構造記述を用いた書誌項目域の自動分割

3C-9

駱 琴, 渡邉 豊英, 吉田 雄二, 稲垣 康善, 柘植 利之
(名古屋大学工学部) (名古屋大学教養部)

1. はじめに

今日、図書館業務は図書目録カードを用いた書誌の管理、閲覧、貸出から、計算機を用いた図書館情報処理システムの下で、より多様な情報サービスが求められている。しかし、多様な情報サービスを実現するには、図書館情報処理システムの機能性に加え、膨大なデータを計算機処理可能な形態に変換し、書誌関連情報として構造化する方法の効率化が重要となっている。従来、図書目録カードは、専門家によって書籍から必要な情報が抽出されて作成されていた。しかし、図書目録カードに記載されている多くの情報は非専門家によっても抽出可能である。書誌情報を自動的に図書館情報処理システムに入力できれば、情報サービスの内容をより向上させることができる。

本稿では、書誌情報を書籍の奥付紙面から自動的に抽出して書誌項目ごとに分類することを研究目的として、書誌情報の項目域を自動的に分割する試みについて報告する。我々の方法は書誌項目域の構成を構造的に記述した知識に基づいて自動的に分割するアプローチである。奥付の様式は発行社、シリーズ書籍などによってそれぞれ異なっており、より汎用的な枠組みで処理する場合に構造情報は手続きと独立である必要がある。

2. 書誌情報の特徴

図書目録カードに記載された情報は必ずしも書誌情報だけではない。管理、整理などのための情報が付加されていて、図書目録カードの様式は日本全国の図書館で統一されていない。しかし、書誌情報として書籍から抽出、記載されなければならない項目は必ず決まっている。たとえば、名古屋大学附属図書館にて用いられている図書目録カードには、洋書で２１項目、和書で２５項目が設定されている。書誌情報に関しては洋書も和書も必要な項目は一緒であり、以下和書に関して検討する。和書に関する図書目録カードに記載されている項目を表１に整理した。

表１ 和書における図書目録カードの情報

分類情報		分類番号，著者記号（図書記号），所在場所
所蔵情報		所蔵部局
登録情報		トレーシング，登録番号，購入日，書店
書誌内容情報	書誌情報	表題，書名，副書名，巻次と年次，著者等，訳者等，版表示，出版地，出版社，出版年，叢書名，定価
	書誌付帯情報	ページ数，冊数，図版表示，大きさ，注記

表１で、書誌内容情報が書籍から抽出された書誌項目であり、このうち１２項目の書誌情報が通常書籍紙面に明記されている。従って、適切な書籍紙面を選択して、その紙面を対象に処理すれば、かなりの書誌データの作成を省くことが可能である。図書目録カードの情報に対して、書誌情報のデータ量は６～７割を占めると思われる。和書の場合、書誌情報はほとんど奥付紙面に集中している。たとえば、図１に示した奥付には８項目が含まれている。

3. 書誌項目域抽出のための構造記述

書誌情報を自動的に抽出するためには、書誌情報が書籍にどのように配置されているかを明確にする必要がある。概略、和書の場合奥付紙面に書誌情報が配置されているが、書誌項目の構成は発行社、各書籍によって少なからず異なっている。従って、一般に定形的に処理することができず、処理対象を同定する手続きが必要になる。すなわち、処理手続きは一定ではなく、知的に実行する必要があり、処理に必要な決定情報は手続きとは独立に定めた方が柔軟性、適用性、拡張性に富む。我々は書誌項目域を同定・抽出するために構造情報を外部から与えて処理する方法を採用した。

書誌項目域の同定・抽出のための構造記述には、次の機能が必要である。

（１）領域の指示は絶対的な大きさ・位置の値ではなく、相対的な値でなければならない。

（２）領域間の関係を構造的に、かつ唯一的に指示できる。

（３）記述内容が単純でなければならない。

（４）少々の相違を吸収できるように、できる限り柔軟に記述できる必要がある。

このような前提条件の下に、階層的に領域の上下左右の関係情報だけで定義する方法を設定した。構造記述は木構造の表現で構成され、中間ノードが各ステップで区別される領域であり、葉ノードが最終的に分割できる項目域、または項目群域である。ノードは終端ノードと非終端ノード４種であり、表２に整理した。各ノードの項目領域の関係は図２のようである。

