情 報 学 基 礎 20－6
（1990. 12. 18）

マルチメディア文献システムにおけるデータの構造化とビュー

高須 淳宏、桂 英史、相澤 彰子、原 正一郎

学術情報センター

マルチメディア文献システムのデータベースにおいて、データベースへのデータの入力を自動化するために、データ入力時におけるデータの構造化とデータベーススキーマの作成を統一的に扱う方法について述べる。本稿で述べる方法は、スキャナーを使って雑誌の目次の画像データを読み込んでからデータベースに入力するまでの一連の作業を自動化するための方法である。目次に関する文法を使用することによって入力データを解析し構造化するとともに、文法からデータベーススキーマを生成することによって、入力データの解析とデータベース化を自動化する方法を示した。本報告の方法によってデータベース入力の自動化の展望が開けるとともに、文法の記述力によって対象データの構造の変化に柔軟に対応できるスキーマ作成が可能であることを示した。

DATA STRUCTURING AND VIEWS OF A MULTIMEDIA DOCUMENT SYSTEM

Atsuhiro Takasu, Eishi Katsura, Akiko N. Aizawa, Shoichiro Hara

National Center for Science Information System

3-29-1 Otsuka Bunkyo-ku, Tokyo 112, Japan

A analytical and integrated method for inputted data and generation of database scheme in multimedia document delivery systems is discussed. This method is applied to the continuous process from scanning contents sheets with image scanner to constructing database. In this paper, we propose a grammatical approach to contents image analysis and data structuring. This method enables to construct database automatically. In addition to that, this method can cover the variation of data structure by means of the syntactical descriptin ability.

1　はじめに

本報告では、マルチメディア文献システムのデータベースにおいて、入力時におけるデータの構造化とデータベーススキーマの作成を統一的に行う方法について述べる。

従来の文献検索システムは書誌情報だけを管理するシステムであったため、サービスは文献の存在や所在情報の検索に限られていた。またマンマシンインタフェースの機能は限られており、システムの操作方法は十分洗練されているとは言い難い。一方データベースを提供する機関においても常に増加するデータをデータベースに格納することは大変な作業であった。そこで、筆者らは、雑誌の目次にある情報を使用した文献検索において、従来の書誌情報サービスに加え全文データをも提供する文献システムの開発を進めている[1]。このシステムは、データベース構築を自動化することと、画像データを駆使することによって優れたマンマシンインタフェースを作成することを目指している。

本報告では、データベースの構築の自動化に関連して、入力データをデータベース用のデータ構造へ変換する方法について述べる。入力データからデータベースを作成するためには、入力データをデータベースに適したデータ構造に変換する必要がある。そこで、対象データに関する知識を文法を用いて記述することによって、データベーススキーマを生成し、構文解析によって得られた構文解析木をデータベース用のデータ構造に変換する。

本報告は5章よりなり、2章で筆者らが開発しているシステムの全体像の概略を示すとともに、このシステムと本報告で述べる方法との関連について述べる。3章では、簡単な目次の例を用いて、入力データの解析の過程を示す。4章では、文法からデータベーススキーマを生成する方法および構文解析木からデータベース用データへの変換方法を示す。5章では、本報告で述べた方法の特徴と今後の課題について述べる。

2　システムの全体構成

筆者らが開発している文献情報システムは、入力系、蓄積系、マンマシン系の3つのサブシステムによって構成される。入力系サブシステムは、スキャナーを使って画像データを取り込み、画像データから目次の書誌情報を取り出し、データベース用に構造化する。蓄積系サブシステムはマルチメディア型のデータベース管理システムであり、目次の画像データと入力系で構造化されたデータを管理するとともに、マンマシン系から出される検索要求を処理する。マンマシン系サブシステムは、目次そのものを利用者のビューとして、目次に対する種々の操作を支援するためのグラフィカルインタフェースである。本報告では、入力系システムにおいて、入力データを解析し、データベースに格納するための構造に変換する方法について述べる。

