4W-5

ＤＴＤマッチングによる大学シラバスの相互変換

平野健太郎　青野雅樹
豊橋技術科学大学

1　はじめに

現在，XML の情報統合が盛んに行われてきている[1，2]．しかし，異なるスキーマをもつ XML データ交換・統合の問題は困難で，普遍な解はない．

本研究では，内容は類似するが，スキーマが異なる XML 化した大学のシラバスに着目して，相互にデータ変換できるシステムを提案し，実験結果を報告する．本システムでは，２つのシラバスの DTD を解析しマッチングを行い，片方のシラバスから，もう片方のシラバスへ変換する．その際，XML の要素名が完全に一致しない場合でも，シソーラスを用い類似要素への変換を実現した．

本稿では，提案する変換手法の概要を述べ，システム構築にあたり問題点となる点を列挙し，実験を行った結果を示す．システムの評価として，生成された変換後の XML と採用される XML を質的に評価できるよう HTML により閲覧させることと，変換する際にどれほど要素名，構成が変わるかの尺度として変換コストを定義し評価する．本システムにより，大学の科目採用及び教員採用戦略，JABEE 対応校との相互比較などへの応用が可能である．また，本システムを応用すればシラバス以外でも，類似する XML であれば相互変換・比較が容易になると考えている．

2　相互変換手法

2.1　概要

本提案システムの概要を図 1 に示す．使用データとして DTD 付帯の大学シラバス XML データを 2 大学分用意する．対象となる XML から類義語データベースとなるシソーラスを生成する．シソーラスに含ませる類似要素の例として「科目名」，「講義名」が挙げられる．また，DTD を解析し，採用する DTD を決定する．この決定は DTD の DOM 木がより深く，より広域なものを採用するようにした．採用する DTD を決定した後（図中では XML 其の 1 を採用），作成されたシソーラス，採用 DTD，XML，不採用 XML を

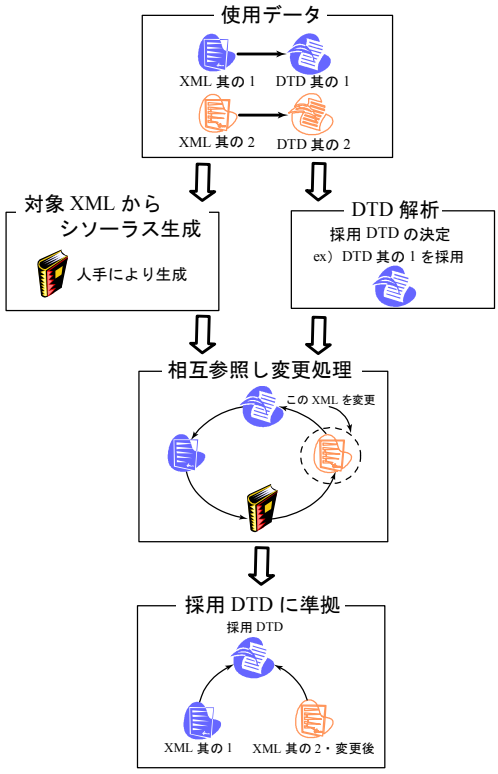

図 1　提案システム概要

使用し，不採用 XML を採用 DTD に準拠するよう変更処理を施す．

2.2　問題点と解決手段

上記システムを構築する際に下記のような問題点があった．A，B は対象となるシラバスデータ（XML＋DTD）である．A を採用シラバスデータ，B を不採用シラバスデータとして考える．

1. A の DTD に要素が存在・B の DTD に不在．
2. A の DTD に要素が不在・B の DTD に存在．
3. 類似要素の出現順序に差異．
4. 類似要素は両者に存在．その要素に子要素が存在，または不在．
5. 類似要素は両者に存在．根からの深さに差異．

1，2 は空要素の生成，対象要素の削除の処理にて解決させ，3 は処理の最終段階で採用 DTD によるソート処理を施した．4 に対しては，採用 DTD により子供要素を生成・削除の処理を行い，5 に対しては，対応要素に子供がいるか判別処理を行い，DTD を解析した際の深さと比較する．採用 DTD の深さが浅い場合，その親要素を削除

Mutual conversion of university syllabus by DTD matching
Kentaro Hirano, Masaki Aono
Toyohashi University of Technology

