I-003

応用規格に基づいたXML文書への変換を可能とする紙文書を対象とした階層的文書変換システム

Hierarchical Document Transformation System Capable for Transforming from Printed Documents to Various Types of XML Documents

石谷康人† 布目光生† 住田一男†
Yasuto Ishitani Kosei Fume Kazuo Sumita

1. はじめに

文書内容へのアクセスが頻繁に生じる，法令集，約款集，規定集，マニュアル，論文，名刺などの文書を紙媒体の状態で管理・運用する場合には，管理スペースがない，閲覧しづらい，文書内容を容易に変更できない，情報を再利用できないなどの問題があった．このような問題を解決する方法として，最近では XML（eXtensible Markup Language）技術をベースとした文書管理ソリューションが提案されている．この場合，既存の文書を XML 化すると共に XML データベースで管理することによって，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・印刷などの様々な応用が可能となり，文書運用コストを大幅に削減することが可能となる．

このような XML 文書管理システムで紙文書を利用する場合には，紙文書を DTD（Document Type Definition：文書型定義）や XML スキーマに基づいた XML 文書（以後，ターゲット XML 文書と呼ぶ）に変換しなければならない．しかし現状では，紙文書のコード化と XML タグ付けには膨大なコストが必要とされている．

本論文では，様々な紙媒体の文書を DTD に基づいた XML 文書に自動変換する新しい方法を提案する．紙文書の内容を XML 文書に自動変換する際，ターゲット XML 文書の構造が複雑である場合には情報不足が生じるために高精度な自動変換を実現することは困難とされている．提案方式ではこのような問題点を解決するために，紙文書からコンテンツ（テキストコード）とその論理構造を復元して DOM（Document Object Model）ツリーを生成する**文書画像理解技術**と，得られた DOM ツリーに対して情報抽出，構造詳細化，構造検証を順次実施して DTD に基づいた XML 文書を生成する**文書構造変換技術**で構成される階層的文書変換方式を採用する．本研究では，階層的な文書変換の過程において，文書モデル，キーワード辞書，文書変換ルールなどの知識を利用することにより，紙文書を複雑な構造を持つ XML 文書に変換することを可能としている．

本論文では，まず，2. で従来の XML 文書変換について概観したあと，3. で提案システムの構成と動作について説明する．そして，4. と 5. で提案方式を構成する文書画像理解と文書構造変換について説明する．6. では，多様な文書を用いた XML 文書変換の実験結果を示し，提案方式の有効性について評価する．

2. 従来の XML 文書変換

紙文書を対象とした場合の従来の XML 文書変換作業は次の工程で構成されている．

ステップ 1：キーボードを介した手作業によるテキスト入力や OCR (Optical Character Reader) を用いた文書内容のテキスト化を行う．

ステップ 2：ワープロやエディタを用いて，テキスト化された文書における語，フレーズ，センテンス，パラグラフなどの文書要素に対して手作業でタグ付けを行う．

ステップ 3：ステップ 2 で作成したタグ付け結果に対して市販のオートタガーや XSLT (XSL Transformations) [1] を適用して文書構造化を行い，応用規格に基づいた XML 文書に変換する．

ステップ 4：ステップ 3 で正しいターゲット XML 文書が得られない場合には，XML 文書専用のエディタを用いてタグ付け結果を編集する．

以上のように，従来の XML 文書変換作業は段階的で繁雑な作業工程で構成されており，膨大なコストを必要とするものであった．また，作業にあたるオペレータには専門的な知識が必要とされていた．

従来の OCR 技術では，ステップ 1 の作業を軽減するために，文書要素の抽出 [14]，文書要素の読み順決定 [3, 20]，OCR 結果のハイパーテキスト化 [18, 21]，レイアウト情報の XML 出力 [4] などの機能が実現されていた．しかし，章節構造，箇条書き構造，表構造，図構造などで構成される階層的な文書構造に基づいて文書内容を自動的に XML 化することはできなかった．

