帰納推論に基づく XML タグ構造分類による XML 検索

駅昌弥 † 大囿忠親 † 新谷虎松 †
‡ 名古屋工業大学大学院 工学研究科 情報工学専攻
e-mail: {eki, ozono, tora}@ics.nitech.ac.jp

1 はじめに

本稿では制約論理プログラミングに基づく帰納推論により，XML のタグ構造を分類し，検索を支援するアプローチについて述べる．一般に，XML で記述されたデータは，スキーマが不完全であることがあるが，それらは例えば，あるタグ A の近くに別のタグ B が存在しやすいといった，何らかの規則性を持つ構造を内在していることが多い．本研究では前述のような XML データに対して同じ情報を表現する構造を，XML データのシンタックスとセマンティクスに基づいて識別し，検索やデータ管理の支援を行うことを目的としている．

本研究は，1) XML 構造のみに着目した述語とタグのデータ型や値域に関する述語の生成，2) 帰納推論により，1 で生成した述語から同じ情報を表現する異なった構造を分類，3) XML Schema で格納，の 3 ステップに分類される．

2 関連研究

本研究に関連する研究をいくつか本章で紹介する．まず，Nestorov が行った半構造データからスキーマの抽出に関する研究 Nestorov 98[3] が挙げられる．本研究は，この研究の考えをベースとしている．

制約論理プログラミングは，論理プログラムに制約解消のメカニズムを組み込んだプログラミング言語のことであり，相対包摂に制約の概念を導入する研究 [2] がある．最後に，本研究の詳細は，Eki 07[1] に記されている．紙面の都合上省略した，述語の生成に関する詳細なアルゴリズムや実験結果についてはこちらを参照してほしい．

3 提案手法

本章では，3.1 節で XML 構造のみに着目した述語とタグのデータ型や値域に関する述語の生成について述べ，3.2 節では，帰納推論により，3.1 節で生成した述語から同じ情報を表現する異なった構造を分類するアプローチについて述べる．本紙では，紙面の都合上 XML Schema で格納を行うプロセスについては省略する．

3.1 述語の生成

まず，XML 構造のみに着目した述語を生成する．ここではリンク構造と値を保持し，元の XML 構造に戻せる (要素の順序は保証しない) ようにする．生成する述語は以下の 7 種類になる．

class(Object, Class). Object の類は Class である．Class は，帰納推論で分類された後，格納される．

link(FromObject, ToObject, Label). FromObject と ToObject は親子関係であり，Label の関係を持つ．

element(Object, Value). Object は，要素 Value を持つ．

attribute(Object, Value1, Value2). Object は，属性 Value1=Value2 を持つ．

root(Object). Object はルートノードである．

branch(Object). Object は枝である．

leaf(Object). Object は葉である．

次に，XML データのデータ型及び値域に関する述語を生成する．本研究では，同じ意味を表現するタグは，データ型や値域が類似していると考えたので，これらの述語を生成した．また，これらは XML Schema のデータ型及び正規表現で表現する．生成する述語は以下の 2 種類になる．

datatype(Object, Type). Object のデータ型は Value である．Type は，例えば 10 進数数字列として decimal 型，文字列型として string 型，真理型として boolean 型などがある．

dataarea(Object, Value). Object の値域は Value である．Value は，文字数や値の範囲を記録する．たとえば数値データで範囲が 1000 以上 9999 以下なら，xsd:minInclusive value="1000", xsd:maxInclusive value="9999" と記述する．

3.2 帰納推論によるマッチング

本研究では代表的な帰納論理プログラミングのシステムである Progol を用い，タグ構造を基に分類規則を解析する．帰納推論に用いる背景知識は，多段階特徴量の階層関係と特徴量間の包摂関係を記述した．ここでは，正例負例の両方のデータを用いた通常の学習と，正例のみ用いる学習を用い，2 種類の分類規則を導出した．

正例負例の両方のデータを用いた通常の学習は，正例のみ用いる学習に比べて制約の強い分類規則になるため，分類されるカテゴリ生成の精度は高いが，学習データと少しでも異なると検出できない．一方で，正例のみ用いる学習の制約は弱いので，学習データと少しだけことなる場合にも対応できるが，カテゴリ生成の精度が悪くなる．

そこで本研究では，正例負例の両方のデータを用いた通常の学習の制約をベースとし，正例のみ用いる学

XML Retrieval Based on Inductive Reasoning
Masaya EKI, Tadachika OZONO, Toramatsu SHINTANI
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya 466-8555 JAPAN

表 1: 実験データ

Test Case	Schema ?	Implisit Structure ?	Instance Filename	Average Instance Size (Kbytes)	Average Number of Nodes
1	Y	Y	testcase1_1-100.xml	22	112
2	Y	N	testcase2_1-100.xml	21	112
3	N	N	testcase3_1-100.xml	22	112

