2Q−7 辞書ハイパーテキスト Hydra(1) — データベース構築手法 —*

内藤 広志, 根本 治朗, 山下 真司, 松山 洋一, 柵木 孝一†
キヤノン(株) 情報システム研究所‡

1 はじめに

ノードとリンクからなる単純なモデルによって文書を表現するハイパーテキストが最近注目を集めている[1]. この技術を用いて新しく文書を作成するだけでなく, 既存のリニアな構造を持つテキストデータをハイパーテキストのノンリニア構造へ変換することは, 過去の資産の有効利用のためにも重要である[2, 3].

そこで我々は, ハイパーテキスト技術の大規模データへの適用例として, 辞書ハイパーテキスト **Hydra** (HyperDictionary Reading Accessories) の開発を進めると同時に, そこから得られた知見に基づいて, 大規模ハイパーテキストデータベースの構築手法について検討したので報告する.

2 データベース構築手法の概要

我々が **Hydra** の開発でとった構築手法を図示すると, 図1のようになる.

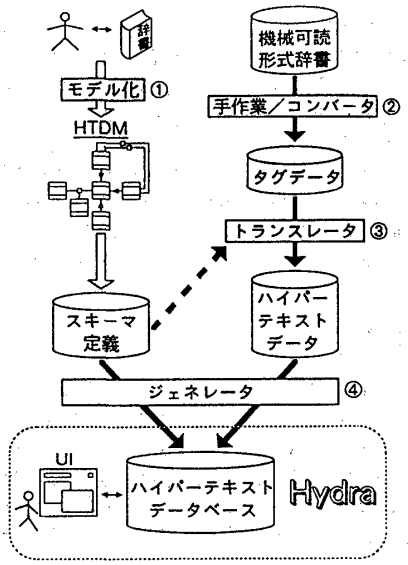

図 1: ハイパーテキストデータベース構築手法

この手法は次の4つのステップからなる.

1. **原データの分析とモデル化**: 原データの構造を分析し, ハイパーテキストデータベースの概念構造をハイパーテキストデータモデル (HTDM) によって表現する.
2. **原データの整形**: 機械可読形式辞書に対し, コンバータ／手作業を用いてタグ付けをする.
3. **ノードへの分割とリンクの付与**: ステップ1で記述された構造に基づいて, トランスレータによってノードへの分割とリンクの付与を行なう.
4. **データベースへの格納**: ジェネレータによって, データベースへの格納を行なう.

この手法の特徴は, データベース設計法や CASE で用いられている構造分析の考えをハイパーテキストの設計に導入したことと, SGML を使用してハイパーテキスト構造を明示的に記述したことである. 特に SGML を採用した理由は, データのシンタクスを明確に記述することに適していることと, 今後 CD-ROM 等のデータ記述の標準として用いられることが期待されるからである.

3 ハイパーテキスト化の処理フロー

本章では **Hydra** の開発を題材に, 既存データのハイパーテキスト化の処理フローについて述べる.

3.1 原データの分析とモデル化

機械可読形式辞書をハイパーテキスト化する際, 単純に既存データに「反義」や「同義」などのリンク情報を付加するだけでは十分でない. フィルタや検索などにより効果的なナビゲーションをするためには, 既存データをコンピュータ処理に適した形に構造化する必要がある.

そこで, 我々はハイパーテキストのノンリニア性が表現可能なハイパーテキストデータモデル (HTDM) を設計し, これを用いて辞書のデータ構造をモデル化した[4].

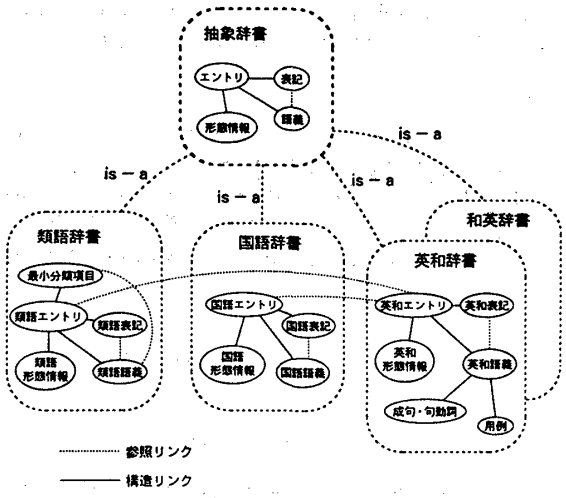

図 2: 辞書モデル

また, ただ単に個々の辞書をモデル化するだけではなく, 様々な辞書構造を抽象化した辞書モデルを考え, 個々の辞書はこれの特殊化したものと考えてモデル化した. これにより辞書データベースへの統一的なアクセス, 複数辞書間のリンク, 複数辞書のマージが可能となる. 図2に抽象辞書モデルと個々の辞書の関係を示す.

---

*The HyperDictionary **Hydra** (1) – Database Constructing –
†Hiroshi Naito, Jiro Nemoto, Shinji Yamashita, Yoichi Matsuyama, Koichi Masegi
‡Information Systems Research Center, CANON Inc.

図 3 に HTDM を用いて国語辞典[5]を記述した例を示す.