一方で，テキスト化された文書を構造化文書に自動変換あるいは半自動変換する方法がいくつか提案されている [2, 6, 15, 16, 17, 19]．これらの方法は，文書変換に必要とされる処理パラメータや文書変換ルールを自動獲得することを大きな特徴としているが，以下に示す問題点があるため実際の業務における様々な文書変換を実現することは困難であった．

問題 1：変換対象が文書中の部分箇所に限定されている

問題 2：文書構造を大幅に複雑化できない

問題 3：多様な文書構造へ柔軟に変換できない

以上から，紙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ことは実現されてなかった．そこで本論文では，これらの問題点の解消を可能とする新しい XML 文書変換方式を提案する．提案方式では，文書画像理解技術と文書構造変換技術を組み合わせることにより，上述したステップ 1〜4 の機能の自動化とステップ間の連携を可能にして XML 変換作業のコストを大幅に削減することを目指す．次節以降に，提案システムの概要とシステムを構成する文書画像理解と文書構造変換について説明する．

† （株）東芝 研究開発センター，知識メディアラボラトリー

3. システム構成

前節で説明した XML 文書変換工程に基づいた階層的文書変換システム（Hierarchical document transformation system）を図 1 に示す．提案システムは，文書画像入力 (Document imaging)，文書画像理解（Document image understanding），文書構造変換（Document structure transformation），XML パーザー（XML parser），文書モデル（Document models），キーワード辞書（Keyword dictionary），文書変換ルール（Transformation rules）で構成されている．

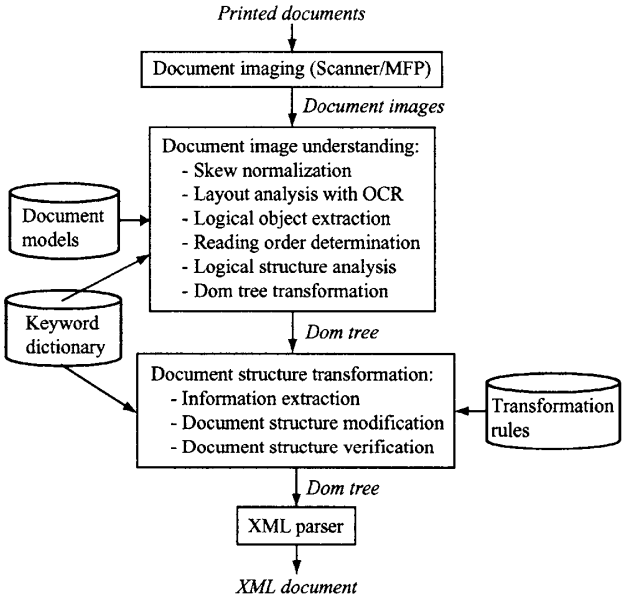

図 1: XML 文書変換システムの構成

本システムではまず，スキャナーや MFP（Multi Function Peripherals）で構成される文書画像入力部により，複数ページで構成される紙文書を複数の文書画像に変換する．次に，文書画像に対して文書画像理解処理を適用することにより，文書画像から読み順通りにコンテンツ（文字認識結果）を抽出すると共に文書論理要素とその階層構造を文書論理構造として抽出し，得られた結果を DOM ツリーに変換する．文書構造変換処理では，文書画像理解処理で得た DOM ツリーに対して，表層表現抽出，表層表現に対する柔軟なタグ付け，構造詳細化，構造検証を順次実施することにより，DTD に基づいた文書構造を有する DOM ツリーに変換する．最後に，XML パーザーを用いて DOM ツリーを XML 文書に変換する．

文書画像理解処理では，入力文書の種別ごとにあらかじめ定義されている文書モデルを利用して，文書固有の論理要素や多様なヘディング記述を有する見出しを高精度に抽出する．文書構造変換処理では，入力文書種別と応用規格の組み合わせごとにキーワード辞書と文書変換ルールがあらかじめ定義されており，同様の文書変換タスクでこれらの外部知識を利用することにより複雑な構造を有する XML 文書への変換を実現する．提案方式では，文書変換ルールはさらに，表層表現抽出ルール，構造詳細化ルール，整合性獲得ルールで構成されている．文書モデルの構成については文献 [12] を，キーワード辞書と文書変換ルールの構成については文献 [7] を参照されたい．

