4E－1

多様な構成を持つ文書を対象とする全文データベースの処理方式

酒井 乃里子　高須 淳宏　安達 淳
東京大学工学部　学術情報センター研究開発部

1 はじめに

科学技術論文を対象とする全文データベースでは、文書の記述に SGML (Standard Generalized Markup Language) を用いることが一般的となっている。SGML は文書を構成する論理要素の決め方を定義するものなので、実際の構成は文書によって様々である。

文書を記述する際の自由度を保持し、個々の文書の論理構成を尊重するためには、各々の DTD による設定は保持しなければならないが、一方データベースとしては、多様な文書を多量に扱うために、統一した処理ができなくてはならない。

以上の問題点を踏まえて、本稿では、各々の文書の固有の論理構成を残したまま、データベースとしてまとめて扱うための手法を提案し、その第一段階として、書誌情報に適用して実際に作成したプロトタイプを紹介する。

2 全文データベースシステムの概要

目指す全文データベースでは、データとなる文書は固有の DTD に基づいて記述されたまま蓄え、ユーザが指定した要素を適宜抽出して、表示や検索などの処理を行う。

また、従来のデータベースにある機能のみならず、文書を画面上で講読し、条件を満たす結果を複数文献に渡ってまとめて表示し、検索結果に基づいてさらに検索を進めるなど、ユーザの論文利用を助ける機能が様々に考えられる。

今回はこれらの機能の核となる、個別の文書の該当する要素を特定し、文法的に処理してデータを抽出する機構を提案する。現在は論文文書のうち、書誌情報を対象としている。この機構の概要は次の通りである。

まずデータベース側で、ユーザに提示する文書の構成を、表示項目として設定する。また DTD ごとに、各項目に対応する要素を定める。ユーザの処理要求や項目の指定に応じて、必要な要素が前述の対応から定まるので、字句解析によりその要素が特定される。この字句解析結果に基づき構文解析を行って、データを抽出する（図 1)。

3 要素を特定する処理

3.1 データベース側の設定

データベース側では、ユーザに提示する標準的なビューとして、表示項目を定める。表示項目は、ユーザが処理要求を出す際の意識に近づけることを基準として設ける。今回は、書誌情報について、著者名・著者所属・タイトル・論文誌名・発行年月・巻号・発行者・ページ・ISSN・キーワード・要旨の 11 の項目を設けた。

3.2 文書データの前処理

SGML では、構成要素をタグで区切る方法で文書を記述する（図 2)。文書の論理構成は DTD ごとに異るので、表示項目に対応させるべき要素も一般に DTD ごとに異る。そこで、予め DTD ごとに表示項目と要素との対応を指定する（図 3)。

```
<論文><論文情報><英題>Co-authoring ofscholarly
papers</英題><英副題>A comparative study on
Japanese and Western papers.</英副題><著者情報>
<Fネーム>Masamitsu</Fネーム><Lネーム>NEGISHI
</Lネーム><英所属>National Center for Science
Information Systems</英所属></著者情報><著者情報>
<Fネーム>Hisao</Fネーム><Lネーム>YAMADA</Lネー
ム><英所属>National Center for Science Information
Systems</英所属></著者情報><英キーワード>Co-authored
papers</英キーワード></論文情報><英要旨><段落>The
number of co-authors for </段落></英要旨>
```

図 2: SGML 文書（部分）

今回のプロトタイプでは、要素の指定を絶対位置（文書の論理構成のトップから見た位置）・相対位置（任意の地点から見た位置）による表現、及び中間を省略するワイルドカード表現の三通りを採用して、柔軟な指定ができるようにした。また、複数の要素を組み合わせて一つの項目とする場合には、DTD においてそれらの要素を含む上位のものを中間の要素として設定する。

A Processing Method for Full-text Database of Documents with Various Structures
Noriko SAKAI[1], Atsuhiro TAKASU[2], Jun ADACHI[2]
[1]Faculty of Engineering, The University of Tokyo
[2]Research & Development Department, National Center for Science Information Systems