し，子要素を親要素の位置に上げる．採用 DTD の深さが深い場合は，親要素を作成しその階下に対応要素を付加させた．

3　実験結果

構築したシラバス変換システムの実験結果を評価するため，以下の 2 種類の評価方法を定義した．

3.1　HTML 閲覧による評価

変換後のシラバスと変換しない採用シラバスとを質的に評価・比較出来るように，両者を入力として閲覧用 HTML ファイルを出力とする評価システムを構築した．図 2 に出力 HTML を示す．これにより，採用シラバスには存在するが，変更されたシラバスには存在しない要素などが一目で分かり比較が容易になる．また，空要素の場合には赤字で null を表示させている．

DTD要素	採用シラバス	変更シラバス
授業科目の区分	専門科目	null
授業科目	数値シミュレーション	数値解析
欧文	Numerical Simulation	null
講義番号	D306	132050
担当教官名	桑原義彦	中内茂樹
対象年次	3年	学部3年次
開講期	前期	2学期
授業時数	60	null
必修選択の別	選択	選択
単位数	3	2
曜日時限	null	null
講義室	null	null

図 2　比較用 HTML

3.2　変換コストによる評価

上記，質的評価に加え，変換する際のコストとして以下のように定義する．ここで，大学のシラバスを X_1，X_2 とし（X_i は XML＋DTD）$X_1 \rightarrow X_2$ の変換を考える．この変換におけるコスト関数を $cost(X_1 \rightarrow X_2)$と表記し，以下の 3 つの変換コストの和であると定義する．

$$\mathrm{cost}(X_1 \rightarrow X_2) = \alpha \sum_{i=1,j=1} \mathrm{cost}(X_1 \rightarrow node[i], X_2 \rightarrow node[j])$$
$$+ \beta \sum_{i=1} \mathrm{cost}(X_1 \rightarrow node[i], null) + \gamma \sum_{j=1} \mathrm{cost}(null, X_2 \rightarrow node[j])$$

ここで，右辺の $cost(a,b)$は，X_1 の DOM 木ノード a を X_2 の DOM 木ノード b に変化するコスト関数とする．1 つ目の項は，X_1 のどこかのノード $X_1 \rightarrow node[i]$と X_2 のどこかのノード $X_2 \rightarrow node[j]$ とで変換があった場合のコストであり，変換コストと定義する．2 つ目の項は，X_1 のノード $X_1 \rightarrow node[i]$に対応する X_2 のノードが見つからず，X_1 側で余剰要素となってしまうコストであり，これを破棄コストと定義する．最後に 3 つ目の項は，2 つ目とは逆に採用 DTD にはあるが，被変換対象 DTD には対応要素がなく，$null$ となるコストであり，これを無効コストと定義する．α，β，γ はそれぞれのコストに対する重みであり，0.5，10，2 とした．ここで，破棄要素の存在がシラバス情報の損失と考えられるため，β を大きくしてある．3 種のコストを計算した結果を表 1 に示す．なお，A，B，C はそれぞれシラバスを表す．B へ変換するコストが小さくなっているのがわかる．このことより，この 3 者間においては B 大学のシラバスが汎用的であると言える．

表 1　変換コスト

	変換	破棄	無効	計
A→B	7	50	24	81
A→C	7	140	16	163
B→A	3.5	110	4	117.5
B→C	3	170	8	181
C→A	5.5	80	32	117.5
C→B	5.5	50	38	93.5

4　まとめ

本稿では，大学シラバスの情報統合のための初期段階として，スキーマは異なるが内容は類似した 2 つのシラバスを解析し，片方に変換するシステムを構築し，実験結果を述べた．2 つのシラバスを変換する際に，シソーラスを作成し要素名が完全に一致せずとも，類義語であれば変換が可能であることを提案した．また，①HTML による質的・直感的評価と②変換コストによる計量的評価の 2 種の評価方法を定義した．①の評価方法によって，より容易にシラバスの比較が可能となり，②の評価方法により，複数のシラバス DTD 中から，量的に良好なシラバス DTD の選別が可能となる．

今回は一方のシラバスをもう一方のシラバスの DTD に準拠するよう変換したが，対応しない要素が削除される（表 1 の「破棄」）ので、情報量の損失を招く．変換コストを利用して，様々な大学のシラバスを統合処理出来れば損失を防ぐことが出来，一般的なシラバス DTD が生成できると考えている．

今後の課題としては，対象シラバスを増やしてのシソーラスデータの拡充や，変換処理の対応強化が挙げられる．また，変換処理から統合処理への下準備として，変換コストの検討が考えられる．

参考文献

[1] 板井久美，高須淳宏，安達淳；“HTML Table 情報の XML による統合，第 62 回情報処理学会全国大会，3W-4，2001

[2] 鬼沢友和，安井浩之，松山実；“XML 文書の統一手法”，情報処理学会データベースシステム研究会，128-18，pp.131-138，2002.7.18