また提案システムでは，各文書画像の処理において，文書論理要素の範囲（論理要素を外接する矩形の位置情報），論理属性，読み順，文字認識結果をそれぞれ修正することが可能となっており，後段の文書構造変換処理に対して処理誤りのないクリーンなデータを供給することが可能となっている．ただし，本システムでは，論理要素の階層的な関係を表す文書論理構造すなわち DOM ツリーの構造を修正することはできない．

4. 文書画像理解

文書画像理解処理は，傾き補正 (Skew normalization) [9]，レイアウト解析＋ OCR (Layout analysis with OCR) [11]，論理要素抽出 (Logical object extraction) [10, 12]，読み順決定（Reading order determination）[13]，論理構造解析（Logical structure analysis）[13]，DOM ツリー変換（DOM tree transformation）[13] で構成されている．以下に各構成要素の概要について説明する．

傾き補正. 入力された文書画像に対して傾き検出を行い，得られた傾き角度を解消するように文書画像に対してアフィン変換を適用することにより傾きを補正する．

レイアウト解析. 傾きが補正された文書画像からテキスト領域，図・写真領域，表領域をそれぞれ抽出する．さらにテキスト領域と表領域から，文字パターンの集合で構成される文字行領域を抽出すると共に，文字行領域に対して文字認識処理 [5, 8] を適用して各文字パターンを文字コードに変換する．

論理要素抽出. テキスト領域と表領域からパラグラフ，章節見出し，箇条書き，表要素，数式，脚注，図表キャプション，ヘッダ，フッタなどの文書論理要素を抽出する [10]．特定の文書に対して文書モデルが定義されている場合には，文書論理要素の抽出結果に対してモデル当てはめを行なうモデルベースト文書理解処理を実施して，タイトル，日付，著者，抄録，文献見出し，文献リストなどの文書固有の論理要素を抽出する [12]．

読み順決定. 文書論理要素に対して水平方向のセグメンテーション (X-Cut) と垂直方向のセグメンテーション (Y-Cut) を再帰的に実施する XY-Cut 処理を適用すると共に，セグメンテーション結果を時系列に並べることにより文書論理要素の読み順を抽出する．

論理構造解析. まず，順序付けがなされた文書論理要素をフラットな木構造に変換する．この木構造において，見出しグループ，著者グループ，章節グループ，リストグループ，表グループ，図グループ，ヘッダグループ，フッタグループなどの論理要素の部分集合を検出すると共にそれらを部分木で記述することにより，階層的な文書構造を生成する．

DOM ツリー変換. パラグラフ，章見出し，箇条書き要素，図，表要素などの論理要素に対して XHTML タグ <p>，<h1>～<h6>，<li>，<img>，<td> をそれぞれ付与し，表の列，表構造，リスト構造などの部分構造に対して <tr>，<table>，<ol>(<ul>) などの XHTML タグをそれぞれ付与する．また，文書固有の論理要素や論理構造に対しては XHTML の要素と class 属性を組み

合わせたタグを付与する．このようにして構成されたXHTMLタグつき木構造をDOMツリーに変換する．

図２に，東芝レビューのフロントページに対して文書画像理解処理を適用した例を示す．

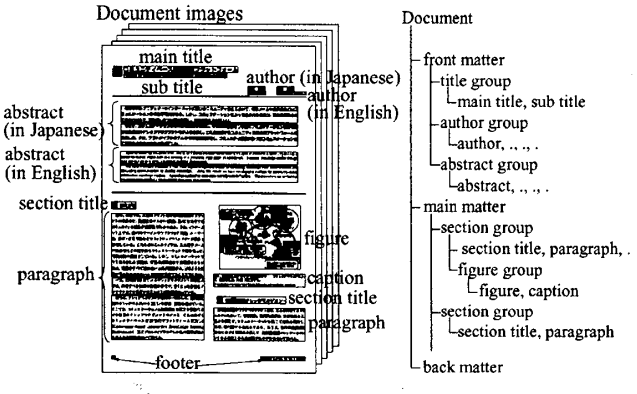

図 2: 文書画像理解の例

5. 文書構造変換

文書構造変換処理は，表層表現抽出（Information extraction），文書構造詳細化（Document structure modification），文書構造検証（Document structure verification）で構成されている [7]．以下に各構成要素の概要について説明する．

