LI_008

同一目的の対話型罫線文書からの標準文書構造の抽出

Extraction of Standard Document Structure from Identical Objective Form Documents

青山 正人†，小迫 政幸†，椋木 雅之†，浅田 尚紀†
M. Aoyama, M. Kosako, M. Mukunoki, and N. Asada

1. はじめに

われわれは，現在急ピッチで各自治体ごとに進められている紙ベースの各種申請書の電子化を支援する電子申請システムための文書構造化技術の開発を進めている [1]．本研究は，自治体の統合も進む中，同一目的 (印鑑登録証明など) の申請文書であるにも関わらず，自治体ごとに様式が異なっている既存文書から標準文書構造を抽出する枠組と文書間距離の導入による客観的な評価基準を提案することを目的とする．本手法により，従来の既存文書を生かしながら，異なる様式の標準化を効率的に進められるだけでなく，県や国レベルでの文書の標準化，さらには，申請された文書からの記入情報の抽出や保存，管理など運用上の効率も大幅に改善されることが期待できる．

2. 対話型罫線文書

対話型罫線文書とは，図1(a)に示すように文書中の欄 (ボックスと呼ぶ) が罫線によって区切られ，情報の記入を通して，文書作成者と利用者の間で記入情報の受け渡しが成立している文書のことをいう．そして文書構造とは，ボックス間の隣接関係によって与えられる記入ボックスへの指示内容を表すものとする．この文書構造は，われわれが従来より提案している書式構造文法を用いて抽出することができる [2]．具体例として，図1(a)の各文書に対して，利用者が記入する領域部分の文書構造を示したものが図1(b)である．

3. 文書間距離と標準文書構造の抽出

異なる様式の同一目的文書では，文書構造の違いだけではなく，ボックス内の指示内容の表現も異なる場合が多い．そこでまず，このような文書構造の本質ではない部分の違いを，次に示す手作業による前処理によって排除する．1) 各対象文書中に含まれる指示内容に無関係な文字を削除する (図1(a)(広島市) 中の「広島市…」など)．2) 文字を挿入するボックスからも指示内容を抽出する (図1(a)(広島市) 中の「住所（居住地）」など)．3) 異なる表現を統一する (例えば,「氏名」,「名前」を「氏名」に統一する)．4) 図1(a)(広島市) 中の「登録者の住所」のような多段階の指示は「登録者.住所」,「住所」に展開するとともに，同一構造をまとめる．

この前処理結果に対し，二つの文書構造を比較評価する尺度として，文書間距離 $d(x,y)$ を，$d(x,y) = \mathrm{xor}(x,y)/\mathrm{or}(x,y)$ と定義する†．ここで，x,y は各文書構造の要素を表し，xor，or はそれぞれ，二つの文書構造の要素に対する排他的論理和，論理和の要素数を意味する．これは，文書構造の要素の並び順を無視した場合のレーベンシュタイン距離 (編集距離) に相当する xor を，文書構造を構成する指示内容数に相当する or で正規化することにより，各文書構造の要素数による影響をなくそうとしたものである．

一方，標準文書構造は，各文書構造の共通部分を抽出した要素 (論理積) と定義する．このとき，すべての対象文書間で共通する文書構造の要素数をカウントし，対象文書数の過半数を超える指示内容から標準文書構造を抽出する．このとき多段階の指示と展開後の指示のカウントが同じ場合は，多段階の指示がかかる方を優先する．

4. 実験

広島県内の7自治体の印鑑登録証明 (広島市，府中町，呉市，熊野町，江田島市，三次市，東広島市) を対象に本手法を適用し，標準文書構造の抽出を試みた．図1は，そのうち広島市，府中町，呉市の文書と文書構造を示したものである．

4.1 標準文書構造の抽出

7文書から文書構造の論理積を取って共通する指示内容を抽出した結果を図2に示す．続いて，多数決に基づいて抽出した標準文書構造を図3に示す．ここで,「登録者の登録番号 (3文書)」と「登録番号 (6 − 3 = 3 文書)」は同数であるが，多段階の指示を優先し,「登録者の登録番号」として抽出している．

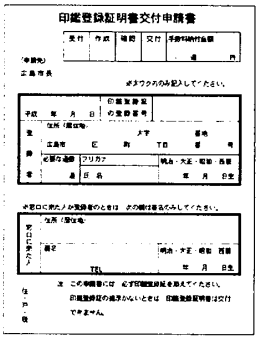
4.2 評価と考察

図3で抽出した標準文書構造と7自治体の文書間で計算した文書間距離を表1に示す．広島市と府中町は文書の見た目は異なっているという印象を受けるが，実際の文書構造では，0.17 という近い距離関係にあることが分かる．一方で，呉市とは，登録者の記入項目の差などから比較的大きな距離関係にあることも分かる．

