半構造化データモデルを利用したXML文書管理システムの試作

5V－5

北野 拓哉　波内 みさ

NEC C&Cメディア研究所

1 はじめに

WWW上の新しいデータ交換形式としてXMLが注目されている．XMLはHTMLのようなWWW上のデータ交換とハイパーテキスト/リンク機能を持ち，またSGMLのような文書型定義(DTD)による要素タグの定義を可能としている．このようなWWW上のXML文書をDTDに基づいてデータベースで管理しようとする場合，DTDごとにスキーマを構築して管理する方法は適さない．XMLユーザは独自の要素タグを定義し利用するため，様々なDTDがXML文書とともにWWW上に発信されるだろう．よってXML文書管理には，構造情報もデータの一つとして取り扱う半構造化データモデル[1]を利用し，データとともに構造情報を管理する方法が適すると考える．問合せ系も各DTDの差異を吸収するような柔軟な仕組みが必要となる．

本研究ではXML文書を各DTDに基づいて要素単位に分解し，半構造化データモデルの一つであるツリー構造で管理するクラス群を実装した．検索では，Lorel[1]などのGeneral Path Expressionによる検索式を実行するメソッド群を用意した．上記メソッドを用いてGeneral Path Expressionのあいまいな論理構造指定に基づく検索プログラムを作成し，XML文書のDTDの差異を吸収した検索実行を可能にした．

また本研究では，実験システムとしてXML電子カルテシステムを試作した．DTDのツリー構造に基づいてカルテの各情報を管理し，症例検索などの複雑な検索要求に応える．

2 半構造化データ格納と問合せ

本システムでは，XML文書中の各要素をツリー構造で構成して管理する．例えば「カルテ」に関するXML文書を例に挙げと，図1に示すように最上位の要素「root」の子要素に要素「カルテ」を挿入し，以下カルテの各子要素を左深さ優先順序のツリー構造で格納する．各要素はオブジェクトとしてメンバ変数とメソッドを持つ．例えば図1の葉の部分に位置する各文字列は，その親要素の文書内容を格納するメンバ変数の値に含まれること

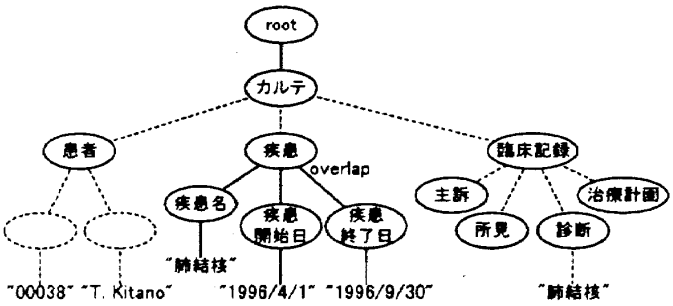

図1　あるカルテに関する要素のツリー構成図

を表しており，また図1の要素「疾患」は，その右下に書かれるメソッド「overlap」を持つことを表している．要素間の線はツリー構造上の親子関係を表し，破線は途中で任意の要素の出現をゆるしていることを意味する．

このように格納されるXML文書に対する検索式を，Lorelなどの General Path Expressionを用いて記述することができる．以下の検索要求を満たす検索式を図2に示す．

「患者"T. Kitano"の疾患"肺結核"に関する"1996/4/1"から"1996/9/30"までの臨床記録を返せ」

```
select CR
from   root.カルテ K, K.#.患者 P, K.#.疾患 D,
       K.#.臨床記録 CR, CR.#.診断 A
where  P.#.%.Like("T. Kitano")
and    D.疾患名 = "肺結核"
and    D.overlap("1996/4/1", "1996/9/30")
and    A.#.%.Like("肺結核")
```

図2　General Path Expressionを用いた検索式の例

General Path Expressionの一つであるワイルドカード#は，#の左の要素(起点要素)から右の要素(目標要素)にいたるツリー上の任意のパス(途中の要素も含む)を表す．図1の点線は図2の#によって表される．同じくワイルドカード%は要素の任意のメンバ変数を表す．また，#.%は任意の要素の任意のメンバ変数を表す．このようなワイルドカードを用いれば，XML文書のDTDが異なっていてもその違いを吸収することができる．

本研究では，図1で示すXML文書を要素単位で格納するクラス群を実装し，図2で示す検索式を実行するプログラムを作成するための検索系API

An XML Document Management System using a Semistructured Data Model,
Takuya Kitano, Misa Namiuchi,
C&C Media Research Laboratories, NEC Corporation.

を提供することを目的とする．

3 XML 文書管理クラスライブラリ

本研究の XML 文書管理クラスは，PERCIO[2]のマルチメディア文書クラスライブラリ(MMDCL)[3]の一部として開発した．図 1 に XML 文書管理クラスの一部のオブジェクト図(OMT 記法)を示す．

