文書の様々な活用を可能にするXML構造化技術

XML Structuring Technology for Various Types of Document Applications

布目　光生　　石谷　康人　　後藤　和之
■ FUME Kosei　■ ISHITANI Yasuto　■ GOTO Kazuyuki

オフィスでは膨大な情報が日常的に扱われるようになった。そのため，目的に応じて必要とする情報に効率よくアクセスし，情報を適切に処理できる技術への要望が高まっている。

そこで東芝は，自然言語処理技術とXML（Extensible Markup Language）技術を文書構造化技術で結び付けることにより，文書から論理要素，論理構造，意味情報を抽出し，それらをメタデータ(注1)として本文に埋め込むことで，文書内容を目的に応じて活用できるXML構造化技術を開発した。更に，紙文書のXML変換，文書カテゴライゼーション（分類），情報検索インタフェースなど，いくつかの応用を実現した。

The dramatic increase in the volume of electronic documents in the office environment has spurred demand for easy access to information resources and for their effective management.

Toshiba has developed an extensible markup language (XML) document structuring technology that facilitates exploitation of information resources corresponding to these needs. Utilizing natural language processing and XML, this technology makes it possible to extract document attributes, such as logical elements, logical structures, and term semantics, and embed them as machine-processable metadata. We have achieved various applications based on this technology, such as a document transformation system from paper to XML, a document categorization system, and an information access interface.

1　まえがき

情報のデジタル化が進んだことにより，情報の生成と流通が容易になった。企業や自治体では特に，情報としての文書が大量に発生しており，それらを目的に応じて効率よく適切に処理することが求められている。

一般的に文書は，書き手が読み手に対して情報を伝えるために作成するものであり，まずは書き手によって文書の役割，内容，体裁などが決められる。しかし，読み手のほうでは様々な目的を持っていることがあり，必ずしも書き手の意図と読み手の目的が合致するとは限らない。このため，書き手による文書の形態と読み手の利用目的が合致しない場合には，書き手と読み手の間にギャップが生じることになる。

東芝は，このような書き手と読み手の間のギャップを埋めるために，文書のXML構造化技術を開発した。このXML構造化技術ではまず，自然言語処理技術を応用して，文書から，タイトル，章と節の見出し，箇条書きなどの論理要素，章や節の構造と箇条書き構造などの論理構造，及び日付，人名，地名，組織名，金額，数量などの意味情報を抽出する。そして，このようにして抽出した情報をメタデータとして取り扱うとともに，国際標準のXML規格に従って文書中にXMLタグとして埋め込む（**図1**）。読み手による文書閲覧，文書管理，情報検索，フォーム処理などにおいて，メタデータを積極的に利用することにより，様々な目的に応じた文書内容への効率的なアクセスと部分情報の再利用が可能になる。

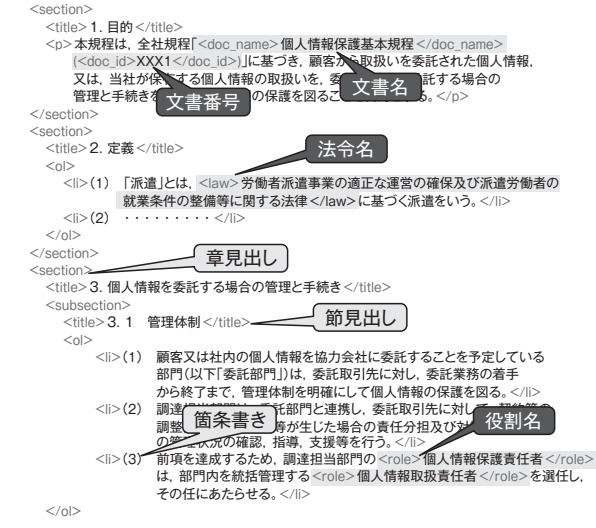


図1. XML構造化の例 — 文書から論理要素，論理構造，意味表現を抽出して，XMLタグとして文書に埋め込む。

Extraction of document logical elements, logical structures, and keywords for XML tags