標準文書構造の抽出を多数決に基づいて行うようにしたことにより，標準文書構造として,「登録者の申請枚数 (3文書)」という多段階の指示ではなく,「申請枚数 (7 − 3 = 4 文書)」という文書構造が抽出された．本手法で用いている文書構造抽出は，意味解析を伴わない形式的な処理によって実現されているが，より多くの文書で採用されている文書構造を抽出することに相当する多数決により,「申請枚数」という，より妥当であると考えられる文書構造を選択できる可能性が示唆された．

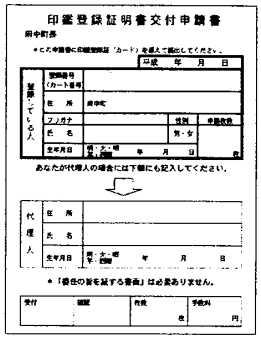
さらに，標準文書構造と7文書間の平均距離は，各文書における文書間の平均距離よりも小さくなっていることが分かる．これは，論理積と多数決による標準文書構造の抽出が，複数文書の重心的な役割を果たす文書構造

†広島市立大学 情報科学部 知能情報システム工学科
†$0 \le d(x,y) \le 1$，距離の公理を満たす．



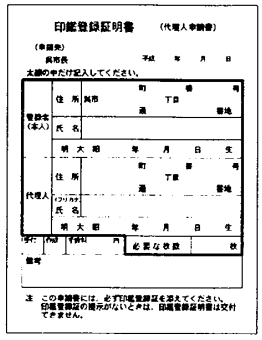
- 平成年月日
- 登録番号
- 登録者
  - 住所
  - 申請枚数
  - フリガナ
  - 氏名
  - 生年月日
- 代理人
  - 住所
  - 氏名
  - 生年月日

(a) 文書　(b) 文書構造



- 平成年月日
- 登録者
  - 登録番号
  - 住所
  - フリガナ
  - 氏名
  - 生年月日
  - 性別
  - 申請枚数
- 代理人
  - 住所
  - 氏名
  - 生年月日

(a) 文書　(b) 文書構造



- 登録者
  - 住所
  - 氏名
  - 生年月日
- 代理人
  - 住所
  - 氏名
  - 生年月日
- 申請枚数

(a) 文書　(b) 文書構造

図 1: 印鑑登録証明と文書構造 (左から広島市，府中町，呉市)

表 1: 文書間距離

	広島市	府中町	呉市	熊野町	東広島市	江田島市	三次市	平均
広島市	—	0.17	0.33	0.22	0.38	0.59	0.52	0.37
府中町	0.17	—	0.44	0.056	0.24	0.63	0.52	0.34
呉市	0.33	0.44	—	0.41	0.55	0.57	0.46	0.46
熊野町	0.22	0.056	0.41	—	0.29	0.62	0.50	0.35
東広島市	0.38	0.24	0.55	0.29	—	0.69	0.55	0.45
江田島市	0.59	0.63	0.57	0.62	0.69	—	0.68	0.63
三次市	0.52	0.52	0.46	0.50	0.55	0.68	—	0.54
標準	0.28	0.11	0.38	0.059	0.25	0.61	0.47	0.31

指示構造	文書数
住所	7
氏名	
申請枚数	
生年月日	
代理人. 住所	
代理人. 氏名	
登録番号	6
登録者. 住所	
登録者. 氏名	
登録者. 生年月日	
代理人. 生年月日	
フリガナ	4
性別	
登録者. フリガナ	
記入年月日	3
登録者. 申請枚数	
登録者. 性別	
登録者. 登録番号	

図 2: 共通指示内容

- 登録者
  - 登録番号
  - フリガナ
  - 氏名
  - 住所
  - 生年月日
  - 性別
- 代理人
  - 氏名
  - 住所
  - 生年月日
- 申請枚数

図 3: 標準文書構造

を抽出することになっていることを示すものであると考えられる．

5. おわりに

本論文では，同一目的の対話型罫線文書から標準文書構造を抽出する手法を提案し，評価尺度として文書間距離を定義した．文書間距離の利用により，標準文書構造が異なる文書における重心的な役割を果たす文書構造となっていることを示すとともに，文書の違いを定量的に評価できる可能性を示した．

謝辞　本研究の一部は，総務省 戦略的情報通信研究開発推進制度 (042308001) の援助を受けた．

参考文献

[1] 浅田尚紀，椋木雅之，青山正人，浅木森友彦: “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,” 信学技報, PRMU2005–111, 2005.

[2] 青山正人，椋木雅之，浅田尚紀: “書式構造文法を用いた対話型罫線文書の解析,” 信学技報, PRMU2005–112, 2005.