レイアウト解析による書誌情報の抽出

永崎 健† 藤尾 正和† 高橋 寿一† 浦野 雄大† 池田 尚司† 福田 安宏‡
(株)日立製作所 中央研究所† (株)日立製作所 公共システム事業部‡

1. 研究の背景

電子文書・紙文書を問わず大量の文献を蓄積・保管し，必要に応じて検索・アクセスを可能とするシステムは，情報爆発の時代[1]における重要な社会インフラである．現在でも，様々な論文を WWW 上で検索できる CiNii や IEEE Xplore 等の学術文献検索システム，膨大な特許を検索できる IPDL 等が，広く利用されている．

文献検索の質を確保するためには，書誌メタ情報を適切に登録・管理することが重要である．一般に，文献検索には Bag of words モデルによる単語検索が広く使われている．しかし，短い単語は得てして数多くの文書にヒットするため，その結果，目的とする文書を探し出すのに時間と手間が掛かる．しかし，文書のタイトル，著者，本文，図，引用等の書誌メタ情報を利用して検索条件を定めれば，より適切に文書を絞り込むことができる．近年，機関リポジトリによる文献管理が提案されているが，書誌メタ情報の適切な登録と管理は重要課題の１つとされている．しかし，メタ情報の付与を人手で行うと，コストと処理時間の点が問題となる．

本稿では，PDF ファイルとして登録された学術文献から，書誌メタ情報を自動抽出する手法に関して報告する．また，PDF 文献のタイプ別による書誌メタ自動抽出の精度について評価実験を行ったので，これを報告する．

2. 書誌メタ解析の課題

書誌メタ解析に関する研究として，東野[2]は文書表紙のレイアウトを構造記述文法で定義し，これを解釈することでタイトルを抽出する手法を提案している．J.Beusekom[3]の手法は，一行目にタイトル，二行目にセンタリングで著者名等のパタンを予め登録しておき，入力文書とのマッチングにより書誌メタ情報を抽出する．Ishitani[4]は，文字列とレイアウト特徴を予めモデル化しておき，マッチングにより論文を XML 化する手法を報告している．また，高須[5]は，OCR 誤りを含む引用文献のテキストから書誌情報を復元する手法について報告している．

文献 1～4 については，いずれも対象文書ごとに細かい指定を必要とするため作業負荷が高く，拡張性に問題がある．また，文書メディアのタイプが書誌メタ解析に与える影響について十分な調査が無い．また，文献 5 は，何らかの手法で引用文献欄にあるテキスト行を抽出した後に，適用される手法を述べたものである．

3. 書誌メタ解析システムの構成

本稿で述べる書誌メタ解析システムは，次の４種類の基本書誌メタ情報を抽出する．

a) 論文のタイトル（日本語・英語）
b) 論文の著者　　（日本語・英語）
c) 論文の要約　　（日本語・英語）
d) 参考文献欄の各引用文献行

上記メタ情報は，文書先頭ページにある表紙系メタ情報（ａ～ｃ）と，文書後尾にある引用系メタ情報（ｄ）とに大別される．

文書の表紙に存在する情報（タイトルや著者等）はレイアウトに特徴があるため（図１），文字種・サイズ・配置等の情報を使って解析ができる．一方，参考文献欄は，開始位置が不定で，ブロック単位の差異が存在しないため，レイアウト情報の利用は限定される．そこで本研究では，表紙系メタの解析と，引用系メタの解析とでアルゴリズムを分けることとした．表紙系メタについては，文字行から得られる特徴を使って正準判別分析で著者・要約等に粗分類し，更に動的計画法（DP）によって順序の整合を取る．引用系メタについては，参考文献欄の中における文書の繰返し構造を解析する．

表紙系メタ解析アルゴリズムの概要を述べる．まず，入力された PDF から文書を構成する要素（テキストと座標）を抽出し，レイアウト再解析を行う．これは例（図２）にあるように，OCR-PDF に記録されたレイアウト情報に問題（行の過剰分割や順序乱れ）があるためである．文書レイアウト解析の手法としては XY 再帰投影法をベースに，PDF 向けへの改良を行った．これにより文字行を読み順に沿って一列に並べる．次に文字行毎に文字サイズ，字種などの特徴を抽出して，これを正準判別分析によって著者部，タイトル部，要約部かのカテゴリに粗分類する．

Document meta extraction system based on layout analysis,: Takeshi Nagasaki, Masakazu Fujio, Toshikazu Takahashi, Takehiro Urano, Hisashi Ikeda and Yasuhiro Fukuda
†Hitachi Central Research Laboratory, ‡ Hitachi Ltd.

