社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－CVIM－145 (13)
2004／9／10

PDLデータの解析による多様な形式の文書からの情報抽出方式の検討

平野 敬[1], 亀代 泰三[1], 岡田 康裕[1], 依田 文夫[1]

[1] 三菱電機株式会社 情報技術総合研究所

概要. ここでは多様なファイル形式の文書から、もれなく内容情報を抽出可能な文書解析方式を提案する。この文書解析方式は、文書を疑似的に印刷処理してプリンタが解釈可能な PDL データを作成し、この内容を解析する。この PDL データの解析処理では、PDL 内部にあるテキストデータを抽出し、イメージデータやベクトルフォントデータを文字認識処理する。これにより電子文書、画像、ＣＡＤ図面等の多様な文書から情報抽出が可能となる。ここでは本方式の概要と評価結果について述べる。

Information Extraction from Various Document Formats Based on PDL Analysis

Takashi Hirano[1], Taizo Kameshiro[1], Yasuhiro Okada[1], Fumio Yoda[1]

[1] Mitsubishi Electric Corporation, Information Technology R&D Center

Abstract. We propose a document analysis method which extracts text information form various document format files. In this method, a PDL (Page Description Language) data file is generated by doing dummy printing process of a document file. In the PDL data analysis, while extracting the text from inside of the PDL data, character recognition process for images is carried out. It allows text extraction without extraction loss from various document files, such as electronic document, image document, and CAD data. The design of this method is presented and experimental results are discussed.

1 はじめに

企業内に蓄積された多量の文書情報を検索・活用するナレッジマネージメントシステムの要求がある。特に、製造業では複数種類の CAD や電子文書、紙文書等、制限のない多様な文書が使用されており、これら多様・多様な文書フォーマットを統一的に扱う必要がある。このようなシステムを実現するには、文書を解析して内容情報を抽出する処理が必要である。

このような文書の内容情報を抽出する手段として、特定の電子文書に対応したフィルタを用意することで、その文書内の文字情報を抽出する方式がある。例えば、Microsoft 社の iFilter[1]を用いることで、Microsoft 社の製品である MS-WORD や EXCEL 等の文書から、高速・簡易にテキスト情報を得ることができる、ただし、このようなフィルタ方式は、一般に文書中のテキスト情報のみを抽出し、文書に埋め込まれたイメージ内の文字や、アウトライン化することで線分として描かれた文字は抽出できない。そのため、文書検索システムに適用した場合、文書上に見えている文字で検索しても、その文書が検索されずに検索もれとなる。また、適用可能な文書のファイル形式が限定され、容易に対象ファイル形式を増やすことができない課題がある。

本課題の対策として、ここでは多種・多様な文書フォーマットから、もれなく情報抽出できる文書解析方式を提案する。本方式では、まず全ての文書を擬似的に印刷処理して、プリンタが解釈できるＰＤＬ（Page Description Language）データに変換する。その後、ＰＤＬデータに含まれるテキストデータを抽出すると共に、イメージで表現された領域に対しては文字認識処理を行い記入された文字を抽出する。また、線分で表現された文字列に対してはイメージに変換した上で文字認識処理を行いテキスト化する。この文字認識処理

では、イメージの内容に応じてカラー・モノクロ画像、活字・手書き文字を自動判別して対象に応じた認識処理を選択実行する。これにより、印刷可能な多種・多様なファイル形式の文書に対応し、文書中の文字をもれなく抽出することが可能となる。また、文書を閲覧する際はＰＤＬデータから作成した各ページのＪＰＥＧ画像を表示するため、クライアントは文書ファイルに対応したアプリケーションを必要とせず、汎用のＷｅｂブラウザのみで閲覧できる特長を持つ。

2 提案する文書解析処理

提案する文書解析処理の処理フローを図１に示す。HTML, XML, TEXT ファイル等のテキストをベースとした文書については、簡単のため、HTML/XML パーサを利用した通常のテキスト抽出処理を利用して情報を抽出する。MS-WORD や EXCEL, AutoCAD 等のバイナリ形式文書に対して、PDL データの解析に基づいた文書解析方式を適用する。以降、各処理の詳細について詳しく説明する。

