電子申請システムのための文書構造化技術に関する研究（042308001）

A Study on Structuring Document Forms for Electronic Application Systems

研究代表者

浅田尚紀　広島市立大学大学院　情報科学研究科

Naoki Asada, Graduate School of Information Sciences, Hiroshima City University

研究分担者

椋木　雅之†　青山　正人†　斉藤 将樹††　山本　武司†††

Masayuki Mukunoki†　Masahito Aoyama†　Masaki Saitoh††　Takeshi Yamamoto†††

†広島市立大学大学院　情報科学研究科

††日本アイ・ビー・エム(株)　ソフトウェア開発研究所

†††（株）ミウラ

†Graduate School of Information Sciences, Hiroshima City University

††Software Development Lab., IBM Japan

†††MIURA Co.

研究期間　平成 16 年度～平成 18 年度

本研究開発の概要

本研究では、各種申請文書を対象に、文書書式の構造化技術の開発と電子申請システムのためのＸＭＬ記述による文書処理支援システムの開発を行った。申請書の機能を対話型文書としてモデル化し、欄の機能分類と書式構造文法を用いた文書構造解析、ＸＭＬによる文書構造記述と電子フォームの自動生成等の技術を開発した。これにより、多様な書式をもつ実在する申請文書を、効率的に電子化して管理・加工することが可能となる。文書電子化支援システムを試作し、広島市・広島県等地元自治体の協力を得て 1388 の実文書を対象に実証実験を行い、多くの文書に対して本システムが適用でき、申請文書の電子化作業の省力化に寄与することを示した。

Abstract

We have developed a structuring method of interactive document forms using the form structure grammar and proposed a XML based description called IDML (Interactive Document Markup Language) designed for the structured documents. We also have developed a prototype system that converts the conventional application forms to the electronic ones with form definition tables. Experiments using 1388 kinds of application forms were performed and the results have shown that the system enables us to process application forms efficiently.

１．まえがき

本研究では、申請文書を対象に、文書書式の構造化技術の開発と電子申請システムのためのＸＭＬ（Extensible Markup Language）記述による文書処理支援システムの開発を目的とする。

電子政府、電子自治体を実現する上で多様な書式をもつ申請書の電子化は最も重要な課題の一つである。文書を電子化することの利点は、従来の印刷文書がインターネットを通じて電子的に表示、送信できることだけではなく、文書の意味、機能、レイアウトに基づく構造化が可能となり、文書の作成、表示、記入、情報抽出の文書処理が省力化、効率化できることにある。一般的な文書の構造解析については従来から研究されてきたが、申請文書を対象とした構造化とその構造記述に基づく電子申請システムの設計についての議論は十分にはなされていなかった。

本研究では、申請文書を対話型文書モデルという新しい観点でとらえることにより、文書に内在する普遍的な対話規則を書式構造文法として記述することを行う。そして申請文書を、欄の指示・被指示情報、欄の論理的な構造情報、欄配置のレイアウト情報、記載された記入情報の４要素で構造化することにより、文書書式のデータベース化、多様な書式の相互変換および標準化、記入情報の検索と再利用、印刷書式から電子申請書式への自動変換、そして印刷申請文書と電子申請文書の相互変換を実現する文書処理支援システムの開発を目指すものである。本稿では申請文書構造化処理の概要と評価実験の結果について述べる。

２．研究内容及び成果

2.1　申請文書の構造化処理

申請文書の構造化処理とは、文書内の記入領域と記入内容に関する指示とを対応付けることにより、申請文書のもつ質問-回答型の構造を抽出する処理のことである。構造化により文書は適切にタグづけされ、計算機によりその内容を容易に管理、保存、加工することが可能となる。

図１に示すように、申請文書には罫線で囲まれた罫線領域とそれ以外の無形線領域が存在する。罫線領域では、最小の矩形である欄が一つの記入領域または指示を表しており、構造化処理の単位となる。一方、無罫線領域では、記入領域は単なる空白であり、指示文字列との位置関係が構造解析において手がかりとなる。このように、罫線領域と無形線領域では、性質が異なるため、それぞれの領域に対して構造化処理を行い、結果を統合するというアプローチをとった(図２参照)。

与えられた対話型文書に対して、まず、幾何情報抽出を行う。幾何情報は、文書中のボックスや文字の位置、大きさ、文字コードの情報からなる。幾何情報は、対話型文書の表示上のレイアウトを表しており、これを用いて文書をディスプレイ上やプリンタなどに同じレイアウトで表示／印刷することができる。

次に、罫線領域には、書式構造文法を用いた構造解析処理を、それ以外の無罫線領域については、ブロック化による構造解析処理を適用する。この際、ユーザが記入場所を記入領域として指定する。指定された記入領域に基づき、

ボックスや文字列の分類を行い、構造解析する。

罫線領域の解析結果を、無罫線領域の一項目として統合することにより、文書全体の構造解析木を得る。構造解析木は、各記入領域とそれに隣接する指示との関係を、階層的に示したものとなる。これを構造情報と呼ぶ。

この構造情報に対して、階層的な指示関係をたどる指示関係解析を行うことで、質問-回答型の内容をもつ指示情報を得る。これにより、罫線文書、無罫線文書、混在型文書のいずれに対しても、文書構造の解析が可能となる。

幾何情報、構造情報、指示情報は、XML を用いて記述する。これに記入情報（電子申請の申請時に利用者が記入する情報）を加えた４つの情報が、電子化された対話型文書のもつ情報となる。