更に大局的な順序整合性を取るために動的計画法を使って，隣接文字行間のカテゴリ分けに矛盾がないかを調べ，これを補正する．

引用系メタ解析アルゴリズムは，上記の再レイアウト解析後に行われる．まず，段組中の文字行を一列に展開し，辞書に登録された先頭ワード・終端ワードを元に参考文献欄の範囲を推定する．先頭ワードとは，例えば「引用文献」「文献」と言った語である．但し，OCRによって誤不読が生じるため，コンフュージョンマトリクス等を考慮して照合を行う．更に，確定した参考文献欄の範囲内において，文字行の配置，先頭の文字並びパタンなどを見て，繰返し構造のサイクルを判断する．

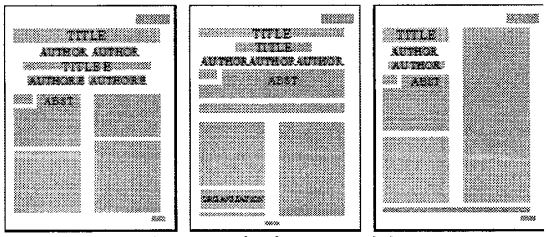

図１．書誌メタの例

４．評価実験

評価実験では３つの情報系学会（情報処理学会論文誌，電子情報通信学会論文誌，DEWS）のPDF文献を，学会毎に100～300ファイル選択して評価用サンプルとして用いた．これらPDF文献は２つのタイプに大別できる．紙文献をスキャンして，OCRでテキストを付与したファイル（OCR-PDF）と，WordやTeX等の電子文書から直接PDF文献を作成したファイル（Word-PDF）である．この２タイプについて，書誌メタ解析の精度を比較評価した．実験結果を表１に示す．

表１により，表紙系メタに関してはOCR-PDF，Word-PDFを問わず適合率が0.98～0.99と安定して高いことが分かる．一方，引用系メタについては，情報処理学会論文誌のOCR-PDFで精度が0.95へと低下した．情報処理学会から選択したサンプルには，他の学会と比べて古い文献も含まれており，そのことがOCR-PDFの品質に影響を与えている．個々のサンプルで抽出精度を調べると0.85～1.0の間で幅があった．これは，ノイズ等の影響でOCRテキストの品質にバラツキが生じ，文字行が密に詰まる傾向にある参考文献欄でのメタ解析に，特に精度に影響が出ているためである．低品質のものは，OCR-PDF上に埋め込まれた文字行情報（並び，配置，順序等）の誤りが大きく，現状のレイアウト再解析でも補正しきれていない．例えば，引用文献でのOCR-PDFの不具合例のうち，図２の左は再レイアウト解析により正しい書誌メタ抽出が可能だが，右は不可能である．

表１．ＰＤＦ文献からの書誌メタ抽出精度

タイプ	対象文献	文書表紙部	引用文献部
Word-PDF	論文誌Ａ	0.99	0.985
	論文誌Ｂ	0.975	0.978
OCR-PDF	論文誌Ａ	0.99	0.95
	論文誌Ｃ	0.98	0.97

※Ａ＝情処論，Ｂ＝DEWS，Ｃ＝信学論

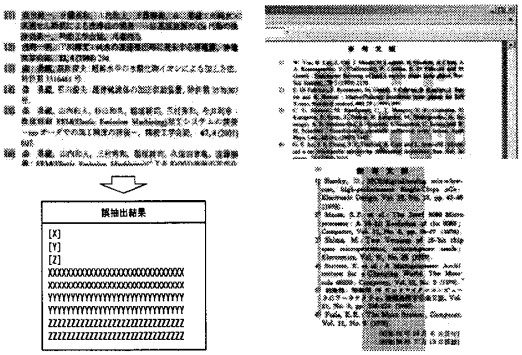

図２．PDF文書のレイアウト問題例

５．おわりに

PDF文献から基本書誌情報を自動抽出するためのレイアウト解析アルゴリズム（表紙系メタの正準判別－DP解析，引用系メタの繰返し構造解析）を開発した．PDF文書として，電子文書由来のWord-PDF，紙文書由来のOCR-PDFの２タイプを考慮して，実サンプルを使って評価実験を行った結果，双方で95%超の抽出精度を得た．

謝辞

本研究に関して議論を賜りました国立情報学研究所・安達教授，相澤教授ならびに研究所の皆様に感謝致します．

参考文献

[1] 喜連川, “情報爆発IT基盤によって人に夢を与えよう”, 人工知能学会誌, vol-21(5), pp.633-634.

[2] 東野, 他, “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”, 信技 PRU86-31, 1986

[3] J. Beusekom, etc, “Distance measures for layout-based document image retrieval”, DIAL2006, pp.232-242.

[4] Yasuto Ishitani, ”Document Transformation System from Papers to XML Data Based on Pivot XML Document Method”, ICDAR2003, pp.250-255

[5] 高須淳宏，他，“テキスト認識エラーモデルによる引用文献文字列からの書誌要素の抽出”,信学論 D-II, J87_D_II(6), pp.1298-1308