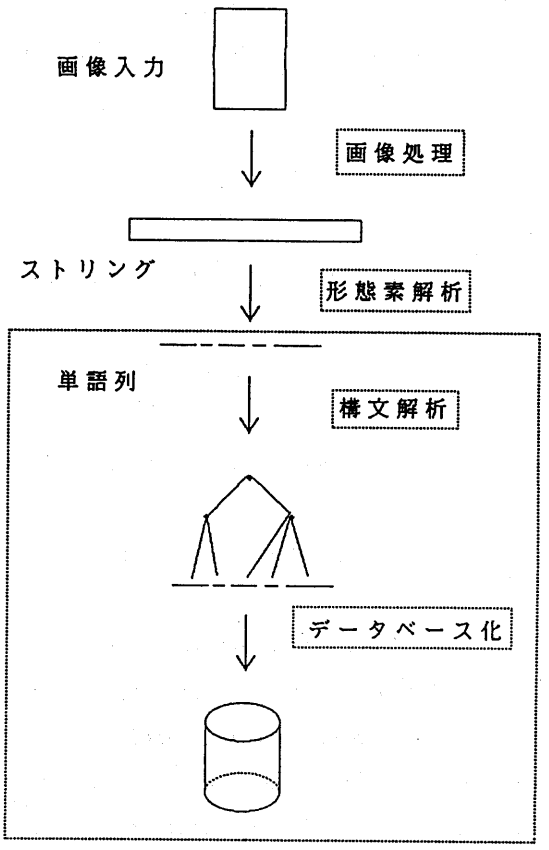


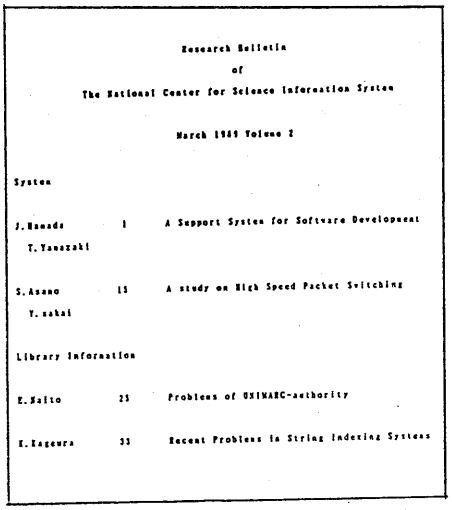
図1　入力システムの処理の流れ

入力系サブシステムでは、まずスキャナーを使って読み込み、画像処理を施してセグメント化を行い、さらに文字領域を取り出し、文字認識を行う[2]。この処理によって目次を左上から右下に水平方向にスキャンした結果得られる文字列が出力される。次にこの文字列に対し、自然言語処理と同様に形態素解析を行って単語単位の切りわけを行う。英文の場合には、画像処理の段階で単語列が出力されることになる。次に目次の文法を利用して構文解析を行い構文解析木を得る。そして最後に構文解析木からデータベースへ格納するためのデータ構造を作り出す。図1に入力系サブシステムの処理の流れを示す。

本報告で述べるのは、図1において入力データの解析からデータベースへの格納まで（破線で囲まれた部分）の一連の処理方法についてである。

3　入力データの解析

入力データを解析して、そのデータのなかにある構造を抽出するためには、対象データの構造に関する情報を前もって表現しておく必要がある。本方式では、言語理論の文法を使うことによって対象データの構造に関する情報を表現する。ここで対象とするのは図2に示されるような雑誌の目次である。



Research Bulletin
of
The National Center for Science Information Systems

March 1985 Volume 2

System

J. Nawada　1　A Support System for Software Development
T. Yamazaki

S. Asano　15　A study on High Speed Packet Switching
Y. Ishai

Library Information

E. Naito　25　Problems of UNIMARC-authority

E. Sagawa　33　Recent Problems in String Indexing Systems

図2　目次の例

以下には図2に示されるような目次の構造を表すための簡単な文法を示す。

開始記号：　S
終端記号：　{hd, mt, au, at, pa, cn}
非終端記号：
　　{AUTHOR, PAPERS, PAPER, BODY, CATEGORY}
生成規則

