

(株) リコー

ソフトウェア研究所

データベース管理システムG-BASEの核はDM-LispなるLispインタプリタである。本論文では、このDM-Lisp上に構築したルール駆動型プロダクションシステムGSC (G-BASE Schema Consultant) の概要を述べる。GSCはユーザが定義したデータベーススキーマを評価し、それを改善するための助言をおこなうことができる。この目的のため、GSCには、G-BASEマニュアルの内容に相当するスキーマ設計のための知識を搭載した。GSCはDBMSの核を兼ねるLispの上で動作しているため、データベースを参照するプロダクションルールを扱うことが容易である。

GSC - A Schema Consultant on a DBMS

Hiroyasu KUROSE, Nobuhiro AJITOMI

Software Research Center
RICOH Co., Ltd

1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112, Japan

G-BASE is a DBMS whose kernel is a Lisp interpreter named DM-Lisp. This paper describes a rule-based production system GSC (G-Base schema Consultant) implemented on DM-Lisp. GSC can examine a database schema made by a user and suggest the user how to refine it. GSC contains knowledge for schema design written in one of G-BASE manuals. Since GSC is implemented on DM-Lisp, a DBMS kernel, it is easy to handle production rules that access to databases.

1. はじめに

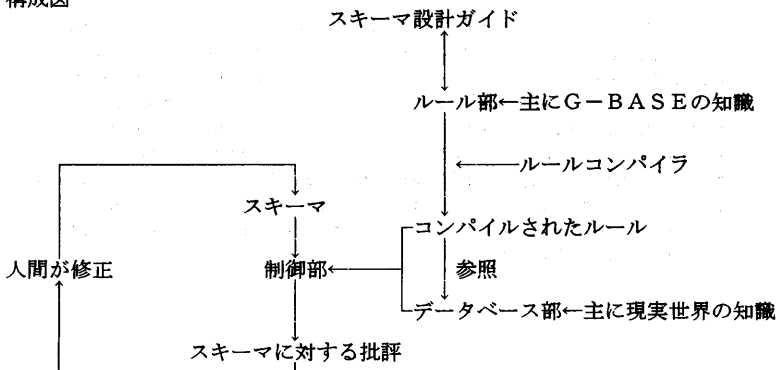
当リコー・ソフトウェア研究所で開発したG-BASE [文献1] は、関係データモデル [文献2] を拡張したグラフデータモデル [文献3] に基づく独特のDBMSである。このG-BASEのカーネルはDM-Lisp (Data Manager Lisp) と呼ばれるLispのインタプリタになっている。DM-LispはCommon Lispに似た言語仕様を持ち、データベースにアクセスする機能を備えたLispシステムである。G-BASEの一部 (検索時の最適化など) はDM-Lispで記述されている。

我々は、このDM-Lisp上に、G-BASE内のデータベーススキーマを診断・評価するためのエキスパートシステム、GSC (G-Base Schema Consultant) を構築した。GSCは与えられたデータベーススキーマを調べ、そのスキーマの問題点を指摘することができる。

DBMS分野におけるAI技術の応用は、自然言語による問い合わせに関するものが中心であるが、ユーザのデータベース設計を支援するシステムもいくつか提案されている。たとえば [文献4] では、スキーマの論理的構造を対話的に作成・改善する方法を論じている。 [文献5] や [文献6] では予想される問い合わせからこの種の構造を自動的に生成しようとしている。また [文献7] のように、DBMS専門家の行動をシミュレートする知的システムを構築しようとする試みもある。より実際的な効果をねらったものとしては、すでに存在するスキーマの評価と改善提案をおこなうシステムがあり、この例としてはCODASYL型DBMSを対象とした [文献8] があげられる。

我々のGSCは [文献8] と同様のアプローチをとっている。GSCの特徴は、第一に (前述のようにDBMSを兼ねるLispシステムの上に実現されたため) DBMSとの結合が自然な形でなされていること、第二にユーザマニュアルと並行に開発されたマニュアルと同等の知識 (その中にはグラフデータモデルに特有の知識も含まれている) を搭載していることである。本稿では、このGSCの概要について報告する。