表層表現抽出. 表層表現抽出では，文書画像理解処理で得たDOMツリーからテキストノードを収集すると共に，キーワード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表層表現を抽出する．キーワード辞書には，文書中から抽出すべき見出し語やキーワードなどがあらかじめ定義されているものとする．さらに，表層表現抽出結果に対して表層表現抽出ルールを適用することにより，文書要素に初期タグと呼ばれる便宜的なタグを付与する．特許公報中の表層表現に対して初期タグを付与した例を図3(a)に示す．

文書構造詳細化. 構造詳細化では，表層表現および文書要素へのタグ付け結果に対して構造詳細化ルールを順次適用することにより，応用規格に基づいた文書構造の詳細化を行う（図3 (b) 参照）．提案方式では，構造詳細化ルールを適用して得た中間的なタグ付け結果に対してさらに構造詳細化ルールをくり返し適用することを可能としている．このように構造複雑化処理を多段に実施することにより，簡単な文書構造を複雑な文書構造に変換することを実現している．

文書構造検証. 入力文書とターゲットXML文書において文書内容の出現順序が異なる場合には，構造詳細化処理では局所的な範囲においてのみ文書型定義に基づいたタグ付けが実施されていると見なすことができる．このため構造検証処理では，構造詳細化処理で得られたタグ付け結果に対して整合性獲得ルールを適用して部分文書構造の並べ替えを行う．さらに，表層表現抽出処理で付与した初期タグや構造詳細化処理で付与した中間的なタグを削除する整形処理を実施する．この結果，文書型定義に基づいた厳密なタグ付け結果が得られることになる．文書構造検証処理で得られたタグ付け結果の例を図3 (c) に示す．

```
<p><_FI>【ＦＩ】</_FI></p>
<p>G06F 17/21 546 Z</p>
<p>530 A</p>
<p>536 A</p>
<p>G06T 7/40 100 C</p>
<p>G06T 7/42 120 A</p>
<p><_req_for_exam>【審査請求】</_req_for_exam>未請求</p>
<p><_number-of-claims>【請求項の数】</_number-of-claims>１８</p>
<p><_filing-form>【出願形態】</_filing-form>ＯＬ</p>
<p><_inventor>【発明者】</_inventor></p>
<p><name><_name>【氏名】</_name>○○　○○</name></p>
<p><address><_address_>【住所又は居所】</_address_>神奈川県川崎市
幸区小向東芝町１番地株式会社東芝研究開発センター内</address></p>
<p><_agent>【代理人】</_agent></p>
<p><_reg-num>【識別番号】</_reg-num>１００００５８４７９</p>
<p><_attorny>【弁理士】</_attorny></p>
```

(a) 表層表現抽出の例

```
<classification-national>
<main-clsf>G06F17/21546</main-clsf>
<further-clsf>G06F17/21530</further-clsf>
<additional-info>G06F17/21536</additional-info>
<additional-info>G06T7/40100</additional-info>
<additional-info>G06T7/42120</additional-info>
</classification-national>
<request-for-examination><_req_for_exam>【審査請求】
</_req_for_exam>未請求</request-for-examination>
<number-of-claims><_number-of-claims>【請求項の数】
</_number-of-claims>１８</number-of-claims>
<jp:filing-form><_filing-form>【出願形態】</_filing-form>ＯＬ</jp:filing-form>
<inventors><_inventor>【発明者】</_ inventor><name>○○　○○
</name><address><_address_>【住所又は居所】</_address_>
神奈川県川崎市幸区小向東芝町１番地株式会社東芝研究開発
センター内</address></inventors>
<agent><p><_agent>【代理人】</_agent></p>
<registered-number><_reg-num>【識別番号】</_reg-num>
１００００５８４７９</registered-number>
<attorny><_attorny>【弁理士】</_attorny></attorny> </agent>
```