S → hd BODY	(1)
BODY → CATEGORY	(2)
BODY → CATEGORY BODY	(3)
CATEGORY →cn PAPERS	(4)
PAPERS → PAPER	(5)
PAPERS → PAPER PAPERS	(6)
PAPER → AUTHORS pa at	(7)
AUTHORS → au	(8)
AUTHORS → au AUTHORS	(9)

図3　目次の文法の例

ここで各終端記号の意味は、hd（header：目次ヘッダ）、au（author：著者名）、at(article title：論文タイトル）、pa（page：掲載ページ）、cn（category name：カテゴリ名）である。目次のヘッダには普通タイトルや発刊年月日などの情報が含まれているため非終端記号として扱うべきであるが、ここでは議論の便宜上、終端記号として扱う。

非終端記号の意味は、BODY（目次本体）、CATEGORY（カテゴリ名とそのカテゴリに含まれる論文の集合）、PAPERS（論文の集合）、PAPER（論文）、AUTHORS（著者の集合）である。

生成規則は、まず、(8), (9)の再帰的な規則で著者の集合を表わしている。(7)は、論文が著者の集合、掲載ページ、論文タイトルより構成されることを表わしている。(5)と(6)の再帰的な規則は著者の場合と同様に論文の集合を表わし, (4)の規則によって論文の集合とカテゴリ名でカテゴリが構成されることを示している。最後に(2)と(3)の規則によって目次本体はカテゴリの集合であること

が定義されている。

ここで示される文法は、図2に示すような目次イメージを解析するための文法である。この文法にさらに規則を付け加えることによって、より複雑な目次も解析することができる。

構文解析については、自然言語処理の分野で多くの方法が報告されているので、ここでは例を示すのにとどめる。議論の便宜上、図2の目次において最初の論文の部分の構文解析の結果を図4に示す。図4の構文解析木は、上記の文法の(7),(8),(9)の規則を適用することによって作られる。

4　データベースの構造への変換

構文解析の結果得られる解析木は木構造になっており、階層モデルやネットワークモデルのスキーマ表現と非常に似た構造表現になっている。もともと文法は文の構造に関する知識を表現しているものであるから、構文解析の結果得られる構造をデータベースに格納するデータの構造に反映することは自然であると考えられる。本章では、文法からデータベースのスキーマを作成する方法と構文解析の結果からデータベースに格納するデータ構造を作成する方法を示す。

4．1　再帰の除去

文法では、集合に関する情報を再帰を使って表わすことになる。例えば、図3に示した文法において、著者の集合（規則(8),(9)）、論文の集合（規則(5),(6)）カテゴリの集合（規則(2),(3)）などは再帰を使って表わされている。この規則の適用の結果は、図4の構文解析木にもあるように繰り返し構造を生成する。しかし、データベースでは、繰り返し構造よりも集合のほうが扱いやすいことが多い。そこでここでは、再帰的な規則から集合を表わす規則への変換方法について述べる。以下では終端記号や非終端記号の右肩に＊をつけることによって集合を表わすことにする。例えばPAPER*は論文の集合をau*は著者の集合を表わす。また（$A_1 A_2$）*のように複数の終端記号や非終端記号に＊を付けた場合は、終端記号や非終端記号の組の集合を表すこととする。以下に再帰的な規則から集合を表わす規則に変換するための方法を示す。

