図書目録カードの認識・理解システム（II）

1K-3

長谷博行，米田政明，酒井充，松田充弘
（富山大学工学部）

1．はじめに

前稿[1]でシステムの概要を述べたが、本稿では項目クラスで表現されている項目規則について説明を加える。更に、システムの一部を用いて行った実験結果について述べる。

2．項目規則

洋書用図書カード認識で用いたプログラムは全てＦＯＲＴＲＡＮで記述したため規則もプログラムとして表わした。本システムの考え方はオブジェクト指向であり、図書カードに記載される各項目もシステムではクラスとして独立に機能し、文字認識された全文字列から自クラスに適合した部分ストリングを抽出する。各項目規則は項目フレーム内に宣言的に記述される。規則の解釈・実行は全項目規則に共通のインタプリタで行なう。そのため規則記述のための簡単な言語を作成した。そのシンタックスを図１に示す。図中の○印は終端記号、□印は非終端記号である。規則は「'」と「'」ではさまれた形をしており、Ｐｒｏｌｏｇにとってはストリングとなる。図でDBは都市名データベースを参照することを意味している。また、strの上から２つはキーワードであり、著者項目では「著」等、版表示項目では「改訂版」等、ページ項目では「ｐ」等がキーワードとなる。２バイトキーワード（日本語）には始めと終わりにそれぞれSIコードとSOコードが挿入されている。また、巻次では「第２巻」とか「第２集」のように部分的な変化がある場合がある。このような変化には「第２(巻｜集)」のように簡潔に記述することができる。更に、特徴的な項目では「第３巻」、「４訂版」、「～出版社」等のようにキーワードの前や間に文字列が書かれる場合がある。このような文字列表現は、図では３番目のstrで表現されていて、例えば数字３文字ならば「N3」、３文字以上５文字以下の日本語を含む文字列ならば「K(3,5)」と記述する。Ａはアルファベット・数字・記号を意味し、Ｘは文字種を限定したいときに有効である。例えば、ページ項目の前ページはローマ数字で書かれるため「X(V,X,I)(1,4)」と表現できる。なお、Ｂは２次矩形単位にカウントすることを表し、＊はワイルドカードで文字列の長さは任意である。Ｈ、Ｅ、Ｔはフラグでありそれぞれ、行の先頭から部分ストリングが始まること、部分ストリングの終端がストリングの終端であること、先頭行が部分ストリングであることを意味する。また、ひとつの項目の条件がＯＲ条件で複数ある場合、規則を併記するだけでよい。記述例を図２に示すが、分類コード項目の条件は「'HN3'」と「'HN3".'N(1,4)'」が併記される。このように、規則はおもにキーワードなど表面的な知識に基づいており、項目によってはかなり不完全な知識しかないものもある。例えば、書名、副書名などはそれだけを見たのでは我々でも同定できないことがある。図２では書名は'HKB1'と記述してあり、これは行の先頭から始まる日本語を含む文字列で１つの２次矩形を表している。このような項目では一般に多くの候補部分ストリングが生成され誤りの原因になりやすい。これは、上位クラスで項目間の整合をとることである程度解決されるが、システムに判断する手だてがなくなるとオペレータの判断に頼ることになる。また、得られた分類パターンの評価scoreを(1)式で与えるが、規則のＡＮＤ条件数R_iは(2)式で定義する。(1)式でＰは分類パターンに含まれる項目の集合である。

$$score = \sum_{i \in P} (c + R_i), \quad 但し、c=10とした \qquad (1)$$

R_i＝キーワードを構成する文字数＋文字列表現数　(2)

例えば、巻次項目の規則'第N(1,2)(巻｜分冊｜集)'をストリングに適用して「第３巻」の文字列が抽出されるとR_iは３であり、「第３分冊」の文字列が抽出されるとR_iは４である。このR_iは前もって与える必要はなく部分ストリング抽出インタプリタで計算される。

次に動作について述べる。項目フレームにインスタンス要求があるとまずその項目フレームのプロダクションルールが働き、もし既に同じ環境下で作られたインスタンスが得られていればそれを返すが、そうでなければその項目の条件を部分ストリング抽出インタプリタへ送り部分ストリング抽出を行いインスタンスを生成する。部分ストリング抽出インタプリタはＦＯＲＴＲＡＮ上で作成する。これは手続きの一種であるため、方法・手段は表面上は見える必要がないと思われるからである。

