# タグ情報を利用した XML 検索システム

松本 亮　　大宮 進　　絹川博之†

東京電機大学 大学院　工学研究科‡

## 1. はじめに

近年、インターネットは大きく発展し、Web サービスや電子商取引などビジネス分野でも盛んに利用されている。その基幹技術として用いられ始めているのが XML[1]とその関連技術である。本研究は、膨大な XML 形式のデータの中から効率よく目的のデータを取得でき、かつ、XML のスキーマに依存しない汎用的な検索システムの構築を目的とする。

## 2. 検索システムの概要

本研究の XML 検索システムは

(1) エンドユーザからの入力を受け付けるインタフェース

(2) インタフェースが生成した検索式を基に検索を実行するエンジン

に分けられる。

XML の検索用言語である XQuery[2]は検索対象が XML 文書であれば、検索式を与えるだけでスキーマに依存することなく検索を行える。しかし、インタフェース部分に関しては、対象とする XML 文書のスキーマによって検索項目やデータ入力部分が変化するため、一つの形式のインタフェースを汎用的に使用することはできない。このため、システム全体として汎用性を持つことは出来なくなっている。そこで、インタフェースを検索対象に応じて生成することでスキーマに依存しない検索システムの構築を目指す。

## 3. 検索要求定義情報

検索対象となる XML 文書に適した検索システムを自動生成するため、以下の 5 情報を定義する。

(1) 検索要求定義

(2) 異種スキーマ文書間のタグ対応定義

(3) 検索 UI 定義

(4) 検索エンジンに適した検索式定義

(5) 検索結果表示形式定義

これらをを「検索要求定義情報」という。

検索要求定義情報の 5 種の各定義において、共通情報に初出定義で定義 ID を付与し、非初出定義内においては定義 ID から当該共通情報を取得し、利用する形を取っている。

これにより、変更への柔軟性と検索要求定義情報編集の効率性が得られる。

検索要求定義情報の各定義について、スキーマの異なる XML 文書、「内閣名簿の XML 文書」と「自由民主党員名簿の XML 文書」からの国会議員情報検索を例に、説明する。(図 1、参照)。

**【内閣名簿の XML 文書例】**

```
<?xml version="1.0" encoding="Shift_JIS"?>
<cabinet>
  <minister>
    <post>内閣総理大臣</post>
    <name>小泉純一郎</name>
    <belong>
      <house>衆議院</house>
      <party>自由民主党</party>
      <district>
        <prefecture>神奈川</prefecture>
        <area>11</area>
      </district>
    </belong>
  </minister>
  <minister>
    <post>総務大臣</post>
    <name>片山虎之助</name>
    <belong>
      <house>参議院</house>
      <party>自由民主党</party>
      <district>
        <prefecture>岡山</prefecture>
        <area></area>
      </district>
    </belong>
  </minister>
</cabinet>
```

**【自由民主党員名簿の XML 文書例】**

```
<?xml version="1.0" encoding="Shift_JIS"?>
<PartyMembers>
  <Member>
    <Personal>
      <Name>小泉純一郎</Name>
      <Birthday>1942.01.08</Birthday>
      <ElectNumber>10</ElectNumber>
      <Prefecture>神奈川</Prefecture>
    </Personal>
    <Belong>
      <House>衆議院</House>
      <Post>総裁</Post>
    </Belong>
  </Member>
  <Member>
    <Personal>
      <Name>山崎　拓</Name>
      <Birthday>1936.12.11</Birthday>
      <ElectNumber>10</ElectNumber>
      <Prefecture>福岡</Prefecture>
    </Personal>
    <Belong>
      <House>衆議院</House>
      <Post>幹事長</Post>
    </Belong>
  </Member>
</PartyMembers>
```

**図 1　異種スキーマ XML 文書例**

### 3.1. 検索要求定義

検索要求に関して、検索内容および対象となる XML 文書群の説明と場所を記述する。

本例の国会議員情報検索では、検索内容は政治家の人物情報検索となる。

### 3.2. 異種スキーマ文書間のタグ対応定義

各検索項目とそれを保持する各 XML 文書内のタグとの関連付けを記述する。XML 文書内のタグのパスについては XPath[3]に準拠した記述を行い、各関連付けにはタグ対応 ID を付与する。

本例の国会議員情報検索では、

*/cabinet/minister/name*（内閣名簿）と、

*/PartyMembers/Member/Personal/Name*（党員名簿）

とが同一の内容である。これらを対応付け、タグ対応 ID(例えば、”meta_001”)を付与することである。

### 3.3. 検索 UI 定義

タグ対応定義によって得られたタグ対応 ID と検索 UI のフィールドを関連付ける。また、利用者が検索キーワードを入力するフィールドの配置やグループ化といったデザイン部分に関しても記述する。各フィールドの親子関係は親子関係を用いて表現し、フィールドののプロパティは属性値を用いて表現する。