構成図



2. 開発の意図

我々がGSCの開発によって意図したのは、第一にデータベース設計者の支援である。データベースの設計、特にスキーマの作成は、DBMSを導入し利用を開始する際に避けては通れないことだが、この作業は適用業務とDBMSの双方に関する詳細な知識を必要とする。GSCの目標は、この種の作業を助け、かかる労力をいくらかでも軽減することである。

特にG-BASEの場合は、グラフデータモデルという他のDBMSには見られないデータモデルを採用していること (具体的には「リンク型」という独特のデータ型を持つ)、便利ではあるが濫用するとシステムに非常な高負荷をあたえる機能が備わっていること (例えば、フィールドひとつの中に数十Kbyteにもおよぶデータを格納できる、等々) などの理由により、計算機資源を効率よく使うスキーマの作成には、DBMS一般に関する知識のほかにG-BASEシステム固有の知識が必須である。我々は、GSCにこのようなG-BASE独特の知識を搭載し、データベース設計者のG-BASEに関する知識・経験不足をできるだけ補うことを試みた。

GSC作成のもうひとつの狙いは、「マニュアルレス」システムの実験である。前述のように、G-BASEのスキーマ設計にあたって知っておくべきことは多いので、G-BASEのマニュアルの一部としてデータベース設計者のために「G-BASEスキーマ設計ガイド」 [文献11] が用意されている。これは確かにユーザのための配慮ではある。しかし、読むべき文書が一冊増えているわけであり、結局ユーザの負担を増やしていることになる。もっとも望ましいのは、マニュアルを読まずとも使うことができる、いわゆるマニュアルレス・システムであろう。我々は、マニュアルの内容を知識ベース化することにより、G-BASEをマニュアルレス化する実験をおこなった。GSCはその実験の成果であり、このシステムの中には前記「G-BASEスキーマ設計ガイド」の内容に相当する知識が含まれている。

なお、GSCの作成には、GSCと並行して（やはりDM-Lisp上に）開発された GIE (G-BASE Inference Engine) システムを利用した。GIEは、if-then型のプロダクションルールを実際に前向き推論をおこなうための (DM-Lispの) プログラムに変換する、RETEアルゴリズム [文献9] に基づくルールコンパイラである。GIEが用いているRETEアルゴリズムには若干の拡張が加えられており、ルールのif部に限量記号 (全称∀及び存在∃) つきの条件式を書くことが可能になっている。

3. システム構成及び入出力

GSCはルール部と制御部、そしてデータベース部から構成されている。

まずルール部は、G-BASEスキーマ設計ガイドの内容をif-thenルールに直したものである。この部分はルールコンパイラGIEのためのルール記述言語 (LispのS式に近い形式) で書かれている。なお、GIE自身はDM-Lispのプログラムである。

データベース部はアプリケーション (G-BASEを適用しようとしている業務) に関する知識を格納したデータベースを参照する部分であり、DM-Lispで書かれている。いくつかのルールのif部ではこのデータベース部を参照している。

制御部はGSCのメインルーチンにあたる。ルール部をGIEに与えると、推論のための一連のDM-Lisp関数が得られるが、これらの関数を呼び出してスキーマ情報を与え、スキーマに関する批評を出力するのが制御部である。この部分はDM-Lispで書かれている。

GSCの通常の使い方は次のようになる。まずデータベース設計者がG-BASEのデータ定義言語 [文献12] を用いてスキーマを作成する。また、業務に関する知識を「カテゴリ」と呼ばれるかたちでGSCのためのデータベースに入力しておく必要がある。(この情報はGSCデータベース部が参照する。) この状態でGSCを動かすと批評メッセージが得られる。この批評にしたがってスキーマを修正した後、再度GSCを起動する。一般には、満足できるスキーマが得られるまで、この「修正→GSC起動」をくりかえすことになる。