このようにＰｒｏｌｏｇで規則のみを記述することによって見通しのよい表現が得られる。

3．実　験

本システムはまだ未完成なので分類パターン生成の部分のみで行った実験について述べる。故に、前稿[1]図１の図書カードクラス、書誌内容領域クラスとその下位クラスの１４クラスを用いて行う。文字認識と領域仮説は既に行ったものとしている。実験結果を図３、４に示す。図３は洋書用カードに対する結果であり、図２における洋書用規則を用いている。この場合は規則に区切り記号法を多用しているため７４個もの分類パターンが生じたが、正しい分類結果は第２位の評価点であった。結果の中で番号の抜けている項目は省略されているものとシステムは解釈している。なお、図中'¥'記号で改行があったことを表している。図４は和漢書用カードに対する結果であり、図２の和漢書用規則を用いている。図４では第１位と第２位の結果がそれぞれ(b)、(c)に示してある。両者の違いは'「工業英語」¥編集部編'の文字列の解釈にあり、途中に改行があったために２通りの解釈が生じた。前者は'編'をキーワードとして捉え、後者は'編集部編'をキーワードとして捉えている。この場合は後者が正解であるが、「編集部」の意味がわからない限り前者も妥当な結果である。これはより高度な判断が必要とされる例である。

4．まとめ

本稿では目録規則[2]から項目規則を作成し、それを使って部分ストリング抽出を行った。更に、それらからつじつまの合う組合せ（分類パターン）を選び考察したところ、正しい結果は評価点も良いことが示された。しかし、図４の結果からキーワード等の表面的な知識の限界も示された。このことはシステムをより高度化にするための鍵になるものと思われる。

謝辞　日頃ご指導頂いています東北大学工学部情報工学科木村正行教授に感謝致します。

[文献] (1)長谷他:"図書目録カードの認識・理解ｼｽﾃﾑ(I)",情処38全大(1989).(2)"日本目録規則",日本図書館協会

An Understanding System for Book Cards(II)
Hiroyuki HASE,Masaaki YONEDA,Mitsuru SAKAI,Mitsuhiro MATSUDA / TOYAMA University

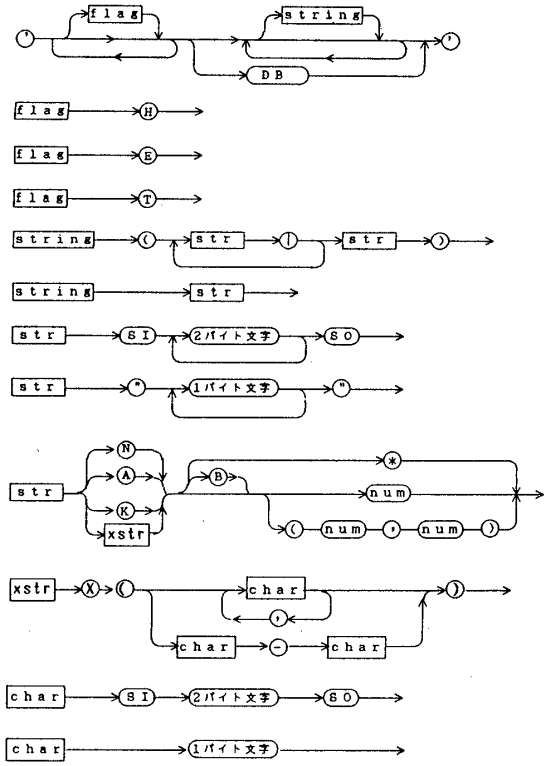

図１　項目規則のシンタックス

```
① 分類コード  : ' HN3'                    ' HN3'
                ' HN3"."N(1,4)'           ' HN3"."N(1,4)'
② 著者コード  : ' HAB1'                   ' HAB1'
③ 図書記号    : ' HAB1'                   ' HAB1'
④ 受け入れ番号: ' HN(1,6)'                ' HN(1,6)'
                ' HN(1,6)"-"N(1,2)'       ' HN(1,6)"-"N(1,2)'
⑤ 標　　目    : ' HTA*'                   ' HTAB1","A*","'
⑥ 書　　名    : ' HKB1'                   ' HA*("."|":"|";"|"/")'
⑦ 副 書 名    : ' KB1'                    ' ":"A*(";"|"/")'
⑧ 巻　　次    : ' N4年度'                 ' ("VOL."|"NO.")N(1,2)'
  ' 第N(1,2)(巻|分冊|集)'                 ' X(0-9)(1,2)'
  ' 昭和N(1,2)年度',' N(1,2)'
⑨ 著　　者    : ' K(2,6)(編著|           ' "/"A*","'
  著|作|編|訳|文|画|絵|
  撮影|作曲|編集部編)'
  ' K(2,6)(等|ほか)共著'
  ' K(2,6)","K(2,6)(著|共著)'
⑩ 版 表 示    : ' 第N(2,6)版'             ' "."A*("/"|".")'
  ' K(1,4)版',' 増補'
⑪ 著者(版)    : ' K(2,6)編著'             ' "/"A*","'
⑫ 出 版 地    : ' HDB'                    ' HDB'
⑬ 出 版 者    : ' KB1(社|出版             ' ":"A*","'
  |出版会|書店|出版センター
  |局|書房|書院|協会|研究所
  |学会|所|堂)'
  '(丸善|有斐閣|化学同人|
  出版者不明)',' KB1'
⑭ 出 版 年    : ' N4'                     ' "C"N4","'
  ' N4"."n(1,2)',' 昭和N(1,2)'            ' N4","'
  ' "c"N4","'
  ' N4(印刷|序|あとがき)'
⑮ ペ ー ジ    : ' HN(1,3)"p"'             ' X(I,V,X)(1,4)","'
  ' HN(1,3)","N(1,3)"p"'                  ' N(2,3)"P.","'
  ' HN(1,3)(冊|丁|枚|軸)'
  ' H"p"N(1,3)"～"N(1,3)',
  ' H図版KB1',' H地図KB1'
⑯ 挿　　図    : なし                      ' ";"A*";"'
⑰ 大 き さ    : ' N2"cm"',                ' ";"N2"CM.","'
  ' N2"×"N2"cm"'                         ' ";"N2"×"N2"CM.","'
⑱ 叢 書 名    : ' "("K*")"'               ' "("A*")"'
⑲ 注　　記    : ' HE'                     ' HE'
```