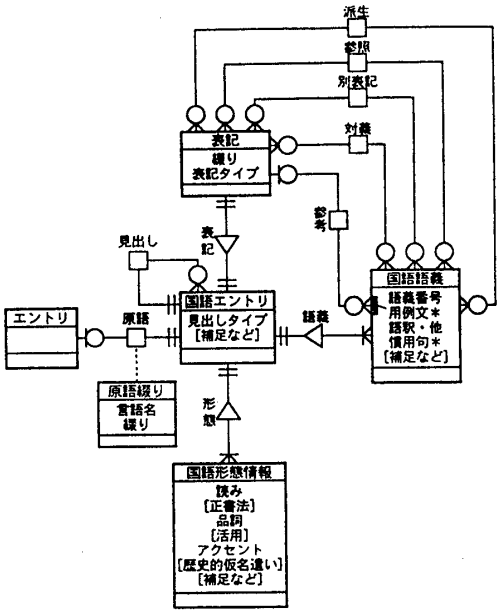

図 3: ハイパーテキストデータモデル

3.2 原データの整形

書籍形式の辞書をそのままコード化した機械可読形式辞書は，一定の文法に基づいて記述されているものの，それは人間が理解できることが目的であり，計算機で処理を行なうには厳密に欠ける[6]．そのために各語彙毎に「アクセント」や「正書法」などの項目に分割し，各項目を識別するためのタグをつける必要がある．これを行なうのがコンバータである．しかし，原データには記述エラーも多いため，手作業で処理しなければならない部分が多い．図 4 は原データの機械可読形式辞書，図 5 はそれをコンバートした結果のタグデータである．タグデータは SGML を用いて記述される．

```
#こと　ば#3［言葉］　（「は」は端の意）
　　1）社会ごとに決まっている、音声による表現（行為）。〔広義
では、文字による表現（行為）をも指す〕「意味の分からない―〔＝単
語〕・推敲の―〔＝文章〕・―を返す〔＝A返答する。B相手の言葉を
反駁（ハンバク）する〕・―〔＝言い方〕が悪い・話し―・書き―・い
なか―」
　　2）¥〔詞〕〔地の文に対して〕〔小説・戯曲の〕会話（文）。
　　……
```

図 4: 機械可読形式辞書[5]

3.3 ノードへの分割とリンクの付与

ハイパーテキスト構造を明示的にするために，トランスレータはタグデータ中の SGML でタグ付けされた項目をノード（オブジェクト），属性，リンクに区別して記述し，ハイパーテキストデータとして出力する．その際，タグデータの構造もHTDM で記述されたノードの構造へ変換される．また，他のノードからの参照ができるためにノードへ名前が付けられる．一方，既存データ中に現れるリンクは，特定のノードそのものを示すものと，特定のキーを持つノード（の集合）を示すものがあるため，それぞれ個別的リンク，関数的リンクとして区別して記述される．図 6 は図 5 のデータをトランスレートした結果のハイパーテキストデータである．

```
<一般の単語>
　<マーク>2
　<見出し>ことば
　<補足>
　　<ACC>3
　　<正書法>言葉
　　<補足的説明>「は」は，端の意
　<語義><小語義>
　　<小語義番号>1
　　<小語義本体>
　　　<語義文>社会ごとに決まっている，音声による表現(行為)。
　　　　〔広義では，文字による表現……
　　　<用例等>意味の分からない〜〔=単語〕……
　　……
</一般の単語>
```

図 5: タグデータ

```
<国語概念 label=ことば-1>
　<形態><国語形態情報>
　　<読み>ことば
　　<正書法>言葉
　　<アクセント>3
　<語義><国語語義>
　　<語義番号>1
　　<語釈・他>社会ごとに決まっている，音声による表現(行為)。
　　　　〔広義では，文字による表現……
　　<用例文>意味の分からない〜〔=単語〕……
　　……
</国語概念>
```

図 6: ハイパーテキストデータ

3.4 データベースへの格納

HTDM で記述されたハイパーテキスト構造を基にデータベースの論理スキーマを設計することができる．ジェネレータは，SGML で明示的に構造が記述されたハイパーテキストデータを読み込み，DBMS に格納する．ジェネレータの処理は使用する DBMS のタイプ（RDB や OODB など）によって異なる．例えば，使用する DBMS が RDB の場合は，ノードやリンクをテーブルに変換し，それぞれに ID を付与し，参照関係を解決した後，DBMS へ格納する．OODB の場合は，ノードをオブジェクトに変換し，リンクはオブジェクトのインスタンス変数にストアする．

4 おわりに

現在開発中の辞書ハイパーテキスト **Hydra** を題材に，既存の大量データをデータベースに取り込む手法について述べた．大量データの処理はデータ表現法やツールの整備によって，作業指針を明確にしたり，自動化することが必要である．今後は作業の各ステップにおけるデータ表現の表現能力を充実させるとともに，ツール群を整備していく予定である．

参考文献

[1] J. Conklin. Hypertext: An Introduction and Survey. *IEEE COMPUTER*, 2(9):17-41, 1987.
[2] D.R. Raymond and F.W. Tompa. Hypertext and The Oxford English Dictionary. *Communications of the ACM*, 31(7):871-879, 1988.
[3] L. Alschuler. Hand-crafted Hypertext - Lessons from the ACM Experiment. In E. Barrett, editor, *The Society of Text*, pages 343-361. The MIT Press, 1989.
[4] 根本治朗 他. 辞書ハイパーテキスト Hydra(2) — データモデル —. 情報処理学会第 4 2回全国大会, 1991.
[5] 金田一京助. 新明解国語辞典. 三省堂, 1989.
[6] 鶴丸弘昭, 内田 彰. 国語辞典からの情報抽出と構造化について. 長崎大学工学部研究報告, 15(24):41-48, 1985.