情 報 学 基 礎 21－5
（１９９１．５．２８）

自然語意味情報を対象にした格構造データベース

*柴崎 真人 *小口 琢夫 **木内 伊都子
* ㈱日立製作所 システム開発研究所
**㈱日立製作所 中央研究所

文書検索システム等の情報検索システムにおいて、蓄積される情報の内容や検索要求の表現にはキーワードのみでは不十分であり、より高度な情報表現が必要である。そこで、文書内容や関連知識などの自然語意味情報を、多ソート論理式で意味付けした格構造モデルで保持する演繹データベースを開発中である。本システムは、対象とする知識を自然語意味情報に限定し、そのための表現機能として、多ソート論理式のルール節の拡張にあたる包含連体修飾というデータ形式と、ソートの集合演算にあたる合成ソートを提供している。

A Case Structured Database for Natural Language Information

*Masato Shibasaki, *Takuo Koguchi, **Itsuko Kiuchi
*Systems Development Laboratory, Hitachi, Ltd.
**Central Research Laboratory, Hitachi, Ltd.
*1099 Ohzenji Asao-ku Kawasaki-shi, 215 JAPAN
**1-280 Higashikoigakubo Kokubunji-shi, 185 JAPAN

In order to represent the contents of documents or retrieval requests for information retrieval systems such as document retrieval systems, it has been recognized that conventional keyword systems are insufficient and a more sophisticated method is required. A deductive database system ,which is based on many sorted logic and case structures of natural language sentenses representing the contents of documents and retrieval requests, is being developed. This system features its representation capability of extended rule clauses and compound sorts.

1．はじめに

現在、文書検索システムの文書内容に関する検索方式としては、商用のものも、研究レベルでも、キーワード方式が中心である[1]。それに対し、キーワードのみでは、蓄積された文書の内容も、ユーザの検索要求も十分に表現できないということで、文書の内容と質問文を意味構造モデルで表現し、両者のマッチングをとることにより、質問文に適した文書を検索する方式の研究もすすめられている[2][3][4]。しかし、これらは、明確に意味付けされた意味構造モデルをもつものではない。我々は、文書内容・関連知識などの自然語意味情報を、多ソート論理式で意味付けした格構造モデルとして保持し、ロジックベースの健全で完全な検索処理を行う「知的検索システム」を現在開発中である。本システムにより、次のようなメリットが得られるものと考えている。

(1)文書内容や書誌事項に関する知識は、検索対象とするだけでなく、検索時に利用することもできる。

(2)検索文内の概念をより詳細な情報に置き換えることによって、検索結果を提供できるので、リファレンスＤＢとしてだけでなく、ファクトＤＢとしての機能も果たす。

本稿では、本システムのデータベース部分を格構造データベース(Case Structured Data Base)とよび、これのデータベースモデルを中心に報告する。ビジュアルインタフェース部分については [5] を参照のこと。

2．基本的考え

インスタンス(個別の事物)のみでなく、クラス(個別の事物を包含する一般的な概念)も記述の対象として含むことにより、継承階層の表現・検索を効率的に実現しようとする拡張Prologがいくつか提案されている(以後、継承階層Prologとよぶ)[6][7][8][9]。格構造データベースのデータベースモデルおよび検索処理の単一化手順はこれらの考え方を多分に継承しているが、次の点が大きく異なる。

(1)各述語の各々の引数が自然語のいずれの格に相当するかを引数の位置によってではなく、格と引数の対として記述することにより示す。

(2)述語にも階層構造をもたせ、また、述語の引数として時制という要素ももたせる。

(3)接合連体・包含連体修飾というデータ形式を提供する。

(4)格の中身は単一のソートまたは定数ばかりではなく、それを集合演算した合成ソートの記述も許す。

(5)変数同志の単一化の際に、変数を型付けしている合成ソートが互いに包含関係にはないが、共通部分集合をもつとき、単一化代入として、変数を型付けしている合成ソート間のglb(greatest lower bound)を直接求めるのではなく、単に、合成ソートを "∩" でつなげる。