表２

記号	ノード種
T	終端ノード
H	水平分割ノード
V	垂直分割ノード
OR	選択ノート
RP	垂直方向の重複ノード

非終端ノードはその子ノードとして、終端ノードまたは非終端ノードを連結する。水平分割ノードは子ノードの領域が全て水平方向に隣合って細分割されることを表し、垂直分割ノードは子ノードの領域が全て垂直方向に隣合って細分割されることを表す。各ノードは同定・分割プロセスに対して並びに従った順序関係を保持し、水平分割ノードの場合は左から右、垂直分割ノードの場合は上から下に順序付けられている。選択ノードは複数の可変な領域関係を表現するノードであり、局所的な構造の相違を指示す

Recognition of Library Catalogging Textures by a Structure Description Method
Qin LUO[1], Toyohide WATANABE[1], Yuuji YOSHIDA[1], Yasuyoshi INAGAKI[1] and Toshiyuki TSUGE[2]
1:Faculty of Engineering, Nagoya University, 2:College of General Education, Nagoya University

るために有効である。従って、通常選択ノード以外は連接ノードである。重複ノードは垂直に繰り返し出現する項目領域を指示する。たとえば、共著の書籍の著者項目の指示に有用である。

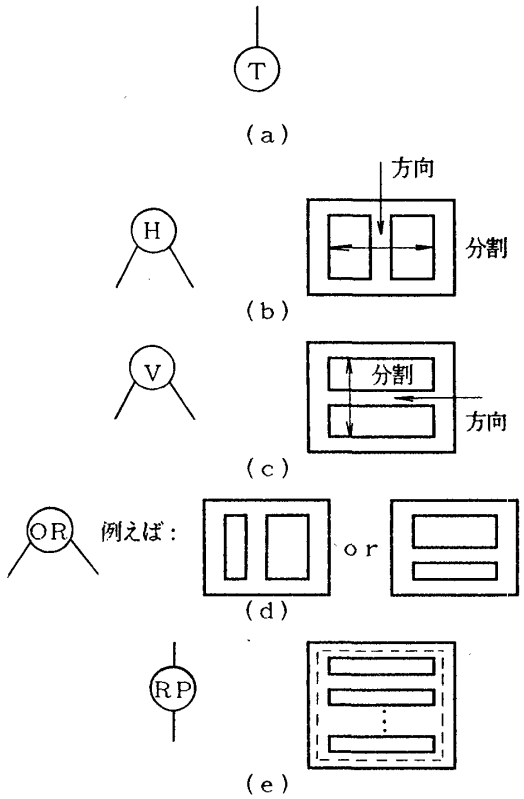


図2　各ノード項目領域の関係

各非終端ノードは4組（MOD，SNUM，OP，CO）から構成される。フィールドMODはノードの種類を表し、ノードに対する各項目領域が水平・垂直の関係を保持している。フィールドSNUMは連結されたノードの個数を指示する。フィールドOPに領域分割するための操作情報が保存されることを表す。すなわち、さらに分割する項目領域の分割操作がその領域間を区別する区切り記号列データに従って定義される。フィールドCOはこの分割操作で求めた領域を記憶するための情報を保持し、次の細項目分割に対して情報を与える。

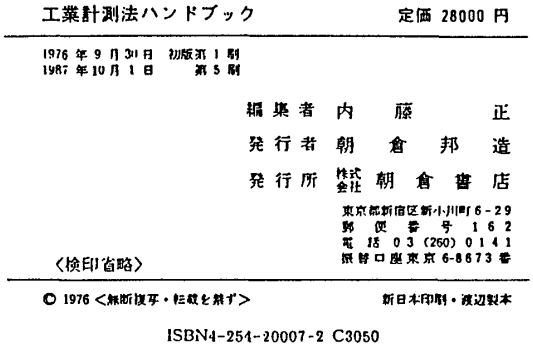
終端ノードで表された項目領域にはまだ複数の書誌項目が含まれる可能性がある。しかし、画像処理ではこれ以上何らかの知識（キーワードなどの書誌情報特有の知識）を前提としなければ不可能である。従って、終端ノードはさらに文字認識などの処理により細分割されて特定の書誌項目が抽出される。