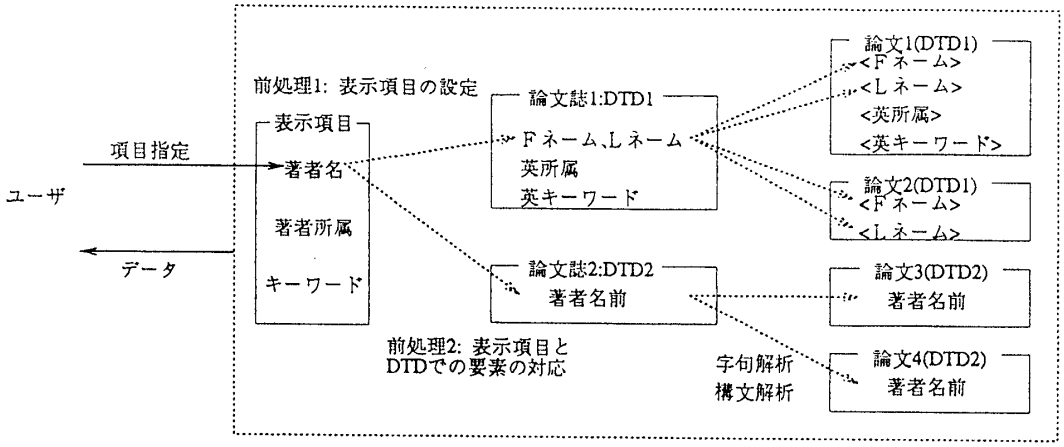

図 1: 多様な DTD による文書からのデータ抽出機構

標準側項目名 (非終端記号名)	各文書の該当する要素(例) (トークン名)
著者名 (注1) (Xauth)	/論文/論文情報/著者情報/Fネーム (Yfnam) /論文/論文情報/著者情報/Lネーム (Ysurn)
著者所属 (注1) (Xaffn)	/論文/論文情報/著者情報/英所属 (Yaffn)
キーワード (Xkywd)	/論文/論文情報/英キーワード (Ykywd)

(注1)これらを含む上位の中間要素Yauthを用いる

図 3: 項目と要素の対応づけ

```
Xauth   :  Yauth
        |  Xauth   Yauth
        ;
Yauth   :  Yfnam   Ysurn   Yaffn
           {
             set_name($1, $2);  /* 著者名を表示用に整形する */
             set_multi($3);  /* 複数回現れるものを表示用に整形する */
           }
        ;
Xkywd   :  Ykywd
           {
             set_multi($1);
           }
        |  Xkywd   Ykywd
           {
             set_multi($2);
           }
        ;
```

図 4: yacc の仕様書

```
著者名    : NEGISHI, M., YAMADA, H.
著者所属 : National Center for Science
Information Systems, National Center for ...
タイトル : Co-authoring of scholarly papers
(A comparative study on Japanese and ...
キーワード: Co-authored papers, Abstracting ...
```

図 5: 処理結果

3.3 データの抽出

まず、文書に対してシーケンシャルに字句解析を行い、前処理で用意した対応表に基づいて、ユーザの指定に必要な要素をトークンとして抽出する。

次に字句解析の結果に対して構文解析を行う。これにはyaccを用いている。データベース側の表示項目を非終端記号として、トークンとの文法的な関係をyaccの仕様書に記述する（図4)。yaccを用いることにより、複数回現れるものや複数要素を組み合わせる対応にも対処できる。構文解析で文法的に受理されると、データが出力される。このデータは表示用に整形する(図5)。

データの整形は、yaccのアクションルーチンで行う。このアクションでは、頻出する処理を汎用ライブラリとして前もって用意しておき、特別な操作を必要としない処理（複数回現れるものは「, 」で区切って並べる、など）は引数にデータを与えるだけで処理できるような環境になっている。

処理結果については、このデータベース固有のDTDを調え、そのタグを付けて出力すれば、結果をより見やすい形に表示することができる。

4 まとめと今後の課題

異るDTDにより記述された文書を、データベースとして多量にかつ画一的に扱って、指定されたデータを抽出する手法について提案し、また書誌情報に適用して作成したプロトタイプを紹介した。

この手法の課題は、yaccの仕様書の記述である。現在は人手により、DTDにおける文書の論理要素とデータベース側の表示項目を意味的に照らし合わせながら作成されているが、これを、DTDでの記述に基づいてある程度自動的に生成することが実用上求められる。今後はこの点について検討したい。

参考文献

[1] Warmer, J. and van Vliet, H., "Processing SGML Documents," ELECTRONIC PUBLISHING, Vol.4, No.1(MAR. 1991).