以下に続く各節では、制御部、ルール部、データベース部のそれぞれについて説明する。

4. 制御部

スキーマ評価のためのif-then形式のルールをルール・コンパイラGIEに与えると、OPS5 [文献9, 10] と同様の前向き推論を行なうための一連のDM-Lisp関数が得られる。このGIEが生成したDM-Lispコード (この部分は現在約300個のDM-Lisp関数から成っている) にスキーマ情報を与え、実際に推論をおこなうのがGSCの制御部である。制御部は比較的単純な部分で、実際にGSCの振舞いを決めるのは制御部自身ではなくGIEが生成したコードの方である。そこで、以下ではGIEについて論ずることとする。

GIEで用いているRETEアルゴリズムには、条件部に通常の述語以外に全称記号 (∀) や存在記号 (∃) を書けるように拡張を施してある。 [文献14] この拡張のおかげで、ルール作成者が一階の述語論理で書ける条件をより自然に記述できるようになり、G-BASEスキーマ設計ガイドの内容をルール化する作業が大変楽になった。

例1

```
(r r1 r2 f1 f2)
(
  (sco-is-record r) (sco-is-record r1) (sco-is-record r2)
  (equal (sco-no-of-field r) 2)
  (not (eq r r1)) (not (eq r r2))
  (sco-is-field f1) (sco-is-field f2)
  (sco-field-in-record f1 r) (sco-field-in-record f2 r)
  (not (eq f1 f2))
  (equal (sco-subcat f1) (sco-subcat (sco-key-of r1)))
  (equal (sco-subcat f2) (sco-subcat (sco-key-of r2))))
(
  (sco-mesg1 "レコード" r "は" r1 "と" r2 "の間のリアルリンクにしましょう。")
  (sco-mesg2 "レコード" r "に属するフィールドは" r1 "と" r2 "の主キーと"
    "同一カテゴリに属していますから、このレコードは" r1 "と" r2
    "の関係を表現する為に作られたものだとして推定されます。"
    "そのような場合はリアルリンクを使用をお勧めします。")
)
```

例1にルールの例を示す。ここで、最初のリストがこのルールの中に現れる自由変数の宣言、次のリストが条件 (if 部)、最後のリストが条件が成立したときに実行されるべきアクション (then 部) をそれぞれ表わしている。例1の意味するところは「フィールド二つから成るレコードでその各々が他のレコードの主キーになっている場合、そのレコードにはいる情報は実リンクで実現すべきである」ということで、これは実リンクの使用法に関するもっとも基本的な戦略である。

ルールの if 部における原子論理式は、GSC 制御部実行の際に、そのまま DM-Lisp インタプリタによって評価される。したがって、DM-Lisp の関数で記述できることはすべてルールの if 部に書くことができる。DM-Lisp にはデータベース (スキーマ情報が格納されているデータ辞書を含む) にアクセスするための関数がいくつか用意されているが、それらを自由にルールの if 部で呼び出せることになる。この機能は、今回のシステムのようにデータベースを参照するプロダクションルールを多用する場合には、大変有用である。また、論理式で書くとは煩雑すぎるような条件については、適当な (その条件を評価するための) DM-Lisp 関数を定義し、ルールの if 部からその関数を呼び出すようなことが可能であり、実際 GSC ではこのような使い方をしている。

GIE はルールをコンパイルする時に、OPS 5 同様、if 部を評価するプログラムの共有化を図っている。すなわち、いくつかのルールの if 部が (あるいはひとつのルールの if 部中であっても) 共通な部分式を持つときには、その部分式を評価するコードを二重に生成しないようにしている。そのほか、最適化のための工夫をいくつかしているが、生成される DM-Lisp プログラムの効率性は、現状では、残念ながらあまりよいものとは言えず、なお改善の余地が残っている。GIE の改良については現在検討中であるが、通常最適化のほかに、個々の原子論理式の成立しやすさや、評価のためのコストを考慮に入れて論理式の評価の順序を決定することも考えている。〔文献 15〕