（注1）　あるデータが付随して持つ，そのデータ自身についての付加的データ。

ここでは，このようなXML構造化のフレームワーク，技術の特徴，及び応用事例について順に述べる。

2　XML構造化がもたらす文書活用の姿

この章では，規程文書を例として用い，XML構造化が目指す文書活用について簡単に述べる。

規程文書は，業務を適切に遂行するための要領やルールを定義した文書であるが，業務が適切になされるために，正確さを重視して硬い表現で記述されていることが多い。また，業務が多岐にわたっている場合には，その内容が大量になってしまうことがある。したがって，いざ利用しようというときには，必要な情報が掲載されている箇所を探しにくく，内容そのものも読みにくいことがある。

このような規程文書をユーザーが利用する場合は，ある目的を持って特定の情報を参照することがほとんどだろう。このような場合，大量の文書から必要とする情報を探し出すとともに，関連する情報も参照しなければならない。そのとき，図1のように規程文書がXML構造化されていれば，データベースの情報管理機能や情報検索機能がそれを有効利用することで，ユーザーの目的に合致した効果的かつ効率的な規程文書の閲覧及び理解が可能になる。

以下に，**図2**を用いて，このような規程文書の活用例について述べる。

(1)　ピンポイントな検索　　論理要素や論理構造に応じて検索対象を絞り込むことにより，適切な箇所を優先的に閲覧することができる（図2の(1)）。

(2)　関連情報の把握　　論理要素間の類似性や論理構造の関連性などから関連する情報をリンク付けすることで，関連情報を一括して閲覧することができる（図2の(2)）。

(3)　重要語を手がかりにしたナビゲート　　統計処理や辞書利用などにより重要語句を検出するとともに，検出した語句を一覧することで，文書内容の理解を促進できる（図2の(3)）。

(4)　文書改廃時のチェック支援　　改廃前後でXML構造化を実施するとともに処理結果の違いを検出することで，用語の不統一などをチェックできるようにして文書の更新作業を支援する（図2の(4)）。

(5)　部分情報の再利用　　XML構造化により得られたメタデータを活用することで，文書間にまたがる関連情報にも効率的にアクセスできる。また，それらの情報を部分的に抜き出して合成することで，文書の効果的な再利用が可能になる（図2の(5)）。

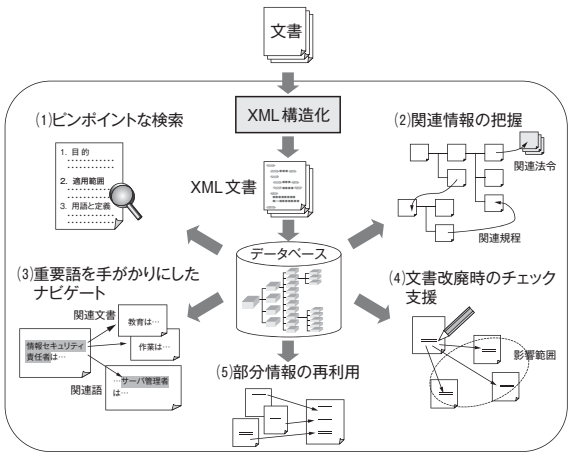


図2．XML構造化がもたらす文書活用の姿 — 文書のメタデータをXMLタグとして文書に埋め込むことにより，絞込み検索，関連情報の把握，ナビゲーション，チェック支援，及び情報再利用が可能になる。

XML applications for business documents

3　XML構造化のフレームワーク

この章では**図3**を用いて，文書のXML構造化を実現するためのフレームワークについて述べる。

このXML構造化では，DOM（Document Object Model）化，形態素解析，文抽出フィルタリング，意味表現抽出，論理構造抽出，文書構造変換などのモジュールによりフレームワークを構成している。それぞれのモジュールをレイヤ（層）として配置し，レイヤ間でDOM形式の情報が共通に流れるようにして，モジュールの自由な組合せにより柔軟に処理できるようにしている。このような仕組みを採用することにより，将来発生するであろう未知の文書フォーマットに対応する場合に，テキスト情報を取り出すためのアダプタを自由に追加できるようになっている。

