4E-1

文書画像理解における論理構造抽出の一方式

屋代寛　村上達也　嶋好博　藤沢浩道
㈱日立製作所　中央研究所

1．はじめに

文書情報を画像データとして大量に蓄積できる文書ファイリングシステムが実用化され、普及している。従来の文書ファイリングシステムにおける検索では、キーワードを用いた検索が主であるが、キーワードだけで文書情報を全て表現することは難しい。したがって、文書の内容からでも検索を可能にすることが必要であり、そのために文書の内容を読み取って自動的に理解し、検索可能なデータ形式に変換する文書理解技術の開発が望まれる。

2．背景

文書の書式を知識として表現する書式定義言語ＦＤＬ(Form Definition Language)を既に開発してきており、この言語を用いて文書の内容を汎用的に理解する基本方式を提案した[1, 2]。この文書理解方式は、文書ファイリングのためのインデックスとして書誌事項を自動的に作成することができるもので、文書の表題や著者名などの書誌事項の項目が抽出対象となっていた。文書の要素間のリンクを結ぶのに必要な文書の論理構造(文書の章・節などの構成)を抽出・理解することはできなかった。

3．対象文書

今回の研究で対象とした文書は文書構造が規則的なものとした。現在多くの学問分野、特に科学技術情報部門では情報の大部分が定期刊行物などの出版物によって伝達される。これらの出版物の内、割り付け構造（ページ上の物理的な配置）が規則的であり、論理構造も明確である学会誌や論文誌などを対象とした。これらの文書は情報の重要度が高く、また、利用価値が高いと考えられる。

4．文書構造抽出システム

文書構造抽出システムでは、まず、文書画像から連結領域(黒画素の塊)を抽出し、その連結領域の外接長方形を求め、これを処理の単位とする。次に、画素単位のパターン解析ではなく、文書画像を外接長方形の集合に置き換えた記号処理を行なう。行間や字間はこの外接長方形の距離で表され、知識ベースに格納されている。また、割り付け構造は長方形領域の包含関係として表現され、雑誌ごとの割り付け構造の変動に対処することができる。このシステムは図１に示すように、領域分離エンジンと書式定義言語解釈部から成る。このうち領域分離エンジンでは、文書画像を入力し、輪郭抽出等の画像処理を行なう画像処理部、テキストに対する文字認識部、長方形の領域に対するソート等の演算を行なう長方形集合演算部から成る。また、書式定義言語解釈部では知識ベースとして格納されている共通文書構造を解釈し、領域分離エンジンに対して演算を命令するとともに、抽出した領域および処理結果の良否を受け取る。

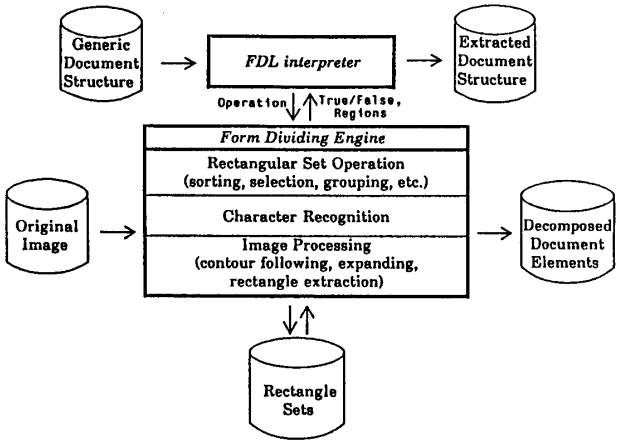

図１．文書構造抽出システム

A Method of Logical Document Structure Extraction
H.Yashiro, T.Murakami, Y.Shima, H.Fujisawa
Central Research Laboratory, Hitachi, Ltd.