2.1 PDL データの作成処理

まず、MS-WORD や EXCEL 等、入力された文書ファイルに対応したアプリケーションを起動し、疑似的な印刷処理を行う。印刷処理を行うと、プリンタドライバが PDL データのファイルを作成する。PDL データの形式は PostScript と PDF が選択可能である（以下の実験では PDF を利用）。ここで、PDL データは印刷処理した際の情報であるため、プリンタに印刷処理を行わせるために必要な以下の情報を持つ。これらの情報は文書の頁単位にアクセスが可能である。

- テキストデータ
  文書中に含まれるテキストの文字コード、各文字の位置・大きさ情報
- 画像データ
  文書中に含まれる画像のバイナリデータ、画像の位置、大きさ情報
- 線分データ
  文書中に含まれる線分の長さと位置情報

2.2 PDL データからのテキスト抽出処理

PDL データの各頁から、内部に含まれるテキストの情報を直接に取り出す。これにより、iFilter と同じく、MS-WORD や EXCEL 中にテキストとして記述された情報を取得する。また印刷時の各文字の位置・サイズも同時に得る。ここで文字の位置・サイズは解像度 300dpi で印刷した場合の座標値で正規化を行う。

2.3 イメージエリアの検出処理

PDL データの各頁に対して、内部に含まれるイメージ群の情報を抽出し、この中から文字認識の対象とするイメージエリアを検出する。具体的には、まずイメージの幅・高さが閾値以下のイメージを除外する。次に複数イメージが重畳されている場合の対策として、重畳されている複数のイメージを、これらを内包した１つのイメージエリアにまとめる（図２）。同様に、一つのイメージが複数のイメージ群に分断されている場合に対応するため、隣接するイメージ群を１つのイメージエリアにまとめる。

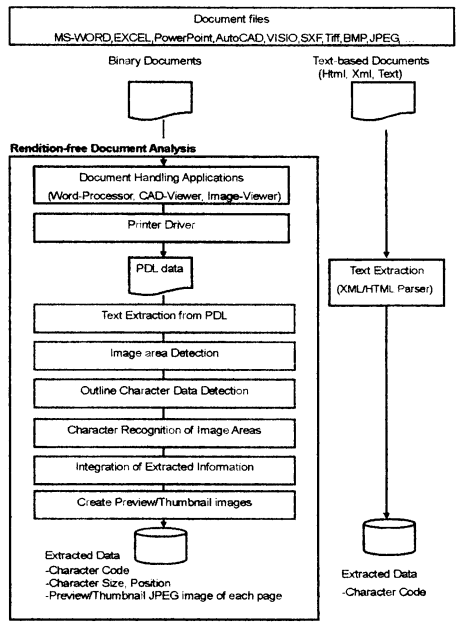

図１　提案方式の処理フロー

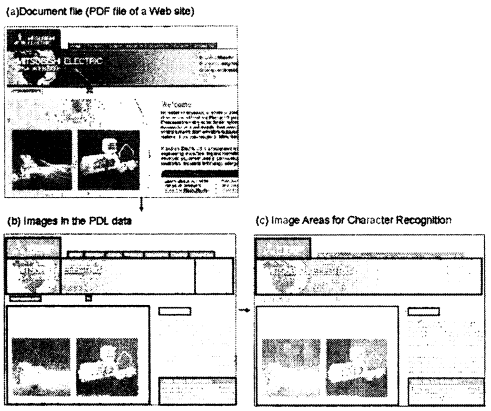

図2 PDLデータからのイメージエリア抽出例
図2(c)に示すように小さなイメージは無視され近接したイメージ群は一つのイメージエリアとして扱う。

2.4 アウトライン化された文字の抽出処理

AutoCADのDWGファイルや、Adobe illustrator等のDTPツールで作成されたPDFファイルでは、見た目はテキストだが、PDL内部では線分を用いたベクトル表現されているアウトライン化された文字がある（図3）。このような文書ファイルからは、テキスト抽出処理やイメージエリアの文字認識処理を行っても文字情報が抽出できない問題がある。ここでは、本問題に対処するため、テキスト抽出処理で文字情報が抽出できず、かつPDLデータ内に線分の情報が含まれる頁については、頁全体をイメージエリアとして文字認識処理を行う処理とした。