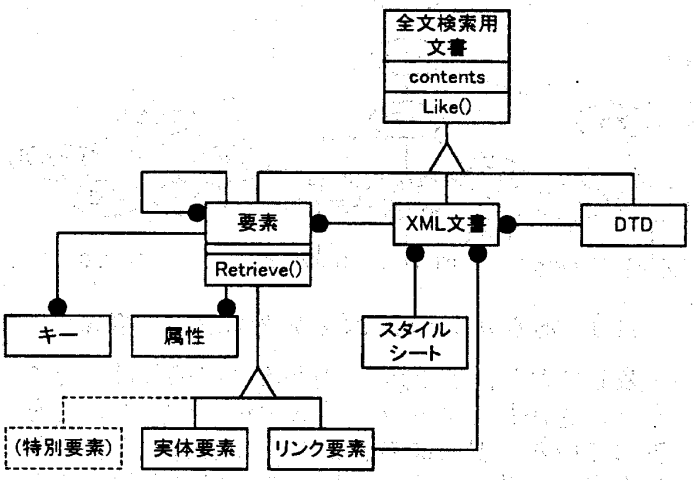

図 3 XML 文書管理クラスのオブジェクト図

XML 文書はクラス「XML 文書」のインスタンスとして格納される．XML 文書中の各要素は分解され，要素間の関係をツリー構造で構成してクラス「要素」のインスタンスとして格納される．図 1 の各要素は，このクラス「要素」のインスタンスとなる．また XML 文書と関係する DTD もクラス「DTD」のインスタンスとして格納される．これらのクラスは，MMDCL のクラス「全文検索用文書」を継承する．各要素の文書内容はメンバ変数「contents」に格納され，その内容をメソッド「Like」を呼び出して全文検索することが可能である．

また本 XML 文書管理クラスでは，特別な要素として実体を有する要素(実体要素)，リンクを有する要素(リンク要素)を管理するクラスをそれぞれ用意した．他にも特別な要素を管理するユーザ定義要素クラスを追加することも可能である．図 1 の要素「疾患」は，メソッド「overlap」を持つユーザ定義要素クラスのインスタンスとして格納される．

ワイルドカードを用いた検索式の実行に関しては，メソッド「Retrieve」を用意した(多重定義も含め数種類ある)．起点要素からメソッド「Retrieve」を呼び出し，引数に目標要素やパスに対する検索条件を指定して検索することが可能である．

4 要素検索高速化手段

ワイルドカード#を用いたツリー上の要素検索では，高速化のために，起点要素から目標要素にいたるパス以外を通ることを避けなければならない．本研究の XML 文書管理クラスでは，図 3 に示す通り DTD に関しても対応する XML 文書と関係づけて管理している．クラス「DTD」はそのインスタンスの DTD の構造を解析して，ある起点要素からある目標要素にいたる要素間のパスを算出する．このパスの情報を基に，本 XML 文書管理クラスの検索系 API は，最小限のツリー上のパス探索で目標要素を捕捉することができる．

5 XML 文書管理クラスの利用例と今後の課題

本 XML 文書管理クラスを用いて，電子カルテ [4]に関する検索システムを試作した(図 4)．カルテの構造に基づく複雑な検索を実行可能にしたこと，また各病院間で用いるカルテの DTD が異なっていても，柔軟にその差異を吸収して相応の検索を実現したことなどが特長である．

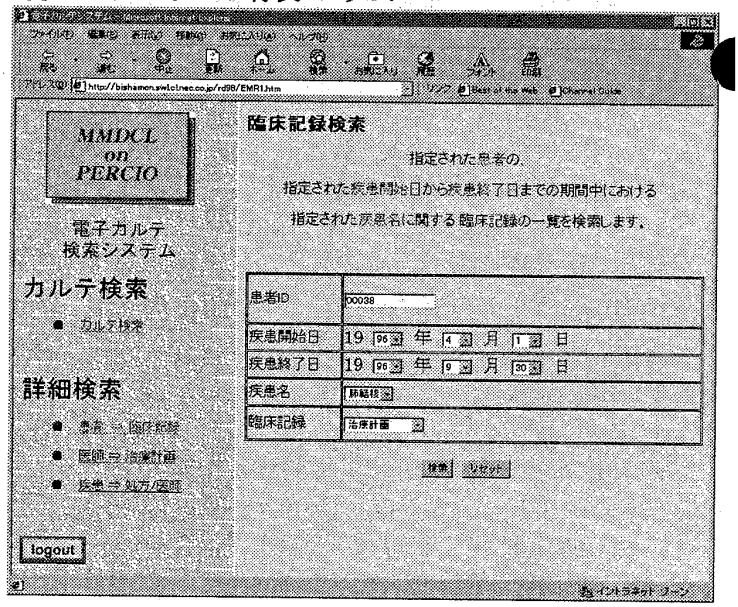

図 4 XML 電子カルテシステム

しかしながら，DTD の差異に柔軟に対応する検索機能の反面，性能面でかなりの犠牲を払っている．インデックス付け，フォルダリング，クラスタリングの技術を用い，また検索アルゴリズム等を改善し，XML 文書検索の更なる高速化を図ることが今後の課題である．

参考文献

[1] Serge Abiteboul et al., "The Lorel Query Language for Semistructured Data", International Journal on Digital Libraries, Vol.1, No.1, pp.68-88, 1997.

[2] NEC, "PERCIO オブジェクト指向データベース管理システム", http://www.ace.comp.nec.co.jp/product/db/percio/p_main.htm, 1996-1997.

[3] Takayuki Saeki, Misa Namiuchi, "Extensible Multimedia Class Library for Object-Oriented Database Systems", International Workshop on Database and Expert Systems Applications, 1998 (to appear).

[4] 羽澄 典宏 他, "電子カルテシステムの全体管理を行う"Medical Manager"の概念設計", 情報処理学会第 57 回全国大会講演論文集, 1998 (予定).