知識ベースには、共通文書構造とレイアウト情報が格納されており、抽出した文書構造の領域と知識ベース内の文書構造とを関連付ける機構を有している。ここで、共通文書構造としては、検索に必要な要素とリンク情報が登録されている。また、レイアウト情報には文書構造を抽出するための特徴が記述されており、たとえば、章題ならば行間が本文領域より広い、段落は字下げがあるなどの情報が格納されている。また、文書の章題・節題には本文で用いているフォントと違うものが用いられており、このフォント情報を知識として利用する。これらの知識を用いて、文書の構造抽出を行なうことを可能としている。知識ベースに格納される文書構造では、各要素の反復や選択などにより、文書構造の共通部分を共通構造で扱えることになっている。反復は章・節など文書の内容によって数が異なる要素を表現する場合に用いられ、選択はページ内に図があるかどうかわからないときに用いられる。

5. 実験

本実験では雑誌「Hitachi Review」(1986年4月号Vol.,35,No.2)を対象とし、400dpi(16本/mm)の解像度で計算機に入力した。本実験では、複数ページにわたる文書を扱い、章題・節題の行間が本文領域の行間よりも広いというレイアウト情報を用いて章・節の抽出を行なった。さらに、字下げ情報を用いて章をパラグラフ単位に、参考文献リストから各参考文献を取だした。また、抽出した結果は木構造データに変換し、各ノードの名前は知識ベース中の共通文書構造の要素名称とその番号で示している（図２参照）。

6. 考察

実験では、文書構造を抽出するためのレイアウト知識を試行錯誤的に設定している。このレイアウトの知識を見出すにはノウハウが必要であり、また、対象文書によっては非常に複雑なものとなる。そのために、パンフレットや一般雑誌のように自由なレイアウトを持つ文書に対しては、本方式の適用は難しい。したがって、適用可能な文書を拡大することが今後の課題である。

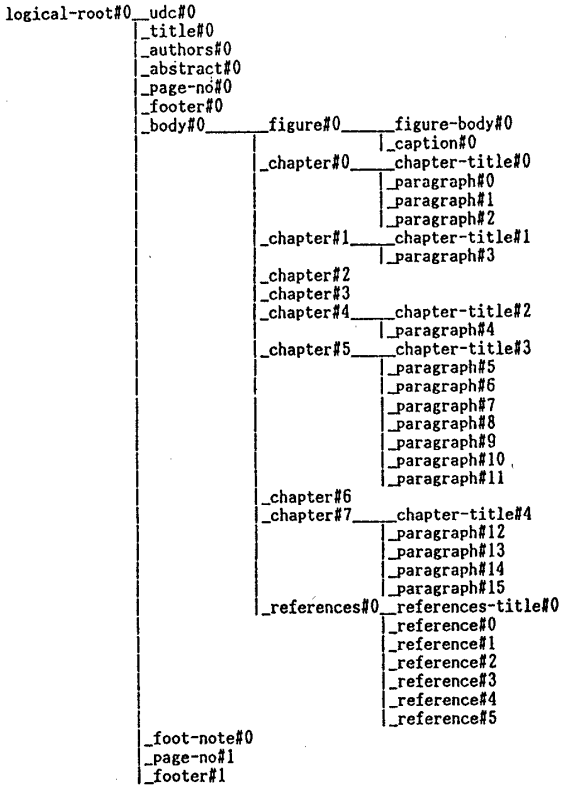

図２. 実験結果

なお、本研究は通商産業省工業技術院大型プロジェクトの一環としてＩＮＴＡＰ((財)情報処理相互運用技術協会)がＮＥＤＯからの委託を受けて、実施したものである。

参考文献

[1] J.Higashino et al., A Knowledge-based Segmentation Method for "Document Understanding,"Proc. 8th Int. Conf. Pattern Recognition,pp.745-748, 1986

[2] H.Fujisawa et al., "Document Analysis and Decomposition Method for Multimedia Contents Retrieval," Proc. 2nd ISIIS'88, pp.239-244, 1988