表層表現抽出と文書構造解析に基づく XML 文書変換システム

布目　光生† 石谷　康人†
Kosei FUME Yasuto ISHITANI

1.はじめに

様々なディジタル･ドキュメントを構造化して XML 文書データベースで管理することにより、文書構造に基づいた検索、閲覧、版管理、セキュリティー管理、自動組版・印刷など様々な応用が可能となったため、文書運用のコストを大幅に削減することが可能となっている。このため、既存の文書を自動的に構造化してタグ付き文書を自動生成する技術へのニーズが高まっており、これまでにいくつかの研究成果が得られている[1-4]。これらの研究では、入力文書が何らかの構造を有していることを前提とすると共に、構造が簡単な文書を複雑な文書に変換することが困難であるという制約があった。現状では、構造を持たないプレーンテキストや最小限の文書構造しか持たない HTML 文書を変換対象とする場合には、文書変換前に手作業によるタグ付けが必要とされている。本論文ではこのような問題点を解決することを目的として、プレーンテキスト、XHTML 文書、XML 文書を既存の応用規格に基づいた XML 文書に自動変換する新しい文書変換システムを提案する。

2.XML 文書変換システムの構成

本論文で提案する XML 文書変換システムを図 1 に示す。本システムは、予め定義されている「表層表現抽出のための知識辞書」と「文書構造化のための規則」を用いることにより、プレーンテキスト、XHTML 文書、XML 文書などの入力文書を文書型定義に基づいた XML 文書（以後、ターゲット XML 文書と呼ぶ)に自動変換するものである。

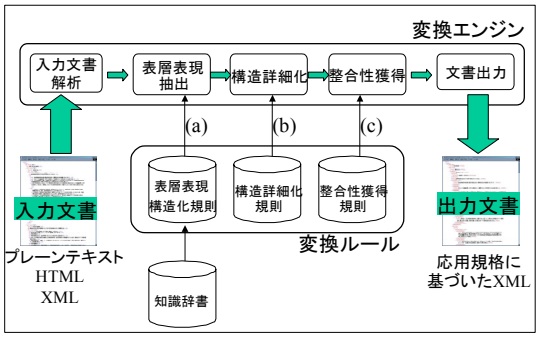

図 1：XML 文書変換システムの概要

まず入力文書解析において、システムに入力された XML(XHTML)文書を XML パーサにより DOM ツリーに変換する。プレーンテキストが与えられた場合には、前処理として XML 宣言などの付与を行なうものとする。次に、表層表現抽出により、DOM ツリーからテキストノードを収集すると共に、知識辞書を用いてテキストノードから文書構造化の手掛かりとなるキーワードを抽出する。これを表層表現抽出結果と呼ぶ。そして構造詳細化処理において、表層表現抽出結果に対して構造詳細化規則を順次適用することによりボトムアップに文書構造の複雑化を行なう。整合性獲得処理では、詳細化された DOM ツリーに対して整合性獲得規則を適用することにより、文書型定義に基づいた DOM ツリーを生成する。最後に、文書出力処理で XML パーサを用いて DOM ツリーを XML 文書に変換する。

3.XML 文書変換アルゴリズム

次に本システムの各処理(a)～(c)の詳細を述べる。

(a)表層表現抽出

まず、入力文書中の表層表現(キーワード)を抽出し、タグを付与する。例えば、医薬品添付文書の場合では、"効能・効果"や"副作用"などの定型的な表現や、"2003 年 8 月改訂"、"pH:7"などの表現を正規表現により抽出すると共に、図 2 に示すように"株式会社"と行頭や行末を手掛かりにして会社名を抽出するなどの手段を用いて、入力文書中から特徴的な表層表現の抽出を行い構造を付与する。

```
<variablelabel>東芝ソリューション<kabu>株式会社</kabu>東京都港区芝浦1-1-1(東芝ビルディング)</variablelabel>
        ↓
<variablelabel><company>東芝ソリューション<kabu>株式会社</kabu></company>東京都港区芝浦1-1-1(東芝ビルディング)</variablelabel>
```

図 2：表層表現に対する構造化例

(b)構造詳細化

次に、表現抽出結果や事前に付与されている構造を手掛かりとして駆動する構造詳細化の変換ルールにより、XML 文書の性質を保持しながら構造の詳細化を行なう。前段で付与された構造情報とその隣接構造を条件として、タグ名の変更や要素の移動といった基本的な操作の他、入力文書に存在しない新規の要素や構造の挿入、入力文書中に繰り返して出現する類似構造の範囲を自動判断してのタグ掛け、変換前後の構造をテンプレート(見本)として与えることによる変換など各種の詳細化を行なう。テンプレート変換では入力文書中の指定箇所を変数で置換可能な他、カウンタ変数の利用による箇条書き番号の自動挿入等が可能である。図 3 は箇条書きを想定するある部分構造 (block-para)に対し、テンプレートとして新たな要素名(item)やラベル名、箇条書き番号要素などを見本構造として準備しておき、該当する構造が入力文書中に出現した場合に、詳細化を行なった例である。

†(株)東芝研究開発センター
知識メディアラボラトリー

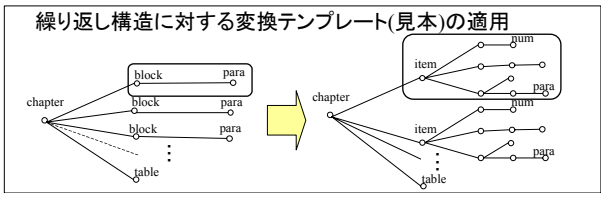

図 3：構造の詳細化例

(c)構造整合性獲得

前段までの処理で得た中間出力文書に対し、最終的に目標文書構造が持つべき制約を充足するために、部分構造の削除や部分構造の並び替えを行い、整合性の獲得（不整合の解消）処理を行なう。その結果、出力文書として目標文書構造の出現制約を満たす文書を得る。

