既存紙文書からＳＧＭＬ文書への変換システムの試作

4Ｊ－9

森口 修，今村 誠，丸田 裕三，鈴木 克志
（三菱電機株式会社 情報技術総合研究所）

１．はじめに

高度情報化社会において増大する情報の中にあって、文書情報の占める割合は高い。従来から OA 戦略の一環として、文書を機械的に効率よく処理することを目的とした文書ファイリングシステムを製品化している。既存文書の大半は紙媒体であるため、文書ファイリングでは紙媒体と電子メディアとの間の相互変換も重要な機能のひとつとなっている。

そのような状況の中で、文書の電子化形式として SGML(Standard Generalized Markup Language: ISO 8879)が最近注目を浴びている。本論文では、既存の紙媒体の文書を SGML 形式に対応した電子メディアへ変換するための方法を検討し、試作した結果について述べる。

２．目的

計算機による情報処理を目的として紙媒体を電子化する場合、以下の観点を重視する必要がある。

(1) 高度な情報処理を可能とする電子化形式
- 文書を管理するための論理構造記述
- 運用上の変更に容易に対応できる柔軟性
- 文書交換・再利用効率向上のための標準化

(2) 電子化に要するコストの最小化
- 文書画像処理、テキスト解析による自動化
- 作業のインタフェースによる効率化

変換先の形式を SGML とすることによって(1)の条件をみたすことができる。(2)の観点からは自動化と使いやすいインタフェースという２つのアプローチを検討する必要がある。

An experimental system for conversion from legacy documents into SGML
Osamu MORIGUCHI, Makoto IMAMURA,
Yuzo MARUTA, Katsushi SUZUKI
Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura, Kanagawa 247, Japan

３．方式の検討と試作

３．１ 方針

従来、非定型フォーマットに柔軟に対応する文書画像処理方式[1]や、文字情報の形態パターンをテキスト解析することにより文書構造を抽出する方式[2]などがある。しかし、すべての種類の文書を解析するための一般的なモデルの構築は困難である。また、文書全体をテキスト解析するには、構造の曖昧さが大きいという問題がある。さらに、SGML においては DTD(Document Type Definition)によって文書の論理構造を定義するため、解析モデルは DTD と対応付けて構築する必要がある。

そこで、DTD を大きく２つの階層に分け、それぞれに文書画像解析およびテキスト解析を対応付けてトップダウンに文書を解析する方針とする。

上位 DTD は文書中の位置や枠で囲まれた領域など文書のレイアウトとの関連性が強いため、上位 DTD に対応した文書画像解析モデルを定義し、レイアウト解析によって上位の論理構造を抽出する。種類が多く、フォーマット変更が頻繁に発生するビジネス文書に柔軟に対応するには、文書画像解析モデルの作成は容易である必要がある。

下位 DTD は複数の上位 DTD から参照され、箇条書きなどの一般的な書式との関連性が強いため、下位の DTD に対応したテキスト解析モデルを定義し、テキスト解析によって下位の論理構造を抽出する方針とする。この場合、テキスト解析は上位の文書構造が解析された後の各論理構造要素に対して局所的に行ない、かつトップダウンに与えられる構造の制約のもとで行なえばよい。

３．２ システム構成

既存文書を SGML 形式に変換するシステムの構成とデータの流れを図１に示す。

文書レイアウトと上位 DTD とを対応付けるイン

タフェースによって作成された対応表をもとに文書画像のレイアウトが解析され、上位 DTD に対応したタグ付けがなされた文字情報が出力される。

修正インタフェースで文字認識結果を確認した後、各領域の文字情報がテキスト解析され、下位の DTD に対応するタグ付けがなされた SGML 文書が出力される。テキスト解析部は現在検討中である。

