4S−3

文法記述によるデータベース入力支援法

高須 淳宏、佐藤 真一、 桂 英史
学術情報センター

1 はじめに

本稿は、学術雑誌の目次画像からデータベースを構築する方法について述べる。文書処理では、その画像の領域分割、各領域の分類、テキスト領域のOCRによるコード化等の処理が行なわれ[1)]。さらに、見出しや本体等領域間の関係を考慮して、文書の論理構造の抽出が行なわれる[3)]。文書から論理構造を抽出する従来の研究は、読者から見た自然な構造を抽出することを目的としており、論理構造の基準が明確でない。そのため、抽出される構造は、抽出法や使用されるヒューリスティクスに依存することがある。本稿では、論理構造をデータベースのスキーマとし、スキーマに基づいた目次画像の論理構造を文法的に記述しデータベースを作成する方法について述べる。

2 目次の論理構造

普通目次は、雑誌タイトルや巻、号などの雑誌自身に関する情報とその雑誌に含まれる文献の情報を含んでいる。各文献は、さらに、著者の集合、文献タイトルなどから構成される。このように、目次のデータは階層的な構造を持っている。そこで、本稿では、nested relationによって、目次データを管理すると仮定する。nested relationのスキーマは、階層構造で表される[2)]。例えば、目次のデータベーススキーマは、

雑誌タイトル、巻、
((著者), 文献タイトル, 掲載ページ)) (1)

となる。ここで、括弧は、その中にあるスキーマよりなる部分関係を表している。例えば、"((著者),文献タイトル、掲載ページ)"は、雑誌に含まれる文献の集合を表す部分関係である。この部分関係には、さらに各文献の著者の集合を表す部分関係のスキーマ"(著者)"が含まれている。

A Database Construction Support Method Based on the Syntactical Description
Atsuhiro TAKASU, Shin'ichi SATOH and Eishi KATSURA
National Center for Science Information Systems

nested relationのタプルは、

1. 各属性値を属性名に置き換える
2. 部分関係 (タプルの集合) をタプルの列に展開する
3. 2. を部分関係に再帰的に適用する

ことによって、属性名をトークンとする文に展開できる。一方、スキーマは、部分関係を正則文法の閉包に置き換えることによって正則表現に変換できる。例えば、(1)式のスキーマは、以下の正則表現に変換される。

雑誌タイトル、巻、
(著者*, 文献タイトル, 掲載ページ)* (2)

この文法は、(1)のスキーマを持つ任意のタプルを受理する正則文法を表している。従って、属性名によって構成される文が与えられた場合、スキーマから得られる文法を使用して文を解析することによりデータベースのタプルが生成できる。

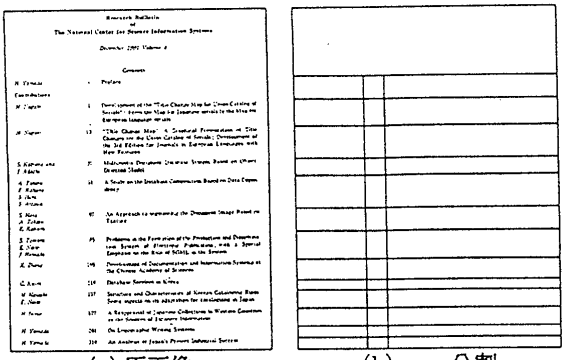
3 レイアウト情報

目次はレイアウトの自由度が高いため、データがスキーマから作成される文法の順序に配置されているとはかぎらない。そこで、文法にレイアウト情報を付加することによって、目次画像上にあるデータを文法の順序に並び変えることを考える。

一般に文書は、直線、段組み、スペースなどによっていくつかの領域に分割されている。そこで、以下に示すareaによる領域分割を考える。

1. ページはareaである。
2. areaを水平または垂直の分割線で2分割できる場合、分割された2つの領域はareaである。

ここで、分割線は、直線または背景色の直線を意味する。areaは、ページから始まり、段落、行、単語など様々なレベルの分割が考えられる。適当なレベルでarea分割を行なうと、areaの並び換えによって2章



(a) 原画像　　(b)area 分割

図 1: 目次ページの area 分割

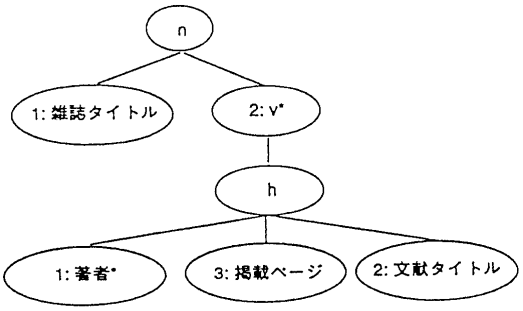


図 2: 文法の木

で述べた文法で受理可能なデータ列の作成が可能になる。例えば、図 1(a) の目次は、図 1(b) のように分割することによって、式 (2) の正則表現で受理可能なデータを作成できる。

そこで、上記の area 分割における、各 area の相対位置を文法に導入する。area 分割は、その分割方法から明らかなように木構造で表すことができる。そこで、各 area の相互関係を以下のノードによって構成される木で表現する。この木で、任意の部分木は、ページの部分領域を表している。

1. 内部ノード　内部ノードには、h, v, h^*, v^*のいずれかのラベルが付けられる。$h(v)$ は、子領域が、水平 (垂直) 方向に並んでいることを表す。$h^*(v^*)$ は、その子領域が、水平 (垂直) 方向に繰り返し現われることを表す。

2. 葉　葉には、スキーマから生成される正則表現の部分正則表現及び葉の順序が付けられる。

図 1の目次に対する文法の木を図 2に示す。同一雑誌の場合、目次の各データは、およそ同一の相対位置にレイアウトされているため、上記の文法の木は、同一雑誌の目次には、適用可能であるものと思われる。目次画像は、まず、文法の木と照合するまで、area 分割される。次に、各 area の内部がデータベースの属性値に相当するブロックに分割される。さらに、各ブロックは分類され、データベースの属性名が割り当てられる。最後に各 area に割り当てられた文法を利用して、解析され、データベースのデータが作成される。この手続きの詳細は、別の機会に報告する。

4　おわりに

本稿では、雑誌の目次画像からデータベースを作成するための方法を目次の論理構造及びレイアウト情報の記述方法を中心に述べた。本稿で述べた方法では、1) 目次の論理構造は nested relation のスキーマによって記述でき、2)area 分割によって目次の論理的データ領域を分割できる、ことを仮定している。nested relation の階層的なデータ構造は、目次に限らず広く一般の文書にも適用可能なものである。また、area 分割もレイアウトによっては、分割できない場合も存在するが、傾きやノイズのない目次では有効であるものと思われる。目次の文法は、基本的にはデータベーススキーマに即したものであるが、レイアウト情報は、多分に area 分割方法に依存するため、文法の木を作成する場合、なんらかの文法作成支援システムが必要になる。今後、レイアウト情報の学習方法について検討する予定である。

参 考 文 献

1) M. Nadler. "A Survey of Document Segmentation and Coding Techniques". *Computer Vision, Graphics, and Image Processing*, Vol. 28, No. 240-262, 1984.

2) H. J. Schek and M. H. Scholl. "The Two Roles of Nested Relations in the DASDBS Project". In *Lecture Notes in Computer Science - Nested Relations and Complex Objects in Databases*, pp. 50–68. Springer-Verlag, 1987.

3) S. Tsujimoto and H. Asada. "Understanding Multi-articled Documents". In *Proc. of 10th ICPR*, pp. 551–556, 1990.