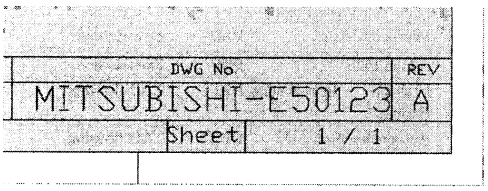

図3 DWGファイルの例
文字列"MITSUBISHI-E50123"はアウトライン化されており、線分で表現されている。

2.5 イメージエリア内の文字認識処理

ここでは、PDLデータをレンダリング処理することで、文書の各頁をBMP画像に変換する（解像度300dpi, Color 8bit depth）。このBMP画像に対して、上記の2.3節と2.4節で抽出したイメージエリア内の文字を認識する。この文字認識処理は、以下の手順でイメージの内部を解析する（図4）。解析の結果、イメージに記述された文字コードと、該文字の位置・サイズを得る。

文字認識処理

① カラー画像を2値化：文字部分を黒、背景部分を白とする2値化を行い、白黒のBMP画像を作成する。
② 文字列抽出：白黒のBMP画像から文字列を抽出する。
③ 活字・手書判別処理：各文字列内部の黒画素方向性の分布を入力として、判別分析法に従い文字列が活字か手書かを判別する。
④ 文字列内文字認識処理：活字文字列に対しては活字の文字認識処理を、手書き文字列については手書きの文字認識処理を適用し、文字コードを得る。

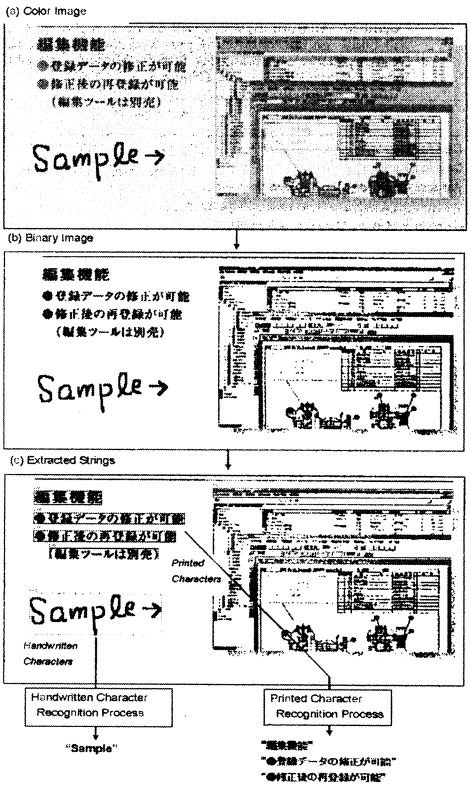

図4 イメージエリアの文字認識処理例

2.6 情報の統合処理

PDL データに含まれるテキスト情報は、飾り文字の部分で重複している場合がある。例えば影付きの文字は、２つのテキストの位置をずらして印字することで実現している場合がある（図 6）。この場合、同じテキストが二重に抽出される問題がある。また、テキストとイメージエリアが重なっている場合、同じ文字列をテキストとして抽出すると共に、文字認識処理も行われてしまい、ここでも同じテキストが二重に抽出される（図 7）。このような問題を抑制するため、ここでは、文字のコードが同じであり、位置と大きさが類似した文字情報を除外する処理を行う。