このように、項目領域の分割手続きは、構造記述によって領域関係を認識して繰り返し終端ノードに到るまで適用される。この場合、事前に得た情報を保持して、細分割が繰り返される。従って、この分割処理は全く処理対象のデータを仮定せずに、適時内部で情報を発生させながら、木構造に沿って繰り返し働く。

4．構造記述の例

前節に述べた構造記述の具体例を示す。書誌情報が含まれている奥付紙面に対して検討する。たとえば、朝倉書店の最近二年間に発行された書籍は概略図1の書式となっている。この書誌情報が記されている領域に、構造記述を適用すると、概略図3のように階層構造として表すことができる。丸印は非終端ノードに、四角印は終端ノードに対応している。

この構造記述を用いて領域分割した結果が図4である。この場合、領域分割のために、直線、行間スペース、空白及び相対の位置関係などに基づいて定義した分割操作OPを6個用いた。ほぼ完璧に領域が分割されている。


工業計測法ハンドブック　　定価 28000 円

1976 年 9 月 30 日　初版第 1 刷
1987 年 10 月 1 日　　第 5 刷

編集者　内藤　正
発行者　朝倉邦造
発行所　株式会社 朝倉書店
東京都新宿区新小川町 6-29
郵便番号 162
電話 03 (260) 0141
振替口座東京 6-8673 番

〈検印省略〉

© 1976 〈無断複写・転載を禁ず〉　　新日本印刷・渡辺製本

ISBN4-254-20007-2 C3050

図1　奥付の例

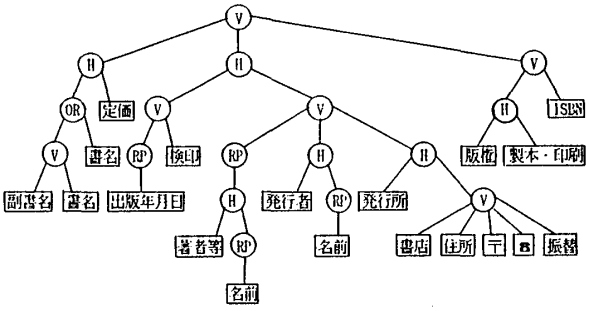


図3　構造記述の例

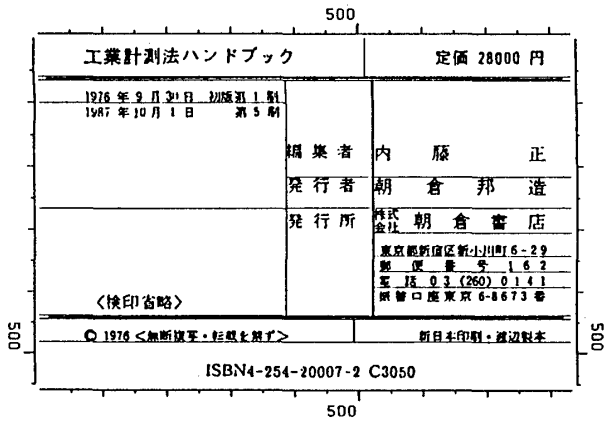


図4　構造記述を用いた領域分割の例

5．おわりに

本稿では、書誌情報の自動抽出を研究目的として、書誌紙面の奥付からの書誌項目域の分割法を報告した。多種多様な書式に対して、それぞれの特性できるだけ依存しない方法として、我々は構造記述法を用いたアプローチを提案し、その適用例を示した。しかし、今後より多くの処理対象に対して、より記述力がある表現法を開発する必要がある。また、本構造記述法は領域分割だけでなく、書誌項目の同定、書誌情報の抽出に対しても有効に働くように、多元知識の同時表現についても検討しなければならない。