例2
(r f)

```
(
  (sco-is-record r) (sco-is-field f)
  (sco-field-in-record f r)
  (sco-virtuallinked f) (not (sco-indexed f)))
)
(sco-mesg1 "レコード" r "のフィールド" f "には索引を付けましょう。")
(sco-mesg2 "仮想リンクの両端のフィールドには索引を付けたほうがよいのです。")
```

最後に、ルール間の競合解消の方法について述べておく。同時にふたつ以上のルールの if 部が成立した場合、その中から実行すべきルールをひとつ選ぶために競合解消の戦略が必要となる。この戦略に関してはLEXやMEA〔文献10〕などがあるが、GIEでは、各ルールにあらかじめ優先度を与えておき、優先度が高いほうを採用するという単純な方法をとった。具体的には、優先度は自然数で与える (より小さい数がより高い優先度を表わす) 仕様になっ

ている。このような方法をとった場合、ルール作成者が各ルールの優先度を定めなくてはならず、特にルールが多いときにはルール作成作業時の手間が大幅が増えてしまうおそれがある。幸い、GSCのルールセットは小規模なものであったため、あまり問題は生じていない。むしろ、ルール作成者がGSCの動作を細部まで制御できるという効果のほうが大きかったようである。

5. ルール部

現在ルール部は約30個のルールから成っている。このルールには「G-BASEスキーマ設計ガイド」〔文献11〕に書かれている内容が記述されている。このルールによって出力される批評は以下のように分類される。

- 1 データベースの正規性に関する勧告
- 2 フィールドに索引をつけることに対する勧告
- 3 フィールドの型に関する勧告
- 4 実リンクに関する勧告
- 5 仮想リンクに関する勧告
- 6 レコード型の重複に関する勧告
- 7 レコード型の主キーに関する勧告

実リンクと仮想リンクはG-BASEに固有のものである。リンクはレコード型間における二項関係である。リンクの概念及び使用法については〔文献1, 3〕に詳述されている。

G-BASEに特有なルールとしては例2がある。これは仮想リンクの両端のフィールドには索引をつけるべきであるというルールを表現している。また例3のようにDBMS一般について適用出来るルールもある。これはあるレコードの全てのフィールドが無制限可変長の場合にメッセージを出力する。全てが無制限というのは設計者の怠慢を意味する場合が多いからである。例3の4行めに出現する“all”は全称記号(∀)を表現している。この条件を論理式で書くと次の形になる。

```
sco-is-record(r) ∧
((∀ f)
  ((sco-is-field(f) ∧
    sco-field-in-record(f,r))
   ⇒ (sco-length-type-of(f) = UNLIMITED)))
```

sco-is-record(r): r がレコードであるという述語
sco-is-field(f): f がフィールドであるという述語
sco-field-in-record(f,r): f が r のフィールドの一つであるという述語
sco-length-type-of(f): f の長さのタイプ属性を得るための関数
UNLIMITED: 無制限可変長を表わす定数

例3

```
(r)
(
  (sco-is-record r)
  (all (f) (imply
    (and (sco-is-field f)
          (sco-field-in-record f r))
    (eql (sco-length-type-of f) UNLIMITED))))
(
  (sco-mesg1 "レコード" r
    "のフィールド全てが無制限可変長というはおかしきありませんか。")
  (sco-mesg2 "フィールドの長さをもう少し真面目に管理しましょう。"))
```

例4

```
(r f)
(
  (sco-is-record r) (sco-is-field f)
  (sco-field-in-record f r)
  (sco-search (sco-subcat f)) (sco-select (sco-subcat f))
  (equal (sco-length-type-of f) UNLIMITED))
(
  (sco-mesg1 "レコード" r "のフィールド" f
    "は長さを制限して索引を付けましょう。")
  (sco-mesg2 "このフィールドは選択性が良いえに検索に使われる可能性が高いため、"
    "索引を付けることが出来れば検索の効率が上がると思われます。"))
```