図２　項目規則記述例（下線部は洋書用規則）

```
903
M31     Manguel, Alberto.
          The dictionary of imaginary places /
        Alberto Manguel & Gianni Guadalupi ;
        ill. by Graham Greenfield ; maps and
        charts by James Cook.  New York :
        Macmillan, c1980.
          438 p. : ill. ; 31 cm.

109759
```

(a)対象図書カード

```
SCORE= 124
 5  HEADING            : MANGUEL, ALBERTO.
 6  TITLE              : THE DICTIONARY OF IMAGINARY PL
                         ACES /
 9  AUTHOR             : /¥ALBERTO MANGUEL & GIANNI GUA
                         DALUPI ;¥ILL. BY GRAHAM GREENF
                         IELD ; MAPS AND¥CHARTS BY JAME
                         S COOK.
12  PLACE OF PUBLICATION : NEW YORK
13  PUBLISHER          : :¥MACMILLAN,
14  DATE OF PUBLICATION : C1980.
15  PAGE               : 438 P.
16  ILLUSTRATION       : : ILL. ;
17  SIZE               : ; 31CM.
```

(b)項目分類結果

図３　洋書用カードに対する結果

```
                JIS ni motozuku eiwa waei gijutsu yogo jiten
                                                        1/2
503.4
K82
1-5     JISに基づく英和・和英技術用語辞典 「工業英語」
        編集部編
        東京  インターブレス  1981.10
        5冊  26cm (工業英語別冊技術用語シリーズ  1～5)
        1  電気・電子・制御・基本
        2  機械・油空圧・溶接・設計
        3  自動車・航空機・船舶・エネルギー
229754-58                          (つぎのカードにつづく)
```

(a)対象図書カード

```
SCORE= 131
 5  HEADING            : JIS NI MOTOZUKU EIWA WAEI GIJU
                         TSU YOGO JITEN
 6  TITLE              : JISに基づく英和・和英技術用語辞典
 7  SUB-TITLE          : 「工業英語」
 9  AUTHOR             : 編集部編
12  PLACE OF PUBLICATION : 東京
13  PUBLISHER          : インターブレス
14  DATE OF PUBLICATION : 1981.10
15  PAGE               : 5冊
17  SIZE               : 26CM
18  SERIES             : (工業英語別冊技術用語シリーズ  1～5)
19  NOTES              : 1 電気・電子・制御・基本¥2 機械・油空圧・溶接・設計¥3
                         自動車・航空機・船舶・エネルギー
```

(b)第１位の結果

```
SCORE= 123
 5  HEADING            : JIS NI MOTOZUKU EIWA WAEI GIJU
                         TSU YOGO JITEN
 6  TITLE              : JISに基づく英和・和英技術用語辞典
 9  AUTHOR             : 「工業英語」¥編集部編
12  PLACE OF PUBLICATION : 東京
13  PUBLISHER          : インターブレス
14  DATE OF PUBLICATION : 1981.10
15  PAGE               : 5冊
17  SIZE               : 26CM
18  SERIES             : (工業英語別冊技術用語シリーズ  1～5)
19  NOTES              : 1 電気・電子・制御・基本¥2 機械・油空圧・溶接・設計¥3
                         自動車・航空機・船舶・エネルギー
```

(c)第２位の結果

図４　和漢書用カードに対する結果