SGML文書データベースへの問合せ言語としてのHyQ

1W－5

今郷　詔

(株) リコー　情報通信研究所

1　はじめに

書誌情報だけでなく文書全体をDB(database)に格納する全文DBシステムが広まってきている。単なる文字列として文書内容をDBに格納するのではなく、文書内容を構造化して格納することによって多様な検索・処理が可能となる。

文書構造を記述する手段としてはSGML (ISO 8879, JIS X4151) が普及しており、アプリケーションから独立した文書表現ができる。またSGMLはハイパーメディア記述言語HyTime (ISO/IEC 10744, JIS X4155) の基盤でもあるので、単なる文書だけではなくハイパーメディア文書記述への拡張も可能である。

そこで我々はSGMLで記述された文書を対象とするDBシステムを作成した[1]。本稿ではこのシステムの問合せ言語について報告する。

2　SGML文書のデータモデル

SGML文書は図1に示すような木構造に対応する。木のノードには2種類あり、文字データを内容とする終端ノードと、他のノードを内容とする非終端ノードがある。

すべての非終端ノードはGI (generic identifier) を一つだけ持っており、GIによってそのノードのクラスが特定される。非終端ノードはその内容とは別に、それぞれのクラス毎に定められた属性 (attributes) を持つ。もしID属性を持っていれば、その値でノードを特定できる。

3　SGML文書DBの問合せ言語

文書DBの作成に用いるDBMSが提供するSQLのような問合せ言語でも、文書DBへの問合せを行なうことができる。しかし、このような言語で問合せを行なうには、文書のスキーマを知らないと問合せを書くことができない。

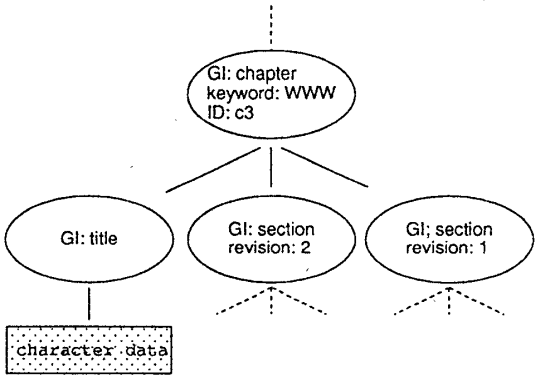

図1: データモデル

また、文書中の情報から導出する目次のような二次的情報は、静的な表現では文書を変更した場合に不整合が起こる恐れがあるので、文書内容に対する問合せとして表現し、文書の利用時に実際の内容を動的に生成するべきである。文書の交換性を高めるために、このような問合せはDBMSに依存しない記述が望ましい。

したがってSGML文書DBへの問合せ言語として、SGMLの知識があれば記述でき、しかもアプリケーションおよびDBMSから独立した言語が望ましい。SQLを拡張してSGMLの構造記述子をサポートする言語も考えられているが[2]、elementの属性やinclusionをサポートしていないので、SGML文書のすべての情報を利用できるわけではない。

我々はこのような条件に合う言語として、HyTimeが規定しているHyQ[3]を利用した。HyQの本来の目的はHyTime文書のハイパーリンクのアンカーを動的に表現することであるが、HyTime文書はSGML文書でもあるので、HyQはSGML文書の問合せにも使用できる。

HyQを用いることによって、次のような問合せを行なうことができる。

- 文書の構成要素 (ノードや属性値など) の取得
- 指定した文字列にマッチする内容を持つノードの取得

Using HyQ as the query language to an SGML document database.

IMAGO Satosi

RICOH Co.,Ltd.

- ノードの特性 (属性値や GI など) をキーとする検索
- ノード間の関係 (指定した親を持つノードなど) による検索

4 HyQ 処理系

HyQ は lisp に似た構文を持ち、問合せ領域として指定したノードリストから、問合せ内容にマッチするノードを返す。問合せ領域として与えたノードリストは、その根のレベルだけを調べる (DOMROOT) か、すべてのノードを調べるか (DOMTREE) を指定できる。

ノードの持つ特性は "Proploc" 文で指定できる。例えば次の問合せは「GI が "section" で属性 revision の値が 2 以上のノード」を検索する。

```
Select(DOMTREE And(
  Eq(Proploc(CAND GI) "section")
  Ge(Proploc(CAND ATTVAL[revision]) 2)))
```

ノードの内容は "Match" 文で指定できる。例えば次の問合せは「GI が "author" で "中村" を内容に含むノード」を検索する。

```
Match(
  Select(DOMTREE
         Eq(Proploc(CAND GI) "author")
  "中村"))
```

我々は、HyQ の構文から省略可能な細かな指定と、location 文の一部などを除き、C++ で処理系を作成した。HyQ テキストを中間形式に変換した後、SGML 文書 DB システムを作成するのに用いた DBMS の問合せ言語を用いて実行する。ただし、この HyQ 処理系は機能評価に主眼を置いているため、実行時の最適化はほとんど行なっていない。

アプリケーションからの呼び出しは、問合せ文とその引数を指定した C++関数で行なう。問合せ結果は、別の問合せ文の引数とすることもできる。

5 評価

HyQ は、SGML 文書のノードの持つ様々な性質を "特性" としてアクセスできる強力な機能を持っており、SGML 文書に対する多様な問合せが可能である。しかし次のような問題があることがわかった。

記述の煩雑さ 問合せの記述方法が直感的でないため、単純に思える問合せでも複雑な記述が必要になる。たとえば、GI が "yyy" のノードのうちで、親ノードの GI が "xxx" のものだけを取り出したい場合は、

```
Select(DOMTREE And(
  Eq(Proploc(CAND GI) "yyy")
  Eq(Proploc(RELLOC(CAND PARENT) GI)
     "xxx")))
```

のように書かねばならない。

特性指定を拡張して、

```
Proploc(CAND GI-extended) "xxx/yyy")
```

のような直感的で簡略な記法が、問合せ文の書きやすさと実行時の最適化処理のために必要であろう。

特性定義の不足 HyQ は文書中の様々な情報を特性としてアクセスする強力な機能を有している。特性として GI や属性値などが定義されているが、それだけでは不足であるので、例えば文書 ID や文書型を表現する特性を追加した。

新たな特性は自由に定義できるようになっているが、交換性を高めるためには何らかの標準化が必要であろう。

6 おわりに

HyQ の処理系を試作することで、SGML 文書 DB に対する問合せ言語として HyQ が十分な機能を持つことを確認した。

今後は、検索速度を向上させるための最適化処理と、実装を省略した部分のサポートを行ないながら、特性定義の拡張を行なう予定である。

参考文献

[1] 岩崎雅二郎. OODB による構造化文書データベースの実装. 情報処理学会研究報告, 94-DBS-99(34):257–260, 1994.

[2] 原 正一郎, 根岸正光, 安永尚志. 文書の構造に着目した全文データベース検索システム. 京都大学大型計算機センター研究セミナー報告, 35:39–56, 1992.

[3] ISO/IEC 10744. *Information technology — Hypermedia/Time-based Structuring Language (HyTime)*, 1992.