文書型定義の統合に基づく構造化文書データベース

1W－6　市川 修　吉川 正俊　植村 俊亮

奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

SGML のような構造化文書では，文書中にその構造に関する情報を記述しているために，従来の内容に基づく検索に加えて，文書の論理的な構造に基づく検索処理を行うことが可能になる．

ここでは，SGML 文書および DTD をプレーンテキストとしてデータベースに格納し，SGML によってタグを付けられた各々の文書に対して，DTD グラフをとおして直接問合せ言語によって検索を行う手段について考える．

2 データベースの概要

SGML のような構造化された文書を扱うために，DTD をオブジェクト指向データーベースのクラス定義に変換した例 [1] もある．しかしこの方法では，異なる型の DTD に対しては，データベーススキーマを定義し直す必要があるために，複数の DTD からなる文書を扱うためには適切ではない．

異なる DTD に対応する構造化文書を統一的に扱うために，作成された SGML 文書 · DTD をそれぞれ一つのクラスに格納する．データベースは，以下のものから構成される．

- SGML 文書：タグ付けされた SGML 文書をそのまま一つのインスタンスとして格納する．
- DTD：各々の DTD (図 1) をそれぞれ，インスタンスとしてそのまま格納する．

```
<!ELEMENT article   [
<!ELEMENT article   --  (title, author+, abstract, section+) >
<!ELEMENT section   --  (title, p*, subsec*) ) >
<!ELEMENT subsec    --  (title, p+, subsec*) >
<!ELEMENT title     --  (#PCDATA) >
<!ELEMENT author    --  (#PCDATA) >
<!ELEMENT abstract  --  (#PCDATA) >
<!ELEMENT p         --  (#PCDATA) >
]>
```

図 1: 文書型定義の例

3 問合せ

DTD 中に保持されている構造情報を利用した問合せの表現および処理のために，DTD に対して，各々のエレメントをノードとして表した有向グラフ (DTD グラフと呼ぶ．：図 2) を作成する．

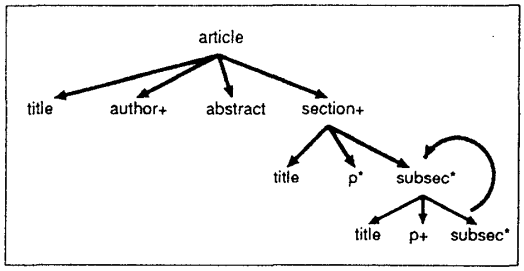

図 2: DTD グラフの例

DTD グラフにおいて，あるエレメント A から，有向枝を辿って到達できるエレメントを，A の下位のエレメントと定義する．逆に，DTD グラフにおいて，あるエレメント X から，有向枝を遡って到達できるエレメントを X の上位のエレメントと定義する．

3.1 構造化文書 (SGML) のための問合せ言語

問合せを行うためには，データベースに格納された SGML 文書に対して直接 SQL を適用する．この SQL の概要は次のとおりである．

- FROM 句には DTD で定義したエレメント名を記述し，その文書の検索される範囲を限定する．SELECT 句，および WHERE 句には FROM 句で記述されたエレメントより下位のエレメントが記述されなければならない．
- 問合せに使用されるエレメント名は，曖昧さがあるものについてはドット記法を用いて記述することが必要である．特に指定がない場合は，FROM 句で定義されたエレメント e の下位のエレメント中で e に最も近いエレメントに対してアクセスされる．
- 繰り返しを伴うエレメントに関しては，何番目のエレメントを検索対象とするかを指定できる．

実際の問合せの例を以下に示す．

Q1. アブストラクトの中に 'SGML' と 'HyTime' を含む <article> の <title> および第一 <author> を見つけよ．

```
SELECT  <title>, <author>[1]
FROM    <article> a
WHERE   <abstruct> contains
            'SGML' and 'HyTime'
```

Q1. のように，繰り返し記号 ("*" および "+") によって定義されるエレメントについては，繰り返された回数を指定することで，繰り返しを伴う問合せも可能となる．

A Database for Structured Documents Based on the Integration of Document Type Definitions
Osamu ICHIKAWA, Masatoshi YOSHIKAWA and Shunsuke UEMURA
Graduate School of Information Science, NAra Institute of Science and Technology (NAIST)

Q2. <article> の中に含まれるエレメントを見つけよ.

```
SELECT  elements
FROM    <article>
```

Q2. において "elements" は，DTD で表現されるエレメントを表すための属性であり，DTD に対して直接アクセスすることで文書の構造に対する問合せも可能である.