データベース部にある情報を用いて起動されるルールもある。データベース部には使用されること多いフィールド(人名、住所、など)の情報が入っている。データベース内の情報はカテゴリと呼ばれており、カテゴリ名という名前前で参照される。カテゴリ名はフィールドの一般的な意味や属性を指示するために用いられる。GSCを用いるユーザはスキーマの各フィールドに対しカテゴリ名を一つ与える。ユーザはフィールドの現実との対応をある程度は知っているため、「personレコードのnameというフィールドは人の名前である」というようなことを指示できる。GSCはこの指示を用いてデータベース部にアクセスし、現実世界の情報を引き出す。データベース内の情報によって、たとえば長さは2文字以上で10文字以下であり、文字型で検索の対象となりやすいということがわかる。この情報を用いてルールが選択され、このフィールドに対する批判が出力される。またこのカテゴリからの情報によってデータ独立性もチェックする。カテゴリを用いたルールとしては例4がある。フィールドの属するカテゴリに選択性と検索性が書かれている。G-BASEでは無制限可変長のフィールドには索引をつけることが出来ない。例4は、選択性と検索性の高いフィールドは長さを制限して索引を付けるべきである、というルールを表現している。

6. データベース部

ここはカテゴリ毎の属性が表となって入る。データベースのスキーマの一般形ははいつているとみなして良い。各々の応用分野毎のスキーマの形態の差はこの知識の変更で吸収できる。ユーザが新しいカテゴリを導入することによって知識を増やすことができる。現在フィールドの属性として、ベースタイプ、長さの上下限、選択性、一意性、検索の条件のなりやすさ、等が設定されている。

この属性のうち、長さの上限などは本来あいまいなものであるが、現在はしきい値を決めてそれより長いものをチェックしている。このしきい値はアプリケーションの種類によって異なってくるが予想される。またしきい値そのものをファジー化することも検討中である。

カテゴリの例として人名のカテゴリを挙げる。この情報を用いて人名のフィールドに関する索引の付加や一意性のチェックを行う。

カテゴリ名	humannam e
ベースタイプ	文字型
長さ	2文字以上10文字以下
一意性	なし
選択性	高い
検索の頻度	高い
親レコード	人間に関するレコード
キーへの成り易さ	非常に成り易い

7. マニュアル

「スキーマ設計ガイド」〔文献11〕を書くことによってルールの整理も同時に行なった。このガイドはルール化を意識して書かれており、構成的にスキーマを作成する部分と、それをより良くするためのアドバイスの部分から成っている。文法やスキーマ作成の実際の作業は「応用プログラマガイド」〔文献13〕に記述されている。設計者は〔文献13〕を読んだ後に〔文献11〕を読めばより良い設計が行なえるようになっていく。

〔文献11〕作成の上で留意した点は、ルールを作りやすいように内容を書くということである。マニュアルのアウトラインの作成法はいろいろあるが〔文献16, 17, 18〕, このガイドはルールの種類ごとに章立てを行い、一つのルールに一節を費やした。作成したスキーマをチェックしたい人はGSCをそのまま使用すればよく、G-BASEのスキーマ設計全般にかかわる知識を得たい人はマニュアルを読めばよい。GSCとマニュアルは相補的な関係にある。

GSC作成はマニュアルレスの実験としては半分成功といえる。確かにマニュアルの内容は項目毎に全てルールにすることが出来た。しかしエキスパートシステムだけでは、ルールを学習することには困難が伴う。マニュアルを完全に廃止するためには、スキーマ作成を学習するためのシステムが必要と思われる。

8. 今後の展開

本節では、GSC/GIEの今後の課題について論ずる。まず、ルールコンパイラGIEについてはルール記述法の強化やそれに伴う推論アルゴリズムの改良が必要である。

- ・現在のところ、if部に書くことができるのは、一階述語論理における式の一部にすぎない。if部の条件式として、より広範囲の論理式を扱えるようにすることが望ましい。