DOM化モジュールは，PDF（Portable Document Format），CSV（Comma Separated Value），XML/HTML（Hyper

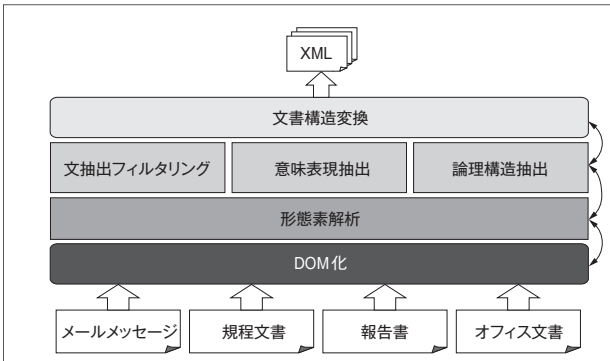


図3．XML構造化フレームワーク — 処理モジュールをレイヤ化することで，モジュールの自由な組合せによる柔軟な処理と新機能の追加が可能になる。

Expandable framework for document structuring

Text Markup Language）などの様々なフォーマットの文書をDOM形式に変換する。この場合，それぞれのフォーマットからテキスト情報を取り出すためのアダプタが必要となる。

形態素解析は，テキスト情報から単語を抽出するとともに，単語に品詞を付与する処理である。この形態素解析の辞書には，典型的な人名，地名，製品名，会社名などの固有名詞も登録されているので，きめ細かい品詞の付与が可能になっている。

文抽出フィルタリングは，後段の処理に不要なハードリターン，スペース，タブなどのコード情報を取り除くとともに，それぞれの文章を抽出する。このとき，不要コード情報を除去したあとに，形態素解析を再度実施することで除去が適切であるかどうかを検証し，それぞれの文章を高精度に抽出している。

意味表現抽出では，品詞や語句の組合せに基づいて，会社名，組織名，人名，地名，URL（Uniform Resource Locator）といった固有表現や，金額，日付，数量などの数値表現などを文章から抽出する。

論理構造抽出は，それまでの処理結果から，見出し，章や節のタイトル，順序付きあるいは順序なし箇条書きなどの文書論理要素を抽出するとともに，章や節の構造及び箇条書き構造といった階層的な論理構造を抽出する。このとき，文書の汎用的な論理構造を表現できるDocBook(注2)というスキーマ（XML文書の形式）に基づいて，論理構造をDOM化する。

文書構造変換では，DocBookスキーマに基づいたDOM形式の文書を，様々なスキーマに従うXML文書に変換する。

4　XML構造化技術の特徴

この章では，XML構造化フレームワークを構成するモジュールのうち，特徴的である意味表現抽出，論理構造抽出，文書構造変換についてそれぞれ述べる。

4.1　意味表現抽出

意味表現抽出は，意味クラス解析[1]と正規表現に基づくパターンマッチング(注3)で構成されている。

意味クラス解析は，あらかじめ作成した500個のルール群を用いて，形態素解析及び文抽出フィルタリングの結果から，114種類の基本固有表現と複合固有表現を抽出する。基本固有表現の抽出では，単語に付与された属性情報から個別の固有表現を抽出する。複合固有表現の抽出では，既抽出の個別固有表現の組合せから複合的な固有表現を抽出する（例：“地名”と“番地”の組合せから，“住所”を抽出する）。

一方，パターンマッチングでは，例えば“個人情報取扱責任者”という意味表現を“/（名詞）+責任者/”というパターンで定義しておき，それを形態素解析及び文抽出フィルタリングの結果に適用することで意味表現を抽出する。