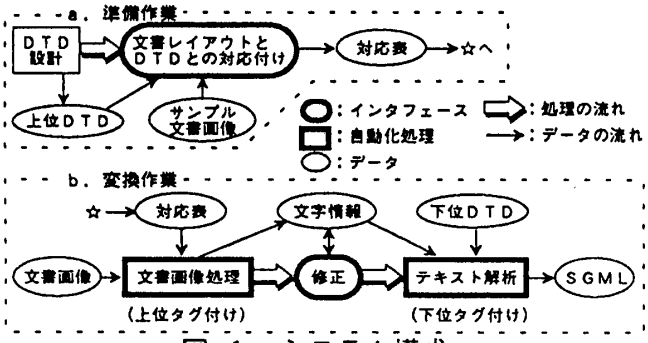

図 1．システム構成

3．3 対応付けインタフェース

文書レイアウトと上位の DTD との対応付けインタフェースを図2に示す。左側に文書画像のサンプルを表示し、レイアウト領域を設定する。同時に、文書画像処理における領域抽出方法、文字認識方法も各領域毎に定義する。右側に DTD で定義された文書構造を木の形で表示し、文書構造要素を選択する。文書レイアウトにおける各領域と DTD における各文書構造要素とを対応付けることによって対応表を作成する。対応付け作業は GUI を用いてビジュアルに確認しながら簡単な操作で行なうため、多くの種類の文書や文書フォーマットの変更に柔軟に対応することができる。

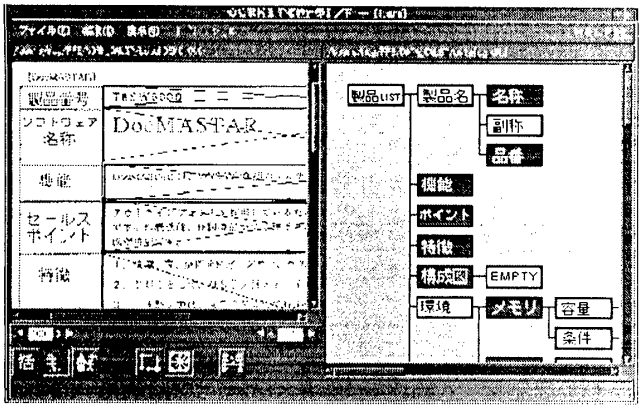

図 2．対応付けインタフェース

3．4 文書画像処理

対応付けインタフェースで作成した対応表をもとに同一種類の文書は一括して文書画像処理される。対応表で指定されたレイアウト抽出方法によって各レイアウト領域が抽出され、上位 DTD で定義された論理構造要素が割り当てられる。各レイアウト領域は文字認識によって文字情報に変換され、この段階では各論理構造要素の仮のインスタンスとなる。

3．5 テキスト解析処理

文書画像解析において上位の各論理構造要素の仮インスタンスとなっている文字情報から、テキスト解析によって下位 DTD で定義された論理構造を抽出し、下位 DTD に対応するタグ付けを行なう。

各領域毎に論理構造を限定することが可能となるため、構造の曖昧性を減少させることができる。例えば領域内の書式が箇条書きであった場合、「組み立て手順」という領域であれば順序構造に、「互換部品」という領域であれば並列構造に限定可能である。

4．まとめ

DTD をその階層によって上位・下位に分割し、それぞれに文書画像解析およびテキスト解析モデルを対応させ、文書構造の上位から下位方向にトップダウンな解析を行なうことによって既存紙文書を SGML 形式に変換する方式を検討した。

上位 DTD と文書レイアウトとを対応させる GUI を試作し、これによって解析モデルの作成が容易となったため、文書の種類に応じた柔軟な対応が可能となった。

今後は、実際の複雑な DTD に対応した SGML 形式への変換のためのテキスト解析モジュール、再帰や繰り返し構造を許容する高度なレイアウト解析との組み合わせを検討する。

参考文献

1) 東野，藤澤，中野，江尻：矩形領域の集合表現に基づく知識表現言語ＦＤＬと文書画像理解への応用，電子通信学会技術研究報告，Vol.86, No.95, pp.11-20(1986).
2) 土井，福井，山口，竹林，岩井：文書構造抽出技法の開発，電子通信学会論文誌，D-II，Vol.J76-D-II，No.9，pp.2042-2052(1993).