作成したシステムのインターフェースを図３に示す。

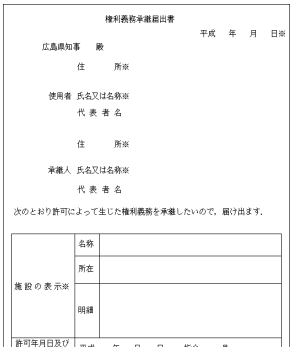

図１　申請文書の例

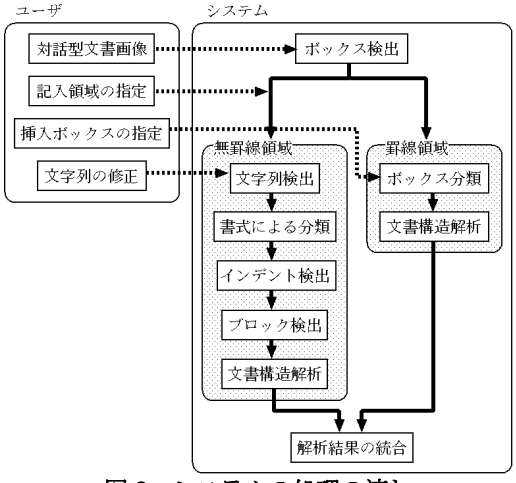

図２　システムの処理の流れ

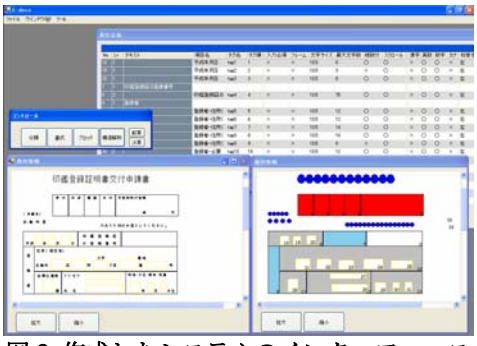

図３ 作成したシステムのインターフェース

2.2 本システムを用いた申請文書電子化の実証実験

(1) 実験方法

作成したシステムの実用性を評価するために、実際に使用されている対話型文書を処理する実験を行った。実験対象は、広島県が既に電子化作業を行った 1388 文書である。

これらの文書について、電子化作業前の PDF ファイルから、本システムを使って記入領域を設定し、同時に、様式項目定義書を作成した。処理結果は、得られた指示関係が適切であるかどうか目視で判定した。自動処理のみで適切な処理結果が得られない場合は、適宜、ボックス種別の修正および文字列抽出の修正を行った。

(2) 実験結果と考察

実験結果を表 1 に示す。1388 文書のうち、1105 文書については、適切な指示関係が得られ、本システムを用いて文書の電子化が行えた。

次に、本システムにより、実際にどの程度の作業効率の向上が得られるか、簡単な評価実験を行った。実験では、罫線文書、無罫線文書、混在型文書の 3 文書に対して、本システムを使った電子化と、広島県等で実施した Adobe Acrobat Professional(以下 Acrobat) を使った電子化作業との作業時間を比較した。Acrobat を使った作業では、様式項目定義書は既に与えられており、それに合わせた記入領域の設定を行った。

被験者は 20 代の学生２名で、本システムの操作は習熟している。Acrobat の操作については、事前に 2 文書に対して電子化作業の練習を行った。結果を表 2 に示す。

この結果より、実作業においても、本システムを使うことで、作業効率が大きく改善すると考えられる。

表１　適用結果

処理不能 111		誤処理 172				
配置の問題	記入かつ指示	幾何抽出失敗	構造解析失敗	階層失敗	成功	計
65	46	115	22	35	1105	1388

表２　作業時間の比較

	被験者 A		被験者 B	
	本システム	Acrobat	本システム	Acrobat
文書１（罫線）	4 分 01 秒	35 分 21 秒	1 分 49 秒	28 分 07 秒
文書２（無罫線）	1 分 23 秒	15 分 27 秒	2 分 00 秒	15 分 52 秒
文書３（混在型）	3 分 14 秒	24 分 01 秒	3 分 13 秒	30 分 44 秒

３．むすび

各種申請文書の電子化を支援するシステムを構築し、実証実験により有効性を示した。今後は、文書構造を考慮して申請文書の新規生成を支援する対話型ワープロへの応用や、IDML により記述された記入済み電子文書の処理など、実用化に向けたさらなる取り組みを進めていく予定である。

【誌上発表リスト】

[1]A.Amano, N.Asada, M.Mukunoki, “Modification table form generation system based on the form recognition”, International Conference on Pattern Recognition(Cambridge, UK)(2004/8/23)

[2] A.Amano, N.Asada, M.Mukunoki, M.Aoyama, "Table Form Document Analysis based on the Dcument Structure Grammar", International Journal on Document Analysis and Recognition, pp.201-213, 2006-06.

[3] 青山正人，小迫政幸，椋木雅之，浅田尚紀，“同一目的の対話型罫線文書からの標準文書構造の抽出”, 情報科学技術レターズ(FIT2006), LI-008 , 2006-09

【申請特許リスト】

[1] 浅田尚紀，椋木雅之，青山正人，“文書構造情報の作成方法（特願 2004-375548）”，申請国日本, 2004/12/27

【報道発表リスト】

[1] “電子申請の新技術開発目指す”，広島経済レポート, 2004/11/1