(6)ソートと定数の階層構造は、包摂(is_a)関係のみではなく、部分全体(part_of)関係も含み、検索の際には、部分全体関係も利用する。

(7)単位を伴う単一の数値または１組の数値も定数・ソートと同様に扱う。

(8)定数・ソート・述語は内部的にはＩＤで管理する。これにより、同義語のカテゴライズや同綴異義語の区別が可能である。

(1)・(2)は本章で、(3)は次章で、(4)・(5)は４章で、(6)・(7)・(8)については５章で説明する。

なお、本稿では、分りやすさを主眼とすることにし、正確な定義は省略し、具体例を中心に説明する。

2．1　概念階層知識

格構造データベースが扱う知識には、概念階層知識と文知識とテンプレートとがある。

概念階層知識は、定数･ソートと述語の階層構造を示す。図１に例を示す。ソートには "@"をつけている。我々は、現実的な観点から、多重継承を許すことにし、また、階層構造が束(lattice)であることを仮定しない。以降の例では図１の概念階層を仮定する。

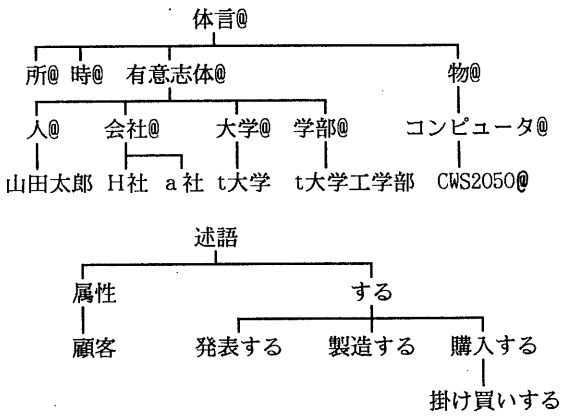

図１ 概念階層の例

表１．　ＣＳＤＢの格ラベル一覧

略記コード	英語名	日本語名
A	Agent	動作主（何が）
O	Object	対象（何を）
R	Recipient	受益者（何に）
S	Start	起点（何処から）
D	Destination	方向（何処へ）
P	Place	所（何処で）
T	Time	時（いつ）
C	Cause	原因（なぜ）
I	Instrument	手段（何で）
U	Purpose	目的（何のために）
M	Measure	数量（いくつ/どの位）
E	Entity	属性対象（～の…は）
V	Value	属性値（である）

2．2　文知識

文知識は、単文･言い換え規則･接合連体修飾･包含連体修飾からなる。本節では、単文と言い換え規則について説明する。接合連体修飾･包含連体修飾については次章で述べる。単文は継承階層Prologのファクト節に、言い換え規則はルール節に対応する。

「ａ社はあるコンピュータをすべての在京新聞で発表した」という内容を考える。これは、多ソート論理式では、

∃Xc／コンピュータ，∀Xn／在京新聞　発表した（ａ社,Xc,Xn）

と表せる。このとき、コンピュータと在京新聞はソートである。この式をスコーレム化すると、

∀Xn／在京新聞　発表した（ａ社,コンピュータ＃１,Xn）

となる。このように意味づけされたものを格構造データベースでは、

＜発表する，過去，ａ社／Ａ，コンピュータ＃１／Ｏ，在京新聞@／Ｉ＞

という形式で表現する。このタイプの知識を単文という。コンピュータ＃１は、システムが内部的につくりだしたコンピュータのソートに属するスコーレム定数であり、ビジュアルインタフェースを介する

ことにより、ユーザからは「あるコンピュータ」としか見えない。一番左が述語、その次が時制である。時制は、現在･過去･未来のうちのいずれかである。そのあとは、格とその中身との対応を示している。格構造データベースでは、表１に示す１３種類の格を用意しており、このうちの使用する格に対してのみ、対応する中身を記述する。格の出現順序に意味はない。

本来、

∀Xn／在京新聞，∃Xc／コンピュータ　発表した（a社,Xc,Xn)

（「a社はすべての在京新聞で何らかのコンピュータを発表した」）

という多ソート論理式も表現できるべきだが、（１）ユーザに限量子の依存関係を指摘させるのは困難、（２）スコーレム関数の導入に伴う出現試験(occur check)の計算の手間を省く　という理由から現バージョンでは、前置する限量子が∃…∃∀…∀という順序の多ソート論理式のみを対象にしている。