4.XML 文書変換事例

図 4 および図 5 に特許公報を目標文書にした場合の変換例を示す。まず図 4 は入力文書であるプレーンテキストから、事前に定義されている表層表現(キーワード)を抽出し構造の付与を行なった結果である。ここでは前処理として XML 宣言を付与する他、文書全体を<root>タグで囲む、各行を<p>で囲むといった処理も施されている。

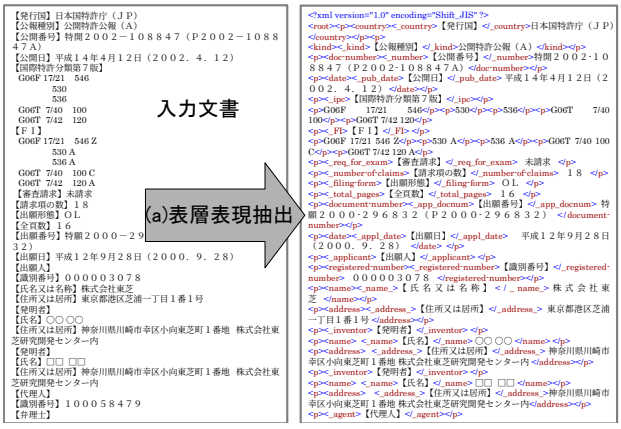

図 4：表層表現抽出結果例

次に、この抽出結果を手掛かりとして構造の詳細化を行なう。図 5 に詳細化結果と出力文書例を示す。

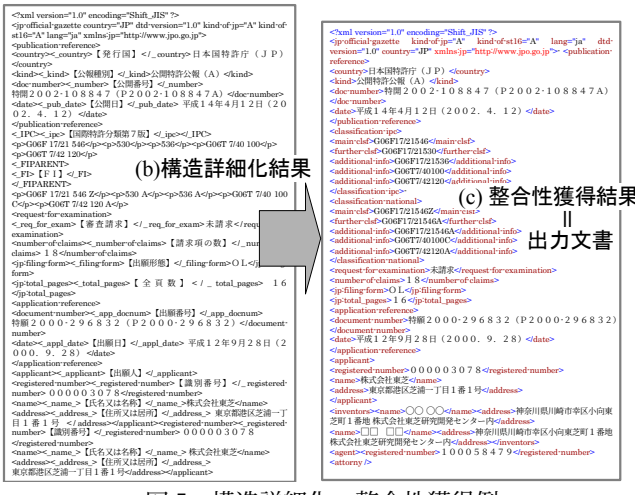

図 5：構造詳細化・整合性獲得例

詳細化処理によって目的の文書構造に対応する要素名付与の他、新規に出現する部分構造や要素の埋め込みなどが行なわれる。その結果を受け、最後に整合性獲得処理が行なわれる。不要なタグや構造の削除処理などを経て、最終的に出力文書を得る。

5.実験

本システムを用いた変換精度の測定結果を表 1 に示す。表には対象文書規模を表す尺度として、各文書を変換するのに必要なルール数、人手作成した正解文書中に含まれるタグ種別数、正解例文書に出現するタグ数も併記した。対象とした入力文書は紙文書からの OCR 出力結果である事務規定文書 8 ページ相当と医薬品添付文書 23 文書、およびプレーンテキストで記述された特許公報文書 5 件の平均値である。ここでの OCR の出力結果とは[5]に述べられている XTHML 形式であり、これには箇条書き構造や章構造などの簡単な構造が付与されている。タグ付け率は、正解文書例中の全タグ数から、入力文書のみでは構造化に必要な情報が無いなどの理由で本システムでは原理的に変換が不可能なものを除外したタグ数を、実際の出力結果で正しく付与されたタグ数で割ったものである。

また評価作業の一環(参考)として、医薬品添付文書 3 ページ相当に対する変換作業時間の測定を行なった。これは紙文書から OCR によって解析した結果を本 XML 文書変換システムにより変換した全作業工程である。また手作業による変換も実施し、変換効率の比較を行なった。その結果、手作業で 242 分を要した変換作業が、本システムを利用した場合には 60 分で完了した。

文書種別	事務規定文書	医薬品添付文書	特実公報
変換ルール数	49	131	107
タグ種別数	36	369	116
正解例文書タグ数	165	612.7	445.3
タグ付け率	94.1	84.2	91.3

表 1：評価対象・評価結果

6.まとめ

本論文では、表層表現の抽出と文書構造解析機能を有する新しい XML 文書変換システムを提案した。その結果、従来手作業で行なわれていた XML 文書の変換工程を自動化することが可能となり、従来の 1/4 の時間で既存の応用規格に基づいた XML 文書を生成することが可能となった。

[参考文献]

[1]E. Kuikka et al. Towards Automating of Document Structure Transformations. DocEng'02 Nov. 8-9,2002.

[2]X.Tang and F.W.Tompa. A High-level Specification Language for Structured Document Transformation. UW School of Computer Science Technical Report, Oct. 2002.

[3]M. Murata. Transformation of Documents and Schemas by Patterns and Contextual Conditions. Principles of Document Processing '96.

[4]酒井 乃里子 他.SGML 文書の論理構造変換手法.情報処理学会論文誌 Vol.39 No.1,Jan. 1998.

[5]石谷 康人.紙文書を対象としたピボット XML 文書に基づく XML 文書変換システム.FIT2003,Sep. 2003.