(b) 構造詳細化の例

```
<classification-ipc>
<main-clsf>G06F17/21546</main-clsf>
<further-clsf>G06F17/21530</further-clsf>
<additional-info>G06F17/21536</additional-info>
<additional-info>G06T7/40100</additional-info>
<additional-info>G06T7/42120</additional-info>
</classification-ipc>-
<request-for-examination>未請求</request-for-examination>
<number-of-claims>１８</number-of-claims>
<jp:filing-form >ＯＬ</jp:filing-form>
<inventors> <inventor> <name> ○○　○○ </name>
<address> 神奈川県川崎市幸区小向東芝町１番地株式会社東芝研究
開発センター内</address>
</inventor>
<inventor> <name> □□　□□ </name>
<address> 神奈川県川崎市幸区小向東芝町１番地株式会社東芝研究
開発センター内</address>
</inventor></inventors>
<agent> <registered-number> １００００５８４７９</registered-number>
<attorny /> </agent>
```

(c) 構造検証の例

図 3: 文書構造変換の例

6. 実験結果

本論文で提案したXML文書変換システムをPC上にソフトウェアとして実装した．本研究では，実際の業務で用いられている紙文書を特定の応用規格に基づいたXML文書に変換する実験を行った．実験では，(1) 文書画像理解の精度，(2) 文書構造変換の精度，(3)XML文書変換時間を計測して，提案方式の有効性を評価した．以下に，それぞれの実験結果について述べる．

文書画像理解の性能評価では，名刺，医薬品添付文書，

技術論文，約款，規定集，法令集などから収集した 300 枚の文書画像を用いて文書論理要素の抽出精度を計測した．その結果，300 枚の文書画像に含まれる 10655 個の文書論理要素のうち 10498 個を正しく抽出することができ，98.5%の処理精度を得ることができた．この場合，文書論理要素の抽出誤りは文字認識誤りとレイアウト解析誤りに起因するものがほとんどであった．このような文書要素の抽出誤りを手作業で修正した結果，論理構造解析処理ですべての文書の木構造を正しく抽出することができた．

XML 文書変換精度の評価では，A4 サイズの紙文書 8 ページ相当の事務規定文書 1 セットと，A4 サイズの医薬品添付文書 5 セット（1 セットあたり 3〜5 ページ相当）を対象として実験を行った．その結果，それぞれ 94.1%と 87.4%の変換精度が得られた．実験では，表層表現の抽出誤りや未抽出，変換ルールの適用誤りや未適用などに起因する文書変換誤りが生じた．表層表現抽出では，本方式で採用したキーワード辞書において，バリエーションを伴う数量や性質などの表現と医薬品名や物質名などの固有表現を適切に扱うことができなかったために表層表現抽出誤りが生じた．また文書変換では，変換ルールをボトムアップに適用して一連の変換処理を実施するようになっていたため，同一の表層表現に対して複数の構造化処理が存在する場合に文書変換誤りが生じた．

本実験では，3 ページ相当の医薬品添付文書 1 セットを対象として，手作業による XML 文書変換の時間と本方式による XML 文書変換の時間をそれぞれ計測した．手作業による XML 文書変換では，紙文書の内容をキーボード入力してプレーンテキスト化する作業に 57 分要し，XML エディタを用いてプレーンテキストを応用規格に基づいた XML 文書に変換する作業に 185 分要したため，合計 242 分かかった．一方，本システムによる XML 文書変換を実施した場合では，文書画像理解と誤り結果の修正によりクリーンデータで構成される DOM ツリーを作成する作業に 7 分を要し，文書構造変換後に誤り箇所を修正してターゲット XML 文書を作成する作業に 16 分を要したため，合計で 23 分かかった．この結果，提案方式を用いることにより，手作業の 1/10 の時間で既存文書をターゲット XML 文書に正しく変換できることが分かった．

7. まとめ

本論文では，法令集，特許公報，約款集，規定集，論文，名刺などの紙文書を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した．本システムは，紙文書（文書画像）から文書内容と文書論理構造を抽出し，DOM ツリーに変換する文書画像理解と，DOM ツリーを応用規格に基づいて構造化してターゲット XML 文書を生成する文書構造変換で構成されている．本方式では，対象文書と応用規格の組み合わせに対して，文書モデル，キーワード辞書，文書変換ルールをあらかじめ準備しておけば，上述した XML 文書変換を自動化することが可能となった．その結果，手作業による XML 文書変換の工程と比較して 1/4〜1/10 の作業時間で既存文書を応用規格に基づいた XML 文書に変換することが可能となった．また，文書内容，XML 技術，XML 応用規格などに関する専門的知識を持たないオペレータでも XML 文書を作成することが可能となった．

参考文献

[1] *XSL Transformations*. http://www.w3.org/TR/xslt.