---

Element based XML Information Retrieval System

Akira MATSUMOTO, Susumu OOMIYA, Hiroshi KINUKAWA†

Graduate School of Engineering, Tokyo Denki University‡

本例の国会議員情報検索では、議員名入力フィールドの表示位置等の設定とタグ対応定義で設定したタグ対応 ID の関連付けを設定することである。

### 3.4. 検索エンジンに適した検索式定義

XML 検索エンジンに渡す検索式を定義する。一般に、検索エンジンによって検索式のフォーマットは異なる。現段階では、タグ対応定義部分の情報と検索式の雛型を利用した穴埋め型の記述となっている。

### 3.5. 検索結果表示形式定義

XML 検索エンジンから帰ってきた検索結果を、ユーザに提示する形式や印刷するために変換する規則を記述するもので、これにはXSL[4]を用いる。

## 4. 検索システムの処理

3 章の検索要求定義情報に基づく検索システムの処理[5]は以下の通りである。

(1) ユーザの検索要求から検索要求定義情報を選出する
(2) 検索要求定義情報内の検索 UI 定義から UI を生成し、表示する
(3) 検索式定義とユーザが検索 UI のフィールドに入力したキーワード、タグ対応定義から検索式を生成する
(4) 生成した検索式を検索エンジンに渡し検索を実行する
(5) 検索結果を取得し、検索結果表示形式定義をもとに整形を行いユーザに検索結果を提示する

以上の処理を図２に示す。

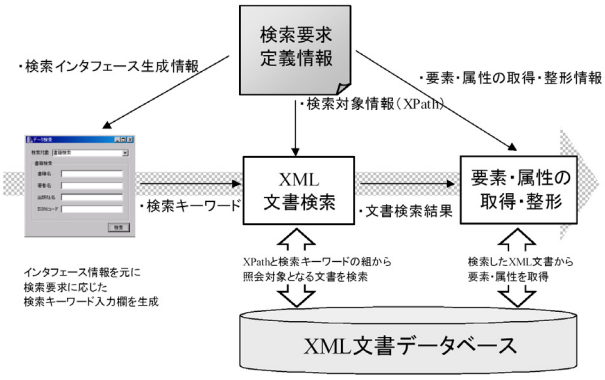


図 2 検索要求定義情報に基づく検索システム

## 5. 考察

現状の問題点、今後の改良点として以下があげられる。

(1) 検索式生成規則の汎用性向上
複雑な構文による問い合わせを受付可能な検索エンジンに対してはテンプレートを利用した検索式生成だけでは十分に機能を利用できない。また、XQuery 等には様々な条件定義が存在するので、構文上可能な範囲でより拡張性の高い記述方式を発案する必要がある。

(2) 出力フォーマット記述の機能追加
XSL は、スタイル記述言語としてだけでなくフォーマット変換の機能も持っている。XML 検索エンジンが返すデータが XML ではない場合や、必要十分でない場合には、情報の抽出・プリフォーマットを行う必要性が生じる。

(3) 検索要求定義情報への検索エンジン定義の追加
検索要求定義情報に、検索エンジン定義を追加することで、検索エンジンに依存する部分に関して分離することが出来れば、よりシステムの変更に対して柔軟に対応することが可能となる。また、それにより既存定義の再利用性も高めることが可能になると考える。

(4) 定義情報記述ツールの開発
検索要求定義情報は、人手でそのすべてを記述するのは現実的ではない。よって、必要な項目に情報を入力するだけで容易に検索要求定義情報を作成することが出来るツールを開発する必要がある。

これらを考慮にいれ、検索要求定義情報のスキーマを作成し、検索システムの実装とフィードバック、編集ツールの開発を順次行う。

## 6. おわりに

XML のスキーマに依存しない汎用的な検索システムを構築する手法として、検索要求に沿った検索インタフェースの生成について設計した。XML は自由にスキーマを定義可能であることから、一意に定まる XML スキーマを期待することは難しい。その上で、検索システムに柔軟性を持たせ、異種スキーマ XML 文書の検索に対応することは意義あることだと考える。

今後は、スキーマ設計をより煮詰めることにより、システムのコンポーネント化および、XML 検索システムのフレームワーク化を目指す。

## 参考文献

[1] World Wide Web Consortium: Extensible Markup Language (XML) 1.0 (Second Edition) W3C Recommendation 6 October 2000 , http://www.w3.org/ TR/2000/REC-xml-20001006

[2] World Wide Web Consortium: XQuery 1.0: An XML Query Language , http://www.w3.org/TR/xquery/

[3] World Wide Web Consortium: XML Path Language (XPath) Version 1.0 W3C Recommendation 16 November 1999, http://www.w3.org/TR/1999/REC-xpath-19991116/

[4] World Wide Web Consortium: The Extensible Stylesheet Language (XSL), http://www.w3.org/TR/1999/REC-xpath-19991116/

[5] 大宮 進, 絹川 博之:XMLタグ情報を利用した検索システムに関する一検討, 情報処理学会 第 64 回全国大会 1Z-04 (2002)