実体-関連モデルに基づくアプリケーションインタフェースの文書画像データベースへの適用

片山 紀生　安達 淳

学術情報センター 研究開発部
〒 112 東京都文京区大塚 3-29-1
{katayama, adachi}@rd.nacsis.ac.jp

現在、電子図書館の一形態として、文書画像データベースと WWW インタフェースの組み合わせが使われている。このようなシステムを、大規模かつ柔軟に構築するためには、動的に HTML 文書を生成し、縮退した形で情報を管理することが必要になる。その場合、文書データの要素をどうやって識別するか、アンカーの情報をデータベースからどうやって導き出すかが問題になる。そこで本稿では、文書画像データベースのアプリケーションインタフェースに、実体– 関連モデルを適用する方法を提案する。特に、一階述語論理に基づく記法を用いることで、記法を簡素化するとともに、データベースの内部構造によらない宣言的なインタフェースを実現する。

An Application Interface for Page Image Document Databases based on the Entity-Relationship model

Norio Katayama and Jun Adachi

Research & Development Department
NACSIS (National Center for Science Information Systems)
3-29-1 Otsuka, Bunkyo-Ku, Tokyo 112, Japan

Recently, the combination of a page image document database and a WWW user interface is widely used for digital libraries. Basically, such system can be constructed by writing all required HTML documents. However, it is necessary to generate them dynamically for flexible implementation. There are two problems for generating HTML documents: how to identify document components and how to retrieve hyperlink information from a database. To solve these problems, we applied the entity-relationship model to the application interface of a page image document database. By employing the first-order predicate logic, we simplifies notations of queries and derivation rules, and achieves an application interface having declarative semantics.

1 はじめに

現在、電子図書館の一形態として、WWW (World Wide Web) と文書画像データベースを組み合わせたものが広く利用されている。すなわち、図書のページを画像データとして電子化し、文書画像データベースに蓄積する。そして、そのユーザインタフェースとして WWW を適用し、WWW のネットワークアクセス機能ならびにハイパーメディア機能を利用するという方法である。このような形態は、全文データよりも画像データの方がデータ量が大きい、全文検索を使用できない、また、まだ発展途上の WWW の技術的制約を強く受けるといった短所がある反面、全文データに比べて電子化コストが低いことや、広く普及している WWW インタフェースを利用できるといった長所があり、電子図書館を構築するための簡便な形態として広く利用されている。

本稿では、そのような WWW を用いたユーザインタフェースを持つ文書画像データベースを構築するための手法として、データベースシステムのアプリケーションインタフェースに実体–関連モデルを適用した例について説明する。

本稿の構成は以下のようになっている。まず、2節で、この研究の背景について説明する。次に 3 節で、本研究が実装したプロトタイプデータベースシステムについて説明する。そして 4節で、本稿が提案する実体–関連モデルに基づくアプリケーションインタフェースについて説明し、5 節で本手法の特徴についてまとめる。

2 研究の背景

電子図書館の一形態として、文書画像データベースと WWW インタフェースの組み合わせが広く使われている。このようなシステムは、更新頻度が少なく、基本的には、雑誌、文献、ページといった表示対象となる要素に対して それぞれの要素に対応する HTML 文書を用意するだけで実現することができる。したがって、原理的には必要な全ての HTML 文書を作成することでシステムを構築することができる。しかし、蓄積データが大規模になったり、蓄積データに対する変更やユーザインタフェースに対する変更に柔軟に対応したり、利用者ごとのカスタマイズ機能を実現したりするには、HTML 文書を予め用意するという形ではなく、動的にあるいは batch 的に HTML 文書を生成するという形を取り、縮退した形で情報を管理することが必要になる。

図 1に、学術雑誌を対象とした文書画像データベースに対する WWW インタフェースの例を示す。この例は、本研究で作成したプロタイプシステムのものであり、雑誌のページ画像を検索・閲覧できるように設計されている。このようなハイパーメディアネットワークを構築する際に問題となるのは、以下の二点である。

- ノード内のアンカーの生成方法
- リンク先のノードの指定方法

銭ら [1, 2] は、質問対リンク (Query Pair) によってこれらの問題を解決している。すなわち、リンク先を指定する質問と、アンカーを指定する質問との二つの質問を対にすることによって、リンクを縮退した形で表現するというものである。この手法は、ハイパーテキストを対象としており、テキスト中の単語から他のノードへのリンクを生成するという用途に有効であることが示されている。ところが、文書画像データベースを対象とする場合には以下のような問題があるために、上記の問題に加えて、文書画像データベースとハイパーメディアとの間のインタフェースをいかに管理するかが大きな課題となる。