- ・変数のデータ型という概念を導入すると、ルールの読みやすさ、生成コードの実行効率ともに向上する可能性がある。どのような形で導入するかを現在検討している。

- ・またルールの二項関係としてルールの優先順位を規定した場合、その二項関係が常に線型順序になるのかどうかというのも問題である。順序関係にならない場合は現行の競合解消ルールを考え直す必要がある。

次にGSCの今後の課題をあげる。現在のところ、GSCルール部のルールはすべてif部でデータ辞書内のスキーマ情報と「カテゴリ」データベース内のアプリケーション情報を参照し、then部で評価メッセージを生成するという形をしている。しかしこれではG-BASEの機能を全て利用しているとは言えない。

- ・G-BASEはデータ自動再構成機能を持っているので、この機能をルールのthen部で呼び出すことができれば自動的にデータベース再構築を行なえる可能性がある。この発展としてスキーマ自動修正システムが考えられる。

- ・カテゴリを与えなかった場合にもなんらかの形でアプリケーションに関する知識を獲得ないしは推論し、動作できることが望ましい。現在のGSCはカテゴリ情報なしでも動作はするが、その場合にはデータの内容にまで言及したメッセージ（たとえば「xは人名ですからy文字をこえることはめったにないはずです」等々）は一切出力できない。

9. おわりに

今回の成果を総括する。我々はDBMSを兼ねるLispシステムDM-Lisp上にデータベース設計を支援するエキスパートシステムGSCを作成し、試用した。その際マニュアル「G-BASEスキーマ設計ガイド」執筆とGSC開発を連係して進め、互いに同等のマニュアルと知識ベースからなるユーザ支援体系を構築することができた。また、GSC作成のためにルールコンパイラGIEをDM-Lisp上に開発したが、その過程でRETEアルゴリズムを拡張し、限量記号を持つ論理式を扱うことに成功した。

参考文献

- 1 “特集/スーパーリレーショナル・データベース・システム”, コンピュトロール17号, コロナ社, 1987
- 2 E. F. Codd, “A Relational Model of Data for Large Shared Data Banks”, Communications of ACM, Vol.13, No.6, 1970
- 3 H. S. Kunii, “Graph Data Language: A High Level Access-Path-Oriented Language”, Ph.D Dissertation, Univ. of Texas, 1983
- 4 上林, 古川, 国分, 黄, “モデル間の変換によるデータベース設計システム”, 情処学会第34回全国大会, 1987
- 5 S. Lanka, “Automatically Inferring Database Schema”, Proc Int Jt Conf Artif Intell, Vol.9, 1985
- 6 溝口, 磯本, 角所, “データベース論理設計支援エキスパートシステム”, ICOT研究論文1984
- 7 川口, 溝口, 山口, 角所, “データベースの論理設計を支援する知的インタビュースystem”, 情処学会知識工学と人工知能研究会, 1986
- 8 馬場, 中野, 池田, “データベース性能診断システム”, 情処学会データベースシステム研究会, 1985
- 9 C.L.Forgy, "Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem", Artificial Intelligence, Vol.19, 1982
- 10 L. Bronston, R. Farrell, E. Kant, N. Martion, "Programming Expert Systems in OPS5", 1985
- 11 “G-BASE スキーマ設計ガイド”, リコー, 1987
- 12 “G-BASE DDL言語仕様書”, リコー, 1987
- 13 “G-BASE 応用プログラマガイド”, リコー, 1987
- 14 安次富他, “DBMSを兼ねるLISPシステム上のエキスパートシステムの構成”, 情報処理学会第35回全国大会, 7Cc-10, 1987
- 15 黒瀬他, “DM-Lispを用いた推論システム(1) — 知識表現 —”, 情報処理学会第36回全国大会, 3R-9, 1988
- 16 海保, 加藤, 堀, 原田, “ユーザ・読み手の心をつかむマニュアルの書き方” 共立出版, 1987
- 17 Brad M. McGehee, The Complete Guide to Writing Software User Manuals, 1984
- 18 高橋, “わかりやすいマニュアルの作成法”, 日経マグロヒル, 1985