[2] H. Ahonen, B. Heikkinen, O. Heinonen, and M. Klemettinen. Printing structured text without stylesheets. In *Proc. of XML Scandinavia 2000*, 2000.

[3] M. Aiello, C. Monz, L. Todoran, and M. Worring. Document understanding for a broad class of documents. *IJDAR*, Vol. 5, No. 1, pp. 1–16, 2002.

[4] O. Altamura, F. Esposito, and D. Malerba. Transforming paper documents into XML format with WISDOM++. *IJDAR*, Vol. 4, No. 1, pp. 2–17, 2001.

[5] S. Ariyoshi. A character segmentation method for Japanese printed documents coping with touching character problems. In *Proc. of ICPR*, Vol. 2, pp. 313–316, 1992.

[6] D. Freitag. Information extraction from HTML: application of a general machine learning approach. In *Proc. of AAAI*, pp. 517–523, 1998.

[7] 布目光生, 石谷康人, 住田一男. 表層表現抽出と文書構造解析に基く XML 文書変換システム. 情報処理学会研究報告, No. 2004-DD-46, pp. 1–8, 2004.

[8] T. Iijima, H. Genchi, and K. Mori. A theory of character recognition by pattern matching method. In *Proc. of IJCPR*, pp. 50–56, 1973.

[9] Y. Ishitani. Document skew detection based on local region complexity. In *Proc. of ICDAR*, pp. 49–52, 1993.

[10] Y. Ishitani. Logical structure analysis of document images based on emergent computation. In *Proc. of ICDAR*, pp. 189–192, 1999.

[11] 石谷康人. データ駆動型処理と概念駆動型処理の相互作用による文書画像レイアウト解析. 情報処理学会論文誌, Vol. 42, No. 11, pp. 2711–2723, 2001.

[12] Y. Ishitani. Model-based information extraction method tolerant of OCR errors for document images. *IJCPOL*, Vol. 15, No. 2, pp. 165–186, 2002.

[13] Y. Ishitani. Document transformation system from papers to XML data based on pivot XML document method. In *Proc. of ICDAR*, Vol. 1, pp. 250–255, 2003.

[14] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Trans. on PAMI*, Vol. 15, No. 7, pp. 737–747, 1993.

[15] E. Kuikka, P. Leinonen, and M. Penttonen. An approach to document structure transformations. In *Proc. of conference on software: theory and practice*, pp. 906–913, 2000.

[16] E. Kuikka, P. Leinonen, and M. Penttonen. Towards automating of document structure transformations. In *Proc. of DocEng*, 2002.

[17] N. Kushmerick, D. S. Weld, and R. Doorenbos. Wrapper induction for information extraction. In *Proc. of IJCAI*, pp. 729–737, 1997.

[18] J. Y. Lee, J. S. Park, H. Byun, J. Moon, and S. W. Lee. Automatic generation of structured hyperdocuments from document images. *Pattern Recognition*, Vol. 35, No. 2, pp. 485–503, 2002.

[19] 建石由佳, 伊東伸泰. 確率文法を用いた文書論理構造の解釈法. 電子情報通信学会論文誌, Vol. J79-D2, No. 5, pp. 687–697, 1996.

[20] S. Tsujimoto and H. Asada. Major components of a complete text reading system. *Proc. of the IEEE*, Vol. 80, No. 7, pp. 1133–1149, 1992.

[21] M. Worring and A. W. M. Smeulders. Content based internet access to paper documents. *IJDAR*, Vol. 1, No. 4, pp. 209–220, 1999.