- 図 1 が示す通り、文書の論理構造と物理構造を反映したリンクを縦横無尽に張る必要がある。
- リンクの種類が多岐に渡るため、相互の依存関係をルールとして記述する必要がある。

例えば、図 2 はページを表示するノードの例であるが、このノードには以下のリンクのためのアンカーが埋め込まれている。

- 前のページへのリンク
- 後のページへのリンク
- 前の記事へのリンク
- 後の記事へのリンク
- 記事の詳細表示へのリンク
- 冊子の目次へのリンク

このような多様なリンクをどのようにして管理するかが、文書画像データベースでは問題となるのである。そこで、これらの問題を解決するために、本研究では文書画像データベースとハイパーメディアとの間のアプリケーションインタフェースとして、実体–関連モデルを適用する手法を提案する。

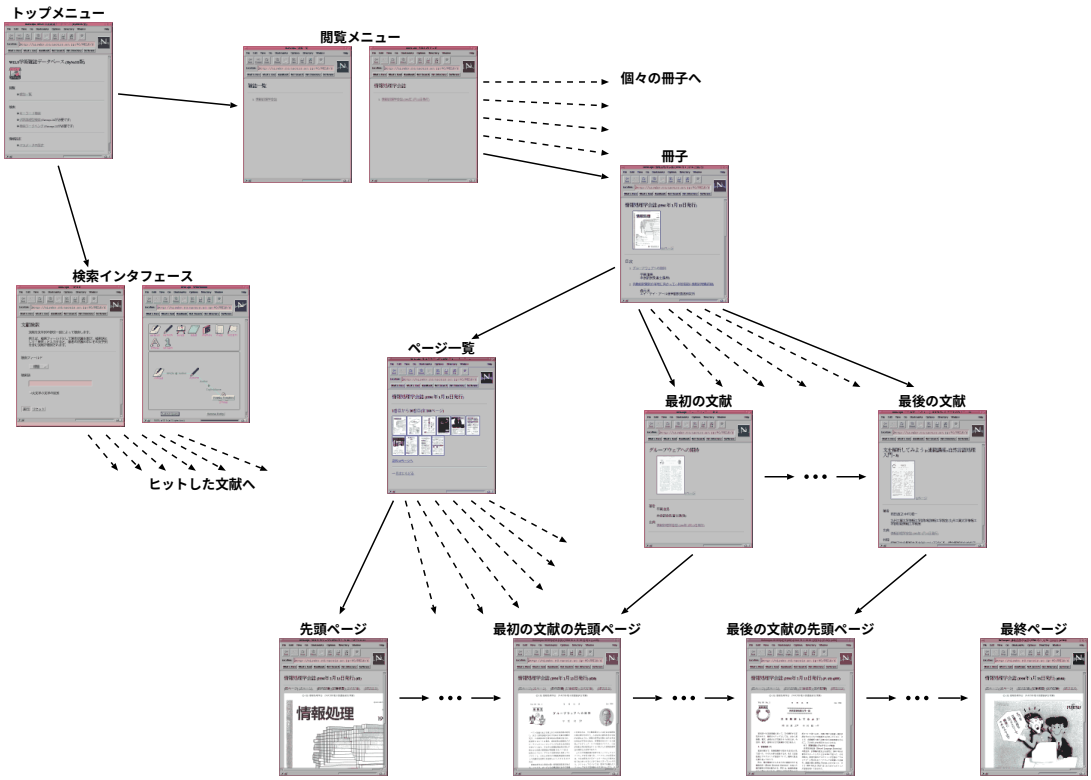

図 1: 文書画像データベースに対する WWW インタフェースの例

図 2: ページを表示するノードの例

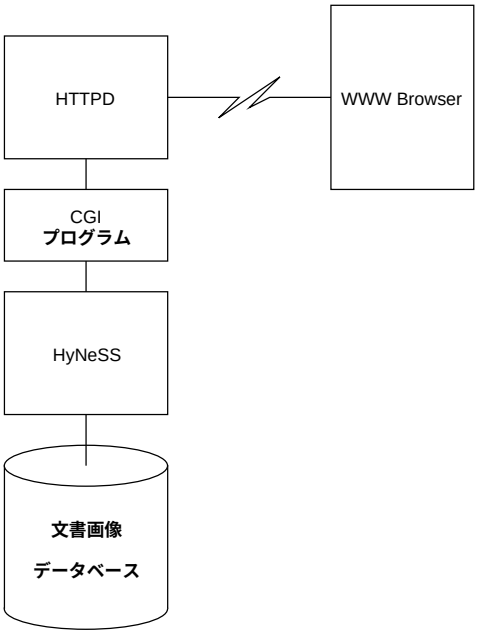

図 3: システム構成

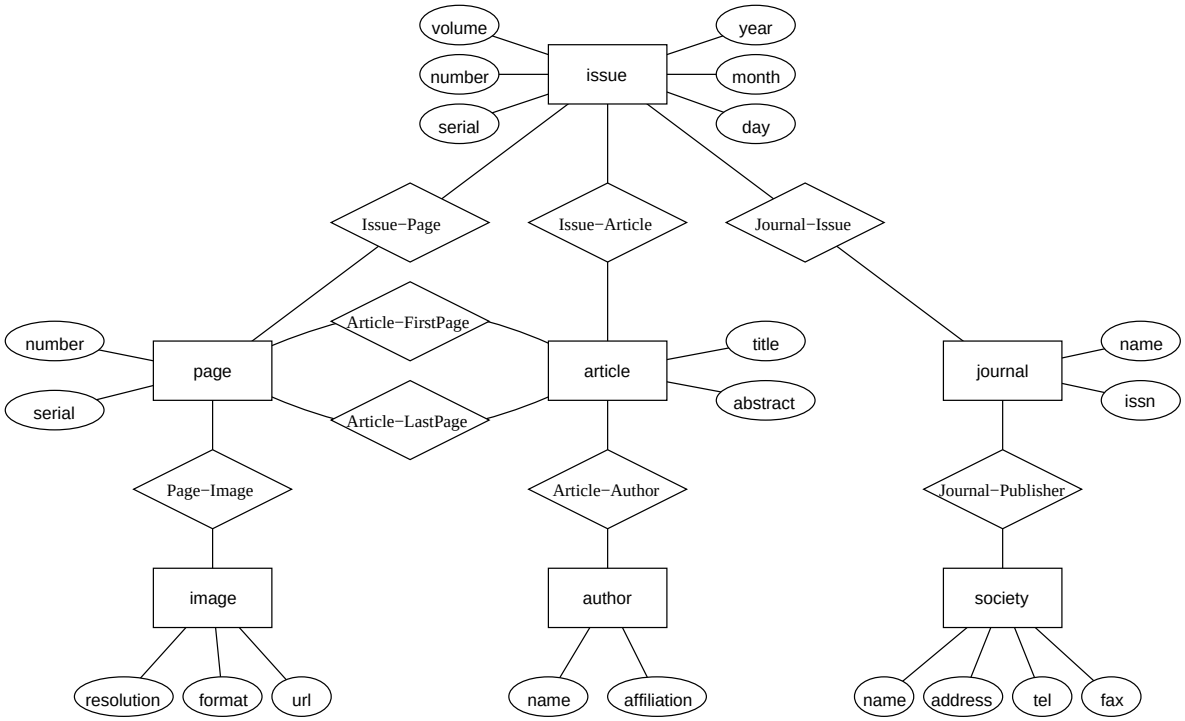

図 4: 文書画像データベースのスキーマ

3 文書画像データベースシステムの構築

3.1 プロトタイプシステムの構成

図 3 に、本研究で実装したプロトタイプシステムのシステム構成を示す。

まず、文書画像データベースには、以下の三種類の情報が蓄積されている。

- 雑誌のページをスキャンしたページ画像
- 雑誌に関する書誌情報
- 文献に関する書誌情報

図 4 にこのデータベースのデータベーススキーマを示す。データベースは、著者らが提案している記憶システム HyNeSS (Hyper-Network Storage System) [3] によって管理されている。HyNeSS は、ハイパーメディアを実体–関連モデルに基づいて蓄積するための記憶システムであり、ノードを実体、リンクを関連によって表現し、特に、関連の管理に一階述語論理を用いる点に特徴がある。

CGI プログラムは WWW ブラウザからのリクエストに応じて HTML 文書を生成する。WWW ブラウザからリクエストがあると httpd が CGI プログラムを起動する。すると、CGI プログラムは HyNeSS を介してデータベースを検索し、HTML 文書を生成するのである。