4.2　論理構造抽出

論理構造抽出は，前段の処理で得られた結果から，まず論理要素を見つけ，論理要素の配置関係から論理要素の階層的な構造を抽出する。論理構造抽出の手順を以下に示す。

ステップ1　　空行数や字下げ数に基づいて，論理要素に相当するブロックを抽出する。

ステップ2　　文書タイトルの定型パターン（例：“報告書”，“明細書”），あるいは章や節のタイトルのヘッダ部（例：“2.”，“第2章”），箇条書きのヘッダ部（例：“(1)”，“・”）などがわかっている場合には，それらをあらかじめパターンとして定義しておき，このようなパターン群を用いることでブロックを特定の論理要素として判定する。

ステップ3　　インデントのレベルや論理要素の種別が同じ場合には，連続する論理要素を階層的にグループ化することで論理構造を抽出する[2], [3]。

ステップ4　　抽出した論理要素と論理構造を，DocBookスキーマに基づくDOM形式として記述する。

4.3　文書構造変換

文書構造変換では，あらかじめ定義された文書モデルを論理構造抽出結果に適用することで，変換対象のスキーマに基づいたXML文書を生成する[4]。この文書モデルは，文書構造の複雑化を可能にする構造詳細化ルールと，スキーマに従った文書構造を生成するための構造検証ルールで構成されている。それぞれのルールは，意味表現，論理要素，及び論理構造を手がかりにして，部分的な構造の変換を達成するように定義されている。文書モデルでは，このようなルールを組み合わせるとともに適用手順を定義することで，様々な構造変換を可能にしている。

実際の文書構造変換タスクでは，一度文書モデルを作成することによって，大量の入力文書を希望するXML文書に一括して変換できるようになっている。

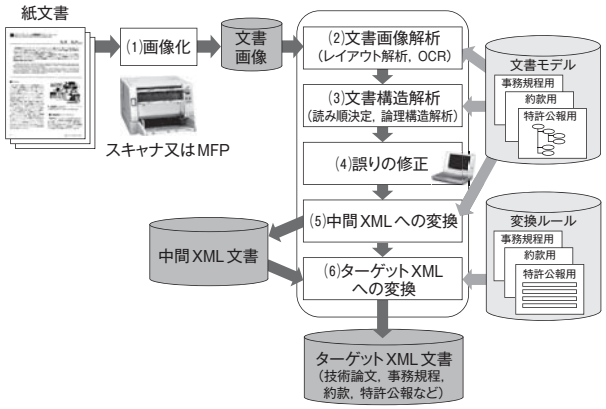
5　XML構造化の応用事例

XML構造化の応用事例として，紙文書のXML変換，文書カテゴライゼーション，及び情報検索インタフェースについて以下に述べる。

5.1　紙文書のXML変換

紙文書をイメージ化したあと，イメージ情報をOCR（Optical Character Reader）によりコード化し，更に，コード情報に対してXML構造化を適用することで，紙文書を既存のスキーマに基づいたXML文書に変換できるようにした[4]。OCR側でもXML構造化を行い[2]，OCRエラーを簡単に修正できるGUI（Graphical User Interface）を提供することで，書籍，学術論

(注2)　XML形式の文書で，タイトル，章や節，段落などの部分構造を定義するための標準的な記法の一つ。
(注3)　二つの文字例のパターンを比較し，両者が同類であるかどうかを調べること。



MFP：Multifunction Peripheral（複合機）

図4．紙文書のXML変換の構成 ― イメージ化した紙文書にOCR処理を適用し，その結果にXML構造化を適用することで，スキーマに応じたXML文書を生成できる。

Configuration of transformation process using optical character reader (OCR) to generate XML documents