Q3. 'SGML' を含む <article> の <subsec> の内容を見つけよ.

```
SELECT  ss
FROM    <article> a, a.<section> s,
        s.<subsec> ss
WHERE   ss contains 'SGML'
```

Q3. のように DTD で中間構造を表すエレメントが，SELECT 句で定義された場合，その下位のエレメントで表される実際のインスタンス (#PCDATA 部) を表すと考える．また FROM 句の s.<subsec> は，エレメント section の中で，subsec を持つものが暗黙のうちに選択されたことを示している[1].

Q4. <subsec> を持つ <section> の <title> を見つけよ.

```
SELECT  s.<title>
FROM    <section> s
WHERE   s contains <subsec>
```

Q4. では SGML 文書中に記述されているタグをテキストの一部とみなしているために，文字列検索と同じように検索することが可能である.

3.2 DTD による問合せの検査

一般に，一つの DTD に対して，それに従った複数の SGML 文書が存在する．そのために，DTD において再帰的に構造が定義されている場合などは，同じ DTD から作成された各々の文書でも，その論理的な構造が異なる場合も存在する．そこで，ある問合せが与えられた時に，ある DTD から作成された複数の文書インスタンスが，その問合せに対して適切であるのかどうかを DTD および SGML 文書から判断し，検索のための文書の範囲を限定する.

DTD(グラフ) と問合せ文を照合することで，ある程度の文書の絞り込みが可能となる．問合せに用いられるエレメントと DTD との関連については，

(1) 問合せに用いられるエレメントが DTD の中にすべて存在する.
(2) FROM 句に 複数のエレメントが定義され，それらがドット表現で繋がる場合，それらを一つのエレメントと同様に扱う．検索範囲は最も上位にあるエレメントに対する範囲と同様である.
(3) 問合せの FROM 句に定義されたエレメントよりも下位でないエレメントが SELECT 句，あるいは WHERE 句で定義されている.
(4) 問合せ文中のドット表記と DTD での上位・下位の関係が一致する.
(5) 問合せ文中で繰り返しを表現している elements[n] に対して，DTD 中の対応するエレメントが繰り返しを表現できる.

の 5 項目について判断することで．DTD から作成される SGML 文書について以下の 3 種類に分類することができる.

i) **DTD から作成される SGML 文書のすべてが適切でない場合**
問合せに用いられたエレメントが DTD に存在するかどうかを判断することであるため，DTD グラフ中の "elements*" を "elements+" とし，"elements?" を "element" と考える．前述した (1)〜(5) に関して DTD グラフを通して検査を行い，どれか一つでも満たさないものがあれば，問合せに使用された DTD から作成される SGML 文書のすべては適切でない.

ii) **DTD から作成される SGML 文書のすべてが適切な場合**
DTD から作成されるすべての SGML 文書を考えなければいけないために，DTD グラフ中のエレメントは，DTD で定義されたものをそのまま用いる．i) と同様に (1)〜(5) に対して検査を行い，条件をすべて満たしていれば，問合せに使用された DTD から作成される SGML 文書のすべてが適切である.

iii) **DTD から作成される SGML 文書の一部が適切な場合**
DTD の中で，"|" が使用されていたり，構造が入れ子で定義されているために，問合せに対して適切な文書とそうでない文書が生じる．ii) と同様の検査を行うが，DTD グラフからだけでは，完全に判断することができなく，実際の SGML 文書に対して検査しなければならない．ただし，選択記号で定義されるエレメントの内の片一方に特有のエレメントが存在する場合や，ネストされた回数についての情報があれば，それらの情報を基にして，ある程度文書の範囲を制限することが可能となる.

4 おわりに

今後は，問合せ能力の拡張および我々のデータベースに合ったインデックスに関して検討を行う.

謝辞：日頃から数々の有益な御討論を頂く植村研究室の皆様に感謝いたします.

参考文献

[1] V. Christophides, S. Abiteboul, S. Cluet and M. Scholl : From Structured Documents to Novel Query Facilities. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pp.313 - 324, May 1994