3.2 HTML 文書を生成する上での問題点

CGI プログラムによって HTML 文書を生成する処理は、「文書データのある要素 (目次、文献、ページなど) が与えられたとき、データベースを検索しその要素を表現するための HTML 文書を生成する」という処理になる。このような処理を実現するためには、次の二つの問題を解決しなければならない。

- 文書データの要素をどうやって区別 (識別) するか？
- 図 4で与えられるスキーマから、どうやってアンカーを埋め込むための情報を導き出すか？

特に、後者については、埋め込むべきリンクの種類が多岐に渡る上に、複数の関連から導出しなければならない場合もあるため、アプリケーションプログラムを効率的に開発するためには、簡潔な方法でリンクを検索できることが求められる。

そこで本研究では、これらの問題を解決するために、文書画像データベースと CGI プログラムの間のアプリケーションインタフェースに、実体–関連モデルを適用する方法を提案する。

4 実体–関連モデルに基づくアプリケーションインタフェース

4.1 アプリケーションインタフェースの概要

本研究が提案するアプリケーションインタフェースは、著者らが提案している記憶システム HyNeSS に基づくものであり、以下の特徴を持っている。

- データベース中の論理的な識別対象 (実体) ごとに識別子を割り振る。
- 実体間の関連を述語として記述する。
- 関連間の導出関係や関連に対する検索を一階述語論理によって記述する。

4.2 実体への識別子の付与

本研究では、実体を識別するための手法として、URN (Uniform Resource Name)[4, 5] に準じたものを使用している。URN は、現在 IETF (Internet Engineering Task Force) で標準化が進められている URI (Uniform Resource Identifier) のひとつで、URL が資源の位置を示すのに対して、URN では位置に依存しない一意な識別子を記述するための枠組を提供する。すでに、実装されているものとしては CNRI の handle が知られており、Cornell 大学が中心になって運営している分散電子図書館システム NCSTRL[6] で既に使われている。

本研究ではこの URN に準じた以下の特徴を持つ識別子を用いている。

- 文書の論理的な対象を一意に識別する。
- 文字列によって表記する。
- 階層性は持たせるが、それ以上のフォーマットを規定しない。

例えば、情報処理学会論文誌 35 巻 1 号に対しては、次のような識別子を割り当てる。

```
/issue/tipsj/35/01
```

また、情報処理学会論文誌 35 巻 1 号の 1 ページめには、次のような識別子を割り当てる。

```
/page/tipsj/35/01/0001
```

このような識別子を導入することによってシステムを構築する上で次のような利点がある。

- URN のように位置に依存しないので、データ独立性が高まる。
- 文字列によって表記しているので、既存の書誌データフォーマットに組み込み易い。また、人間がある程度解釈できるのでわかり易い。
- 階層性を持っているため、名前空間を独立な領域に分割していくことで、名前の一意性を持たせやすい。
- 階層性以外のフォーマットを規定していないので、適宜ローカルなコンベンションを決めることができる。

4.3 述語によるリンクの記述

HTML 文書を生成するためには、文書データの論理要素 (実体) 間の関連を検索することが必要である。これらの関連については、オブジェクト指向モデルのように属性とみなしたり、関数型モデルのように関数とみなす方法も考えられるが、これらは関連を一方向の写像として表現しており、文書構造のように双方向性のあるリンクを表現することには適していない。そこで、本研究では実体間の対称性をよく反映すること、また、一階述語論理に基づく導出機構を利用できることから、関連を述語によって表現する。

例えば、冊子と文献の間の関連 `Issue-Article` は、述語によって次のように表現する。

```
[ Issue=$issue, Article=$article ]
```

一般的には、述語は `Issue-Article(X, Y)` というように、述語名の後に値を並べて記述することが多いが、本研究では述語の対称性を活かすために、項の順序に意味を持たせず、それぞれの項にラベルを置く形をとっている。すなわち、括弧の中はラベルと値の組の並びであり、`Issue`, `Article` がラベル、`$` で始まっている `$issue`, `$article` が変数である。

述語による表現の対称性は、リンクを検索する際に顕著に現れる。例えば、この述語は、次のように、ある冊子 (`/issue/tipsj/35/01`) が掲載している文献を検索するのに用いることもできるし、

```
[ Issue=/issue/tipsj/35/01, Article=$X ]
```

逆に次のように、ある文献 (`/article/tipsj/35/01/01`) を掲載している冊子を見つけることにも利用できる。

```
[ Article=/article/tipsj/35/01/01, Issue=$issue ]
```

このように、一つの述語を対称的に使えることが、リンクを述語によって表現することのメリットであり、アプリケーションプログラムにおいて一貫した記法を提供することができる。