表 2: 実験結果

Test Number	Test Case	Evaluation Vakue	Average Constraint Count	Average Precision(%)	Average Recall(%)
1	1	0.8	11	91.0	100
2	1	0.6	11	78.8	100
3	1	0.4	11	78.6	100
4	2	0.8	21	80.2	74.1
5	2	0.6	21	64.7	82.3
6	2	0.4	21	59.7	83.9
7	3	0.8	35	38.6	25.6
8	3	0.6	35	35.4	32.5
9	3	0.4	35	22.7	49.2

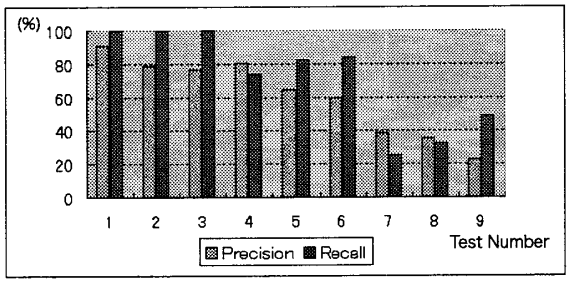

図 1: 実験結果のグラフ

習の制約による検出結果を統合することで検索結果を得ることで，検索精度を改善するアプローチをとっている．

4 実験と結果

実験に用いたデータは，NewsML や web の log など 81 種類のユニークな XML データから生成された 300 種類の XML データを用いた．実験は，表 1 に示すデータの種類で，それぞれ 100 種類の XML データの実験結果の平均を取得した．ここでは，XML データのノードが，約 110 ノードになるように 300 種類の XML データを生成した．表 1 の 'Schema?' は，その XML データが DTD などのスキーマを持つことを示している．'Inplicit Strucutre?' は，その XML データが implicit な構造を持つことを示している．implicit な構造は，世の中で最も多いと思われる XML データであり，スキーマの存在しないが規則性を持ったデータの集合である．implicit な構造を持たない XML は，複数の異なった XML データからランダムに取得した集合体であり，規則性はほとんど見られない．Evaluation Value は，正例負例の両方のデータを用いる割合であり，0.8 の場合は例負例の両方のデータが 8 割，正例のみのデータが 2 割という意味になる．ここでは，各 Test Case に対して，Evaluation Value の値が 0.8, 0.6, 0.4 の 3 パターンの実験を行った

実験結果を，表 2，そのグラフを図 1 に示す．表 2 の Test Number と，図 1 の横軸の Test Number は対応している．また，縦軸は再現率と適合率を示している．この実験ではスキーマを持つ Test Case1 は，非常に高い再現率であるが，Evaluation Value が 1 以下の時は 100%にならなかった (表の 2 の結果は，有効数字三桁で表示してある)．その理由としては，定義の仕方のよって似通った構造が存在してしまう点が挙げられる．本研究では検索の際に，元のスキーマを参照していないので，この要素が増えるとスキーマを持っていても再現率が低下する．また，再現率は，タグの値域などにも依存してしまうので，単一のスキーマを持っている XML データ集合に対してはスキーマと XML データを一緒に格納し，検索を行う既存の XML データベースを用いたほうが効率がよい．

スキーマを持たないが implicit な structure を持つ Test Case 2 は，Evaluation Value を高めに設定したときは再現率が約 74%，低めに設定したときには適合率が約 60%に落ちている．本研究の主な目的はこの Test Case2 であり，今後このケースの再現率をできるだけ近づけると同時に，精度の高い適合率を示す手法の考案が必要である．スキーマを持たなく，implicit な structure も持たない Test Case 3 は，Test Case 2 に比べさらに多くの制約が生成された．しかし，XML データは基本的に規則的な構造を持つものであり，それに対してデータ検索や交換などを行うのが普通であるので，本研究ではこのケースは想定する必要がないと考える．

5 まとめ

本稿ではスキーマが不完全であるような XML データに対して，同じ情報を表現する構造を XML データのシンタックスとセマンティクスに基づいて識別し，検索やデータ管理の支援を行うアプローチについて述べた．その際導入した手法では，制約論理プログラミングに基づく帰納推論により，XML のタグ構造を分類している．本研究は，特徴の差異が比較的大きなタグ構造に対しては高精度な分類や検索を実現するが，微小な差異を示すタグ構造に対する有効性の検証は今後の課題である．

参考文献

[1] Masaya Eki, Tadachika Ozono, Toramatsu Shintani, 'On an XML Database System Based on Constraint Logic Programming', WorldComp ICAI'07, 2007.

[2] Fumio Mizoguchi, Hayato Ohwada, 'Constraint relative least general generalization for inducing constraint logic programs', New Generation Computing, pages 335-368, 1995.

[3] Svetlozar Nestorov, Serge Abiteboul, Rajeev Motwani, 'Extracting Schema from Semistructured Data', SIGMOD'98, pages 295-306, 1998.