次に、「ある会社が製造した物を購入した人を、その会社の顧客という」という内容を考える。これは多ソート論理式では、

∀Xc／会社,∀Xh／人,∀Xt／物

製造した(Xc,Xt)∧購入した(Xh,Xt)→顧客(Xc,Xh)

と表せる。このように意味づけされたものを格構造データベースでは、

＜顧客，現在，会社@１／E，人@２／V＞

← ＜製造する，過去，会社@１／A，物@３／O＞

＜購入する，過去，人@２／A，物@３／O＞　　S１

という形式で表現する。このタイプの知識を言い換え規則という。会社@１の@１は、２つの会社@１が同一事物であることを意味するものであり、システムが自動的に番号付けして付与する。

２．３　テンプレート

テンプレートは各々の述語に対して、とりえる格とその型付けを示す。これは下位の述語にも継承される。例えば、購入するという述語のテンプレートは、

＜購入する，有意志体／A，物／O，所／P，時／T，物／U，＞

のように表現されている。システムは、テンプレートに適合しない文知識や質問文はうけつけない。

２．４　質問文

質問文は継承階層Prologのゴール節に相当する。「H社の顧客である人は？」という質問を考える。「H社の顧客は人である」という内容は、多ソート論理式では、

∃Yh／人　顧客(H社,Yh)

と表わされ、これの否定は、

∀Yh／人　～顧客(H社,Yh)

である。格構造データベースでは、先の質問文を、

←＜顧客，現在，H社／E，人@／V＞

という形式で表現する。

２．５　単一化

以後はＰｒｏｌｏｇの用語に準じて、単一化方法を説明する。選択式と入力節とが単一化に成功する基準は以下のすべての条件を満足するときのみである。

(1)選択式の述語が入力節のヘッド部の述語と同じか、より上位である。

(2)選択式の時制と入力節のヘッド部の時制とが等しい。

(3)選択式に記されているすべての格に対して、次の条件を満足する。

選択式の格と同一の格が入力節のヘッド部に存在し、ソートをそのソートに属する定数の集合、定数を単位集合とみなしたとき、両者の格に対応する定数またはソートが包摂関係にあるか、共通部分集合をもつ。

言い換え規則Ｓ１と、Ｓ２･Ｓ３の単文があるとき、「Ｈ社の顧客は？」という質問に対する検索処理の一例を示す。

＜製造する，過去，Ｈ社／Ａ，CWS2050@／Ｏ＞　　Ｓ２

（「Ｈ社はすべてのCWS2050を製造した」）

＜掛け買いする，過去，山田太郎／Ａ，CWS2050＃１／Ｏ＞　　Ｓ３

（「山田太郎はあるCWS2050を掛け買いした」）

最初のゴール節は、

←　＜顧客，現在，Ｈ社／Ｅ，人@／Ｖ＞　　Ｇ１

である。このゴール節はまず、Ｓ１を入力節として、代入を｛会社@１→Ｈ社／Ｅ，人@→人@２／Ｖ｝とすることにより、単一化に成功し、新たなゴール節

←　＜製造する，過去，Ｈ社／Ａ，物@３／Ｏ＞　　Ｇ２

＜購入する，過去，人@２／Ａ，物@３／Ｏ＞　　Ｇ３

を得る。次に、Ｇ２を選択式とするとき、Ｓ２を入力節として、代入を｛物@３→CWS2050@／Ｏ｝とすることにより、単一化に成功し、新たなゴール節

←　＜購入する，過去，人@２／Ａ，CWS2050@／Ｏ＞　　Ｇ４

を得る。最後に、選択式Ｇ４に対して、Ｓ１を入力節として、代入を｛人@２→山田太郎／Ａ｝とすることにより、単一化に成功し、空節が導かれ、Ｇ１の人@／Ｖに対する合成代入｛人@→山田太郎／Ｖ｝が解となる。

３．連体修飾

我々は、自然語意味情報をより直接的に表現するために、接合連体修飾と包含連体修飾という２種類のデータ形式を提供する。

３．１　接合連体修飾

「ａ社が製造したコンピュータをｔ大学が購入した」という内容を考える。これは、「ａ社が製造した」コンピュータの集合と、「ｔ大学が購入した」コンピュータの集合との間に共通部分があることを意味する。多ソート論理式では、