（1）再帰的な規則を見つける

規則中に以下の条件を満たす規則の組を見つける。

$L \rightarrow R_1 R_2 \cdots R_n L$　　①

$L \rightarrow R_1 R_2 \cdots R_n$　　②

ここで、Lは非終端記号を，またR_k（$1 \leqq k \leqq n$）は終端記号または非終端記号を表している。

（2）集合を表わす規則に変換する

（1）の規則①②をL→（$R_1 R_2 \cdots R_n$）*にまとめる。

上の変換方法は右再帰の規則のみを変換するが、左再帰（$L \rightarrow L\ R_1 R_2 \cdots R_n$）についても同様の変換を施す。

この変換を図3の文法の生成規則に適用すると以下の規則が作られる。アンダラインを付与した規則が変換されたものを示している。

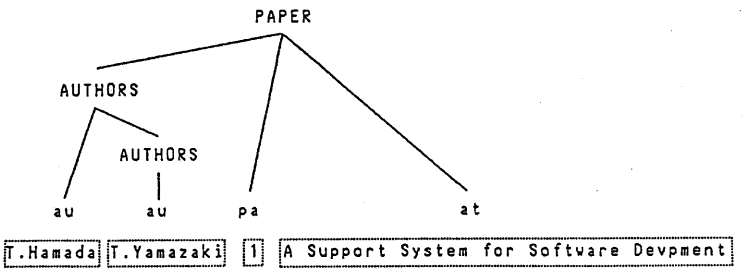


図4　目次データの構文解析

S → hd BODY (1)

BODY → CATEGORY* (2)

CATEGORY →cn PAPERS (3)

PAPERS → PAPER* (4)

PAPER → AUTHORS pa at (5)

AUTHORS → au* (6)

図5　集合を使用した生成規則

この変換は構文解析木の上では、繰り返し構造から集合への変換となる。図4の構文解析木は図6の構造に変換される。

4．2　スキーマへのマッピング

文法は対象データの構造的特質を記述することによって言語のクラスを定義している。そこで、ここでは文法に記述されている構造的特質を使ってデータベーススキーマを作る方法について述べる。なおここで示す方法は、文法が文脈自由文法で記述されていることを仮定している。

構文解析は、文を分解してその構成要素と構成要素間の関係を引き出す処理と考えることができる。この処理では、終端記号は基本構成要素を表す。また、非終端記号はその基本構成要素との組合せにより複合的な構成要素をなす。そこで、データベーススキーマへのマッピングでは、まず、終端記号を実体のクラスに、非終端記号を実体を組み合わせて構成される複合的な実体のクラスに対応づける。すると生成規則は、複合的な実体の構造を記述していると見ることができる。

本方式では実体のクラスを抽象データ型で表すことにする。実体のクラスを単なるデータ型ではなく抽象データ型としたのは、検索の段階でデータの選択条件の柔軟な記述を可能にするためである。例えばタイトルは複数の単語から構成されるため、検索の場合にはタイトル中の単語に注目して検索することが考えられる。従ってタイトルの抽象データ型には文字列の部分列マッチングを行う手続きを付加する。一般に、ここで使用する抽象データ型には、実体の選択条件を処理するための手続きが付加されているものとする。特にマルチメディアデータに対しては、画一的な選択条件の処理では対応することが難しい。そこで、実体のクラスごとに選択処理方法を定義することを可能にし、それを柔軟なシステム構築に役立てる。

一方、複合的な実体は、その実体を構成する要素へのポインタの組で表す。すると複合的な実体のクラスは、ポインタの組の集合つまりポインタによって構成される関係によって表される。複合的な実体の構造は生成規則によって表されている。そこで、各規則L → $R_1R_2\cdots R_n$に対応してn項関係を作成する。第k項は、R_kが終端記号の場合はR_kに対応する抽象データ型のインスタンスへのポインタとし、R_kが非終端記号の場合はR_kを表わす関係のタプルへのポインタとし、R_kが集合の場合はポインタの集合とする。関係は非終端記号に対応して作成されるのではなく、生成規則

