

知識ベース管理システム KAPPA のデータ・モデル

— NF2 を基にしたモデル —

金枝上教史*¹ 三石彰純*² 加藤隆賢*³

*1: 新世代コンピュータ技術開発機構, *2: 三菱電機, *3: エルゴソフト

1. はじめに

知識ベース管理システム Kappa [1] においては、種々の知識処理システムで使用する複雑で大量のデータを効率よく蓄積・管理するため、NF2 (Non First Normal Form) を基にしたモデルを最下層にもつ4階層のデータ・モデルを採用している。非正規型モデルを基本的なデータ・モデルとすることで、複雑な構造として表現される知識の格納と操作を容易にしている。本発表では Kappa が採用しているモデル階層について説明する。

2. データ・モデルの階層

Kappa におけるデータ・モデルは図1に示すように、外部モデル、通信モデル、基本モデル、原始モデルの4階層からなり、ディスク上の Kappa ファイルには原始モデルと同じ内容で構造を持たないストリングを格納する。

通信モデルを除く3階層はそれぞれのレベルのテーブルを定義するものであり、通信モデルは外部モデルから基本モデルへの変換を定義するものである。

外部モデルは意味ネットワークやフレームなどを表現したり操作する機能を提供し、ユーザが機能を拡張することも可能である。項を基本的なデータとしてサポートしており、単一化による検索をサポートする。

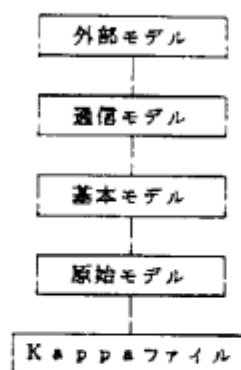


図1. Kappa のモデル階層

通信モデルはローカル/リモートPSIの通信機能を持ち、データベースの物理的な所在地を意識せずにアクセスできるようにする。

基本モデルは原始テーブルをそのまま見せる非正規型テーブル、原始テーブルの機能を制限した関係型(正規型)テーブル、静的な結合関係をサポートする結合型テーブル、静的な関係をサポートする分割型テーブル、概念辞書のような構造データを扱う構造型テーブルなどを提供する。

原始モデルは非正規型モデルにもとづいたデータ構造とデータ操作を提供する。

3. 非正規型モデルの必要性

Kappa のデータモデルを決定するのに当たって、Kappa 開発の直接のきっかけとなった自然言語処理システムにおける電子化辞書、概念辞書の構造や、定理証明支援システムにおける数学の定義、定理、証明などの構造を調査し、要求される機能を整理した。その主な要求項目を列挙すると、

- ・膨大な量(数十万件)のデータが格納できること
- ・階層構造をもつデータが効率よく格納できること
- ・比較的高速に検索できること
- ・データの構造変化に対応できる柔軟性をもつこと
- ・データの型として項を扱えること(単一化による検索を含む)
- ・非常に長いテキストを格納できること

などである。これらの条件から Kappa の原始モデルとして非正規型モデルを採用したのは以下の理由による。

①知識処理システムにおいて扱うデータ(知識)は複雑な構造をもっており、正規型では簡単に表現できない。無理に表現すると正規化のためにデータが幾つかのテーブルに分割され、検索時に結合演算が不可避となる。

②非正規型モデルでは属性として属性の集合をもつことや、属性の値として領域の部分集合をもつことができるため、実世界の複雑な対象を非常に素直な形で表現することができる。

③非正規型モデルは正規型の拡張と考えられ、正規型モデルの体系をベースとして使うことができる。

④非正規型では数学的な扱いが複雑になるが、正規型と非正規型のスキーマ変換や非正規型に対する代数についての研究が進んできている。

A Data Model for Knowledge Base Management System KAPPA
- A Model based on NF2 -
Atsushi KANAEGAMI¹, Akitoshi MITSUISHI², Takahiro KATO³
*1: ICOT, *2: Mitsubishi Electric Corp., *3: Ergo-soft Corp.