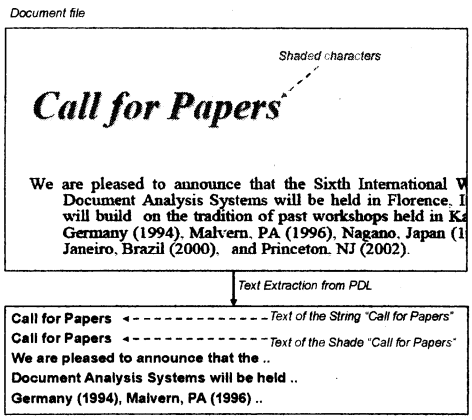

図６　飾り文字からのテキスト抽出

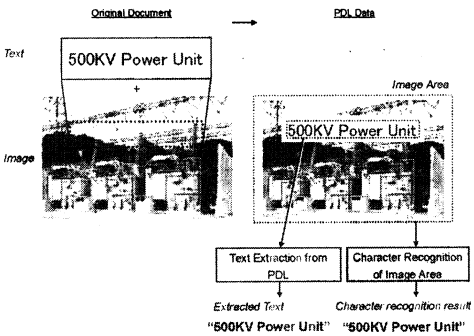

図７ テキストとイメージが重畳した箇所からのテキスト抽出

2.7 プレビュー/サムネイル画像作成処理

PDL データをレンダリング処理した結果から、各頁のプレビュー/サムネイル用の JPEG 画像を作成する。これらの JPEG 画像を用いることで、Web ブラウザベースの文書検索処理において、文書ファイルの閲覧に必要な CAD 等のアプリがインストールされていないクライアントからも、文書の内容を閲覧できる。

2.8 文書解析結果

以上の処理が完了すると、文書の各頁に対して、文字コードと、文字コードの位置・サイズ、およびプレビュー/サムネイル用の JPEG 画像を得る。

3 実験結果

本章では、実際の文書ファイルを用いた評価結果について記す。

3.1 実験用データ

表 1 に示す 16 種類の文書ファイルを評価データに用いる。これらの文書は社内・インターネットで集めた技術資料である。

表 1 実験に用いた文書データのリスト

Data No.	Document file name/Application	File extension	Type
1	Microsoft MS-WORD	DOC	Electric doc.
2	Microsoft EXCEL	XLS	Electric doc.
3	Microsoft PowerPoint	PPT	Electric doc.
4	Microsoft Visio	VSD	Electric doc.
5	Adobe PDF	PDF	Electric doc.
6	Drawing Interchange Format	DXF	CAD drawing
7	Autodesk AutoCAD	DWG	CAD drawing
8	Autodesk AutoCAD	DWF	CAD drawing
9	AutodeskInventor	IDW	CAD drawing
10	Tag Image File Format	TIFF	Paper doc.
11	BMP/DIB	BMP	Paper doc.
12	Fuji Xerox DocuWorks	XDW	Electric doc.
13	Scadec exchange format *1	P21 SFC	CAD drawing
14	Justsystem 一太郎	JTD	Electric doc.
15	Justsystem 花子	JHD	Electric doc.
16	Deneva Canvas8	CNV	Electric doc.

*1: CAD data exchange format specified by MLIT(Ministry of Land, Infrastructure and Transportation) in Japan

3.2 情報抽出結果

評価データに対する情報抽出処理を行った結果を表 2 に示す。表中の“TXT”電子的なテキストデータから抽出した情報、“IMG”が画像から抽出した情報、“OLN”はアウトライン化された文字から抽出した情報を示す。本結果から、多様な文書ファイルから内容情報が抽出できていることが分かる。ただし、画像から抽出した情報と、アウトライン化された文字から抽出した情報については、その情報抽出率が約 80%と低い。これは、電子ファイルに貼り付けられた画像の多くが、アプリケーションの操作説明用にウインドウのキャプチャ画面を貼り付けたカラーの縮小画像であり、文字認識が困難であることが原因である。また、CAD 図面においてレイアウトの複雑さにより文字列の抽出精度が低くなったことが、アウトライン化された文字からの情報抽出率を下げた。

表 2 情報抽出率

Data No.	Number of text lines in the document *Nc* (Correct Data)			Number of extracted text lines by this Method *Ne*			Information Extraction Rate *Ne/Nc* (%)		
	TXT	IMG	OLN	TXT	IMG	OLN	TXT	IMG	OLN
1	202	125	0	202	79	0	100	63	-
2	870	18	0	870	14	0	100	77	-
3	76	45	0	76	30	0	100	67	-
4	51	0	0	51	0	0	100	-	-
5	76	78	0	76	69	0	100	88	-
6	60	0	128	60	0	112	100	-	88
7	0	0	226	0	0	176	-	-	78
8	0	0	55	0	0	34	-	-	62
9	80	0	0	80	0	0	100	-	-
10	0	136	0	0	124	0	-	91	-
11	0	38	0	0	38	0	-	100	-
12	59	9	0	59	6	0	100	67	-
13	359	0	0	359	0	0	100	-	-
14	86	0	0	86	0	0	100	-	-
15	137	13	13	137	0	0	100	0	0
16	0	0	350	0	0	294	0	0	84
Total	2419	462	772	2419	360	616	100%	78%	80%