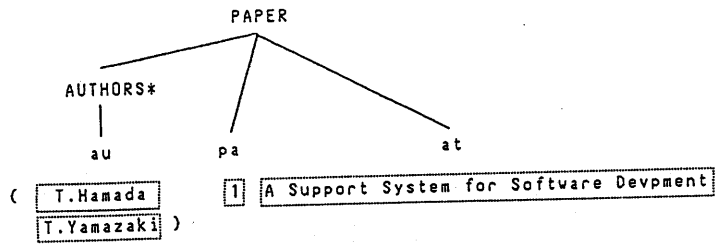


図6　集合規則を使用した構文解析

に対応して作成されることに注意してほしい。そのため、１つの非終端記号が複数の関係によって表現されることもある。この理由については４．３節で述べる。

このように変換することによって文法は、抽象データ型とポインタを構成要素とする関係によって作られるスキーマにマッピングされる。例えば、図５の文法では、６つの抽象データ型と６つの関係で構成されるスキーマが作られる。

構文解析によって作られたデータは、以下の手続きに従ってデータベースに格納される。

・構文解析木のリーフは、対応する抽象データ型の集合に格納する。

・構文解析木の節は、その節を展開する規則に対応する関係に格納する。関係の値はその子ノードに対応する関係のタプルや抽象データ型へのポインタとなる。図６の構文解析木に対応するデータベース内のデータを図７に示す。

ここで述べたスキーマは概念レベルでのスキーマであり、実装にあたっては関係モデルや階層モデル、オブジェクト指向モデルなどに変換する必要がある。

４．３　構造の変種

図５に示される文法では各非終端記号の展開の規則は１つしかないため、構文解析の結果は同一の構造となる。しかし実際のデータには、構造が微妙に変形したものが多く存在する。例えば、目次の場合では掲載論文にシーケンシャル番号が割り振られている雑誌もある。このような雑誌のデータでは、図６と図８を比較すればわかるようにデータ構造が微妙に変化している。本方式では、このような構造の変種を容易に扱うことができる。

以下例を用いて構造の変種への対応を示す。まず文法を次のように修正する。終端記号としてsn（sequencial number）を、生成規則として

PAPER → sn AUTHORS pa at　(5')

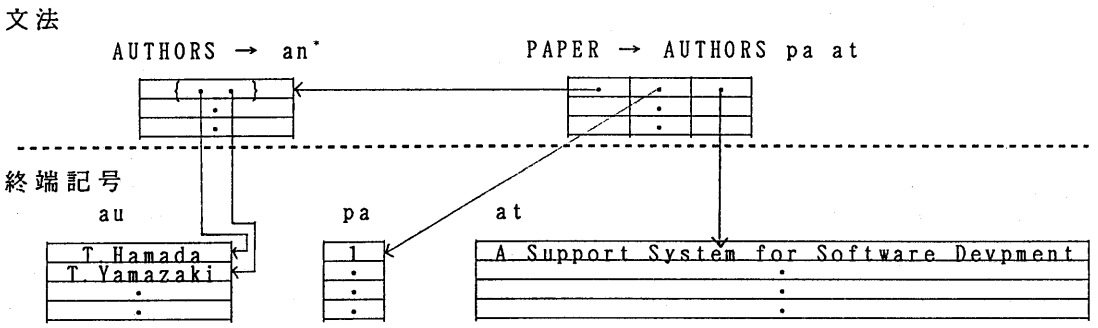


図７　データベース内のデータ

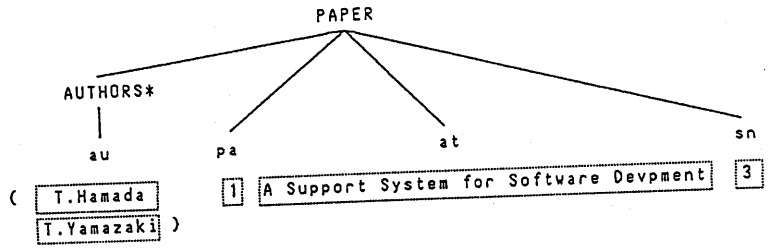


図８　構造の変種

を追加する必要がある。このように修正した文法を用いて、シーケンシャル番号のついた目次を解析した結果が図8に示された構造である。

一方データベーススキーマは、（5'）の規則を追加したことによって、非終端記号PAPERに対応して2種類の関係が作られる。4．2で複合的な実体に対応する関係を非終端記号ではなく生成規則に対応して作ったのは、このような構造の変種に対応できるようにするためである。