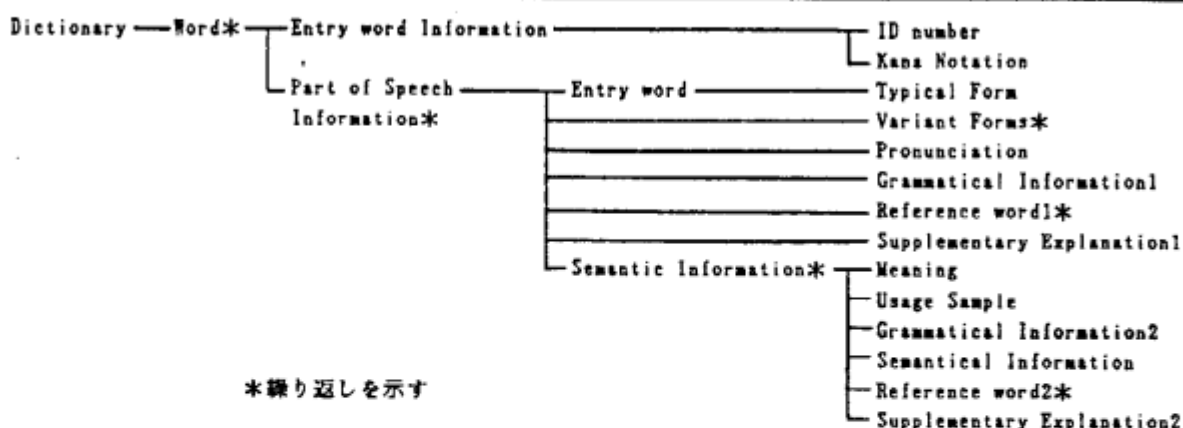


図2. 電子化日本語辞書のデータ構造例

4. 原始モデルの構造

原始モデルにおけるテーブルを原始テーブルと呼び、スキーマとレコード集合の対で定義する。スキーマは属性の集合であり、属性には以下の4種類がある。

- ①単純属性
- ②繰り返し属性
- ③集団属性
- ④繰り返し集団属性

単純属性はアトミックなデータ、すなわち整数、浮動小数点数、ストリング、空値、項*注などを属性値としてとり、繰り返し属性はアトミックな値の集合を属性値としてとる。集団属性は他の属性から帰納的に定義される構造型データを属性値としてもち、繰り返し集団属性は集団属性の属性値の集合を属性値としてもち、繰り返し属性、繰り返し集団属性の個々の属性値をその属性値のオカレンスと称する。KAPPAではテーブルにおけるレコードの順序が意味を持たないように、属性におけるオカレンスの順序にも意味を与えていない。従って、繰り返し属性には複数の値を持ち値間の順序に意味を持たないデータを格納する。一方、1つの属性に複数の値があり、値の順序に意味がある場合には順序を示す別の属性を付加して繰り返し集団属性として表現する。

図2に電子化日本語辞書のデータ構造例の木表現を示す。図から明らかなように階層的で複雑な構造を持っている。図3には図2で示した辞書を原始テーブルで表現し、1つの語の情報を1つのレコードで表現した例を示す。図において、右側が単純属性または繰り返し属性でアトミックな値をもっている。左側は集団属性または繰り返し集団属性で、それら自身はアトミックな値をもたない。非正規型モデルを採用することによって、階層的で複雑な構造を持つデータが素直に表現できている。

*注: 原始テーブルで提供するデータ型としての項には単一化による検索はできない

集団 / 繰り返し集団属性		単純 / 繰り返し属性	
Word	Entry word Information	Entry word	ID number
			Kana Notation
	Part of Speech Information*	Semantic Information*	Typical Form
			Variant Forms*
			Pronunciation
			Grammatical Information1
			Reference word1*
			Supplementary Explanation1
			Meaning
			Usage Sample
			Grammatical Information2
			Semantical Information
			Reference word2*
			Supplementary Explanation2

*は繰り返し集団 / 繰り返し属性

図3. 原始モデルで表現した辞書データ

5. まとめ

KAPPAにおけるモデルの階層と非正規型の原始モデルについて述べた。試作システムではテーブルの検索、読み、更新の機能を実現し、テーブル間の和、差、結合などの演算は本格システムにおいて実現する。現在、テーブル及びテーブル操作の数学的定式化を進めており、正規型モデルとの間の構造変換などが今後の課題である。

参考文献

- [1] 横田, 内田, 濱口, “知識ベース管理システム KAPPAの構想”, 情報処理学会第33回全国大会, 5M-5, 1986.