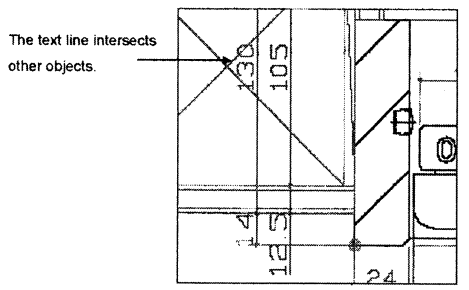

図８　情報抽出失敗例

3.3 文書検索システムへの適用

本情報抽出処理を、文書検索システム[2][3]における文書登録処理に適用した。この文書検索システムでは、文書を頁単位で検索処理し、検索でヒットした文字の位置情報を利用してプレビュー画像上に赤枠を描いて表示する。これにより、検索されたキーワードの位置を素早く見つけることができる（図 9）。本システムでは、ページ単位に検索を行い、検索でヒットした文書のプレビュー用 JPEG 画像を表示する。また、検索でヒットした検索キーワードを赤枠で囲んで表示する（図 10）。

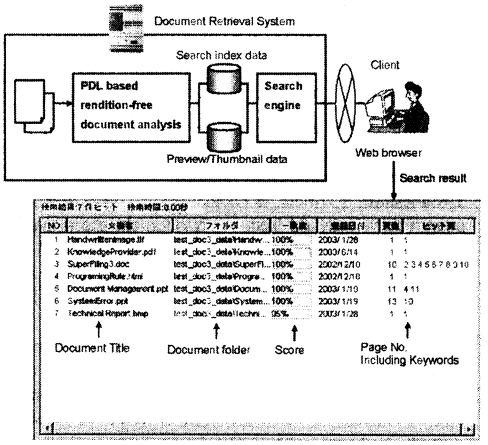

図９　文書検索システム構成.

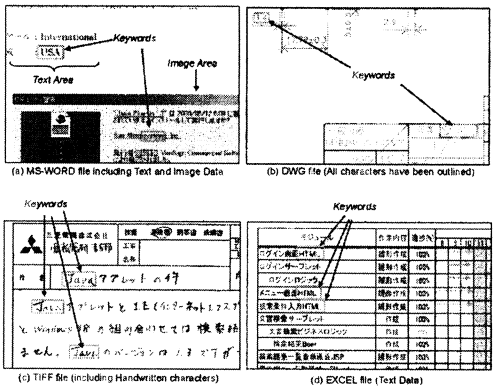

図１０　検索された文書の例

4 まとめ

本論文では、多種・多様なファイル形式の文書から、もれなく情報抽出可能な文書解析方式を提案した。16 種類の文書ファイルに適用し、その有効性を確認した。ただし、評価結果から電子ファイルに貼り付けた画像や、CAD 図面におけるアウトライン化した文字列からの情報抽出率が約80%と低い問題がある。今後は、本問題に対する対策について検討を行う予定である。また、帳票や文書画像に対するレイアウト解析技術[3][4]をベースとして、多様な文書ファイルから抽出した情報を構造化する技術についても検討を進めたい。

参考文献

[1] Microsoft MSDN site: http://msdn.microsoft.com/library/default.asp?url=/library/enus/indexsrv/html/ixrefint_9sfm.asp

[2] T.Kameshiro, T.Hirano, Y.Okada and F.Yoda: "A document retrieval method from handwritten characters based on OCR and character shape information," Proc of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01), pp.597-601, Seattle, USA, Sept. 2001

[3] T.Kameshiro, T.Hirano, Y.Okada and F.Yoda: "A document image retrieval method tolerating recognition and segmentation errors of OCR using shape-feature and multiple candidates," Proc of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99), pp.681-684, Bangalore, India, Sept. 1999

[4] R.Casey, D Ferguson, K. Mohiuddin, and E. Walach, "Intelligent Forms Processing System", Machine Vision and Applications, Vol.5, No.3, pp.143-155, (1992)

[5] T.Hirano, Y.Okada and F.Yoda: "Field Extraction Method from Existing Forms Transmitted by Facsimile", Proc if the Sixth International Conference on Document Analysis and Recognition (ICDAR'01). Pp.738-742, Seattle, USA, Sept.2001