文書型定義 (DTD) の類似性を利用した XML 文書の検索手法の提案

3T－3　小野 智弘　西山 智　小花 貞夫
(株)KDD 研究所

1. はじめに

近年、インターネットやイントラネット上で文書を記述、交換するための言語として XML (eXtensible Markup Language)[1] が注目されている。XML は HTML と異なり、構造を持った 文書を記述するためのタグを用いることにより、文書を一まとまりではなく細かい要素の単位で記述・管理することを可能とする。また、タグはユーザが自由に定義して使用できるため、インターネットやイントラネット上では、構造上異なっているが意味的に類似したタグ定義 (DTD: Document Type Definition) を持つ XML 文書が散在することとなる。筆者らは、このような類似の XML 文書をユーザが DTD の差異を意識せずに効率的に検索するための手法を提案し、その実現例を述べる。

2. XML データベースの現状と課題

これまでに、XML で記述された文書を格納し、検索するためのデータベースが幾つか発表されている[2]。これらの大部分は、DTD を RDBMS や OODBMS 等のデータ形式へ変換し、XML 文書を格納・検索する方式を採用している。問い合わせ言語は XML-QL[4]や、XQL[5]等が W3C(World Wide Web Consotium) で検討されている。図 1に 2 種類の DTD と、それに基づいた XML 文書、検索式の例を示す。(a) は paper, title, author, date の各タグ（タグの名前を要素名と呼ぶ）を定義している DTD で、paper が残りの 3 つを含むことを示している。(b) は XML 文書の表現で、paper を起点（ルート要素名）とする DTD に従っていること、値が、"SAMPLE TITLE"、john、1103 であることを示している。(c) は XML-QL で記述した検索式で、paper をルート要素名とし、author の値が john である XML 文書から title の値を取得することを示している。

DTD については、医療分野での電子カルテ用 DTD や、金融分野の取引き文書用 DTD 等のように、特定の産業分野で扱うものを共通化し、それに従って文書を記述／交換しようとする動きがある[2]。このような分野では決められた共通の DTD が利用されるため、上記の XML データベースが有効となる。

一方、XML 文書は HTML 文書に代わってインターネットやイントラネット上で幅広く利用され、格納されることが期待されている。この場合、全ての利用者間で共通の DTD が利用されるのではなく、情報発信者が独自に定義／拡張した DTD を用いて文書を記述すると

Retrieval Method for XML documents based on similarity of DTD(Document Type Definition)
Chihiro ONO, Satoshi NISHIYAMA and Sadao OBANA, KDD R&D Labs. Inc.

(a)DTDの例1
```
<!ELEMENT paper(title, author, date)>
<!ELEMENT title(#PCDATA)>
<!ELEMENT author(#PCDATA)>
<!ELEMENT date(#PCDATA)>
```

(a')DTDの例2
```
<!ELEMENT article(Title, page, writer)>
<!ELEMENT Title(#PCDATA)>
<!ELEMENT page(#PCDATA)>
<!ELEMENT writer(#PCDATA)>
```

(b)XML文書の例1
```
<?xml version="1.0" ?>
<!DOCTYPE paper
  SYSTEM "http://www.a.b.c/pap.dtd">
<paper><title> SAMPLE  TITLE </>
<author>john</><date>1103</></>
```

(b')XML文書の例2
```
<?xml version="1.0" ?>
<!DOCTYPE article
  SYSTEM "http://www.a.b.c/pap.dtd">
<article><Title> SAMPLE  TITLE </>
<page>123</><writer>john</></>
```

(c)XML-QLで記述した検索式の例1
```
WHERE <paper>
        <author> john </>
        <title> $a </>
      </> IN "www.a.b.c/bib.xml"
CONSTRUCT $a
```

(c')XML-QLで記述した検索式の例2
```
WHERE <article>
        <writer> john </>
        <Title> $a </>
      </> IN "www.a.b.c/bib.xml"
CONSTRUCT $a
```

図 1: DTD、XML 文書、検索式の例

考えられる。このような環境で利用する XML 文書は、類似した異なる DTD に基づいて記述されたものが多くなる。このような場合に従来の XML データベースを利用すると、必要な値があると思われる全ての DTD の文書に対して別々に検索式を記述する必要があるため、効率的でない。図 1を例にとると、XML データベースでは paper と article で定義される文書は異なったものであるため、paper と article それぞれに対して図 1(c)(c') のように問い合わせる必要がある。

これまでに、異なる DTD に対する柔軟な対処方式として、ワイルドカードを含む曖昧なパス指定を導入することにより構文の異なる DTD の差異を解消する方式[3]が検討されている。ところが、DTD の要素名の配置（順序）の差異のみに対応しており、要素名の差異には対応していないため、十分でない。

3. DTD の類似性を利用した文書検索手法の提案

ユーザに DTD の要素名の差や配置を意識させずに XML 文書を効率的に検索するために、類似した異なる DTD を持つ XML 文書をユーザからの一つの検索要求で一括して扱う仕組みを提供することを提案する。具体的には、1) 事前に対象となる DTD を絞り込むための索引を類似検索を利用して構築し、2) 検索時に検索パラメタを類似検索を利用して拡張する。

[カテゴリ索引の構築]

同一の DTD に従う全ての XML 文書の同一のタグに対応する値は同じ意味上の分類 (カテゴリ) に属する (例えば< author >というタグに対応する値はいずれも「著者」を表す) という仮定に基づき、DTD のあるタグに対応した要素の集合を特徴付ける「カテゴリ名」を、実際の XML データベース内の要素の値からシソーラスを

