刊行物目次解析のための書誌情報表現の分析

1K－2

丸川勝美　嶋　好博　藤澤浩道
(株)日立製作所　中央研究所

1．まえがき

情報伝達の媒体として，「紙」は大きな役割を果たしており，紙の情報を管理し付加価値をつけることが必要であり，「紙」から電子化文書への媒体変換が行なえることが望まれる．特に，刊行物の目次には検索用データとなる書誌情報などが記述されており，書誌情報を自動的に抽出することで，検索可能な対象を容易に増やすことができる[1], [2].

目次画像から書誌データを抽出するには，様々な情報の中から必要な項目を識別するレイアウト理解の技術が必要である．特に，多種の刊行物を扱う場合，書誌情報の表現パタンに応じ，物理的な書誌情報の表現知識(モデル)が異なることはやむを得ないとしても，同一対象に対しては一つのモデルで対処する必要がある．

本報告は，目次画像から書誌情報を抽出することを目指し，様々な目次を収集し，書誌情報の表記パタンの分析を行い，その結果について述べる．

2．分析対象刊行物

書誌の表記パタンの分析を行うため，97種の雑誌を収集した．収集した刊行物は索引付けの点で重要なものを選択した．これには，科学技術関連の雑誌が94%と多くを占め，特に，大学などの紀要や試験所などの紀要や報告書が41%を占めた．

3．目次のレイアウト

目次には，書誌，書誌を整理するための「分類名」，その他色々な情報が含まれ，これらが様々な規則で表現され，レイアウトされる．例えば，書誌がシングルカラムやマルチカラムで表されたり，また書誌が複数種の異なる表記規則で表されたりするなどである．書誌がシングルカラムでレイアウトされる刊行物の割合は全体の69%であった．

本報告は，書誌がシングルカラムで記述された刊行物を対象とし，そこで用いられる書誌の表記パタンの分析結果を示す．

4．書誌と「分類名」との表記関係

書誌を分類するために，「分類名」が記載される場合がある．これは目次画像から書誌データを抽出する際のノイズとなる．「分類名」が存在する刊行物は54%で，「分類名」がセンタリングされたり，字下げされたり，5種類のパタンに分類できる．一方，「分類名」が存在しない刊行物は全体の46%を占める．

5．書誌の表現

5.1　書誌の構成要素

ここでは，書誌情報を必須な基本構成要素と刊行物に応じて利用される準構成要素に分ける．書誌の骨格を構成する基本構成要素は表題，著者名，開始頁番号，デリミタであり，準構成要素は副題や著者の所属などである．そして，書誌の構成は書誌の骨格である基本構成要素の構成パタンを調べることで検討できる．

5.2　基本構成要素による書誌の構成パタン

書誌を構成するパタンは9種類に分類できる．この分類は，画像の走査と同様な方法により，文字列の整列順として捉え行った．以下，表題をT，著者名をA，開始頁番号をPで表す．また，最後の数値はそのパタンが占める割合を示す．

(1) T - A - P．Tは左詰め，Pは右詰め．51%.

(2) A - T - P．Aは左詰め，Pは右詰め．19%.

(3) T┬P / A．Tは左詰め，Pは右詰め．3%.

(4) P - T - A．Pは左詰め．1種はAは右詰め．もう1種は全体で左詰め．3%.

(5) A - P - T．Pを軸に，Aは右詰め，Tは左詰め．1%.

(6) T - A' - P．Tは左詰め，Pは右詰め．(1)にて，前後のデリミタを軸に，Aを垂直下方向に記述(A')．10%.

(7) T - A'' - P．Tは左詰め，Pは右詰め．(1)にて，前後のデリミタを軸に，Aを垂直上下方向に記述

Logical Elements Representation on Table of Contents of Journals
Katsumi MARUKAWA, Yoshihiro SHIMA and Hiromichi FUJISAWA
Central Research Laboratory, Hitachi, Ltd.

(A''). 9%.

(8) A' - P - T. Pを軸に，A'は右詰め，T'は左詰め．(5)にて，後のデリミタを軸に，Aを垂直下方向に記述(A'). (5)にて，前のデリミタを軸に，Tを垂直下方向に記述(T'). 1%.

(9) T - P - A''. Pを軸に，Tは右詰め，A''は左詰め．前のデリミタを軸に，Aを垂直上下方向に記述(A''). 1種，1%.

以上のように，頻度の高いパタンは(1), (2)で，(1)の変形が(6), (7)であり，(1)と同一視できる．(1),(2),(6),(7)を併せたパタンは全体の90%を占める．

5．3　各基本構成要素の表現

(1)　表題

表題は単数もしくは複数行にわたって表記され，その記述は垂直方向に2次元的に変化する．そして，表題が最後に表記されるA - P - T, A' - P - T'を除き，表題の最右端の文字が書誌の右端になることはほとんどない．

(2)　著者名

著者名は垂直方向に2次元的に記述される．そして，その表現は3種類存在し，一つは1行で表記される場合．そして残り二つは複数行にわたって記述される場合で，個々の著者を1行ずつ記述するものと，複数の著者単位で複数行に記述するものである．さらに，これらは右揃いや左揃いでの表現がある．また，表題と同様，著者名が最後に表記されるP - T - A, T - P - A''を除き，著者名の最右端の文字が書誌の右端になることはほとんどない．

(3)　開始頁番号

開始頁番号は数字のみにより表記され，数字単体の場合とこれを括弧「( )」で囲む場合がある．

5．4　基本構成要素間の接続パタン

表題・著者名・開始頁番号の基本構成要素を繋げるデリミタは，実線，改行，点線，スペース，コロン，括弧(開始頁番号を表記する左括弧)，罫線の7種類である．接続に用いられるデリミタは，5．1の表記パタンにより，個々の特徴が現われる．例えば，(2)A - T - Pでは接続に用いられるデリミタはコロンと点線の組みである．

5．5　基本-準構成要素間の表記パタン

(1)　表題

連接する準構成要素は副題，項番，「分類名」，記号の4種類が存在する．副題が連接する割合は67%である．

(2)　著者名

連接する準構成要素は所属そして記号(「他」や「司会」など)，日時の3種類が存在する．連接する所属の表記は2種類あり，著者名の前に表記される場合と，著者名の後に表記される場合である．

準構成要素が連接する割合は16%である．準構成要素が連接するパタンはT - A - Pに対し82%であり，P - T - Aが9%，A - T - Pが9%である．

(3)　開始頁番号

連接する準構成要素は3種類存在し，同一巻を通しての開始頁番号，そして雑誌中の分野別毎の開始頁番号，各書誌の終了頁番号である．これらはいずれも数字のみもしくは「( )」で囲んで表記される．

準構成要素が連接する割合は6%である．

6．　二値化処理で対処可能な刊行物の割合

二値化処理で対処可能な刊行物は全体の81%で，不可能な刊行物が15%，不明な刊行物が4%である．

7．　あとがき

本報告は，目次画像から書誌情報を抽出するため，書誌の表記パタンの分析結果について述べた．今後，この結果に基づき，方式の開発を行う．

謝辞

本分析に御協力頂いた国立国会図書館殿に感謝致します．

参考文献

[1] 丸川，外，“文書イメージ情報の抽出/検索方法の一検討，” MIRU'96, pp. 283-287 (1996).

[2] K. Marukawa et al., “Document Retrieval Tolerating Character Recognition Errors,” Pattern Recognition, Vol .30, No.8, pp. 1361-1371 (1997).