4．4　検索におけるユーザビューについて

本方式で作られるデータベーススキーマは、ある意味でデータベース作成者のビューと位置づけることができる。作成者のビューがデータベースの内部の構造に反映されるのは従来の方法と同様である。ここでは、マンマシンインタフェースとユーザビューについて簡単に触れる。

筆者らは、ユーザビューとして目次そのもの、つまり目次の画像データを表示システムとして提供することを考えている。一方本報告で述べた方法は、基本的には、目次を文字データに変換して蓄積する方法である（ロゴや写真などは画像データとして格納され、画像データ用の抽象データ型の操作を使って操作される）。そのため、本方式で作成される文字データに画像データを対応づけ、検索処理には文字データを使用し、ユーザへは画像データを提供できるようにしなければならない。従ってデータベーススキーマを拡張して画像データも格納できるようにする必要がある。画像データを追加する方法はいくつか考えられるが、これは実装上の問題であり、ここではこれ以上議論しない。

目次全体の画像データをユーザビューとした場合、ユーザは目次上の任意の実体（文法の終端記号に対応する）に対して条件を記述することになる。これはデータベースにおいては、まず抽象データ型のクラスに対して選択を行い、条件にあったインスタンスからポインタを逆に辿って関連する実体を求めるといった処理を行う必要がある。ポインタを逆に辿るためのメカニズムについても実装の問題であるためここではこれ以上議論しない。検索系と関連したこれらの問題については、物理スキーマの作成方法として他の機会に報告する。

5　おわりに

本報告では、まず目次データを自動的にデータベースに入力することを目的とし、入力データの解析に文法を利用することを述べた。さらに、その記述された文法からデータベーススキーマを生成し、解析した入力データをそのスキーマに変換する方法について検討した。この方法によって、データベース入力の自動化への展望が開ける。また、これまでのデータベースモデルでは構造の変種を統一的に扱うことが難しかったが、本報告の方法では容易に構造の変種に対応できることを示した。

今後は以下の課題について取り組みたい。

・物理スキーマの実現方法

本方式で示したスキーマは概念的なものであり、ユーザビューや処理効率などを考慮した物理的なスキーマを作成する方法について検討していない。

・マルチメディアデータの本格的な扱い

本報告では、画像や音声などのマルチメディアデータに対する操作を詳しく検討しなかった。現在パターン認識の分野で画像データのパターンを文法によって記述することが試みられており、画像データに対する文法の記述が有効であることが示されている[4]。今後は画像や音声などのマルチメディアデータの構造解析にも文法を適用することによって、マルチメディアデータを本格的に扱うことを試みる。

・目次に関する文法

本報告では、目次の文法とその解析方法については詳述せず、本方式を明らかにするための例示にとどめた。現在、目次イメージの内容（書誌的属性）に及ぶ文法記述どそれに基づく解析方法について基礎的な検討の途上にある。

［参考文献］
[1]E.Katsura et.al, "An Approach to Electronic Contents Services Based on a Multimedia Document System", Proc. of 14th Int. Online Information Meeting 1990
[2]原他、「目次イメージのセグメント化と文字認識」、第４回人工知能学会全国大会論文集、1990
[3]J.Hopcroft, J.Ullman 「オートマトン 言語理論 計算論Ⅰ・Ⅱ」 野崎他訳 共立出版
[4]H.Bunke, A, Sanfeliu, "Syntactic and Structural Pattern Recognition - Theory and Applications", World Scientific 1990