∃Xc／コンピュータ 製造した(ａ社,Xc) ∧ 購入した(ｔ大学,Xc)

と表わされ、スコーレム化すると、

製造した(ａ社,コンピュータ＃１) ∧ 購入した(ｔ大学,コンピュータ＃１)

となる。格構造データベースでは、２つの単文

＜製造する，過去，ａ社／Ａ，コンピュータ＃１／Ｏ＞

＜購入する，過去，ｔ大学／Ａ，コンピュータ＃１／Ｏ＞

の合成として表現する。

３．２　包含連体修飾

「ｔ大学が購入したコンピュータはすべてａ社が製造した」という内容を考える。これは、「ａ社が製造した」コンピュータの集合の中に、「ｔ大学が購入した」コンピュータの集合が含まれることを意味する。「ｔ大学が購入した」コンピュータが空ではないと考えるなら、多ソート論理式では、

(∀Xc／コンピュータ 購入した(ｔ大学,Xc) → 製造した(ａ社,Xc)) ∧

(∃Xc／コンピュータ 購入した(ｔ大学,Xc))

と表わされる。これをそのまま格構造データベースで表現するなら、

＜製造する，過去，ａ社／Ａ，コンピュータ@１／Ｏ＞

← ＜購入する，過去，ｔ大学／Ａ，コンピュータ@１／Ｏ＞　Ｓ４

＜購入する，過去，ｔ大学／Ａ，コンピュータ＃１／Ｏ＞　Ｓ５

となるが、格構造データベースでは、この他に、Ｓ４とＳ５から導かれるＳ６もデータとして保持しておく。

＜製造する，過去，ａ社／Ａ，コンピュータ＃１／Ｏ＞　Ｓ６

ただし、Ｓ４が入力節として選ばれて単一化に成功したならば、Ｓ５とＳ６が入力節として選ばれることがないようにこれらにマスクをかけ、入力節としてのＳ４の処理が終われば、マスクをはずす。

これにより、演繹推論を行わずに、「ａ社がコンピュータを製造した」ことを導くことができる。

４．合成ソート

格構造データベースでは、格の中身は単一の定数またはソートのみではなく、次のような構造(合成ソートとよぶ)を許している。ここで、ｇijはソート、ｃijは定数またはソートである。（ｉ，ｊ，n_1，…，n_mは整数）

$(g11-(c12+\cdots+c1n_1))\cap\cdots\cap(gm1-(cm2+\cdots+cmn_m))$

合成ソートは、ソートをそのソートに属する定数の集合、定数を単位集合とみなして、それらに集合演算を施したものである。

単一化のルールは次のとおり。

格Ｃにおいて、合成ソートｅ1,ｅ2に対して、

①ｅ1のすべての要素がｅ2に含まれるとき

単一化に成功し、単一化代入は〔ｅ2→ｅ1／Ｃ〕となる。

②ｅ2のすべての要素がｅ1に含まれるとき

単一化に成功し、単一化代入は｛ｅ1→ｅ2／Ｃ｝となる。

③上のどちらでもなく、かつ、共通部分集合をもつならば、

単一化に成功し、単一化代入は｛ｅ1→ｅ1∩ｅ2／Ｃ，ｅ2→ｅ1∩ｅ2／Ｃ｝となる。

④上のいずれでもないならば、単一化に失敗する。

５．部分全体関係･数値情報･同義語

５．１　部分全体関係

格構造データベースでは、定数とソートの階層構造を構成する関係として、包摂関係だけでなく、部分全体関係も利用している。これは、部分全体関係を利用した次のような検索が有用であると考えるからである。

＜購入する，過去，t大学工学部／Ａ，CWS2050＃１／Ｏ＞　　Ｓ７

（t大学工学部がＣＷＳ２０５０を購入した）

Ｓ７の単文と、t大学工学部がt大学の部分であるという部分全体関係があるとき、「CWS2050を購入した大学は？」という質問に対して「t大学」が解として期待される。

検索処理の過程では、上記の質問を表すゴール節、

←　＜購入する，過去，大学@／Ａ，CWS2050@／Ｏ＞

を選択式とし、Ｓ７を入力節とする。このとき、代入｛大学@→t大学／Ａ，t大学工学部→t大学／Ａ，CWS2050@→CWS2050＃１｝により、単一化に成功すればよい。２．５節で述べた、単一化の成功の基準(3)に次をつけ加える。