4.4 一階述語論理によるリンクの導出

HyNeSS では、一階述語論理に基づいた記法で、リンクを検索したり、導出することが可能になっている。例えば、冊子 `/issue/tipsj/35/01` が掲載している文献のタイトルは、下の質問によって検索できる。

```
[ Issue=/issue/tipsj/35/01, Article=$article ],
[ Article=$article, Title=$title ]
```

また、ルールを定義することによって新たな述語を定義することも可能であり、例えば、文献と著者名の関連を表す `Article-AuthorName` という述語は、スキーマ上の `Article-Author`, `Author-Name` から下のように定義される。

```
[ Article=$article, AuthorName=$name ] :-
     [ Article=$article, Author=$author],
     [ Author=$author, Name=$name]
```

この記法の優れている点は、この記述だけで文献→著者名の関係だけでなく、著者名→文献の関係も双方向に記述できている点である。さらに、HyNeSS の特徴として、四則演算のような手続き型の関連も述語として宣言的に記述できるようになっており、例えば、ページの前後関係は、ページの冊子ごとの通し番号 (serial number) から下のように導出できる。

```
[ PrevPage=$prev, NextPage=$next ] :-
  [ Issue=$issue, Page=$prev ],
  [ Issue=$issue, Page=$next ],
  [ Page=$prev, Serial=$serialPrev ],
  [ Page=$next, Serial=$serialNext ],
  [ Sum=$serialNext, Value1=$serialPrev, Value2=1 ]
```

このように実体間のリンクを対称的に検索したり導出できることは、検索方向によらない一貫した記法を提供できるので、アプリケーションプログラムを書く上で大きな利点となる。

5 結論

本稿が提案しているアプリケーションインタフェースの特徴をまとめると以下のようになる。

- 文書の構成要素を、URN に準じた一意な識別子によって識別する。
- 関連を述語によって記述し、リンクの検索方向によらない一貫した記法を提供する。
- 一階述語論理に基づいた宣言的な記法で、リンクを検索したり、新たな関連を定義できる。
- 導出ルールを一階述語論理で記述することにより、ひとつのルールで双方向のリンクを導出できる。
- 四則演算など手続き型の処理も述語として宣言的に記述できるので、手続き型の処理を含む導出についても、ひとつのルールで対称的にリンクを導出できる。

これらの特徴から、本稿が提案する手法は、アプリケーションプログラムを開発する上で以下の利点を持っている。

- 検索方向によらない一貫した記法を提供することで、記法の種類を減らすことができる。
- ひとつのルールで双方向のリンクを導出することで、ルールの数を減らすことができる。
- URN に準じた一意な識別子と、一階述語論理に基づく宣言的な記法を用いることにより、データベースの内部構造から独立したアプリケーションプログラムを開発できる。

これらの利点は、大規模な文書画像データベースを構築する際に、特に有効であると考えられる。

謝辞

本研究は日本学術振興会産学共同研究支援事業からの助成を受けました。また、実験データは、学術情報センター電子図書館プロジェクトに協力頂いている情報処理学会のデータを使わせて頂きました。ここに深謝致します。

参考文献

[1] 銭 晴, 谷崎正明, 原伸一郎, 田中克己, 「ハイパーテキストデータベースシステム TextLink/Gem におけるオブジェクトとスキーマの動的、段階的な構築機能」, 信学技法 DE92-39 (1993).

[2] Qian, Q., Tanizaki, M., and Tanaka, K., "Abstraction and Inheritance of HyperLinks in an Object-Oriented Hypertext Database System TextLink/Gem," *IEICE Trans. Inf. & Syst.*, E78-D, 11 (Nov 1995) pp. 1343–1353.

[3] 片山紀生, 高須淳宏, 安達 淳, 「ハイパーメディア型のネットワーク構造に基づく記憶システムの設計と実装」, 情報処理学会研究報告, 95-DBS-104 (Jul. 1995), 57–64.

[4] Sollins, K., and Masinter, L., "Functional Requirements for Uniform Resource Names," RFC1737 (Dec 1994).

[5] The URN Implementors, "Uniform Resource Names, a Progress Report," D-Lib Magazine (Feb 1996).

[6] Lagoze, C. and Davis, J., R., "Dienst: An Architecture for Distributed Document Libraries," *Communications of the ACM* **38**, 4 (Apr 1995) 47, `http://cs-tr.cs.cornell.edu/` .