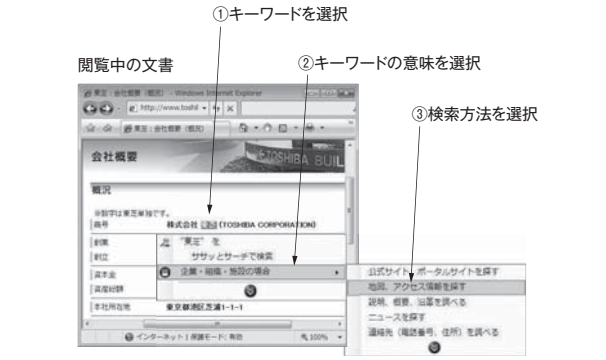


図5．情報検索インタフェースの例 ― 閲覧文書において，キーワードをマウスでなぞると，キーワードの意味とその検索方法のメニューが表示される。

Example of search interface chaining

文，名刺，公文書，規程・約款，特許明細書，医薬品添付文書などの紙媒体を，効率よくXML文書に変換できるようにした（**図4**）。

5.2　文書カテゴライゼーション[3]

キーワードと文書種別の組合せで構成される辞書を用いることで，入力文書群を文書種別に応じて高精度に分類する，モデルベースト文書カテゴライゼーションを実現した。この事例では，入力文書において論理要素と論理構造を抽出し，その結果に対して辞書を適用する際に文書モデルにより適用箇所をコントロールすることで，高精度な文書カテゴライゼーションを実現している。

5.3　情報検索インタフェース[1]

Webブラウザやオフィスアプリケーションで閲覧している文書において，気になる語句があれば，それをマウスやペンなどのポインティングデバイスでなぞるだけで，関連情報を的確に検索できるインタフェース技術を実現した。このインタフェース技術では，ユーザーがキーワードをなぞったとき，キーワードの意味を特定するとともに，その意味に応じた検索方法の候補をユーザーに提示することを特徴としており，ユーザーは提示された検索方法を選ぶだけで関連情報を的確に検索できるようになっている（**図5**）。

6　あとがき

文書から意味情報，論理要素，論理構造を抽出し，それらをXMLで記述するとともにメタデータとして本文に埋め込むことで，文書内容を目的に応じて閲覧，理解，及び再利用できるようにする，XML構造化技術について述べた。この技術は，当社の伝統的な自然言語処理技術と世界標準のXML技術を独自の文書構造化技術で結び付けたもので，これによって，文書を適切に活用するための一つのソリューションを提供できるようになった。

今後は，XML構造化と応用事例を拡張し，必要となる要素技術を追加していくことで，ユーザーがより効果的かつ効率的に文書情報を活用できるようになることを目指す。

文　献

(1) 鈴木　優，ほか．連鎖検索インタフェース“ササッとサーチ$_{TM}$”．東芝レビュー．**62**, 12, 2007, p.54－57.
(2) Ishitani, Y., et al. "Document transformation system from papers to data based on pivot XML document method". Proc. ICDAR2003. Edinburgh, Scotland, UK, 2003-08, IAPR. p.250－255.
(3) Fume, K., et al. "Model-based document categorization employing semantic pattern analysis and local structure clustering". Document Recognition and Retrieval XV. Proc. of the SPIE. San Jose, California, USA, 2008-01, SPIE. p.68150R－68150-8.
(4) 布目光生，ほか．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．情報処理学会研究報告．**2004**, 97, 2004, p.1－8.

布目　光生　FUME Kosei
研究開発センター 知識メディアラボラトリー研究主務。
文書構造化技術及び知識抽出技術の研究・開発に従事。
人工知能学会，情報処理学会会員。
Knowledge Media Lab.

石谷　康人　ISHITANI Yasuto, D.Eng.
東芝ソリューション(株) IT技術研究所 研究開発部主任研究員，工博。ビジネスインテリジェンス技術の研究・開発に従事。
IEEE，電子情報通信学会，情報処理学会会員。
Toshiba Solutions Corp.

後藤　和之　GOTO Kazuyuki
東芝ソリューション(株) IT技術研究所 研究開発部研究主務。情報検索，文書の自動分類，ナレッジマネジメントなどの技術開発に従事。情報処理学会会員。
Toshiba Solutions Corp.