選択式の格と同一の格が入力節のヘッド部に存在し、そのヘッド部の格の中身が定数であり、それの部分全体関係による上位概念（全体概念）で、選択式の格の合成ソートに包摂される定数ｗが存在する。このとき、格をＣ、入力節のヘッド部の格の定数をｔ、選択式の格の合成ソートをｅとするとき、単一化代入は｛ｔ→ｗ／Ｃ，ｅ→ｗ／Ｃ｝。ただし、入力節のヘッド部の述語と格の組合せが、本項目の適用を受けるものと定義されている場合に限る。

５．２　数値情報

格構造データベースでは、文知識あるいは質問文において格に数値情報を表現することができる。数値情報は、単一の数値または数値範囲と、数値の単位からなる。数値情報における限量子の意味は、∀については、その数値範囲に属するすべての値についてその文知識（あるいは質問文）の内容が成り立つことであり、∃については、その数値範囲に属するある値についてその文知識（あるいは質問文）の内容が成り立つことである。次に例を示す。

＜購入する，過去，t大学工学部／Ａ，[1985,1990]年＃２／Ｔ，CWS2050＃１／Ｏ＞

（t大学工学部は 1985年から1990年の間にＣＷＳ２０５０を購入した）

５．３　同義語

格構造データベースでは、定数･ソート･述語に対して概念ＩＤを付与し、概念階層知識･文知識･テンプレート･検索文においては、概念ＩＤによりそれぞれを識別している。それぞれの概念ＩＤとそれに対応する文字列との対応は別に管理する。同義語には同一の概念ＩＤを与えておくことにより、表記が異なっていても同一視することができる。また、同綴異義語については、それぞれの意味ごとに概念ＩＤを与えることにより、区別して扱うことができる。同様の考えは、［10］にもみられる

６．おわりに

格構造データベースは、自然語意味情報を表現しやすくするために、連体修飾･合成ソート･部分全体関係･数値情報･ＩＤによる概念の管理を導入した格構造データベースモデルをもつ演繹データベースである。本データベースは、リファレンスＤＢとしてだけでなく、ファクトＤＢとしての機能も果たすものであり、より高度な検索機能をもつ情報検索システムの要素技術になりうるものと考えている。

［１］拝原正人: 日本語文献データベースへの知的アクセス: 電子情報 信学会誌, 72,7, pp. 797-806, (1989).

［２］杉山健司, 秋山幸司, 伊吹潤, 川崎正博, 内田裕士: 自然言語理解に基づく情報検索システム IRIS: 情報処理学会,自然言語処理研究会資料, 58-8,(1986).

［３］稲垣博人: 文書内容検索システムにおける現状と課題 －文書入力型内容検索システムを中心に－: 情報管理, 33,2, pp. 123-135,(1990).

［４］福永博信, 齋藤珠喜: 全文探索と多様な表現: 情報処理学会第39回全国大会, 1G-7, pp. 682-683,(1989).

［５］Fujisawa, H., Kiuchi, I., Koguchi, T., Kondo, H.: A Visual Interface for a Personal Information Base using a Concept Network: Proc. of The second International Symposium on Database Systems for Advanced Applications, Tokyo, (1991).

［６］赤間清: PAL:継承階層を扱う拡張Prolog: 情報処理学会論文誌, 28,4, pp. 322-329,(1987).

［７］Huber, M., Varsek, I.: Extended Prolog for Order-Sorted Resolution: Proc. of '87 International Symposium on Logic Programming, pp. 34-43,(1987).

［８］Montini, G.: Efficiency Considerations on Built-in Taxonomic Reasoning in Prolog: Proc. of IJCAI-87, pp. 68-75,(1987).

［９］Ait-Kaci, H., Nasr,R.: LOGIN:A Logic Programming Language with Built-in Inheritance: J. Logic Programming, 1,3,pp. 185-215,(1986).

［１０］McSkimin, J.R., Minker, J.: A Predicate Caluculus Based Semantic Network for Deductive Searching: Associative Networks, N.V. Findler (eds.), pp. 205-238, Academic Press, (1979).