利用した類推により導出する。この導出した「カテゴリ名」を索引鍵とし、それに対応する実際の「DTD」を値とする索引(以降、カテゴリ索引と呼ぶ)を構築する。類推には既存の類似検索手法[7]を用いる。あるいは、実際のXML文書ではタグの要素名がその値のカテゴリを表現していることが多いため、要素名をそのままカテゴリ名としてカテゴリ索引を構築しても良い。

[検索時のパラメタの拡張]

ユーザからの検索要求時に、検索パラメタの類義語をシソーラス等を利用した類似検索により取得し、その類義語全てを検索鍵としてカテゴリ索引を用いて対応するDTDを得ることにより、ユーザの検索要求に関連するDTDを得る。この類似検索では、システムがどの範囲までを類似と解釈するかに応じて適切な類似検索手法[7]を採用する。

[システム構築時の留意点]

本検索手法を実現するシステムを構築するに当たり、ユーザからの入力形式(ユーザがDTDを知っている場合はDTDに従った一つの検索式を入力させ、DTDを知らない場合にはパラメタ全てを曖昧検索の対象とする)や、ユーザへの出力形式(統一するなら、XMLデータベースからの応答を合成する必要があり、そうでなければ収集結果をそのままユーザへ返却する)等を考慮する。

4. システムの例

上記の方針に従ったシステムは幾つか考えられるが、以降では、ユーザの曖昧な検索要求から実在する複数のDTDに基づくXML文書への検索要求へ変換する「仲介モジュール」(図2)の構築例を示す。

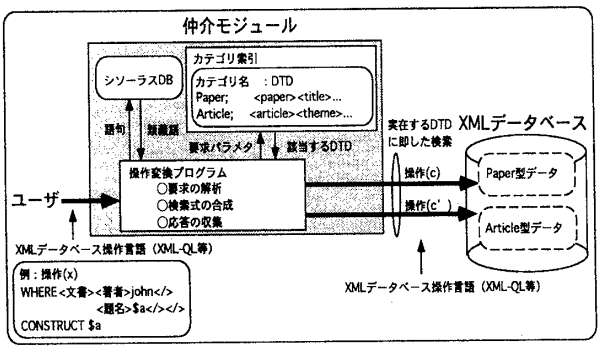

図 2: 仲介モジュールの例

例えば、著者がjohn氏である本の題名を知りたい場合、ユーザは全くDTDを知らなくても、検索式の書式に従って操作(x)(図2左下)を仲介モジュールへ発行すると、仲介モジュールがそれぞれpaperで定義される文書に対して図1(c)の操作を発行し、articleで定義される文書に対して(c')の操作を発行する。

4.1 仲介モジュール

仲介モジュールは、XMLデータベースのDTDの情報を保持する「カテゴリ索引」、与えたキーワードの複数の類似語を出力する「シソーラスDB」と、それらを利用する「操作変換プログラム」から構成される。ここで用いるシソーラスDBはQZS Dictionary Server[6]等の既存品を用いる。ユーザ、仲介モジュール、XMLデータベース間のインタフェースは、利用するXMLデータベースの提供する操作言語(例えば、XML-QL等)を用いる。

4.2 検索操作の内部処理

仲介モジュールの内部処理を操作(x)から操作(c),(c')への導出を例に以下に示す。

(1) 要求の解析:
検索操作のパラメタより、ルート要素名('文書')、条件の対象となる要素名と値('著者'と'john')、抽出対象となる要素名('題名')を抽出

(2) 検索式の合成:

(i) ルート要素名の類義語をシソーラスDBを用いて検索(=文、本、paper..)

(ii) 上記の類義語をルート要素名とするDTDをカテゴリ索引から検索(=paper, article)

(iii) 検索したルート要素に対応する他の要素名をシソーラスDBを利用してカテゴリ索引から抽出(著者→< paper >< author >, < article >< writer >)

(iv) 各型に対する検索式を合成し、XMLデータベースへ送付(式=(c),(c'))

(3) 応答の収集:
XMLデータベースからの複数の応答を収集してユーザへ返却

5. おわりに

本稿では、ユーザにDTDの差異を意識させない類似のXML文書を検索する手法を提案し、解決例を示した。今後は、実際に仲介モジュールを実装し、有効性の評価を行う予定である。また、本稿ではDTDが既知の状態での蓄積した文書の検索手法について論じたが、本手法の、DTDが既知とならないインターネット上の文書の収集(検索エンジン等)への適用も検討する予定である。最後に日頃御指導頂く(株)KDD研究所 村谷 拓郎 所長 および 鈴木 健二 副所長に感謝します。

参考文献

[1] W3C, "Extensible Markup Language(XML) 1.0", http://www.w3.org/TR/REC-xml, 1998

[2] "特集 XML" 日経バイト pp.114-131, 1999.1

[3] 北野他, "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", 第57回情処全大, 1998

[4] W3C, "XML-QL: A Query Language for XML", http://www.w3.org/TR/NOTE-xml-ql, 1998

[5] W3C, "XML Query Language(XQL)", http://www.w3.org/TandS/QL/QL98/pp/xql.html, 1998

[6] "QZS Dictionary Server", http://www.qzc.co.jp/qzsth.html

[7] 幡鎌他 "ナレッジマネジメントへ向けて", 人工知能学会, Vol13 No.6. pp.52-59 1998